

Real-Effort Tasks in Experiments: The Task Choice Matters

Christian Waloszek

Dissertation ETH No. 27144

PEC Dissertation Series, No. 1, 2021

© Copyright 2021

Christian Waloszek

DISS. ETH NO. 27144

Real-Effort Tasks in Experiments: The Task Choice Matters

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Christian Waloszek

Diplom-Physiker, Freie Universität Berlin

born on 26.08.1985

citizen of Germany

accepted on the recommendation of

-

Prof. Dr. Marko Köthenbürger

Prof. Dr. Petra Schmid

2021

Table of Contents

	Page
List of Figures	ii
List of Tables	iii
Abstract	1
Introduction to the Thesis and Dissertation Overview	3
Overview of the Research Agenda	8
Chapter 1: Approaches to Measuring Effort and Classifying Real-Effort Tasks	12
1.1 Approaches in the Study of Effort Provision: Chosen Effort vs. Real Effort	13
1.2 A Classification Based on the Realism of the Task	16
1.3 A Classification Based on the Required Subject Qualities: Skills and Personality Traits .	20
1.4 Other Classifications of Real-Effort Tasks	22
1.5 Choose a Task That Suits Your Research Question	25
Chapter 2: Developing Real-Effort Tasks	31
2.1 Introduction	31
2.2 Shortcomings of Real-Effort Tasks	33
2.3 Why Do Subjects Make an Effort to Complete Tasks? – A View From Motivational Psychology	37
2.4 Criteria and Practices for Designing Real-Effort Tasks	58
2.5 Applying the Criteria and Practices: The Novel Single-Slider Task	91
2.6 Conclusions	93
Chapter 3: Comparing Real-Effort Tasks	97
3.1 Introduction	97
3.2 The Real-Effort Task Survey	103

3.3	Research Design and Methodology	113
3.4	Experimental Results	125
3.5	Conclusions	142
Chapter 4:	Determinants of Real-Effort Task Performance	147
4.1	Introduction	147
4.2	Research Design and Methodology	152
4.3	Experimental Results	175
4.4	Conclusions	191
Chapter 5:	Conclusion and Discussion	196
Appendix A:	Appendix to Chapter 2	199
A.1	The Single-Slider Task: Instructions in English and German	199
Appendix B:	Appendix to Chapter 3	202
B.1	Details on Experimental Design and Procedure	202
B.2	Real-Effort Task Survey: Survey Structure and German Version	262
B.3	Additional Figures and Tables	265
B.4	Supporting Documents	303
Appendix C:	Appendix to Chapter 4	310
C.1	Experimental Design	310
C.2	Additional Figures and Tables	317
Colophon		343
	Reproducibility of Results: Information on R Environment and Packages	344
	The Single-Button Task	346
References		347

List of Figures

Figure Number		Page
1	Tasks in real-effort experiments	4
2	Applications of real-effort tasks	7
1.1	Approaches of effort provision	13
1.2	Mundane realism	17
1.3	Realism and usefulness of output in real-effort tasks	18
1.4	Saliency of the usefulness of the output and the degree of realism of a task	21
1.5	Specific and generic real-effort tasks	27
2.1	Purpose- and activity-related incentives in the extended version of Heckhausen's Advanced Cognitive Motivation Model	39
2.2	Propositional logical version of Heckhausen's Advanced Cognitive Motivation Model	45
2.3	Purpose-related and activity-related-incentives	47
2.4	Incentive residing in the activity for exemplary tasks	50
2.5	Skill-demand balance and the flow channel model	54
2.6	Sequence of action phases based on Heckhausen's Rubicon Model	56
2.7	Design criteria and design practices for fabricating and implementing real-effort tasks	60
2.8	The wire-loop game	65
2.9	Likelihood of entering the flow state in relation to task difficulty and skill level	67
2.10	The single-slider task	92
3.1	Development of the real-effort task survey	104
3.2	Countermeasures to non-monetary incentives	106
3.3	Survey items and the design criteria and practices on which they are based	108
3.4	Experimental procedure	118
3.5	Task selection, task grouping, and planned orthogonal contrasts	121
3.6	Plot of regression coefficients for all motivational survey items	129
3.7	Plot of regression coefficients for all effort-related survey items	132
3.8	Differentiate tasks along three dimensions	133
3.9	Personal-hit list and real-effort task survey as summated rating scale	140

4.1	Determinants and course of motivated action	149
4.2	Schematic representation of the experimental procedure	153
4.3	Selection of seven real-effort tasks employed in the experiment.	155
4.4	Diagnosis scheme to capture forms of motivations and motivation problems	164
4.5	Purpose- and activity-related incentives in the extended version of Heckhausen's Ad- vanced Cognitive Motivation Model with links to Rheinberg's diagnosis scheme	165
4.6	Score distributions of the subjects for the selection of tasks	175
4.7	Correlation between subjects' performance across tasks	177
4.8	Superimposed scores conditional on subjects' qualities	178
4.9	Superimposed scores conditional on subjects' motivations	179
4.10	Superimposed scores conditional on subjective performance assessment	189
4.11	Score distributions for the math task conditional on subjects' self-assessed numeracy skills and performance satisfaction	190
A.1	Single-slider task: Task preparations – checking subjects' left- or right-handedness	199
A.2	Single-slider task: Instructions	200
A.3	Single-slider task: Task page	200
A.4	Single-slider task: Results page	201
B.1	Experimental procedure (detailed)	203
B.2	Data collection and anonymization process	207
B.3	Instructions - Information sheet for the participants (Teilnehmerinformation)	210
B.4	Instructions - Consent form for the participants (Einverständniserklärung)	211
B.5	Instructions - Introduction: Welcome screen	212
B.6	Instructions - PANAVAKS scale	218
B.7	Instructions - Anagram task: exemplary anagrams	226
B.8	Instructions - Anagram task: Trial round	227
B.9	Instructions - Dual-2-back task: Possible colors and positions of squares	229
B.10	Instructions - Dual-2-back task: Color sequence	229
B.11	Instructions - Dual-2-back task: Position sequence	230
B.12	Instructions - Dual-2-back task: Color and position sequence	230
B.13	Instructions - Dual-2-back task: Control question 2	231
B.14	Instructions - Dual-2-back task: Control question 3	231
B.15	Instructions - Dual-2-back task: Control question 4	232
B.16	Instructions - Introduction to the real-effort tasks: General task instructions	235
B.17	Instructions - Introduction to the real-effort tasks: Control questions	235
B.18	Instructions - Introduction to the real-effort tasks: Snake trial round	236
B.19	Instructions - Introduction to the real-effort tasks: Ready for the tasks	236
B.20	Instructions - Real-effort tasks: Confirm task-understanding	237
B.21	Instructions - Real-effort tasks: Count-down timer	237
B.22	Instructions - Real-effort tasks: Button to switch to outside option "Snake"	238
B.23	Instructions - Real-effort tasks: Break-page	238
B.24	Multiplication task (Dohmen & Falk, 2011): Instructions	240

B.25 Multiplication task (Dohmen & Falk, 2011): Instructions	241
B.26 Multiplication task (Dohmen & Falk, 2011): Task page	241
B.27 Multiplication task (Dohmen & Falk, 2011): Results page	242
B.28 Word-transcription task (modified from Kephart, 2017): Instructions	243
B.29 Word-transcription task (modified from Kephart, 2017): Instructions	244
B.30 Word-transcription task (modified from Kephart, 2017): Task page	244
B.31 Word-transcription task (modified from Kephart, 2017): Results page	245
B.32 Code-transcription task (Kephart, 2017): Instructions	246
B.33 Code-transcription task (Kephart, 2017): Instructions	247
B.34 Code-transcription task (Kephart, 2017): Task page	247
B.35 Code-transcription task (Kephart, 2017): Results page	248
B.36 Word-encryption task (Erkal et al., 2011): Instructions	249
B.37 Word-encryption task (Erkal et al., 2011): Task page	250
B.38 Word-encryption task (Erkal et al., 2011): Results page	250
B.39 ab-Typing task (Berger & Pope, 2011): Instructions	251
B.40 ab-Typing task (Berger & Pope, 2011): Task page	251
B.41 ab-Typing task (Berger & Pope, 2011): Results page	252
B.42 Ball-catching task (Gächter et al., 2016): Instructions	253
B.43 Ball-catching task (Gächter et al., 2016): Instructions	254
B.44 Ball-catching task (Gächter et al., 2016): Task page	254
B.45 Ball-catching task (Gächter et al., 2016): Results page	255
B.46 Instructions - Personal hit-list	256
B.47 Instructions - Final motivational questionnaire (screenshot)	257
B.48 Instructions - Payment screen	260
B.49 Instructions - “Thank you for your study participation”-screen	261
B.50 Instructions - Waiting screen with Snake (optional)	261
B.51 Grouping of the items of the real-effort task survey	263
B.52 Survey implementation in the laboratory software oTree	264
B.54 Distribution of survey responses for all tasks and survey items	266
B.53 Percentage responses for all items of the real-effort task survey	270
B.55 Distribution of task indices for all tasks	275
B.56 Responses to real-effort task survey conditional on order of tasks in the task sequence	276
B.57 Visualization of the consistency of the subjects’ perception of the tasks	289
B.58 Factor analysis for the real-effort task survey	291
B.59 Distribution of the responses to the Personal Hit-List for all tasks	294
B.60 Correlation of the real-effort task survey as summated rating scale and the Personal Hit-List	296
B.61 Mean task ratings for clusters based on subjects’ rating of the multiplication task	297
B.62 Mean task ratings for clusters based on subjects’ rating of the single-slider task	298
B.63 Mean task ratings for clusters based on subjects’ rating of the ball-catching task	299
B.64 Approval by the ethics commission of ETH Zurich	303
B.65 Approval by the ethics commission of ETH Zurich (cont.): page 2.	304

B.66	Approval by the ethics commission of ETH Zurich (cont.): page 3.	305
B.67	Consent from the data protection officer of the University of Hamburg	306
B.68	Consent from the data protection officer of the University of Hamburg (cont.): page 2.	307
B.69	Financial support for the laboratory experiments: Research grant from the MTEC Foundation	309
C.1	List of characterization questionnaires	312
C.2	List of characterization questionnaires (cont.): page 2.	313
C.3	List of characterization questionnaires (cont.): page 3.	314
C.4	List of characterization questionnaires (cont.): page 4.	315
C.5	List of characterization questionnaires (cont.): page 5.	316
C.6	Score distributions for the selection of real-effort tasks	318
C.7	Ranking of subjects according to their “normalized total score”	320
C.8	Correlation between subjects’ “normalized total score” and their score in each task	320
C.9	Correlation between subjects’ qualities	322
C.10	Correlation between subjects’ motivational characteristics	323
C.11	Correlation between the items of the real-effort task survey	324
C.12	Pearson-correlations for task scores and responses to the real-effort task survey	326
C.13	Superimposed scores conditional on subject’s task perception	327
C.14	Compare the performance of subjects with and without outside option available	338
C.15	Compare the performance of subjects with outside option available, but did not use it, and without	339
C.16	Schematic representation of the machine learning approach	341
C.17	The single-button task	346

List of Tables

Table Number	Page
1.1	Subject qualities that are required for a task or that promote success in the task 23
1.2	Skill- and personality-based classification of tasks 24
1.3	Generalizability of the results obtained 29
2.1	Frequently employed outside options 79
3.1	Real-effort task selection employed in the experiment 115
3.2	Descriptives for the motivational items of the real-effort task survey 126
3.3	Regression estimates for all motivational survey items (simple model) 128
3.4	Descriptives for the effort-type items of the real-effort task survey 130
3.5	Regression estimates for all effort-related survey items (simple model) 131
3.6	Multiple comparisons with the highest mean to identify a favorable task 137
3.7	Multiple comparisons with the lowest variance to identify a favorable task 138
4.1	Skill and personality traits considered decisive for subject performance in the task se- lection 162
4.2	Motivational dimensions regarded as crucial for subject performance in the task selection 170
4.3	Task performance conditional on all potential subjects' characteristics, including controls 182
4.4	Task performance conditional on all subjects' motivations, including controls 185
4.5	Task performance conditional on all subjects' characteristics (skills, personality) and motivations, including controls 187
B.1	Experiment content and author contributions 217
B.7	Final motivational questionnaire 257
B.8	Orthogonal contrast-coding scheme 267
B.9	ANOVAs for the fixed effect "task" for each motivational survey item 267
B.10	ANOVAs for the fixed effect "task" for each effort-related survey item 268
B.11	Greenhouse-Geisser estimates 268
B.12	Regression estimates for all motivational survey items with adjusted p-values (simple model) 271

B.13	Regression estimates for all effort-related survey items with adjusted p-values (simple model)	272
B.14	Calculations for Hsu's multiple comparisons with the best (for means)	274
B.15	ANOVAs for the fixed effects "task" and "index" for each motivational survey item . . .	278
B.16	ANOVAs for the fixed effects "task" and "index" for each effort-related survey item . .	278
B.17	Regression estimates for all motivational survey items (complex model)	279
B.18	Regression estimates for all effort-related survey items (complex model)	280
B.19	Assess the impact of additional regressors for the motivational items with the AIC-criterion	282
B.20	Assess the impact of additional regressors for the effort-related items with the AIC-criterion	282
B.21	Assess the impact of additional regressors for the motivational items with an ANOVA .	283
B.22	Assess the impact of additional regressors for the effort-related items with an ANOVA	283
B.23	Detailed factor loadings and factor correlations	292
B.24	Outside option usage pattern	293
B.25	Activities subjects stated as strongly disliked activity and strongly liked activity	295
B.26	"Lovers" and "haters" of tasks according to the Personal Hit-List	300
B.27	Average rating of all tasks for the "lovers of a particular task"	300
B.28	Average rating of all tasks for the "haters of a particular task"	301
C.1	Descriptive statistics of all real-effort tasks (normalized scores)	318
C.2	Descriptive statistics for the control variables	328
C.3	Share of subjects that re-read the instructions for each task	330
C.4	Task performance conditional on all subjects' qualities, including controls	331
C.5	Task performance conditional on all subjects' motivations, including controls	332
C.6	Task performance conditional on all subject characteristics (qualities and motivations), including controls	333
C.7	Task performance conditional on subjects' characteristics for selected tasks	335
C.8	Task performance conditional on the availability of an outside option	336
C.9	Outside option usage for specific tasks	337
C.10	Session info on R environment	344
C.11	Required packages	345

Acknowledgments

Without the help of numerous people, this work could not have been realized. First and foremost, I would like to thank the participants of the laboratory experiments, without whose participation this thesis would not have been possible in the first place.

I will always be most indebted to my supervisor Marko Köthenbürger. With Marko, I had an ideal mentor for this work, who granted me a high degree of encouragement and freedom, coupled with feedback and support at all times. Furthermore, I would like to thank Petra Schmid for agreeing to co-supervise my thesis. Through her perspective from psychology, she gave my work in many ways new impulses, which enriched it greatly. I thank her very much for her valuable support and encouragement.

I would also like to thank the members of the Chair of Public Economics at ETH Zurich who have accompanied me during this period. Mohammed Mardan, Michael Stimmelmayr, and Federica Liberini deserve special mention for their constant support.

Furthermore, I would especially like to thank Christoph Schulze for countless discussions on behavioral economics and designing experiments. I have benefited from his challenging views and thought-provoking ideas in many ways.

Special mention should be made of the very cooperative collaboration in planning and conducting the laboratory experiments with Olaf Bock and the staff of the WiSo Research Laboratory at the University of Hamburg. I would also like to give a particular word of thanks to Alexander Sandukovskiy for his support in the technical implementation of the laboratory experiments. The eager testing of

the lab software by Christopher Klenk, Christian Netzeband and Jocelyn Schmidle also deserves to be mentioned. In addition, I would like to thank the MTEC foundation for generous support in funding the experiments. Without this backing, the studies contained in this thesis could not have been carried out to their extent.

My thanks further go to Falko Rheinberg, Steve Heinke, Christian Waibel, Gunda Johannes, Jan Schmitz, and Klaus Krippendorff for their advice on the content of this thesis. I would like to thank Lilian Gasser and the ETH Zurich Seminar for Statistics for their detailed consultation on methodological questions. For their proofreading work, I thank Dieter Waloszek, Aled Seys-Llewellyn and Carlos Gonzales.

I would also like to thank the people who accompanied me outside the academic world. In particular, I would like to mention my siblings Lena Frey and Sonja Rupp and my grandmother Lilli Jundt, as well as Christopher Klenk and my rowing friends. To the latter I owe a debt of gratitude that I could or had to leave the office at least some days a week at reasonable times. Their mental support also outside the rowing boat kept me afloat time and again.

And finally, I thank my parents for their far-reaching and loving support along my way, and for making me the person I am. My greatest thanks go to my girlfriend Maja and my son Mads Emil – they were always there for me, and always believed in me.

Dedication

To my parents

Abstract

Real-effort tasks are widely used in experimental research to study effort provision. A great variety of tasks exists, each of which has its own properties. The tasks are employed in different areas of application. The tasks carry different properties, which involve advantages and disadvantages depending on the specific area they are applied. This thesis is about the choice of task and its implications. It is structured as follows.

The *first chapter* introduces the broader topic and provides background information on measuring effort in experiments. As real-effort tasks greatly differ in their design, several ways of classifying them are presented. The task classification includes their *degree of realism*, the *extent to which the output produced is useful*, and the *skills and character traits* required to perform a particular task well.

The *second chapter* provides an overview of the literature that criticizes the design and implementation of real-effort tasks. As a synthesis of the corresponding literature, a series of *design criteria* are presented. The criteria aim to improve *experimental control* while maintaining the *greater realism* of real-effort measurements compared to stated effort. To achieve this, design practices are presented in order to enhance experimental control over the effort-cost function to ensure that *voluntary effort provision* is kept to a minimum and over the output-production function to ensure that *actual effort is required* to complete the task.

To evaluate and compare tasks with regard to these aspects, the *third chapter* introduces a new methodology, the *real-effort task survey*. The survey is filled out by (prospective) study participants and determines their subjective perception of the task design. This is crucial because only they themselves can judge i) to what extent a task motivates them to make voluntary efforts and ii) how strenuous it is for them. Furthermore, the results of a first application of the survey are presented, comparing seven

frequently used task types.

To shed light on the impact of task properties on effort measurements, *Chapter four* examines the influence of subject characteristics on individual performance. To this end, the study presented in Chapter three contains several additional elements to characterize the study participants. Using methods of motivation diagnostics and machine learning, abilities, personality, and motivation are found to explain a large part of the variation in the subjects' observed effort.

Chapter five concludes this thesis, sums up the results and contributions, puts them into relation, and provides an outlook on prospective research.

To conclude, this work aims to raise awareness of the various properties of tasks, their differences, and their varying suitability for a given application. The thesis makes several conceptual and methodological contributions and serves the practitioner to classify, design, select, and implement tasks. In summary, tasks are not simply neutral and interchangeable. This is why the choice of task is vital and must match the research question being investigated.

Introduction to the Thesis and Dissertation Overview

Real-effort tasks are frequently used in experimental and behavioral economics and related social sciences to investigate effort provision. In the last decades, a wealth of tasks has been introduced.¹ Each of them has different properties and resembles real work to a greater or lesser extent (see Figure 1).² They all have advantages and disadvantages that only really become apparent in the light of their application – i.e., in relation to a concrete research question and the chosen study design. To what extent the task properties influence the results in a desired or undesired way depends on how well the task matches the object and approach of research.

¹A first tabular summary of real-effort tasks compiled by Christina Gravert, Assistant Professor at the Department of Economics at the University of Copenhagen, has been circulated via the Experimental Methods Discussion Google group of the Economic Science Association in early 2014 (Gravert, 2014). Charness et al. (2018) present an extensive tabular summary of tasks giving examples of experiments using stated and real effort. Grouping real-effort tasks in several categories, their table portrays the multitude and diversity of tasks proposed in the literature.

²The exemplary tasks mentioned in Figure 1 were used, for instance, in the following studies. *Catching falling balls*: Gächter et al. (2016); *cracking walnuts*: Fahr & Irlenbusch (2000); *door-to-door fundraising for a non-profit*: Gneezy & List (2006); *entering book codes into a library database*: Gneezy & List (2006), Hennig-Schmidt & Sadrieh (2010), Corgnet, Hernán-González, Kujal, et al. (2015), Charness et al. (2016); *moving one or more sliders*: see Section 2.5 and Gill & Prowse (2015).

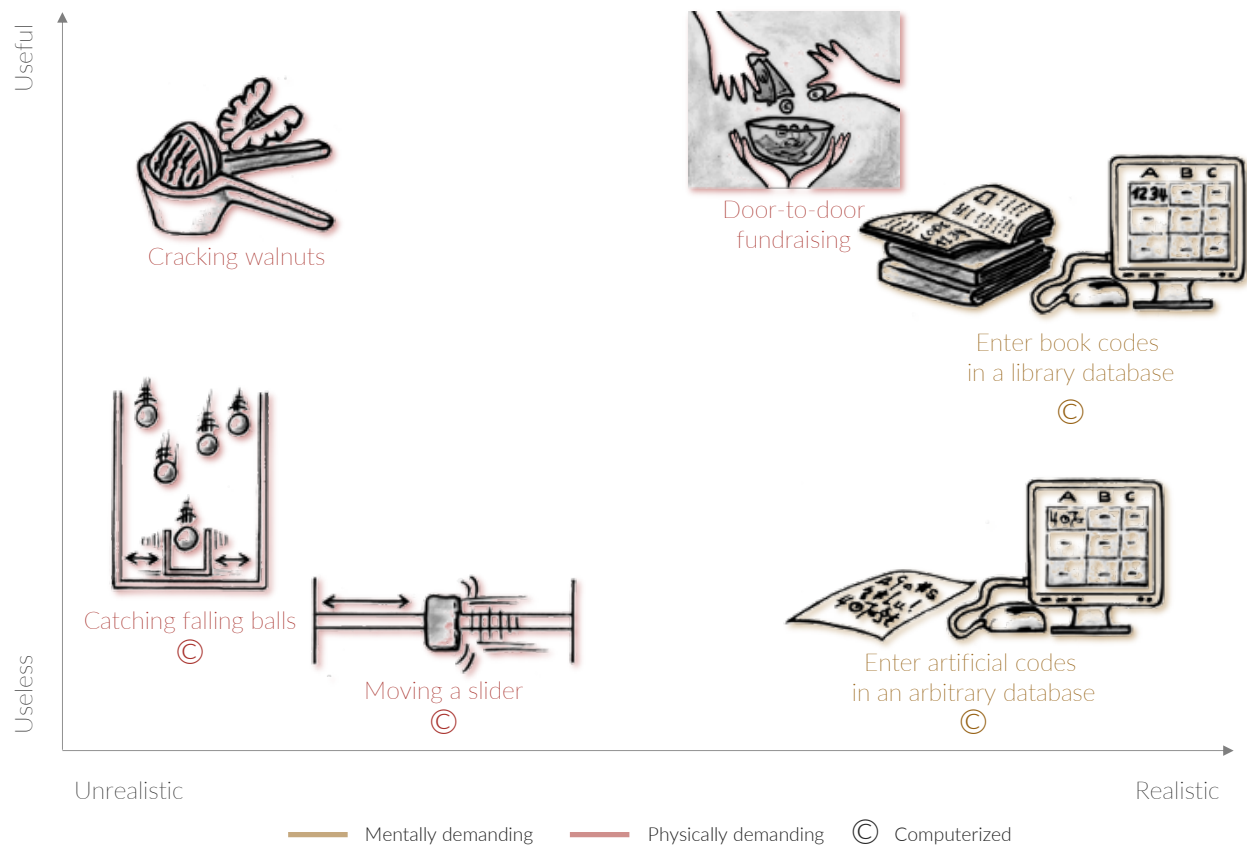


Figure 1: **Tasks in real-effort experiments:** Experimentalists usually employ *real-effort* tasks to study subjects' effort provision or to endogenize endowments in experiments. The tasks differ in terms of the *usefulness of produced output* (catching falling balls vs. cracking walnuts), their *degree of realism* (moving a slider vs. entering random codes into a database), whether they are *cognitively* or *physically demanding* (entering book codes into a library database vs. door-to-door fundraising for a non-profit organization), whether they are *computerized* or *performed manually* (cf. those previously mentioned). “*Useful tasks*” generate an outcome *and* this outcome is valuable, i.e., is meaningful to the subject, the experimenter, or a third party. Tasks are “*realistic*” if subjects can perform them (or similar tasks) outside the laboratory to earn money. One can customize the data entry task in terms of its usefulness by using either real book codes or random, artificial codes. Similarly, some tasks can be adjusted in their degree of realism. Adaptations of this kind allow the task and the salience of its usefulness and realism to be tailored to the research question.

Broadly speaking, three different applications for real-effort tasks in experimental research can be distinguished: In a *first application*, researchers want to examine a particular effect in isolation and measure effort provision conditional on specific treatments as accurately as possible (e.g., the effect

of stake sizes with regard to piece rates).³

In a *second application* the subjects have to make a genuine effort to generate funds in preparation for a later investment decision or pro-social decision, e.g., in a dictator game, ultimatum game, or public-good game. The researchers want the study participants to accumulate the initial endowment without knowing anything yet about the later economic decision that is actually being studied. To this end, the acquired earnings should be as equal as possible for all subjects, so that they have similar starting conditions for their further decision making.⁴

In a *third application* the researchers want to approximate a real working situation as closely as possible, whereby all effects induced by the actual work task are ideally reflected in the experiment.⁵

Implementing a particular task in the aforementioned applications entails certain advantages and disadvantages. However, these are possibly associated with specific effects on the effort measurement or downstream investigations. As a starting point for examining the consequences in each setting, the best-possible experimental situation is assumed: The subjects are randomized well across the different treatments of the study such that the design is fully balanced in terms of subjects' characteristics (e.g., demographics, skills, personality traits, motivation). Influences due to sample bias can, therefore, be ruled out.⁶ Since effort made is difficult to observe and directly measured, it is commonly equated

³The role of stake-sizes are examined in [Ariely, Gneezy, et al. \(2009\)](#), [Corgnet et al. \(2016\)](#) and [Houy et al. \(2016\)](#) and in combination with pro-social incentives (working for a charity) in [Imas \(2014\)](#). For gender effects in effort provision conditional on different incentive schemes (fixed payment, piece-rate payment, team production and tournament) see [Bortolotti et al. \(2016\)](#), [Niederle & Vesterlund \(2007\)](#), [Nabanita et al. \(2013\)](#) and [Mascllet et al. \(2015\)](#).

⁴Most commonly subjects work (unawarely) in preparation of a pro-social decision situation ([Fahr & Irlenbusch, 2000](#); [Bonein & Denant-Boèmont, 2015](#); [Bosman & Winden, 2002](#); [Dutcher et al., 2015](#); [Nikiforakis et al., 2012](#); [Rutström & Williams, 2000](#); [Reinstein & Riener, 2009](#); [Carpenter et al., 2014](#)) or a later investment decision ([Fochmann et al., 2012](#); [Corgnet, Hernán-González, Kujal, et al., 2015](#)). In both cases, the aim is to curtail/diminish any "house money"-like effects, whereby subjects who have received an endowment as a windfall (from the experimenter) behave more benevolently or risk-takingly than with money they have really "earned" ([Hoffman et al., 1994](#); [Corgnet, Hernán-González, Kujal, et al., 2015](#); [Reinstein & Riener, 2009](#); [R. Thaler, 1985](#); [R. H. Thaler & Johnson, 1990](#)).

⁵The principal-agent model is widely used in contract theory and labor supply theory to study incentive effects prevalent in the real work environment ([Carpenter & Huet-Vaughn, 2017](#)). A large number of experimental investigations with real-effort tasks to test this theory have been conducted. For example, experiments regarding taxation and the "Laffer"-curve have been performed by [Fochmann et al. \(2013\)](#), [Lévy-Garboua et al. \(2009\)](#), [Sutter & Weck-Hannemann \(2003\)](#) and [Swenson \(1988\)](#).

⁶Many experimental studies deviate from the described ideal experimental settings. This means that additional biases are introduced, for example, due to small sample sizes or non-random participant assignment to treatments.

with the output produced, i.e., the score obtained by the subjects in the task.⁷

Consider the first experimental application outlined above. Suppose the execution of a task requires particular abilities or personality traits: Those who possess them to a greater extent will benefit, thus having an easier time completing the task. If the study participants possess these qualities to varying degrees, they will incur different costs. As a result, the effort ultimately measured will depend on the predisposition of a subject. In terms of the second application of tasks, if the earnings generated by providing effort in the task depend on individual subject characteristics, any subsequent investigations employing the endowment would likewise contain a bias.

In a linear world, all this would not pose a problem. Relevant subject characteristics like demographics, skills, or preferences can be inquired from the study participants before or after completing a task. In any subsequent analysis, they can be controlled for in a straightforward manner. However, if the subjects' characteristics do not enter linearly, the situation changes. Mere differencing is no longer sufficient such that the subject-specific effects can no longer be easily deducted in the analysis. For these reasons, it is helpful to use a task in the first two above-mentioned fields of application that *does not introduce any bias*. Such tasks are mostly generic, outright simple, and can be coined "neutral tasks." Specifying, refining, and comparing them forms an essential part of this work. Research conducted in the realm of the first and second application of real-effort tasks somewhat aims for *task neutrality* in a sense that it is *independent* of the choice of a (neutral) task.

In the third application described previously, the situation is different. The goal is to approximate a real working environment with its actual work activity as closely as possible in order to achieve *realism*. Thus, any arousal of feelings and motivations as well as prerequisites of worker qualities that are present in the original work environment and activity, are ideally mimicked in the study. To achieve this *congruence*, the study design and the chosen task must reproduce these subtleties. Clearly, the measured effort in this task application may well be influenced by the characteristics of the subject (and this is intended to be the case to achieve this conformity). However, it is highly preferable that an informed task-selection is made to enable the congruence between the *task and lab setting* and

⁷However, there are exceptions: In the literature employing "creative" tasks, [Kachelmeier et al. \(2008\)](#) distinguish between *quantity* and *quality*, and [Laske & Schröder \(2017\)](#) additionally also *originality* of produced output. [Abeler et al. \(2011\)](#) and [Noussair & Stoop \(2015\)](#) utilize *waiting time* as a measure of effort. [Bortolotti et al. \(2016\)](#) use the *number of mistakes made* as an indirect measure of the effort expended, whereby more mistakes are equated with less effort made to complete the task.

reality, i.e., the *topic of research*.

Such tasks that possess latent properties that people associate with actual work can be coined “realistic tasks.” Which task properties might be crucial to emulate and match a real work situation will be explored in more detail in the following chapters. Figure 2 summarizes the three applications of real-effort tasks, the goal of a task in each of them, and the most appropriate task type for each application.

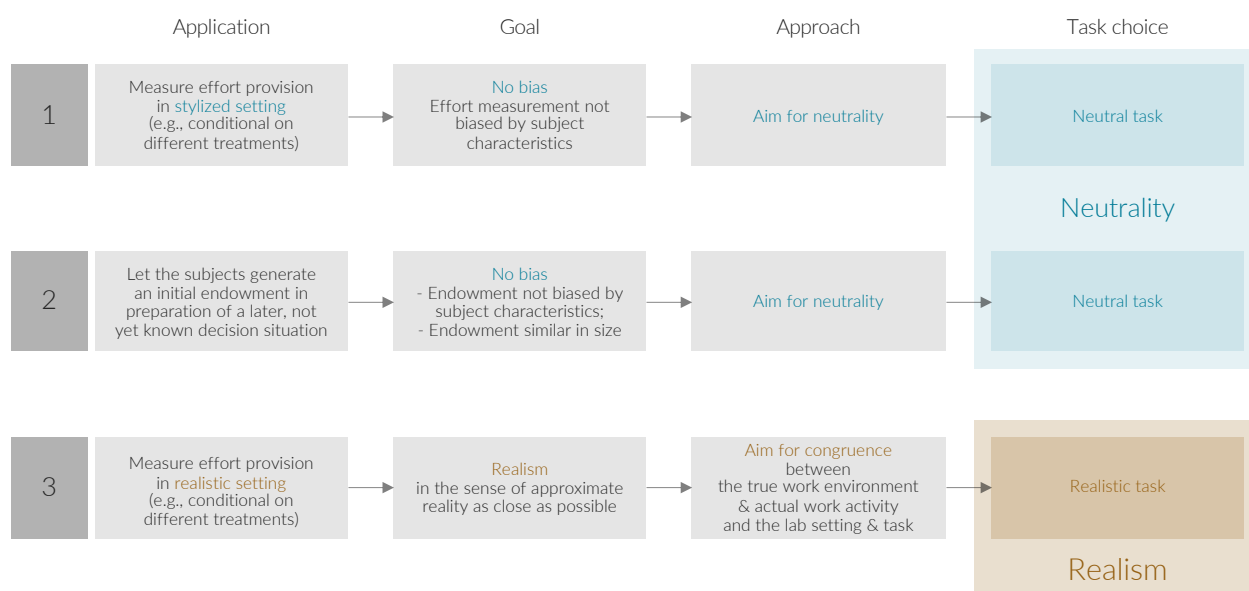


Figure 2: **Applications of real-effort tasks:** The tasks allow performing effort measurements for a wide range of research purposes, with different research questions and contexts. Tasks can be very simplistic to abstract from the situation as a whole and to eliminate all unwanted side effects (application 1). Or they can also try to reproduce reality as accurately as practicably possible to deliberately include the aforementioned influences (application 3). The tasks are further employed to let the subjects accumulate their “own” financial resources, which they will have to use in subsequent socio-economic decisions (application 2). With the goal of *no bias* in applications one and two, neutral tasks are suitable to achieve **neutrality**; with the somewhat contrasting goal of *realism* in application three, realistic tasks are preferred to achieve **congruence**. Whether realism can actually be achieved ultimately hinges on whether or not a congruence between laboratory and reality can be established. This, in turn, depends strongly on the laboratory setting and the choice of the (specific) task.

Overview of the Research Agenda

As noted initially, the extent to which the advantages and disadvantages of a task entail implications in the outlined task applications depends on how well the task's design and the research purpose match. The *choice of the task* is, therefore, of vital importance for research involving real-effort tasks. This thesis offers several alleys to facilitate the search for a suitable task for a particular research purpose. For this, the first two chapters aim to deepen the understanding of the scope of task properties by providing a novel classification of tasks (Chapter 1) and a set of design criteria to develop, select and implement tasks (Chapter 2). Subsequently, chapters three and four aim to illuminate the influence of task properties on the effort measurement by presenting experimental evidence for the differences in the design of tasks in terms of how study participants perceive them (Chapter 3) and the extent to which the subjects' characteristics determine individual task performance (Chapter 4). Each one of the chapters is described in more detail below.

Chapter 1: Approaches to Measuring Effort and Classification of Real-Effort Tasks. To narrow down the topic and establish a common ground, the two common approaches to studying effort provision – *chosen effort* and *real effort* – are presented. Thereafter, ways to classify real-effort tasks based on their *degree of realism*, the *usefulness of produced output*, and *abilities and personality traits* required to succeed in task completion are introduced. Furthermore, the proposed neutral and realistic tasks are described, also with regard to the discussed *fields of application*. Suggestions are offered to researchers who are unsure which approach to adopt in their research. Moreover, the chapter also discusses the extent to which experimental results may generalize across different settings and tasks.

Chapter 2: Developing Real-Effort Tasks. Experimentalists employ real-effort tasks to induce psychological effects that are present in the field environment also in the laboratory setting. However, any increase in *realism* compared to stylized, chosen effort goes hand in hand with a substantial *loss in control over the costs and benefits* of effort provision. Combining the best of both worlds would be ideal, i.e., having tasks with great realism *and* great control. After discussing a variety of shortcomings of real-effort tasks, the second chapter presents a comprehensive set of *design criteria* to approach this

ideal. These criteria aim at increasing experimenter control over effort cost and output production. In addition, practices for designing and implementing tasks to achieve these objectives are described. Some design practices can be adopted in [any application](#) of real-effort tasks (task-independent design practices); address the specific design of a task (task-dependent design practices). Overall, the criteria and practices can help experimenters to develop new tasks or to select an existing task and to improve its implementation. Finally, based on these findings, a novel task is proposed (the [single-slider task](#)).

[Chapter 3: Comparing Real-Effort Tasks.](#) When discussing the shortcomings of tasks in the previous chapter, it was pointed out that extrinsic incentives for the provision of effort may be substituted by “intrinsic incentives,” such as joy in performing the task or the desire to please the experimenter. Furthermore, tasks can be more or less strenuous and require different types of effort (e.g., physical or mental). To evaluate and compare tasks in these terms, the chapter introduces a new methodology, the *real-effort task survey*. It is conducted among study participants to determine their perception of any tasks under consideration. After completing a given task, subjects rate it according to how much and what type of effort it requires and how it is designed to assess whether it initiates voluntary effort. With regard to the [first and second application area of tasks](#), the survey allows for identifying a task that is perceived by subjects *similarly* and as *not motivating* and as physically or mentally sufficient *demanding*. Concerning the [third task application](#), the survey analogously facilitates identifying a task that matches the specific environment of the research question being addressed (which might require a non-neutral setting). The chapter also presents a first application of the survey to compare seven distinct tasks. The selection represents task types frequently used in the literature and is based on the classification proposed in Chapter 1. Substantial differences across tasks are found: according to the subject’s perception of the tasks, these vary considerably along motivational dimensions and in terms of types and amounts of effort demanded. Besides, [Hsu \(1996\)](#)’s method of *multiple comparisons with the best* is employed to differentiate the tasks. Regarding the criteria for neutral tasks, the method identifies the single-slider task as the most favorable in the task set.

[Chapter 4: Determinants of Real-Effort Task Performance.](#) The previous chapter revealed that tasks i) differ significantly along motivational dimensions, and ii) demand different amounts and types of effort according to subjects’ assessments. The former finding suggests the presumption that a moti-

vating task design may evoke voluntary effort provision. Moreover, the latter indicates that different tasks require different abilities to perform well, as previously argued in Chapter 1. This chapter continues in this line of thought and assesses whether individual task performance depends on further factors beyond the effort exerted. For this, the above-discussed laboratory experiment contained several additional elements to examine to what degree the *subjects' personality, abilities, and motivation* drive their performance. Differences in these *subject characteristics* were evaluated through frequently used psychological questionnaires, regularly carried out in studies in behavioral economics and psychology. The actual behavior of the participants was assessed using the previously mentioned diverse set of real-effort tasks. Subjects are found to score very differently across the selection of tasks. To determine which subject characteristics actually affect task performance, various methods are employed, including approaches from motivation diagnostics and supervised machine learning. Depending on the task, the subjects' abilities, personality, and motivation account for notable to substantial portions of the individual differences in performance. The results suggest that the task choice is decisive for experimental outcomes since individual performance – and thus the “measured effort” – is strongly task-dependent.

The findings are relevant for experimental researchers who use tasks in their research and for readers of the literature on real effort, especially those who want to compare results across tasks. The following considerations can be derived from the results for choosing a task in the [applications](#) described above.

In the first and second application area, completing a task ideally depends as little as possible on the subjects' characteristics. Thereby, there is no longer a strong imperative to control for them in a later analysis. Consequently, the aim is to find a task that is as “neutral” as possible. However, suppose it is unavoidable to use a task that hinges on certain subject characteristics, e.g., to relate to the literature. In that case, it should at least be manageable and straightforward to control for them.

In application three, approximating an actual work situation, the reasoning differs fundamentally, and the aim is *not* to control for subject characteristics. Instead, the congruence between the experiment and the real world is the primary consideration that guides the choice of task. The goal thereby is to ensure that “the right subject characteristics” are decisive for the execution of the task, i.e., the same that are relevant in the real environment in order to reproduce it as accurately as possible.

[Chapter 5: Conclusions and Discussion](#). The final chapter of this dissertation summarizes the various contributions and findings, places them in relation to each other, and offers an outlook on possible future research.

1

Approaches to Measuring Effort and Classifying Real-Effort Tasks

The previous chapter briefly outlined [three applications](#) of real-effort tasks. For each of them, conceptual concerns were expressed, and possible ways forward were identified revolving around *neutral* and *realistic* tasks. This chapter provides background information on the measurement of effort provision in general and gives further details on the remedies discussed.

To begin with, the two common approaches to studying effort provision – *chosen effort* and *real effort* – are compared and contrasted. Then, a typology of real-effort tasks is presented to distinguish them whether they produce a useful outcome and whether they are very similar to a real workplace. Furthermore, tasks can be classified based on their demand for skills or particular character traits. This additional level of task differentiation extends the existing typology, and neither impairs nor limits it.

The presented approach is, thereafter, compared with classifications suggested in the literature. After a brief discussion of the generalizability of experimental results with neutral and specific tasks, some final remarks conclude the section.

1.1 Approaches in the Study of Effort Provision: Chosen Effort vs. Real Effort

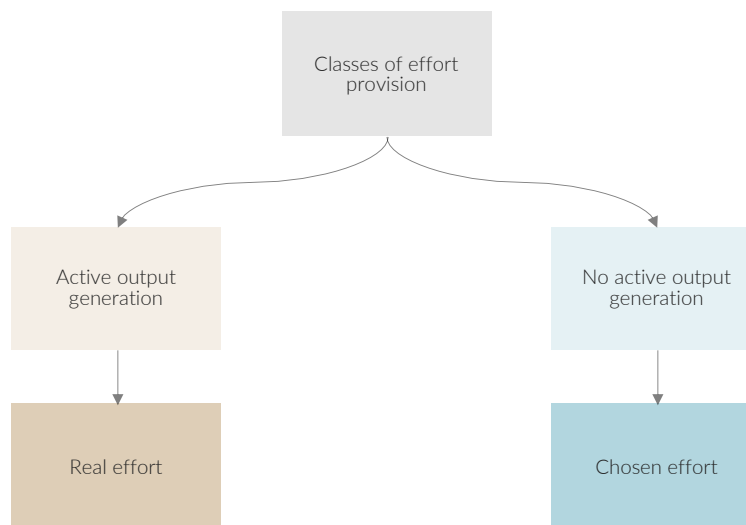


Figure 1.1: **Approaches of effort provision**

Two approaches are common in the study of effort provision: “chosen effort” and “real effort.” *Chosen effort* is also referred to as stylized or stated effort. There, study participants are confronted with an *abstract choice situation*: Instead of providing actual effort, subjects have to select an “effort level” from an explicit *effort-cost table*. Their payoffs derive from an artificial monetary cost function, where a higher (chosen) effort reduces the experimental earnings of a subject. (However, it may increase the payoffs of other parties, e.g., if the study aims to reflect a firm-worker relationship, for instance that of an “employer” or an “employee colleague.”) In contrast, *real effort* involves a genuine exertion over a sustained period of time.¹ In the following, arguments for each approach are presented.

¹Apart from the implementation introduced by Dutcher et al. (2015), the timing of decisions naturally differs for the two approaches to measure effort.

The key advantage of chosen effort is its *supreme control*. The *cost-of-effort* are effectively induced by the experimenter, who has complete authority over the shape of the function (Carpenter & Huet-Vaughn, 2017; Dutcher et al., 2015). This rigid control makes it possible to tailor the effort-cost function to a particular real work situation's specific features, focusing on the essential and abstracting from minor details. The resulting firm relation between experimental design and underlying theory enables the researcher to state and test specific hypotheses and to derive a better "understanding for when and perhaps why individual behavior differs from those theoretical predictions" (Dutcher et al., 2015, p. 3). Moreover, stated effort leaves no room for distortion of effort choices due to individual ability or learning (Brüggen & Strobel, 2007, p. 232). The researcher can also decide how effort costs should vary across study participants and the course of the experiment (Gill & Prowse, 2015, p. 1). Finally, chosen effort experiments are far less elaborate and require much less time to conduct than real-effort experiments.

With no real effort being exerted to produce output, *preferences over outcomes* are essentially induced in chosen-effort experiments (Carpenter & Huet-Vaughn, 2017, p. 2). The authors further emphasize that in addition to the reward, the "actual" costs of effort are also expressed in monetary terms. This means that labor provision is highly abstracted and generalized – completely removed from any specific work environment. It is also not accompanied by any *blood, sweat, or tears*. Dijk et al. (2001, p. 189) add that real work also includes a social dimension and "involves effort, fatigue, boredom, excitement and other affectations not present in the abstract experiments" with chosen effort. Carpenter & Huet-Vaughn (2017, p. 3) expand on this, highlighting that the effort-cost function induced in stylized effort does not necessarily adequately capture these integral components and essential features of work. Montmarquette et al. (2004, p. 1380) further point out that many stated-effort experiments suppose an equivalence between "intention of contribution *and* effort," i.e., stated effort as *effort intended to be made* and *effort actually provided*, as well as between "disutility of effort *and* money," i.e., *disutility by providing effort* and *disutility due to lower-income*. However, these studies do not provide sufficient evidence for these cognitive or psychological equivalencies. To conclude, it may come as no surprise that Gneezy & List (2006, p. 1366) question whether "the behavior of laboratory subjects, who are asked to choose an effort or wage level (by circling or jotting down a number) in response to pecuniary incentive structures, [is indeed] a good indicator of actual behavior in labor markets."

As previously outlined, real effort requires *actual effort provision*, which is mostly physically or mentally demanding. This indeed renders the experiments less abstract and artificial (Charness & Kuhn, 2011, p. 5). Besides, the more realistic effort provision is, the more it provokes psychological effects such as feelings of attachment and entitlement (Fahr & Irlenbusch, 2000; Nikiforakis et al., 2012; Rutström & Williams, 2000). With the ambition that behavior observed in an experiment more accurately mirrors actual behavior at work, real effort intends to expand the realism and authenticity of laboratory studies (Carpenter & Huet-Vaughn, 2017, p. 17). Among others, Gill & Prowse (2015, p. 1) go further and claim that real-effort improves external validity.

However, many real-effort tasks are very abstract and do not resemble a proper workplace closely.² More severe is the loss of experimental control associated with the provision of genuine effort (Carpenter & Huet-Vaughn, 2017; Charness et al., 2018). This aspect substantially diminishes the advantages of real-effort over stated-effort experiments. Dutcher et al. (2015, p. 2) highlight that “the nature of the cost function is unknown and typically not under the control of the experimenter.” The authors continue to state that a reduced connection to theory permits for less precise predictions and makes it harder to recognize any divergence from these. Gill & Prowse (2015, p. 4) emphasize the heterogeneity in effort costs across study participants due to unobserved differences in skills required by the task. Brüggem & Strobel (2007, p. 236) expand on this and point out that “except for the plausible assumption that the individual cost of effort is increasing and convex, the functional form remains unclear.” It may even be that some tasks are perceived as so entertaining by certain study participants that it is not entirely clear whether they will incur any costs for completing them. Finally, since effort is difficult to observe directly, the assumption of a one-to-one mapping to produced output is inevitable and necessary for a rigorous equilibrium analysis (Bortolotti et al., 2016, p. 63).

Comparing subjects' behavior in both approaches to measuring effort, Dutcher et al. (2015, p. 2) stress the “importance of matching cost functions across contexts.” The authors indicate that this will most likely lead to very similar results if the induced cost function in the chosen-effort approach coincides with the true costs borne by subjects in the real-effort task. If this is not the case, results will – not surprisingly – differ. Likewise, if the actual costs incurred by the subjects in performing a

²This drawback can be mitigated by the choice of task, as explained in detail in Section 1.5.

particular task do not take the form assumed in the present model, the behavior will deviate from any theoretical predictions (Dutcher et al., 2015, p. 2).

Despite the limitations discussed, real effort certainly seems more authentic and credible than highly stylized, chosen effort. For this reason, the theme of this thesis is to optimize and refine the usage of real-effort tasks in such a way that, with greater control and sufficient precautions, real effort is superior to chosen effort. Since genuine effort can be delivered in various ways, the next section introduces a typology of tasks.

1.2 A Classification Based on the Realism of the Task

Real-effort tasks add *mundane realism* to laboratory experiments (Carpenter & Huet-Vaughn, 2017, p. 2). First and foremost, this entails that *actual effort* is provided. Second, *behavioral* or *emotional responses* may more likely be triggered than in chosen effort (Fahr & Irlenbusch, 2000; Ku & Salmon, 2012). Third, the task may produce an *output of use and relevance outside of the laboratory* (Dutcher et al., 2015). Fourth, the task may *aim to mirror* or even *represent a work task*. However, tasks that do not correspond to actual work situations are equally conceivable, and there may be good reasons to implement one of them. Real-effort tasks may vary significantly along these four “dimensions of mundane realism” (see Figure 1.2 for an illustration). The first dimension is evident and given for all tasks, while the second is more subtle and requires further (experimental) investigation. Therefore, the following concentrates on the third and fourth dimensions to provide a general basis and explain why the later sections of the thesis focus on *useless, unrealistic* real-effort tasks.

The extent to which a real-effort task resembles an actual work task or even a work environment determines how *realistic* it is. One can imagine a continuum of tasks ranging from *realistic tasks*, that represent actual work being carried out in a laboratory environment, to *unrealistic tasks* that have nothing in common with the world of work. Apart from that, *useful tasks* could be defined as producing tangible output with relevance outside of the laboratory, while *useless tasks* yield nothing worthwhile (Dutcher et al., 2015, p. 4).³

³In terms of later terminology used in the “Design Criteria” proposed in Section 2.4, a task is considered *purposeful* if it has an outcome, i.e., produces a (tangible) output and is thus not purposeless, and is in addition *meaningful*, if this

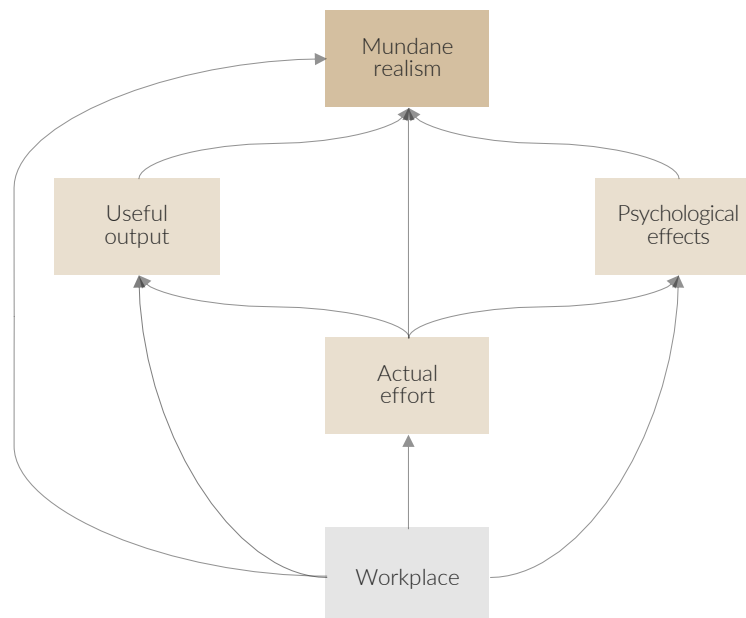


Figure 1.2: **Mundane realism:** The purpose of real-effort tasks is to add realism to laboratory experiments (Carpenter & Huet-Vaughn, 2017, p. 17). The figure illustrates key components to achieve this.

Realistic tasks that mirror a real-work situation very closely do not necessarily have to lead to a useful outcome. For example, a data-entry task can be implemented as a useful task by providing subjects with actual data that they enter, thus helping to build a real database. The task can be modified so that it degenerates into a useless task by providing irrelevant or artificially created data to the subjects. Similarly, real data can be used, but the services provided by the subjects do not serve any subsequent purpose. In any case, the purpose of the task intended by the experimenter must be 1) communicated clearly enough that all subjects can unambiguously understand it, and 2) communicated evidently enough that all subjects can perceive it and act upon it. The conveyance and conspicuousness of the purpose are decisive for whether the subjects ultimately perceive the task as useful or useless.⁴ Figure 1.3 illustrates the typology described and provides examples of 1) “useful” and “useless” and 2) “realistic” and “unrealistic” tasks.

outcome has some value that persists outside of the laboratory. *Useful tasks* are thus both purposeful and meaningful.

⁴For a related discussion, see Smith and his precepts of economic experiments (1982, 2010). The author lists *salience of rewards* to be of utmost importance to the function and success of an experiment (Smith, 1982, p. 931; 2010, p. 132).

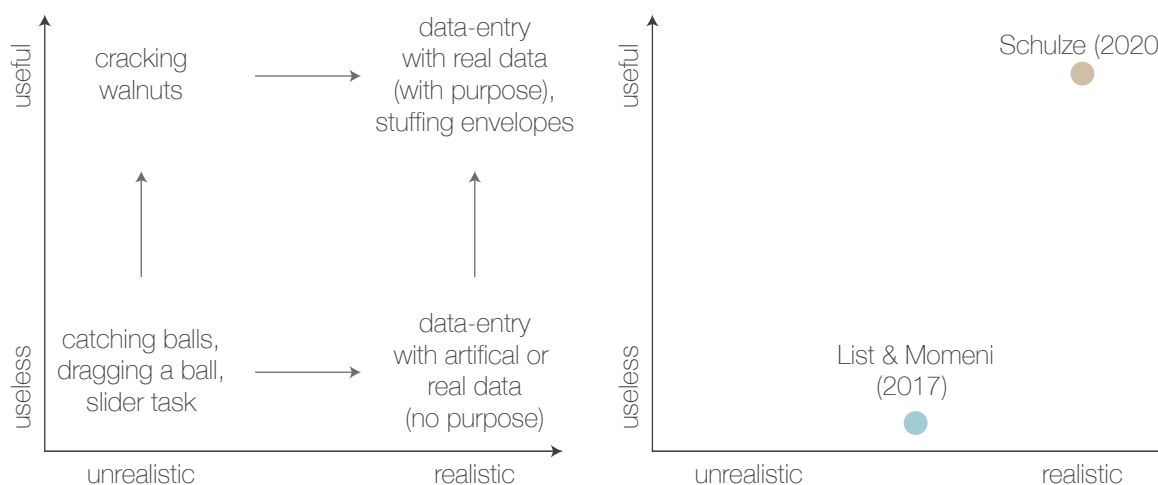


Figure 1.3: **Realism and usefulness of output in real-effort tasks:** A) tasks can be differentiated in particular along two dimensions. *Realistic tasks* closely mirror actual work tasks. *Useful tasks* yield an output that has a value outside of the laboratory. The “cracking walnuts” task of [Fahr & Irlenbusch \(2000\)](#) does indeed provide valuable and delicious results. However, it is rather uncommon that the typical subjects (university students) would take on a similar job to earn a living (at least in most Western countries with experimental laboratories). In contrast, data entry and stuffing envelopes represent standard student jobs or even office tasks. The “dragging a ball”-task from [Heyman & Ariely \(2004\)](#), the ball-catching task from [Gächter et al. \(2016\)](#), and the slider task from [Gill & Prowse \(2015\)](#) are both useless and unrealistic. B) To demonstrate the importance and relevance of these task dimensions, consider the results of [List & Momeni \(2017\)](#) (in petrol blue), which could not be reproduced by [Schulze \(2020\)](#). To show the presumed effect (CSR increases employee shirking), [List & Momeni \(2017\)](#) decided on a task that was relatively easy to implement on Amazon’s online platform Mechanical Turk and was, to a certain extent, realistic – but above all *useless*. [Schulze \(2020\)](#) (in brown) chose a task that had a much closer connection to the research question. The author finds that the shirking/cheating results of [List & Momeni \(2017\)](#) cannot be confirmed if one employs a task that is not only (more) realistic but also *useful*.

1.2.1 Realistic vs. Unrealistic Tasks

The advantage of a realistic task over an unrealistic one is i) that it genuinely corresponds to an existing work task, and ii) that the cost function may capture crucial aspects of the actual workplace. Through a more authentic work environment, mundane realism is greatly increased. This major advantage, at the same time, embodies its most significant limitation and drawback: implementing a true work task in a laboratory environment may increase external validity; however, generalizability is reduced as any results obtained may hold only for the specific task employed in the study (see also [Dutcher et al., 2015](#)).

Unrealistic tasks, on the contrary, do not resemble a real work situation very closely, such that the effort-cost function most likely does not capture the field appropriately. External validity is not given, and results may bear limited expressiveness beyond the laboratory. As [Dutcher et al. \(2015, p. 3\)](#) stress, one would have to back up the tacit/implicit claim that the effort-cost function of an abstract, unrealistic task involves similar costs as an actual task at the workplace (this is rarely done in the literature as exemplified by the authors).⁵ However, generalizability across tasks may not be so restricted, i.e., the results may not be so rigidly confined to a particular task. In principle one would otherwise have to expect heterogeneous behavior and hence observe very diverging results. This, therefore, means that generalizability is strongly related to the design properties of the individual task.^{6,7}

1.2.2 Useful vs. Useless Tasks

When comparing useful and useless tasks, one can expect that the former are more likely to generate psychological effects such as feelings of attachment or entitlement and trigger experimenter demand effects in the form of active participation.⁸ Despite their appealing intuition, such frequently expressed claims of an “emotional connection to effort choices” remain unsupported assertions until evidence is provided ([Dutcher et al., 2015, p. 4](#)). To investigate the impact of usefulness and realism of a task on experimental results, [Dutcher et al. \(2015\)](#) compare two types of real effort (useful and useless) and stated effort. With an elegant design, the authors ensure that effort costs are comparable across treatments, while utility functions may differ. In the treatment with the realistic, *useful* task, subjects are informed that the business data they enter serves a research project. In the treatment with the realistic, *useless* task, study participants record the same Reuters data but without any further information. [Dutcher et al. \(2015\)](#) observe the same outcomes across treatments. The authors

⁵As example, [Benndorf et al. \(2014\)](#) employ the rather unrealistic “counting numbers” task. Subjects are given a lattice with zeros and ones and their task is to count the frequency of the number one (i.e., add all ones). Referring to the task, [Benndorf et al. \(2014, p. 3\)](#) state that “the task is tedious and may thus adequately resemble work effort.” The authors leave open, however, if everything that is tedious resembles work, or whether every form of work is indeed tedious.

⁶[Carpenter & Huet-Vaughn \(2017, p. 17\)](#) survey the literature of real-effort tasks from 1997 to 2016 and conclude that most experimental studies employed unrealistic tasks. They strongly argue in favor of employing realistic tasks to enhance the realism of laboratory studies. The authors further emphasize that the step from an unrealistic to a realistic task is often small, at least to give the task a putative goal (p. 10).

⁷See Chapter 2 for a closer examination of the design properties of real-effort tasks.

⁸Conversely, useless tasks can induce corruptive behavior such as fraud in some individuals ([Schulze, 2020](#)).

conclude that *if any* differences in behavioral responses are observed across different approaches and types of effort provision, these can most likely be attributed to differences in the cost-of-effort functions and timing of decisions – and “not [to] the nature of the effort itself” (Dutcher et al., 2015, p. 13). Regrettably, Dutcher et al. (2015) do not indicate whether the subjects are provided with an outside option. Lacking any alternative activity, subjects may engage in active participation, i.e., complaisantly obey and perform the task. Moreover, if a task produces an outcome that *appears to contain a purpose*, some subjects will complete the task irrespective of the remuneration. In each of the two real-effort treatments, the same financial data had to be entered. In both cases, the data most probably conveyed the impression that they were important and useful for the researchers. Put differently, the *purposelessness* of the task in the *realistic, useless real-effort task treatment* was not sufficiently salient such that subjects believed they completed a task that contained a meaning.⁹ If one were to compare a *mind-numbing, useless, unrealistic task* and a *captivating, useful, realistic task*, greater differences in experimental results would be expected (see also Schulze, 2020 for a similar task comparison).

Conversely to task application area one and two, useful and realistic tasks are employed in task application area three as elaborated subsequently. The task’s usefulness and realism must become sufficiently salient to the subjects from the instructions such they can fully grasp and consider it in their decision making.

1.3 A Classification Based on the Required Subject Qualities: Skills and Personality Traits

Beyond the attributes of real-effort tasks described above, there is another, more subtle dimension along which they vary: Tasks can require different skills and personality traits to complete them successfully.¹⁰ This dimension is orthogonal to the classification of tasks described earlier. It extends the scope without restricting the former differentiation. To shed light on this further dimension, a

⁹See Smith (1982) and Smith (2010) for a related discussion of salience in terms of rewards in economic experiments.

¹⁰The influence of skill bias has also been recognized for example by Gill & Prowse (2015). Lezzi et al. (2015) observe that anxiety, risk and gender have varying effects on different tasks.

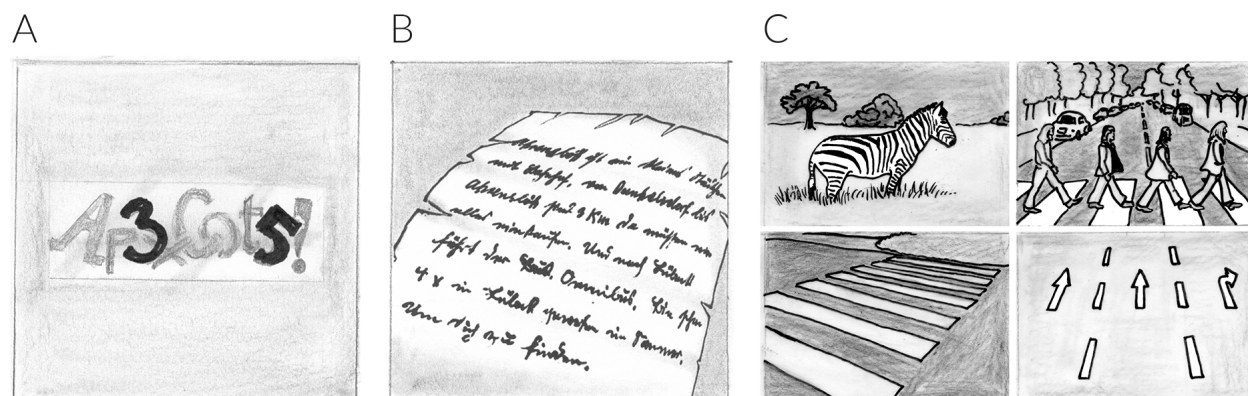


Figure 1.4: **Salience of the usefulness of the output and the degree of realism of a task:** For some tasks, their *usefulness* and *realism* – as well as their *salience* – can be adapted to the respective research question. For example, text recognition tasks can involve A) transcribing computer-generated text that is difficult to read (e.g., random word-letter combinations in “CAPTCHAs,” as in [McMahon, 2015](#)), or B) transcribing passages of poorly preserved and hardly readable historical literature (cf. [Augenblick et al., 2015](#)). The one task is immediately perceived as *artificial and useless*, the other as *genuine and useful*. Fig. C demonstrates that this does not necessarily have to be so obvious. In this image recognition task, subjects have to check whether images actually contain specific information (crosswalks). Since it requires little skill, it can serve as a simple visual task for [task application areas one and two](#). But one can also use the task to support, test, or confirm the pattern detection of artificial intelligence-based algorithms in order to train them. However, this purpose is not necessarily clearly evident. Thus, when the task is used in [application area three](#) to this end, the task’s usefulness and realism must be made sufficiently clear to the subjects by the instructions so that they can fully grasp it and take it into account for their actions.

comprehensive review of the experimental literature on real-effort tasks was prepared (available on request). The survey includes both studies that employ *computer-based* and *non-computer-based* tasks. During the screening of the tasks and their implementation, the skills and personality traits required of the subjects were compiled.¹¹ The extent to which they are needed varies from task to task. For example, some tasks may require mental abilities to perform sophisticated calculations, solve word grids, or memorize images or shapes. Other tasks do not involve cognitive load but demand physical fitness and agility or even vivid creativity.

Tasks may require self-control to resist bodily or mental impulses or external temptations, but also in the sense of required willpower to complete a task. Several causes can profoundly affect and un-

¹¹The compilation is partially inspired by the *Cattell-Horn-Carroll theory* on the structure of cognitive abilities ([Schneider & McGrew, 2012](#)). Moreover, the distinction between skills and personality traits is not entirely straightforward for some dimensions. However, a precise decomposition is not of further relevance for the subsequent treatment. Henceforth, they are jointly referred to as *subject qualities*.

dermine the determination of a subject. Possible triggers are, for instance, pauses in the course as well as the monotony of the task itself. If pauses for reflection are possible, the temptation to think about the sense of the task increases (“what is the point of it all?” “do I really want to do this?”). Tasks differ in the degree of automation and thus in the necessary capacity to maintain focus and attention. When the task is so monotonous that hardly any mental effort is required, the wandering mind likely begins to question the sense of the task.¹² Furthermore, completing a strenuous task for a prolonged period of time calls for grit, persistence, and sustained interest. Tasks that provide feedback at regular intervals, such as an interim result, require less stamina. Finally, personal ambition and performance motivation can spur the provision of effort, regardless of any incentive. Table 1.1 documents the compiled skills and personality traits required for real-effort tasks according to the reviewed part of the literature.

Considering a particular task, one can examine how it relates to the subject qualities just presented. However, for further comparison and classification of tasks, such a detailed analysis is not necessarily required in its entirety. Therefore, five categories of tasks were defined based on the qualities demanded from the subjects. This pragmatic approach serves to provide a better overview of real-effort tasks. Furthermore, the grouping forms the basis for the data analysis in Chapter 3.

Finally, the tasks summarized in the literature review were assigned to the five categories based on the greatest agreement. The classification of tasks according to subject qualities is presented in Table 1.2 for an exemplary selection of tasks.

1.4 Other Classifications of Real-Effort Tasks

In the economics literature, [Charness et al. \(2018\)](#) and [Carpenter & Huet-Vaughn \(2017\)](#) both introduce new taxonomy to distinguish different types of tasks.¹³ In contrast to the approach presented here, the authors concentrate more on the action performed in the task and give far less considera-

¹²However, if the task is “involving,” i.e., the subject is occupied by the task to such an extent that the monotony of the task is pushed into the background, then relatively little self-control is required to carry it out. As an example, weight lifting seems very simple and monotonous, but is so very physically demanding that the former aspect is overshadowed.

¹³The categories presented in this work were developed independently of these authors, not least because of misalignments in approach and interpretations.

Table 1.1: **Subject qualities that are required for a task or that promote success in the task:** Tasks differ along with a variety of dimensions, in particular the extent to which they demand physical or mental abilities and which personality traits facilitate task fulfillment. The listed subject qualities constitute an integral part of the empirical analysis conducted in Chapter 4.

Subject qualities	Description
Physical abilities	<i>Coordination, flexibility and endurance^a</i>
Prerequisites to concentrate and to maintain focus	<i>mental ability and personality</i> to concentrate (self-control, self-regulation, patience, perseverance, self-efficacy), <i>physical capacity and energy</i> to focus; (potentially) challenging properties of the task include interruptions or small pauses during the task, monotony or automatic nature of the task and time pressure
Vigilance	Paying high attention and yet be prepared to observe the unexpected ^b
Cognitive abilities	<ul style="list-style-type: none"> • <i>Short reaction time and decision speed:</i> the immediacy with which a subject must react to a given incentive or task • <i>Quantitative reasoning</i> • <i>Common knowledge</i> • <i>Spatial awareness</i> as well as <i>abstraction and association</i> (e.g., for structure finding) • <i>Short-term memory:</i> the capacity to grasp and remember information in the immediate consciousness and to access and use it shortly afterward • <i>Long-term memory:</i> the capacity to memorize information and to retrieve it (skillfully) later in the thought process • <i>Language fluency and speech comprehension:</i> involves vocabulary and meanings and reading and writing skills
Performance motives	Personal ambition evoking (voluntary) effort provision

^a The tasks used in the literature rarely require physical strength.

^b As an example of “in-attentional blindness,” consider the “invisible gorilla” study by [Simons & Chabris \(1999\)](#), which examined subjects’ capacity to notice surprising events.

Table 1.2: **Skill- and personality-based classification of tasks:** Tasks can be assigned to five different categories, which reflect the abilities and character traits that are primarily required to perform the task well. In Chapter 3, the classification serves to identify a set of tasks that is as diverse as possible. It also enters into the analysis.

Category	Selected examples/exemplary reference
Quantitative and analytic reasoning	Calsamiglia et al. (2013), Dohmen & Falk (2011), Gneezy et al. (2003), Niederle & Vesterlund (2007), Sutter & Weck-Hannemann (2003), Dijk et al. (2001), Vandegrift & Brown (2003)
Language (fluency) and verbalizing	Charness & Grieco (2014), Dickinson (1999), Eckartz et al. (2012), Jones & Linardi (2014)
Memory and knowledge	Erkal et al. (2011), Nikiforakis et al. (2012), Benndorf et al. (2014), Kephart (2017), Winter et al. (2012)
Mechanical	Swenson (1988), Berger & Pope (2011), DellaVigna & Pope (2016), Gill & Prowse (2015), Heyman & Ariely (2004)
Entertainment/game	Augenblick et al. (2015), Gächter et al. (2016)

tion to the required subject qualities. For example, [Carpenter & Huet-Vaughn \(2017\)](#) group all tasks involving *typing* into a single category. This way of categorizing may neglect that different “typing tasks” request dissimilar skills and vary in the demand for cognitive and physical effort. For example, in the “typing task” from [Dickinson \(1999\)](#) subjects repeatedly had to transcribe a long paragraph of text. In contrast, in the “typing task” from [Berger & Pope \(2011\)](#), subjects had to alternately press the letters “a” and “b” on the computer keyboard as fast as possible, which is physically quite demanding. Hence, apart from the very different degree of realism, usefulness, and complexity of the tasks, the difference in physical and cognitive requirements could hardly be more striking. On a similar note, “solving anagrams” may resemble a “puzzle” ([Carpenter & Huet-Vaughn, 2017](#)) or a “creativity task” ([Charness et al., 2018](#)). However, more striking and characteristic is the intense focus on language and verbalization skills, which are required of the study participants and distinguish the task. In summary, [Carpenter & Huet-Vaughn \(2017\)](#) and [Charness et al. \(2018\)](#) introduce classifications of tasks that appear somewhat arbitrary and are based on a blend of task types, task properties, and required skills.

The use of tasks in the field of psychology goes back a very long time. At the beginning of the

20th century, will psychologists (e.g., N. Ach) as well as labor psychologists used various types of tasks. For example, Kurt Lewin used a classification for his work in the 1920s, which differentiated tasks as follows: Tasks with a pre-defined, specific end (“*terminal actions*”) were distinguished from those that were of an endless nature (“*series actions*,” e.g., winding an infinite number of beads on a string). The former could be oriented towards the solution of a given problem (terminal actions *with problem character*, i.e., solving puzzles) or not (terminal actions *without problem character*, e.g., painting a checkerboard pattern). In addition, tasks were evaluated according to their productiveness, as is also done in the present work. For further evaluation and differentiation of tasks, a closer examination of the psychological literature is worthwhile and highly recommended.

1.5 Choose a Task That Suits Your Research Question

While Carpenter & Huet-Vaughn (2017) strongly argue in favor of realistic tasks with useful output to achieve mundane realism, Dutcher et al. (2015) warn and remind that the results are restricted in their generalizability. Rather than field relevance, the nature of the cost function should be decisive for the choice of effort measurement. Therefore, real effort may only prove beneficial if the cost function induced in stated effort cannot accurately capture the crucial features of the actual work setting. “The real effort task must be therefore chosen due to it possessing properties very similar to the field situation of interest and it must also be made clear the domains to which those results do and do not apply” (Dutcher et al., 2015, p. 14). To reframe and extend the argument of the authors: One thing is whether the *task is realistic*, whether the *produced output is useful*, and whether the *task requires specific skills or favors particular personality traits*; another thing is whether the task is *compatible and aligned with the research question*.

Suppose it is of minor importance for a given research question to replicate a real workplace. In that case, an *unrealistic* task seems more suitable since these are usually easier to implement in the laboratory and are more time-efficient. In fact, a realistic task may be undesirable in some instances, as it would entail too many unintended side effects (consider applications 1 and 2 discussed in the introduction). Moreover, unrealistic tasks nevertheless generate a certain feeling of entitlement towards an earned endowment due to the actual effort involved, which is a genuine advantage compared to

stated effort (cf. Dutcher et al., 2015). If the task is also *useless*, this would further mitigate the impact of non-monetary incentives on effort provision, including experimenter demand effects.

If a research question requires that mundane realism is achieved, a *realistic* task with *useful output* serves the purpose (see application 3). Ideally, the task is closely resembling a real work situation and requires the appropriate skills and personality traits. It may even incorporate an actual job in the laboratory. Notably, the task must correspond to the working environment that is of interest. For example, clerical tasks may be applicable to model white-collar workers (e.g., data base search as in Linardi & McConnell, 2011; or data entry tasks as in Corghnet, Hernán-González, Kujal, et al., 2015; Gneezy & List, 2006; and Hennig-Schmidt & Sadrieh, 2010; whereby the effort made can even raise funds for a charity as in Charness et al., 2016), whereas mechanical tasks could be used to mirror labor of blue-collar workers (e.g., stuffing envelopes as in Falk & Ichino, 2006; Hennig-Schmidt & Sadrieh, 2010; Konow, 2000; and likewise to the benefit of charities through fund-raising mail campaigns in DellaVigna et al., 2016).

In an exaggerated presentation, one could distinguish between two extreme cases: *general* and *specific* research questions. For each of these cases, there are more or less suitable real-effort tasks.

Tasks to address *general research questions*

- should be unrealistic;
- must have a useless output (else they may induce intrinsic motivation);
- must not demand any particular subject qualities (otherwise, there is heterogeneity in effort costs, which results in skill bias, for example).

Tasks that fulfill these properties are very *generic* and mostly outright simple.¹⁴ They are suitable in the [first and second task applications](#) discussed in the thesis introduction and allow for *neutrality* and an unbiased measurement of effort. This task type could be especially helpful and valuable for basic

¹⁴In the introduction of the thesis these tasks were referred to as *neutral* tasks. With the broader background presented in this chapter, the concept could be refined and lead to the term *generic* tasks, which is more precise in some respects. The term *neutral* tends to emphasize the independence of subject qualities and the interchangeability of tasks. Conversely, the term *generic* stresses the lack of realism, the uselessness of the generated output and the suitability for general research questions. Both terms have their advantages in their respective contexts. From now on, they are used synonymously, whereby the term is chosen in each case that more strongly highlights the task properties of interest for the particular situation.

research to fathom or prove postulated general mechanisms.

Specific research questions are certainly more “task-dependent,” and require a task that is specifically tailored to the use case. Such *specific tasks*

- must be realistic;
- should have a useful output;
- the demand in subject qualities must suit the topic of research (the task must mirror any heterogeneity in effort costs or bias that exists in the real environment).

These specific tasks need to reflect the object of study truthfully and accurately in all essential aspects and, therefore, are mostly much more complex. Such tasks are well suited for the [third task application](#) and enable *realism*. This means that by congruence with reality, the context and the resulting sensitivities are captured, which actually affect the provision of effort (see [Dutcher et al., 2015](#) for a discussion of matching cost functions). Both types of tasks are illustrated in Figure 1.5.

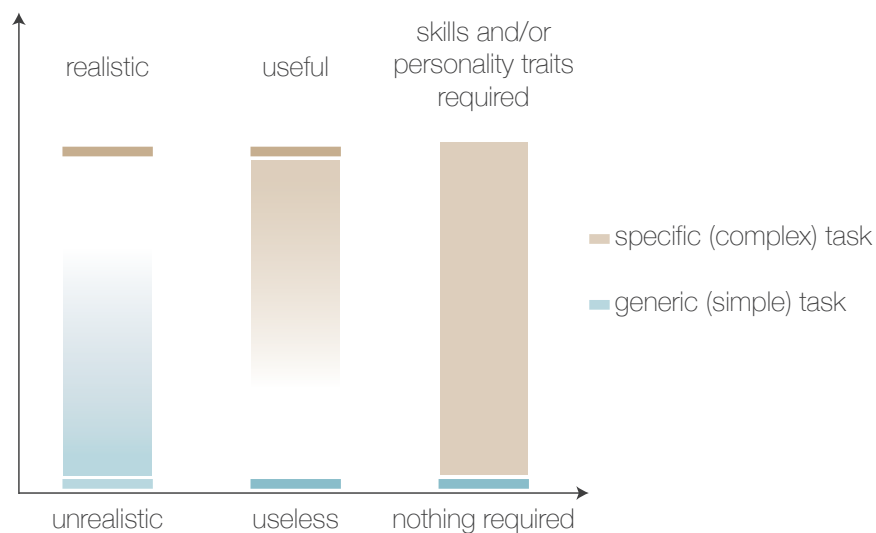


Figure 1.5: **Specific and generic real-effort tasks:** For *general research questions*, (simple) [generic tasks](#) permit for *neutrality* and an unbiased measurement of effort. In the case of *specific research questions*, (complex) [specific tasks](#) allow for *realism*, i.e., reference to and proximity to reality, which in turn allows capturing the context and resulting sensitivities that influence the provision of effort. For this reason, they need to be tailored or chosen to suit the respective research question addressed.

In any case, a task’s properties must be (clearly) evident so that the study participants can understand the meaningfulness of the task and become truly aware of it. Only if these are well communicated

do they influence the subjects' behavior in an intended way and to the intended extent – or not. The extreme cases presented, i.e., *generic* and *specific tasks*, serve to illustrate and contrast this issue, while in practice, a continuum (of tasks) exists along the different task properties.

Since the properties of the task must above all correspond to the research question to be addressed, it is quite conceivable that deviations from the two extreme cases mentioned are necessary. To achieve such a matching, [Eriksson et al. \(2017, p. 627\)](#) deliberately seek a task which does *not* contain any ability-bias: “Using a task in which a low performance would signal low cognitive ability would have generated more embarrassment in case of exposure, which we wanted to avoid for ethical reasons.” Similarly, [Niederle & Vesterlund \(2007, p. 122\)](#) examine gender differences in the decision to enter a tournament and, therefore, look for a task in which effort provision does not differ significantly across incentive schemes.

The choice of task, whether generic or specific, defines and understandably confines the scope of the results obtained. If the scope is to be extended, it is reasonable to repeat the laboratory experiment with a more generic or specific task, respectively. In this way, it can be shown that the results obtained can be easily generalized and can be sustained beyond the specific circumstances (see [Table 1.3](#)). However, researchers may also find that the opposite is true and hence showcase the limitations of prior results.¹⁵ Reporting such results is, therefore, very much needed and essential to better understand and place the robustness and scope of theories and experimental results.

Nevertheless, the framing of the study should not be neglected either, i.e., the embedding of the task in the experiment.¹⁶ A *concrete framing* can further increase realism, but at the expense of generalizability; However, *general framing* can be at the expense of making the experiment more difficult for subjects to understand. Beyond the argument of [Dutcher et al. \(2015\)](#) of a matching cost function to establish a connection between theory and experiment, a matching utility function to incorporate the context of the research topic would, therefore, be appropriate.

In conclusion, a study's research question primarily determines the choice of a task. Suppose the type of effort modeled has a direct effect on the behavior of subjects. In that case, the choice of the task

¹⁵Reconsider the study from [Schulze \(2020\)](#) who could not reproduce of [List & Momeni \(2017\)](#) when employing a useful task. Moreover, [Lezzi et al. \(2015\)](#) find that fear, risk and gender have quite different implications for different tasks. This is not surprising as [Dutcher et al. \(2015, p. 5\)](#) point out, since “different cost functions yield different results.”

¹⁶Consider also the earlier remarks on salience in economic experiments.

Table 1.3: **Generalizability of the results obtained:** The repetition of a laboratory experiment with a more generic respectively more specific task permits to investigate whether the results are only valid under certain circumstances or whether they can be generalized.

Tasks in Experiment 1 and 2	Confirm results?	Findings of Experiment 2 in relation to findings of Experiment 1
Generic task \Rightarrow specific task	Yes	Confirms the results also for the task tailored to the research question: increases/underlines the robustness of the results
	No	Indicates a false-positive of previously obtained results for the generic task
Specific task \Rightarrow generic task	Yes	A generalization of the results, which were collected for the task adapted to the research question, is possible
	No	It is not possible to generalize the results that were obtained for the task tailored to the research question: The scope of the findings is, therefore, limited

is crucial, and it is, therefore, appropriate to restrict oneself to tasks that derive from the environment one wishes to model (see also [Dutcher et al., 2015](#)). As [Cooper & Kagel \(2016, p. 274\)](#) point out "..., while it is clear that subjects in a laboratory choosing numbers to represent effort are performing a substantially different task than a worker planting trees in British Columbia, it is not so clear that one of these cases is closer than the other to the situation of stock-brokers working in a Boston office. All settings have specific elements which may affect behavior."

Referring back to the [three areas of application of tasks](#) discussed at the outset of this thesis, one can conclude: If it is essential to capture the field environment, implementing a realistic task with an output of relevance outside of the laboratory is helpful; if however mirroring the field is not an issue, it is sensible to employ an unrealistic task that has no useful outcome, to mitigate side effects. Only then should aspects such as practical implementability, comprehensibility for study participants, and relation to the literature play a role.

The remainder of the thesis focuses on this latter type of tasks, i.e., *unrealistic, useless* real-effort tasks. The next chapter proposes a set of "design criteria" that are derived from the literature to facilitate the development, selection, and implementation of tasks. Although not further specified,

the recommendations presented also apply in general to more realistic or useful tasks, perhaps even more so for these.

2

Developing Real-Effort Tasks

2.1 Introduction

The previous chapter introduced the two common approaches to studying effort provision: *chosen effort* and *real effort*. Both approaches have advantages: For chosen effort, experimentalists stress its *greater level of control*, which is of particular importance when testing theory; in real effort, the *genuine effort* provided allows for *greater realism* and better reflects the real world of work (Carpenter & Huet-Vaughn, 2017; Charness et al., 2018).

This chapter aims to reconcile the trade-off between experimental control and realism. To this end, a set of *design criteria* is proposed to address the shortcomings observed in tasks. To meet the criteria and achieve task improvement, a set of *design practices* is presented. Finally, a novel task designed according to the criteria is offered as an example of applying these practices.

In order to draw attention to the scope and potential consequences of shortcomings of tasks, reference is made to shortcomings discussed in the literature. They involve effort costs on the one side and output production on the other. The later proposed design criteria and practices frequently take the shortcomings as direct starting points.

Completing a task is associated with effort. However, the reasons for making an effort can be manifold. Therefore, the next section builds on the findings of motivational psychology to explore why subjects exert themselves to complete a task. The insights gained will help to guide and structure the subsequent investigations of this and later chapters.

To find remedies for the discussed shortcomings, a representative part of the literature on *real effort* was sifted. As a synthesis thereof, a set of design criteria for tasks is proposed. To address these criteria and overcome the shortcomings identified, a set of design practices to improve tasks was compiled. These break down into *task-dependent* and *task-independent practices*. The former directly target the inherent properties of a given task; the latter are more general in nature and can be applied to any measurement of real effort regardless of the choice of task.

Carpenter & Huet-Vaughn (2017, p. 5) point out that, “*chosen effort experiments* [emphasis added] seem «cleaner» and more «powerful» in a modest sample than in real-effort experiments where the cost-of-effort function is not assigned and underlying parameters for ability and other possibly confounding factors can vary by worker.” Several measures have been suggested in the literature to this end. However, there has been neither a consolidation nor a comprehensive discussion of these measures. This chapter aims to fill this research desiderata and aims at making *real-effort task experiments* «cleaner» and more «powerful».

The design criteria and accompanying comprehensive set of design practices help to enhance control over the individual cost-of-effort function and the output-production function. They further support the development of new tasks, the selection of a more suitable task, and the implementation of tasks in *various applications* generally. To give an example of a task that builds on the design practices introduced and meets the criteria, a novel task is presented. The task involves repeatedly moving a single slider from one side of the computer screen to the other and is, therefore, called the *single-slider task*. It is utterly boring, very tedious, and meaningless for both the experimenter and the subject. Designed as a neutral task, the single-slider task is well suited for the *first and second task application*

mentioned at the beginning of this thesis.

The remainder of the chapter is structured as follows: Section 2.2 discusses *shortcomings of real-effort tasks and their consequences*; Section 2.3 examines, from the perspective of motivational psychology, why subjects make an effort to accomplish a task; Section 2.4 presents a synthesis of the literature on real effort in the form of a set of *design criteria* and *design practices* to provide pathways for improving tasks; Section 2.5 offers a task, the *single-slider task*, designed to comply with them; finally, Section 2.6 concludes and discusses the contributions of the chapter.

2.2 Shortcomings of Real-Effort Tasks

With regard to the *applications of tasks* discussed at the beginning of this thesis, the literature lists a multitude of shortcomings. These are particularly evident and plausible for application one, where the goal is to measure [a pure treatment effect on] effort provision unbiased by any subject characteristics in a highly stylized setting. For illustration, suppose an additively separable utility function that increases in (monetary) compensation for produced output and decreases in effort cost. Each of these components is examined below.

Real-effort measurements go along with a *significant loss in control over the effort-cost function*, for several reasons. *First* of all, individual effort provision may depend on both skills and personality. For example, a task might require particular cognitive capabilities such as linguistic or mathematical skills and analytical thinking, resulting in a clear *ability bias*.¹ At first glance, physical tasks may seem less prone to such distortions; however, they may require specific motor skills and stamina, which also favor certain individuals.

¹Ariely, Gneezy, et al. (2009) compare two different stake sizes for two tasks, one mentally demanding (*number-adding task*), one physically demanding (*pressing-keys task*). In the cognitively demanding task, the authors observe very frequent poor performance for the larger stake size (“choking under pressure”), but much less so in the simple physical task. Interestingly, those that had the greatest performance increase in the number-adding task choked the most in the pressing-keys task. Ariely, Gneezy, et al. (2009, p. 463) state that the “individual level variation suggests that the factors leading to choking under pressure include not only individual characteristics, but also task-specific characteristics.” Regarding the likely relevant subject characteristics, see e.g., Berger & Pope (2011), who finds that self-efficacy drives task performance in the pressing-keys task. Other tasks, in turn, have different properties and require quite different subject qualities (skills and personality traits). For example, for a number adding task, Bartling et al. (2009) observe that subjects with lower risk aversion or greater self-confidence, along with those who possess the skills required for a task, self-select into a competitive treatment.

Second, if not carefully designed, repetitive tasks can be susceptible to *learning effects* in that within-subject experimental designs suffer from an individual-specific bias (Araujo et al., 2016; Benndorf et al., 2014). Randomization of treatments in between-subject designs permits mitigating such learning effects. However, the first noted individual-specific fixed effects due to variation in abilities may better be countered in within-subject designs. This presents the researcher with a diametric choice of how best to design the experiment.

Third, there may be subjects participating in the study who find value in a task and, therefore, voluntarily exert effort. In the literature on motivational psychology, such engagement in an activity due to the activity itself is attributed to *activity-related incentives*, which are rooted in the execution of the activity (see the discussion in the next section and Figure 2.1). If subjects enjoy performing a task, it can undermine extrinsic motives and make it difficult to determine their impact on effort provision (Erkal et al., 2017).² In an extreme case, subjects perceive the task as so enjoyable that they will not respond to the experimentally induced incentives but exert full effort regardless (Araujo et al., 2016). On the contrary, if the task is too grueling and frustrating, subjects may be unwilling to provide any effort (Araujo et al., 2016) (in the motivational psychology literature, this is referred to as a *volition deficit*).

In the study presented in Chapter 3, 92% of the participants reported that a main reason to complete the given tasks was money. However, 12 out of 248 subjects viewed this quite differently and indicated that this was less or not the case for them. Furthermore, 96 out of 248 subjects confessed that they wanted to meet the expectations of the experimenter. This underlines the role of experimenter-demand effects as a *fourth* confounding factor. Zizzo (2010, p. 75), these describe “changes in behavior by experimental subjects due to cues about what constitutes appropriate behavior.” To what extent they influence experimental results is still debated in the literature. There is evidence that subjects do not care about the welfare of the experimenter (Frank, 1998) or reciprocate benevolence of the experimenter (Abeler et al., 2011).³ In contrast, in studies that compare

²In the experimental study presented in Chapter 3, 116 of 248 subjects (47%) moderately to strongly agreed that the completion of seven tedious and demanding tasks was “fun.”

³In Frank (1998) the experimenter credibly announces that he will burn any payoffs not realized by subjects in an ultimatum game. However, receivers do not accept lower thresholds and payoffs are, therefore, annihilated resulting in a cost to the experimenter. Abeler et al. (2011) study reference-dependent preferences in a real-effort experiment and do not find evidence for reciprocal behavior towards the experimenter in a control treatment with an additional lump sum payment.

different incentive systems, subjects make a non-negligible effort, even if they are remunerated by a lump sum payment. In one of these studies conducted by [Mascllet et al. \(2015\)](#) the participants of the fixed wage treatment were asked for their motivation to exert effort. Most subjects selected “I find it normal to expend effort when being remunerated,” indicating a *sense of duty* ([Mascllet et al., 2015, p. 18](#)). Also [Fleming & Zizzo \(2015\)](#) find support for experimenter-demand effects in experiments and further point to social pressure. For an extended discussion of experimenter-demand effects, see [Zizzo \(2010\)](#). A lack of desirable alternatives may also lead to *active participation*, as pointed out by [Lei et al. \(2001\)](#) and [Corgnet, Hernán-González, & Schniter \(2015\)](#).

Taken together, these factors utterly diminish the control of the researcher over the cost-of-effort function. While demographics and measures of evidently task-related abilities are frequently used as controls, experimenter demand, learning effects, and the influence of activity-related incentives are largely neglected.⁴ In fact, curiosity, enjoyment, and personal ambition towards completing a task can significantly impact performance, are to a great deal subject-specific, and might even change during the course of the experiment. This can go so far that the effect of monetary incentives is almost entirely obscured by the presence of non-monetary incentives and is, therefore, hardly observable ([Erkal et al., 2017](#)). Accordingly, using a task that limits such influencing factors is crucial to obtain usable and informative data. Providing ways and means to mitigate the impact of non-pecuniary incentives and to gain control over the effort cost function is the first objective of this chapter and is covered in the first four subsections of Section 2.4.

The next Section focuses on the second component of the utility function and deals with *output production* and the most likely output-contingent remuneration of the study participants. First of all, the effort exerted by the subjects may not translate into output. Such an *inelastic effort response* may occur when the production function is not sufficiently sensitive to (changes in) effort ([Araujo et al., 2016](#)). If the task design or its implementation limits the range of possible effort levels, ceiling effects might be observed. Boundary solutions may further occur independently of task enjoyment if the cost-of-effort is too low or too high ([Araujo et al., 2016](#)). Coined “choking under pressure,” subjects may be incapable of providing effort if the task is too demanding in light of the stake size

⁴Exceptions include, for example, for activity-related incentives [Dijk et al. \(2001\)](#), [Mascllet et al. \(2015\)](#) and [Giusti & Dopeso-Fernández \(2018\)](#), and regarding learning effects [Benndorf et al. \(2014\)](#) and [Araujo et al. \(2016\)](#).

(Ariely, Gneezy, et al., 2009; Baumeister, 1984; Pokorny, 2008). Moreover, an incentive effect, which is small with respect to uncontrolled variations in the task (i.e., due to activity-related incentives, additional purpose-related incentives, individual skills, or learning effects), results in *statistical insignificance*. Therefore, the second objective of this chapter is to provide strategies to improve the (sensitivity of the) output-production function.

The implications of the described shortcomings naturally vary for the different [applications of the tasks](#). In terms of application two, having to earn an endowment actively reinforces the subjects' sense of entitlement to it (Carpenter & Huet-Vaughn, 2017).⁵ The authors continue by pointing out that heterogeneous unobserved abilities will lead to differing endowments of the subjects, resulting in omitted variable bias, which, however, can be remedied by appropriate randomization (p. 5). As already noted in the introduction to this thesis, the situation changes if the characteristics of the subjects do not enter linearly. Mere differencing is no longer sufficient to mitigate subject-specific effects. Accordingly, the level of initial endowment *can* influence subject behavior, as in Torgler (2002), who observes that having a higher initial endowment translates into having a greater tax morale.

Concerning application one and three, the following considerations can be made. If the measured effort enters the model as the *dependent* variable, the regression results and hypothesis tests are unaffected. Yet, the coefficient of determination R^2 is likely higher for a task that better suits the application: For [application one](#), a *generic task* which is unrealistic, generates useless output and does not demand any particular subject qualities; for [application three](#), quite contrarily, a *specific task*, which is realistic, generates useful output, and any demanded subject qualities must suit the topic of research. However, if the provided effort enters the model as an *independent* variable, the described shortcomings may translate into substantial measurement errors rendering regression coefficients and hypothesis tests invalid.

⁵This is to diminish “house money” alike effects, whereby subjects behave more benevolent or risky respectively with money they were endowed with by the experimenter than with money they really “earned” (see e.g., Hoffman et al., 1994; Corgnet, Hernán-González, Kujal, et al., 2015; Reinstein & Riener, 2009; R. Thaler, 1985; R. H. Thaler & Johnson, 1990).

2.3 Why Do Subjects Make an Effort to Complete Tasks? – A View From Motivational Psychology

Individuals derive utility from performing a work activity in a real working environment, e.g., enjoyment or adversity. The same applies to the completion of a task in a laboratory setting. The standard model to describe effort provision (initially) contains a utility function separating the (monetary) reward and the cost of effort of the input. For example, given a flat payment scheme and actual effort cost (i.e., the payoff is already fixed regardless of the outcome produced), then subjects should not expend any effort at all. And yet, even under these experimental conditions, a not insignificant amount of effort can be observed (e.g., [Masclét et al., 2015](#)).⁶ The assumption that the subjects' behavior in the laboratory is based solely on monetary incentives, therefore, falls somewhat short. Instead, non-monetary factors may be present that motivate effort provision – apart from pecuniary incentives.

This circumstance is increasingly recognized and taken up in the economic literature ([Bowles & Polanía-Reyes, 2012](#); see, e.g., [DellaVigna et al., 2016](#)). For example, [Carpenter & Huet-Vaughn \(2017\)](#) point out that subjects may exert substantial effort due to intrinsic motivation, a feeling of obligation, or to gratify or please the experimenter. [Benabou & Tirole \(2003\)](#) propose a model for effort provision to underpin and illuminate the matter, distinguishing between *extrinsic motivations*, *intrinsic motivation*, *reputation*, and *(self)-signaling*. Extrinsic motivations are founded in monetary incentives that are in some way brought to a subject from outside. Conversely, intrinsic motivation is related to its inner drives. For the present case, intrinsic motivation appears rooted in an internal drive to perform the specific activity.

However, a problem is that the conception, definition, and usage of the term “intrinsic motivation” have varied considerably over time.⁷ [Rheinberg & Vollmeyer \(2012, p. 153\)](#) give an overview and

⁶[Masclét et al. \(2015, p. 19\)](#) find that “women exert significantly more effort than men under a fixed wage and that such positive effort is primarily motivated by a sense of duty, enjoyment of the task and distaste of boredom.”

⁷[Masclét et al. \(2015, p. 17\)](#) point out that “the notion of *intrinsic motivation* embraces various concepts in the economic literature including self-esteem, pride in one's work, enjoyment of the task, fairness considerations and a sense of duty to honor one's contractual obligations (Deci, 1975; Baron, 1988; Kreps, 1997; James, 2005; Ellingsen and Johannesson, 2008; Kuhnen and Tymula, 2012).” [Rheinberg & Engeser \(2018\)](#) provide an overview of the historical development of the term; see also [Rheinberg & Vollmeyer \(2012\)](#), p. 149-153.

critically note that the reference point for what should be (called) “intrinsic” is continually changing: “Sometimes it is the activity itself, sometimes the theme of the action, sometimes the (object of interest) and sometimes the person or the self.”⁸ Due to its inflationary use, coupled with imprecise conceptualization and multifaceted application, Rheinberg & Vollmeyer (2012) recommend that the respective circumstances be examined more closely, instead of using the vague term “intrinsic motivation” and possibly overloading it with yet another point of view. To better understand the underlying reasons for the subjects’ actions in a given case, one would have to delve deeper and reconstruct the entire motivational process. Following this recommendation, this section takes a side step and introduces methods of motivational psychology, which is concerned with elucidating those “mechanisms that energize and direct behavior” (Rheinberg & Engeser, 2018, p. 581). These will provide a comprehensive basis for disentangling the “plot situation” the subject is confronted with in the laboratory. The insights gained enable a better understanding of what motivates subjects to make an effort in a given task and what driving factors might be involved.⁹

2.3.1 The Advanced Cognitive Motivation Model

Why do subjects perform the tasks they are posed in an experiment? What motivates them to provide effort? In motivational psychology, Rheinberg & Vollmeyer (2012, p. 15) define motivation as “activating orientation of current life pursuits toward a positively evaluated goal state.” The extended version of Heckhausen’s *Advanced Cognitive Motivation Model* of Rheinberg (1989) provides a systematic and methodologically sound way to present and analyze motivational phenomena. The model depicts the general structure of goal-oriented behavior and consists of “the perceived *situation*, a possible *action*, the *outcome* of this action and the *consequences* that the outcome of the action will bring about with a certain probability” (Rheinberg & Vollmeyer, 2012, p. 132).¹⁰ It distinguishes i) an *expectation layer*, ii) the *subjective episode structure*, and iii) an *incentive layer*.

⁸Translation by the author. Rheinberg et al. (2003, p. 3) further remark that the term “intrinsic motivation” is value-charged, which becomes evident once the term is applied to things like *aggression* or *addictive behavior* instead of just for positive things like *motivation to learn* or *interest* (for example, see Rheinberg & Manig (2003) why it is “fun to create graffiti”). In contrast, the terminology used in the following does not imply any value-related connotations.

⁹The approach presented below is in some ways similar to Erkal et al. (2017) and Eriksson et al. (2017), but is much more nuanced and solidly grounded in motivational psychology.

¹⁰Translation by the author.

Incentives enter the model in two ways: As *activity-related incentives*, which reside in the execution of the activity, and through the consequences of its outcome as *purpose-related incentives*. “The strength of a person’s current motivation, i.e., [the] tendency to act, depends on three types of expectancies, as well as on the incentives in place” (Rheinberg & Engeser, 2018, p. 591). These are described in detail below with reference to the situation of the subject in the laboratory.

The following presentation was prepared with strong reference to the motivational psychological considerations of Rheinberg & Vollmeyer (2012), and their essential findings were transferred here to a behavioral economic application. Paraphrases are provided with direct page references in the following form: (R&V, p.page). The complete model is depicted in Figure 2.1.

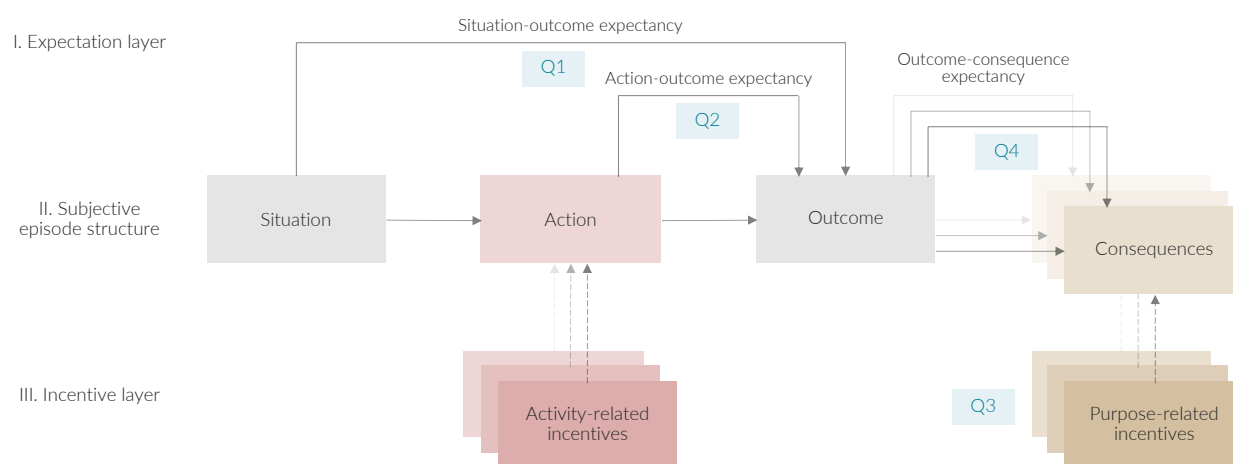


Figure 2.1: **Purpose- and activity-related incentives in the extended version of Heckhausen’s Advanced Cognitive Motivation Model** (adapted from Rheinberg, 1989, p. 104): The model provides a theoretical framework for the systematic analysis of a given *plot situation*. Applied to the present experimental setting, a subject may carry out an **action** (real-effort task) to achieve a desired **outcome** (high score) to reap implied **consequences**. These can include material rewards as well as self-evaluation consequences or other evaluation consequences. Each of them is expected to materialize with a certain probability and carries its own **purpose-related incentive**. An effort is ultimately only made if an outcome can be achieved that is worthwhile given its likely consequences (Rheinberg & Vollmeyer, 2012, p. 140). The action-motivating consequences are reaped only after the activity is completed and the outcome is attained. Besides, a subject may perform an activity by virtue of the activity itself since the task provides pleasure (fun, joy, flow). These incentives residing in the execution of an activity are called **activity-related incentives**. To illustrate the direct relationship between the later described *propositional logical version of the model* (see Figure 2.2), the connection between both models is indicated (Q1-Q4).

Whether motivation actually translates into action depends on whether a desired outcome *can* or

must be achieved through one's own actions (R&V, p. 131).¹¹ In the model, one's assumption about whether or to what extent one's actions *can* lead to the outcome is referred to as *action-outcome expectancy* (R&V, p. 132).^{12,13} In accordance with the concept of the *internal locus of control* of Rotter (1966) it revolves around what the individual has or believes to have under control (R&V, p. 137).

In some circumstances, it may also be that greater efforts are not even necessary to achieve the outcome. In other circumstances, fate seems to have predestined the outcome, so any action would make no difference to change it in any way. This expectation of what will happen regardless of whether one acts or not is termed *situation-outcome expectancy* and is known at least for common situations (R&V, p. 132).¹⁴ Subjects are likely to have such expectations for tasks they have previously encountered in the same or similar form inside and outside the laboratory.

For instance, this would be the case with familiar tasks where one already knows for sure that, e.g., due to preparatory work that has already been done, one's high level of competence, or support from others, the desired outcome will materialize almost by itself without one having to intervene: The situation proceeds effortlessly towards achieving the goal. In terms of measuring effort, this may be the case, for example, with puzzle tasks, general-knowledge quizzes, or word-formation tasks, which are also used in the literature (Winter et al., 2012; Eckartz et al., 2012; Jones & Linardi, 2014; Wozniak et al., 2014). If the solutions to the problems posed by the experimenter are already known to the

¹¹(R&V, p. 131) is used synonymously for (Rheinberg & Vollmeyer, 2012, p. 131) as defined previously.

¹²The subjective expectation of being able to bring about a certain outcome through one's own actions actually comprises two components: First, one must be certain that the action leads to the outcome (aptly described by Bandura, 1977 as the *outcome efficacy* of an action); second, one must be certain that one can perform the action oneself (R&V, p. 137). Thus, one can be quite convinced that a particular action leads to the achievement of a goal, and still not act because one believes that one cannot accomplish it (R&V, p. 138). This can have *motivational reasons*, e.g., because the activity is very repulsive and, therefore, has a negative inherent incentive (see Section 2.3.5 below); but it can also have *cognitive reasons*, e.g., because the person does not know how to act due to a lack of cognitive skills (R&V, p. 138). Finally, it may even have *physical reasons*, e.g., the person is convinced that a certain action will lead to the attainment of the desired outcome, but still does not act because she believes – due to a lack of physical abilities – that she will not be able to accomplish the task (whether this lack of physical abilities actually exists is another matter). In summary, “one must believe that a certain action can bring about an outcome with a high degree of certainty, and must also believe that one can carry out this action” (R&V, p. 139). This differentiation is certainly interesting from a theoretical perspective. Professor Rheinberg revealed in a personal exchange, however, that it is not very helpful and enlightening in order to describe the behavior of people in everyday life: They rarely bother to think about how they can do something if they do not have the necessary skills (personal communication, Nov. 2020).

¹³In research on achievement motivation, this is referred to as “probability of success” (R&V, p. 132).

¹⁴Rheinberg & Engeser (2018, p. 591) provide the example of a traffic light that turns from red to green (desired outcome) regardless of whether one honks or not (action).

subject, little effort is required to complete the task successfully.

The opposite situation occurs when the task is literally impossible for the subject to fulfill, e.g., because the necessary skills or expertise are lacking or impossible to possess. Making an effort appears to be pointless since the task's outcome, i.e., a deficient performance, seems as already determined.¹⁵ With regard to commonly used tasks, insufficient language proficiency, writing skills, or mathematical ability could be the reason why a subject finds a task almost insuperable. Moreover, in tasks that contain a certain element of chance so that any effort made may be nullified, a similar scenario could arise. If the subject judges the probability of this event occurring to be too high, it again appears pointless to make any effort since the outcome of the task (low earnings) seems likewise predetermined. In both cases, the subject has a *high* situation-outcome expectancy ("outcome is given"), which leads to a *low motivation* to complete the task.

In contrast, the opposite applies to the *action-outcome expectancy*: the more the subject believes that her effort will actually produce an outcome ("I can achieve a higher score"), the *greater her motivation* will be to make an effort in the task.

The attractiveness of an outcome itself is determined in the model by its *consequences*. In terms of subjects completing tasks in the laboratory, these obviously involve external rewards, mostly in some form of monetary compensation provided by the experimenter. However, building on [J. Heckhausen & Heckhausen \(2018\)](#), possible consequences further include outcome-dependent *self-evaluation* (e.g., feeling of success, sense of pride in one's efficiency),¹⁶ *other-evaluation* (e.g., social recognition by the experimenter or peers), and the *approach to higher goals* (e.g., maintenance of the attitude of always being diligent).^{17,18}

¹⁵A task that is difficult – but not impossible – for the subject to accomplish due to lack of physical or mental ability would "only" lead to a low *action-outcome expectancy*. Nevertheless, it would also give rise to a low motivation to perform the task.

¹⁶The relationship between the implicit motives of classical motivational psychology (e.g., *achievement motive*, *power motive*, *affiliation motive*) and the consequences, which trigger motivational incentive effects, will be discussed below (e.g., given a strong achievement motive, self-evaluation consequences have a high incentive).

¹⁷If the minimum score of previous subjects is given as a reference value, the achievement of this outcome represents some kind of *factual evaluation consequence*, which can also have an incentive effect and thus promote motivation.

¹⁸ Some of these motivation-inducing consequences are also found in the economic literature. For example, the drive to fulfill (believed) expectations is discussed by [Zizzo \(2010\)](#) and termed *experimenter-demand effects* (see also brief description in Section 2.2); for the aspiration to be well regarded by others (*image motivation*, *social recognition*) see, e.g., [Ariely, Bracha, et al. \(2009\)](#), and [Kosfeld & Neckermann \(2011\)](#), and for "positive feeling[s] from doing meaningful work, adhering to a social norm of working hard, or signalling prosociality" (*warm glow*), see e.g., [DellaVigna et al. \(2016, p. 2\)](#);

An outcome can entail several of these consequences. Each of them carries a *purpose-related incentive* that reflects how important the consequence is to the subject. The strength of the incentive of each of the consequences naturally varies from one individual to another. For example, the appreciation of monetary rewards probably depends on an individual's wealth, income opportunities outside the laboratory, upbringing, and socialization. Depending on how homogeneously a subject pool is set up, the motivation-promoting incentive effect can accordingly vary greatly from study participant to study participant (with regard to the [areas of application of tasks](#), this can be both beneficial and detrimental to the researcher's goals).

As in all models in line with expectation-value theory, whether a purpose-related incentive ultimately has a motivational effect also depends on the probability of occurrence of the respective consequence – provided that the desired outcome has been achieved. The various consequences of the outcome each occur with a different probability (if it materializes at all). But instead of actual probabilities, only the “individually expected probabilities” influence the subject's actions. They enter the model as *outcome-consequence expectancy*.¹⁹

How desirable an outcome is, depends on the combination of the importance of its consequence(s) and the probability of its occurrence, provided the outcome is achieved. Put differently, whether a subject eventually does become active depends on how much the subject appreciates the consequences of the outcome (its purpose-related incentives) and the subjective expectation that the consequences actually occur if the outcome is attained (their outcome-consequence expectancies) (R&V, p. 133). Thus, each of the consequences unfolds its own motivation-promoting effect, which together form the motivational potential of the outcome. Applied to the subject in the laboratory, the model suggests that a subject's tendency to act increases the more certain the outcome entails consequences with a high incentive value, and the more this outcome depends on the subject's individual actions rather than resulting from the course of events alone. Thus, to exert effort in a task, a

The desire to do something that is beneficial for others also received a fair amount of attention (*pro-social motivation*) see, e.g., [Ghatak & Mueller \(2011\)](#), [Imas \(2014\)](#), [Charness et al. \(2016\)](#). However, the treatment is certainly more nuanced in motivation psychology, where a distinction is made between the outcome (task score), the consequences (e.g., rewarding), and the individually perceived likelihood of realization of the consequence (outcome-consequence expectancy). The interplay of the latter then determines the strength of the resulting purpose-related incentives.

¹⁹The outcome-consequence expectancy is also referred to as *instrumentality*, which more explicitly describes how much the (desired) outcome is seen as an “instrument” to bring about the (intended) consequence.

subject must be certain that:²⁰

1. no points (the desired outcome) will accumulate in the absence of effort;
2. the amount of points achieved can be sufficiently influenced by one's own actions;
3. the points will certainly have consequences (e.g., points are converted into real currency for payments);
4. the consequences (i.e., payment and any others) are sufficiently important to the subject.

If any of these conditions is not met, the subject will not exert (much) effort. Thus, one can distinguish four qualitatively different forms of motivational withdrawal: Exerting effort is viewed by the subject 1) as pointless, since unnecessary, 2) as ineffective for earning points, or 3) that the points earned have no assured consequences, or 4) these consequences are not considered worthwhile.

2.3.2 A Propositional Logical Version of the Model

Beyond the previously provided approach to the model, the model's *propositional logical form* opens up another perspective. This format presented in [H. Heckhausen & Rheinberg \(1980\)](#) provides a quicker understanding of the model and allows easier applicability of it to individually assess a subject's motivation (see [Figure 2.2](#)). To this end, the three expectancies in the model must be captured as well as the types and incentives of the consequences that the individual associates with the outcome of the action (R&V, pp. 134–135). With the collected information, this version of the model allows predicting when a subject will attempt to complete a task and if not, for what motivational reasons. To illustrate the direct relationship between both versions of the model, the connections to the propositional logical version of the model are indicated in the above depicted extended version of Heckhausen's Advanced Cognitive Motivation Model ([Figure 2.1](#)).

Concerning real effort measurements, the propositional logical version of the model is exemplary applied to a participant, who shall solve a multiplication task. Suppose the subject does not feel in a position to influence the outcome (level of the final score) because it appears to her to be predestined by the situation (Q1). She thus has a high *situation-outcome expectancy* and will not even try to

²⁰Based on the presentation in [Rheinberg & Engeser \(2018\)](#), p. 592.

make an effort. If, on the contrary, the score does not seem to be predetermined, the subject may wonder whether her effort is actually and adequately affecting the score (Q2). In case her *action-outcome expectancy* is high, she assesses there is enough time to complete the mentally demanding multiplication task. Next, the question arises whether the fruits of the labor are worth the effort (Q3: *purpose-related incentives*). In addition to the monetary rewards, the increased self-esteem that comes from having solved four tricky math problems or the subjects' pride in being on a (public) list of top-performing study participants could trigger additional effort. As for the *outcome-consequence expectancy* (Q4), if she is motivated by achievement and then reaches a high score, it will fill her with pride. Based on previous experience in laboratory experiments, the receipt of the monetary reward in return for the points achieved seems virtually certain to her. However, the subject could wonder whether the list of top-performing study participants will actually be made public by the experimenter, thus nullifying the motivational incentive effect of this consequence.

Although the expectancies and consequences depend on the objective conditions of a situation, their interpretation and evaluation are subject to the individual. The preceding exemplary assessments for a subject who is to perform a multiplication task may look entirely different for *another participant* as well as for *another task*. The example illustrates that self-evaluation consequences have a pronounced motivational effect on subjects *with a strong achievement motive* (R&V, p. 136). This also applies to higher goals.^{21,22}

²¹Among the implicit motives, motivational psychology further distinguishes the *power motive* and the *affiliation motive*. These can also play a non-negligible role in real-effort experiments, e.g., when subjects compete with each other in a task or instead jointly make an effort to produce a public good. For both motives other consequences are likely influential, e.g., other-evaluation (R&V, p. 136). In general, further influences of motives are conceivable, e.g., on the different expectancies in the model; however, a connection has not yet been empirically established.

²²To decompose the motivational process entirely in its elements, Rheinberg (2004) developed a scheme for motivation diagnostics that defines different classes of *motivational forms* and *motivational deficits* (for a summary, see Rheinberg & Vollmeyer, 2012). It goes beyond the propositional logical version of the Advanced Cognitive Motivation Model presented in Figure 2.1 and includes incentives residing in activities. The study participants' motivation to make an effort in a task can be captured even more precisely with the help of the diagnostic scheme. In Chapter 4, the scheme is applied to determine subjects' motivation for effort provision in each of seven different tasks.

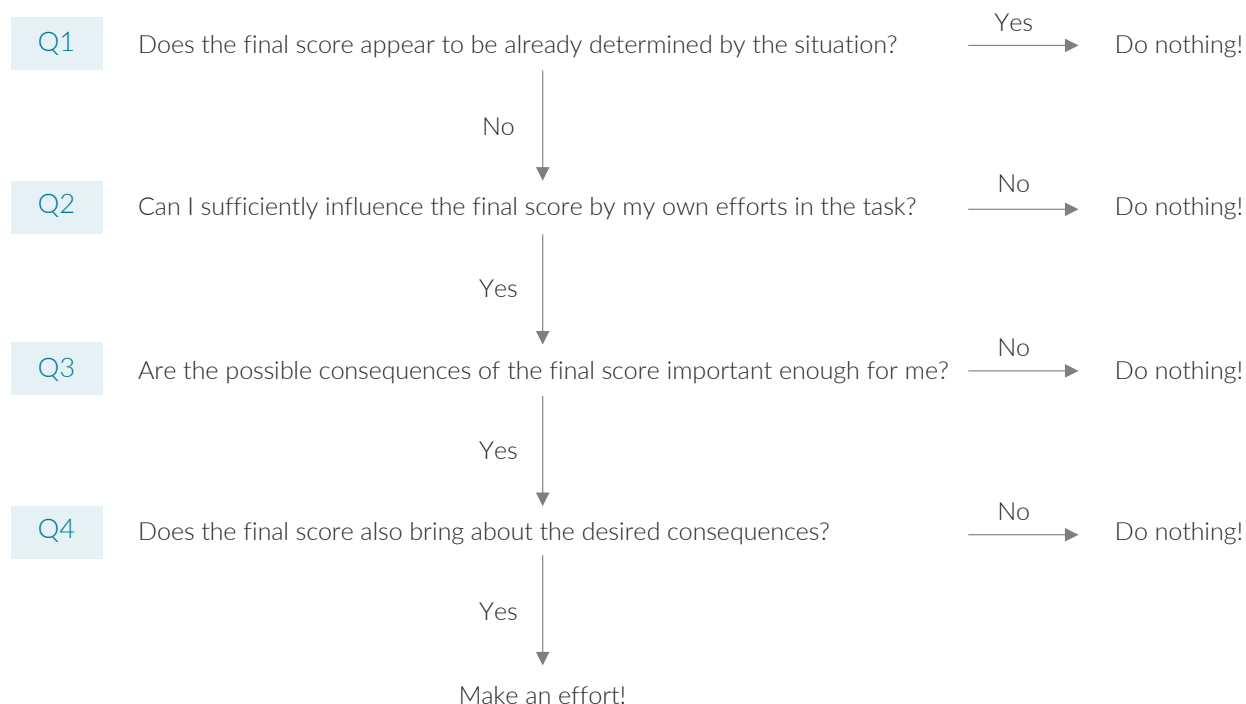


Figure 2.2: **Propositional logical version of Heckhausen's Advanced Cognitive Motivation Model** (adapted from [H. Heckhausen & Rheinberg, 1980, p. 19](#)): Enables an individual assessment of the motivation of a subject to perform a task. For this purpose, several elements of the Advanced Cognitive Motivation Model are queried: the three expectations and the incentives of the consequences attached to the outcome of the action (R&V, pp. 134–135).

2.3.3 Extending the Advanced Cognitive Motivation Model

The elements of Figure 2.1 described so far resemble the original version of the *Advanced Cognitive Motivation Model* of Heckhausen. The therein assumed *purpose rationality* of the motivational structure causes an action to become appealing *only because its outcome entails attractive consequences*. Applied to the situation in the laboratory: The subject only makes an effort because an outcome is possible that seems worthwhile in view of its probable consequences (R&V, p. 140).

According to [Rheinberg & Vollmeyer \(2012, p. 141\)](#), the structure of the plot situation, in which the purpose succeeds the action both temporally and functionally, often occurs in everyday situations. However, people also tend to undertake certain actions because they simply enjoy performing them, regardless of what the outcome and consequences may be. The authors continue that the incentive thereby resides in the *execution of the activity itself* – and not in the consequences of it. As examples

of activities that are not directly related to a specific goal, [Rheinberg & Vollmeyer \(2012, p. 144\)](#) mention certain sports (skiing, surfing, ...), motorcycling, making music, or programming. Characteristic for these activities is, for example, the *pure joy of perfect and harmonious movement* or the *switching off and getting absorbed in the activity*, which is also referred to as “flow state.” Thus, the incentive lies only in performing the activity and not in the consequences of an achieved final outcome (R&V, p. 145). Referring to similar conceptual distinctions from the early 20th century, [Rheinberg \(1989\)](#) labels these *activity-related incentives* and introduces them as an extension of Heckhausen’s model.

In a way, the “purpose of the activity” consists of “feeling good” *during* the activity (R&V, p. 141). To maintain the “in itself” highly rewarding execution state, one would like to pursue the activity “for as long as possible” – rather than merely yearning for its end. (Intermediate) outcomes can even prove undesirable, as [Rheinberg & Vollmeyer \(2012, p. 141\)](#) illustrate using the example of a skier who, after enjoying the downhill run, reaches the valley station and has to queue up in the freezing cold to take the lift back up. In contrast, *purpose-related incentives*, as the authors continue, are effective and encourage efforts even though they can only be realized *after* the activity has been completed and the outcome has been obtained: The fruits of labor can only be reaped when the work is done. Although some of these joyful activities can involve very undesirable consequences (e.g., smoking, base jumping, and wingsuit flying), they are nevertheless pursued with great commitment at times. In contrast, certain activities possess a negative execution incentive but bring about desirable consequences (e.g., brushing one’s teeth). Therefore, some people occasionally tend to avoid them as tackling them requires a fair amount of overcoming (see Section 2.3.5). This can even be the case if they are absolutely convinced of the benefit, necessity, and importance of the action to achieve the outcome and reap the consequences. Nevertheless, the activity is refrained from or at most rudimentarily carried out. Those who approach the activity with less reluctance have to put forth significantly less or no overcoming to get it done.²³ [Rheinberg & Engeser \(2018\)](#) provide a multitude of examples to illustrate the difference between both types of incentives. A subset is summarized in Figure 2.3, which displays all conceivable combinations of them (pure purpose-related activities and pure activity-related activities are shown in the center column and middle row, respectively).

²³[Rheinberg & Engeser \(2018, p. 593\)](#) discuss this in relation to exam preparation, which can be very challenging for some students, even though they are aware of the fruitfulness and usefulness of learning.

		Purpose-related incentives Incentive of outcome consequences				
		Positive	Neutral	Negative		
Activity-related incentives Incentives residing in activities	Positive	① Cooking dinner for friends	⑤ Skiing, surfing, motorcycling, reading	② Over-eating, strong smoking, excessive drug use, extreme sports (high risk)	Activity completion is fun	Positive execution incentive
	Neutral	⑥ Search for the key to unlock the basement and turn on the heating	⑦ (Nothing is done)	⑧ (Nothing is done)	Activity is neither fun nor hassle	No execution incentive
	Negative	③ Doing the dishes, repairing a bike, cleaning the house, ironing shirts	⑧ (Nothing is done)	④ (Nothing is done)	Inherently unpleasant activity	Negative execution incentive
		Outcome of activity has highly attractive consequences	Outcome has neither attractive/unattractive consequences	Outcome of activity has very unattractive consequences		
		Positive consequential incentive	No consequential incentive	Negative consequential incentive		

Figure 2.3: **Purpose-related and activity-related-incentives and their combinations:** The center column displays *purely activity-related incentives* without any consequences (5); the middle row features *purely purpose-related incentives* involving activities that are neither pleasant nor unpleasant (6). A pleasant activity may lead to desirable consequences (1) or undesired consequences (2). Likewise, an unpleasant activity may lead to desirable consequences (3) or undesired consequences (4). If a plot situation contains an action that provides neither a consequential incentive nor an execution incentive, there is no incentive to perform the activity (7). If one of them is not present and the other is negative, there is a negative incentive and the individual will not carry out the activity (8). Example for (3), *Doing the dishes, to have a clean and usable kitchen, to host friends*: The incentive of the consequence is so strong that the aversive activity is (actually) carried out. Example for (6), *searching for the key to unlock the basement and turn on the heating*: The incentive of activity lies almost exclusively in the consequences of the outcome. If both types of incentives prevail, it is not yet empirically established how they interact, i.e., whether they mutually promote or impede each other. The provided examples were mostly taken from Rheinberg & Engeser (2018) and Rheinberg & Vollmeyer (2012).

Rheinberg & Vollmeyer (2012, p. 143) emphasize that individuals are guided both by activity-related and purpose-related incentives, albeit to varying degrees and depending on the context. One might take pictures on vacation to keep the moment in fair memory, fuel one's blog or Instagram feed to impress friends, gather material for one's travel agency's next newsletter to sell trips better afterward, or simply because one is a passionate photographer.²⁴ Consequently, people can perform the same activity for very different reasons, because they are driven by different consequences. Regarding the latter activity-related incentives, Rheinberg & Engeser (2018) remark that they cover a wide spectrum and go far beyond the psychological needs for autonomy and competence that *self-determination theory* identifies for making an activity attractive:

“There is no doubt that both these motivational systems are extremely important. Passionate hobby enthusiasts refer to them repeatedly when interviewed about the incentives that induce them to engage in their leisure time activities (Rheinberg, 1993). However, besides these two, several other incentives also play a vital role. These include the excitement of exposure to risk (e.g., extreme sports or illegal graffiti spraying) or unusual physical sensations (e.g., riding a roller coaster or motorcycling), being at one with nature (e.g., hiking or mountaineering), and so on (Rheinberg, 1993, 1996; Stops & Gröpel, 2016).” (Rheinberg & Engeser, 2018, p. 584).

This also illustrates that the quality of incentive-providing experiences varies greatly across activities, as does the breadth of the respective spectrum of activity-related incentives (see also Rheinberg & Engeser, 2018, pp. 597–598).

For the case that both types of incentives occur in a single plot situation, it is not yet clear how they combine and interact, especially if they carry different signs (R&V, p. 143; cf. the upper right and lower left quadrant in Figure 2.3). Moreover, motivation can change in the course of the activity, e.g.,

²⁴**Examples of purpose-related incentives in photography:** *Keeping the moment in good memory:* Consequence are the feelings of happiness that are triggered when one picks up the pictures of the past summer vacation on cold winter days. Thus, taking pictures is guided by the anticipated possibility of reactivating affectively toned positive experiences in the future by means of a visual stimulus (= the photo taken), i.e., by purpose-related incentives. *Pictures for blog to impress friends:* Consequence is feelings of pride (self-evaluation) about the great blog as well as recognition from friends (other-evaluation) each of which carries an incentive; *Gathering material for the next newsletter:* Consequence is the future income, which provides a material incentive. **Example of activity-related incentives:** *Passionate photographer:* Incentive lies in the execution of the activity itself (positioning and photographing an object or capturing the moment, later post-processing and cropping in the photo lab/at the computer).

enjoyment of an activity (activity-related incentives) can subsequently trigger an incentivizing self-evaluation consequence, in the sense of pride in one's efficacy (purpose-related incentives). However, in laboratory experiments, it is more likely that the external reward initially motivates effort, but that this effort is then perpetuated by emerging activity-related incentives (aroused as the activity is performed).²⁵

As an interim assessment and application, the following remarks can be made: In most research involving real effort, subjects complete tasks to earn money (sooner or later in the experiment). Thus, the plot situation includes a purpose-related incentive in the form of monetary rewarding. Whether activity-related incentives are deliberately part of the experiment or whether an attempt is made to avoid them instead, depends on the [different application areas of tasks](#).

In the case of the second task application, it seems irrelevant at first whether or not the subjects enjoy working on an initial endowment by completing an assigned task. However, it will vary individually whether a subject takes pleasure in the activity or even gets into a flow state. As a result, the subjects will incur different costs for task performing and, therefore, are likely to accumulate varying endowments. The argumentation in the application area one is analogous.

In application area three, the reasoning differs: If the activity, *in reality*, leads to enjoyment or flow among some workers, this must also be considered in the laboratory and reproduced accordingly.

Since this thesis focuses on the first and second application area, the left column of the Figure 2.3 is of primary interest, where the *incentive of the outcome consequences is positive*. Starting from this, Figure 2.4 goes a step further and abstracts from the various possible purpose-related incentives (see Section 2.3.1), and concentrates on the incentives residing in the execution of the activity. The fictitious illustration depicts the intensity of activity-related incentives for six subjects in three types of tasks (*cognitively demanding, physically demanding, entertaining*). The figure illustrates that the strength

²⁵ Another conceivable possibility is that an initial processing of the task is aroused by "curiosity" or "interest." If the focus of the incentive is on a specific object rather than a specific execution component (as in activity-related incentives), one refers to it as *interest* (cf. Rheinberg & Engeser, 2018, pp. 583–584). In the present case, subjects could thus have an interest in mathematical puzzles, read books about them, attend lectures on them, or exchange ideas with others about them in online forums. However, this does not mean that they have to be particularly good at solving mathematical puzzles or that they particularly enjoy solving them. Mathematical puzzles simply interest them. This particular form of motivation focusing on a specific object is of secondary importance for further considerations and will, therefore, not be explored further. *Curiosity*, when satisfied, may result in a state of "reduced uncertainty," which can also provide a certain degree of incentive. However, it is the activity itself, i.e., the state of searching for answers, that drives curious people and is both sought and enjoyed. It is examined in more detail in Section 2.4.1.1 in the context of the design criteria.

of activity-related incentives varies greatly from task to task and for a given task can vary considerably from one individual to another. The tasks of interest for the first and second areas of application either have no or only negative activity-related incentives. However, if the participants perceive the activity as too repulsive, they may lack the necessary will to complete it. To ensure that the tasks fall into this range, additional clarification is necessary. The following thus addresses *activity-related incentives and flow* (Section 2.3.4) and *aversive activities and volition* (Section 2.3.5) in more detail.

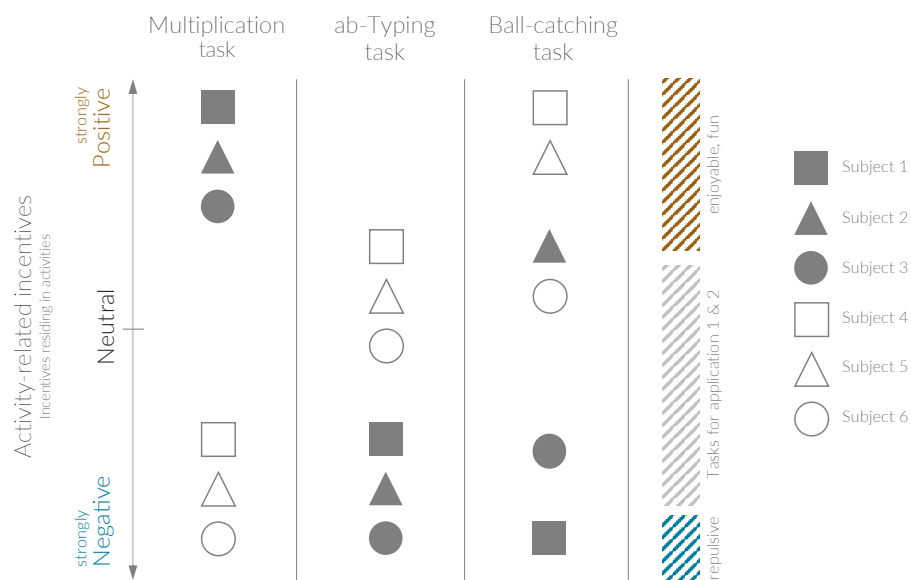


Figure 2.4: **Incentive residing in the activity for exemplary tasks:** The fictitious illustration summarizes the intensity of activity-related incentives for six subjects in three different tasks (the cognitively demanding multiplication task of [Dohmen & Falk \(2011\)](#), the physically demanding ab-typing task of [Berger & Pope \(2011\)](#), and the entertaining ball-catching task of [Gächter et al. \(2016\)](#)). The tasks relevant to the first and second area of application ideally have no or only negative activity-related incentives. Nevertheless, the activity must not be perceived as too *aversive* by the subjects. Otherwise, some may not have sufficient willpower to carry it out.

2.3.4 Incentives Residing in Activities and Flow

Motivation research in this vein has a long tradition. Thus, [Rheinberg & Vollmeyer \(2012, p. 146\)](#) refer to [Bühler \(1922\)](#), who introduced the principle of “functional pleasure” (Funktionslust) according to which activities are performed for their own sake and optimized in their course, especially in playful contexts. They further make reference to [Duncker \(1941\)](#), who discusses the appeal of “dynamic joys” that can be experienced while driving a car or motorcycle, doing sports, or playing certain

games. Although these often (ultimately) lead to some sort of goal or outcome, the authors continue that the incentive in these activities does not rest in their consequences. Instead, fun and enjoyment in approaching the outcome and any challenges that come with it prove motivating, performance-enhancing, and rewarding. Regarding the design of tasks, this aspect seems to be of particular importance for those containing strategic elements and, therefore, generate a certain amount of tension and thrill. Special precautions are necessary to ensure that such tasks are not perceived by some subjects as one of the “purpose-free games” just described.

What could make the state of engaged working towards a not yet achieved performance goal inherently attractive? Taking up this question, [Rheinberg & Vollmeyer \(2012\)](#) apply the preceding considerations to *achievement motivation*. Beyond the joy over success, which is in the foreground in the classical achievement motivation theory, activity-related incentives may evolve by experiencing one's efficient optimal functioning on the way to a challenging goal, which makes one forget space and time (R&V, p. 148). In support of this, the authors refer to research using the intellectually challenging in-basket exercise often used in assessment centers for recruitment of management positions. This research area finds that highly achievement motivated individuals become more absorbed by the task than those with lower levels of achievement motivation ([Engeser & Vollmeyer, 2005](#)). As an example from experimental economics, consider the study from [Araujo et al. \(2016\)](#), who examine the slider task from [Gill & Prowse \(2015\)](#). In preparation for their study, the authors test the task in a trial with ten rounds. There was no monetary incentive for them to make greater efforts and improve their performance. Nevertheless, this simple, meaningless task could arouse a sense of ambition in the researchers [sic.], who report that each of them had an inner, deepest desire to exceed their personal best ([Araujo et al., 2016, p. 11](#)).²⁶

The essence for economic research is that achievement motivation will certainly play a role in real-effort experiments and is hard to prevent in its entirety. It is, therefore, sensible to be aware of these influences and to adapt the study design and the task in such a way that such interfering factors are

²⁶[Araujo et al. \(2016\)](#) draw attention to this issue in the discussion of their study to illustrate the shortcoming of the task of having either a ceiling effect or a production function that is not sufficiently sensitive to variations in the exerted effort. If one continues the authors' thought, the study participants could be equally absorbed in the activity and in surpassing past performance. Any external incentive intended to motivate their efforts would become ineffective in this case.

minimized if necessary.²⁷

The flow state was briefly mentioned earlier. It can be described as a “state of self-reflection-free complete absorption in a smooth running activity,” “in which process and concentration succeed as if by themselves and without volitional effort” (R&V, p. 153/p. 177).²⁸ This implies that if a subject were to enter the flow state during a task, her effort costs would approach zero. Furthermore, greater flow mostly results in higher performance (Engeser et al., 2005; Engeser & Rheinberg, 2008). Interestingly, the flow state is not only observed during attractive leisure activities or challenging work – but even during primitive activities like simple computer games (Vollmeyer & Rheinberg, 2003: *Pacman*; Rheinberg & Vollmeyer, 2003: *Roboguard*). Unfortunately, similar tasks are employed in laboratory experiments to measure effort (consider, e.g., Augenblick et al., 2015; who let subjects perform a simplified version of *Tetris* or the *Ball-catching task* from Gächter et al., 2016). Since flow can severely undermine the effort measurement in the first application area of tasks, it requires further consideration.

In case the activity requires considerable skill to perform, flow is more frequently observed among those who have it. Nevertheless, a significant flow effect on performance is observed regardless of skill level (see Rheinberg et al., 2003; and Engeser et al., 2005). The question arises whether the “matching of ability and demands” described by Csikszentmihalyi (1975) as a central component of flow always promotes it. As it turns out, this is very much dependent on the task and its *consequences*: If scoring on a task does not have serious consequences (games played at home), the balance of ability and demand is conducive to flow; on the other hand, if the performance has serious consequences (e.g., exams during studies), the highest flow rates are achieved when the ability *exceeds* the demands (Rheinberg & Vollmeyer, 2012, pp. 156–157; see also Engeser & Rheinberg, 2008 on moderating effect of (perceived) *importance* of the activity and Figure 2.5B).

Rheinberg & Vollmeyer (2012, p. 158) further emphasize that there are not only differences between subjects as to *whether* there is a balance between their individual abilities and the demands of a task, but also *when* there is a balance between them (see also Figure 2.5A). For the multiplication task

²⁷The following Section 2.4 describes ways to discourage achievement acting for its own sake (e.g., it might prove helpful if the task is tedious, toilsome, tiring, and incredibly dull).

²⁸Translation by the author. The state was first described and characterized by Csikszentmihalyi (1975). For a closer description of flow, see Rheinberg & Vollmeyer (2012, pp. 153–165), who also list its defining components (R&V, p. 154).

of [Dohmen & Falk \(2011\)](#), for example, this implies that not everyone with high mathematical skills necessarily and inevitably enters the flow state when performing the task. To provide an explanation, [Rheinberg & Vollmeyer \(2012\)](#) draw a connection to the risk-choice model of the previously discussed achievement motivation. Depending on the individual level of the achievement motive, there are great differences in the confidence in one's own success, which can have a motivation-enhancing or -inhibiting effect: A match between the demands of the task and one's skills can spur achievement motivated individuals to higher performance and facilitate their transition into flow; in contrast, those motivated by failure become anxious and stressed, so that the flow state becomes virtually unreachable (R&V, p. 158; see [Figure 2.5](#)).²⁹ For individuals motivated by achievement, not only the *pride in one's efficiency* as an anticipated sense of achievement becomes stimulating, but also the *complete absorption in the activity* (R&V, p. 159).

Furthermore, the authors note that a clear *goal orientation* of an activity facilitates the transition into the flow state (R&V, pp. 163–164).³⁰ However, if the goal of one's own actions is not clearly defined and present, attention and processing capacity are required, which distracts and prevents flow (R&V, p. 161).³¹ The authors continue that flow is, therefore, especially observed in activities that are free of interruptions. One might be inclined to say that only qualified experts with the necessary (technical) expertise can perform activities without interruption and thus are able to experience flow. In retrospect to the experiences of subjects during the games *Pacman* and *Roboguard* mentioned at the beginning of the section, flow can *also* be experienced without prior knowledge – albeit just for simple activities. The *degree of matching* between the demands of an activity and one's abilities is what defines whether an activity is *challenging* (R&V, p. 162). How *doable* a challenge is and to what extent one is *willing to accept* it (\Rightarrow level of the performance motive) then determines whether the situation is flow-enhancing and thus performance-enhancing or not. This leads to the final element of this excursion into motivation psychology: *volition*.

²⁹See also [Rheinberg et al. \(2003\)](#) and [Engeser & Rheinberg \(2008\)](#) for empirical evidence.

³⁰Interestingly, feelings of happiness and flow do not directly go hand in hand: Often, feelings of happiness do not arise until after the activity is completed, when the burden of concentration and deep involvement falls off (R&V, pp. 163–165).

³¹[Rheinberg & Vollmeyer \(2012\)](#) thus draw connections between flow and the concept of action control of [Kuhl & Beckmann \(1985\)](#).

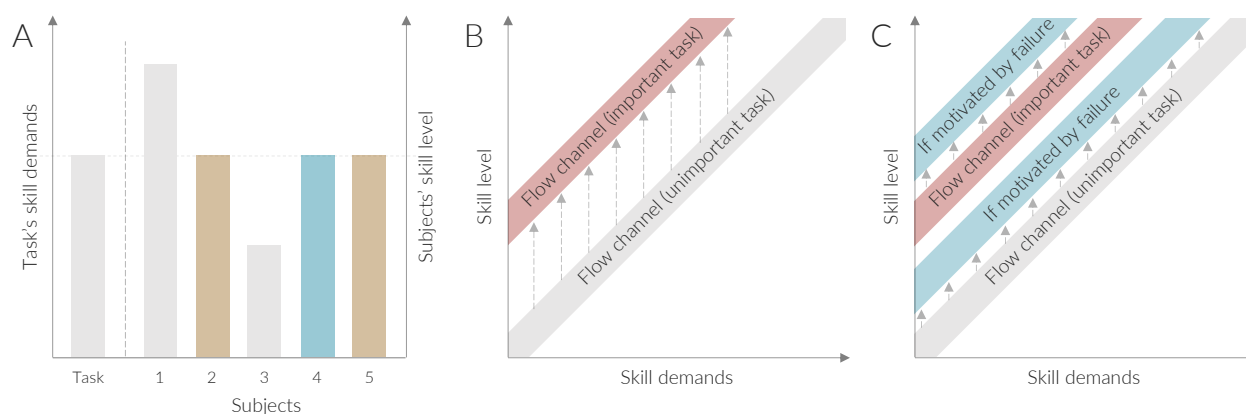


Figure 2.5: **Skill-demand balance and the flow channel model:** The extent to which a balance of demands and skills is conducive to flow depends on several factors, including the *importance* and the *complexity* of the task (cf. Rheinberg & Engeser, 2018, pp. 605–606). (A) With increasing task complexity, the skill demands increase (height of the bar to the left). Subjects 1 and 3 are unlikely to get into flow because they either exceed or do not meet the skill requirements of the task. Subjects 2, 4, and 5 could, in principle, enter flow since their abilities are in balance with the task's demands. However, while the skill-demand balance spurs on subjects **motivated by achievement (2 and 5)**, those **motivated by fear of failure (4)** are overwhelmed and cannot enter the flow state. (B) and (C) show adaptations of the *Flow channel model* (modified after Csikszentmihalyi, 1975): “Importance” here refers to the magnitude of the consequences sought or prevented. Concerning laboratory experiments, this could represent a particularly large reward, for example. (B) If the task is “unimportant” (i.e., a small remuneration), the balance of demands and abilities is flow-promoting (the channel illustrates this condition of a *skill-demand balance*). However, if the task is “important” this is not the case and flow is more likely to occur when the skills exceed the demands. In order to achieve flow in “complex tasks,” the mastery of certain basic skills is necessary, whose execution must first be sufficiently automated (*expertise effect of flow*). Those who do not possess them cannot attain the state. Higher skill, therefore, allows more frequent transition into the flow state, especially for important tasks. (C) **Individuals motivated by failure** further need a certain “capability buffer,” as they become anxious and stressed when there is a balance between demands and skills, making flow virtually unattainable.

2.3.5 Aversive Activities and Volition

In the previous sections, the motivation evaluation process was described, which serves to select the activity to be carried out before becoming active. After thorough consideration and if sufficiently motivated, an *intention* is formed, and the next step is to put the activity into practice. This may be easy for activities one enjoys performing, but it becomes hard for those one does not like very much. The latter may nevertheless lead to desirable consequences, i.e., have important or particularly

worthwhile implications, or ward off dire consequences (R&V, p. 177).³² To perform and endure such *aversive activities* until the goal is reached, *volition* or *will* is necessary (see Figure 2.6 for an illustration). Overcoming internal and external resistance and impairments to carry out what one has set out to do can be perceived as strenuous; this effort, however, does not correspond to that of performing an activity itself (R&V, p. 178). The authors mention as an example that someone who willingly dances the night away in a disco may feel less exhausted, despite the physical exertion, than an anxious person who has just completed a first bungee jump and had to overcome considerable inner resistance. A closer examination of the volitional processes involved in the provision of effort would allow a more profound and more precise determination of why subjects engage in (aversive) tasks even for a prolonged period. However, a comprehensive treatment would go beyond the scope of this thesis.³³ The following remarks briefly point to aspects that are essential for the topic and subsequent chapters.

Willpower to execute activities is especially needed for aversive activities, but also for activities that only yield little reward. To master this kind of activities successfully, *action control* is needed to shield the current intention to act from other motivational tendencies (i.e., motivations for alternative actions; R&V, p. 182). The predisposition of action control varies (see also Kuhl, 1983, who distinguishes between *action-oriented* and *state-oriented individuals*). Those with higher action orientation find it easier to remain on the right track despite resistances, interruptions, or failures in the course of action, but also competing temptations, until the goal is reached (R&V, p. 183). The authors continue that these individuals are not only able to formulate clearer and more complete intentions, but they also have more strategies for successful action control at their disposal.³⁴

Joyful activities hardly require willpower and shielding from distracting influences; however, for aversive activities, the opposite is true. Suppose a task that is perceived by many subjects as very aversive

³²As an example of an activity that has worthwhile consequences, but which is associated with fear, disgust or pain for many people, Rheinberg & Vollmeyer (2012, p. 194) refer to a visit to the dentist. In this case, the “positively evaluated goal state” toward which the current life pursuit is directed lies in the avoidance of the undesirable event “caries.” According to Rheinberg & Engeser (2018, p. 579), such an “avoidance motivation” may have different qualities than a pure “approach motivation.”

³³E.g., for a discussion of motivational competence and volitional competence, see Rheinberg & Vollmeyer (2012).

³⁴Rheinberg & Vollmeyer (2012, pp. 183–184) provide a summary of Kuhl’s findings on the characteristics of individuals with high action orientation and on strategies for “volitional action control.” Meta-motivation processes likewise serve to imagine and paint attractive consequences in order to trick oneself and increase one’s own motivation (R&V, p. 181). Consciously formulated, specific intentions with clear instructions for action can reduce the cognitive load once the appropriate situation occurs and can be conducive to the realization (R&V, p. 192).

is used in an experiment. In that case, there is a chance that instead of measuring exerted effort in relation to financial incentives, the action control abilities of the subjects are assessed instead.

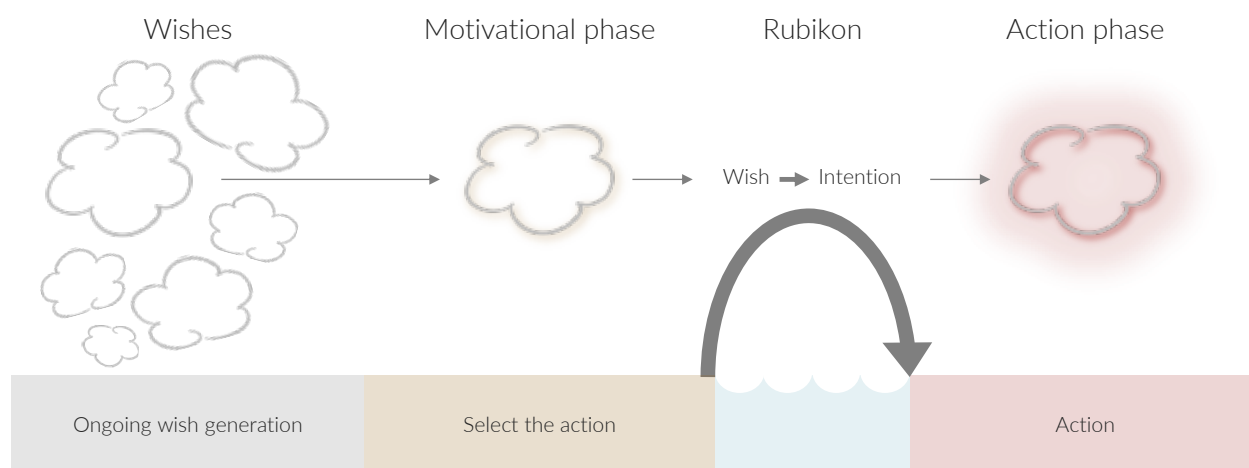


Figure 2.6: **Sequence of action phases based on Heckhausen's Rubicon Model** (modified after J. Heckhausen & Heckhausen, 2018, p. 358): In the previous sections, the motivation evaluation process was described, which serves to select the action to be carried out. More generally, people carry around a lot of desires or wishes for possible actions at any given time. Eventually, a subset of these actually enter the evaluation process in the **motivation phase** in an appropriate situation. If an activity proves to be worthwhile to be carried out (see Figure 2.1 above), an *intention* for its realization must be formulated. Once this is decided, *the die is cast* (the **Rubikon** has been crossed, which gives the model its name). In the following **action phase**, volitional processes ensure and safeguard the execution of behavior until the intended outcome is reached.

2.3.6 Concluding Remarks

Meanwhile, it has become clear that “effort” is neither easy to grasp, nor can it be viewed in isolation. Instead of dangling freely, it is *accompanied by*, is *embedded in*, or, even more, is *the result of* motivational and volitional processes. Thus, if one attempts to measure how much effort someone makes, one inevitably measures a blend of “these factors.”

The extended version of Heckhausen's *Advanced Cognitive Motivation Model* served as a starting point for a systematic analysis of the plot situation of a subject in the laboratory. The investigation demonstrated the importance of differentiating between the *outcome* of an action and its *consequences*. For the study participants, the latter are likely to comprise more than just the number of points achieved multiplied by a piece-rate. A wide range of non-pecuniary incentives may induce effort provision and

are likely to conceal the effects of monetary incentives. With a deeper understanding of the motivational and volitional process behind the provision of effort, the next section focuses on how to rule out activity-related incentives and purpose-related incentives other than pecuniary incentives, e.g., due to self-evaluation or other-evaluation.

2.4 Criteria and Practices for Designing Real-Effort Tasks

At the onset of this thesis, several [applications of tasks](#) are presented. Chapter 1 emphasizes that the task must be chosen to match the application and consistent with the present research question. However, selecting a suitable task is not straightforward: Section 2.2 points out a multitude of shortcomings observed across tasks; Section 2.3 reveals that the reasons for making an effort can be manifold. Besides, tasks can be implemented in different ways, and common standards for experiments with real effort are yet missing.

This section presents a comprehensive set of design criteria to facilitate the development, selection, and implementation of tasks to address this need. In preparation for the compilation of criteria, a representative share of the literature on real effort has been reviewed.³⁵ The proposed criteria can broadly be grouped as follows (see also Figure 2.7):

1. *Curb activity-related incentives and flow;*
2. *Curb undesired purpose-related incentives;*
3. *Skills and character traits should be irrelevant;*
4. *No learning effects;*
5. *Elastic effort response;*
6. *Statistical significant results.*

The first four criteria aim at reducing the impact of non-pecuniary incentives by attempting to gain more control over the cost-of-effort function. The latter two criteria seek to improve the significance of results and, therefore, (mainly) target the output production function. Each of the listed criteria can be broken down into *sub-criteria*. These are then addressed with a set of *design practices*. Some of them may appear evident or trivial – e.g., using a trial period to mitigate learning. However, they may easily be neglected in the process of designing and planning an experiment – despite their great importance.

³⁵The literature review process has benefited greatly from tables of real-effort tasks assembled by [Gravert \(2014\)](#), [Charness et al. \(2018\)](#) and [Winter \(2017\)](#).

The design practices can be implemented independently of one another. They have to be taken into account and applied against the background of the respective research question and the particular experimental design. Consider, for example, studies in the [third area of application of tasks](#) that aim to reproduce a work situation in which people perform a task they enjoy. In this case, there is no reason to adopt design practices that reduce employees' motivation to complete the task. Similarly, gender-specific differences in performance could form an integral part of a research topic. The task to be chosen must evoke them accordingly to reflect these circumstances.

Some design practices are more general and can be implemented in any real-effort experiment. They are, therefore, referred to as *task-independent design practices*. Conversely, ways and means that specifically target the properties of tasks are termed *task-dependent design practices*. Figure 2.7 summarizes the design criteria with both types of practices to meet them.³⁶

Each design criterion (possibly with its sub-criteria) and the respective practices are described in the following. References and examples from the literature are given to substantiate their relevance and impact. Some of the design practices are capable of meeting several of the design criteria (or their sub-criteria). These are described only when they first appear unless further clarification is needed when they also apply for another criterion.

Diligent randomization of subjects to treatments and, when feasible, inclusion of a *control group* are further conducive to addressing several of the design criteria. Since these constitute elementary components of laboratory experiments, their listing and further discussion is redundant.

The presented criteria are directed primarily towards effort measurements with *unrealistic* tasks producing *useless* output in [the first and second application of tasks](#). Despite this, they can also be applied to *realistic* tasks in [application three](#) if not indicated otherwise.

³⁶Task-dependent practices will also play a crucial role in the remainder of this thesis: Section 2.5 introduces a novel real-effort task, designed to meet them as much as possible; Chapter 3 proposes a survey to evaluate tasks according to a selection of the task-dependent practices to identify a preferred task for experiments in the [first and second area of application of tasks](#).

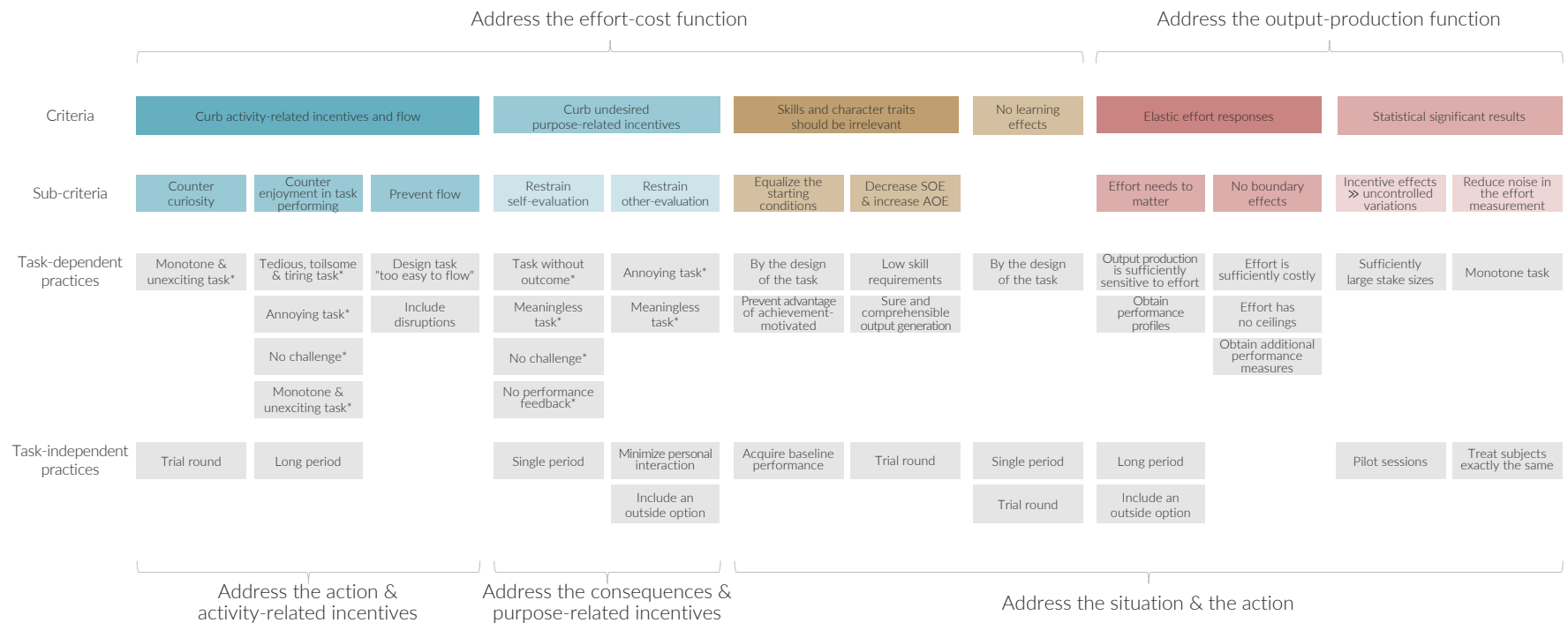


Figure 2.7: **Design criteria and design practices for fabricating and implementing real-effort tasks:** The first row lists the six design criteria. **Criteria in petrol blue** aim to counter “voluntary effort” provision due to i) *activity-related incentives* (curiosity, task-enjoyment) and subjects entering the easy flow state, and ii) *purpose-related incentives* other than pecuniary incentives, e.g., due to self-evaluation (feeling of success, sense of pride in one’s efficiency) or other-evaluation (social recognition by the experimenter or peers). **Criteria in brown** target influences from skills and learning. **Criteria in red** address the output-production function. Some of the criteria can be divided into *sub-criteria* (see the second row). Furthermore, the third and fourth rows list *design practices* to address each of the criteria. Thereby, the third row contains *task-dependent practices*, which directly affect the design of a task (marked with a “*” are those that serve as the basis for the real-effort task survey introduced in Chapter 3). The fourth row includes *task-independent practices* that can be implemented in *any* task. The grouping brackets below indicate to which structural elements of the **extended version of Heckhausen’s Advanced Cognitive Motivation Model** the criteria and practices can be associated. Certain practices meet multiple criteria (or their sub-criteria). Subsequently, they are described only as they first occur, unless a further explanation is required.

2.4.1 Curb Activity-Related Incentives and Flow

Taking a perspective from motivational psychology, the previous section emphasizes that subjects can perform a task simply out of curiosity or enjoyment of the activity. Many experimentalists are aware of this problem and have developed a number of more or less boring tasks (or deliberately adopted such tasks for this reason). Concerned that voluntary provision of effort could influence their results, they explicitly emphasize that a monotonous and tedious task was intentionally and carefully chosen for the particular experiment.³⁷ However, few authors give reference to how well the given task fulfills these criteria, why these specific criteria were selected, and how it compares to other tasks.³⁸ In short, there is no common agreement on which properties of a task can mitigate activity-related incentives. In the following, design practices are outlined how a task can be constructed and implemented in such a way that *curiosity*, *joy*, and *flow*, can be greatly reduced.

2.4.1.1 Counter Curiosity

Some people do certain things simply out of curiosity – and that includes newer activities. Admittedly, this does not have to apply to all subjects. However, a significant share of study participants will initially perform *any activity* that the experimenter “feeds to them” (even if it is somewhat aversive or does not yield large returns).

Although curiosity has as a consequence the state of “reduced uncertainty,” which can offer a certain incentive, it is the *activity itself* that drives a curious person, i.e., the state of searching for solutions, which is sought and enjoyed. Curious people even specifically seek out situations that offer this state of search and enlightenment, e.g., the cautious look into the neighbor’s garden.^{39,40} Curiosity is thus *activity-centered*. It can be undermined as follows.

³⁷For example, [Corgnet et al. \(2011, p. 12\)](#) express that “by using a long, repetitive and effortful task we ensure that individual performance is mostly driven by effort considerations. We do so because our main objective is to test standard predictions of incentives theory while abstracting from confounding factors such as intrinsic motivation.”

³⁸Chapter 3 takes a first step in this direction and examines and compares a set of seven tasks.

³⁹The possibility to attribute curiosity to *activity-related incentives* was confirmed by Professor Rheinberg in a personal exchange (personal communication, Dec. 2020).

⁴⁰Curiosity is not to be confused with *interest as described earlier*.

Trial round. Subjects who perform a task only out of curiosity may have an innate drive to try out new activities. However, this initial spark is likely only transient in nature. The inclusion of a *trial round* in real-effort experiments allows subjects to become familiar with the task, its properties, and its technical implementation. Any anticipation and enthusiasm for the task are thus greatly reduced before the actual effort measurement begins. Possible learning effects during the completion of the task can be additionally mitigated to a certain degree.

Monotone and unexciting task. Using a *repetitive and non-exciting* task may further diminish task-performing out of curiosity. For example, [Bortolotti et al. \(2016\)](#) (sorting and counting coins) and [Berger & Pope \(2011\)](#) (ab-typing task) deliberately employ monotone tasks to reduce the likelihood that subjects derive utility from completing the task.

2.4.1.2 Counter Enjoyment in Task-Performing

Tedious, toilsome, and tiring task. Regardless of the size of monetary incentives, the subjects' effort can remain high if they enjoy performing a task. A simple way to reduce any performance out of sheer pleasure in the task is to make it "triple t": *tedious, toilsome, and tiring*. If a task is boring and not entertaining, subjects will be much less inclined to perform it out of fun or joy. A task that is laborious and fatiguing deprives any pleasure and enthusiasm to complete it. Conversely, a "task that is enjoyable to the subjects would blur the line between labor and leisure and make interpretation of the results much more difficult," as [Dickinson \(1999, p. 650\)](#) notes. [Abeler et al. \(2011\)](#) and [Benndorf et al. \(2014\)](#) certify the counting-zeros task to be both boring and tedious. [Bonein & Denant-Boèmont \(2015\)](#) attest the slider task from [Gill & Prowse \(2015\)](#) to be very dull.

No challenge. Referring to [McClelland \(1999\)](#), [Rheinberg & Engeser \(2018, p. 599\)](#) note that the *activity-related incentive of achievement motivation* "resides in the experience of «doing better for its own sake» ([McClelland, 1999, p. 228](#)) – a kind of «consummatory experience» that is characteristic of achievement motivation." The authors continue that *feelings of competence* experienced while executing the task are accompanied by *full immersion* in that task, just like in the case of flow. Thereby, the degree to which the own skill level matches the skills required by a task determines whether the

task represents a *challenge* for the subject or not. If there is a balance between skill-demand and own skills, the flow state becomes permissible (see Section 2.3.4). The activity-related incentive to achieve goes beyond the general components flow and additionally provides the pleasure of one's *optimally-efficient functioning on the way to a challenging goal* (Rheinberg & Engeser, 2018, p. 601). Therefore, to undermine incentives of this kind, it is advantageous to design the task so that it *does not present or entail a challenge*: If there is no challenging goal, there cannot be any joy in pursuing that goal. The design practice to counter achievement motivation is, therefore, akin to the one against flow and points in the same direction: Make tasks so simple that they can be done by anyone and do not pose a challenge to anyone (see Section 2.4.1.3).⁴¹

Long period. Prolonging the *duration of effort provision* acts beneficial in a similar way, as it is harder to maintain motivation and exert maximum effort for a sustained amount of time (self-control and concentration become increasingly costly). In line with this, Corghnet et al. (2011) state that “given the limited duration of laboratory experiments, the use of long and laborious tasks are necessary to create boredom and fatigue” (p. 13).⁴²

In fact, a prolonged task duration might be essential to reveal crucial details: In a field experiment, Gneezy & List (2006) find support for gift-exchange in a data entry task for a university library as well as in door-to-door fundraising at first; however, the observed effect is transient and after a fraction of the task duration effort levels become indistinguishable across treatments.

Nevertheless, if subjects are obliged to complete a task over a very long period of time, their abilities or character traits may become more apparent. For example, consider a tedious and toilsome task that requires an overly high level of concentration. Exaggerating the duration is likely to lead to the

⁴¹Recall the study by Araujo et al. (2016) examining the slider task by Gill & Prowse (2015): Over ten rounds, the goal of “outperforming previous rounds” was enough to elicit significant effort from the researchers [sic.]. Therefore, the design practice *single-period* recommends conducting only one round of effort provision and avoiding intermediate feedback.

⁴²The “required” task duration to evoke the mentioned psychological reactions in the participants certainly depends on the type of task and the design of the experiment. It is thus reasonable to perform pilot sessions and obtain feedback from the participants to gauge the time of duration. In light of recent literature, a period duration of five minutes as implemented by Niederle & Vesterlund (2007) seem to represent a lower bound. Most frequently, task durations are between ten and 20 minutes, as in Nikiforakis et al. (2012), Dohmen & Falk (2011) or Eriksson et al. (2009). Effort measurements for more than 60 minutes are not uncommon (see Dickinson, 1999; Falk & Ichino, 2006). In this case, they are often subdivided again in intervals of ten to 20 minutes. This has already been used by Swenson (1988) to enable subjects to recover from a tedious, physically demanding task, or more recently, for example, by Corghnet et al. (2016) and Corghnet, Hernán-González, & Schniter (2015). However, providing performance feedback during breaks is not recommended, as discussed below.

effort measurement, also reflecting the subject's action control up to a certain extent (see [Kuhl & Beckmann \(1985\)](#) and Section 2.3.5).

2.4.1.3 Prevent Flow

In Section 2.3.4, flow is discussed as a state of total immersion in an activity in which effort is hardly associated with cost. At first it may seem bizarre to discuss flow in the context of real-effort tasks. However, if one realizes i) "how easily" individuals enter the (performance-enhancing) flow state, ii) which factors are conducive to a transition, and iii) that these factors are not infrequently present in tasks, then the situation changes. One becomes even more aware of the importance of avoiding such factors in tasks to prevent flow when one realizes that the computer game industry deliberately works with elements of this kind to provide the player with a flow experience (see e.g., [Sweetser & Wyeth, 2005](#)).

Several conditions were identified that are necessary for subjects to enter flow, including: (i) a clear goal; (ii) one's abilities must match the demanded abilities if the activity is unimportant; or (iii) one's abilities must exceed the demanded abilities if the activity is instead important; iv) if the subject is motivated by failure, in both cases the subject additionally needs an "ability buffer" to have the confidence to be able to accomplish the task (see also [Rheinberg & Engeser, 2018](#)). Figure 2.8 illustrates how relatively small variations in the design of the digital version of the wire-loop game can give rise to flow-promoting side effects. In the following, design practices to prevent subjects from entering the flow state are described.

Design task "too easy to flow". Figure 2.5 in Section 2.3.4 exemplifies the *skill-demand balance* as a necessary condition for flow. In addition to the increase in skill demand associated with increasing *task complexity*, the impact of *task importance* was also considered, which can have an unfavorable effect on entering the flow state. The modified version of the flow channel model depicted in the figure illustrates that if a task is *important*, one's ability level must exceed the task's ability requirements for flow to be accessible. In the laboratory, one's own performance is of "great importance" insofar as it determines the payoff – which is, after all, the initial reason for participating in the study in the first place.

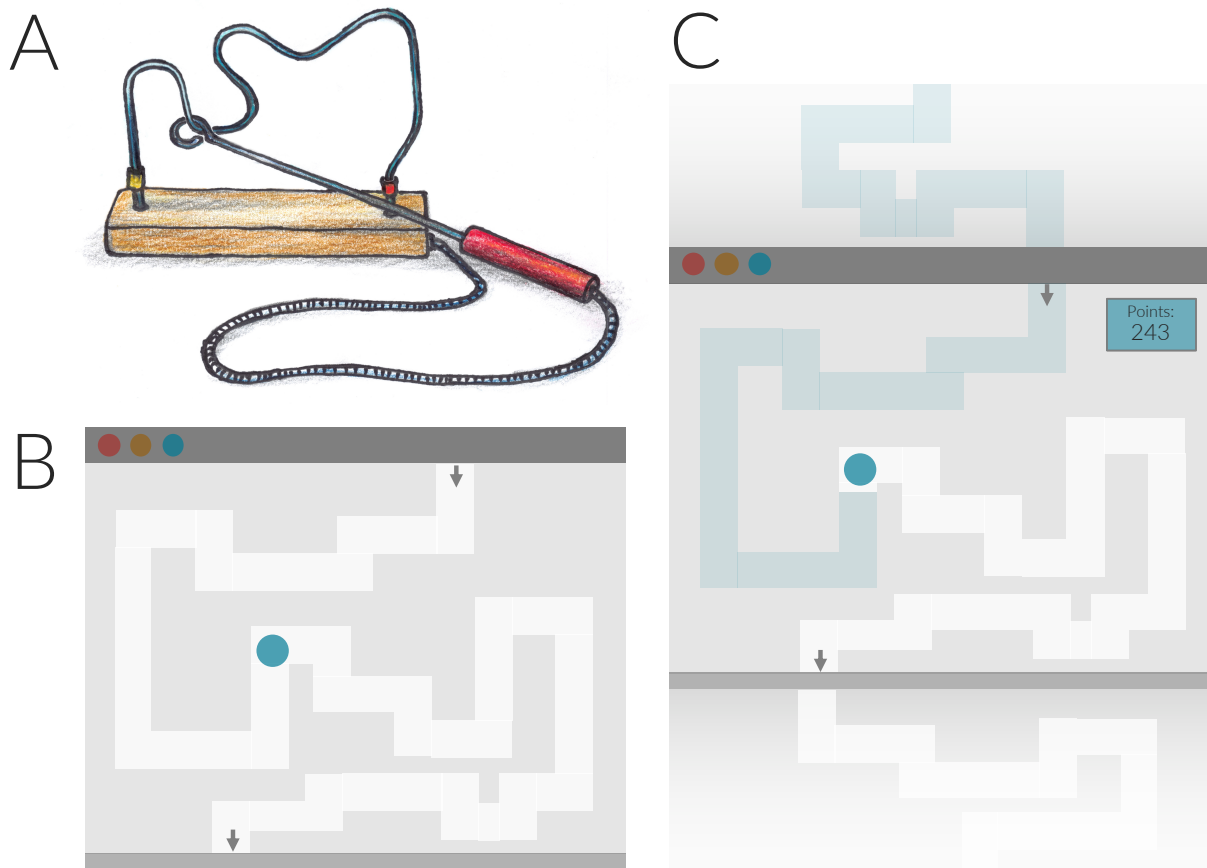


Figure 2.8: **The wire-loop game:** A) in the original version, the task is to move a metal loop along a winding wire without the loop touching the wire. Loop and wire are connected to a power source to form a closed circuit when contact occurs. Upon touching the wire, a loud alarm tone sounds. B) In a digital version, the wire is replaced by a narrow, winding path along which an object (e.g., a circle) must be dragged without touching the sides. As in the offline version, the game has a fixed start and endpoint that is visible and provides an attainable goal. The degree of difficulty changes both with the twist of the wire (path) and the diameter of the loop (circle). Sophisticated hand-eye coordination is advantageous in either variation. C) Alternatively, one could imagine an endless version in which the path is continuously extended and replenished. In this “infinite version,” flow is expected to occur more frequently than in the “finite version.” Because flow is performance-enhancing, the average performance is likely higher (Engeser et al., 2005; Engeser & Rheinberg, 2008; cf. Vollmeyer & Rheinberg, 2003). For demonstration purposes, two unfavorable ways to deliver feedback are depicted (see Section 2.4.2.1): i) the number of points achieved so far is indicated, and ii) the successfully completed part of the path is highlighted in color. They can promote self-evaluation consequences (feeling of success, pride in one’s own efficiency), which provide a purpose-related incentive for some of the subjects and motivate their efforts. Renouncing these two (needless) elements leaves the task plain and dull and without any outcome.

However, if subjects are motivated by fear of failure, they need the aforementioned *ability buffer*: only if their own skill level greatly exceeds the task's ability requirements can they achieve flow. Thus, a task with a tenable skill demand that is not too high and not too low would enable a larger number of subjects to enter the flow state – at least in theory (see Figure 2.9).

However, since the cost of effort approach zero in the case of flow, the goal in [task applications one and two](#) is just the opposite: To prevent subjects from entering the flow state while ensuring that they all have similar starting conditions. The latter is easily achieved by assigning an extremely difficult task, which exceeds the skill levels of all subjects such that virtually nobody is able to accomplish it.⁴³ However, to what extent this actually measures “effort,” or perhaps rather stubbornness or even apathy due to being over-challenged, remains an open question.

Alternatively, one can choose a relatively-easy-to-complete task, such that *anyone could do it*, and *no one is challenged enough to enter the flow state*. To get an idea of how basic and unchallenging the task needs to be, consider that [Vollmeyer & Rheinberg \(2003\)](#) observe flow even in such simple activities as the game of *Pacman*. Thereby, flow has a strong positive effect on the performance of the participants.⁴⁴

Include disruptions. Section 2.3.4 emphasized that flow is only possible if the goal of the action is perfectly clear. Only if this is the case, no further attention and processing capacity is required and the goal can be pursued with great determination and dedication. Therefore, flow is predominantly observed in activities that are free of interruptions, leaving no room for musing. One possibility to obstruct a transition into or the remaining in the flow state is thus to deliberately *introduce disruptions* in the task sequence. This can create space for rethinking and questioning the goal of one's own actions.

However, careful consideration is required when including interruptions, as the ability to suppress such thoughts becomes more and more demanding as their frequency increases. As mentioned above,

⁴³See [Heyman & Ariely \(2004\)](#) for an application of this kind, where “effort” was equated with the “time spent on trying to solve an unsolvable task until giving up.”

⁴⁴The authors observe a smooth sequence of the task (as in the flow condition) even with low task difficulty. However, the absorbedness typical of flow (and, therefore, presumably the gains in productivity) is only achieved at intermediate task difficulty ([Rheinberg & Vollmeyer, 2003](#); [Vollmeyer & Rheinberg, 2003](#)). “Under-challenge” will thus primarily affect absorbedness during the task. One should note that the subjects were paid a fixed amount for their participation in these studies, so a fear of failure does not play a role.

there is a tendency for the measurement of subjects' effort to degenerate into a measurement of their action control (see again Kuhl & Beckmann (1985) and Section 2.3.5).

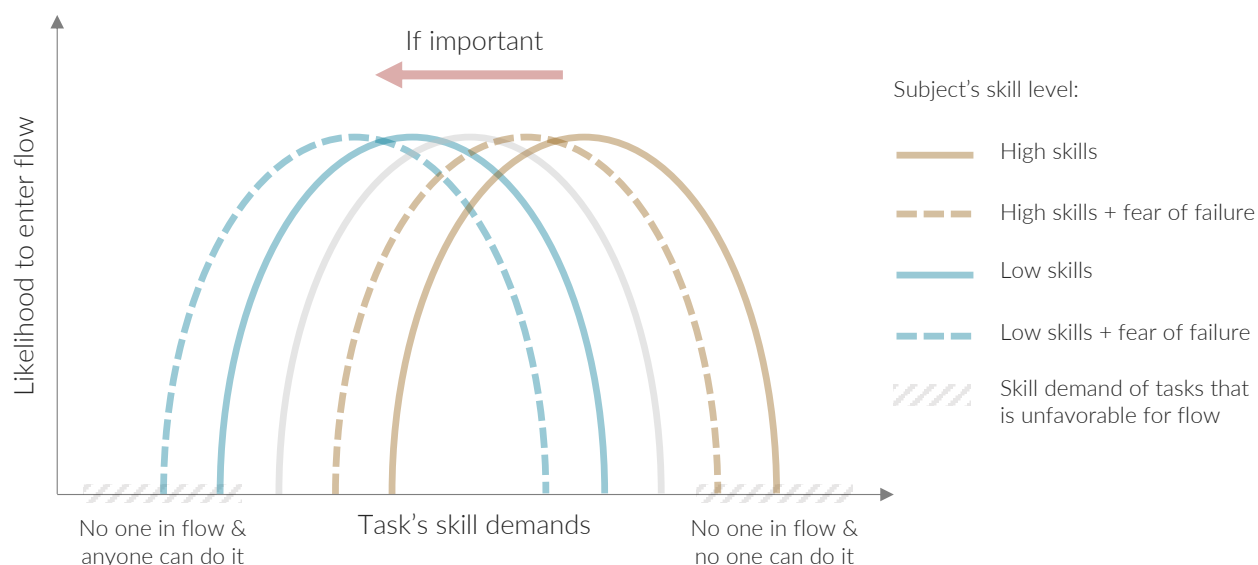


Figure 2.9: **Likelihood of entering the flow state in relation to task difficulty and skill level:** Several studies find an inverse U-shaped pattern for subjects who report experiencing flow contingent on task difficulty (Rheinberg & Vollmeyer, 2003; for simple tasks, see Vollmeyer & Rheinberg, 2003). The figure illustrates the influence of the *individual skill level*, *motivation by failure*, and the *importance of the task* on the *possibility or probability of reaching the flow state*. Due to the skill-demand balance as a necessary condition for flow, subjects with **higher skill levels** are more likely to enter flow on more complex tasks, while subjects with **fewer skills** achieve this only on simpler tasks. For important tasks, one's skill level must exceed the skill requirements of the given task for flow to be feasible. Consequently, all the curves depicted shift to the left with **increasing task importance**. Since the goal in **task applications one and two** is to derive tasks that prevent flow, tasks that are very simple, can be done by anyone, and do not pose a flow-promoting challenge to anyone are a good choice (see the dashed area to the left).

The goal in real-effort tasks is fairly straightforward: earn points (for one's benefit or the benefit of others)! Softening this goal to prevent the transition to the flow state is not necessarily beneficial, as it could bring unwanted and undesirable side effects. Nevertheless, one can seek to attenuate the salience of the goal by attempting to let certain elements fade into the background (see Figure 2.8 for an example). However, approaches of this kind primarily aim to reduce the influence of self-evaluation consequences and other-evaluation consequences and will, therefore, be discussed in the next section.

2.4.2 Curb Undesired Purpose-Related Incentives

As noted in Section 2.3.4, an individual's motivation may change during the course of an activity. For example, motivation may initially start from a certain degree of curiosity regarding the activity, which then evolves into enjoyment in performing the activity (both activity-related incentives). Furthermore, *self-evaluation consequences*, such as a “feeling of success” and “pride in one's own efficiency” can be triggered during the process. These outcome-dependent future internal states unfold a motivationally effective incentive. Besides, “approval-seeking” toward the experimenter and peers represent *other-evaluation consequences* that provide further purpose-related incentives and may likewise encourage effort.⁴⁵

Moreover, Rheinberg & Engeser (2018, pp. 587–588) describe the differences between individuals motivated by the consequences of *mastery-goal orientation* and *performance-goal orientation* in the context of learning motivation. Applied to the situation of subjects performing tasks in the laboratory, the following results. *Mastery-goal orientation* can be characterized as a “self-evaluation consequence” with a *personally set benchmark*. Subjects with such a trait will commit themselves, because they want to excel in the task “for themselves,” which is why their goal is to acquire or improve possibly needed skills. Conversely, *performance-goal orientation* constitutes an “other-evaluation consequence,” whereby individuals orient themselves to a *social benchmark*. Subjects motivated in this way exert themselves to demonstrate their proficiency because they want to prove their superiority in ability and expertise over others.

For subjects with performance-goal orientation, the task does not even have to be integrated into a multiplayer game to spur them on. It is sufficient for them to meet another subject in the hallway outside the laboratory after the experiment to brag about their performance (hoarding of this kind

⁴⁵Related non-pecuniary incentives discussed in behavioral economics were described earlier. These include, for example, “warm glow,” which DellaVigna et al. (2016, p. 2) describe as a “placeholder for any motive that increases a worker's utility from exerting effort on behalf of the employer, independent of the returns generated for the employer. This could be a positive feeling from doing meaningful work, adhering to a social norm of working hard, or signalling prosociality.” In terms of “experimenter-demand effects” described in Section 2.2, consider Charness et al. (2013) who note that subjects may feel obliged to perform specific actions to meet certain expectations of the experimenter. Masclet et al. (2015, p. 18) expand on this, noting that the potentially prevalent “authority relationship between participants and experimenter [...] may simply reflect a *sense of duty* [emphasis added] and mirror the field setting in which this type of vertical relationship exists between employer and employee.” For a more detailed discussion of loyalty, commitment, and contract fidelity to the experimenter in laboratory experiments, as well as implications of any experimenter demand on external validity, see Zizzo (2010).

was observed frequently after the sessions of the experiment described in Chapter 3).

2.4.2.1 Restrain Self-Evaluation

Task without outcome. One way to convey the purposelessness of a task is if it does not lead to *any tangible outcome*, such that any exerted effort does not produce “anything.” In terms of the *model* introduced in Section 2.3, the points accumulated in the task constitute its sole outcome. At the end of the experiment, they are converted into real money, which provides the desired purpose-related incentive. To approximate this, [Abeler et al. \(2011\)](#) implement a boring, pointless task to be confident that it implies actual effort costs for the study participants. The futility of the task must be salient to all subjects and for the entire course of the task – to ensure that they are aware of it at all times. At best, it becomes evident already while reading the instructions of the task. Furthermore, tasks with this property also discourage goal setting, i.e., targeting and then striving for a specific score. And, if there is no outcome, this also deprives any achievement-motivated action of its basis. After all, achievement-motivated action is outcome-oriented by its very structure and moves towards a specific goal, in which it derives its incentive to act. If there is no attainable goal, there is no basis for achievement motivation (see also Section 2.3.4, Section 2.4.1.2 and [Rheinberg & Engeser, 2018](#)).

Meaningless task. If a task nevertheless produces some more or less tangible outcome, then it possesses a purpose. In this case, one can further distinguish if the outcome is *meaningful*, in which case it has a value outside the laboratory, or not, and thus is *meaningless*.^{46,47} If the output is of value, this can bring consequences, which in turn carry an incentive and spur effort over and above monetary incentives. As an example, the completion of a task may provide benefit to the subject herself (cracked walnuts may be retained), the researcher (data entry into a database), or a third party (donations collected for a non-profit organization).⁴⁸ In *task application area three*, this may be of value to generate realism or answer the specific research question. They are, therefore, crucial for

⁴⁶A task that is both *purposeful* (has an outcome) and *meaningful* (this outcome has a value outside of the lab) can be termed *useful* (see also the discussion in Section 1.2.2 on the usefulness of output produced in tasks).

⁴⁷This design practice serves both to restrain *self-evaluation* and *other-evaluation*.

⁴⁸The “cracking walnuts” task of [Fahr & Irlenbusch \(2000\)](#); entry of books and articles into library or research database ([Corgnet, Hernán-González, Kujal, et al., 2015](#); [Gneezy & List, 2006](#); [Hennig-Schmidt & Sadrieh, 2010](#)); door-to-door fundraising ([Gneezy & List, 2006](#)) or folding letters for fundraising mailings ([DellaVigna et al., 2016](#)).

the experimental design and can hardly be neglected. However, in [task application areas one and two](#), such supplementary influences are much less sought or appreciated. To avoid these influences, a meaningless task whose output has no value outside the laboratory, i.e., which does not entail beneficial consequences in favor of the study participant, the experimenter, or third parties, resembles a good choice. In this line, [Abeler et al. \(2011, p. 5\)](#) deliberately employ a simple counting task, which is “clearly artificial and [generates] output of no intrinsic value to the experimenter ...[to minimize] any tendency for subjects to use effort in the experiment as a way to reciprocate for payments offered by the experimenter.”

No performance feedback. Feelings of success or pride in one’s efficacy, as well as a mastery-goal orientation, can be nourished by performance feedback given during task performance. This means that subjects who can monitor and track their performance while completing the task are spurred to develop personal ambitions, set goals, and strive to outperform themselves. Performance feedback in this context refers to *any* form of tracking and displaying so far produced output. These include, in particular, scoreboards and progress bars that indicate progression and potentially a clear end or upper limit to be reached.

If one chooses to implement a hands-on task, such as stuffing envelopes, the output produced is inevitably observable. In contrast, computer-based tasks can usually be designed so that subjects do not need to be informed about their current score.⁴⁹ Two unfavorable ways of giving feedback, which may work as nudges and motivate performance, are exemplarily shown in [Figure 2.8](#) in terms of the wire-loop game: The current score is presented, and the share of the wire that has already been mastered is highlighted.

Single period. Finally, to curb any motivation to beat one’s performance in the previous round(s), subjects at best complete a task *only once*. Alternatively, if the experimental design does not permit conducting only a single round, subjects may receive information about their performance within

⁴⁹(Continuous) Relative performance feedback has received attention in the literature on non-monetary performance incentives ([Charness et al., 2013](#); [Eriksson et al., 2009](#); [Falk & Ichino, 2006](#); [Fu et al., 2015](#)). Notwithstanding the impact of live rankings against the progress of an average subject or peer in a group, however, the forms of performance feedback addressed here are much broader and go beyond them. Explicitly meant here is *any* form of feedback or information about the subject’s current score.

the task only after all rounds of the task have been completed. Providing interim feedback instead would counteract any attempts to prevent efforts due to mastery-goal orientation and achievement motivation.

2.4.2.2 Restrain Other-Evaluation

Annoying task. If a task is sufficiently meticulous and grueling, subjects will be reluctant to perform the task solely to please the experimenter. Tasks have been developed or deployed particularly with this purpose in mind (Augenblick & Rabin, 2015; Charness et al., 2013; Masclet et al., 2015, 2015; Neyse et al., 2014). However, if the task is perceived too cruel, subjects might refuse to complete it, as Araujo et al. (2016) point out (see also the discussion on *volition and aversive activities* in Section 2.3.5).⁵⁰

Minimize personal interaction between experimenter and subjects. To further counter experimenter-demand effects, Masclet et al. (2015) suggest to minimize any personal interaction between experimenter and subjects. For example, subjects may only receive written instructions (printed or on the computer screen) or are played pre-recorded audio instructions.⁵¹

Cox & Sadiraj (2019) implement a double anonymous (or “double-blind”) payoff protocol: When entering the laboratory, subjects draw an envelope from N identical-looking sealed envelopes, each containing a numbered key to a personal payment box; subjects are informed to keep their key number private and may only enter it in the respective payment form (on the computer or printed); at the end of the experiment, the experimenter uses this information, linking key-numbers to payoffs, to fill the single payment boxes; subjects may then, one-by-one, collect their earnings by unlocking their personal payment box confidentially. Given the study participants’ privacy during payment and their anonymity in the experiment itself, which may further mitigate experimenter-demand effects,

⁵⁰To protect study participants from overly punishing and torturous tasks, economics journals increasingly require formal approval by university ethics committees.

⁵¹Such measures also help to establish a more professional environment so that subjects take the experiment seriously. They also contribute to the standardization of the individual sessions of an experiment (reducing noise) as well as experimental studies as a whole, which facilitates the reproducibility of results.

this procedure appears very suitable for economic experiments.⁵²

Include an outside option. Once subjects have entered the experimenter's realm, they are bound to complete his experiment and any grueling duties that come with it. In the absence of any desirable alternatives, all they can do is to sit around idly and twiddle their thumbs, in order to run down the clock. In short, with nothing better to do than adhering to the experimenter's wishes, the subjects' opportunity cost-of-working are approximately zero.⁵³

In light of this, it is not surprising that experimental findings for the relationship between effort provision and piece-rate incentives are somewhat mixed and not as unambiguous as standard economic theory suggests.⁵⁴ Furthermore, observing non-negligible effort provision under flat payment schemes emphasizes that non-pecuniary incentives may significantly influence behavior.⁵⁵

To unravel the impact of non-pecuniary incentives, [Erkal et al. \(2017\)](#) increase the opportunity cost of working by offering study participants three different outside options. The authors find effort levels to be significantly lower in all three treatments than in the control treatment. A paid alternative activity turns out to be most effective in diminishing the influence of non-pecuniary incentives on behavior and allows the recovery of pecuniary incentive effects ([Erkal et al., 2017](#)).

[Corgnet, Hernán-González, & Schniter \(2015\)](#) examine how much time subjects devote to an outside option (surfing the Internet) under two different payment schemes. Only when access to the Internet was available, individual payment showed stronger effects than team payment.⁵⁶ This indicates

⁵²In some experiments, subjects can pose comprehension questions to the experimenter. In this case, confidentially is only assured if the number of the workplace of a subject does not equal the number of her private payment key. Thus, when entering the laboratory, subjects could separately draw a seat number and a blank, closed envelope containing the key.

⁵³Some may argue that there are plenty of wonderful job opportunities and other pleasurable things to do beyond the laboratory. However, locked up in the experimenter's dungeon any of these "outside options" appear very intangible and far from reach. Hence, they barely matter to or may affect the dynamics inside.

⁵⁴[DellaVigna & Pope \(2016\)](#) find a positive relation; an inverse *U*-shaped relationship between effort levels and stake-size is found by [Pokorny \(2008\)](#), [Ariely, Gneezy, et al. \(2009\)](#); conversely, [Gneezy & Rustichini \(2000\)](#) find a *V*-shaped relationship; [Eckartz et al. \(2012\)](#) and [Araujo et al. \(2016\)](#) do not find any pronounced incentive effects at all. However, there are differences in the experimental design between these studies, such as the choice of task and the duration of effort provision.

⁵⁵See, for example, [Charness et al. \(2013\)](#) and [Masclot et al. \(2015\)](#).

⁵⁶[Corgnet, Hernán-González, & Schniter \(2015\)](#) observe that subjects devote 11.9% of their time to surfing the web when paid individually compared to 28.5% under team payment. The outside option, therefore, influenced output production differently across incentive schemes. Notably, individual production was similar whether the Internet was available or not.

that offering alternative activities is indispensable to unveil incentive effects – in particular in brief experiments employing undemanding tasks (Corgnet, Hernán-González, & Schniter, 2015). Without the outside option available, non-pecuniary incentives most likely dominate, such that no incentive effects are observed.

Lei et al. (2001) examine the role of speculation in the formation of bubbles in an experimental asset market. They find that a large share of the trading activity that leads to bubbles and crashes can be accounted to a lack of alternative activities available to study participants. Trading behavior driven by boredom, routine or urgency to act may lead to decision errors and thus carries an adverse effect on earnings. The authors postulate an *active participation hypothesis*. Rephrased for a general setting, it would read that subjects perform a work task a) because there is no other activity at their disposal, and b) because the experimental protocol encourages them to do so, i.e., promotes participation in a given manner (Lei et al., 2001). Rather than sitting around and doing nothing, subjects perform the task – even if it is not beneficial to their ultimate earnings.

To generalize, a lack of alternative activities may foster active participation in experiments such that subjects provide effort to please the experimenter or merely because they do not have anything better to do. According to Carpenter & Huet-Vaughn (2017), this may “inflate output numbers and bias treatment effects” to an extent that the literature is only just beginning to investigate.

Bearing in mind that many real-effort experiments aim to study labor supply and principal-agent relationships, *zero opportunity costs* for effort provision also do not resemble real-work environments very closely. In these, on-the-job leisure activities and distractions are ubiquitously present. A survey conducted revealed that an average employee spends more than eight hours per workweek on activities not related to work (OfficeTeam, 2017). Of this, nearly one hour per day is used on personal mobile devices, mostly to access private email and surf the web (e.g., access social networks, sports websites, online shopping, and more). Non-work related activities pursued at work may further include reading magazines and newspapers, playing games, taking extended coffee or smoking breaks, engaging in office gossiping or distracting colleagues.⁵⁷

However, production remarkably decreased under team payment when the outside option became available (Corgnet, Hernán-González, & Schniter, 2015).

⁵⁷Not all non-work related activities performed by workers necessarily come with a loss in productivity. For the need to take breaks to recover from work and resources depleting tasks, to regain focus or creativity, see, for example, Trougakos

Erkal et al. (2017) point out that employees can decide how to allocate their time at work in the majority of workplaces. As described, outside options in real work environments can take many forms, and they alter employee behavior in different ways. Omitting alternative activities in the experimental design would mean to neglect these influencing factors and their impact on how individuals perceive and respond to incentives at work, which is particularly relevant in [task application area three](#), which aims for realism.

Outside options in experimental research, therefore, serve two main purposes: *First*, they offer study participants an alternative activity and, thereby, increase the opportunity cost of providing an effort. Enabling the subject to do something else than following the experimental protocol makes the latter less salient and mitigates experimenter-demand effects in the form of believed expectations (\Rightarrow [applications one and three](#)). *Second*, outside options enrich the experimental setting to resemble an actual workplace more closely by offering a simple distraction or something more pleasurable or entertaining to do than completing an effortful task (\Rightarrow in particular [task application area three](#)). If subjects cannot perform their work task at the same time (or any more), outside options also decrease the benefit subjects may derive from task completion. In short, including alternative activities in experiments can increase the opportunity cost of providing efforts to mitigate active participation while enhancing the “mundane realism” of the experiment (Corgnet et al., 2016). It is thus surprising that only a minority of real-effort experiments include an outside option yet (Carpenter & Huet-Vaughn, 2017).

Just as many different tasks have been introduced over time, so too have various outside options been used. Swenson (1988) was among the first to use outside options to allow subjects to pursue “active leisure.” Subjects could choose to read magazines of different genres or play a video game or trivia card game. Among the first to experimentally assess how much time people devote to leisure and work, Dickinson (1999) examined on- and off-the-job leisure in a study on labor supply and work intensities by permitting study participants to leave the laboratory early. In an experimental study on individual and team incentives and peer pressure, Corgnet et al. (2011) include a simple clicking task to simulate a continuous influx of income irrespectively of provided work effort.⁵⁸ Table 2.1 lists the

& Hideg (2009).

⁵⁸Corgnet et al. (2011) include surfing the Internet as a further alternative activity. Allowing for peer monitoring makes shirking as prominent as working. This increases production drastically while reducing Internet consumption significantly compared to team production without the ability to monitor peers.

outside-options commonly used in the literature, some of which may seem more, some less tempting from a subject's perspective. This leads to four aspects to consider when incorporating an outside option into an experiment:

1. How *realistic* is the outside option, i.e., would this or a similar alternative also be available to employees in a real-work situation?
2. Is the outside option of *similar value to all* subjects?
3. Is it *implementable* in a given research design?
4. How *effective* is the outside option to *reduce the impact of non-pecuniary incentives* on behavior?

Table 2.1 includes two columns indicating the *degree of realism* and whether a *homogeneous valuation* by subjects is presumed. For example, the *incentivized time-out button* represents a somewhat abstract form of leisure, as highlighted by [Corgnet, Hernán-González, & Schniter \(2015\)](#).⁵⁹ Real employees engage in various activities for private purposes during work time, as mentioned earlier. However, they typically do not receive any fixed payment for taking a break and do nothing. With relatively little relation to the real workplace, this quite frequently employed outside option appears rather far-fetched. On the contrary, surfing the web may be considered a “real on-the-job leisure activity” since most employees may access the Internet through private mobile devices at any time such that it resembles a “relevant feature of real-world organizations” ([Corgnet et al., 2016, p. 2927](#) [sic.]), as also outlined above. To assure that “surfing the Internet” is of similar value to all study participants, as few as possible access restrictions could be applied, such that subjects may access personal emails, social media, instant messaging, or websites for sports, news, entertainment, and online shopping – just as in real life. If instead newspapers or magazines are offered as an alternative activity, the selection must be diverse enough to cover a wide range of interests and tastes.

The *implementability* is not considered as a separate third column, since all of the mentioned alternatives can be incorporated in laboratory experiments nowadays. Therefore, the research design itself and not technical difficulties are the main limiting factors. For example, permitting subjects to leave the experiment early is not an option if their presence is required for successive parts of the experiment (e.g., if after earning an endowment in an effortful task, subjects are matched in a multiplayer

⁵⁹[Erkal et al. \(2017\)](#) further note that the 25 time-out seconds incentivized with 0.10 € used by [Mohnen et al. \(2008\)](#) instead represent “inactivity,” in contrast to pleasurable *active* breaks at a real workplace.

game). [Corgnet, Hernán-González, & Schniter \(2015\)](#) question to which extent this alternative reflects a real work setting. Moreover, the authors note that due to the “lack of control over subjects’ activities and desired alternatives outside the laboratory, heterogeneity in quitting behaviors has been difficult to interpret” (p. 286). [Cooper & Kagel \(2016\)](#) mention that few study participants actually decide to leave and waive (potential) payments, as publicly walking out of the experiment may come with psychological discomfort.⁶⁰ Hence, they purposely conduct their experiment online on [Amazon Mechanical Turk](#), a crowd-sourcing website for labor. Conversely to the laboratory, the authors could thereby access a real labor market with “real outside options” – in which quitting naturally occurs.⁶¹

The fourth aspect demanded from a potential outside option is its *effectiveness in mitigating the impact of non-pecuniary incentives* on effort provision. However, studies comparing different alternatives are rare. [Erkal et al. \(2017, p. 529\)](#) compare three different outside options and find that offering a paid alternative activity is more effective in mitigating the influence of non-pecuniary incentives than an incentivized time-out button or the possibility for subjects to quit the laboratory prematurely.⁶² At the time of writing, no comprehensive comparison of the outside options listed in [Table 2.1](#) was available. Nevertheless, as the examples at the beginning of this section demonstrated, the inclusion of any alternative activity seems preferable in any case, rather than offering none at all.

If one decides to include a particular outside option in an experiment, this can be done in several ways. Thereby, the following implementation features can be considered:

1. A *trial round* to explore the alternative;
2. The option to *multitask*;

⁶⁰[Carpenter & Huet-Vaughn \(2017\)](#) note that the first study participant to leave the laboratory may set peer effects in motion. Few studies actually report that subjects quit the experiment prematurely. However, none of the studies reviewed showed a sharp decline in participation rates.

⁶¹Readers nevertheless interested in incorporating the possibility for subjects to leave the laboratory early are referred to [Dickinson \(1999\)](#), who makes two important remarks: *First*, to remove any sense of indebtedness or duty, as discussed at the beginning of the section on experimenter-demand effects, study participants should be obliged to provide a minimum amount of effort before they may leave the experiment (e.g., in the experiment from [Dickinson \(1999\)](#) subjects had to transcribe at least three paragraphs); *second*, the wage paid by the experimenter must be sufficiently high such that subjects do not abandon the experiment as early as possible to pursue a different working opportunity to generate income. This also guarantees that pecuniary incentives are the reason for subjects to remain in laboratory; if they decide to quit it reflects that they have supplied their optimal amount of work ([Dickinson, 1999](#)).

⁶²In [Erkal et al. \(2017\)](#) the “paid alternative activity” was to complete the same task as the work task just with a piece-rate instead of competing for a prize. The outside option, therefore, merely represents a change in the monetary remuneration scheme and any associated non-pecuniary incentives, such as (non-)eagerness to compete.

3. If multitasking is *not* available, the option to *resume working on the task after switching to the outside option*.

Section 2.4.1.1 described that some subjects might be curious to try out tasks. Analogously, there may be a desire to learn about the outside option, unless the outside option is well known to all study participants, such as surfing the Internet. Alternatively, a *trial round of the outside option* might likewise mitigate the issue.

Depending on the work environment to be modeled, one may enable subjects to *multitask* between the task and the outside option, or not. In Erkal et al. (2017), the screen is split between the work task and the paid alternative activity such that subjects may alternate between both at any time. Similarly, one could envision subjects to complete a typing task, as in Dickinson (1999), or a mechanical task, like the “ab”-typing task from Berger & Pope (2011), while being able to watch YouTube videos simultaneously. In contrast, Corgnat et al. (2016) disable multitasking by letting subjects actively switch between the work task and surfing the Internet. Another way to prevent multitasking is to purposely choose an outside option that subjects cannot make use of while performing the work task. In particular mechanical tasks, like the typing task from Swenson (1988) or cracking walnuts as in Fahr & Irlenbusch (2000), may render it technically unfeasible to perform both activities at the same time even if permitted. With or without the choice to multitask, subjects *actively choose* to pursue the outside option. However, one may argue that both implementations differ in salience and level of distraction of the alternative activity.

If one decides to disable multitasking, one further needs to choose whether to allow subjects to *switch to the outside option only once or back and forth between both*. Corgnat et al. (2016) explicitly separate the working task and the outside option but allow subjects to effortlessly and quickly shift between both in order to track how much time they spend on each activity.⁶³ In Bonein & Denant-Boèmont (2015), subjects can also surf the Internet as a leisure activity. However, to assess their self-control and ability to resist a tempting alternative, subjects cannot return to the effort task to earn further points.

⁶³To examine if subjects work longer or harder in presence of an outside option, it is advisable to allow subjects to a) switch among work task and alternative activity seamlessly, b) track how much time the subject spends on each and c) employ a task with an output production function, which is sensitive to the effort, such that differences in effort provision become apparent.

To summarize, the absence of outside options may amplify output produced in the task and bias treatment effects (Carpenter & Huet-Vaughn, 2017). Therefore, alternative activities may be incorporated into laboratory experiments to mitigate active participation. Moreover, these may (aim to) replicate on-the-job leisure activities to resemble real work situations more closely, particularly in [task application area three](#).

Yet, more research is needed to determine which outside option is most appropriate for their purposes, i.e., in terms of *realism*, *comparable value to all subjects*, *implementability*, and *effectiveness in mitigating the impact of non-pecuniary incentives* (both [activity and purpose related](#)). Besides, the question of which implementation features of the outside option are most beneficial in this respect also deserves more attention.

Table 2.1: **Frequently employed outside options:** Exemplary selection of outside options used in the literature together with their *degree of realism* and whether they are of *similar value to all* study participants (specified on four levels: No/(No)/(Yes)/Yes).

Alternative activity	References	Comment	Realism	Similar valuation
Surf the Internet	Bonein & Denant-Boèmont (2015), Corgnet et al. (2011), Corgnet et al. (2016), Houser et al. (2017), Kessler & Norton (2016), McMahon (2015)	Cyberslacking is very prevalent in the modern workplace, see the study by OfficeTeam (2017)	Yes	Yes
Read personal documents or magazines	Masclat et al. (2015)	Subjects may be informed upon signing up for the experiment or later reminded by email to bring along paperwork they would like to read. If subjects forget to do so, their alternative activity is less intense.	Yes	(Yes)
Read newspaper or magazines	Charness et al. (2013), Corgnet, Hernán-González, & Schniter (2015), Eriksson et al. (2009), Rey-Biel et al. (2018), Swenson (1988)	Becoming less prominent; depending on the range of print media offered, the individual valuation of subjects may vary	Yes	(Yes)
Watch pre-selected popular YouTube videos	Hayashi et al. (2013)	Within the bounds of a universities rule to Internet access (individual valuations may vary according to access restrictions)	Yes	(Yes)
Play a game	Swenson (1988)	Computer games (Tetris, Snake, Minesweeper, etc.) ^a or card games (Trivia, etc.)	Yes	(No)
Press an incentivized time-out button	Blumkin et al. (2012), Eckartz (2014), Erkal et al. (2017), Mohnen et al. (2008)	Paid inactivity ^b is quite untypical and a very abstract form of "leisure activity"	No	Yes
Alternative activity to earn money	Corgnet et al. (2011), Dijk et al. (2001), Erkal et al. (2017)	A dissimilar incentivized task or the same task yet under a different incentive scheme	No	Yes
Leave the laboratory early	Abeler et al. (2011), Dickinson (1999), Erkal et al. (2017), Falk & Huffman (2007), Rosaz et al. (2016)	Heterogeneity in quitting behavior: missing control over alternatives activities	(No)	(No)

^a Even the ball-catching task from Gächter et al. (2016), which has been perceived as enjoyable by 67% of the subjects in a study presented in Chapter 3, could serve as an outside option. It has the advantage that the experimenter can track how much "fun-effort" subjects provided in the alternative activity;

^b In the real workplace, some might be inactive while getting paid. However, getting rewarded for being inactive is not very common.

2.4.3 Skills and Character Traits Should Be Irrelevant

Apart from measurement errors that may emerge from omitting the impact of non-pecuniary incentives, an additional source may be the approximation of genuine effort by output. The commonly assumed one-to-one correspondence between provided effort and output produced supposes a homogeneous production function and a homogeneous cost function *across all individuals*. However, study participants differ not only in their motivations – but also in skills and personality. Thus, substantial heterogeneity in effort costs and production capabilities may arise. Therefore, a further criterion for designing tasks is that *cognitive and physical abilities* and *personality* should not determine task performance ultimately. This concern has been recognized in the literature, and several authors proclaim in reference to the particular task used that skills are irrelevant to the fulfillment of that task. Corresponding evidence is, however, rarely provided.⁶⁴ There is also very little literature to date on which subject characteristics may compromise effort measurement in which task (types). Addressing this gap is the focus of Chapter 4, which aims to identify the determinants of effort expended for various tasks. At this point, a number of design criteria and practices to decrease the above-mentioned heterogeneity across study participants are presented.

2.4.3.1 Equalize the starting conditions

Typical real-effort tasks require a variety of abilities reaching from mathematical skill to creativity (see also Section 1.3). When a task requires a particular one of these, subjects who possess that ability to a greater degree incur lower effort costs. With regard to the [previous consideration on skill-demand balance](#), it turned out that for tasks requiring greater skill, subjects with higher skill levels are more likely to enter the flow state. Beyond their ability advantage, these competent subjects may additionally benefit from the performance-enhancing effects of flow.

A similar argument may apply to tasks that require certain character traits such as self-control, perseverance, or patience. Confidence in being able to accomplish a task can also vary by gender, as shown

⁶⁴For example, consider [Rey-Biel et al. \(2018, p. 8\)](#) who note that their “task is inspired by the data entry task of Gneezy and List (2006). Our task is also similar to [Abeler et al. \(2011\)](#), where participants had to count the number of zeros in tables that consist of 150 randomly ordered zeros and ones. Such tasks are mainly effort-related and not skill-related, i.e., success in such a task is mainly attributed to hard work more than to individual skill.”

by Günther et al. (2010), who examine the wage gap between men and women and find that results are highly task dependent. Accordingly, women respond less to competitive incentives in “masculine tasks,” as men do in “neutral tasks,” and more strongly than men do in “feminine tasks.” The authors interpret their findings to suggest a “stereotype threat explanation,” such that “women tend not to compete with men in areas where they (rightly or wrongly) think that they will lose anyway – and the same holds for men, although to a lower extent” (p. 395). When studying gender differences, the nature of the task thus appears decisive.

Concerning application area one and in particular area two, it is worthwhile to ensure that all study participants have similar starting conditions already *by the design of the task*. That is, the task is deliberately designed so that task completion does not require greater skill, favor a particular personality, or reward prior knowledge. In this regard, simple tweaks in the experimental implementation of a task may already help equalize the preconditions between study participants. For example, if a task requires higher concentration, differences in the ability to concentrate can be mitigated by providing earplugs to participants in the experiment. Based on participant feedback in a trial study, these were provided to study participants in the experiment presented in Chapter 3 and were highly appreciated by the subjects.

Section 2.3.4 discusses the greater confidence in success of *subjects motivated by achievement*. A match between the skill demand of a task and the own skills can spur these individuals to higher performance and facilitate their transition to the flow state, in which their effort cost approach zero. Conversely, *individuals motivated by fear of failure* become worried, stressed, and afraid of failing and do not get anywhere near flow in the first place. To level the playing field and *prevent an advantage of those motivated by achievement*, one can choose tasks for which none of the study participants have prior experience or can (more quickly) develop “greater confidence in their own success.” To achieve this, one can resort to tasks that are so simple that even those who are not confident of their success can solve them readily and thus perform well (if they want to). This is particularly the case for tasks that make only low demands on one’s abilities so that they are feasible for all subjects, and none of them becomes fearful of failure.

To account for any remaining individual differences, it is useful to obtain information about the effort level of each subject when there is no (monetary) incentive to perform. This “*baseline performance*”

can then later be used to normalize the results of subsequent treatments, as applied, e.g., by [Imas \(2014\)](#).

2.4.3.2 Decrease situation-outcome expectancy and increase action-outcome expectancy

In Section [2.3.1](#), it is pointed out that if subjects have a *high situation-outcome expectancy*, they believe that the outcome is predetermined by the situation, and their actions will not make a difference: fate is given, and thus any effort is futile. Examples are provided of how a *high situation-outcome expectancy* can occur when subjects believe i) that they do not have sufficient skills to complete the tasks successfully or ii) that the outcome is the result of chance and beyond their control. In either case, this leads to a low motivation to complete the task.

The following considerations can be made regarding both cases: If a *task requires only low skills*, subjects are less likely to simply give up on it; the same applies if the *output generation is sure and follows comprehensible procedures*. This emphasizes the importance of a *trial round*. By allowing subjects to get to know and become familiar with a task, they can observe that their efforts actually *do produce an outcome* (despite their lower skill level or any chance elements). The high situation-outcome expectancy can, therefore, be substantially appeased and reduced by the practice round.⁶⁵ Becoming familiar with the task procedure may also increase subjects' *action-outcome expectancy*, making them feel more capable of completing the task and, consequently, achieving earnings (see also [Heyman & Ariely, 2004](#)). Another benefit of a practice round is that, if sufficiently long, it reduces confounding effects due to curiosity and learning in the subsequent effort measurement, especially for repetitive tasks.

⁶⁵In economics terminology, the trial round allows subjects to update their prior assumptions about the task.

2.4.4 No Learning Effects

In the literature, multiple tasks are reported to be highly susceptible to learning-by-doing effects when repeated over several rounds. For example, Gill & Prowse (2015) report a 16% learning effect in their *slider task* over ten repetitions; Benndorf et al. (2014) identify performance increases of 28% in the *word-encryption task* by Erkal et al. (2011) over three repetitions, and of 29% in the *counting-zeros task* by Abeler et al. (2011) over four repetitions; Wozniak et al. (2014) observe non-negligible learning effects both in a math task and a word-formation task. Since subjects evidently gain experience in executing a given task, a *baseline treatment* may permit to control for any improvements in task-performing. Tasks may also be *constructed in a way to obstruct or suppress learning*. Finally, a *practice round*, to familiarize subjects with the task, and letting subjects complete the *task only once* may help mitigate strong initial learning effects.

Include a control group in the treatment design. To control for any improvement in task performance with time, Carpenter & Huet-Vaughn (2017) recommend to include a *baseline treatment*. They further demand to randomize and subsequently control for the treatment order in within-subject designs in order to differentiate between treatment and learning effects.

By the design of the task. A number of authors advise *designing tasks* in a way to *prevent learning effects from the onset*, such that task performance does not improve (much) with time (Abeler et al., 2011; Benndorf et al., 2014; Dickinson, 1999; Fu et al., 2015; Heyman & Ariely, 2004). For example, Benndorf et al. (2014) are able to reduce learning effects in the *word-encryption task* of Erkal et al. (2011) from 29% to 8% by reshuffling the encryption table in addition to the word being encrypted. The authors further note that learning behavior in their double randomization task does not differ by gender, gradually slows down and eventually stops. Heyman & Ariely (2004) implement an elegant experimental design devoid of any possibility for learning. Subjects have to solve a *number adding puzzle* and could familiarize and become comfortable with the task in four trial rounds. Thereafter, subjects receive an *unsolvable puzzle*. The authors employ the amount of time, subjects spent until giving up, as the ultimate effort measure.

Dickinson (1999) find in a pilot experiment for a *transcription task* that had to be performed repeatedly

over several days that subjects were able to recall significant portions of the text being transcribed. In the final study, subjects had to transcribe a new paragraph of similar difficulty on each of the successive days. This example documents the importance of conducting pilot experiments (including asking participants for feedback) to uncover possible learning effects due to the task design. In simple, monotonous tasks, learning effects tend to be small (consider the single-slider task presented in Section 2.5).

Single period of effort provision. It was pointed out earlier that completing a task only once suppresses any motivation to exceed one's performance in the previous round(s). An additional advantage of an effort measurement in *one period only* is the lack of learning effects across rounds. In case that the experimental design requires several periods, [Mohnen et al. \(2008\)](#) propose to crop off the periods at the beginning and the end. Such truncated periods can then be used both as an individual measure of ability for every participant and to examine how learning effects, fatigue, and declining concentration and motivation affected the results. The approach should work well on the *word-encryption task with double randomization* introduced by [Benndorf et al. \(2014\)](#), who report that subjects quickly learn at the beginning of the task, but then stop to improve their performance further.

2.4.5 Elastic Effort Response

Research in the [area of task application one and three](#) investigates how individual effort varies under different incentive schemes and in different contexts. For this purpose, the generated output, which usually serves as a measure of the effort expended, must be sufficiently elastic across the incentives under consideration. By way of example, [Giusti & Dopeso-Fernández \(2018\)](#) observe that effort provision depends on stake size under performance pay, but much more so in a less challenging task. However, whether this finding results from the (nature of *this* challenging) task or the subjects' peculiarities remains an open question.⁶⁶ More elaborately, it is unclear if effort provision in difficult tasks is inelastic with respect to stake sizes because subjects *do not raise* or *cannot raise* their output. In the first case, subjects may make no attempt for motivational or volitional reasons or due to mental or physical limitations. In the latter case, the production function of the task is rather insensible to increases in effort. Thus, even if subjects could or would like to generate higher output, a stronger effort would not help much.

For the task to accurately capture the subjects' actual effort, their *effort must matter*, and *boundary effects must not limit its provision*. Ways to achieve these sub-criteria are described below.

2.4.5.1 Effort Needs to Matter

Output production is sufficiently sensitive to effort. In the *multiplication task* of [Dohmen & Falk \(2011\)](#), each individual calculation takes a relatively long time, so the resolution of the effort expended in the task is not particularly fine-grained. Especially if the overall task duration is short, it will be hard to distinguish subjects from each other in terms of effort exerted (a subject who has almost succeeded in solving three equations is equated with a subject which has barely solved two equations). In order to discriminate subjects sufficiently well on the basis of their exerted effort, the task must be *sufficiently sensitive to effort*. This implies that an increase in input (effort expended) must lead to an increase in output *and* that even small changes in effort must lead to changes in output, so that a fine-grained resolution of the effort measurement becomes possible. "Solving CAPTCHAs" by [McMahon](#)

⁶⁶Summarizing the results of several literature reviews, [Kachelmeier et al. \(2008\)](#) find that incentive effects are predominantly task and setting specific. In light of the *ibid.* comments, this hardly comes as a surprise.

(2015) represents a task in which one can adjust the task difficulty and thus the required effort almost seamlessly. Increased tilting, overlapping, and blurring of the numbers and letters contained in the CAPTCHAs makes reading increasingly tricky and raises the effort required to solve the task (see Figure 1.4 for an illustration).

Obtain performance profiles. Measuring the same output for three different subjects does not mean that they all exerted themselves to the same degree: i) a very capable subject might have found the task interesting at first and produced a high output, but then got bored and switched to the outside option; ii) a very dedicated but less able subject produced low output throughout the duration of the task; iii) another subject had anxiety about the task and thus aversion towards it. When she finally overcame these, she eventually achieved the same output as the other two – but under high time pressure and effort.

Despite these differences in how much subjects exert themselves in general and over time, the *same output* is observed for all three subjects. The arguments raised so far in terms of differences in *skill level*, *motivation*, and *volition* suggest that the design and implementation of a task can influence the research results collected with it. For each of the exemplar subjects, one can come up with a modification of the task that may lead to divergent findings (e.g., i) remove the outside option, ii) use a simpler task, iii) use a less aversive task).

One solution to overcome this is to employ a task that is incredibly dull and monotonous, and thus tedious, toilsome, and tiring (but the task may still not be so grueling that it becomes aversive for some subjects and requires overcoming to complete).⁶⁷ An alternative is to use a task that allows *recording the subjects' performance profile* over time. The detailed information gathered thereby provides an indication of how comparable the data collected for the individual subjects are, or to what extent the task design and implementation determined the output generation of the subjects. Besides, performance profiles could also be interesting for the actual analysis.⁶⁸ To capture informative performance

⁶⁷Corgnet, Hernán-González, & Schniter (2015) and Eckartz (2014) demonstrate that incentive effects increase with the presence of *leisure activities or a paid outside option*. If a useless task with the described properties is adopted and the subject nevertheless performs the task over the entire, long period of time, thus disregarding the outside option, this “must” basically be due to incentive effects.

⁶⁸In particular, if the study design may lead to transient psychological effects, e.g., due to an unexpected, sudden pay raise, it may be of interest to choose a task that allows subjects' performance to be tracked over time.

profiles, the effort measurement must be sufficiently fine-grained. For example, the *ab-typing task* of [Berger & Pope \(2011\)](#) enables a quasi-continuous effort measurement and thus a high resolution in measuring effort.

Long period. If the effort resolution of a task is somewhat coarse, an *extended task period* is helpful. A prolonged task duration allows obtaining a smoothed performance profile in which transient effects of increased effort due to curiosity or social desirability are attenuated and eventually are no longer noticeable (see also discussion earlier in Section [2.4.1.2](#)).

2.4.5.2 No Boundary Effects

Another sub-criterion in designing tasks is to avoid boundary effects. To achieve this, the **exertion of effort must be sufficiently costly**.⁶⁹ Otherwise, the subjects will expend the maximum effort regardless of any cost. To conduct a valid effort measurement, it is further recommended to design the task in such a way that **no ceiling effects** occur. This means that the range of possible effort levels is not restricted in any way. Considering the previous discussion about [choking under pressure](#), this also implies that the piece rate may not be set too high. Otherwise, subjects will fail at the task because they are as if paralyzed by the high amount at stake.

For some tasks and experimental setups, **additional performance measures** can be obtained without much expense. For example, [Abeler et al. \(2011\)](#) use both the number of tables for which subjects correctly counted the number of “1s” and the time subjects decided to work on the task. For instance, in other tasks, one could use the total number of answers submitted in addition to the number of correct answers to obtain a more precise measure of effort.

⁶⁹[Fahr & Irlenbusch \(2000\)](#) further note that effort must also be sufficiently costly to truly generate a sense of entitlement to the rewards.

2.4.6 Statistical Significant Results

The output that subjects produce in a task must be easily measurable, and the task must allow for substantial variance in output production (see the case for [fine-grained effort resolution](#) and [no boundary effects](#)). Beyond these basic task properties, further precautions are helpful to obtain insightful and statistically significant results.

2.4.6.1 Incentive Effects Are Large in Relation to Uncontrolled Variation

To induce preferences, incentives must be *monotonic*, *salient*, and *dominant* (Smith, 1982). That is, i) subjects must prefer more to fewer rewards, and without satiation, ii) their rewards must depend on their individual actions (and possibly those of other subjects) and be prominent and comprehensible, and iii) changes in their utility mainly derive from their rewards, such that all other influences are negligible. Smith's "precepts of economic experiments" are frequently referred to in the experimental literature. However, how large monetary stakes must be for the reward structure to dominate the subjective costs (or values) associated with completing a (particular) task remains rather unresolved, as the remark by Carpenter & Huet-Vaughn (2017) may illustrate:

"... without knowing the parameters of participants' actual utility functions, it may be hard to calibrate the incentives of a real effort experiment – setting piece rates, for example is often a "crap shoot." Should you pay them one cent per keystroke, ten cents or ten dollars? Imagine that although participants have heterogeneous costs of effort in a given task, the functions are all relatively flat (despite being increasing and convex). Without knowing this, it could easily be the case that piece rates or other marginal incentives are set too high or too low and everyone either works as hard as possible or as little as possible. Unless the experimenter can identify the incentive "sweet spot" treatment effects will be artificially negligible by design." (Carpenter & Huet-Vaughn, 2017, p. 6)

To achieve that incentive effects are large in relation to uncontrolled variation in task performance, the [stake sizes must be sufficiently large](#) (appropriate to the task at hand). Besides, very low piece rates invariably result in low payoffs, making any effort to generate output seem futile to the subjects.

Conversely, very high stake-sizes may lead to [choking](#), as noted earlier ([Ariely, Gneezy, et al., 2009](#); [Corgnet et al., 2016](#); [Pokorny, 2008](#)). A careful adjustment of the incentive scheme based on [pilot sessions](#) seems indispensable. Reporting the results of these (task-specific) testing in the Appendix adds towards achieving more clarity on the (non)monotonic relationship between monetary incentives and performance and represents a valuable contribution to our field. This will help build a repertoire of tasks for which the “incentive sweet spot,” as termed by [Carpenter & Huet-Vaughn \(2017\)](#), is known. In particular, those tasks that have an incentive sweet spot that is “affordable” (from an experimenter’s perspective) and encompasses a wide enough range to allow variation in the level of stakes will be beneficial for research.

2.4.6.2 Reduce Noise in the Effort Measurement

As mentioned earlier, unsystematic variation in effort provision may have various sources, including activity-related incentives, purpose-related incentives other than monetary rewards, individual skill level, or learning. Beyond the previously mentioned design measures to mitigate these sources of confounding, it is reasonable to [treat the subjects precisely the same](#) so that they would, at least in principle, have the same effort and production capacities. This means that the subjects are confronted with the same experimental procedure and the same technical infrastructure.⁷⁰ If on-screen instructions are not sufficient or spoken instructions are preferred for experimental design aspects, prerecorded audio instructions help keep across-session variation low. They bear the additional advantages that i) the experimenter can be sure that all study participants have heard the complete instructions and in the identical, reproducible way, and ii) all subjects in the session know that all other subjects have received the same instructions.

Furthermore, the extent to which repetitions of a task are alike can introduce noise into the effort measurement. This becomes quickly apparent if one compares the text transcription task of [Dickinson \(1999\)](#), in which various long text passages have to be transcribed, with the ab-typing task of [Berger & Pope \(2011\)](#), in which the keys “a” and “b” have to be pressed alternately on the keyboard. [Monotonic tasks](#) thus have the advantage that the task is exactly the same in every repetition. Therefore,

⁷⁰The subjects’ workstations (cubicles) are ideally surrounded by sound-absorbing walls to reduce visual and acoustic distractions, and are uniformly equipped with the same configuration of PC, mouse, and keyboard.

the variation in task execution is close to zero (for each subject as well as to a similar extent across subjects).

2.5 Applying the Criteria and Practices: The Novel Single-Slider Task

This section introduces a new real-effort task, the *single-slider task*. It was deliberately designed to conform to the previously outlined task-design criteria and practices. The computerized task belongs to the class of [useless](#) and [unrealistic](#) tasks since it neither produces any tangible, meaningful output nor resembles any activity of real working life.

The procedure of the single-slider task is extremely simple. On the computer screen, the subjects are presented with a line containing a slider whose position can be changed by clicking on it and then moving it along the line with the mouse cursor. The initial position of the slider is at one end of the line. To score points, the subject must move the slider from one end to the other, as illustrated in [Figure 2.10](#). When the slider is correctly positioned at the other end of the line, the line lights up green to signal and confirm that a point has been scored. After that, the slider automatically jumps back to the beginning of the line, and the task starts over again. With the laboratory software oTree, the task can also be performed analogously on a tablet. In this case, any movements are performed with the finger (tap and drag). Before the task begins, the subjects are asked whether they are left- or right-handed. The direction of movement of the slider is adjusted accordingly, i.e., from left-to-right for right-handers and vice versa for left-handers.⁷¹

As noted above, the single-slider task was intentionally designed to abide by the [design criteria and practices](#). Therefore, it possesses a number of desirable properties with respect to [task application areas one and two](#), which are considered below.

To curb activity-related incentives, the task follows in the “tradition of using mind-numbing tasks” as formulated by [Heyman & Ariely \(2004, p. 790\)](#). To avoid curiosity, the task is *genuinely unexciting and very repetitive*. To discourage any enjoyment in task-performing, the task is utterly *tedious, toilsome, and tiring*. Moreover, it is *not challenging* at all wherefore subjects are not expected to enter the flow state. To mitigate the impact of potential self-evaluation consequences, the task *neither generates a valuable outcome nor provides stimulating feedback*. To address other-evaluation consequences that

⁷¹Detailed English and German instructions for the task can be found in [Appendix A.1](#).

"links-rechts" Schieberaufgabe

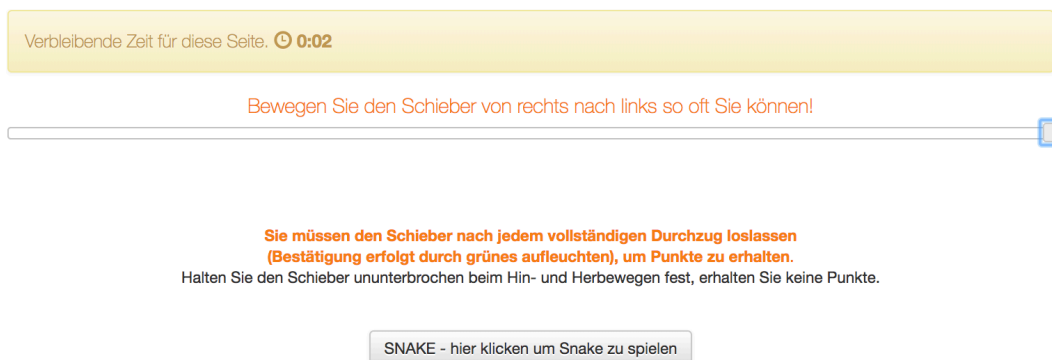


Figure 2.10: *The single-slider task*: To score points, subjects must repeatedly move a slider from one end of the line to the other. Before the subjects reach the instructions for the task, they are asked whether they are left- or right-handed in order to adjust the direction of sliding accordingly. The implementation depicted includes the possibility to switch to the game *Snake* as an outside option.

may further provide purpose-related incentives and likewise encourage effort, the task is designed to be on the edge of being *grueling* and *pointless* in every respect. Greater engagement in the single-slider task for “approval seeking” thus becomes rather improbable.

By design, *neither the individual ability nor character traits are expected to impact individual effort costs significantly*. Furthermore, the task basically leaves *no room for learning*.

The single-slider task is *sensitive to effort*, such that an increase in exerted effort seamlessly translates into an increase in output. In other words, the production function is sufficiently fine-grained to track any effort that is provided by a subject. To avoid boundary effects, *provision of effort comes with considerable costs* (physical through the movement of the slider, and mental through the meticulous course of motion). Also, by design, the task does not contain any upper limits on effort provision, which could result in ceiling effects.

Finally, the task proves rather easy to implement in experiments. As it is outright simple, the task can very easily be explained to and conceived by subjects and does not require any previous knowledge. A short explanatory note is nevertheless displayed below the task while it is being performed. To prevent any difficulties in task handling, it is recommended to include a practice round. To ensure comparability of results, subjects should further be equipped with the exact same technical infrastructure and input devices within an experiment (i.e., identical computer, screen, and mouse or the same type of tablet). Code for the laboratory software oTree and participant instructions for the

single-slider task are available to interested researchers upon request.

To summarize, considering the single-slider task in terms of the design criteria and practices, it satisfies these as far as possible. Apart from the fact that it can only be performed digitally, no major disadvantages are initially apparent. In the study presented in Chapter 3, subjects asserted that the task is very strenuous and physically demanding, confirming that it provides a valid measure of effort. Thereby, the properties of the task favor a reduction of unsystematic variation in the effort measurement. In consequence, the replicability of results obtained with the task improves in comparison to similar tasks. As a useless, unrealistic task, it thus proves to be very suitable for laboratory experiments conducted in the [task application areas one and two](#).

2.6 Conclusions

Completing a task involves effort. Yet, the reasons for providing effort can be manifold. Section 2.3 addresses this and sheds light on why subjects exert effort to complete a task from a motivational psychology perspective. Triggered by different activity-related incentives and purpose-related incentives, the motivations can vary greatly for different tasks and in their diverse application areas.

Depending on the goal of the latter, the desirability of non-monetary incentives varies considerably. As described in the introduction to this thesis, the goal in [task application area three](#) is realism. To achieve congruence between the laboratory and a real work environment, non-monetary incentives present in the latter must not be absent from the experiment to influence the behavior of study participants.

The situation in [task application areas one and two](#) differs substantially. In these, the goal is an unbiased effort measurement or unbiased endowment, which is similar in size for all subjects. Various recommendations in this regard can be found in the literature to approach the “neutral tasks” preferred for these purposes. For example, [Eral et al. \(2017\)](#) point out two approaches to diminish the impact of non-pecuniary incentives: *directly* by making the task less attractive and enjoyable, i.e., to decrease the benefits subjects obtain from task-performing, and *indirectly* by raising the opportunity cost of

working.⁷² A comprehensive examination has been missing so far and is proposed in this chapter. As a synthesis of the literature on real effort, a set of *design criteria* and *design practices* is offered to enhance existing tasks and assist in developing new tasks. The criteria and practices allow expanding the range of tasks with greater control over the effort cost and output production.⁷³ Without prioritizing among them, the criteria allow for a more comprehensive view on and better assessment of empirical results obtained using real-effort tasks.

To provide an example of the application of the design criteria and practices, a novel task is presented that adheres to them. The [single-slider task](#) is particularly developed to provoke as little heterogeneity in effort costs and production capabilities as possible.

For task application area one and two, the presented criteria and practices suggest that extremely simple and monotonous tasks should be employed. However, some may argue that such tasks are too far removed from real work. Nevertheless, they can serve to reveal some very thought-provoking matters of fact. As an example, [Giusti & Dopeso-Fernández \(2018\)](#) compare two tasks with different levels of difficulty and skill requirements. The authors find that “consistent with the literature on reference point shifting (Chen and Rao, 2002; Baucells et al. 2011), the effect of a decrease in piece rate, which follows a previous increase is negative and much stronger than the effect of an increase following a decrease” (p. 31).

As mentioned in Section 1.5, a (hasty) generalization of such interesting results obtained in the laboratory should be abstained from (here, for example, transferring the findings to the fruit-picking industry, where seasonal wage fluctuations are prevalent, coupled with recommendations such as “Refrain from increasing the piece rate if maintaining the increase is not feasible in the long run.”). In order to be able to [extend the scope of the results](#) obtained in the laboratory for possible interest-

⁷²The first approach is embraced in Section 2.4.1, which aims at curbing activity-related incentives and flow; the latter recommendation is picked up in Section 2.4.2 in the context of outside options to mitigate undesired purpose-related incentives. Further suggestions can be found, for example, in [Charness et al. \(2018\)](#) and [Corgnet, Hernán-González, & Schniter \(2015\)](#). Since providing guidance on task design and implementation is not the focus of these contributions, their coverage is, therefore, somewhat more limited.

⁷³Up to now, the choice of tasks with greater control over the cost of effort and output production is fairly limited. [Gächter et al. \(2016\)](#) introduce a ball-catching task specifically to fill this gap. However, the task exhibits several disadvantages: It is perceived by subjects as entertaining (see Section 3.4) and suffers from a non-negligible skill-bias, as it in essence represents a simple optimization problem (see Section 4.3). Conversely, the hand-gyrometer task employed by [Imas \(2014\)](#) appears more promising. However, this task is also much more complex and time consuming to implement, as it requires special equipment and most likely a one-to-one attendance of experiment staff.

ing real-world applications, replication in a field experiment is indispensable. Due to the significantly greater complexity and costs involved, the findings obtained in the laboratory are nevertheless of great interest. A preceding, further corroboration and substantiation of the laboratory results by repeating the experiment with other tasks is certainly expedient.

Understandably, certain trade-offs between the different criteria and practices are necessary when designing and implementing tasks. For example, by having an interruption-free, endless nature, one can try to generate or emphasize *aimlessness* and *meaninglessness* in a task (recall the endless variant of the [wire-loop game](#)). On the other hand, the transition to the flow state is favored when the activity can be performed without any disruption. Deliberately built in interruptions or pauses can try to hinder this so that the subject gets the occasion to question the meaningfulness of the task. This, in turn, *hampers the transition into flow*. A weighing up of the criteria and practices conditioned by the [application area](#) and the research question is inevitable.⁷⁴

Irrespective of the choice of the task and its implementation, it must also be ensured that the subject pool is selected to match the [task application area](#) and research question. Thus, in retrospect of the [remarks on motivational psychology](#), the subjects should possibly have a somewhat similar background, or not, depending on the application. This becomes even more evident when using real-effort tasks in field experiments to bridge the gap between laboratory and reality, i.e., real labor markets. Yet, the [scope of the results obtained in each case](#) is evidently limited.

A large part of the literature on real effort was reviewed in preparation of this first compilation of design criteria and practices for real-effort tasks. However, the list is, to some extent, subjective and not entirely exhaustive. Moreover, depending on the task application and experimental design used, some task properties may have a greater influence on experimental results than others. Consequently, certain design criteria and practices may be more relevant and have a greater impact on the accuracy (or even validity) of effort measurement, respectively. However, this study does not attempt to weigh each criterion and practice against the others. Further research is desirable in this regard to more accurately assess their importance and, if necessary, prioritize among them.

⁷⁴A compromise between the two practices would be, for example, (across all participants) randomized interruptions during the course of the task, whereby the display of any spurring intermediate results should be explicitly avoided. Instead, the [display of a short pause indication to stretch and recover](#) from the task could be considered.

Despite the improvements in measuring effort that can be achieved by applying the criteria and practices, certain tasks may in general not be recommended for research, in particular in terms of [task application areas one and two](#). Some of them require only a minimal amount of effort to complete. They are, therefore, hardly suitable as a measure of effort. For other tasks, the subjects' abilities, personality, and motivation may have a strong impact on the measured level of effort. In such tasks, the individual costs for exerting effort may vary significantly across subjects. Accordingly, observed effort levels or generated endowments may carry a systematic bias that cannot necessarily be offset by simple randomization (see also the [discussion in the introduction to this thesis](#)).

Therefore, further investigation is warranted to shed light on these task properties and to identify tasks suitable for experimental research. This forms the focus of the remainder of this thesis and will be addressed in two steps: Chapter [3](#) seeks to confirm that tasks differ along motivational dimensions and in terms of the type of effort they involve; thereafter, Chapter [4](#) aims to pinpoint determinants of effort exerted in order to find tasks that represent sound measures of actually exerted effort.

3

Comparing Real-Effort Tasks

3.1 Introduction

A large variety of tasks have been introduced in the experimental literature. They differ in particular in two aspects: *First*, how much voluntary effort they induce, such that subjects exert effort irrespective of the incentive; *second*, whether earnings can be achieved without greater effort. The design of a task in this regard may have the following implications for the effort measurement when the task is used in [application areas one and two](#). In addition to monetary incentives, subjects may be motivated by [activity-related incentives](#) and undesired [purpose-related incentives](#), although to varying degrees, so that a certain amount of noise can hardly be avoided. However, some tasks are more susceptible to this and may yield a biased mapping of the incentive to effort provision. Moreover, experimental results will be distorted if, for some study participants, the completion of the task requires little effort

due to an advantage in [skill level or personality](#). Further, if the task, in general, does [not require significant effort to complete](#), the task may not be a good measure of effort altogether. The choice of task, therefore, has substantial implications. To examine and evaluate tasks in relation to the outlined dimensions, a new methodology is introduced in this chapter.

Section 2.3 presents a view from motivational psychology why subjects may provide effort in tasks. To recap, in addition to monetary incentives, effort provision can be motivated both by activity-related and purpose related incentives. These reside in the performing of the activity and the consequences of its outcome. In any given task, both activity-related and (even several) purpose related incentives may occur at once. Altogether they form a motivating potential that encourages (voluntary) effort in addition to any monetary incentive. The strength and extent of these additional incentives varies greatly between different tasks. Even more, their influence on subjects' actions – i.e., their effort generation – varies considerably across individuals. Consequently, the outcome produced by any subject is motivated by an individual cocktail of the monetary incentive and additional activity-related incentives and purpose related incentives.

In task application area three, this may be intended to generate congruence with real work in order to answer a specific research question. However, for addressing generic research questions in task application areas one and two, the influx of influences can be highly undesirable. As described earlier, accounting for these unintended additional effects is not necessarily feasible or easily accomplished. Therefore, the objective in task areas one and two is to select a task that does not induce additional, undesirable activity- and purpose-related incentives, or to eliminate or reduce them as much as possible by applying the [design practices](#) proposed in the previous chapter.

As noted above, several additional activity-related incentives and purpose related incentives may be present and act at once. To get an impression of which of these incentives are at work and to what extent they may motivate effort, this chapter introduces a new tool, the *real-effort task survey*. This survey is based on a subset of the design practices to address non-monetary incentives. The survey asks subjects after they have completed a task how they perceived it. More specifically, the survey indirectly queries the subjects' assessment of the extent to which the design practices have been fulfilled. By aggregating the assessments of all study participants, one can gain an impression of the extent to which a task actually fulfills the design criteria or whether subjects are motivated by activity-

and purpose-related incentives.

As mentioned above, tasks also differ greatly in the extent to which they require effort to complete. Therefore, the survey contains two additional items for subjects to assess how physically and mentally demanding a task is.

Overall, the survey allows estimating the (expected) influence of non-monetary incentives caused by the task design in advance of a study, as well as to what degree the task actually demands effort from the subjects and thus represents a suitable measure of effort. To ensure that this approximation also holds for the participants of the ultimate study, the subjects completing the survey should ideally come from the same population as the former. Collecting the survey for several tasks with the same group of subjects allows to compare the tasks and to settle upon one for a final study design. The survey's assessment of motivational influences beyond money may further be useful to control for these in a study.

The real-effort task survey can be used to examine any type of task, both computerized and manual. To implement the survey, subjects of the population of interest perform the task in return for money and fill in the survey subsequently. To compare multiple tasks, subjects process one at a time and complete the survey after each task. The task comparison can be refined by the method of *multiple comparisons with the best* (MCB) of Hsu (1996).¹ Applied to the subjects' assessments, it can determine a task that most closely matches the design practices underlying the survey.

Besides, a greater consistency among subjects' assessments of the tasks is beneficial: With a greater consensus of ratings for the motivational items, the influences of activity- and purpose-related incentives can be expected to be of similar magnitude; analogously, a greater similarity of the subjects' ratings of the effort-related items indicates that the task will be similarly effortful for them. For tasks for which there is only a low agreement between the subjects' ratings, the opposite is true: Greater differences in motivational influences, or how strenuous the task is for individual subjects, can be expected. This is accompanied by an increased noise in the effort measurement (*ceteris paribus*). Again, the MCB method proves useful in broadening and deepening the task comparison: Applying the method a second time to the variances rather than the means allows one to compare the consis-

¹A closer description of the method can be found in the [results section](#) when it is applied, as well as in Appendix [B.3.2.4](#) in detailed form for one of the items of the survey survey.

tency between subjects' ratings; in this way, tasks can be identified for which one can expect only a comparatively small amount of noise to be present in the effort measurement.

The real-effort task survey was first applied to compare tasks in June 2018. Two hundred and forty-eight students participated in a laboratory experiment and completed a diverse set of tasks. The tasks were selected based on the classification of tasks by ability and personality presented in Section 1.3. To reflect the wide range of tasks available, one or two tasks were included from each category described. To give an outlook on the results, the subjects viewed the selected tasks as very different. This holds both for the motivational and effort dimensions of the survey. Thus, subjects perceive some tasks as more motivating, others as less motivating, and that they differ in the type and amount of effort required. Moreover, subjects' perceptions of one and the same task displayed a high degree of heterogeneity. Finally, support for the validity and the reliability of the real-effort task survey is found.

To my best knowledge, a comprehensive survey to assess *activity-related and purpose related incentives* and *effort-type* of a real-effort task has not yet been presented in the economic literature. Consequently, an objective evaluation and rigorous comparison of tasks along these dimensions are still missing. Merely, a number of authors pose questions to control for "intrinsic motivation" (e.g., [Dijk et al., 2001](#); [Giusti & Dopeso-Fernández, 2018](#); [Mascllet et al., 2015](#)). Others try to assess whether a sense of duty, pleasing the experimenter, or boredom is driving subjects' efforts ([Mascllet et al., 2015](#)) or how much effort subjects effectively exert, whether they feel stressed or get exhausted (e.g., [Dohmen & Falk, 2011](#)). These approaches tend to assess the state of mind, attitude and condition of the subject, rather than the properties of a task. Providing an objective evaluation of tasks is also not the specific target of these studies. Consequently, their treatment is far less detailed and does not detract from the approach presented herein.

Possibly closest to the approach in this chapter in the economic literature are the studies by [Fu et al. \(2015\)](#) and [Giusti & Dopeso-Fernández \(2018\)](#). [Fu et al. \(2015\)](#) study the impact of feedback on tournament performance both for groups and individuals. These authors define three criteria to decide among a selection of three tasks.² Their criteria touch upon three crucial dimensions of real-effort

²To assure that ability neither varies too much among study participants nor changes over time, [Fu et al. \(2015\)](#) demand that their task is *easy* and *does not require prior knowledge*. These requirements resonate with the design criteria defined for real-effort tasks proposed in Section 2.4. Second, the authors require that the task *precisely records the effort provided* (see

tasks; however, they do not go into particular depth. For example, concerning activity-related incentives, the criteria proposed by [Fu et al. \(2015\)](#) are insufficient to provide a comprehensive picture of the additional incentives at work. Furthermore, their criteria are also not embedded in a larger theoretical argument, nor do they paint a full portrait of the properties of tasks (they neither account for influences arising from purpose-related incentives, skills and personality, or learning, nor do they address the statistical significance of the results). Finally, [Fu et al. \(2015\)](#) do not use any particular methodology to obtain an objective task evaluation but rather provide a situational comparison of their task set based on their own three criteria.

[Giusti & Dopeso-Fernández \(2018\)](#) use a sub-scale of the *Intrinsic Motivation Inventory* (IMI) developed by [Ryan \(1982\)](#) in the context of *self-determination theory* (SDT), which has been employed in many adaptations over time. Yet, the scale is rather general and not specifically built to address real-effort tasks. The “interest/enjoyment” sub-scale used by the authors contains seven items that assess whether a subject perceives an activity enjoyable, boring, or interesting. The goal is to find out whether the activity is “intrinsically motivating” and, therefore, voluntarily performed. The dimensions captured by the IMI are important and are reflected in items of the real-effort task survey. However, as discussed in Section 2.3, the scope and concept around the term “intrinsic motivation” as used by the SDT is rather unfortunate and restrictive.³

Using the terminology introduced in the *extended version of Heckhausen’s Advanced Cognitive Motivation Model* by [Rheinberg \(1989\)](#) allows reasoning far beyond this: The items of the real-effort task survey are based on a variety of design practices which are specifically designed to eliminate *activity-related incentives* and *purpose-related incentives* or the *consequences that trigger them*. While items of the IMI cover a part of the activity-related incentives, purpose-related incentives are not considered by the IMI at all. Moreover, since [Giusti & Dopeso-Fernández \(2018\)](#) aim to study two different tasks in terms of incentive effects, comparing tasks in general and as detailed as in the present study is not

also Section 2.4.5). The effort-sensitivity of a task cannot be tracked through a survey completed by study participants. Both the dispersion in the output produced by the subjects and the resolution of the piece-rate for each unit of output produced are more viable sources of information in this regard (see the [design practice](#) in terms of sufficiently sensitive and fine-grained output production). Third, [Fu et al. \(2015\)](#) demand that their task is *sufficiently tedious* – which likewise resembles one of the presented [design practices](#) and is also found in the real-effort task survey.

³For a critical review of the term “intrinsic motivation” in self-determination theory, see [Rheinberg & Engeser \(2018\)](#).

their focus.⁴

Much more closely related approaches can be found in the motivational psychology literature with the *Questionnaire on Current Motivation* (QCM) by Rheinberg et al. (2001) and the *Flow Short Scale* (FSS) by Rheinberg et al. (2003).⁵ As the name of the QCM suggests, the questionnaire aims to determine a subject's current motivation with regard to an activity. Thus, its 18 items are designed to query how subjects perceive themselves in relation to an activity. The survey provides an evaluation of the *subjects' current attitude toward the task*. In contrast, the real-effort task survey seeks to evaluate the *properties of the activity*. On a similar note, the FSS inquires how subjects feel during the performing of an activity and assesses the different components of flow (see, e.g., Rheinberg & Engeser (2018) for a brief account). Like the QCM, the FSS inquires about indicators of a subject's functional state rather than the design specifics of an activity. Furthermore, although subjects entering the flow state are a concern in the design of tasks, activity-related incentives in terms of curiosity and enjoyment of the task, as well as undesirable purpose-related incentives, likely prove to be far more influential on subjects' propensity to exert effort. Therefore, to keep the real-effort task survey short and concise, it focuses on these additional incentives that may be present and motivate effort.⁶

The current study intends to contribute to the literature mainly in two ways. *First*, a new methodology to evaluate tasks is introduced. In practice, the real-effort task survey can serve as a tool i) to contrast tasks in terms of their fulfillment of countermeasures to non-monetary incentives and the type of effort required;⁷ ii) to identify a favorable task in conjunction with the method of multiple comparisons with the best by Hsu (1996); and finally, to confirm that a task possesses particular properties, for which can then be controlled in later analyses. In this way, the real-effort task survey can be helpful

⁴Giusti & Dopeso-Fernández (2018) compared a challenging arithmetic task (counting down from a very large number by repeatedly subtracting a fixed amount from the previous one) and a less challenging counting task (counting the number of occurrences of a specific letter in a passage of text).

⁵In the German-language literature on motivational psychology, these scales are known as the *Fragebogen zur aktuellen Motivation* (FAM) and *Flow-Kurzskala* (FKS).

⁶Having subjects evaluate the design practices to *prevent flow* is also far from straightforward, as noted in the description of the survey development in Section 3.2.

⁷The items of the real-effort task survey are based on the *design practices* proposed in Section 2.4 to *counteract additional activity-related and undesirable purpose-related incentives or the consequences that trigger them*. Hereafter, these design practices with their purpose are referred to synonymously as "countermeasures to non-monetary incentives." Thus, a survey item assesses the degree to which a task satisfies a particular countermeasure.

for researchers to check (established) experimental results across tasks.

Second, a set of frequently used task types is contrasted by applying the survey in a laboratory experiment. Based on subjects' assessments, the survey provides insights on the tasks' strengths and weaknesses: It reveals which of the tasks are more likely to lead to voluntary effort due to additional activity-related and purpose-related incentives and whether they require greater effort. Among the selection of tasks, the MCB method identifies the *single-slider task* to match the attributes of a favorable task in terms of [task application areas one and two](#).

The remainder of the chapter proceeds as follows. At first, the real-effort task survey is introduced in closer detail in Section [3.2](#). Thereafter, the research design and methodology employed in the study are outlined: Section [3.3.1](#) describes the experimental design, including the selection of tasks and the experimental procedure; Section [3.3.2](#) details the empirical strategy that builds on the distinct properties of the tasks in the selection. Thereafter, Section [3.4](#) presents the results of the survey-based task comparison. Section [3.5](#) concludes the chapter and offers alleys of future research.

3.2 The Real-Effort Task Survey

This section introduces the *real-effort task survey*, a novel tool for evaluating and comparing tasks. The survey is built upon a subset of the design criteria presented in Section [2.4](#). It inquires about subjects' perceptions of i) specific design properties that relate to design practices to counteract [activity-related incentives](#) and [purpose-related incentives](#), and ii) the nature and extent of effort required of study participants. The survey is applied in a (laboratory) experiment to determine the subjects' assessments of one or more tasks. Collectively, these (subjective) assessments provide the researcher with a comprehensive evaluation of a single task or multiple tasks.

To ensure that the survey provides valuable data, various provisions were made in its design. The next section provides details on the [survey design and item choice](#). Besides, specific steps are recommended when [conducting the survey](#). Thereafter, the [validity, reliability, and robustness of the results](#) to be obtained with the survey are only briefly addressed, as they are discussed at length in the results section of this chapter. Finally, possible [applications of the survey](#) are outlined.

Survey design and item choice. As a synthesis of the literature on real effort, several design criteria are presented in Section 2.4. The development of the real-effort task survey follows an explorative approach based on the design practices to meet these criteria. As previously noted, the compilation of design practices can be divided into *task independent* and *task dependent practices*.⁸ The former practices are universally valid and can be applied to any task, wherefore they are not suitable for evaluating or comparing tasks. The latter practices are concerned with the concrete design and implementation of a task and thus provide a starting point for task evaluations (Figure 3.1 visualizes this procedure).

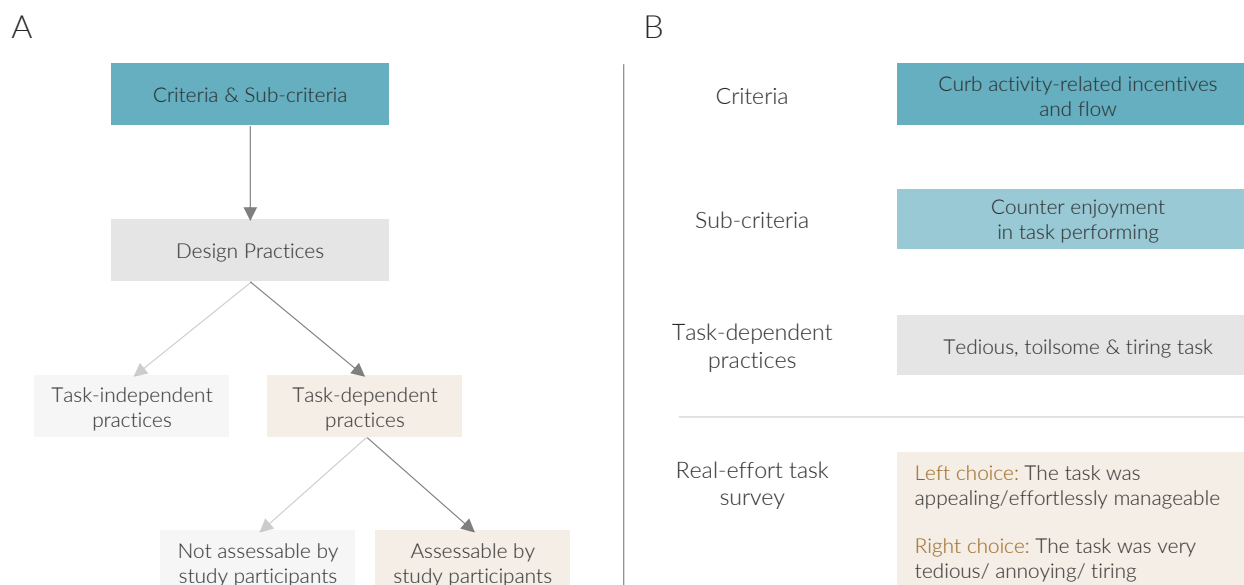


Figure 3.1: **Development of the real-effort task survey:** A) Procedure for developing the survey starting from the [design criteria and practices](#). B) Example illustrating the development of and relationship between a design criterion and a final survey item.

When designing surveys, a compromise must be made between *length* and *measurement accuracy and informativeness*. The length of the survey defines its duration, which in turn conditions its applicability and ease of use. How well the survey items show sufficient discriminatory power and whether they capture the construct(s) of interest determines the informativeness and measurement accuracy of the survey.

In this trade-off, priority is given to brevity in order to counteract possible respondent fatigue when the

⁸See also Figure 2.7 for an illustration of the division into *task-independent* and *task-dependent* design practices.

survey is used in within-subject designs. Thus, for practicality and usability, the survey is condensed to a subset of the aforementioned task-dependent practices, namely those that could be assessed by the study participants. The reason for this is the following: Ultimately, it is the study participants who work on a task and not the researchers; how they truly perceive the task design is probably best judged by themselves. Furthermore, an objective evaluation is only possible to a limited degree, making the detour via subjective individual assessments necessary. Similar approaches can be found in the doctor-patient relationship as well as in drug research (e.g., self-evaluation by patients whether or not they feel better by using certain manual therapies or taking certain drugs). Nevertheless, the responses of the study participants likely provide information about whether the task design has an influence on their behavior, i.e., to what extent the conception of a task may motivate voluntary effort.

The choice of design practices addressed in the real-effort task survey essentially centers around two of the design criteria: *curb activity-related incentives* and *curb purpose related incentives*. The former can be further broken down into three sub-criteria: i) *counter curiosity*, ii) *counter enjoyment in task performing*, and iii) *prevent flow*; those of the latter into iv) *restrain self-evaluation* and iv) *restrain other-evaluation* (see Figure 2.7 for an illustration of the criteria and sub-criteria and the design practices to address them).

The design practices to *prevent flow* are somewhat difficult for study participants to assess. “Disruptions” may be incorporated in various ways in different tasks, making them hard or impossible to compare. Also, disruptions may be implemented so that subjects do not consciously become aware of them. Similar considerations apply for “designing a task to easy to flow.”⁹ For these reasons, the design practices to prevent flow are not explicitly assessed by the real-effort task survey. Researchers interested in additionally capturing components of flow are referred to the *Flow Short Scale (FSS)* by Rheinberg et al. (2003).¹⁰ Therefore, in order to keep the investigation concise and to the point, the survey focuses on the design practices that address the remaining sub-criteria mentioned. The

⁹Designing survey items for the design practice “designing tasks too easy to flow” would further require that subjects must be able to a) assess whether or not they entered the flow state during task completion, b) report which task properties facilitated or hindered their entry into flow.

¹⁰As noted in the [literature discussion in the introduction](#), the FSS provides an excellent measure to examine whether subjects are likely to enter the flow state, which has been used in many different studies and research contexts. However, in contrast to the real-effort task survey, the scale does not assess properties of the activity, but rather asks about indicators of the subject’s functional state. The present study refrains from querying the FSS in addition to eliciting the real-effort task survey to keep the processing time for the subjects short and to avoid further inconveniencing them.

motivational dimensions covered by the survey are illustrated in Figure 3.2.

Criteria	Curb activity-related incentives and flow		Curb undesired purpose-related incentives	
Sub-criteria	Counter curiosity	Counter enjoyment in task performing	Restrain self-evaluation	Restrain other-evaluation
Task-dependent practices	Monotone & unexciting task	Tedious, toilsome & tiring task*	Task without outcome	Annoying task
		Annoying task	Meaningless task	Meaningless task
		No challenge	No challenge	
		Monotone & unexciting task	No performance feedback*	

Figure 3.2: **Countermeasures to non-monetary incentives:** Design practices proposed in Chapter 2 to counter additional activity-related and purpose related incentives. The practices represent the core aspects that the survey aims to cover in motivational terms.

The development of the survey does not follow a classical approach in which, starting from a large pool of items, their number is successively reduced by applying a specific methodology. Instead, an item is constructed for each of the above-outlined motivational dimensions that the survey is to cover. For example, to address the design criterion *curb activity-related incentives*, one of the *design practices* suggests to “make a task tedious, toilsome, and tiring.” To capture this, the item “Was the task appealing/effortlessly manageable or very tedious/annoying/tiring?” is included in the survey. If the subjects eventually perceive a task as tedious, this indicates that the task is actually tedious given the multitude of responses from the large number of subjects (see Figure 3.1.B). It further suggests that the task provides little additional activity-related incentive.

In addition to the motivational items, the survey contains two items to elicit the type (mental or physical) and the amount of effort demanded by a task. All survey items are listed in Figure 3.3, which also depicts their grouping along with the design criteria and practices.¹¹ An illustration of the experimental implementation of the survey with the original wording can be found in Appendix B.2.

¹¹To clearly delineate the constructs that the survey measures, four dimensions can be distinguished: The survey asks about the fulfillment of the design practices intended to prevent activity-related and purpose-related incentives, and about the extent to which the task is physically or mentally demanding. Figure B.51 in the Appendix summarizes this intended structure of the survey.

In the experimental implementation, the eight items of the real-effort task survey are arranged vertically one below the other. Each of the items consists of two opposing statements to express contrary perceptions about a given task.¹² Items are scored on a horizontally arranged response scale with seven levels, which allows for fine-grained measurement. The rating levels are evenly spaced and equipped with integer anchors. Yet, verbal labels are not attached to all rating levels. The endpoints of the scale do not precisely correspond to opposite choices. Nevertheless, the response scale contains a neutral category centered in the middle of the seven levels. Given these features, the scale should be considered as a *discrete visual analogue scale* and not a Likert-type scale, according to J. S. Uebersax (2006).^{13,14} Due to the integer anchors, however, the *intervals between the response levels* are mapped by the raters, i.e., the subjects, onto a “mental number line” and are regarded as actually equal (Harpe, 2015). Since the presentation of the survey as a whole “strongly implies that raters should regard the rating levels as exactly or approximately evenly-spaced,” the raters’ assessments can be treated as interval-level data (S. Uebersax, 2000, p. 1).

If the survey is employed in repeated-measures designs, the order of items should be retained (neither randomized after each task nor the poling reversed). This approach allows subjects to become more familiar with the survey to improve the sincerity and accuracy of their responses (see also the [discussion below](#)).

¹²Various survey formats were considered when designing the survey. In the widespread Likert scale format, subjects have to state their agreement or disagreement with particular statements. Conversely, *semantic differential scales* or *discrete visual analogue scales* offer subjects a one-dimensional range between two extremely opposing adjectives or statements. This enables subjects to express their opinions very precisely. For this very reason, the survey is deliberately designed as such a scale with bipolar verbal anchors. It allows the subjects’ assessments of the tasks along the constructs covered to be determined very accurately.

¹³The terminology of rating scales is not very clear-cut in the literature and may also vary across disciplines. According to Harpe (2015), scales with numbered choices that have verbal anchors only attached to their poles can be referred to as “*numerical rating scales*.” Conversely, J. S. Uebersax (2006) terms scales containing a set of bipolar descriptors as “*discrete visual analogue scales*.” Such scales are used, for instance, in epidemiological and clinical research to measure the intensity or frequency of symptoms. As an example, patients are asked to indicate their pain on a continuum between the two statements “no pain” and “worst pain imaginable.” In the case of a visual analogue scale (VAS) the pain is marked on a line; for a discrete visual analogue scale (DVAS), patients tick one of 11 numbered, evenly spaced boxes. According to Desselle (2005), scales with two polar adjectives or short descriptions as bipolar descriptors are called “*adjectival rating scales*” or “*semantic differential scales*.” The authors note that these scales are mostly used to assess opinions, attitudes, values, and beliefs. The items of the real-effort task survey contain two *bipolar adjectives* or *opposing descriptive statements*, such that *semantic differential scale* or *adjectival rating scale* appears as the most appropriate term. If one wishes to emphasize the discrete nature of the numerical anchoring, *discrete visual analogue scale* can be used instead.

¹⁴This detailed consideration of the scale may seem less necessary at first. However, the reasoning behind it is that depending on the design of a scale, parametric methods may or may not be used for analysis. As will be described later, this is possible for the scale type of the real-effort task survey.

According to participant feedback, the survey is simple and clearly worded, yet precise; it uses appropriate language that is easily understood by the usual target group (students in experimental laboratories).¹⁵ These properties are particularly important for long-lasting within-subject designs where various tasks are compared. They ensure that the subjects do not lose too much attention and that their concentration is not depleted. Adjustments to the survey wording may be appropriate if the survey is conducted with a substantially different subject pool.

Item	Left choice	Right choice	Activity-related incentive	Purpose-related incentive	Effort type
1	was fun	I did not enjoy it	Task enjoyment in general		
2	gave me a target/ performance measurement that spurred me on	did not give me any feedback	No challenge	No performance feedback	
3	aroused my curiosity/ was entertaining	was very uninteresting/ boring	Monotone & unexciting task		
4	was appealing/ effortlessly manageable	was very tedious/ annoying/ tiring	Tedious, toilsome & tiring task	Annoying task	
5	appeared to be meaningful	seemed pointless		Meaningless task	
6	was physically easy	was physically demanding/ exhausting			Physical effort
7	was mentally easy	was mentally demanding/ exhausting			Mental effort
8	produces something/ achieves a goal'	produces nothing/ has no measurable result		Task without outcome	

Figure 3.3: **Survey items and the design criteria and practices on which they are based:** The survey contains six items to evaluate tasks along a range of motivational dimensions. To the right of each survey item, the *design practices* captured by the item are listed. These practices aim to address the *design criteria* to *curb activity-related and purpose related incentives*, as noted above. Furthermore, the survey elicits the amount and type of effort required by a task (*physical, mental*).

Recommended survey implementation. Putting the real-effort task survey into practice, subjects first perform the task of interest to earn money, and then fill out the survey. If the goal is to compare multiple tasks, the survey can be used in both a between- and a within-subject design. In either implementation, the number of required participants scales with the number of compared tasks.

Within-subject designs have the advantage of taking individual-specific effects into account. In this

¹⁵An initial draft of the survey was pretested in a pilot study with 46 participants (four sessions) in April 2018. As in the actual study, the subjects were recruited from the students of the University of Hamburg. The participants of the pre-test expressed that it uses clear language, that the items are formulated precisely and are easy to understand.

case, the subjects complete all tasks, and after each individual task, they fill out the identical survey. This allows to evaluate the tasks item by item, that is, along each survey dimension. To accommodate for order effects, the sequence of tasks must be randomized. However, it is recommended to keep the sequence of survey items the same for all tasks under study. This will ensure that subjects do not misread a survey item and inadvertently provide an unintended response. Key elements of the proposed survey implementation are explained in more detail below.

- **Suggested survey procedure:** For the subjects to be able to evaluate a task correctly, they must be able to gain sufficient experience with it. Therefore, the following sequence is suggested for experiments: First, subjects should have enough time to read the task instructions thoroughly; in a subsequent trial round (~ 30 seconds), subjects can then familiarize themselves with the task; to clear up any remaining ambiguities, they should then have the opportunity to reread the instructions. A duration of at least five minutes is recommended for the following phase of effort provision (see also Section 2.4). This gives the study participants sufficient time to become fully acquainted with the task in order to be able to assess it conscientiously.
- **Order of items in within-subject designs:** Randomizing the sequence of survey items, reversing their poling, or inverting their wording are common strategies to account for order effects. In within-subject designs comparing several tasks, subjects complete all tasks and encounter the survey after each of them. Any of the mentioned changes could confuse subjects or might even remain unobserved and increase the noise in the data (also, “disagreement with a negatively worded item is not necessarily the same as agreement with a positively worded item and vice versa,” see [Desselle, 2005, p. 8](#)). For these reasons, it is recommended to retain the order *and* poling of survey items in repeated-measures designs (see also [Rosenberg et al., 2018](#), who advocate a consistent arrangement of items to make the task assessments less mentally stressful for the subjects). This approach also allows subjects to become more familiar with the survey to improve the sincerity and accuracy of their responses and, as a result, reduces white noise.¹⁶

¹⁶It has been mentioned in the literature that subjects have a tendency to prefer the left side of a response scale, i.e., the categories that are listed “first” ([Chan, 1991](#); [Friedman et al., 1994](#)). This left-side bias would lead to a general shift of responses to the left for all survey items and for all tasks. Since this shift is task-independent, the overall results and the task comparison remain unchanged, but only if the *items’ poling* is not randomized.

In the current setup, all survey items are coded so that higher response values are beneficial from the researcher's perspective with respect to [application areas one and two](#). Related to the example mentioned at the beginning, this means that subjects perceive a task as more tedious, which, due to the large number of subjects, suggests that the task is indeed tedious and indicates that the task provides little additional activity-related incentive. The same applies to the amount and type of effort required.

- **Task sequence in within-subject designs:** In the case of repeated-measures designs, it is advisable to limit the number of tasks to be differentiated to no more than seven. Otherwise, the experiment will be extremely repetitive, lengthy, and tiresome – especially if the set of tasks is without much variety.¹⁷ Furthermore, it is recommended to randomize the order of tasks to control for order effects and increased boredom and fatigue. For these reasons, the experiment conducted also includes a break page after the completion of each task. The page encourages the study participants to briefly rest and stretch themselves to recover (see Figure [B.23](#)).
- **Subject pool:** A sampling error may occur if the participants of the task comparison do not belong to the same population as the subjects of a subsequent experiment with the finally selected task (consider, for example, that a task comparison is conducted with adults, while the actual study is later carried out with children). Therefore, essential characteristics of subject groups should not vary; otherwise, the extent to which activity-related and purpose-related incentives motivate voluntary effort is likely to differ. Similarly, the degree to which a task is demanding can vary widely for different groups of people.

Outlook on survey analysis, survey validity and reliability, and robustness of results. The survey is developed as a compilation of individual items, which are examined separately. However, (parts of) the survey can also be considered as an aggregated scale containing multiple items. This applies in particular to the items for assessing the design practices to curb activity-related and purpose related

¹⁷As a reference, the experiment conducted with seven tasks and a period of effort provision of five minutes per task lasted about $2\frac{1}{4}$ hours. About an hour of this was devoted to filling out a larger number of questionnaires to determine subject characteristics (see Chapter 4).

incentives. A task should be physically or mentally demanding, but not necessarily both. Combining both effort items into an “overall effort” dimension is, therefore, not advisable. Similarly, a certain amount of information is lost when the effort items and the motivation items are aggregated into an overall measure. However, simplicity and practicability may justify certain exceptions.

The validity and reliability of a survey can best be assessed on the basis of actual evidence. Section 3.4 discusses the results of a first application of the survey and addresses both of these aspects in detail. An exploratory factor analysis largely confirms the [intended structure of the survey](#) (see Section [B.3.3.1](#)). The factor analysis further reveals that the number of motivational items could be reduced by one. However, the presented experimental results are robust to these changes. If a shorter version of the survey is desired, the number of items may, therefore, be reduced by this particular item. Yet, scale purification by eliminating items may influence construct validity, such that the survey does not measure any longer what it sets out to measure. Since the result of the factor analysis in this regard are also not particularly striking, the analysis continues with the survey as it was initially designed. Robustness checks are provided in Appendix [B.3.2.5](#).

A frequently raised concern in survey research is that some subjects are inclined to make socially desirable responses ([Desselle, 2005, p. 9](#)). This is less of an issue in the proposed experimental setup for several reasons. *First*, the survey does not evaluate any construct that would provide an incentive for the subjects to respond in a way that would make them “look better” in the eyes of the researcher. *Secondly*, the survey instructions and the survey are impartial about the tasks and do not imply any norm in how to rate them.¹⁸ *Third*, the survey items are the same for all tasks, and they are also presented in identical order. Taken together, this means that subjects are hardly inclined to answer the survey in a way they might think the experimenter would want them to. In other words, there seems to be little temptation for the subjects to attempt to fathom the purpose of the study.¹⁹ To

¹⁸The choice of integers as anchors for the levels is, to some degree, arbitrary. One might argue that the subjects could interpret lower values as “better” – or just the opposite. It is true that during an experiment, subjects tend to contemplate the purpose of the study. However, when subjects are tired and exhausted from a five-minute repetitive task, there seems little likelihood that they will give much thought to whether it might be in the researcher’s best interest to give a particular task a higher or lower rating. A more truthful assessment of their experience with the task seems more probable. Researchers who wish to disguise the purpose of the study further may instead use letters of the alphabet as anchors. However, this is accompanied by a loss of the advantages of the “mental number line,” as discussed above (see also [Harpe, 2015](#)).

¹⁹The participants of the pilot study also reported that the survey did not contain any influential wording or provoked socially desirable responses.

nevertheless conceal its aim, the items of the survey are framed as neutral and objective as possible. For this same reason, the numerical labels of the rating levels range from “1 to 7,” instead of from “-3 to +3” (see also [Schwarz et al. \(1991\)](#) for the impact of the choice of integer anchors on subject behavior).

Survey applications. The real-effort task survey can be used by experimental researchers when they want to i) provide evidence that a task evokes activity- and purpose-related incentives or requires a particular type of effort, ii) control for such task properties in the analysis of a study, iii) compare a selection of tasks and identify a task favorable for research with the help of the multiple comparisons method of [Hsu \(1996\)](#). Due to its general nature, the survey can be utilized to compare both [realistic and unrealistic](#), as well as [useful and useless](#) tasks.²⁰ Moreover, it is applicable to computerized just as much as to manual tasks. Also, in field settings without technical infrastructure, it can be implemented in the form of a pen-and-paper version. Finally, the survey can also be helpful in comparing and validating experimental results across different tasks.

Code and instructions for the real-effort task survey for the experimental laboratory software [oTree](#) are available to interested researchers upon request.

²⁰For a discussion of different types of effort measurement, see Section 1.

3.3 Research Design and Methodology

3.3.1 Experimental Design and Procedure

This section describes a first application of the *real-effort task survey* to evaluate and compare tasks in a laboratory experiment. In a non-competitive work environment, study participants provide effort in a diverse selection of tasks to earn money. The tasks differ in the amount and type of effort they require. After each task, the subjects fill in the survey to reveal their subjective perception of the task. In the following, details on this choice of tasks are provided in Section 3.3.1.1. Thereafter, Section 3.3.1.2 presents a brief portrait of the experimental procedure.²¹ Finally, the empirical strategy is described in Section 3.3.2.

3.3.1.1 Real-Effort Task Selection

Over the years, a wide range of tasks has been proposed by experimental researchers (see Section 1). To reflect this diversity, this study aims to compare as diverse a selection of tasks as possible. Therefore, a variety of tasks used in the literature were reviewed and, based on the task classification by ability and personality traits introduced in Section 1.3, a set of tasks with high degree of heterogeneity was identified. While other combinations are conceivable, this compilation was deliberately chosen for several reasons:

1. They show substantial variety, i.e., they differ from each other in terms of
 - a. their implementation,
 - b. the personality and abilities they require of the study participants;
2. According to my assumptions, heterogeneity in effort provision conditional on task type and subject characteristics can be well identified.

The latter point represents an essential aspect of Chapter 4, which aims at proving that people provide

²¹Appendix B.1.1 provides a detailed account of the experimental design and procedure.

effort based on their character traits and capabilities. With these points in mind, the following set of real-effort tasks is chosen:

- *Solving multiplication problems* (Dohmen & Falk, 2011);
- *Transcription task* (Kephart, 2017);
- *Word-encryption task* (Benndorf et al., 2014; Erkal et al., 2011; Nikiforakis et al., 2012);
- *Pressing-keys task* (Berger & Pope, 2011; DellaVigna & Pope, 2016; Swenson, 1988);
- *Single-slider task* (Waloszek, mimeo 2019, see Section 2.5);
- *Ball-catching task* (Gächter et al., 2016).

Transcription tasks can be constructed in several ways and ask subjects to either write *actual words* or *randomly generated letter and number combinations*. In the latter case, one can distinguish between “memory tasks,” if the number of letters or digits is less than seven, and “transcription tasks,” if it is greater or equal to seven. The transcription task of Kephart (2017) is used in the experiment in two ways: as a *memory task*, where subjects had to transcribe random letter and number combinations with four digits; as a *transcription task*, wherein study participants had to copy German foreign words, which are rarely used in the daily language and had on average nine and a half letters.²²

The final selection of tasks studied in the experiment can be assigned to the task categories presented in Table 1.1 according to their demands on skills and personality traits. This category assignment is summarized in Table 3.1, together with a brief description for each of the tasks.²³ Each of the tasks is described in more detail in the instructions provided to the study participants (see Appendix B.1.4.4).

²²The foreign words employed in the task were obtained from the following websites: *99 Wörter aus der gehobenen Sprache für spannendere Blogtexte* on 15.04.2018, and from *120+ bildungssprachliche Adjektive* on 15.04.2018.

²³The indicated number of citations was recorded from Google Scholar in February 2019. Apart from the typing task, this number refers to the publication that originally introduced the task. The typing task is first introduced by Swenson (1988) and recently taken up again by Berger & Pope (2011), for which the number of citations is given. In general, the specified number of citations can have different reasons. It provides only limited information about the use of the task introduced in the publication in further works. All tasks were introduced in the current version in the last ten years. In addition, some tasks have been modified in the meantime so that later improved versions may have received greater attention and references in the literature.

Table 3.1: Real-effort task selection employed in the experiment

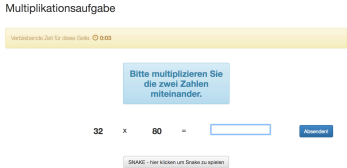


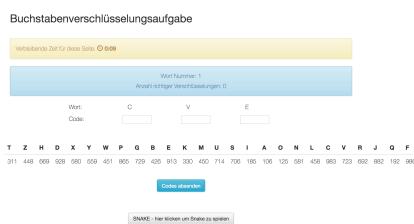
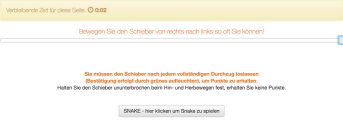
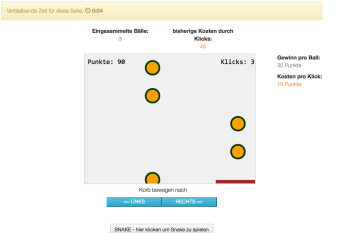
Illustration	Task	Reference	Citation	Description
Quant. and Analyt. Reasoning				
	Multiplication task	Dohmen & Falk (2011)	656	Multiplying two 2-digit numbers, demanding mathematical skills and concentration
Language and Verbalizing				
	Word-transcription task	Waloszek modified from Kephart (2017) (unpublished)	-	Transcribing long, German foreign words that are rarely used in everyday language. The task requires language and writing skills
Memory and Knowledge				
	Code-transcription task	Kephart (2017)	-	Transcribing random letter and number combinations with four digits to challenge short-term memory only
	Word-encryption task	Erkal et al. (2011)	113	Encoding a given "word" (three-letter combination) with an encryption table containing three 3-digit numbers for each letter of the alphabet

Table 3.1: Real-effort task selection employed in the experiment (continued)

Illustration	Task	Reference	Citation	Description
Mechanical				
<p>*A oder B* Tastendruckaufgabe</p> 	ab-Typing task	Swenson (1988)	116	Pressing two keys ("a" and "b") alternately on the computer keyboard
<p>*links-rechts* Schieberaufgabe</p> 	Single-slider task	Waloszek (mimeo 2020)	-	Moving a slider from one end of a line to the other [similar to the "dragging a ball" task of Heyman & Ariely (2004)]
Playful and Entertaining				
<p>Balleinfangaufgabe</p> 	Ball-catching task	Gächter et al. (2016)	20	Moving a tray to catch falling balls. The task combines the tangible action of catching balls with induced material effort costs

3.3.1.2 Experimental Procedure

The computerized experiment was programmed in *Python* using the laboratory software *oTree* (Chen et al., 2016).²⁴ It was conducted at a large public university in Germany on four consecutive days within one week of June 2018.²⁵ All internal review board procedures were followed.²⁶ Two hundred forty-eight study participants were recruited using the experiment management system *hroot* (Bock et al., 2014). The sample consisted mostly of university students (57% were females; 80.2% were Germans, 4.4% were Europeans, and 15.3% were non-Europeans; subjects average age was 26 years).²⁷ They were majoring in a variety of fields (less than 30% studied law or economics).

The subjects were divided into twelve experimental sessions (13-28 subjects per session), with none of them participating in more than one session. The study followed a within-subject design, such that all participants completed the same experiment content. The experiment included five steps, out of which two were administrative (colored in gray in the experimental procedure depicted in Figure 3.4 below). The actual laboratory study consisted of three steps:

1. the filling in of (standardized) psychological questionnaires (subsequently termed characterization surveys);²⁸
2. an incentivized task combined with a brief ex-post survey (the *real-effort task survey*), and
3. the completion of two concluding questionnaires.

Step number two was repeated seven times, i.e., once for each task. The content of the experiment

²⁴Although performance losses are relatively unlikely to occur in modern experimental laboratory facilities, they cannot be ruled out. For example, the freezing of computer screens has been reported when a large number of participants perform the same task accessing the same online code fragment. For this reason, the laboratory software *oTree* was deliberately chosen to conduct the experiment, as it is very robust in this respect. Latency times and other problems with the software were not observed during the study. However, the experiment did also not involve any interactions between the subjects.

²⁵The experiment sessions took place at the experimental laboratory of the University of Hamburg: [WiSo-Research Laboratory](#), Faculty of Business, Economics and Social Sciences Universität Hamburg, Von-Melle-Park 9, 20146 Hamburg, Germany. Contact person: Olaf Bock, Phone: +49 40 23 952-3759, E-Mail: olaf.bock@WiSo.uni-hamburg.de.

²⁶The study adheres to the ethical standards of the WiSo-Research Laboratory, and proof of consent from the local authority (data protection officer of the University of Hamburg) has been acquired. Furthermore, approval by the ethics commission of ETH Zurich has been obtained (see Appendix B.4.1).

²⁷See also Table C.2, which summarizes descriptive statistics for several control variables.

²⁸These psychological assessments are part of Chapter 4 and will not be discussed further at this point.

was the same across all sessions. Solely the order of the characterization surveys and the sequence of tasks were randomized across sessions.^{29,30}

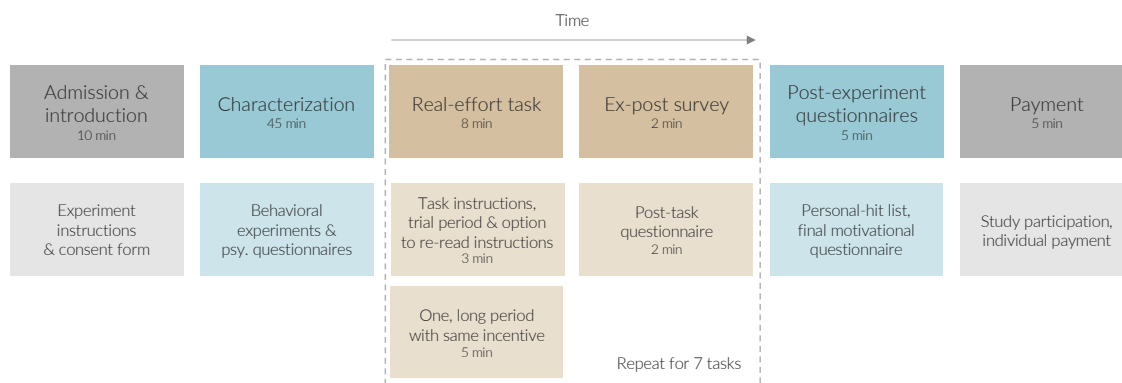


Figure 3.4: **Experimental procedure:** The experiment contains five steps. The *first* and *last steps* are administrative. The *second step* contains a psychological assessment, which forms the basis of Chapter 4. The *third step* represents the focus of this chapter, in which the subjects complete and assess a selection of tasks. After receiving general instructions, subjects fill out control questions. Then subjects perform the tasks, which vary in the type of effort required (physical/mental). After each task, the subjects are asked to fill out the *real-effort task survey* to capture their view of the task. All subjects complete all seven tasks. Solely the order of tasks and characterization surveys was randomized across groups of, on average, ten subjects. In the *fourth step*, subjects complete two concluding questionnaires to record their overall perception of the tasks as well as their motivation for participating in the study.

Piece-rate incentives induced effort provision in the tasks, whereby, conditional on the task design, the stake sizes varied.³¹ After each task, subjects were informed about the number of points they had collected. At the end of the experiment, these were converted to Euro at a conversion rate of 1000 points = 1 €. After completing the final questionnaire on their overall task perception and their motivation to participate in the study, subjects were individually and discreetly paid their earnings.

²⁹More precisely, each session was split into two groups. Each group was assigned a randomly generated session configuration containing a distinct task and survey order. The twelve sessions, therefore, yielded a total of 24 groups with differing session configurations. Notably, the number of subjects *per group* ranged from five to 15. The experiment was conducted in a double-blind procedure so that neither the local experimenters from the WiSo-Research Laboratory nor the study participants knew of any group assignment.

³⁰The experiment employed standard methodology frequently used in behavioral and experimental economics. Participants were not deceived, personal identities were not revealed, and the study did not contain any physical or body-related intervention. Negative consequences on the participants' mental or physical well-being are not expected from participation in the experiment. Before each experiment session, every participant was asked to provide written consent (the respective consent form is enclosed in Appendix B.1.3).

³¹Based on a pilot experiment, piece rates were adjusted to ensure similar average payments for all tasks.

The average duration of an experimental session was slightly over two hours. Total payoffs accordingly accumulated to, on average, 21.80 € for completing all seven tasks, including a show-up fee of 5 €. The experimental procedure is illustrated in Figure 3.4. A detailed account of its elements and their timing is provided in Appendix B.1.1.

3.3.2 Empirical Strategy

3.3.2.1 Hypotheses

As outlined at the outset, this study aims to resolve the following research question:

According to the task assessments of study participants, *do tasks differ?*

The question is addressed in a laboratory experiment with the help of the real-effort task survey, which reveals the subjective perception of the study participants for the examined tasks along *motivational dimensions* and in terms of the demanded *type of effort*. From the general research question, one can derive the following set of hypotheses (one for each item of the survey) regarding subject's perception of the seven tasks:

For a given item $q \in [1, 8]$ of the real-effort task survey:

- H_0 : Subjects perceive the seven tasks in the same way such that there is equality between mean responses to the given survey item q : $\mu_{t_1}^q = \mu_{t_2}^q = \dots = \mu_{t_7}^q$ for $t \in [t_1, t_7]$
- H_A : Subjects perceive the seven tasks as being different such that the mean responses to the given survey item q differ: $\mu_{t_i}^q \neq \mu_{t_j}^q$ for at least two tasks.

A simple omnibus ANOVA can be used to test this. However, it is highly likely that “some task” will be perceived as different from the others, such that one can be sure to find a significant difference among the means. A more refined approach to examine where any differences stem from considers the properties of the tasks. As described in Section 3.3.1.1, the selection of tasks was compiled

to show substantial heterogeneity. Despite this, certain tasks display design similarities and can be grouped together. Based on these groupings, a set of six *contrast hypotheses* to distinguish the tasks was defined a priori. Each of these contrasts examines whether two specific subsets of the task selection differ along a particular survey dimension. This is done by assessing whether the means of the outcome variable are significantly different for both subsets, i.e., by comparing the average population means of subjects' task perception for both groups of tasks. For example, contrast C_1 examines whether there is a significant difference between *cognitive and memory tasks* and *mechanical and playful and entertaining tasks*. Similarly, contrast C_5 tests whether *mechanical tasks* (represented by the ab-typing task and the single-slider task) differ significantly from *playful and entertaining tasks*, i.e., the ball-catching task. The planned orthogonal contrasts allow for a fine-grained, step-by-step differentiation of the tasks in the selection. The complete set of contrast hypotheses is as follows:

- **Contrast C_1 :** *Cognitive tasks & memory tasks vs. mechanical tasks & playful and entertaining tasks*

$$H_{01} : \kappa_1 = -\frac{1}{4}(\mu_{math} + \mu_{words} + \mu_{codes} + \mu_{encrypt}) + \frac{1}{3}(\mu_{ab} + \mu_{slider} + \mu_{balls}) = 0, H_{A1} : \kappa_1 \neq 0$$

- **Contrast C_2 :** *Cognitive tasks vs. memory tasks*

$$H_{02} : \kappa_2 = -\frac{1}{2}(\mu_{math} + \mu_{words}) + \frac{1}{2}(\mu_{codes} + \mu_{encrypt}) = 0, H_{A2} : \kappa_2 \neq 0$$

- **Contrast C_3 :** *Compare cognitive tasks only (multiplication task vs. transcribe-words task)*

$$H_{03} : \kappa_3 = -\mu_{math} + \mu_{words} = 0, H_{A3} : \kappa_3 \neq 0$$

- **Contrast C_4 :** *Compare memory tasks only (transcribe-codes task vs. word-encryption task)*

$$H_{04} : \kappa_4 = -\mu_{codes} + \mu_{encrypt} = 0, H_{A4} : \kappa_4 \neq 0$$

- **Contrast C_5 :** *Mechanical tasks vs. playful and entertaining tasks*

$$H_{05} : \kappa_5 = -\frac{1}{2}(\mu_{ab} + \mu_{slider}) + \mu_{balls} = 0, H_{A5} : \kappa_5 \neq 0$$

- **Contrast C_6 :** *Compare mechanical tasks only (ab-typing task vs. single-slider task)*

$$H_{06} : \kappa_6 = -\mu_{ab} + \mu_{slider} = 0, H_{A6} : \kappa_6 \neq 0$$

These task comparisons are performed along each dimension of the real-effort task survey, i.e., separately for each survey item. The selection of tasks and their grouping, in conjunction with the there-

upon derived set of planned orthogonal contrasts, are illustrated in Figure 3.5. The associated contrast coding matrix combining the orthogonal contrasts can be found in Appendix B.3.2.1.

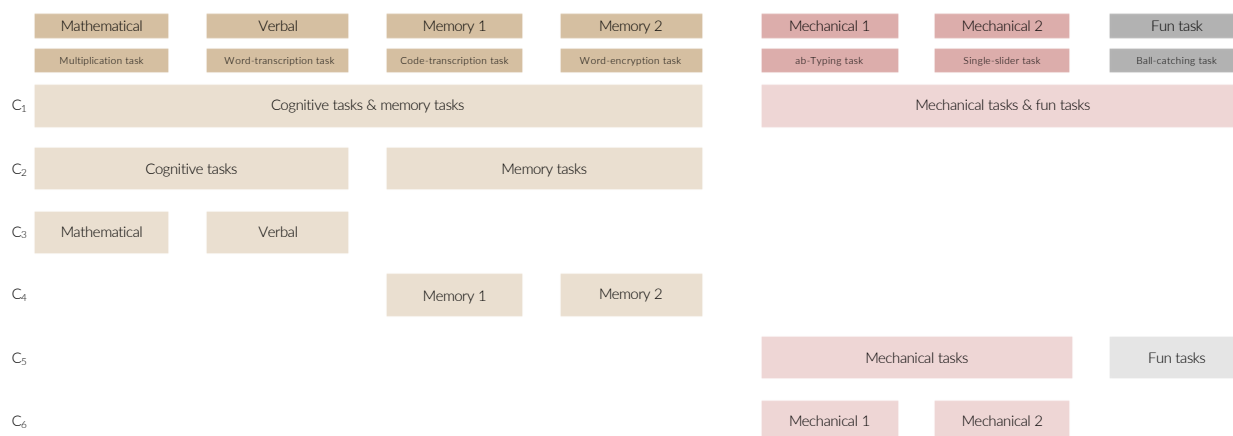


Figure 3.5: **Task selection, task grouping, and planned orthogonal contrasts:** Prior to the analysis, subsets of the task selection were defined based on task similarities. These form the basis for the planned orthogonal contrasts. In each of the planned comparisons, two of the pre-defined subsets of the tasks are compared. For a given survey item, if there is strong evidence that the mean responses differ in the two subsets, then the subjects perceived the respective tasks as different from each other. For example, contrasts 2 tests if *cognitive tasks* differ from *memory tasks*. The regression coefficient provides an estimate of the difference in subjects' perceptions of the two subsets of tasks (in relation to the response scale). The task comparisons are separately performed for all survey dimensions, i.e., item-by-item.

3.3.2.2 A Simple Model

The analysis of the survey items with *planned orthogonal contrasts* allows to differentiate between tasks and to evaluate to which degree subjects' perception of tasks differs. To reframe these main interests, the analysis aims to answer the following questions along each dimension of the real-effort task survey:

- I) Is there a *task effect*? That is, is there an *effect* of task on the subject's task assessment?
- II) How large is the *variability between different subjects* with regards to the *general task liking level*?

To investigate these questions along each survey dimension, the following baseline specification of a linear mixed-effects model is estimated (separately for each survey item q):³²

$$Y_{ts}^q = \mu^q + \alpha_i^q + \delta_s^q + \varepsilon_{ts}^q \quad (3.1)$$

Subjects face each task and the subsequent survey only once. Consequently, one response Y_{ts}^q is recorded per subject s for each survey item q and task t ($s \in [1, 248]$, $t \in [t_1, t_7]$, $q \in [1, 8]$). μ^q serves as a global mean for survey item q . To address the first and main question, the model contains a fixed effect for the task α_i^q with seven levels (one for each task). To further assess the variation between different subjects, *subject* is treated as random to have the fixed effects to be population averages. The random effect of subject δ_s^q , therefore, resembles a “general task liking level of subject s ,” i.e., how much subject s likes real-effort tasks “generally.” The interpretation of the model is as follows (in terms of any particular survey item). There is an “average task perception” (across the whole population) in relation to the seven different tasks given by the fixed effect α_i^q . However, each subject may deviate from this “preference profile.” This random deviation of an individual study participant consists of a general shift δ_s^q (“subject-specific general task perception” or “general task liking of a subject”), with regard to this specific survey dimension.

3.3.2.3 Structure of the Collected Data, Assumptions Made Concerning the Empirical Analysis and the Data Analysis Approach

The subjects’ responses to each of the survey items represent the main study variables. They allow to contrast the task selection and to assess the variability of subjects’ perception of the tasks. Subjects indicated their agreement with each of the eight survey items (statement pairs) on a response scale with seven choices: (1) complete agreement with the statement on the left end of the scale; (7) full agreement with the statement on the right (see also Appendix B.2 for an illustration of the experimental implementation). The measured categorical variable resembles an ordinal approximation of an underlying, continuous variable.

³²The data are approximately normally distributed, and homogeneity of variance is given, see Section 3.3.2.3 below. However, the observations are not independent and demand a linear mixed-effects model rather than a linear regression model as the appropriate approach.

In the conducted study, seven tasks were compared with the help of the survey. Therefore, the nominal, independent variable *real-effort task* contains seven levels. Since all study participants completed the survey for all tasks, the design is *balanced*, i.e., there is an equal number of observations in all levels of the factor “real-effort task.”

For a given subject and task, the responses to the survey items are dependent. This violates the *assumption of independence* and requires replacement by the *assumption of sphericity*. The *Greenhouse-Geisser’s* estimate represents a conservative measure to adjust the degrees of freedom (and hence the *p*-values) accordingly. The empirical analysis further involves multiple comparisons, which requires adjustment of *p*-values. However, the observed *p* values are so unambiguous that adjustments due to the repeated-measures design and due to the multiple-hypothesis testing do not alter the results obtained (see Appendix B.3.2.3 for an exemplary assessment of the changes in *p*-values due to both). For this reason, unadjusted *p*-values are reported in the subsequent analysis.

In the empirical analysis, each item of the survey is examined individually.³³ Each “subject” is treated as a block wherefore *subject* is added as a random variable as previously outlined. The order of tasks (including the subsequent real-effort task survey for the respective task) was randomized and is thus controlled for (24 groups, each with a different sequence of tasks). Hence, it is not included in the analysis as a control. Whether the randomization procedure worked sufficiently well is addressed in the Appendix.³⁴

The analysis assumes that the rating levels of a particular survey item are evenly-spaced and supposes that the scaling is comparable both across items and across tasks. The survey representation was designed to allow for these assumptions (see also the discussion in Section 3.2). To emphasize the key aspects: *First*, the items (statement pairs) of the survey are laid out in a vertical grid. *Second*, the numbered radio buttons of each response scale are arranged horizontally evenly between the poles. The resulting grid structure assures that distances between radio buttons appear comparable for each item and across items. These presuppositions are crucial to treat the ratings as interval data

³³When the survey is considered as a *summated rating scale*, i.e., several items of the survey are aggregated into a single score, this is stated explicitly.

³⁴In Appendix B.3.2.5, descriptives are provided to examine whether the assumption that *randomization of tasks across groups is sufficient to mitigate order effects* may hold. Moreover, an extension to the simple model (3.1) is presented in Appendix B.3.2.5. In this more complex model, the *position of the task sequence* is added as a fixed effect to control for order effects.

(S. Uebersax, 2000), and to be able to aggregate survey responses (across items or tasks).

There is a discussion in the literature about whether the data of numerical rating scales can be analyzed with parametric methods. If the response scale contains at least five levels and their integer anchors are equally spaced, the obtained data can be treated as interval-level data (S. Uebersax (2000), Harpe (2015), Kerlinger & Lee (2000); the response scale of the real-effort task survey contains seven levels, and the full range of rating options was utilized by the study participants). Consequently, parametric methods can be applied (Desselle, 2005; Harpe, 2015).³⁵ Because effect sizes must be directly interpretable, I follow this approach and proceed with parametric tests.³⁶

The linear mixed-effect model is fit for each survey item by restricted maximum likelihood (REML) using the analysis software *R*.³⁷ The model assumptions were visually examined using Tukey-Anscombe plots and q-q plots. According to the q-q plots, the observed data stem from distributions that are rather short-tailed (both for the residuals and the random effects). Therefore, one can safely assume normality in the subsequent analysis, and any obtained estimates are considered as rather conservative.³⁸ The residuals were examined with Tukey-Anscombe plots (both for the fitted values as well as the explanatory variable), and i) fluctuate randomly around zero and ii) are independent. The expectation of the error is thus zero, which indicates that there is no systematic error. Further, the variance of the error term does not increase with the response variable (task rating in the survey), giving support to the assumption of a linear model. The next section summarizes the results of the empirical analysis.

³⁵For the related, single Likert items, which have verbal anchors attached to each response level, it is usually agreed that if an item contains more than five points (*five or more categories rule*), parametric tests can be used (see Johnson & Creech, 1983; Norman, 2010; Sullivan & Jr, 2013; Zumbo & Zimmerman, 1993).

³⁶*Ordinal Regression* yielded similar results in terms of *p*-values.

³⁷The regression analyses were performed in *R* using the *lmer*-function from the *lme4*-package (Bates et al., 2015; Kuznetsova et al., 2017).

³⁸The sampling distribution is also approximately normal due to the large number of subjects ($N = 248 \gg 30$).

3.4 Experimental Results

In its first application, the real-effort task survey was used to compare a selection of seven distinct tasks. The main result of the laboratory study is that the subjects perceive the tasks as very differently designed. This applies to both the *motivational dimensions*, i.e., the design practices to counter activity-related and purpose related incentives, and the *effort dimensions* (physical and mental). In addition, subjects' perceptions of a given task exhibited a high degree of heterogeneity. In the next sections, these results are presented more elaborately. They are then considered together to gain insights into how physical or mental demands might translate into activity-related and purpose related incentives. Furthermore, a multiple comparison procedure is employed to determine a “favorable task” regarding [application areas one and two](#) from the researcher's point of view. Finally, supporting evidence on the validity of the survey is provided, including the results of an additional survey termed “personal-hit list” (Rheinberg, 1989). Robustness checks are provided in [Appendix B.3.2.5](#).

Unless otherwise indicated, in all figures and tables, survey items are coded so that higher scores represent greater agreement with the design practices against non-monetary incentives or a greater demand for effort.³⁹

3.4.1 Compare Tasks Along Motivational Dimensions

Table [3.2](#) summarizes the mean values for the motivational items of the real-effort task survey.⁴⁰ The font color of the stated values for each survey item varies between [brown](#) and [petrol blue](#) according to their agreement with one of the two-item statements. Substantial differences are observed in terms of subjects' perception of the tasks for all motivational survey items. The figure highlights the variety in the design of tasks used in experimental research. The large differences perceived by the students in the design of the tasks with regard to motivation-relevant design aspects give a first impression of the extent to which different motivational influences are present in the tasks.

³⁹As per the [exposition below](#), tasks with higher scores may be “favorable” in terms of task application areas one and two from the perspective of the researcher.

⁴⁰Further illustrations of the survey responses are provided in [Appendix B.3.1.2](#). [Figure B.53](#) summarizes the responses to the motivational dimensions of the real-effort task survey for all seven tasks. The detailed distributions of the two-way data of the entire survey are depicted in [Figure B.54](#).

Table 3.2: **Descriptives for the motivational items of the real-effort task survey:** A survey item consists of two opposing statements marking the endpoints of a response scale with seven levels. The **left statement** corresponds to a positive assessment of a task (agreement to this is coded with 1), the **right statement** is a negative assessment of the task (agreement is coded) with 7. The table summarizes the mean responses with standard deviations. The font color of the reported values reflects the degree of agreement with the statements. Each item of the questionnaire is based on a specific design practice. The *higher the mean*, the more likely it is that the design practice underlying the item is met, and thus any non-monetary incentives are mitigated. Conversely, the *lower the mean value*, the more probable voluntary effort provision is.

	Survey choices		Cognitive tasks		Memory tasks		Mechanical tasks		Fun tasks
	Left (1)	Right (7)	Multipli- cation	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
Q1	... was fun	... I did not enjoy it	3.96 (2.23)	2.54 (1.62)	3.71 (1.84)	3.49 (1.87)	4.3 (2.04)	5.53 (1.72)	3.27 (1.67)
Q2	... gave me a target/ performance measurement that spurred me on	... did not give me any feedback	3.37 (2.03)	2.94 (1.83)	3.56 (1.91)	3.12 (1.64)	3.09 (1.89)	4.47 (1.99)	3.37 (1.69)
Q3	... aroused my curiosity/ was entertaining	... was very uninteresting/ boring	3.73 (1.91)	2.95 (1.61)	3.92 (1.83)	3.68 (1.86)	4.5 (1.9)	5.56 (1.68)	3.56 (1.71)
Q4	... was appealing/ effortlessly manageable	... was very tedious/ annoying/ tiring	4.49 (1.8)	2.88 (1.62)	3.63 (1.77)	3.71 (1.76)	4.46 (2.01)	5.08 (1.94)	2.82 (1.63)
Q5	... appeared to be meaningful	... seemed pointless	3.56 (1.99)	3.66 (1.91)	4.4 (1.89)	4.42 (1.83)	5.71 (1.47)	5.92 (1.51)	4.48 (1.81)
Q8	... produces something/ achieves a goal	... produces nothing/ has no measurable result	3.35 (1.85)	3.5 (1.89)	4.06 (1.89)	3.95 (1.84)	4.46 (1.88)	4.94 (1.96)	3.94 (1.85)

A simple omnibus test in one-way analysis of variance (ANOVA) yielded a significant difference in means for all items of the survey (results are summarized in Table B.9 in the Appendix). Therefore, it was possible to proceed with the [planned comparisons](#) defined earlier and analyze the survey item-by-item.⁴¹

Table 3.3 summarizes the results of the regression analysis. The [linear mixed-effects model](#) is fit by restricted maximum likelihood (REML) estimates of the parameters. To illustrate the interplay between the regression model and the planned comparisons and in their application to each survey item, the estimates and confidence intervals are depicted in Figure 3.6 (for two selected survey items). For most of the motivational dimensions, strong evidence is found against the (respective) null hypothesis that the particular subgroups of tasks, as defined by the planned contrasts, do not differ. Put differently, for a given survey item q of the real-effort task survey, the selection of tasks can well be distinguished by orthogonal contrasts. Considering the regression estimates, tasks differ substantially along the motivational dimensions.

⁴¹The employed orthogonal contrasts are described in closer detail in Section 3.3.2.1 (see also Figure 3.5). The contrast coding matrix is provided in Appendix B.3.2.1.

Table 3.3: **Regression estimates for all motivational survey items (simple model):** The table contains the results of individual regressions for the six motivational items of the survey. Each item (column) assesses a different design practice to address activity-related and purpose related incentives. For each regression, each estimate (row) resembles a different contrast, which compares two subgroups of the task selection (see Figure 3.5). The regression estimates are visualized in Figure 3.6, which also contains the statement pairs for two survey items. The coefficients are given in relation to the intercept with the same scaling of the response scale. The model includes a fixed effect for the *task* and a random effect for *subject*. A more complex model controlling for the [position of the task in the task sequence and the group](#) is provided in Appendix B.17. As previously outlined, adjustments with the *Greenhouse-Geisser's* coefficient to account for the repeated-measures design yield no major changes in terms of *p*-values. Similarly, this holds for multiple comparisons. Consequently, the stated *p*-values are neither adjusted for sphericity nor for multiple-hypothesis testing ($p \leq 0.001$ in bold).

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	No outcome
(Intercept)	3.828*** (0.000)	3.417*** (0.000)	3.986*** (0.000)	3.869*** (0.000)	4.590*** (0.000)	4.027*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.609*** (0.000)	0.680*** (0.000)	1.664*** (0.000)	0.765*** (0.000)	2.331*** (0.000)	1.257*** (0.000)
C2: Cognitive → Memory	0.351*** (0.000)	0.192* (0.041)	0.458*** (0.000)	-0.020 (0.832)	0.800*** (0.000)	0.583*** (0.000)
C3: Cognitive: Math → Words	-0.706*** (0.000)	-0.216** (0.001)	-0.393*** (0.000)	-0.804*** (0.000)	0.050 (0.409)	0.075 (0.211)
C4: Memory: Codes → Encryption	-0.109 (0.111)	-0.222*** (0.001)	-0.117 (0.062)	0.038 (0.569)	0.008 (0.895)	-0.056 (0.344)
C5: Mechanical → Fun	-1.098*** (0.000)	-0.277*** (0.000)	-0.980*** (0.000)	-1.301*** (0.000)	-0.891*** (0.000)	-0.507*** (0.000)
C6: Mechanical: ab-Typing → Single Slider	0.615*** (0.000)	0.690*** (0.000)	0.526*** (0.000)	0.310*** (0.000)	0.107 (0.080)	0.236*** (0.000)
N (subjects)	248	248	248	248	248	248

Note: *p*-values are not adjusted for multiple hypothesis testing.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

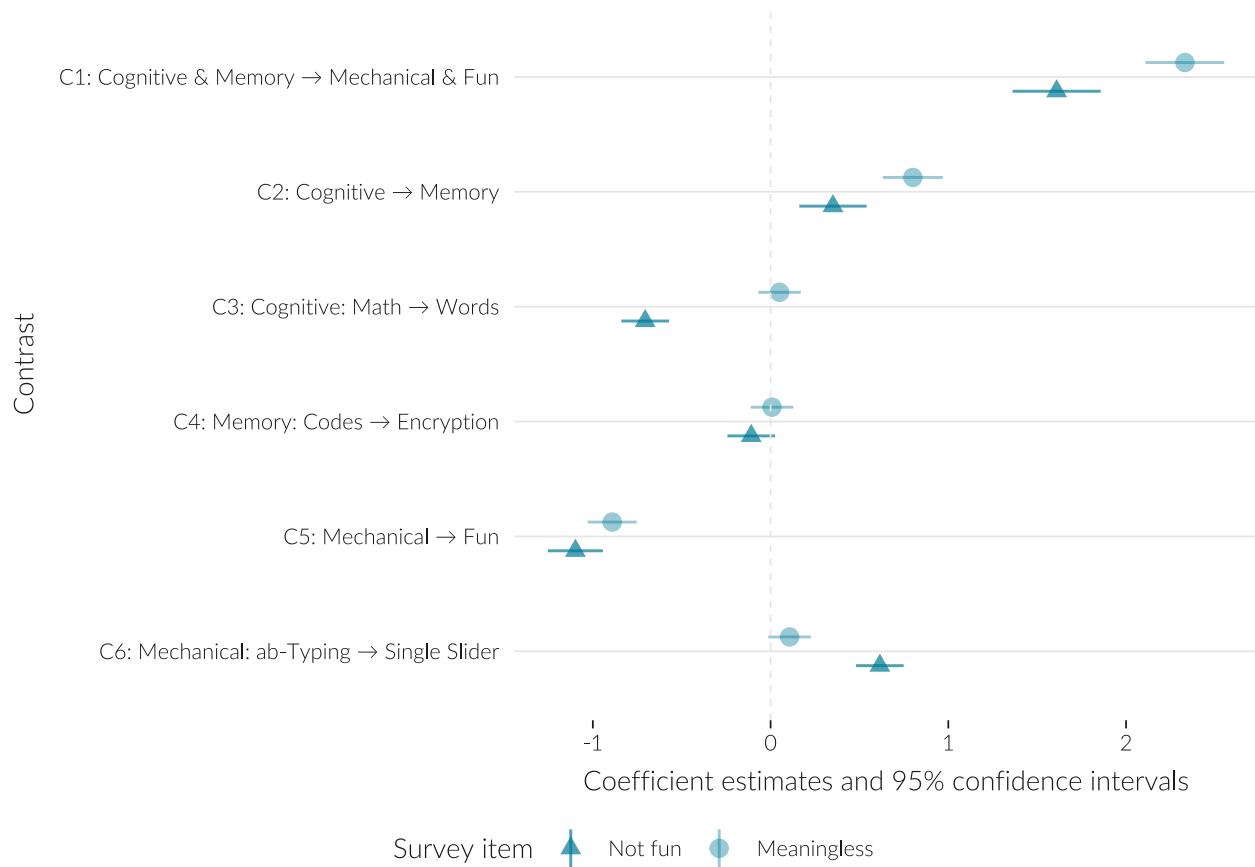


Figure 3.6: **Plot of regression coefficients:** Estimates and respective 95% confidence intervals for all orthogonal contrasts and selected motivational survey items. The first item reveals that subjects perceived the *mechanical tasks* as much less enjoyable than the *playful and entertaining task* (mean ratings differed significantly across both task subsets -1.11 , $p < 0.001$). *Cognitive task and memory tasks* appear to subjects substantially more meaningful than *mechanical and playful and entertaining tasks* (2.34 , $p < 0.001$). And comparing the cognitive tasks, the *multiplication task* was less enjoyable, aroused less curiosity, and was more tedious and tiring than the *word-transcription task* (-0.72 , -0.40 , -0.80 ; all $p < 0.001$).

3.4.2 Compare Tasks Regarding the Demanded Effort

Table 3.4 summarizes the mean values for the effort-type items of the real-effort task survey. The tasks differ strongly in their demand for physical and mental effort and can broadly be grouped accordingly: (1) **mentally very demanding tasks** (*multiplication task*), (2) **moderately mentally demanding tasks** (*word-transcription* and *code-transcription task*, *word-encryption task*), (3) **physically very demanding tasks** (*ab-typing task* and *single-slider task*), and (4) **neither mentally nor physically demanding tasks** (*ball-catching*

Table 3.4: **Descriptives for the effort-type items of the real-effort task survey:** Subjects evaluate the tasks on a response scale with seven levels, whereby higher values correspond to a greater demand for effort. The table depicts mean response levels with standard deviations. For the detailed distributions of the two-way data, see Appendix B.3.1.

	Survey choices		Group 1		Group 2		Group 3		Group 4
	Left (1)	Right (7)	Multipli- cation	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
Q6	... was physically easy	... was physically demanding/ exhausting	2.5 (1.84)	2.47 (1.62)	2.73 (1.69)	2.6 (1.63)	4.72 (1.87)	4.25 (2.05)	1.9 (1.31)
Q7	... was mentally easy	... was mentally demanding/ exhausting	4.94 (1.76)	2.82 (1.56)	3.32 (1.81)	3.44 (1.8)	2.43 (1.85)	2.79 (2.03)	2.21 (1.37)

task). This classification closely matches the task similarity grouping that forms the basis for the contrast coding scheme (see Figure 3.5).

A one-way repeated measures ANOVA reveals significant differences in the mean survey response across the seven real-effort tasks for both effort related survey items (physical effort and mental effort).⁴² Again, the [simple mixed-effect model](#) is fit to the data. Table 3.5 shows the estimates for the two regressions of the subjects' responses to the two effort-related items on the [orthogonal contrasts](#). Figure 3.7 visualizes the following results. The first contrast C_1 compares *cognitive and memory tasks* to *mechanical and playful and entertaining tasks*. Compared to the former, the latter tasks are more physically demanding (1.81, $p < 0.001$), but less mentally demanding (-1.97 , $p < 0.001$). C_2 : Moving from the *cognitive tasks* (multiplication and word-transcription task) to the *memory tasks*, the demand in mental effort decreases (-0.51 , $p < 0.001$). C_3 : Among the cognitively demanding tasks, the *multiplication task* demands more mental effort than the *word-transcription task* (-1.06 , $p < 0.001$). C_4 : The memory tasks do not differ in their effort demands. C_6 Comparing only the mechanical tasks, the *ab-typical task* requires slightly more physical effort (-0.24 , $p < 0.001$), while the *single-slider task* is mildly more mentally demanding (0.17, $p < 0.05$). Both tasks demand much more effort than the *ball-catching task* (C_5 : physical effort -1.72 , $p < 0.001$; mental effort -0.26 , $p < 0.001$).

⁴²See Table B.10 in the Appendix for the results of the ANOVA.

Table 3.5: **Regression estimates for all effort-related survey items (simple model):** Subjects' responses to the two effort-related survey items are assessed by a linear mixed-effects model. The model contains a fixed effect for the *task* and a random effect for the *subject*, and is fit by restricted maximum likelihood (a more complex version of the model that controls for the [position of the task in the task sequence and the group](#) can be found in Appendix B.18). The table shows the parameter estimates for separate regressions for the effort items on the [planned contrasts](#). Each contrast compares two [subsets of the task selection](#). In each regression, the contrasts are applied separately to the outcome variable (subjects' responses to a given survey item). The estimates are visualized for all contrasts in Figure 3.7, jointly with the wording of the survey items.

	Physically demanding	Mentally demanding
(Intercept)	3.025*** (0.000)	3.134*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.799*** (0.000)	-1.979*** (0.000)
C2: Cognitive → Memory	0.183* (0.050)	-0.500*** (0.000)
C3: Cognitive: Math → Words	-0.016 (0.807)	-1.060*** (0.000)
C4: Memory: Codes → Encryption	-0.067 (0.314)	0.060 (0.362)
C5: Mechanical → Fun	-1.726*** (0.000)	-0.265*** (0.001)
C6: Mechanical: ab-Typing → Single Slider	-0.234*** (0.000)	0.179** (0.007)
N (subjects)	248	248

Note: p-values are not adjusted for multiple hypothesis testing.

* p < 0.05, ** p < 0.01, *** p < 0.001

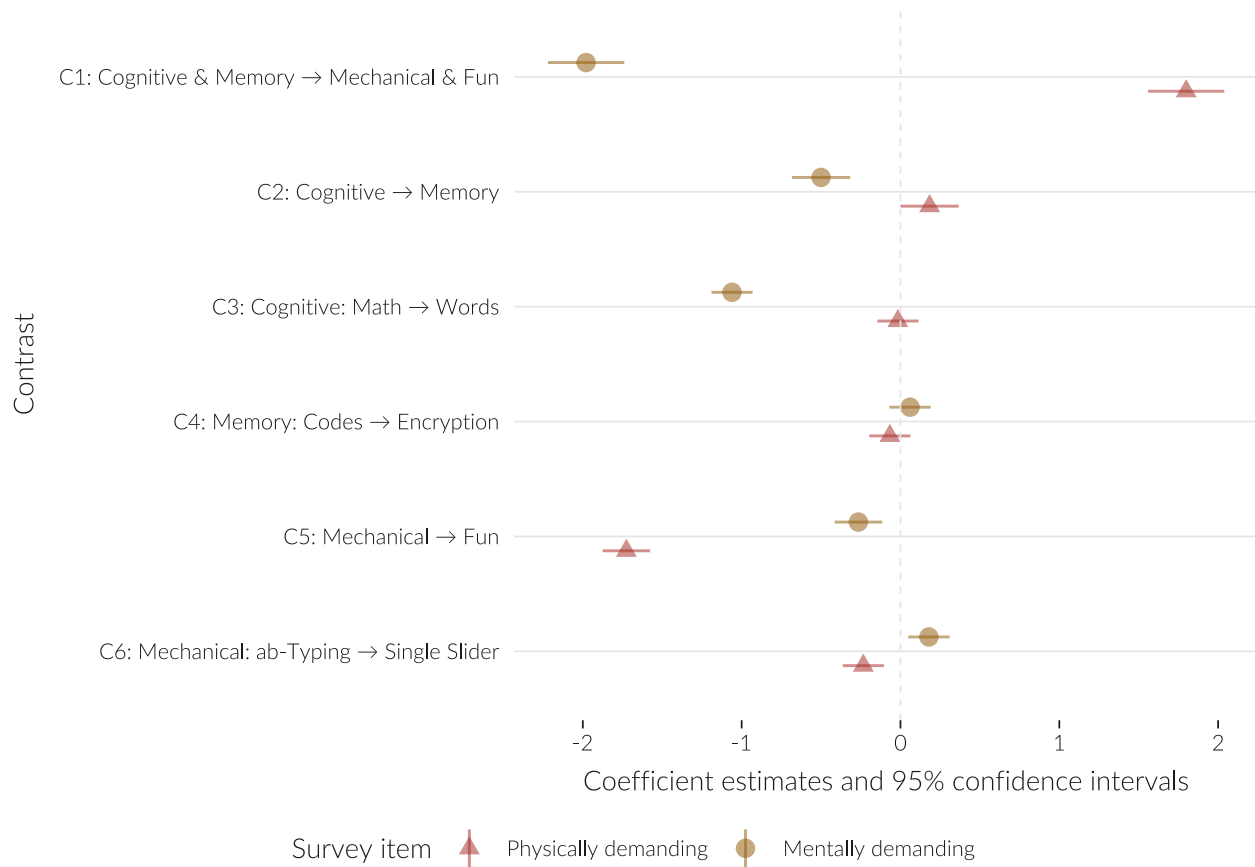


Figure 3.7: **Plot of regression coefficients:** Estimates and respective 95% confidence intervals for all orthogonal contrasts and all effort type survey items.

3.4.3 Combine the Findings

The subjects perceive the tasks to be very different along both the motivational and effort dimensions of the survey. Figure 3.8 combines these findings and places them into relation. To this end, the motivational dimensions of the survey are aggregated into a single score. The figure highlights the distinct difference between (1) mentally highly or (2) mentally moderately demanding tasks and (3) physically highly demanding tasks. The latter group of tasks is rather dull and repetitive and is, therefore, perceived by the study participants as much less pleasant and enjoyable than the former. A fourth group of tasks is neither physically nor mentally challenging; it mirrors an existing game and is perceived as a rather pleasurable activity. What is true for the ball-catching task also holds to some extent for the word-transcription task: while it is one of the more cognitively demanding tasks, it is

highly varied and interesting and, therefore, the most fun. The figure gives a first overall picture and allows to recognize possible relations between type and amount of demanded effort and motivation. The observed results are in marked contrast to [Carpenter & Huet-Vaughn \(2017\)](#), who state that arithmetic tasks, decoding, and typing tasks are not intrinsically interesting.⁴³ Agreement between both studies is given for tasks which [Carpenter & Huet-Vaughn \(2017\)](#) refers to as “computer tasks,” which include the single-slider task and which they consider uninteresting.

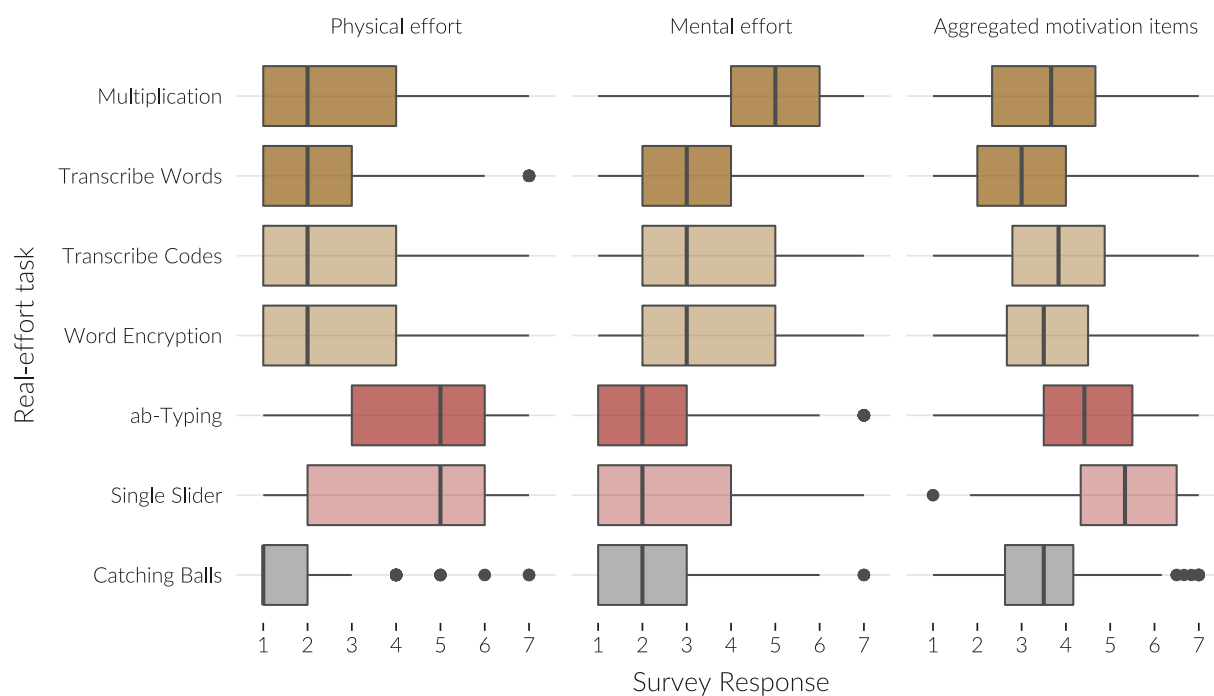


Figure 3.8: **Differentiate tasks along three dimensions:** The motivational dimensions of the survey are treated as a summated rating scale and are aggregated into a single score (*mean*); both effort dimensions are non-aggregated. Consequently, the width of the confidence intervals cannot be compared across the three dimensions of the survey. Higher median values (centered line in the middle of each box) correspond to a less motivating task and more effort demanding task. A smaller box and shorter whiskers indicate a greater congruence in the subjects' perception of a task. Tasks are color-coded by the *effort type* they demand: (1) mentally very demanding tasks and (2) moderately mentally demanding tasks in shades of brown, (3) physically very demanding tasks in shades of red, and (4) neither mentally nor physically demanding tasks in grey. The figure offers a first indication of how the type and amount of effort required can translate into or affect the subjects' motivation.

⁴³[Carpenter & Huet-Vaughn \(2017\)](#) introduce a typography of commonly used real-effort tasks and present it in a tabular summary. The table contains an example for each task type, usage frequencies in the literature, and an assessment by the authors as to whether the output of the task is typically useful or interesting in itself.

3.4.4 Identify a Favorable Task for Experimental Research

If the conception and design of a task give rise to additional non-monetary incentives, subjects may provide effort irrespective of the monetary incentive. If earnings can be achieved without greater effort, the task may not be a good measure of effort in the first place. Thus, a “favorable task” in terms of [application areas one and two](#) i) *does not motivate subjects to make a voluntary effort*, and ii) *requires a substantial effort to be completed*. Since people are different, the fulfillment of these attributes will vary from person to person in each task. Yet, the more similarly motivating the task is, the less pronounced is any motivational bias. Likewise, the more similarly strenuous a task is for the subjects, the lower the ability bias.⁴⁴

The new methodology presented in this chapter allows to investigate how *subjects perceive* tasks regarding these “attributes of a favorable task.” The survey determines the *subjective assessment* of the tasks by the study participants. The *Multiple Comparisons with the Best* (MCB) method of [Hsu \(1996\)](#) can be applied to these assessments to identify a task from the examined selection that most closely matches the attributes described.⁴⁵

The multiple comparison method analyzes the differences between level means. It determines the factor levels with the highest mean (the “best”), those that cannot be distinguished from these “best,” and those that differ significantly from the best (see also [Kuehl, 2000](#); [Oehlert, 2010](#)). In the present case, one obtains a different mean for each task and survey item. The “best” is defined as the *highest mean survey response*, which, depending on the survey item, corresponds to a very demotivating or very strenuous task. For each survey item, the subset selection procedure yields the best tasks and those that do not differ significantly from them. The tasks in this “group of best tasks” are not significantly different from each other but significantly better than the other tasks. The comparison is based on the construction of simultaneous confidence intervals, each for the difference between the mean of a certain level and the best mean of the remaining levels. When a confidence interval has zero as its endpoint, there is a statistically significant difference between the respective mean values.

⁴⁴In this regard, the discussion of *neutral task* in the introduction to this thesis and *generic tasks* in Chapter 1 may also be considered.

⁴⁵A ranking based on the mean survey responses for each item may serve as a simple alternative for exploratory analysis. However, this approach will not take the magnitude of the (differences of the) means into account.

Since there are seven tasks, there are $k = 7 - 1 = 6$ comparisons to the best mean.⁴⁶

Multiple comparison tests can be performed on means, but also on variances. Thus, the analysis is carried out in two ways: Firstly, as a comparison of the *mean values* in order to examine which tasks are perceived as most strenuous and most demotivating (“best” is defined as *highest mean*). Secondly, as a comparison of the *variances* to assess which tasks are perceived most similarly in both dimensions (“best” is defined as *lowest variance*).⁴⁷

The results of the *multiple comparisons of mean survey responses* are summarized in Table 3.6 for all tasks and survey items. For most motivational survey items, the single-slider task is identified as the “best task.” For survey item 5 it forms the “group of tasks that are the best” together with the ab-typing task. These tasks do not differ significantly from each other but are significantly better than the remaining tasks. The ab-typing task turns out to be most demanding in physical effort (“best” along item 6). Nevertheless, the single-slider task is not far from being included in a group of best for this survey item. Finally, the multiplication task is identified as “best” in terms of mental effort demand.

Beyond the first application of the MCB method to compare the means of the subjects’ ratings, the method can be applied a second time to compare the *consistency between the subjects’ ratings*, i.e., the variances. More precisely, the reasoning is as follows: The more similarly motivating a task is perceived, the more likely it is that the task is actually equally motivating for all study participants, so that any motivational bias is less pronounced. Likewise, the more a task is perceived as similarly strenuous by the subjects, the more likely it is that the task will actually be similarly effort demanding for all of them, so that less ability bias can be expected. In this application to variances, Hsu (1996)’s MCB method constructs a confidence interval for the difference between the value of the variance of a given level and the best variance for the remaining levels. Thereby, “best” is defined as the *lowest variance* in survey responses for a given item and task. When a confidence interval has zero as its endpoint, there is a statistically significant difference between the respective variances.

Table 3.7 summarizes the results for all tasks and survey items. By far the most agreement in subjects perception is given for the ball-catching task. Also, the word-transcription task and the single-slider

⁴⁶The family-wise error rate is fixed at $\alpha_E = 0.05$, and the individual error rate α_C is adjusted accordingly to achieve it.

⁴⁷Appendix B.3.2.4 provides more details on the calculations of the multiple comparison procedure and provides an example for the last survey item.

task show a lower variance for several questions. In general, the variances do not differ as strongly as the means, as can also be seen in the illustrations of the distributions and tables of the data (see Tables 3.2 and 3.4 and Figure B.54 in the Appendix).

To summarize, the subjects perceive the tasks to be very different (cf. median values in Figure 3.8 and comparison of means in Table 3.6). Furthermore, the subjects perceive a given task very differently (cf. width of boxes and whiskers in Figure 3.8 and comparison of variances in Table 3.7). Some tasks appear much less motivating to the subjects than others (e.g., *ball-catching task* compared to the *ab-typing task* and the *single-slider task*). In addition, certain tasks are perceived more similarly by the subjects (e.g., the *word-transcription task* opposed to the *multiplication task*). The heterogeneity and greater variance in responses for different tasks is further exemplified in Appendix B.3.2.7, providing illustrations of the congruency in subjects' task assessments. Superimposed contingency tables of "inter-rater agreement" are visualized as heat maps for each task and the different dimensions of the real-effort task survey.⁴⁸

⁴⁸The reasons for observing more or less agreement between study participants can be manifold. Appendix B.3.2.7 discusses this in more detail with reference to the literature on *inter-rater reliability* or *inter-rater agreement measures*.

Table 3.6: **Multiple comparisons with the highest mean to identify a favorable task:** The MCB method creates simultaneous confidence intervals (CIs) for all seven tasks to determine which task performs the “best” in terms of highest mean survey responses. The CIs indicate whether a task is “best” (*lower bound* of CI is zero), whether tasks are “insignificantly different from best” (CI contains zero), and whether tasks are “significantly different from best” (*upper bound* of CI is zero). For five out of six motivational items, the single-slider task is perceived significantly worse (higher mean) than any other task as their CIs are entirely below zero. For these items, the task is identified as the “best.” For the fifth motivational dimension (*meaningfulness*), there is no evidence to indicate a significant difference between the single-slider task and the ab-typing task because both of their CIs contain zero. Together they form the “set of best tasks in terms of meaninglessness.” With regard to demand in physical effort, the ab-typing task is “best,” since again, the CIs for all other tasks are completely below zero. According to Hsu (1996), a lower bound approaching zero indicates that the task is close to the best. This holds true for the single-slider task, whose CI is much closer to zero than that of the remaining tasks. Thus, the task just missed being included in the “group of best tasks” in terms of physical demands. In terms of mental effort, the multiplication task is perceived as the most demanding task.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	Physically demanding	Mentally demanding	No outcome
Multiplication	No (-1.96,0)	No (-1.49,0)	No (-2.21,0)	No (-0.98,0)	No (-2.75,0)	No (-2.61,0)	Yes (Best) (0,1.88)	No (-1.97,0)
Transcribe Words	No (-3.37,0)	No (-1.92,0)	No (-2.99,0)	No (-2.59,0)	No (-2.65,0)	No (-2.64,0)	No (-2.5,0)	No (-1.82,0)
Transcribe Codes	No (-2.2,0)	No (-1.29,0)	No (-2.02,0)	No (-1.84,0)	No (-1.9,0)	No (-2.37,0)	No (-2,0)	No (-1.26,0)
Word Encryption	No (-2.42,0)	No (-1.73,0)	No (-2.26,0)	No (-1.76,0)	No (-1.89,0)	No (-2.5,0)	No (-1.88,0)	No (-1.37,0)
ab-Typing	No (-1.61,0)	No (-1.76,0)	No (-1.44,0)	No (-1,0)	Yes (-0.6,0.17)	Yes (Best) (0,0.85)	No (-2.9,0)	No (-0.86,0)
Single Slider	Yes (Best) (0,1.61)	Yes (Best) (0,1.29)	Yes (Best) (0,1.44)	Yes (Best) (0,0.98)	Yes (-0.17,0.6)	No (-0.85,0)	No (-2.54,0)	Yes (Best) (0,0.86)
Catching Balls	No (-2.65,0)	No (-1.49,0)	No (-2.38,0)	No (-2.65,0)	No (-1.83,0)	No (-3.21,0)	No (-3.11,0)	No (-1.38,0)

Table 3.7: **Multiple comparisons with the lowest variance to identify a favorable task:** In a second application, the MCB procedure serves to identify the tasks that perform best in terms of least variance. The CIs *now* indicate whether a task is the “best” (*upper bound* of the CI is zero), whether tasks are “insignificantly different from the best” (CI contains zero), and whether they are “significantly different from the best” (*lower bound* of the CI is zero). The *ball-catching task* is most frequently among the “set of best” tasks (five times) and twice the “best” task (for physical and mental effort) in terms of variability in task perception. The *word-transcription task* and the *single-slider task* are four and three times, respectively, among the “set of best” tasks. The remaining tasks are once or twice among this set.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	Physically demanding	Mentally demanding	No outcome
Multiplication	No (0,2.71)	No (0,1.81)	No (0,1.44)	No (0,0.98)	No (0,2.18)	No (0,2.05)	No (0,1.59)	Yes (-0.34,0.42)
Transcribe Words	Yes (-0.53,0.23)	No (0,1.06)	Yes (-0.62,0.15)	Yes (-0.41,0.36)	No (0,1.86)	No (0,1.29)	No (0,0.94)	Yes (-0.18,0.58)
Transcribe Codes	No (0,1.13)	No (0,1.36)	No (0,1.13)	No (0,0.89)	No (0,1.78)	No (0,1.51)	No (0,1.79)	Yes (-0.2,0.57)
Word Encryption	No (0,1.24)	Yes (-0.55,0.22)	No (0,1.26)	No (0,0.83)	No (0,1.57)	No (0,1.3)	No (0,1.74)	Yes (-0.42,0.34)
ab-Typing	No (0,1.92)	No (0,1.29)	No (0,1.4)	No (0,1.78)	Yes (-0.52,0.25)	No (0,2.17)	No (0,1.94)	Yes (-0.22,0.55)
Single Slider	Yes (-0.05,0.72)	No (0,1.68)	Yes (-0.15,0.62)	No (0,1.49)	Yes (-0.25,0.52)	No (0,2.87)	No (0,2.62)	No (0,0.84)
Catching Balls	Yes (-0.23,0.53)	Yes (-0.22,0.55)	Yes (-0.05,0.72)	Yes (-0.36,0.41)	No (0,1.5)	Yes (Best) (-1.29,0)	Yes (Best) (-0.94,0)	Yes (-0.34,0.43)

3.4.5 Survey Validation

In the following, evidence is provided that the real-effort task survey is a valid and reliable measure. As a first step to verify the validity of the survey, one can consider the extreme cases in the regression analysis for the motivational survey items (see Table 3.3 and Figure 3.6). As intuitively expected, *mechanical tasks* and *playful and entertaining tasks* differ greatly from cognitive tasks (*first contrast*). On the contrary, the *memory tasks* do not differ significantly from one another (code-transcription task and word-encryption task; *fourth contrast*). In addition, the task classifications derived from the data presented in Table 3.4 and Figure 3.8 are consistent with the proposed grouping based on task similarities (compare Figure 3.5). Both of these findings provide support to the consistency of the survey. Moreover, an explorative factor analysis for the survey reflects the *intended survey structure* very well, which substantiates its content validity (see Appendix B.3.3.1).

To examine the concurrent validity of the real-effort task survey, its experimental findings are compared with a second measure that determines the “popularity” of the tasks. The measure was implemented according to the approach of Rheinberg (1989) to create an *individual ranking scale* termed “personal-hit list” (PHL). His approach is mainly employed in motivational psychology within the diagnosis of incentive qualities to assess the (relative) strength of activity-related incentives. Here the scale was used to elicit subjects’ overall impressions of the tasks once they had completed all tasks. To construct the individualized scale, subjects were asked to name an activity which they currently enjoy performing (e.g., “swimming in the lake”) as well as one they dislike to do (e.g., “cleaning the dishes”). Moreover, they had to state a neutral activity centered between both extremes (see Figure B.46 in the Appendix for illustration of the experimental implementation of the scale).⁴⁹ By construction of the individual ranking scale, subjects may hold the same view of two tasks and assign them an identical position on the scale. This particular design does not permit to obtain a complete ordering of individual preferences across tasks. However, assuming that the reference points of the scale are similarly placed and distant from another for all study participants, the aggregation of the individual views is likely to resemble a *general perception* or the *popularity* of any given task. The left graph in Figure 3.9 depicts this popularity measure for all tasks in the selection. The right graph summarizes the

⁴⁹Table B.25 lists the activities subjects stated as “strongly disliked activity” and “strongly liked activity”; Figure B.59 depicts the distribution of the responses to the personal-hit list for all tasks.

data for the real-effort task survey. Subject's responses are aggregated across all eight survey items for each task.⁵⁰ To facilitate comparison between the two measures, the aggregate scores for the real-effort task survey were recoded so that higher values in this composite measure correspond to a *more positively perceived task or a lower demand for effort*. Both diagrams show great similarity, which is confirmed by correlation analysis (see Appendix B.3.3.3). The individual ranking scale reaffirms the overall findings of the real-effort task survey and, thereby, corroborates its concurrent validity.

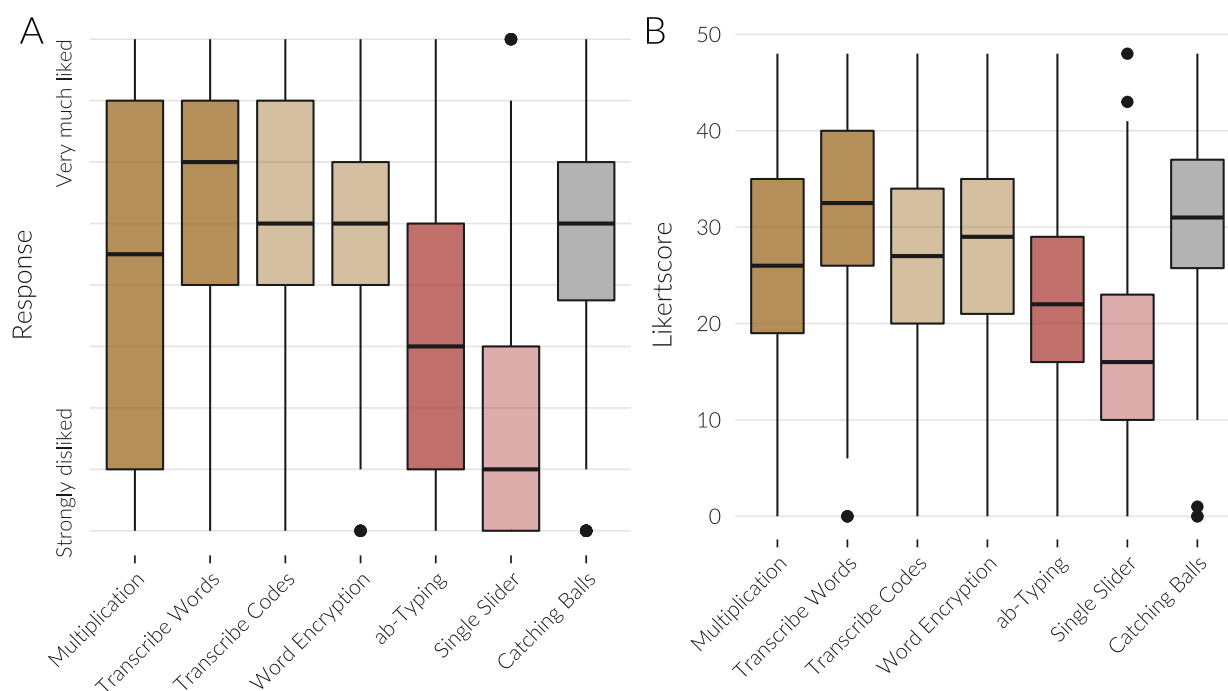


Figure 3.9: **Overall task impression based on a subjective scale (personal-hit list, left) and combined task perception based on responses to the real-effort task survey (right):** A) *Personal-hit list*: After completing the set of tasks, subjects had to rank these on a subjective scale, following Rheinberg (1989). By aggregating subjects' overall impression of the tasks, one can determine the "general perception" or "popularity" of the tasks. The single-slider task turns out to be the least popular task in the set. B) *Real-effort task survey*: The responses to all survey items are aggregated into a single score for each subject and task (aggregate scores are recoded to allow comparability with the PHL). This provides a composite measure of subjects' perception of a task. The composite measure highly correlates with the personal-hit list (see also Figure B.60 in the Appendix), giving further support to the validity of the real-effort task survey. For both measures, the variability in responses is larger for the multiplication task than for any other task.

⁵⁰The aggregate scores for the real-effort task survey summarizes the eightfold amount of data (eight survey responses per subject and task) than for the personal-hit list (one observations per subject). Consequently, the interquartile ranges (size of the boxes in the box plot) for the survey scores are mostly smaller than for the personal-hit list.

The internal consistency of the real-effort task survey was examined with Cronbach's Alpha (Cronbach, 1949). Overall, the survey shows good reliability, $\alpha = 0.86$. For the subscale of motivational items, reliability is excellent, $\alpha = 0.90$. Internal consistency reliability is further substantiated by examining the correlation between survey items for a particular task. As expected, a high correlation is observed across the motivational survey items, but not across both effort items.

Finally, it is important to note that, due to selection bias, the results obtained are only meaningful for the sample at hand, i.e., students of the University of Hamburg. Their distribution in cognitive and physical abilities as well as in preferences for particular tasks cannot readily be transferred to other subject groups.

3.5 Conclusions

At the beginning of the chapter, two main aspects are highlighted, along which tasks differ: *first*, the extent to which they motivate voluntary effort so that subjects perform regardless of any monetary incentive; *second*, whether earnings can be achieved without any “blood, sweat, and tears.” To investigate and evaluate tasks in this respect, this chapter presented a new tool for distinguishing tasks, named the real-effort task survey. The survey inquires the subjective perception of the study participants regarding the design of tasks along motivational dimensions and with regard to the effort required. The insights gained allow an initial evaluation of i) whether a task is actually strenuous, and thus can serve as a measure of effort, and ii) how susceptible it is to induce voluntary effort by additional non-monetary incentives, and, therefore, can lead to biased results.

In a first application, the survey was used to compare a selection of seven distinct tasks. On their own, the responses of the individual subjects represent only a subjective assessment of a task. Taken together, the two hundred and forty-eight task assessments form a comprehensive evaluation of the properties of a task. Thus, a greater degree of agreement in attributing a particular property to a task renders it more likely that the task actually does possess that property. The recorded responses to the real-effort task survey i) testify the heterogeneity among the selection of tasks, ii) exemplify how differently subjects perceive tasks, and iii) give an indication of how much motivation tasks can evoke in study participants, which can confound any effort measurement.

The survey identifies several potential sources of bias among the compared tasks.

- The word-encryption task and the ball-catching task are likely to be completed by study participants due to inherent task enjoyment (see survey items 1, 3 and 4 in Table 3.2). Conversely, the multiplication task, the ab-typing task, and, in particular, the single-slider task are noticeably less enjoyed.⁵¹

⁵¹The study also included an outside option. 84.3% of the study participants were given the possibility to cancel the current task in order to switch to an alternative activity. However, they could not resume the task to earn more money. After the end of the current task period, they proceeded to the next task just like the remaining subjects. Rather few subjects decided to make use of the outside option. Subjects are most inclined to do so for the single-slider task (4 subjects) and particularly for the multiplication task (20 subjects) (see also Appendix B.3.3.2). This provides further evidence that the subjects have strong feelings about these tasks. However, whereas the single-slider task appears to be

- The multiplication task and the word-encryption task are perceived by the subjects as much more useful than the remaining tasks: they appear to be *meaningful* and to *have an output* (see survey items 5 and 8 in Table 3.2). This, in turn, may lead to effort being exerted due to induced *activity-related incentives* and *purpose-related incentives rooted in self- and other-evaluative consequences* (e.g., subjects exert effort because they assume the task serves a particular purpose for the researcher).
- A lower variance indicates that if there are any spillover effects from the non-monetary incentives on effort provision, they are at least similar for all subjects. The multiplication task is highly regarded by some of the subjects and strongly disliked by others (high variances observed for most motivational items of the survey, see Table 3.2 , as well as for the personal-hit list, see Figure 3.9). Since the task is, therefore, perceived very heterogeneously, it is very likely that the motivational bias is larger for the multiplication task than for other tasks (see Appendix B.3.4 for a more detailed examination of the individual preferences of the study participants).
- The multiplication task (mental effort), the ab-typing task, and the single-slider task (both physical effort) involve considerably more exertion than the remaining tasks. Generally speaking, simple, physical tasks seem to require less skill than mentally demanding tasks, and are, therefore, less prone to ability bias than the latter. The multiplication task, as a representative of the latter type, is perceived by the subjects as highly demanding in terms of mental effort and thus most likely yields biased results.
- Assuming that subjects' perceptions of the tasks are valid, this means that the more a task is perceived as demanding, the more effort is indeed required to complete it. Accordingly, it is debatable whether tasks that are *not* perceived as strenuous actually require considerable effort – and represent an appropriate measure of effort provision. Among the tasks examined in the study, the ball-catching task most likely does not capture effort very well. Given the responses to the survey, the effort measurement obtained with this task might rather reflect the enjoyment of

generally unpopular (in the personal-hit list 56.9% indicate that they strongly disapprove the task), the multiplication task is perceived very heterogeneously by the study participants (only 28.2% strongly dislike the task).

the task and the cognitive abilities of a subject, as the task resembles a “challenging” optimization problem.

- By relating subjects’ perception of the effort and motivational dimensions, one gains first insights into how the conception of tasks and their demand for physical or mental effort might translate into motivation. Put differently, how the subjects perceive certain task designs as motivating and, therefore, tend to make voluntary efforts.

Based on subject’s task assessments, [Hsu \(1996\)](#)’s method of multiple comparison with the best was used to identify a favorable task.⁵² Along several motivational and effort-related survey dimensions as well as with regard to the degree of agreement among subjects along these, the single-slider task is identified by the method. The task is perceived by subjects as least motivating in terms of most motivational items of the survey. Therefore, it appears unlikely that the task is performed simply due to any activity-related or purpose related incentives; the obtained results are thus likely unbiased by additional non-monetary incentives. Finally, the subjects attribute the task a high demand for physical exertion, which indicates that it actually does measure effort.

The findings are corroborated by the results of the *Personal Hit-List* by [Rheinberg \(1989\)](#), which made it possible to record the subjects’ overall impression of a task and to obtain their “general perception.” The single-slider task turns out to be clearly the least popular.

The presented methodology for comparing tasks and the conducted experimental study bear some limitations. Generally, the scope and informative value of a survey are restricted by practicality and usability. The real-effort task survey is designed to be completed by experimental subjects. Understandably, these can only assess certain dimensions of a task. This, to some extent, limits the scope of the survey and, in consequence, subsequent judgments and comparisons of tasks. In particular, task-dependent design practices relating to the output production function cannot be evaluated by a novice. Therefore, these practices can equally not be analyzed through a survey filled in by study

⁵²Initially, a more sophisticated measure for assessing consistency between subjects’ judgments was sought (see Appendix B.3.2.7 for a brief review of the literature and measures of inter-rater agreement). A stimulating exchange with Professor Krippendorff between December 2019 and February 2020 led him to develop a new measure suitable for my purposes. It relates to the familiar measure of inter-rater agreement termed *Krippendorff’s Alpha*. However, an adequate treatment of the new approach would go beyond the scope of this thesis. Especially since the already presented method of multiple comparisons of Hsu, applied to the variances, also offers some deeper insights.

participants. Such crucial dimensions of real-effort tasks require the judgment of an expert and should be examined independently. As an example, the sensibility of experimental results obtained with a specific task towards stake sizes demands separate investigations.

Similarly, the survey is not capable of providing detailed information about which task actually tracks exerted effort more closely and precisely. To examine and increase the accuracy in the measurement of subjects' exerted effort levels, an in-depth monitoring of strain-and-stress indicators with the help of medical devices (blood pressure, neuroimaging, etc.) is needed. Using such techniques, however, to examine which task most accurately describes the actual effort involved is very costly and technically elaborate, and goes beyond the scope of this work.⁵³

The set of [design criteria](#), which form the bases of the survey, are derived from a systematic evaluation of the real-effort literature.⁵⁴ Nevertheless, the survey's scope is constrained in the sense that the criteria on which it is based are the result of a somewhat subjective selection; alternative survey formats and item choice are, therefore, conceivable. As described, the survey was designed with practicality and usability in mind; therefore, any extension may come at the cost of these.

A final limitation of the study concerns the randomization procedure. To counter order effects, the succession of tasks and, therefore, the subsequent filling-in of the survey was randomized across 24 groups. Consequently, the parsimonious model presented in Section [3.3.2](#) does not account for any order effects. To examine this potentially imperfect randomization strategy, Appendix [B.3.2.5](#) presents an extension to the simple model. Therein, both order effects (in terms of the position of a task in the task sequence) and group belonging (variability between different groups with regards to the general task liking level) are considered. As it turns out, this complex model does not provide a better fit of the data than the simple model permits.⁵⁵

⁵³A subjective assessment by the subjects about their actually exerted effort levels may be added to the survey to provide complementary insights.

⁵⁴As mentioned in the previous chapter, the design criteria and design practices are not prioritized in any way. In very special cases, Hsu's method of multiple comparison with the best may not help to obtain a satisfactory ranking of the considered tasks. For example, one can imagine the following situation in which two tasks both belong to the "set of best tasks" and are thus equally good across all but two motivational dimensions: One task is substantially more *tedious* than the other, which conversely is completely *meaningless*. The researcher now faces a trade-off between the two respective design practices to counter non-monetary incentives: Should the more tedious task be chosen, to address *activity-related incentives*, or the more meaningless task, to mitigate *purpose-related incentives*? Under these "extreme" circumstances, the multiple comparison procedure may not be able to determine favorable tasks.

⁵⁵One could further consider a repetition of the study in which the order of the survey items is randomized. For reasons outlined in Section [3.2](#), it is not expected that the accuracy and validity of the results will improve.

For the practitioner, the purpose of the real-effort task survey is twofold: *First*, given a particular task the survey allows to obtain a comprehensive evaluation of the properties of the task. In [task application areas one and two](#), the survey can thus be used to control for additional non-monetary incentives and whether the task was physically or mentally demanding for a subject in a later analysis.⁵⁶ In [task application area three](#), the survey can help identify tasks that meet specific requirements, such as demanding a particular level of effort, encouraging voluntary effort to a desired level, or being perceived by subjects similarly or not. *Second*, given a set of tasks, the survey items allow to compare the tasks in terms of the task-dependent design practices to counter activity-related and purpose related incentives and to assess whether they are more mentally or physically demanding. The MCB method from [Hsu \(1996\)](#) can be applied to identify a favorable task for experimental research: with the *mean survey responses* it allows to determine a task that subjects perceive most negatively along the motivational survey items and most demanding in terms of required effort.⁵⁷ If subjects' task ratings are consistent (small *variance of the responses*), motivational influences and effort demands can be expected to be the same for them. The survey is not limited in its application to computer-based tasks, and manual tasks may also be examined.

The students' perceptions of the tasks revealed major differences in their design with regard to motivational aspects. The findings provide a first taste of the extent to which design-related motivational incentives have an effect on effort provision in addition to monetary incentives for different tasks. Thus, it seems relatively likely that tasks that are perceived as highly motivating will be performed by subjects quite irrespectively of the incentive scheme. The survey can be helpful in detecting these tasks, which may be more susceptible to yielding confounded experimental results. The next chapter continues this line of thought and seeks to determine how much of the variation in effort expended can be explained by non-monetary incentives and individual characteristics.

⁵⁶In laboratory experiments, it is likely that the external reward initially motivates effort. In the course of the task, however, this effort is then partly sustained by emergent activity-related incentives and undesired purpose-related incentives that are triggered when the activity is performed. Such a change in the form of motivation in the course of the task is by no means far-fetched. Therefore, applying the survey also to tasks where non-monetary incentives are not particularly obvious at first constitutes a useful precaution.

⁵⁷Only if tasks require considerable effort to complete and this from all study participants, one can obtain a fine-grained mapping of any monetary incentive to provided effort.

4

Determinants of Real-Effort Task Performance

4.1 Introduction

Since effort in itself is difficult to measure, it is commonly approximated by the *performance in the task*, i.e., by the output produced or the observed score. However, performance may involve more than just “intentional effort,” which is the exertion that people deliberately and consciously make to complete a task. This raises the question of *which factors contribute to a “measured effort” beyond incentive effects*.

According to the general model of motivation psychology, an individual’s *current motivation* to make an action is a compound of her needs, motives and goals, and the situation she faces which entails a set of opportunities and incentives (J. Heckhausen & Heckhausen, 2018, p. 4; see also Rheinberg & Vollmeyer, 2012, p. 70). Applied to real-effort task experiments, a subject’s current motivation to

provide effort results from the interaction between her individual mindset, the given task and setting, as well as a set of incentives (see Figure 4.1). This theoretical approach is discussed in more detail in Chapter 2. In addition to the financial incentives, non-monetary incentives could be at play and trigger voluntary effort.

Besides, Gill & Prowse (2015, p. 4) point out that unobserved differences in *abilities*, which are demanded by a task, may result in heterogeneity of the effort costs across subjects and, therefore, lead to a skill-bias. Section 1.3 introduced a novel classification of tasks. It extends the argument from Gill and Prowse to also include *personality traits* to be equally decisive to subjects' performance (also consider Lezzi et al., 2015). Because if certain personality traits are beneficial in a task, the subjects who possess them will have lower effort costs. The person-situation interaction is, therefore, much more complex than the simple model described above suggests: Tasks require specific skills *and* personality traits, and the subjects who enter the laboratory possess these to a greater or lesser extent. The individual predisposition of both can, in turn, have spill-over effects on the subject's current motivation (and thus again on their actions).

In light of the *applications of the tasks* mentioned at the beginning of this thesis, the consequences of these influencing factors become apparent (see also the discussion in Section 2.2). In the first and second applications, the objective is to choose a task that is as "neutral" as possible so that completing the task depends as little as possible on subject characteristics. In this way, one can dispense with the need to control for these in the later analysis. However, should it be inevitable to use a task type that depends on particular characteristics – e.g., to provide a relation to the literature – then it is at least helpful if it is easy and uncomplicated to control for the characteristics.

In application three, the argumentation differs fundamentally. The goal is no longer to reduce the influence of the subject characteristics to a minimum or to control for them. Instead, the congruence between the experiment and reality is the main criterion for the choice of the task. Ideally, the same conditions that prevail in a real work situation are reproduced – including any aroused feelings and motivations or dependencies on abilities or personality. All this aims at reaching mundane realism and increasing external validity. To make this possible, the same subject characteristics as those observed in the real situation must be decisive for the task execution in the experiment. In the run-up to the experimental investigations, it is, therefore, essential to examine and determine the applicable

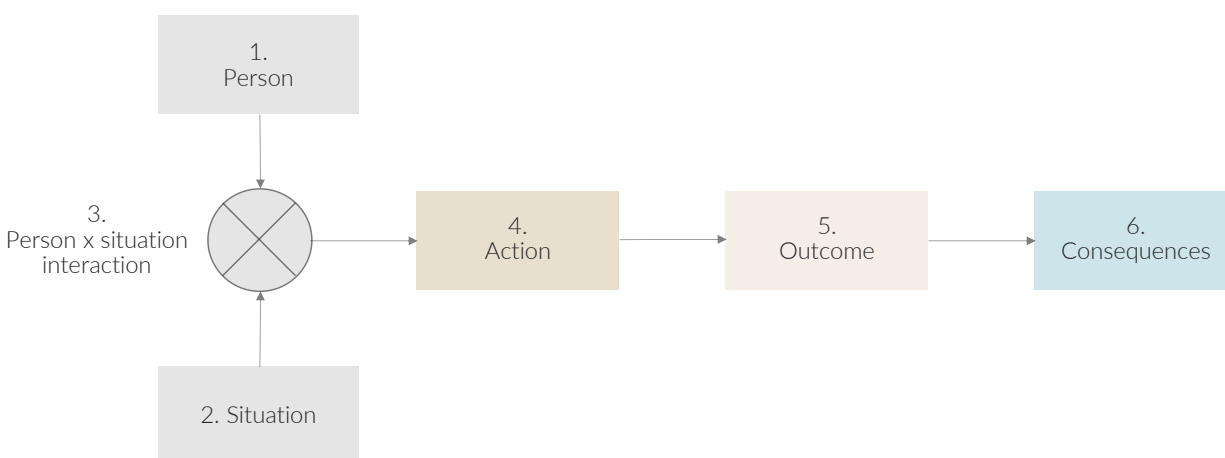


Figure 4.1: **Determinants and course of motivated action:** according to the general model of motivation psychology (adapted from J. Heckhausen & Heckhausen, 2018, p. 4). The *study participant* (1) enters the laboratory with a given set of *needs, motives, and goals*. The experiment in which she participates defines the *situation* (2), providing her with opportunities and potential incentives. More precisely, the subject faces a given real-effort task, and her efforts in the task are remunerated with a fixed piece-rate. The *interaction between the subject and the situation* (3) leads to the subject's *current motivation*, which, in turn, determines her *action* (4), i.e., her effort in the task. The *outcome* (5) of her efforts are the experimental points she collected. These in turn can have various *consequences* (6), including the material benefit of the earnings, but also self-assessment and evaluation by others. However, the situation in the laboratory is somewhat more complex than the simple model suggests: The task requires certain skills or personality traits or favors the performance of those who have them. This, in turn, affects the current motivation and actions of the subjects, who become aware of it and thus are more motivated and more easily handle the task – or not and, therefore, struggle to cope with it.

and relevant characteristics in order to subsequently identify a matching task that takes these into account.

The considerations regarding the choice of a task, in the various applications, make it clear that the subjects' characteristics invariably have an influence and entail consequences in real-effort experiments. An investigation of the relationship between the subjects' characteristics and their performance is, therefore, essential for all three applications of tasks, whether i) to compare tasks with regard to the presence of subject characteristics to identify a (preferably) neutral task, ii) to examine which subject characteristics need to be checked later on if it is known that some of them have an influence, iii) to ensure that subject characteristics are actually present if this is necessary to replicate a real work situation as closely as possible in order to achieve congruence.

As a first step in this direction, this chapter examines effort responses, to a given incentive, in terms

of a variety of subject characteristics for a set of different tasks. More precisely, the measured effort is compared to an individuals' *qualities* (i.e., *abilities* and *personality*) and *motivation*. Thereby, subjects' effort is approximated by produced output. To this end, the laboratory experiment described in Chapter 3 contains several additional elements. Subjects are characterized in terms of their qualities and motivation with a range of psychological measures. Next, they complete the tasks. Subjects' performance is then compared to the characteristics, separately for each task.

Different approaches are adopted to characterize the subjects with regard to their qualities and motivation. The *task classification* presented in Section 1.3 serves to systematically distinguish existing tasks to form a compilation of tasks with a high degree of heterogeneity (see Section 3.3.1.1). Here, the classification also helps in the systematic selection of subject qualities that are likely to contribute to the measured effort (see Table 1.1 and Table 1.2). Since the tasks are chosen to be as extensive as possible, these determinants naturally vary considerably from task to task.

With the help of *motivational diagnostics*, psychologists seek to decipher and record the activating target orientation of behavior and the underlying personal and situational conditions in a structured way (Rheinberg, 2004). This is to support desired behavior or to reduce or redirect undesirable behaviors. To facilitate this, Rheinberg (2004) introduces a diagnosis scheme that builds on the extended version of Heckhausen's *Advanced Cognitive Motivation Model* presented in Chapter 2. The scheme allows to decompose a given situation one step at a time in order to identify the existing forms of motivation and problems of motivation. Rheinberg (2011) discusses a variety of measures to assess each step of the diagnosis scheme. A selection of these measures is used to investigate the motivation of subjects to make an effort in the tasks. Several analytic approaches are pursued to examine subjects' performance in each task in terms of their individual characteristics. To explain actual past behavior, a simple linear regression analysis is performed (both separately and jointly) for subjects' qualities and their motivation. In order to predict the performance of prospective subjects in future laboratory experiments, a machine-learning algorithm is trained and the model with the highest predictive accuracy is determined.

The literature contains rather few investigations related to the current approach. In a pre-test of their study, Huang & Murad (2017) compare five real-effort tasks to obtain a task which is very dissimilar from a number-adding task. They settle upon the circle task from Hollard et al. (2016) since it does

not show any correlation with the mathematical task in terms of subject's performance, their perception of the difficulty of the task, and their retrospective choices of submitting past performance to comparative pay (Huang & Murad, 2017, p. 7). As in this work, the authors are able to differentiate tasks based on information provided by the subjects; however, in a much less extensive and less sophisticated way.¹ Much closer to the present study is the approach presented by Lezzi et al. (2015). They compare chosen effort with three real-effort tasks in a contest game (the slider task from Gill & Prowse (2015), a number counting task as in Abeler et al. (2011), and number adding task). The authors characterize subjects in terms of anxiety, risk aversion, and gender and find that these factors play a different role for each task they investigated. Thus, their study provides first insights into how decisive differences in task properties and subject characteristics are for experimental results. Despite this valuable finding, the characterization of the subjects made by Lezzi et al. (2015) is nevertheless rather limited. Moreover, the range and variability of their tasks (two mathematical tasks and one generic task, which is, however, susceptible to learning)² are very limited compared to the task selection of the thesis. Hence, the present study represents a comprehensive extension of the investigation of Lezzi et al. (2015). It contains both a more extensive and profound rationale for the selection of the tasks and for the approach taken to characterize the study participants.

Regarding the investigation of subjects' motivations for performance by methods of motivational diagnostics, the literature citing Rheinberg (2004) was reviewed. The author's diagnosis scheme is usually applied to assess motivation on a *case-by-case* basis and then moving through the scheme *step-by-step*. The approach taken here is based on applying the scheme *in its entirety at once* and *for many study participants simultaneously*. Similar approaches could not be found in the literature and could also not be recited by Professor Rheinberg.³ Besides, the diagnostic scheme was originally developed to explain an occurred phenomenon *ex post*, *looking back in time*, in a condition-analytical way. Instead, it is now applied *forward looking* in an attempt to make *ex ante* predictions with the help of a

¹As noted in the description of the experimental procedure of the study by Huang & Murad (2017, p. 45), subjects perception of the difficulty of a task is assessed with a single item only. However, the seven-point scale contains no verbal anchors or other instructions on how to interpret and use it. Also, the sample on which the findings reported by the authors are based is much smaller than in the present study (18 vs. 248).

²See also Section 2.2 and Section 2.4.4 for a discussion of learning effects in tasks.

³Personal communication, Oct. 2020.

machine-learning approach.⁴ This means that on the basis of the observed behavior of a majority of the study participants, the actions of the remaining subjects are inferred.

This chapter is structured as follows. First, the research design and methodology employed are presented: Section 4.2.1 describes the experimental design, reviewing the selection of real-effort tasks and introducing the questionnaires used to characterize the study participants in terms of their qualities (skills & personality) and motivation; thereafter, the empirical strategy is presented in Section 4.2.4. Subsequently, the results are detailed in Section 4.3, followed by concluding remarks (4.4).

4.2 Research Design and Methodology

4.2.1 Experimental Design and Procedure

The study presented in Chapter 3 contains several additional elements to investigate the effort provision conditional on subject characteristics across a set of diverse tasks. Except for the administrative part, the experiment consists of three steps: The *first step* and *third step* serve to characterize the subjects, while the effort measurement is collected in the *second step*. In the *first step*, subjects complete various questionnaires to acquire detailed information about their qualities (personality traits and abilities) and motivations (motives, needs). To counter potential bias and spill-over effects, the order of these characterization measures is randomized.

In the *second step*, study participants complete the tasks. These are selected based upon the classification presented in Section 1.3 to show great heterogeneity and, therefore, differ largely in the amount and type of effort they demand (see also Section 4.2.2). Since the study is conceived as a within-subject design, all subjects perform all tasks. To counteract order effects and to mitigate spillover effects, their order is likewise randomized. After each task, subjects fill in the real-effort task survey to elicit their subjective perception of the task. Certain items of the survey serve, as part of the

⁴A related, but somewhat less laborious approach is implemented by Rheinberg (1989). Instead of examining “all subjects at once,” the author very elegantly analyzes the subjects one after another to make verifiable, model-based predictions for each individual case. Similar to the present work, Rheinberg (1989) characterized students with the help of a questionnaire. Using the information obtained, the author went through the extended version of Heckhausen’s *Advanced Cognitive Motivation Model* (see Chapter 2) for each student to make a prediction about their performance in a later examination.

approach based on motivational diagnostics, to assess the subjects' motivation for providing effort. For this purpose, in the *third step*, the subjects further complete a final questionnaire that asks about the subjects' motivation for completing the task.

With the intent to determine individual effort solely dependent on the incentive and the subject characteristics, the subjects' behavior is not additionally manipulated in any other way. However, the outside option included in the study was not available to a small share of the participants to evaluate the effect of this *instrument to curb intrinsic motivation* for the different tasks.⁵

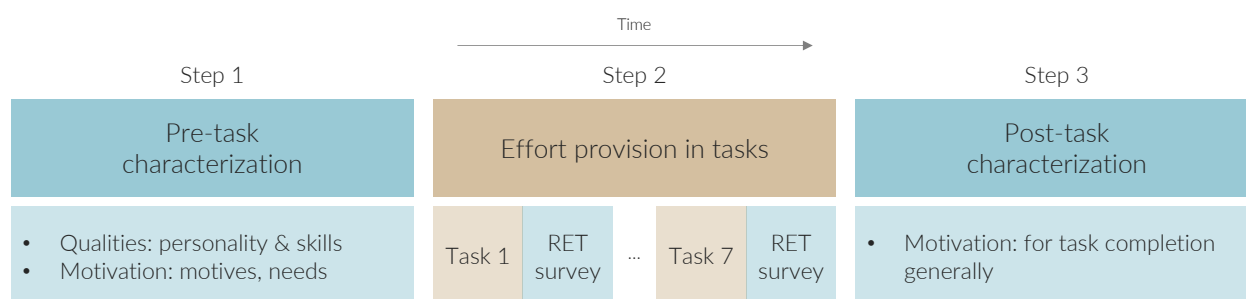


Figure 4.2: **Schematic representation of the experimental procedure:** Experimental contents relating to the characterization of the subjects are colored in petrol blue and regarding the effort provision in brown. The subjects' qualities and general motivations are assessed in *step 1*, which includes self-assessment surveys, questionnaires that indirectly elicit constructs, and simple ability tests. In *step 2* the subjects complete the set of tasks; their efforts are remunerated with a piece rate that varies according to the task design. After each task, subjects fill in the real-effort task survey to elicit their subjective perception of the task. In the concluding *step 3*, a self-assessment survey explicitly asks the subjects about their motivations for the effort they put into the tasks.

In the following the methodical approach of the experiment is described in detail. First, the step about effort provisioning is discussed in greater depth, with the diverse selection of tasks being shortly reviewed (4.2.2). Next, the approach chosen to determine the subject characteristics is described (B.1.1). It includes an overview of the qualities and motivations of the subjects that are considered crucial for their performance, and the scales by which they are measured. The experimental procedure was previously introduced in Section 3.3.1.2. A complete description, including the experiment's chronological sequence, can be found in Appendix B.1.1.

⁵See also Section 2.4 for a discussion of outside options.

4.2.2 Step 2: Effort Provision

To compile the set of seven real-effort tasks examined in the experiment, a large body of the experimental literature was surveyed, including both *computerized* and *non-computerized/manual* tasks. The classification presented in Section 1.3 proposes five distinct categories to which the reviewed tasks are subsequently assigned. One or two computer-based tasks are selected from each category to reflect the wide variety of tasks available to researchers. The set, therefore, shows a great diversity along with a multitude of dimensions. These include, among others, the demand in *physical skills and endurance, concentration, attention, quantitative and analytic reasoning, working memory, language skills, and verbalizing abilities*. Consequently, heterogeneity in effort provision – conditional on task type and subject characteristics – can systematically be identified. Figure 4.3 below reiterates the set of tasks employed and provides references for each task.^{6,7} The tasks are implemented in the laboratory software oTree (Chen et al., 2016).⁸

The effort provision part of the experiment contained several recurring elements that were well thought out and deliberately included in the experimental procedure. Before the study participants can begin with the first task, they have to complete a number of control questions to ensure that they have fully understood the instructions. Only after providing the correct answers they could proceed with the tasks to earn money. The following elements are present in each task in order to establish a consistent procedure and thereby minimize disturbing influences: *i) a trial round (15 sec) to familiarize*

⁶A comprehensive description is given in Section 3.3.1.1, containing more information on how the selection of tasks is made and how the tasks can be assigned to five different categories that reflect their main properties. Further details on the tasks are provided in the instructions for the study participants listed in Appendix B.1.4.4.

⁷Note that a considerable proportion of subjects knew some of the tasks due to previous participation in laboratory experiments. When explicitly asked at the end of the experiment, 86 subjects indicated that they were familiar with some of the experiment's content. In a comment field, they were able to provide further details of their prior knowledge on which the following estimates are based. 5% of the subjects were formerly confronted with the *multiplication task*; roughly three times as many knew of the *ab-typing task* (15%), most likely in a related study conducted in Hamburg in December 2017; only very few subjects had been exposed to the *word-encryption task* so far (2%). In contrast to, e.g., word puzzles or anagrams, which can be quickly retrieved once they are known, the mentioned tasks seem to be less problematic per se due to significantly lower training effects. Furthermore, basically none of the subjects had previously encountered the *word- and code-transcription task*, the *single-slider task*, or the *ball-catching task*. The subject, which stated that it had formerly completed a sliding task likely referred to the task from Gill & Prowse (2015), in which subjects have to position a larger number of sliders as accurately as possible at a fixed position (see Section 2.4 for a discussion of this task).

⁸Author contributions see Table B.1.

with the task; ii) the *option to re-read the task instructions* after the trial round to avoid and clarify misunderstandings; iii) a *countdown before the start* of the task to ensure that the subjects are focused and ready to begin; iv) a *long duration* of the task (5 min) to overcome transient effects of effort provision out of simple curiosity; v) an *alternative activity* to generate opportunity costs of effort provision;⁹ and vi) a *reminder screen* after completion of the task encouraged the study participants to take a short break to recover and regain focus.

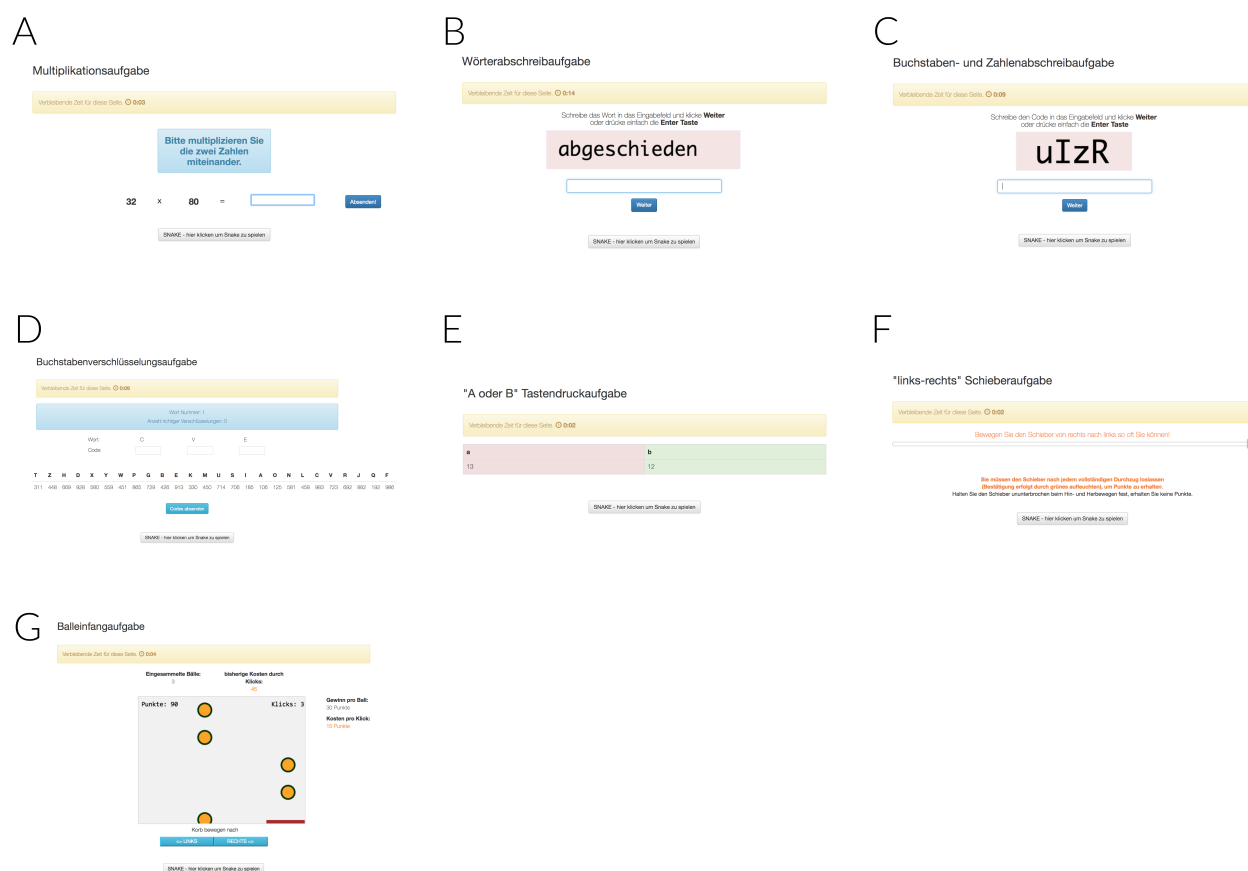


Figure 4.3: **Selection of real-effort tasks employed in the experiment:** A) multiplication task (Dohmen & Falk, 2011), B) word-transcription task (Waloszek modified from Kephart, 2017, unpublished), C) code-transcription task (Kephart, 2017), D) word-encryption task (Erkal et al., 2011), E) ab-typing task (Berger & Pope, 2011), F) single-slider task (Waloszek, mimeo), and G) ball-catching task (Gächter et al., 2016).

⁹An outside option was offered to 209 of the 248 study participants. Subjects could cease to earn points by switching to the game "Snake" whenever desired during the task duration. However, they could only collect points again in the next task, i.e., after the current task has elapsed. This is clearly communicated in the instructions for the task part and also verified in the control questions. The subjects are also given the opportunity to play the well-known computer game for a short time to alleviate curiosity about the outside option.

4.2.3 Steps 1 and 3: Characterization

The real-effort task classification presented in Chapter 1 argued that certain subject *qualities* may influence its performance in real-effort tasks. More precisely, tasks require a certain personality and skills to perform the task well. Moreover, non-monetary incentives may further motivate task performance (see Chapter 2 for further discussion). To further explore this, the chapter examines a wide range of tasks to establish which subject characteristics are most likely to be decisive for individual performance. A variety of characterization measures to assess subjects' *qualities* as well as their (broader) *motivations* to complete a task was sifted. The final selection of measures consisting of surveys and ability tests is discussed separately for both dimensions.

As noted earlier, the typification and systematization presented in this chapter do not claim to be completely comprehensive. Instead, it should provide a first starting point for future research.

The final set of characterization measures was implemented in the experimental laboratory software oTree to collect them as part of the computer-based experiment. For each of the measures, an analysis routine was programmed based on the scoring method used in the original publications, which are referred to in the two summary tables below.¹⁰

4.2.3.1 Step 1: Skills and Personality Traits

The previous section discussed the set of tasks examined in the experiment. The choice was made to achieve the greatest possible heterogeneity with regard to the task properties described in Table 1.1 of Section 1.3. For each of the tasks, their essential properties were identified.¹¹ A considerable part of the relevant literature from the fields of psychology and behavioral sciences was screened to find suitable surveys to investigate the tasks in this regard.¹² Of the compiled list of potential characterization measures, the promising ones were included in the study. In the following, the measures are

¹⁰The code for the laboratory software oTree as well as the analysis scripts for the free statistical software *R* are available to interested researchers upon request.

¹¹The classification presented in Section 1.3 discussed further subject characteristics that can be decisive for task performance, including *short reaction time and decision speed, spatial awareness as well as abstraction and association, and long-term memory* (see also Table 1.1). However, these are of minor relevance to the tasks included in the present task comparison.

¹²An extensive tabular synthesis summarizing the literature search is available on request.

described in more detail and preliminary considerations are made as to which tasks they are likely to be relevant. This information is summarized in Table 4.1.

Physical Abilities. The ab-typing task and single-slider task are classical mechanical tasks. Their demands on physical abilities (especially agility and endurance) can hardly be compared with those of the remaining tasks. To evaluate the typing skills and physical dexterity in the use of computers, the subjects were asked whether they were able to use touch typing.¹³ The study participants expressed their self-assessment on a four-point scale ranging from “not at all” to “very well.”

Prerequisites to Concentrate and to Maintain Focus. Prerequisites to concentrate and to maintain focus cut across mental abilities and personality traits. These include *grit*, *perseverance*, and *patience*.¹⁴ These are discussed more elaborately in the following.

Stamina and a *sustained interest* are required to perform an exhausting task for an extended period. However, the need for determination and stamina varies considerably across tasks and can be undermined in several ways. For example, *pauses within the task flow* can act as triggers for subjects to be seduced into thinking about the task’s sense. That appears to be the case for the multiplication task after solving each subtask. Likewise, a *monotonous task flow* can challenge the willpower of the study participants. Both mechanical tasks (ab-typing and single slider) are, by nature, very automatic and require little mental effort to perform. As a result, the wandering mind begins to question their purpose. The tasks are also much less “exciting” and “rousing” than the ball-catching task, which, despite its somewhat limited variety, captivates the subject far more. To assess the determination, persistence, and sustained interest of the subjects, the German version of the *Grit-Scale* of Breyer & Danner (2015) is used.¹⁵ The scale consists of nine items, which are phrased as statements. The subjects indicate their agreement with them on a response scale with five levels, ranging from (1) “not

¹³This is also known as the “10-finger typing system” and is referred to in German as “10-Finger-Tastschreiben” or “Maschinschreiben.”

¹⁴*Self-regulation*, *self-control* and *self-efficacy* are likewise of great relevance to maintain focus, which is why data was also collected for these constructs with appropriate surveys. Some of them are included in the approach based on *motivational diagnostics* presented below (Section 4.2.3.2). Others do not enter the empirical analysis of subjects’ qualities due to their conceptual overlap with the above-mentioned constructs in order to avoid collinearity of the considered measures.

¹⁵The original English version of the scale was developed by Duckworth et al. (2007).

at all” to (5) “to a very high extent.” Exemplary items include: 4.) “I can cope with setbacks,” 6.) “I am good at resisting temptation,” and 7.) “I finish whatever I begin.” The composite scoring is calculated as the average of all items, taking into account the negative coding of some of them. Higher scores are indicative of greater perseverance and passion for long-term goals. Comparing the design of the tasks in the selection, the single-slider task is particularly strenuous (see also Section 2.5). It is, therefore, expected that the grit score will contribute to the subjects’ performance in this task.

Patience and frustration tolerance are arguably key to mastering monotonous, tedious tasks. [Vischer et al. \(2013\)](#) propose an *Ultra-Short Survey Measure of Patience* (PATs), which consists of a single question only. It asks subjects to indicate their general level of impatience: “Are you generally an impatient person, or someone who always shows great patience?” Subjects’ responses are recorded on a scale with eleven levels ranging from (0) “very impatient” to (10) “very patient.”

Vigilance could eventually play a role in the ball-catching task to a certain degree. However, sound measurement methods for assessing the subjects’ alertness and attentiveness are rather elaborate to implement. For reasons of practicability, the study, therefore, does not include any measure to evaluate this subject quality.^{16,17} Furthermore, *physical condition* and *energy levels* can have a negative impact on and even be decisive for individual performance. However, they are relatively difficult to determine. For instance, subjective measures were included in the study to assess the subject’s alertness. As the measurements are not expected to provide precise information, they are not included among the main study variables.

¹⁶The *Dual-2-Back Task* described further below might serve as an indirect measure of vigilance. Although this “characterization task” is mainly aimed at testing subjects’ short-term memory, a worse performance in it might also indicate lower attention.

¹⁷Subjects were also asked to report their *state of mind* at various stages of the experiment using the PANAVA-KS scale from [Schallberger \(2005\)](#). The scale allows to assess subjects’ *positive activation* (PA), *negative activation* (NA), and *valence* (VA). The PA-subscale, which was measured immediately before the completion of the tasks, could be used as a measure for subjects’ *initial* alertness and attention. Future advanced analyses could include this subscale. A more differentiated treatment, however, would aim to study the alertness and attention of the subjects in a more timely manner. For example, one could apply the *Experiential Sampling Method* (ESM) introduced by [Csikszentmihalyi & Larson \(1987\)](#), where the measurement is taken *throughout the execution of the task* to capture the current state as accurately as possible. The subjects would be interrupted at regular intervals from the beginning to the end of the task to fill out the PANAVA-KS questionnaire. If the subjects would do this for all tasks considered, their experience of the tasks could be compared much further (see [Rheinberg et al. \(2003\)](#) for a discussion of the ESM and [Aellig \(2004\)](#) for an exemplary study on the application of the method to analyze flow experiencing during mountain climbing).

Cognitive Abilities. The *Subjective Numeracy Scale* (SNS) from [Fagerlin et al. \(2007\)](#) is employed as a measure of *quantitative reasoning*. The scale asks the study participants for their subjective assessment of their mathematical abilities and their passion for mathematical problems. The *ability subscale* with three items is considered as part of the analysis of subjects' qualities.¹⁸ Higher scores indicate a greater perceived numeracy.

Crystalline intelligence constitutes a part of general intelligence and refers to the entire knowledge that people acquire in the course of their lives and the ability to use it to solve problems ([Schipolowski et al., 2014](#)). It is evaluated using the *Short Scale of Crystalline Intelligence* (KKI) of [Schipolowski et al. \(2014\)](#), which comprises 12 items (Short Scale of Crystalline Intelligence). These cover declarative knowledge from the natural sciences, humanities and social sciences. Each item offers four possible answers, only one of which is correct.¹⁹ The score of the scale is obtained by adding the correct answers. It provides a measure of cognitive performance in the sense of the ability to access a broad spectrum of knowledge.

If a subject scores well in either the KKI or the SNS, they can be considered "cognitively competent" and, therefore, likely to be better at the mathematical task.²⁰ In addition, the KKI score is expected to be relevant to both cognitive tasks and memory tasks.

The *Dual-2-Back Task* introduced by [Jaeggi et al. \(2010\)](#) allows examining subjects' *short-term memory*. In this "characterization task," colored squares are displayed for a blink of a second at varying positions

¹⁸The three items of the *ability subscale* of the SNS are scored on a six-level response scale. Verbal anchors are placed only at the ends of the response scale: (1) "Not at all good" and (6) "Extremely good." Exemplary items include 1.) "How good are you at working with fractions?" and 3.) "How good are you at figuring out how much a shirt will cost if it is 25% off?"

¹⁹The abbreviation is derived from the original German name of the scale, *Kurzskala kristalline Intelligenz*. Example items include 3.) "What is amber made of?" a) volcanic magma, b) fossil resin (correct), c) silicates, or d) crystals, and 9.) "What is the 'Nibelungenlied'?" a) famous poem by Friedrich Schiller, b) Greek legend handed down from antiquity, c) national anthem of Switzerland, d) medieval heroic epic (correct).

²⁰*False positives* cause this screening method to identify a subject as competent even though the subject is actually not. It results in a systematic overestimation of the estimators. However, this case is extremely unlikely and, therefore, rather irrelevant. Conversely, *false negatives* cause the method to identify the subject as incompetent when the subject is, in fact, competent. This case leads to a systematic underestimation of the estimators, which does not harm the analysis and, therefore, does not pose a serious problem. For the SNS, the risk of false positives is slightly higher because the questionnaire asks the subjects for their self-assessment, and they do not have any incentive to respond truthfully. In the KKI, the subjects also have no incentive to answer the general knowledge questions honestly – but they must first know the correct answers.

and with varying colors.²¹ The subjects must memorize the color and position of the squares and confirm by pressing the corresponding key when either of them was the same as two steps back. Since the task is quite complex, the subjects had to answer several control questions before they were granted a trial round. Prior to starting the task, they were also given the opportunity to re-read the task instructions (about half of the subjects made use of this option). A performance metric named *Jaeggi-score* is finally calculated for both recorded dimensions, color and position.²² To facilitate the analysis, the *Jaeggi-scores* obtained for the color and the position stimulus were combined into a joint measure. Higher values of this holistic outcome measure imply a better working memory and a superior visual search ability.

In the *transcribe-codes task* and the *word-encryption task*, subjects have to transcribe codes consisting of numbers and letters with a length of less than seven characters. These tasks represent typical challenges for the working memory of the subjects. In contrast, the *transcribe-words task* constitutes more of a transcription task, since the German foreign words that the subjects had to transcribe were, on average, nine and a half letters long (see also Section 3.3.1.1).

An *anagram task*, as first employed by Ammons & Ammons (1959), served to gain insights into the *language proficiency* of the subjects. In this task, the subjects had to generate anagrams from the letters of a specified word. For this, they had to break down the word into its individual letters and rearrange them to form new words. As an example, possible anagrams for the word “pigeon” include: one, open, gin, pine, pig, pie, gone. The subjects had to find anagrams for six words, each consisting of six letters.²³ The words were of similar difficulty, i.e., the words were taken from everyday language and a similar number of anagrams could be generated from each word. For each of the words the subjects had a processing time of 30 seconds. To ensure that the subjects had fully understood this second “characterization task,” they again had to complete a trial round and were also able to re-read

²¹An illustration of the task together with its instructions is given in Appendix B.1.1 (see Figure B.9 to Figure B.12).

²²A *Jaeggi-score* is calculated separately for each of the dimensions as follows: $J_i = \frac{TP}{TP+FP+FN}$. The *true positives (TP)* represents the number of correct reactions to a given stimulus, the *false positives (FP)* the frequency of mistaken reactions, since no stimulus was actually present, and the *false negatives (FN)* the missing reactions to a given stimulus. Unlike common performance metrics such as precision, accuracy and recall, the *Jaeggi-score* does not contain *true negatives (TN)*. The reason for this is that it is not clear whether the respondent *intentionally* or *unintentionally* did not react to the absence of a stimulus.

²³A task description with more information about the words used in the experiment can be found in Appendix B.1.1 (see Figure B.7).

the task instructions before the start of the task (0.11% of subjects decided to do so). The number of correct anagrams found for each word were finally added up to obtain a measure of the subjects' language fluency and linguistic ability.

All personality and skill-related constructs covered in the study are summarized in Table 4.1, along with the corresponding characterization measures. For each task, it is indicated which of the constructs are regarded as crucial for the subjects' performance.

Table 4.1: **Skill and personality traits considered decisive for subject performance in the task selection:** For each task, the relevant constructs are specified, which form the basis for the empirical analysis.

Construct	Description	Multipli- cation	Transcribe Words	Transcribe Codes	Word Encryp.	ab- Typing	Single Slider	Catching Balls	Reference
Physical abilities									
Typing skills	Inquiry of the typing skills of the subjects: Ability to use 10-finger typing		low	low					Own measure
Prerequisites to concentrate and to maintain focus									
Grit	Examine stamina, persistence and sustained interest	mid				mid	high		Breyer & Danner (2015) Grit Scale
Patience	Serves as a general measure of patience and frustration tolerance	low				high	high		Vischer et al. (2013) Ultra Short Survey measure of patience
Cognitive abilities									
Quantitative reasoning	Inquire the subjective assessment of math skills and enthusiasm to solve math problems	high							Fagerlin et al. (2007) Subjective Numeracy Scale
Crystalline intelligence	Query the state of the general knowledge of the subjects	high	mid	mid	mid				Schipolowski et al. (2014) Kurzsкала kristalline Intelligenz
Short-term memory	Assess working memory capacities, i.e. subjects capacity to quickly grasp and store information and to recall it shortly after			high	high				Jaeggi et al. (2010) Dual-2-back task
Language proficiency	Evaluate subjects' language fluency and linguistic skills		high						Ammons & Ammons (1959) Standard Anagram Task

4.2.3.2 Step 3: Motivations

To analyze the motivation of the subjects in the experiment to make an effort in the tasks, a diagnostic scheme developed by Rheinberg (2004) is used. The scheme takes the extended version of Heckhausen's *Advanced Cognitive Motivation Model* presented in Chapter 2 as a starting point and allows to systematically dismantle and analyze a particular plot situation step by step. To demonstrate the direct connection between model and schema, the former is displayed again with the linkages clearly highlighted (see Figure 4.4 and Figure 4.5).²⁴ The scheme is usually applied to diagnose the prevailing form of motivation or any motivation problems on a case-by-case basis.²⁵ This is done by moving through the diagnostic scheme from top to bottom by answering each of the questions in turn. For the sake of simplicity, the scheme offers only two opposing response options in each step, even though, in reality, there is, of course, a continuum of gradations.

In order to fully capture and characterize the motivation (or lack of motivation), Rheinberg (2004) advises going through the entire diagnostic scheme since several forms of motivation or motivational problems may be present simultaneously, each of which promotes or impedes the considered behavior. In terms of these, Rheinberg (2004) distinguishes as forms of motivation: *self-initiative, spontaneous activity; externally controlled activity; self-controlled target activity; self-controlled target activity*. Conversely, the author identifies a *complete motivation deficit, incentive deficits, effectiveness deficit, volition deficit* as motivation problems. Each of them is described in more detail below when discussing the corresponding question of the scheme. Rheinberg (2011) discusses a variety of measures to assess each step of the scheme. In the following, both the scheme and the measures are presented.

²⁴Diagnosis scheme: Personal translation by the author, since an English version is yet unavailable in the literature.

²⁵For a more detailed description, see Rheinberg (2004), Rheinberg (2006), and Rheinberg (2011).

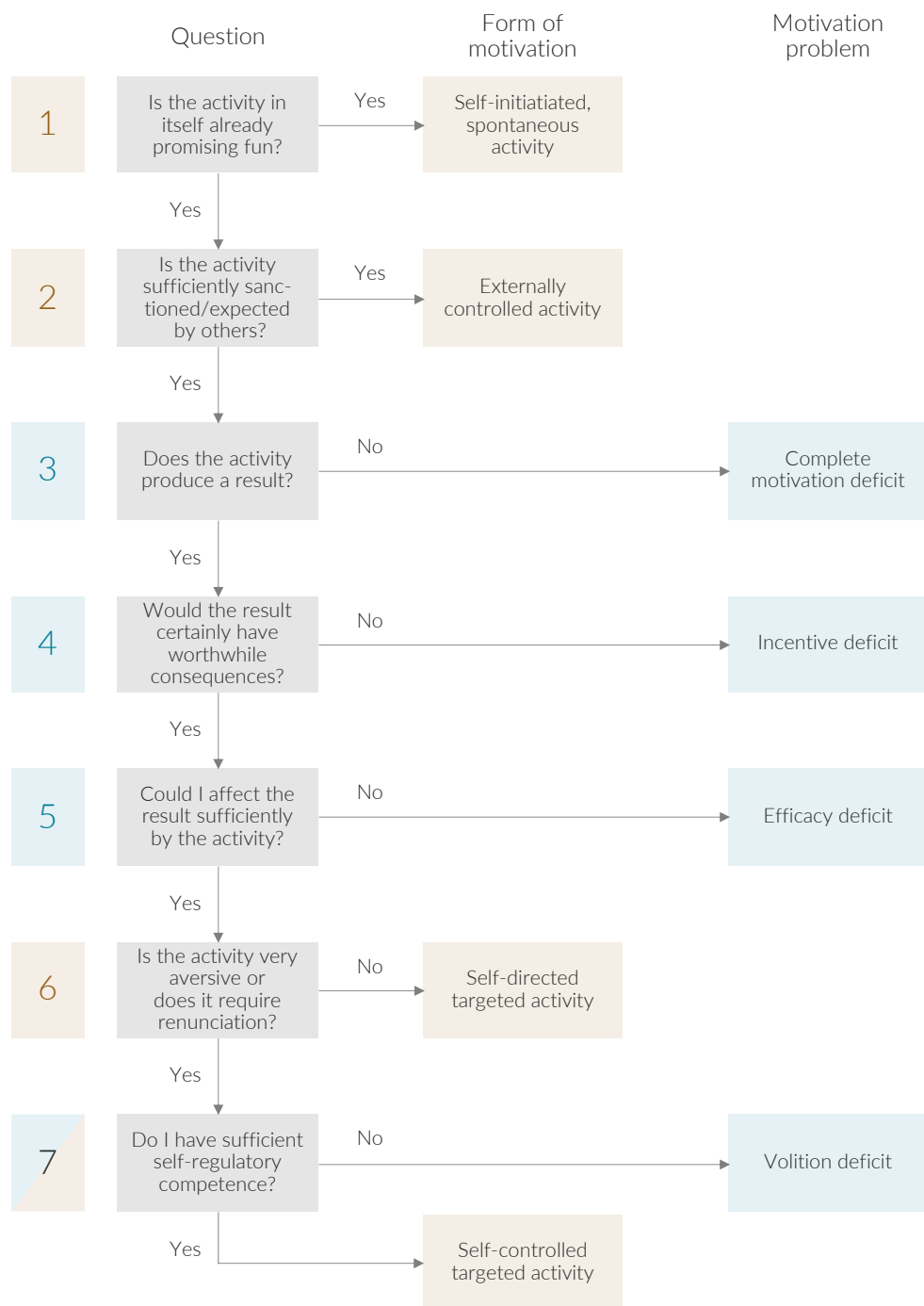


Figure 4.4: **Diagnosis scheme to capture forms of motivations and motivation problems** (adapted from Rheinberg, 2004, p. 24): The scheme serves as a blueprint to assess subjects' motivation in performing the tasks. The scheme builds on the extended version of Heckhausen's *Advanced Cognitive Motivation Model* presented in Section 2.3. To be able to refer to the model and to illustrate the relation between model and scheme, the model is again depicted in Figure 4.5. For the application of the scheme, all of its questions are addressed in sequence.

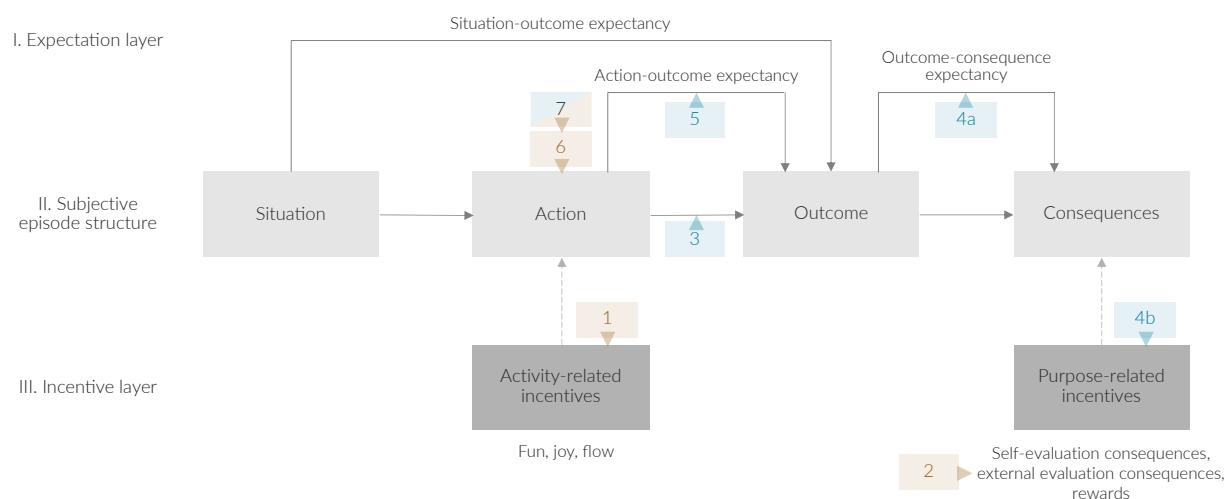


Figure 4.5: **Purpose- and activity-related incentives in the extended version of Heckhausen's Advanced Cognitive Motivation Model with links to Rheinberg's diagnosis scheme** (modified after Rheinberg, 1989, p. 104): The model was presented in Chapter 2.3 to examine why subjects provide effort and complete the tasks in real-effort experiments. It provides an analysis structure to specify the motivation in a given plot situation, in particular allowing to distinguish different forms and deficits of motivation. As explained in Section 2.3.1, the latter may trace back to one or more of three different types of expectations, or else can be due to inadequate or unfavorable incentives; unfavorable incentives may originate in the consequences, e.g., "it does not pay," and/or in the execution of the activity, i.e., "the execution is repugnant" (cf. Rheinberg & Engeser, 2018, pp. 593–594). To show the relation between both model (after Rheinberg, 1989, p. 104) and scheme (after Rheinberg, 2004, p. 24) their connection is indicated (motivation forms and motivation deficits).

Q1. Assessing activity-related incentives. To assess the incentives residing in the performance of an activity, the first item of the diagnosis scheme asks, *Is the activity in itself already promising fun?* This form of motivation is characteristic of self-initiated, spontaneous activity. The relative attractiveness of the tasks and relative task preferences are assessed with the *Personal Hit-List* by Rheinberg (2004). The scale serves as a measure of (relative) strength of activity-related incentives (see Section 3.4.5 for a description of the scale).²⁶

Section 2.3 discussed at greater length that individuals may feel an incentive i) from performing an activity or ii) from the purpose or consequences of an activity. The *Incentive-Focus Scale* (German: *Anreiz-Fokus Skala*, AFS) from Rheinberg et al. (1997) is employed to distinguish *task* and *purpose*

²⁶The response scale of the Personal Hit-List has nine levels, ranging from (1) "strongly disliked activity," via a (5) "neutral activity" to a (9) "strongly liked activity." Lower values imply greater preference for this activity, wherefore the activity is eventually carried out spontaneously and self-initiated.

motivation, i.e., focus on the activity or focus on its purpose. The scale contains two subscales for “activity centring” (TZ) and “purpose centring” (ZZ) with ten items each.²⁷ One score is obtained for each of the two subscales by summing the responses for all subscale items. A higher score in the respective scale corresponds to stronger “activity-centering” or “purpose-centering” of the subject.

Q2. Assessing externally controlled incentives. To assess whether the form of motivation arises from an externally controlled activity, the second question in the scheme asks *if the activity sufficiently sanctioned/expected by others?* After the subjects had completed all tasks, they had to complete a final motivational questionnaire, which concluded this part of the experiment. One of the items asked the subjects if they wanted to fulfill any expectations of the experimenter, e.g., the willingness to work. Subjects’ answers are scored on a response scale with seven levels, spread between the integer anchors (1) “do not agree at all” to (7) “agree fully.”²⁸

Q3. Diagnosis of desired outcomes. The third question of the diagnosis scheme asks *does the activity produce a result?* More elaborately, it inquires whether the activity leads towards a goal and, if so, whether this motivates the activity. If it does not, a complete motivation deficit may arise. The real-effort task survey presented in the previous chapter contains an item that asks subjects whether the task gave them a target or performance measurement that spurred them on. This survey item serves to evaluate whether this type of motivation problem prevails.

Q4. Assessing outcome dependent incentives. The fourth item in the diagnosis scheme poses the question of whether *the result would certainly have worthwhile/rewarding consequences.* The motivation problem examined here is whether there is a deficit of incentives. In order to assess whether the

²⁷The items are presented as contrasting pairs of statements (one from each subscale), which need not be mutually exclusive. Subjects respond to each item on a Likert scale with four levels ranging from (0) “not applicable at all” to (3) “applies exactly.” Exemplary statement pairs include (A: activity statement, B: purpose statement) A) When I decide on a task, I tend to focus more on whether or not I like the activities involved. B) When I decide on a task, I tend to focus on what results can be achieved and what consequences they may have. A) If I do not enjoy an activity in itself, then the potential results of this activity can hardly make me do it. B) If it is obvious that an activity does not contribute anything to achieving the desired results, then I will hardly ever perform it, even if in itself, it may be attractive. A) In case of doubt, my motto is: “Fun before benefit!” B) In case of doubt, my motto is: “Benefits before fun!” (translations by the author).

²⁸An illustration of the questionnaire with the instructions, the precise formulation of all items, and the response scheme is provided in Appendix B.1.1 (see Table B.7 and Figure B.47).

results obtained would entail rewarding consequences, the subjects were asked whether it was a motivation for them to earn a lot of money. In addition, they were asked whether they considered themselves to be hard-working people who wanted to show commitment. Both items were also part of the concluding survey after the tasks. They allow to assess the influence or role of extrinsic monetary incentives and whether subjects wanted to cultivate their self-image due to both peer-demand effects and self-demand effects.²⁹

To further assess whether the outcome would carry worthwhile consequences, the *Achievement Motives Scale* (AMS) is employed to assess the explicit motive relevant to the situation, i.e., the subjects' achievement motive. Personal ambitions may play a role in all tasks in the selection and may result in voluntary effort provision. The shortened *Achievement Motives Scale* of Engeser (2005) allows shedding light on the general motivation of the subjects to be successful. The scale contains two subscales, "hope for success" and "fear of failure," with five and six items, respectively.³⁰ The "net hope" score of the scale combines the results of both subscales and is intended to reflect personal ambition. Higher values imply confidence in success, while lower values indicate fear of failure.

Q5. Self-efficacy expectations. For motivation to translate into action, one needs to have faith that 1) a certain action will bring about the desired outcome and 2) in being able to perform that action (see *action-outcome expectancy* in the *Advanced Cognitive Motivation Model*). To address both of these dimensions, the third question of the diagnosis scheme asks, *Could I affect the result sufficiently by the activity?* The motivation problem which the fifth question in the scheme targets is an efficacy deficit.³¹

Locus of control of reinforcement is measured according to Rotter (1966) with the scale Internal-External

²⁹Responses were scored in the same format as reported above for question 2 of the scheme.

³⁰The subjects indicate their choices on a Likert scale with four levels, each of which has a verbal anchor ((1) "does not apply at all" to (4) "fully applies"). Exemplary items include for the subscale "hope for success" 4.) "I am appealed by situations allowing me to test my abilities" and for the subscale "fear of failure" 7.) "Even if nobody would notice my failure, I'm afraid of tasks, which I'm not able to solve."

³¹An alternative and, according to Prof. Jutta Heckhausen, a more clear translation of the term efficacy deficit would be "control deficit" (Private communication with Prof. Jutta Heckhausen and Prof. Falko Rheinberg, October 2020). However, since the chapter also deals with related but distinctly different concepts such as self-control and locus of control, the rather different sounding and written term is used here to emphasize the distinction.

Locus of Control by Kovaleva et al. (2014).³² Internal locus of control (ILC) refers to the extent to which an individual is convinced that she can control events and perceives them as a consequence of her own behavior; external locus of control (ELC), on the other hand, refers to the extent to which she regards events as fate, chance or under the control of others over whom she has virtually no influence Kovaleva et al. (2014).

The authors continue and state that “generalized control belief is a lasting, cross-contextual expectation that is linked to one’s self-image, knowledge of the world and the sum of all learning experiences and thus has a superordinate function for goal-oriented action” (p.3, English translation by the author). The scale contains four items, two for each dimension of the scale.³³ For each subscale, ILC and ELC, a score is obtained by averaging the responses to the survey items. If a subject has a high score on the ILC subscale, internal locus of control is given, and she believes that her own behavior determines the progression of events; if she has a low value for the ELC subscale, external locus of control is present, and she feels that her own actions have little or no influence on the course of her life (Kovaleva et al., 2014).

Q6. Aversiveness of the activity. The sixth question of the scheme asks whether the *activity is very aversive or does it require renunciation/sacrifice?* A self-directed targeted activity may prevail as motivation form. To judge whether performing the tasks involved any aversion or sacrifice, the final questionnaire asked subjects if they had more fun playing the game Snake than completing the tasks.³⁴

Q7. Assessing self-regulation and volition. It was pointed out in Section 2.3 that flow is only possible when the goal of the action is clear, and the entire attention is focused on the execution of the activity (and thus indirectly on the achievement of the goal). To maintain this state of maximum concentration and absorption in the activity, it is necessary to suppress all distractions. For this, a high measure of action control is necessary. The question therefore arises: *Do I have sufficient self-regulatory competence?* The motivation form examined here is that of a self-controlled targeted

³²The German designation of the scale is “Intern-Extern-Kontrollüberzeugung-4 (IE-4)” (English translation by the author).

³³Exemplary items include for the subscale *internal locus of control* 2.) “If I work hard, I will succeed.” and for the subscale *external locus of control* 4.) “Fate often gets in the way of my plans.” Responses are recorded on a Likert-scale with five response anchors, ranging from (1) “doesn’t apply at all” to (5) “applies completely.”

³⁴Responses were scored in the format described previously for question 2.

activity.

As a measure of self-regulation ability, the *Action Control Scale* (ACS-90) explores whether subjects, after failure, tend to be trapped in the negative situation (state orientation) or manage to direct their thoughts and plans towards actions to overcome it (action orientation) (Kuhl, 1990). The survey thereby serves to identify individuals who are susceptible to impairments due to state orientation. In the present case, these study participants may not necessarily have a deficit in motivation to complete a task but lack the (conscious) volition to perform it. To address the specific circumstances, only the subscale “*Performance-related action orientation vs. volatility*” (AOP) of the ACS is considered.³⁵ It examines to what extent someone is “absorbed” by an activity without attention being distracted from its execution, e.g., because too much thought is given to the goal to be achieved or possible alternatives (Kuhl, 1990).³⁶ According to the authors, a low AOP score indicates that the person switches to other activities prematurely and has a tendency to “actionism”; on the other hand, a high AOP score reflects that the person is very much involved in activities that he or she finds interesting. The AOP score, therefore, serves as a measure of (intrinsic) activity centering. According to Kuhl (1990), “activity-centred people are more optimistic, motivated, ambiguity-tolerant and efficient in solving complex problem situations (Kuhl, J. & Wassiljew, I. 1985).”³⁷

The complete set of motivational dimensions considered decisive for all real-effort tasks is reported in Table 4.2.

³⁵The original German name of the scale is *Handlungskontrolle nach Erfolg, Mißerfolg und prospektiv* (HAKEMP) from which only the subscale *Handlungsorientierung bei (erfolgreicher) Tätigkeitsausführung* (HOT: Tätigkeitszentrierung) is considered.

³⁶Each of the twelve survey items describes a situation, and the subjects must choose one of two response alternatives, one action-oriented (AO) and one state-oriented (SO). To avoid positional effects, the AO and SO items are distributed equally to the first and second response alternatives. Exemplary items are: *When I'm working on something that's important to me: A) I still like to do other things in between working on it or B) I get into it so much th I can work on it for a long time; When I am busy working on an interesting project: A) I need to take frequent breaks and work on other projects or B) I can keep working on the same project for a long time.* In both items, choice B) corresponds to performance-related action orientation. To calculate the AOP-score of a subject, the frequency a subject chose the action-oriented response alternative is calculated, such that scores range between zero and twelve. According to Kuhl (1990), study participants can be classified as state-oriented for AOP-scores between zero to nine and action-orientated between ten and twelve.

³⁷Translation by the author.

Table 4.2: **Motivational dimensions regarded as crucial for subject performance in the task selection:** For each construct, the data collection measure used is listed along with its reference. The listed constructs form the basis of the empirical analysis for all tasks.

Construct	Description	Reference
Q1. Capturing activity-related incentives		
(relative) strength of activity incentives	Subjective evaluation how attractive or aversive an activity is; evaluation takes place relative to other activities in an individually specified reference frame.	Rheinberg (2004) Personal-hit list (PHL)
Task vs. purpose motivation	Assess whether the subjects are motivated by the execution of the activity or the consequences/purpose of the activity.	Rheinberg et al. (1997) Incentive-Focus Scale (IFS)
Q2. Capturing externally controlled incentives		
Experimenter demand effects	Inquire whether subjects wanted to meet any expectations of the experimenter, e.g. commitment.	Final Motivational Survey: Item 3 (own measure, inspired by Wild et al. (1995))
Q3. Diagnosis of desired outcomes		
Task experience: qualitative assessment of task attractiveness	Ask the subjects whether the task gave them a goal/performance measurement that spurred them on.	Real-effort task survey: Item 2 (see Chapter 3)
Q4. Capturing outcome dependent incentives		
Outcome dependent incentives	To evaluate whether the results achieved would have worthwhile consequences, the subjects were asked if earning a lot of money was a motivation for them and if they considered themselves to be hard-working people who wanted to show commitment.	Final Motivational Survey: Items 2 and 4 (own measure, see ref. above)
Achievement motive	Examine the subjects' general motivation to succeed. The questionnaire records the hope component (HE) and the fear component (FM) of the achievement motive, the difference of which, "net hope", is included in the analysis.	Engeser (2005) Achievement Motives Scale (AMS, shortened version)
Q5. Self-efficacy expectations		
Locus of control	Assess whether the subjects tend to be more inclined towards an internal or external locus of control of reinforcement, i.e., if they feel that their actions have any influence on the course of their lives.	Kovaleva et al. (2014) Internal-External Locus of Control (IE-4)
Q6. Aversiveness of the activity		
Aversiveness and renunciation	Ask the subjects whether they enjoyed the game Snake more than completing the tasks.	Final Motivational Survey: Item 6 (own measure, see ref. above)
Q7. Capturing self-regulation and volition		
Self-regulation: Action-versus state-orientation	Record the self-regulation ability of the subjects, and thus their frustration tolerance, distractability, and capacity for initiative.	Kuhl (1990) HAKEMP (ACS-24)

4.2.4 Empirical Strategy

The questionnaires in part 1 of the experiment on the qualities of the subjects and subsequently the questions in part 3 on the general motives for performing the tasks determine the basic disposition of a subject, i.e., how does someone think and act in general. The empirical analysis examines how these general tendencies (in terms of abilities and personality traits as well as general motivations) affect the performance of the study participants. The main study variables are 1) the (individual) level of effort provided in each of the seven tasks, 2) the subject's characteristics (as elicited through a set of characterization surveys), and 3) the (individual) responses to the final motivational survey. As commonly assumed, the exerted effort is approximated by performance, i.e., produced output.

A description of the sample was provided in Section 3.3.1.2 in Chapter 3. Subjects demographics are further summarized in Table C.2.

In the following, the steps taken to prepare the data for analysis are described as well as the two analysis approaches based on linear regression modeling (4.2.4.2) and machine learning (4.2.4.3).

4.2.4.1 Data Preparation

In terms of subjects' performance in the tasks, *z-scores* served to detect outliers in the data. The scores of three subjects in the single-slider task are far above those of the other study participants. Since they collected up to five times as much as the average participant, fraudulent behavior cannot be ruled out. In the remaining tasks, outliers in the lower range are observed, i.e., subjects who scored no or just very few points. It is unclear why these subjects scored rather poorly, e.g., i) whether they could not perform the task well because they lacked the necessary skills, ii) whether they did not like to make an effort because they did not enjoy the task, or iii) whether they did not want to make an effort for any other reason. Nevertheless, these observations constitute valid data points, such that they are not excluded from the data set.

4.2.4.2 Regression Analysis

For each task, Table 4.1 lists the set of subject qualities that are expected to contribute to their performance with the corresponding survey measures to evaluate them.³⁸ Similarly, Table 4.2 lists constructs and survey measures for the motivational characterization of subjects based on motivation diagnostics. Both of these sets of survey measures serve as the basis for the empirical analysis. In order not to run into endogeneity problems, they are considered in separate regressions at first. In each case, subjects' performance is regressed on all respective measures separately for each task. At the end, the survey measures to assess the subjects' qualities and motivation are considered jointly in order to obtain a comprehensive picture.

Certain subjects were not able to fill in some of the characterization surveys in time. These incomplete observations are excluded from the regression analysis.³⁹ As for question 6 in the diagnostic scheme only data are available for those subjects who had the possibility to switch to the alternative activity Snake, the data analysis contains only their observations.

In Chapter 3, it was argued that the tasks included in the set could be grouped according to their properties. The resulting grouping structure was subsequently also confirmed by the data.⁴⁰ This grouping of tasks serves as useful in the empirical analysis.

The subjects' scores are min-max normalized to allow for comparisons of the estimates between the tasks. Yet, the characterization variables are not rescaled and remain in their respective units. The reason for this is that the scale and scoring methods of the individual characterization measures differ significantly. The number of items varies greatly (one single item for the *Ultra-Short Survey Measure*

³⁸Only a subset of HAKEMP survey items were intended to be elicited in the study (those for the HOP and HOT subscales). During the empirical analysis, it became prevalent that the order of items of the employed German version of the survey did not match the order of items in the original English publication. The items of the HOP and HOT subscales were selected with the original ordering of items in mind, which differed from the actual ordering. Consequently, only eight instead of twelve items were elicited from each of the two subscales (while eight items of the HOM subscale were elicited unintentionally).

³⁹These observations pose less of a problem for the machine-learning approach since missing values can be substituted via *knn*-imputation, as discussed below.

⁴⁰According to subjects' responses to the real-effort task survey, the tasks require substantially different amounts of physical and mental exertion; furthermore, study participants perceived the task as more or less motivating (see the findings of the previous Chapter 3.4 for more details).

of *Patience* (PAT), 36 items distributed across three subscales for the *Action Control Scale* (ACS)), as does the structure of the response scale (two opposing statements at the ACS survey, 11 levels for the PAT) and scoring schemes (summing or averaging).⁴¹ Normalization of each individual scale could give the impression that they are all based on the same underlying framework. A consideration with reference to the scale structure and evaluation schema used in each case appears more appropriate. Notwithstanding this, the effects of changing a regressor by one unit can be assessed well across the different tasks, which is anyway more informative. For any given task, the magnitude of the coefficient specifies by how much the mean subject score changes given a one-unit shift in one of the characterization variables, while the other variables in the model are held constant.

4.2.4.3 Machine Learning Approach

The previous analysis aims to explain the subjects' performance by a set of characteristics based on a linear regression approach. Supplementarily supervised machine learning is applied to allow for predictions about (unseen) observations based on previous information. More precisely, it means predicting the performance of certain subjects based on their qualities and motivations using an algorithm that is trained with the observations of other subjects, i.e., their respective qualities, motivations, and performances. As before, the scores obtained by the subjects serve as outcome variables and the same set of characteristics as input variables. Based on a set of training data, the machine learning algorithm is trained, i.e., a regression function is estimated. This function is then used to predict the outcome, i.e., subjects' scores, on a set of test data. These are unseen observations to the algorithm developed, which permits for *generalization*. The performance of the derived regression model is then assessed by comparing the actual with the predicted subject scores. This procedure is applied separately for each task and for several common modeling techniques, to derive an approach that predicts unseen data the best. Figure C.16 in the Appendix summarizes the machine learning approach applied, which is described in more detail below.

For the training and testing of the prediction model, the “hold out” technique is used. In the first step, the data are split into a *training set* and a *test set*. Thereafter, the training set is further divided

⁴¹Consider, for example, the Incentive-Focus Scale (AFS), which contains two subscales with ten items each. These are scored on a Likert scale with four levels ranging from “not applicable at all” (0) to “applies exactly” (3). A score is obtained for each subscale by summing the subjects' responses of the individual subscale items.

using *k-fold cross-validation* to develop and train a set of different models. The *modeling approaches* included ordinary least squares (OLS) and its penalized versions (lasso regression, ridge regression, and elastic-net regression), as well as random forests (RF), partial-least squares (PLS), *k*-nearest neighbors (KNN), support-vector machine (SVM) and generalized-boosted regression modeling (GBM). For each of the approaches, the following procedure is applied: Across five repeats of cross-validation folds, the model's tuning parameters are dissected.⁴² Next, the different parameter constellations are compared using the commonly employed performance metric *root-mean-squared error* (RMSE). The best performing combination of tuning parameters is selected as the final model (for the respective modeling approach). This model is then trained on the complete training set, i.e., on all observations contained in the training set). The final regression model is then confronted with unseen observations of the test set to verify whether it generalizes well to new data. Its performance is assessed by comparing the RMSE for the training set to the RMSE of the test set (each comparing actual values with predicted outcomes). The procedure thus provides a final model for each different modeling approach along with a performance evaluation. This set of final models is then compared in terms of predictive accuracy to assess which regression model generalizes best to new data.

As noted earlier, some subjects could not complete a specific characterization survey in time, which resulted in incomplete observations. To yet be able to include these observations in the machine-learning-based analysis, *k-nearest neighbor imputation* (knn-imputation) is used. Using the Euclidean distance as the distance metric, *k* close by neighboring points of the missing value are identified (excluding any other missing values). Since the incomplete observation's actual value is likely similar to those of its neighboring observations, their mean value provides a decent estimate. It is inserted instead of the missing value in the data set. Consequently, all observations can be included in the analysis.

⁴²OLS does not contain any tuning parameters. Yet, the estimated regression functions and the value of the performance metric vary across different cross-validation folds as the observations contained in each train and validate set differ across folds.

4.3 Experimental Results

4.3.1 Descriptive Analysis

All study participants completed all of the seven examined tasks and in a randomized order. Figure 4.6 presents the score distributions for each of them. The piece rates were calibrated based on a pilot trial to ensure that incentive effects and average payments were similar for all tasks. The scores are min-max normalized to allow for comparability between the tasks, accounting for any remaining variability due to differences in piece rates. The distribution of scores differs considerably from task to task. For the subjects who scored highest in the multiplication task (4th quartile) and lowest in the ab-typing task (1st quartile), their performance in the remaining tasks is also displayed as an overlaid scatter plot. Scoring well or poorly in one task does not necessarily imply performing similarly in other tasks.⁴³

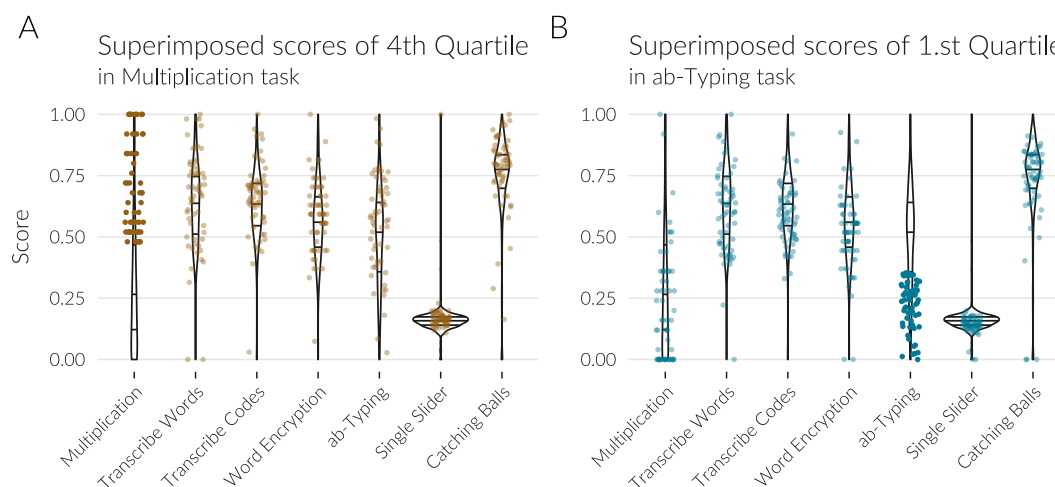


Figure 4.6: **Subjects' performance in and across the selection of tasks:** The violin plot summarizes the score distributions for the seven real-effort tasks. The performance of A) the best performing quartile in the multiplication and B) the least performing quartile in the ab-typing task are superimposed on the scoring distributions in the rest of the tasks. Subjects' performance in one task has little predictive value for their performance in another task.

To further explore this, Figure 4.7 presents the correlations of the subjects' performance across the different tasks. In general, only relatively weak correlations are observed. It demonstrates that the

⁴³See also Figure C.7 in the Appendix, in which the subjects are ranked according to their overall performance. It illustrates visually that the subjects perform very differently in individual tasks.

subjects perform rather unevenly in the tasks. Furthermore, it indicates that the tasks most likely measure “different types of effort.” Especially the multiplication task and the catching-balls task show particularly low correlation values and thus stand out. The remaining tasks show a gradual similarity. In particular, the two transcription tasks require similar performance, as is to be expected.

In addition, the score distributions are provided for each task on the diagonal of the figure. The distributions vary considerably: The low variance of the single-slider task is noticeable; in contrast, the ab-typing task and the multiplication task have a higher variance such that the scores are much more dispersed (see also Appendix C.2.1 for a detailed discussion). Researchers who want to use the multiplication task must be aware that the task already produces a distribution of scores that deviates strongly from a normal distribution independently of other additional influences (such as different incentives).

4.3.2 Regression Analysis

This chapter’s theme is that subjects’ qualities and motivations can influence task performance over and above incentive effects. The following figures aim to approach this visually. Figure 4.8 presents the scoring distributions for the set of tasks, superimposing subjects’ performance with A) high or low quantitative reasoning skills, and B) very good or very poor short-term memory. Study participants with higher abilities achieve higher scores than the average participant for the multiplication task. Conversely, subjects with a poor short-term memory perform worse than average in the transcription tasks. To provide a first idea of the influence of subjects’ motivation on their effort provision, Figure C.13 depicts subjects’ performance with, with the score distributions of subjects with A) a high or low achievement motive, and B) a desire or not to meet any experimenter expectations superimposed. In tasks where the average score of these subgroups differs from the average score of the total population, this factor appears to have an impact on subject motivation.

In summary, the figures suggest 1) that specific skills and personality traits may be advantageous for completing specific tasks, but are less supportive for others; and 2) that for some tasks, those subjects who perceive a task very well (or badly) also score accordingly good (or bad). For the tasks studied in the experiment, a number of features were identified that are expected to be relevant to the subjects’

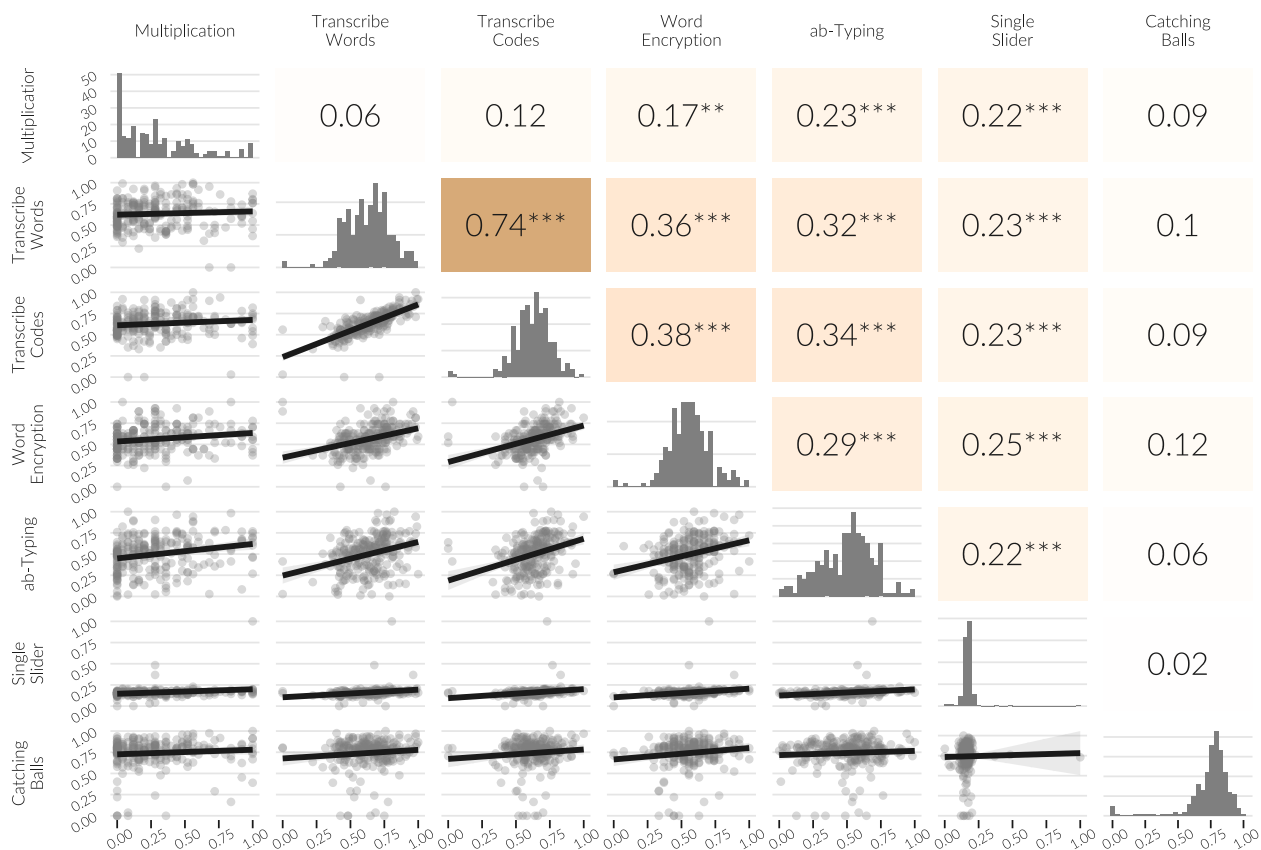


Figure 4.7: **Correlation between subjects' performance across tasks:** The tasks are arranged vertically and horizontally according to their task type (see also Table 3.4): (1) *mentally very demanding tasks* (multiplication task), (2) *moderately mentally demanding tasks* (transcription of words and codes tasks, word-encryption task), (3) *physically very demanding tasks* (ab-typing task and single-slider task), and (4) *neither mentally nor physically demanding tasks* (ball-catching task). Along the diagonal line, the score distribution of the study participants for each task is displayed. The upper triangle lists the correlation values with significance for each pairwise combination. The lower left triangle contains the corresponding correlation plots. In general, the correlation level is rather weak. The exception to this is the two transcription tasks, which, as might be expected, ask for similar skills.

performance (see Section B.1.1). These mainly cover two dimensions: subjects' qualities, i.e., skills and personality traits, and subjects' motivations. In a first step, both dimensions, along which the tasks vary, are examined separately for each task (in Section 4.3.2.1 for skills and personality traits and in Section 4.3.2.2 for motivations). As anticipated, the extent to which these characteristics play a role varies between tasks. Finally, both dimensions are considered jointly in Section 4.3.2.3. Several control variables are included in the analysis to take account of confounding factors.

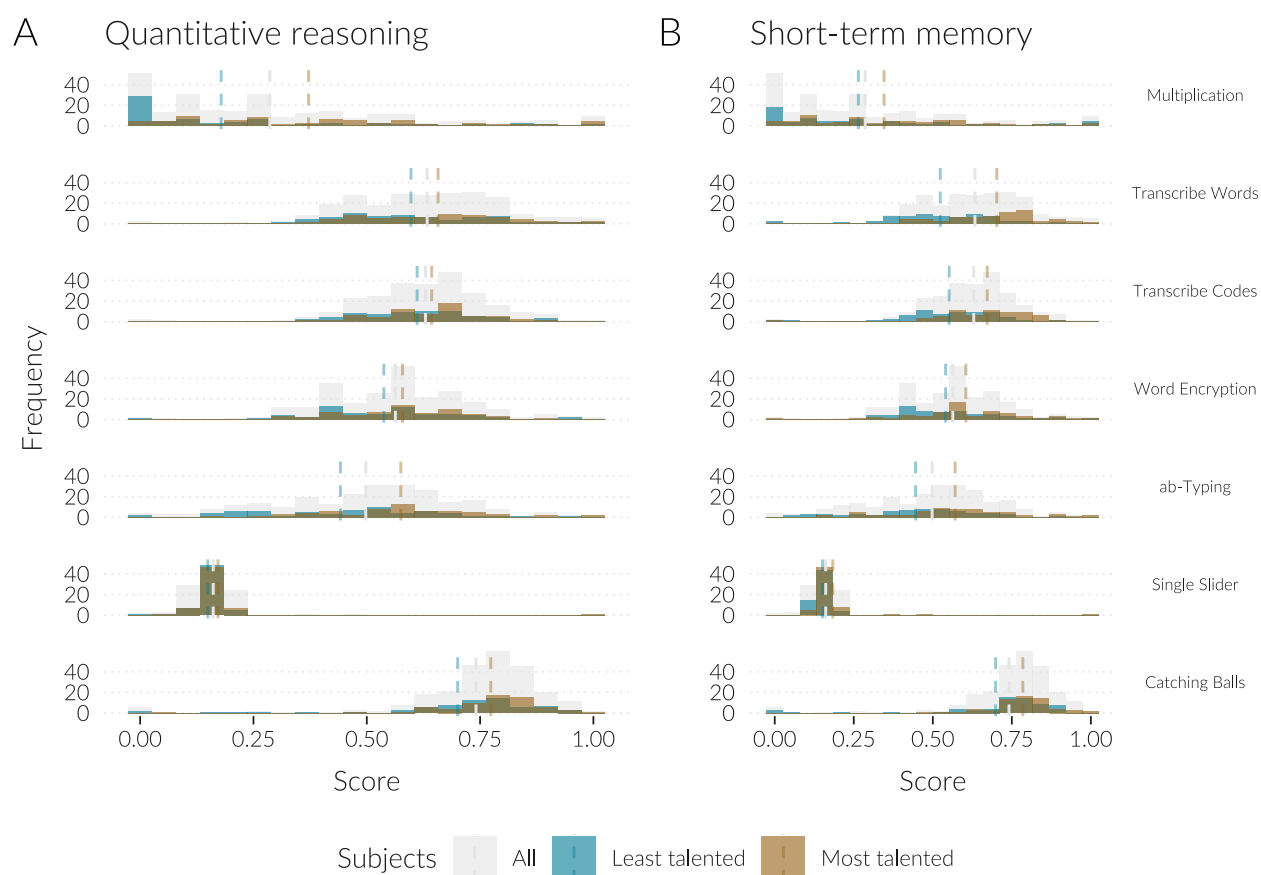


Figure 4.8: **Superimposed scores conditional on subjects' qualities:** Frequency distribution of scores for all tasks with those of the first and last quartile in terms of A) *quantitative reasoning* and B) *short-term memory* superimposed. A clear difference in the score distribution of the “worst” and the “best” from the overall distribution indicates that the ability is decisive for the respective task. Quantitative thinking, as one might expect, is most likely crucial for the multiplication task, but probably not for the transcription and memory tasks. In contrast, short-term memory is most likely essential for these.

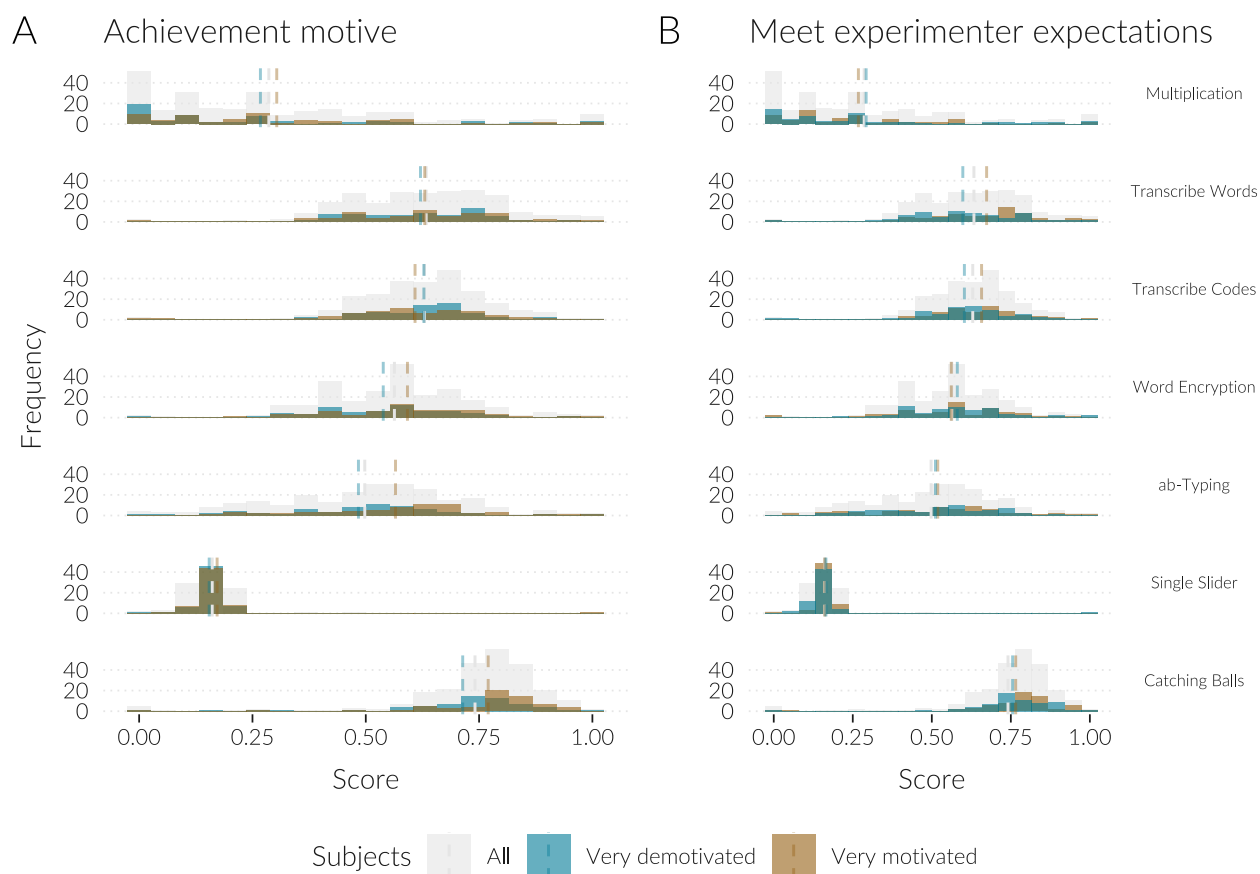


Figure 4.9: **Superimposed scores conditional on subjects' motivations:** Frequency distribution of scores for all tasks with those of the first and last quartile in terms of A) *task performance motivation* and B) *meeting experimenter expectations* superimposed. A strong difference in the score distribution of the “**very motivated**” and the “**least motivated**” of the overall distribution indicates that this particular motivational factor proves influential for the respective task.

4.3.2.1 Skills and Personality Traits

Table 4.1 listed the subject qualities considered relevant for the set of tasks, along with the surveys and ability tests employed to investigate them. The correlations between these characterization variables collected by these measures are rather weak (see Figure C.9 in the Appendix). The survey measures used to collect data on subjects' abilities and personalities are thus found to vary satisfactorily. Multicollinearity is, therefore, of secondary importance for the subsequent analysis. The linear regression analysis results in terms of all subject qualities are summarized in Table 4.3 for the set of

tasks.^{44,45} Several general observations can be made. *First*, only a subset of the explanatory variables eventually seems relevant for the subjects' performance. In other words, some of the variables seem to dominate the others, meaning that if anything is significant, it is always the same variables. These are, in particular, subjects' typing skills, their language proficiency and their short-term memory. *Second*, the grouping of tasks described in Chapter 3 can again be observed. Subjects' performance in the *mentally demanding tasks*, which include the multiplication task, the transcription of words and codes tasks, and the word-encryption task, is highly determined by cognitive skills. Conversely, in the remaining *non-mentally demanding tasks*, performance is not "cognitively determined." Both task groups are considered in closer detail in the following.

Finally, the share of subject performance that can be explained by the explanatory variables varies greatly between the tasks. Especially the "*cognitively determined tasks*" display a high adjusted R^2 ; it is lowest for the ball-catching task and the single-slider task.

The mentally demanding tasks are strongly skill-dependent, which is expressed by large coefficients that are highly significant. These include quantitative skills for the multiplication task, knowledge of touch typing for both transcription tasks and the encryption task, and language proficiency and short-term memory for the word-transcription task.^{46,47} For the word-transcription task, *short-term memory* was not included among the set of prior considerations. However, the findings clearly suggest that it does play a significant role in the task.

Concerning the non-mentally demanding tasks, only a few explanatory variables are found to be

⁴⁴The pre-considerations regarding advantageous abilities and personality traits summarized in Table 4.1 can be regarded as a prior in the regression analysis. Table C.7 in the Appendix compares them with the full set of characterization variables (with and without control variables) for a selection of three tasks (multiplication task, word-transcription task, and single-slider task).

⁴⁵The regressions included a set of control variables (Table C.4 explicitly states the estimates for the controls). A discussion of the control variables is provided in Appendix C.2.2.2.

⁴⁶Self-assessed knowledge of touch typing served as measure of typing skills and physical dexterity in the use of computers. The survey item offered four response levels ("Not at all" to "very well") and was coded as an ordered factor. The linear term of the corresponding categorical outcome variable is positive and significant for the word-transcription task, the code-transcription task, and the word-encryption task. The remaining terms (quadratic, cubic) are negligibly small and mostly non-significant. A positive effect of touch typing can, therefore, be observed.

⁴⁷Subjects' short-term memory is assessed with the Dual-2-Back Task. For the word-encryption task, it is found to be significant only at the 10%-significance level, yet with a relatively large coefficient. This strong influence on the subjects' performance comes as no surprise: The Dual-2-Back Task challenges subjects' memorizing skills, but also contains distinct elements of a visual search. Both features of the ability task are also found in the word-encryption task.

significant, wherefore the share of explained variation in performance is relatively low. As noted above, *touch typing* serves as a proxy for physical dexterity in the use of computers. It is, therefore, not surprising that touch typing is essential for the single-slider task, as it demands higher skills with the computer mouse. However, this does not apply to the ab-typing task, since it is very easy and straightforward to type the letters “a” and “b.”

Crystalline intelligence, which constitutes a part of general intelligence, has an adverse effect on performance in the ab-typing task. The finding suggests that the senselessness and mind-numbing nature of a task may actually inhibit the provision of effort.

For the single-slider task, the coefficient for grit is significant and noticeable in its size. However, contrary to the expectations expressed earlier, it is negative. Consequently, greater perseverance does not seem to be very helpful in coping well with this tedious, toilsome and tiring task. Since this is rather counterintuitive, the relationship between *grit*, perception of the task as *pointless* and *task performance* was further investigated. If an interaction effect between the former two is added, the coefficient for persistence is no longer significant, indicating that the effect is much more subtle than it first appears. Grouping subjects by their degree of perseverance reveals that, 1) subjects with a *low* level of grit perceive the task as “somewhat” pointless, but irrespectively put effort in the task and perform very well in the task, 2) subjects with *high* level of grit perceive the task similarly pointless and also perform rather well in the task, and 3) subjects with a *medium* level of grit perceives the task as “very” pointless and performs worst. Comparing the group with medium and high determination, the latter seems to be able to fade out the senselessness of the task and overcome any aversion to perform well and earn money. The subjects in the group with a medium level of grit seem not to be capable of this.

Table 4.3: Task performance conditional on all potential subjects' characteristics, including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.191 (0.221)	0.459*** (0.000)	0.604*** (0.000)	0.449*** (0.000)	0.472*** (0.000)	0.126*** (0.001)	0.374*** (0.000)
<i>Physical abilities</i>							
Touch typing	0.048 (0.246)	0.112*** (0.000)	0.116*** (0.000)	0.099*** (0.000)	0.022 (0.457)	0.006 (0.564)	-0.023 (0.390)
<i>Prerequisites to focus and to maintain concentration</i>							
Grit	-0.043 (0.150)	-0.010 (0.489)	-0.019 (0.152)	0.016 (0.337)	0.014 (0.520)	-0.006 (0.419)	0.021 (0.275)
Patience	-0.005 (0.446)	-0.006* (0.058)	-0.002 (0.464)	-0.004 (0.315)	0.000 (1.000)	-0.002 (0.376)	-0.003 (0.493)
<i>Cognitive abilities</i>							
Quantitative reasoning	0.063*** (0.000)	0.004 (0.497)	0.003 (0.569)	0.007 (0.382)	0.011 (0.222)	0.002 (0.556)	0.016* (0.060)
Common knowledge	-0.008 (0.315)	0.004 (0.278)	0.000 (0.959)	-0.007 (0.119)	-0.010* (0.092)	0.002 (0.308)	0.006 (0.272)
Short-term memory	0.023 (0.627)	0.073*** (0.002)	0.035 (0.103)	0.057** (0.039)	0.045 (0.190)	0.008 (0.504)	0.040 (0.201)
Language proficiency	0.000 (0.846)	0.006*** (0.000)	0.004*** (0.000)	0.003** (0.017)	0.004*** (0.005)	0.001** (0.011)	0.002 (0.160)
Num.Obs.	248	248	248	248	248	248	248
R2	0.205	0.473	0.364	0.207	0.271	0.119	0.114
R2 Adj.	0.154	0.438	0.323	0.156	0.224	0.062	0.057
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: p-values are not adjusted for multiple hypothesis testing.

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

4.3.2.2 Motivation

Table 4.2 lists the survey measures to assess the motivational characteristics of the subjects. Correlations between subjects' motivations are listed in the Appendix: Figure C.10 covers the general motivational survey items and Figure C.11 depicts the correlations for the items of the real-effort task survey exemplarily for the ball-catching task. The items of the final motivational survey show a moderate to large positive correlation. *Activity centring* is (as expected) negatively correlated with *purpose centering* as well as with the *performance motive*. The global motivations examined in the survey appear to be more interconnected, with a central element being the self-image of being a hardworking person and striving to show commitment. The remaining motivational survey items are rather weakly correlated.

The results of the linear regression analysis in terms of motivations according to Table 4.2 are presented in Table 4.4 (the controls included in the regression are discussed in Appendix C.2.2.2 with Table C.5 further listing the coefficients of the controls). The measures employed to assess the various questions of the diagnosis scheme employ different response schemes. These were implemented according to the original publications and vary in number of levels and poling. A direct comparison of the estimates is therefore not advised; the interpretation of the estimates also has to consider the original scale and its polarity. For an easier reading of the regression table, scales with a polarity contrary to the default are marked with an asterisk (*).

- **Q1. Capturing activity-related incentives:** Activity-related incentives appear to be at play in the cognitive tasks, but also in the ball-catching task. This is indicated by the highly significant and substantial coefficients for the *Personal Hit-List**, which is particularly noticeable for the multiplication task.
- **Q2. Capturing externally controlled incentives:** Experimenter demand effects are likely at play in both transcription tasks, for which a significant coefficient is observed.
- **Q3. Diagnosis of desired outcomes:** Subject who rate a task in the second item of the *real-effort task survey** to possess a target or performance measurement that spurred them on performed

better in the multiplication task and the ab-typing task. Interestingly, the relationship is reversed for the word-transcription task.

- **Q4. Capturing outcome dependent incentives:** Subjects who scored well on the multiplication task appear to be highly driven by monetary incentives.⁴⁸ Subjects who indicated that they had a hardworking attitude performed slightly better on the word-transcription task. Interestingly, cultivating a positive self-image has no effect on any of the other tasks. Higher confidence in success as assessed by the *Achievement Motives Scale* results in a better performance in the ball-catching task. Surprisingly, a negative coefficient is observed for the transcribe-words task for the measure.
- **Q5. Self-efficacy expectations:** The extent to which subjects feel capable of controlling events and perceives them as the consequence of their actions appears vital for a good performance in the multiplication task and the code-transcription task. Both tasks require skill: mathematical ability for the multiplication task; familiarity with the computer keyboard for the code-transcription task, so that the codes can be transferred without constantly shifting the gaze between the keyboard and the screen. Awareness of one's abilities seems to increase one's sense of control over one's actions and the situation.
- **Q6. Aversiveness of the activity:** Subjects who preferred the tasks to the outside option Snake performed better in the ball-catching task. This measure of whether the tasks were highly aversive did not predict performance for any of the other tasks.
- **Q7. Capturing self-regulation and volition:** Contrary to what one would expect, state-oriented subjects perform better on the multiplication task than action-oriented subjects, although the latter are known to handle difficult problem situations better.

⁴⁸Compared to the other tasks, however, this task is rather coarse-grained in its effort resolution: calculations take a relatively long time and are rewarded with a high piece-rate.

Table 4.4: Task performance conditional on all subjects' motivations, including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.387* (0.083)	0.644*** (0.000)	0.761*** (0.000)	0.582*** (0.000)	0.792*** (0.000)	0.246*** (0.001)	0.998*** (0.000)
<i>Q1. Capturing activity-related incentives</i>							
PHL	-0.043*** (0.000)	-0.028*** (0.000)	-0.012** (0.021)	-0.016** (0.011)	-0.001 (0.919)	-0.006** (0.014)	-0.013** (0.018)
Activity focussed	-0.006 (0.188)	0.002 (0.482)	-0.002 (0.400)	-0.004 (0.190)	-0.007* (0.072)	-0.001 (0.501)	-0.006* (0.075)
Purpose focussed	-0.007 (0.120)	0.001 (0.783)	-0.002 (0.558)	0.001 (0.791)	0.002 (0.637)	-0.002 (0.290)	-0.008** (0.018)
<i>Q2. Capturing externally controlled incentives</i>							
Meet experimenter expectations	0.001 (0.942)	-0.020*** (0.001)	-0.011** (0.048)	-0.007 (0.254)	0.001 (0.938)	0.000 (0.983)	-0.002 (0.730)
<i>Q3. Diagnosis of desired outcomes</i>							
no target	-0.020** (0.022)	0.016** (0.015)	-0.005 (0.369)	-0.003 (0.657)	-0.015** (0.048)	-0.006** (0.016)	0.005 (0.518)
<i>Q4. Capturing outcome dependent incentives</i>							
Earn a lot of money	0.027** (0.040)	-0.001 (0.870)	-0.009 (0.273)	0.011 (0.240)	0.001 (0.897)	0.006 (0.165)	0.009 (0.311)
Diligent attitude	0.005 (0.661)	0.013* (0.084)	-0.001 (0.829)	0.009 (0.277)	0.005 (0.554)	0.000 (0.938)	-0.001 (0.858)
Achievement motive	-0.025 (0.207)	-0.018 (0.167)	-0.024** (0.049)	-0.005 (0.704)	-0.002 (0.901)	-0.001 (0.864)	0.023* (0.098)
<i>Q5. Self-efficacy expectations</i>							
Internal locus of control	0.074*** (0.007)	0.021 (0.235)	0.036** (0.030)	0.012 (0.535)	0.006 (0.776)	-0.002 (0.823)	-0.018 (0.336)
External locus of control	0.014 (0.573)	-0.014 (0.378)	-0.011 (0.483)	0.008 (0.642)	-0.011 (0.618)	-0.012 (0.138)	-0.003 (0.855)
<i>Q6. Aversiveness of the activity</i>							
Snake more fun than tasks	0.006 (0.552)	0.001 (0.916)	-0.003 (0.656)	-0.003 (0.693)	-0.007 (0.410)	0.000 (0.886)	-0.022*** (0.002)
<i>Q7. Capturing self-regulation and volition</i>							
Action orientation (AOP)	-0.020** (0.025)	0.005 (0.403)	0.007 (0.181)	0.009 (0.150)	-0.003 (0.646)	0.001 (0.721)	0.005 (0.451)
Num.Obs.	205	205	205	205	205	205	205
R2	0.416	0.317	0.251	0.197	0.274	0.169	0.232
R2 Adj.	0.359	0.251	0.179	0.120	0.204	0.089	0.158
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: p-values are not adjusted for multiple hypothesis testing.

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

4.3.2.3 Examine the Subject's Characteristics and Motivation Jointly

Finally, both dimensions influencing subject performance, subjects' skills, and personality traits, and their motivations, were included in the linear regression analysis. The results are depicted in Table 4.5, again containing a set of control variables (Table C.6 in the Appendix includes the estimates for the controls).⁴⁹

For the multiplication task, certain subject qualities become significant that did not play a role in the simpler model: Touch typing becomes significant with a large estimate, and patience also appears to affect performance, with the coefficient only just significant. At the same time, the large coefficient for quantitative reasoning shrinks. The motivational measures found to be crucial for the multiplication task all remain significant with coefficients of comparable size.

For the word-transcription task, grit becomes significant. Again there is only limited change in the significance and the coefficients for the motivational measures. For the code-transcription task, meeting experimenter demand is no longer significant, while the remainder of the previously significant coefficients is yet significant. For the word-encryption task, short-term memory is surprisingly no longer significant when subjects' qualities and motivation are considered. As before, activity-related incentives as tracked by the Personal Hit-List remain highly significant with a large coefficient.

Already before, few predictors were significant for the ab-typing task and the single-slider task. For the ab-typing task, the coefficient for having a target remains at a comparable size; however, it is just yet significant.

For the ball-catching task, the coefficients for whether the task is perceived "as fun" and "more fun than the game snake" remain significant. Achievement motivation is no longer significant, as are quantitative reasoning skills.

The explanatory power of the model varies greatly between the different tasks: As an estimate of the amount of variance the model accounts for, the adjusted R^2 ranges from 0.165 for the single-slider task to 0.496 for the word-transcription task.

⁴⁹Note that moderate correlation is found for *grit*, i.e., the score of the *Grit Scale (BISS8)*, and *performance motives*, i.e., the "net hope" score of the *Achievement Motives Scale (AMS-NH)* ($r(242) = 0.51, p < 0.001$). The latter is a combination of its two sub-scales "hope for success" and "fear of failure" and aims to reflect personal ambition.

Table 4.5: Task performance conditional on all subjects' characteristics (skills, personality) and motivations, including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.398 (0.133)	0.481*** (0.001)	0.686*** (0.000)	0.524*** (0.004)	0.786*** (0.000)	0.221*** (0.005)	0.891*** (0.000)
<i>Physical abilities</i>							
Touch typing	0.082** (0.035)	0.095*** (0.000)	0.111*** (0.000)	0.089*** (0.001)	0.023 (0.467)	0.001 (0.928)	-0.014 (0.606)
<i>Prerequisites to focus and to maintain concentration</i>							
Grit	-0.026 (0.488)	-0.036* (0.079)	-0.022 (0.273)	-0.011 (0.653)	-0.029 (0.341)	-0.011 (0.324)	0.008 (0.746)
Patience	-0.012* (0.079)	-0.004 (0.266)	-0.001 (0.729)	-0.006 (0.172)	-0.001 (0.876)	-0.003 (0.140)	-0.007 (0.118)
<i>Cognitive abilities</i>							
Quantitative reasoning	0.029** (0.049)	0.007 (0.327)	0.006 (0.365)	0.009 (0.296)	0.008 (0.479)	0.006 (0.162)	0.012 (0.212)
Common knowledge	0.004 (0.588)	0.003 (0.528)	0.003 (0.402)	-0.006 (0.240)	-0.010 (0.110)	0.004 (0.112)	0.006 (0.290)
Short-term memory	-0.034 (0.460)	0.062** (0.015)	0.028 (0.263)	0.046 (0.133)	0.057 (0.139)	0.010 (0.472)	0.019 (0.573)
Language proficiency	0.001 (0.562)	0.007*** (0.000)	0.005*** (0.000)	0.004*** (0.002)	0.005*** (0.006)	0.002*** (0.006)	0.002 (0.153)
<i>Q1. Capturing activity-related incentives</i>							
PHL	-0.037*** (0.000)	-0.013** (0.012)	-0.012*** (0.007)	-0.013** (0.034)	-0.005 (0.418)	-0.008*** (0.001)	-0.014** (0.016)
Activity focussed	-0.008* (0.098)	0.001 (0.639)	-0.003 (0.186)	-0.005 (0.136)	-0.007* (0.090)	-0.001 (0.356)	-0.006 (0.104)
Purpose focussed	-0.009* (0.073)	0.002 (0.428)	-0.002 (0.412)	0.001 (0.810)	0.002 (0.539)	-0.002 (0.250)	-0.008** (0.015)
<i>Q2. Capturing externally controlled incentives</i>							
Meet experimenter expectations	0.006 (0.525)	-0.012** (0.021)	-0.004 (0.375)	-0.002 (0.711)	0.002 (0.772)	0.002 (0.451)	0.000 (0.939)
<i>Q3. Diagnosis of desired outcomes</i>							
no target	-0.021** (0.020)	0.009* (0.099)	-0.003 (0.496)	-0.004 (0.541)	-0.013* (0.084)	-0.006** (0.018)	0.003 (0.636)
<i>Q4. Capturing outcome dependent incentives</i>							
Earn a lot of money	0.024* (0.063)	-0.006 (0.369)	-0.011 (0.120)	0.008 (0.340)	-0.002 (0.821)	0.003 (0.501)	0.004 (0.656)
Diligent attitude	0.005 (0.678)	0.016*** (0.009)	0.000 (0.955)	0.008 (0.282)	0.005 (0.556)	0.000 (0.967)	-0.003 (0.699)
Achievement motive	-0.027 (0.200)	-0.016 (0.176)	-0.024** (0.033)	-0.001 (0.922)	0.009 (0.613)	0.001 (0.843)	0.020 (0.174)
<i>Q5. Self-efficacy expectations</i>							
Internal locus of control	0.080*** (0.004)	0.014 (0.344)	0.031** (0.034)	0.005 (0.776)	-0.001 (0.968)	0.000 (0.964)	-0.016 (0.423)
External locus of control	0.006 (0.820)	-0.011 (0.407)	-0.010 (0.455)	0.009 (0.604)	-0.011 (0.620)	-0.008 (0.289)	0.002 (0.919)
<i>Q6. Aversiveness of the activity</i>							
Snake more fun than tasks	0.007 (0.544)	0.002 (0.676)	0.000 (0.935)	-0.003 (0.700)	-0.006 (0.465)	0.002 (0.642)	-0.021*** (0.006)
<i>Q7. Capturing self-regulation and volition</i>							
Action orientation (AOP)	-0.020** (0.029)	0.005 (0.355)	0.006 (0.232)	0.007 (0.246)	-0.005 (0.507)	0.001 (0.770)	0.003 (0.631)
Num.Obs.	205	205	205	205	205	205	205
R2	0.462	0.562	0.483	0.343	0.347	0.275	0.278
R2 Adj.	0.380	0.496	0.405	0.242	0.247	0.165	0.168
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: p-values are not adjusted for multiple hypothesis testing.

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

4.3.3 Additional Findings based on Subjects' Subjective Performance Assessment

In the following, the notion that tasks can have (de)motivating effects on study participants is further substantiated. First of all, the goal-setting in tasks is discussed based on reference points that are oriented towards previous experiences. Subsequently, tasks are examined that discourage at the mere sight of them and thus have a deterrent and exertion-inhibiting effect. In the course of this, the subjects' switching behavior to the outside option *Snake* is briefly addressed (see also Appendix C.2.3.1).

After each task, subjects had to indicate whether they were satisfied with their performance on a scale ranging from (1) "not at all" to (7) "completely." Figure 4.10 presents the score distributions for all tasks with the scores for the subjects who indicated a very high and a very low satisfaction with their performance superimposed. The subjective assessment of their performance corresponds rather closely to the subject's actual performance (see also Figure C.12, which reports Pearson-correlations for subjects' scores and their subjective performance assessments in each task in the last column). This applies especially to the cognitively highly demanding tasks, but also to the two memory tasks and the ball-catching task. Considering that the subjects have no *specific* reference point against which to compare their performance, this result is both surprising and remarkable. However, previous experience with similar activities inside and outside the laboratory could guide expectations of their performance and thus serve as a rough reference. In contrast, no or only a very weak correlation between the physically demanding tasks and the subjects' subjective performance assessment is observed. This could point to the extraordinary nature of these tasks and the resulting lack of any notion of a reference point whatsoever. As a result subjects cannot set themselves (precise) goals, such that self-evaluation consequences can no longer give rise to purpose-related incentives and induce voluntary effort.

As far as the multiplication task is concerned, the subjects can most certainly draw on previous experience with similar tasks from school. It seems reasonable to assume that the subjects have some sort of reference point for such tasks to compare their performance with. The observed strong correlation between the subjective performance assessment of the subjects and their actual performance supports this assumption ($r(246) = 0.67, p < 0.001$; this value is by far higher than the correlation val-

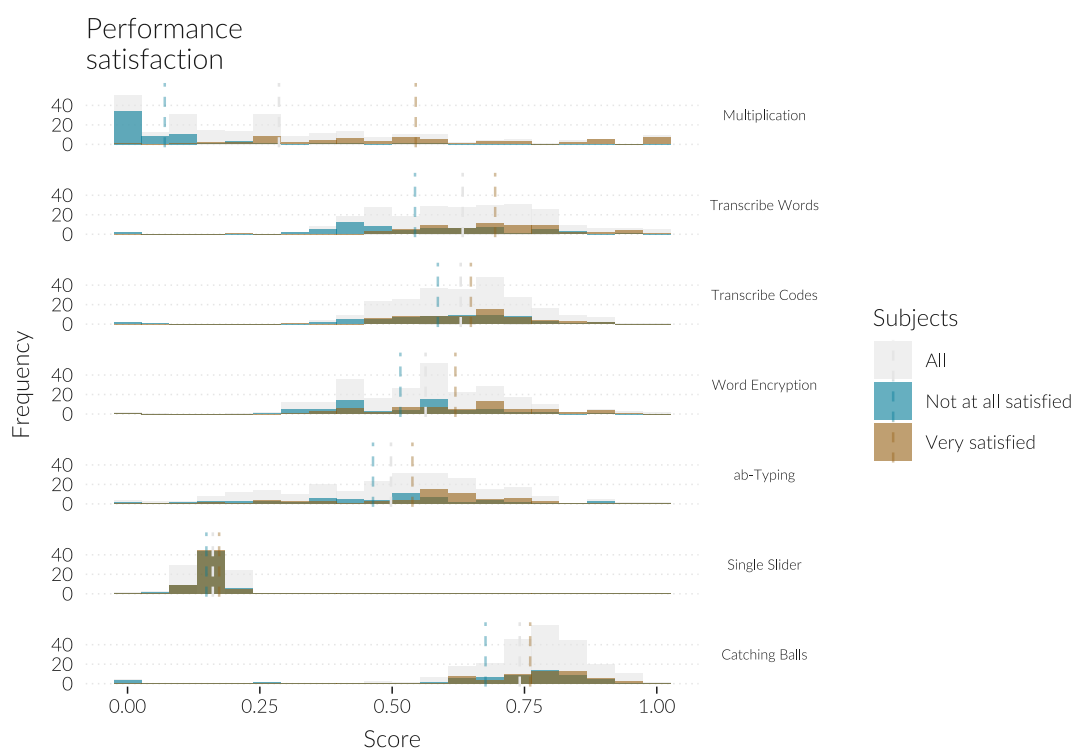


Figure 4.10: **Superimposed scores conditional on subjective performance assessment:** After completing a task, subjects were asked whether they are content with their performance. Possible answers ranged from (1) “not at all” to (7) “completely.” The figure presents the score distributions for all real-effort tasks with those of the respondents in the first and fourth quartile to this post-task subjective performance assessment superimposed.

ues of the remaining tasks). However, the above-mentioned previous experiences, which could date back to the subjects’ school days, may have both positive and negative associations and may subsequently evoke and trigger certain emotions and reactions. The question arises whether the reference point, which was established (long) before the start of the task and was brought into the mind at its beginning, can have a (de)motivating effect.⁵⁰ To illustrate this, consider the effect of the sight of a rowing ergometer or just a simple pull-up bar. Depending on physical fitness and general enjoyment of sports activities, the first glance at the sports equipment may arouse personal ambitions – or have a strong deterrent effect. Therefore, it would hardly be surprising if the performance subsequently

⁵⁰As part of the subject characterization at the onset of the experiment, subjects were asked to assess their mathematical abilities. This ex-ante assessment correlates weakly with the later performance in the multiplication task ($r(246) = 0.31$, $p < 0.001$), see also Figure 4.11, which plots subjects’ scores in the multiplication task against their mathematical abilities and color-coded by their subjective performance assessment.

delivered was linked to the first impression.⁵¹ Just like a pull-up bar, a real-effort task could unfold a motivating or demotivating effect on the study participants even before starting the task and only on the basis of their first impression or previous experience. This motivating or demotivating effect may then substantially influence or even determine their performance in the task.^{52,53} Therefore, the choice of the task seems to be all the more important.

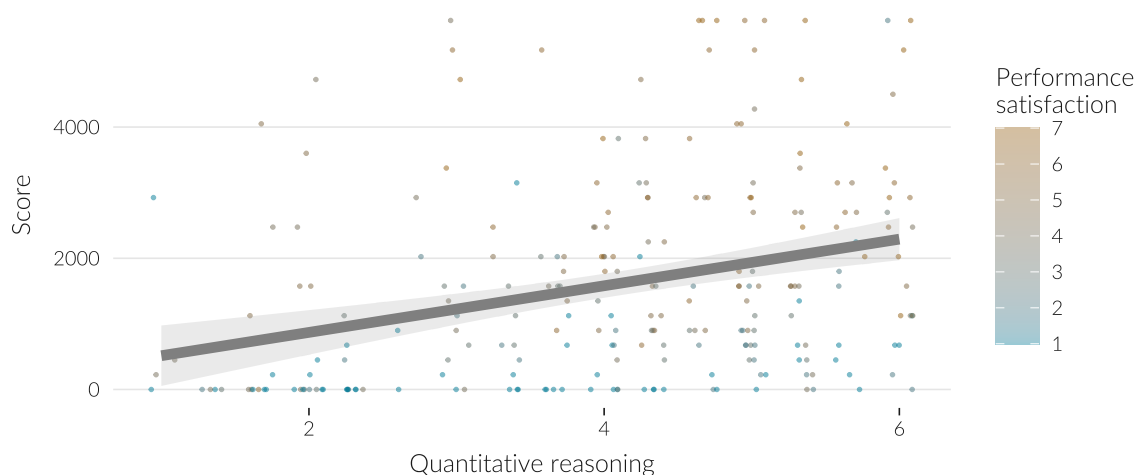


Figure 4.11: **Score distributions for the math task conditional on subjects' self-assessed numeracy skills and performance satisfaction:** At the onset of the experiment, subjects had to evaluate their mathematical skills. This ex-ante assessment is weakly correlated with the subsequent performance in the task ($r(246) = 0.31$, $p < 0.001$). Subjects ex-post assessment of their task performance is strongly correlated with their actual scoring ($r(246) = 0.67$, $p < 0.001$). Finally, a weak correlation is also observed between the subjects' subjective assessment of their mathematical skills and their evaluation of their task performance ($r(246) = 0.29$, $p < 0.001$).

⁵¹Rowing ergometers can be adjusted to different levels of difficulty. In principle, therefore, the experimental situation could be adjusted in such a way that everyone, regardless of whether they are athletically predisposed or not, can perform well. Nevertheless, individual performance is likely to follow the fitness level due to the demotivating first impression of the sports equipment.

⁵²On a related note, the multiplication task is the only task from the set for which *subjects' perception of the tasks* shows a greater correlation to their achieved score. The finding indicates that activity-related incentives are a crucial driver in subjects' performance in this task (see also Figure C.12 in the Appendix, which reports correlations between subjects' performance and their survey responses for all tasks).

⁵³The majority of study participants had the opportunity to switch to an alternative activity during the task, i.e., the well-known computer game *Snake* (see Appendix C.2.3.1). The change to this activity was irreversible so that the subjects could only earn money again when the time for the current task had elapsed, and the next task was due. The findings suggest that subjects switched to the outside option primarily for two reasons: 1) they were overburdened and demotivated by the actual task to be completed, or 2) they were tired and bored of the task. The former is observed for the multiplication task, the latter for the single-slider task. The multiplication task registers the largest number of subjects who give up the task to play the game *Snake*. Its strong demotivating effect is thus once again confirmed.

4.4 Conclusions

In the economic literature, it is commonly assumed that people *can* respond homogeneously in effort provision with respect to incentives. However, people are different by nature – they have different character traits and abilities and vary in their motives, needs, and goals. Therefore, the assumption of homogeneous responses to a particular incentive scheme for a given task appears to be a rather crude simplification. To shed light on this issue, this chapter studied effort responses to a given incentive structure – across a set of different tasks –, *conditional* on subject characteristics. More precisely, the controlled laboratory environment allowed to assess whether subjects' measured level of effort depends on *abilities*, *personality*, and *motivation*. For this, subjects were characterized with a range of psychological questionnaire measures. Next, their performance was assessed in a set of diverse tasks. The performance was then compared to the characteristics, separately for each task. The results indicate a considerable and not negligible influence of the subject's abilities, personality, and motivation on the effort measurement, although much less so for simple and generic tasks. The chapter offers an insight into the variation of these factors across different tasks and consequently, their relevance in general. This provides a first impression of possible implications for research results obtained with the tasks used (or comparable ones).

Since this is a research desideratum, any typification and systematization put forward here does not claim to be exhaustive but should at best serve as a starting point for further research.

The score distributions varied considerably among the set of tasks. Depending on the purpose of a task, certain forms of mean dispersion, kurtosis, and skewness may be advantageous. Consequently, different tasks are suitable for different applications. Suppose the participants are to be equipped with similar initial endowments for the further course of the experiment. In that case, tasks that promote a score distribution with only a small variance are beneficial. Among the examined tasks, such a distribution can be observed for the code-transcription task, the single-slider task, and the ball-catching task.

If, on the other hand, the subjects are to be differentiated from each other according to their efforts in as high a resolution as possible, then tasks that produce a score distribution with large variance are preferable. This is the case for the ab-typing task and less so for the word-transcription task and the

word-encryption task (although in the present group of tasks, the multiplication task has the greatest variance, this can be attributed primarily to differences in the subjects' mathematical abilities and only then to their efforts). In preparation for a study, researchers are encouraged to clarify whether a task has a desired resolution, i.e., an intended spread in performance. Conducting pilot studies provides useful to examine this decisive task property in advance (see also Section 2.4.5 for related design practices).

Comparing the study participants' performance across the various tasks, it becomes apparent that they scored very differently on each task. Piece rates of the tasks are comparable such that none of the tasks pays off more than the other. Assuming that the subjects are willing to make the same effort in all seven tasks, the obtained results suggest that further factors are at play. The analyses carried out in terms of the study participants' qualities and motivations provide initial insights.

Since the skills and personality traits were surveyed of the subjects before the tasks were completed (and also because these subject qualities can be assumed to be "fixed," at least for the duration of the experiment), any observed correlation provides support for a causal relationship between the quality in question and the subjects' performance in this task (although a third confounding variable may be present).

For certain tasks, subjects already have an existing reference point based on previous experience with this or similar tasks. This reference point may be motivating for some subjects and encourage them to set goals, i.e., to exceed their previous performance. However, for other subjects, the reference point can be negatively connotated and recall previous bad experiences. It can thus have a demotivating effect or even deprive them of any motivation.

To circumvent the issue, tasks can be used for which subjects are likely to have little or no prior experience and thus no reference point (with either positive or negative connotations). In addition, tasks that are discouraging by their appearance alone can be avoided. Exemplary tasks that go along with these suggestions are the ab-typing and the single-slider task. In the absence of a reference point, goal setting can act less as a motor for purpose-related incentives and thus induce voluntary effort. The tasks are mind-numbing but yet not so repulsive that they evoke shame and disgust or lead to powerlessness and mental surrender.

In terms of physical skills and computer use, the ability to type with ten fingers proves to be an

important and revealing measure for several of the examined tasks. In order to be able to assess its influence on performance more accurately, an explicit typewriting test could be carried out in follow-up studies.⁵⁴

Moreover, people likely have preferred directions of movement on the computer screen, i.e., they can easier and quicker move the cursor in some directions than in others. To take this into account, the subjects were asked whether they were left- or right-handed before they arrived at the single-slider task (roughly 0.07% of the subjects indicated they are left-handed). Based on this information, the sliding direction was adjusted. At the end of the experiment, the subjects could provide comments. A review of these revealed that the study participants' left-right-handedness did not necessarily coincide with their preferred hand for the computer mouse. Consequently, their preferred direction for cursor movements did not correspond to the ex-ante defined settings, resulting in these subjects having worse starting conditions, which increased the noise in the data. The following approach may allow circumventing this: Instead of asking subjects to state whether they are left- or right-handed, they could be provided directly with a single-slider with the instruction to slide it from left to the right and vice versa. Afterward, one could ask the subjects if it was easier or came more naturally to move the slider in one direction than the other. For the effort measurement, the sliding direction could then be fixed according to the preferred direction of cursor movement, proceeding with the usual task instructions.

Tasks in which colors play a role pose an additional challenge, especially for people who have difficulty differentiating colors. In the present study, this was observed for one of the characterization tasks.⁵⁵ If a task requires discrimination of colored objects, one solution would be to use colors easily recognized by color-blind people. In addition, eyesight could be checked to control for color blindness. Since some people are not even aware of their visual limitations, a simple color test provides more valuable data than a self-assessment. Of course, it is easiest to use a task in which colors do not play a central

⁵⁴Consider for example the type-writing test provided by [Tippenakademie](#).

⁵⁵The Dual-2-Back Task implemented following [Jaeggi et al. \(2010\)](#) allows assessing subjects' short-term memory. As described in Section [B.1.1](#), colored squares are shown for a blink of a second at different positions on the screen, and subjects have to recall their color and positions. Subjects were asked whether they have difficulties in differentiating colors. The twelve subjects who indicated that they are "slightly color-blind" achieved a score of 0.9 (SD = 0.15) in the Dual-2-Back Task, whereas the remaining study participants with no problems in sight scored 1.02 (SD = 0.4). A Welch two-samples t-test showed that the difference was statistically significant, $t(20.97) = 2.52$, $p < 0.05$.

role.

A number of control variables were additionally included in the analysis to take confounding factors into account. These cover gender, age, nationality (German or not), relationship status, and knowledge of experimental content. No substantial changes to the results are observed by adding in the control variables. Strong gender effects were present in the ab-typing task and less pronounced in the single-slider task. Controlling for gender is thus recommended for applications of these tasks.

Figure C.12 in the Appendix reports Pearson-correlations between subjects' scores and their assessment of a task with the real-effort task survey. Subjects' performance in a task has hardly any influence on their perception of the task. This finding applies to all tasks except the multiplication task and, to a certain extent, the ball-catching task.⁵⁶ The subjects thus seem to make a comparably objective assessment of a task's properties. Thus, unless subjects are severely demotivated due to a lack of skills, the survey proves to be a useful tool for evaluating tasks.

Section 2.4.2.2 discusses that adding an outside option may be vital to observe incentive effects. In the present experiment, however, the embedding of an alternative activity might have been counterproductive in order to demonstrate the effect that was intended to be observed. The reason for this is that precisely those subjects who are most influenced by the outside option, in fact, abort the task. Due to their switching to the game Snake, exactly the variation the study intended to observe is lost. This means that instead of observing how certain subjects "torture" themselves with a task over its entire duration and perform poorly, only their performance is observed until they switch to Snake prematurely.

Moreover, the outside option may not have been appealing enough to increase the opportunity cost of effort: At the end of the experiment, 53.6% of the participants indicated that they liked the game Snake; however, only 12.5% of them declared that they enjoyed playing Snake more than completing the tasks. Follow-up studies could include alternative activities that genuinely raise the opportunity cost of efforts, such as the option to leave the laboratory prematurely, as realized by Erkal et al. (2017).

⁵⁶In the multiplication task, the subjects' *lasting impression* of the task is basically determined by their performance in the task (see the correlation values between their perception and performance in each task in Figure C.12). Conversely, their performance in the task depends on their abilities and personality traits, their *previous experience* with this or similar tasks, but also on their *first impression* of the task in the run-up to it. In the end, this could even be decisive. This means that mathematically savvy subjects are stimulated to be zealous, while mathematically non-savvy subjects are deterred and demotivated. Their opposing views of the task as expressed ex-post in the survey are, therefore, not surprising.

At the end of the study, subjects were asked for the main reason for their participation in the study. Most subjects stated that they came primarily for the money (79.4%). However, another 14.5% said they participated out of general curiosity and another 6% out of personal interest in experimental economic research. For the given sample, one-fifth of the study participants may not or may only partially be steerable in their actions by monetary incentives. Another indication in this direction is that, after having completed seven tedious and demanding tasks, almost half of the subjects stated that they enjoyed doing so.⁵⁷ The view that “subjects rush into the lab just for the money” thus falls somewhat short.

Certainly, incentive effects can be observed in the laboratory. The results of this chapter suggest that this depends very much on the choice of task. For some tasks, skills, personality, and motivation combined may end up being more decisive for performance than any intended effort. Follow-up studies are desirable, for example, to investigate for the considered tasks to what extent the effort increases with the stake size.

⁵⁷46.8% of the subjects responded to the statement that the fulfillment of the task was “fun” with moderate to strong agreement.

5

Conclusion and Discussion

To summarize, this thesis provides insights “why” and “how” the choice of real-effort task matters. It acts as a practitioner’s guide to classifying, designing, choosing, and implementing real-effort tasks. With this, the thesis may contribute to the development of standards in the use of real-effort tasks and help streamline experimental practices to improve the comparability of results.

A central tenet was: *Know thy task*. Designing a task is not straightforward. A task should cost effort, but if it is challenging, some subjects get spurred on while others become frustrated. That is, be aware of what is involved with the choice of task in relation to the experiment and its results.

Therefore, this thesis resembles a sort of journey along the topic of real-effort tasks. It gives an overview of available tasks and their applications, sheds light on the properties of tasks, and points out issues to consider when choosing and implementing a particular task. It also presents a new task evaluation tool that can assist in the selection of tasks for any of the aforementioned applications.

Throughout the thesis, several concerns are raised regarding the design properties of tasks. Finally, an example is given of how these can influence the effort measurement, allowing a first impression of what tasks may measure in addition to the intentional effort.

This thesis showed that real-effort tasks differ considerably in i) their conception and design; ii) the extent to which they induce non-monetary incentives or privilege certain skills or personality traits; and iii) the type and amount of effort they demand. The tendency for the effort measure to be subject to influences beyond incentives varies from task to task. Also, it seems evident that the construct differs that tasks are evaluating. Already the first descriptive findings of Chapter 4 drew attention to the question of whether tasks really measure the same kind of “effort.” The performance of the participants varied too widely across the considered tasks. Hence, the question arises whether there are several types of “effort” and, if so, which or how many. A deeper consideration of the concept of “effort” is thus needed.

Even more, in light of the discussion on motivational psychology in Chapter 2, the concept of “performance” in return for monetary incentives may need to be reconsidered in some respects. Take as an example a professional soccer player for whom a plethora of performance measures are available: number of goals scored, meters run per game, number of successful passes – just to name a few. Which performance measure is the most adequate? Moreover, if one were to raise the player’s salary fivefold in order to study incentive effects, would his performance increase by the same amount? Would Pelé have scored five times as many goals per game, ran five times as far, and hit five times as many successful passes? Even a soccer player of this stature would hardly have been able to achieve such an increase in performance. Of course, a linear increase in effort and thus in the performance achieved is not necessarily to be expected and is somewhat exaggerated. But even a tripling of efforts would barely be possible. Therefore, it may seem reasonable that in quite a few situations, the observable incentive effects turn out to be rather small. As outlined earlier, other factors could prove to be much more influential on performance, e.g., how often Pelé is substituted even though he thinks he played very well, how well he gets along with the coach and with his teammates, and how often he is served his favorite dish for lunch.

A further discussion of different disciplines in this direction would be desirable and could lead to numerous insights. In addition to motivational psychologists, behavioral economists, and labor

economists, even biologists and philosophers might want to add to the debate. In the end, the benefit of this work may be that, like a good pass in soccer, it can improve the results of others.

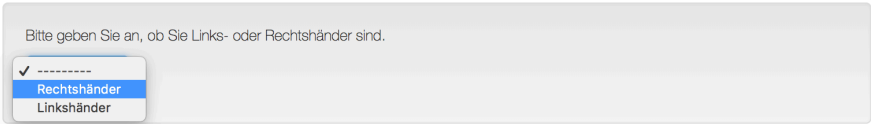
A

Appendix to Chapter 2

A.1 The Single-Slider Task: Instructions in English and German

First, subjects are asked whether they are left- or right-handed.

"links-rechts" Schieberaufgabe



Bitte geben Sie an, ob Sie Links- oder Rechtshänder sind.

✓ -----
Rechtshänder
Linkshänder

Mit dem Klick auf * **Weiter** * gelangen Sie zur Anleitung für die Aufgabe.

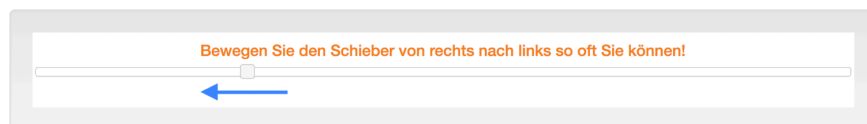
Figure A.1: **Task preparations – checking subjects' left- or right-handedness:** Query whether the subjects are left- or right-handed before receiving any information about the task. The direction of movement of the slider is adjusted accordingly behind the scenes.

Thereafter they are given detailed instructions on how to perform the task and earn payoffs.

"links-rechts" Schieberaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Bewegen eines Schiebers Geld verdienen. Ihre Aufgabe ist es mit der Maus den **Schieber von rechts nach links** zu ziehen.

Illustration der Aufgabe:



Der Schieber wurde bereits über die Hälfte nach links bewegt. Kommt man mit dem **Schieber am Ende der Linie an**, springt dieser wieder an den **Anfang der Linie zurück und bestätigt damit die erfolgreiche Bewegung**.

Ziel der Aufgabe ist es, in einer vorgegebenen Zeit **so oft wie möglich den Schieber von einem zum anderen Ende der Linie** zu bewegen.

Dabei ist zu beachten:

- Sie erhalten **10 Punkte** für **jede vollständige Bewegung des Schiebers von rechts nach links**;
- Für **unvollständige Bewegungen des Schiebers** erhalten Sie **keine Punkte**.
- **Sie müssen den Schieber nach jedem vollständigen Durchzug loslassen (Bestätigung erfolgt durch grünes aufleuchten), um Punkte zu erhalten.**
- Halten Sie den Schieber ununterbrochen beim Hin- und Herbewegen fest, erhalten Sie **keine Punkte**.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.

In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure A.2: Instructions for the task

"links-rechts" Schieberaufgabe

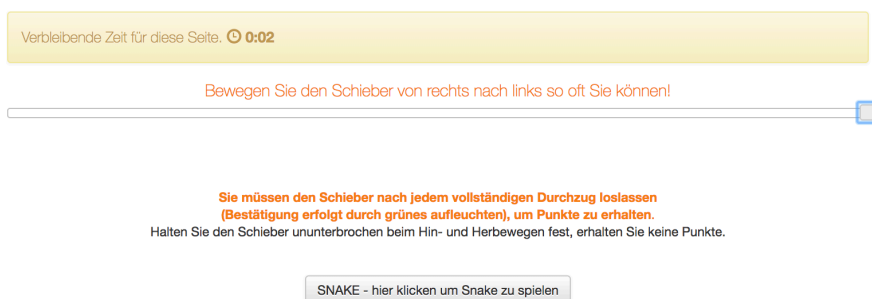


Figure A.3: **Single-slider task:** To earn points, subjects have to move the slider from one end of the line to the other. They must click on the slider, move the mouse cursor along the line to its end and only then release the mouse button. To confirm that a slider movement has been successfully completed and a point has been scored, the line lights up in green. Subjects can familiarize themselves with the task for 15 seconds in a trial round.

"links-rechts" Schieberaufgabe: Ergebnisse

Vollständige Bewegungen von rechts nach links: 2

Anzahl vollständige Schiebewegungen	Anzahl an gesammelten Punkten
2	20 Punkte

In dieser Aufgabe haben Sie somit **20 Punkte** gesammelt (10 Punkte für jede vollständige Bewegung des Schiebers von rechts nach links). Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure A.4: Results page

B

Appendix to Chapter 3

B.1 Details on Experimental Design and Procedure

B.1.1 Experimental Procedure Details

The experiment follows a within-subject design so that the subjects complete all seven tasks, one after the other. Subjects face each task and the subsequent real-effort task survey only once. The timing of the experiment is illustrated in Figure [B.1](#). Each experimental session followed this same protocol. The elements of the experimental procedure are described in the following.

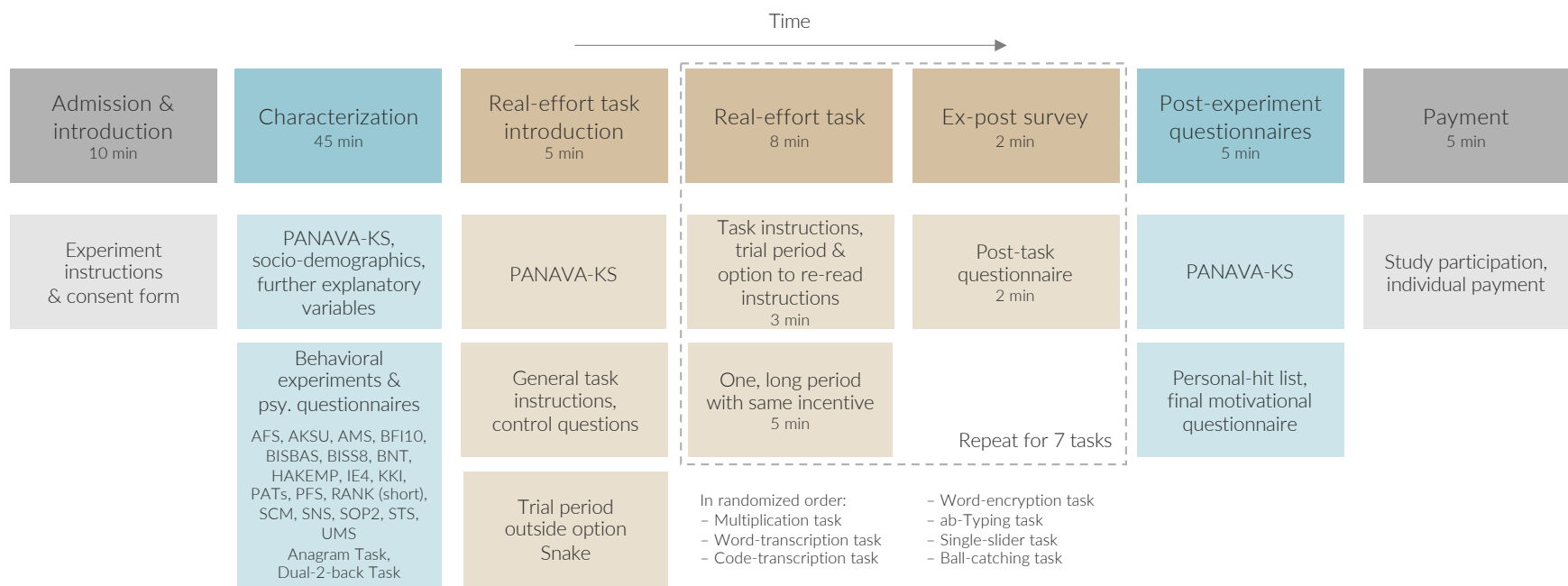


Figure B.1: **Experimental procedure (detailed)**

(1) Experiment Instruction and Consent Form

As a first step, the participants were welcomed and the set-up of the experiment was explained. They were provided with detailed information on the duration of the experiment and the type of information to be revealed in the questionnaires. After exposing the experiment's framework, all participants were free to decide whether they wanted to participate in the study or not. All subjects taking part in the experiment were asked to complete the consent form.

(2) Characterization

The second step of the experiment is the subject of Chapter 4. It aims to show how effort responses to a particular incentive scheme vary according to individual characteristics for different tasks. Therefore, the participants were asked to answer standardized questionnaires that are frequently used in the field of psychology and behavioral economics. The characterization questions seek to determine the degree of certain character traits and skills, such as self-control, attention, cognitive abilities, and literacy, as well as the individual level of motivation.

(3) Real-Effort Task and (4) Ex-post Survey

In the main part of the experiment, subjects successively completed a number of real-effort tasks. These differ significantly in the abilities they demand from the subjects and can be grouped in five categories as follows:¹

- *Quantitative & analytical reasoning*: multiplication task;
- *Language & verbalizing*: transcription task *with words* (word-transcription task);
- *Memory & knowledge*: transcription task *with random letter combinations* (codes-transcription task), word-encryption task;
- *Mechanical*: alternately pressing-keys task (ab-typing task), single-slider task;
- *Entertainment*: ball-catching task.

¹Further details on the task selection and the reasoning behind are provided in Section 3.3.1.1.

First, the subjects were given a detailed introduction to this part of the experiment. Several control questions subsequently had to be answered to ensure that the subjects understood them. Subjects in the treatment with an outside option also had the opportunity to familiarize themselves with it (see further information below). For each task, subjects were given a short trial period of 15 seconds. Afterward, they have five minutes to provide effort. Subjects received points that were directly related to their individual performance in fulfilling the task (piece-rate incentive). At the end of the experiment, all earned points are converted from the experimental to the real-world currency with a conversion rate announced at the outset of the experiment.

After each task, subjects had to fill out a brief questionnaire to elicit their perception of the respective task as well as potential motives and attitudes concerning their effort provision.² To account for spill-over effects between subsequent tasks, the order of tasks is randomized across groups. However, tasks from a given category may demand the same mental or physical resource (e.g., concentration). To reduce such depletion effects, tasks from the same category may not follow upon another.

If subjects become bored of a task they may chose an outside option and play the well-known computer game “Snake.”³ However, by choosing to switch to the outside option, subjects also choose to stop their effort provision and subsequent collection of earnings for a given task. As soon as the respective task’s time has elapsed, subjects may revert to providing effort and earning money by completing the next task. Importantly, subjects’ behavior was not manipulated in any kind of way such that individual effort provision may be observed only conditional on incentives and subject characteristics.

(5) Post-experiment Questionnaires

After completing all tasks, participants were asked to fill out two questionnaires. To compare the entire selection of tasks, subjects were explicitly asked to rank these following [Rheinberg \(1989\)](#)’s method of a “personal-hit list.” A final questionnaire elicited potential motives and attitudes concerning their study participation and effort provision overall. For those subjects who had the opportunity to switch to the alternative activity *Snake*, the questionnaire also contained two items regarding it.

²The survey is described in closer detail in Section 3.2.

³This feature was available to 209 of the 248 study participants. A control treatment with 39 subjects completed the same experiment without the possibility to switch to an outside option.

(6) Payment

In the last step, the payments to the participants were made confidentially. The payoff to each participant consisted of a fixed payment (5 € show-up fee) and additional effort related payoff. As the experiments lasted on average 2 hours and 10 minutes subjects' earnings accumulated to on average 21,80 € (payments ranged from 15,70 € to 29,30 €). The wage rate, therefore, approaches 10,00 € per hour, as recommended by the guidelines of the WiSo-Research Laboratory.

The experiment was designed such that i) no participant was disadvantaged from the outset and ii) a fair reimbursement for his or her time and contribution was granted. Therefore, each subject ex-ante had the same probability of being part of a specific session configuration. Moreover, to assure that the final payoffs are comparable across session configurations, the incentive schemes were adjusted to keep payoffs similar across tasks. To this end, a pilot experiment was conducted in April 2018. Piece-rates were gauged such that subjects would earn approximately 2.50 € in each task.

To avoid experimenter effects and assure internal validity, several precautions were taken:

- *interactions between subjects and experimenter were kept at a minimal level*: subjects were seated and enclosed in cubicles. Standardized instructions were displayed on-screen and played as a pre-recorded audio file. The study participants knew that a research assistant was present in the laboratory; However, they did not directly interact with each other apart from entering the laboratory or if the subjects had a comprehension question regarding the experiment's content.
- *external research assistants*: the experiment execution was commissioned to the WiSo-Research Laboratory. The research assistant present in the laboratory was not involved in the research project. Hence, she is "blind" to the experiment, i.e., neither knows the purpose nor the hypotheses of the study.
- *within-subject design*: each subject completed *all* tasks and subsequently the real-effort task survey. Also, subjects could participate in the experiment only once.

B.1.2 Subject Recruitment, Data Collection, and Anonymization Process

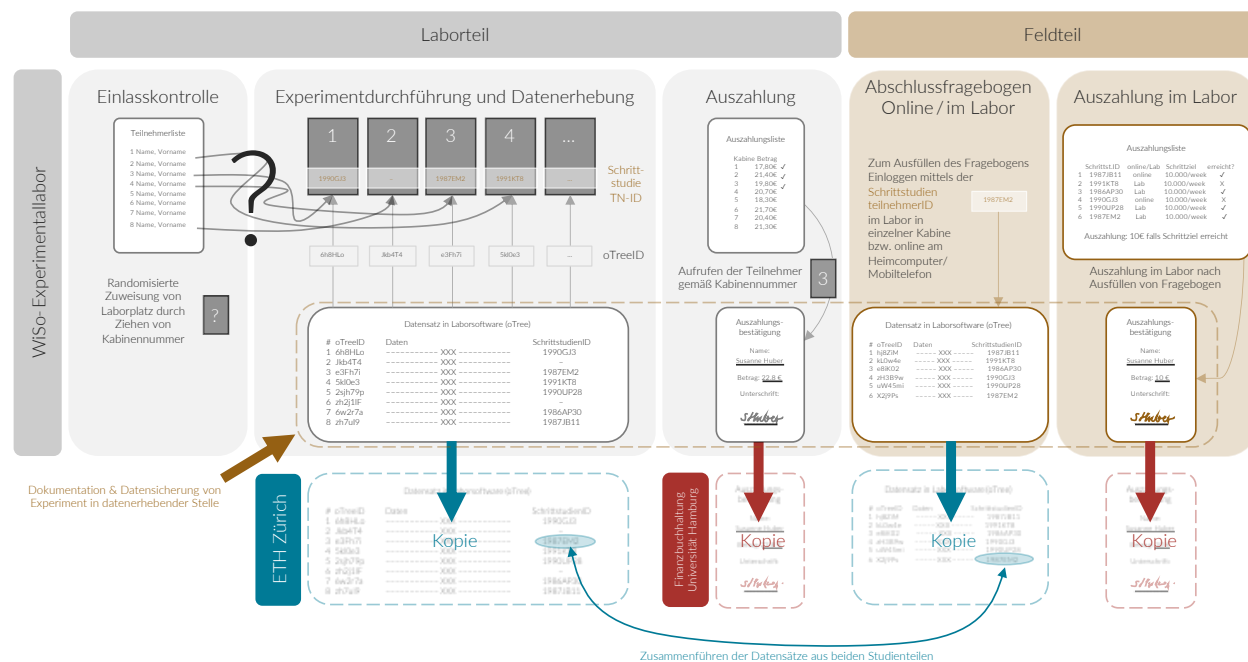


Figure B.2: **Data collection and anonymization process:** At the end of the experiment, the subjects could decide to participate in a small field study (step study). This independent study investigates the extent of the students' physical activity over the course of a week in relation to various incentive schemes. The figure illustrates the precautionary measures taken to ensure the study participants' privacy in both studies and across them.

B.1.2.1 Recruitment of Participants; Selection and Eligibility Criteria

Study participants were recruited through the subject pool from the WiSo-Research Laboratory.⁴ The pool currently contains 1,600 active members, which are mostly students from the University of Hamburg. Registration for the pool is voluntary, and subjects must be at least 18 years old. Besides aiming at an equal gender distribution in the sample there were no further selection or exclusion criteria. Since the WiSo-Research Laboratory staff performed the study's administration, I, as the researcher, did not interfere with the participants' recruitment, admission to the laboratory, and payment after the experiment.

⁴Recruitment was conducted using the experiment management system *hroot* (Bock et al., 2014).

B.1.2.2 Privacy and Anonymity in Decision-Making

The design of the experimental laboratory permits that participants remain anonymous towards each other, assuring complete privacy and unbiased decision-making. In the written instructions, participants were informed that only I, as the researcher, will learn about their decisions, but in an anonymized way (see instructions included in Appendix [B.1.4](#)). In detail, the data was truncated such that even I, as the researcher, am unable to match individual decisions with participant identities to further assure and protect both their anonymity and privacy.

B.1.2.3 Data Collection and Handling

For data collection and handling, a particularly cautious procedure (see Figure [B.2](#)) was chosen. Simultaneously, the amount of personal data collected was kept as low as possible, and all data is anonymized, allowing for a maximal amount of privacy for the participants. To assure the study participants' anonymity and privacy, the WiSo-Research Laboratory in Hamburg was commissioned to act as a "data-collecting body." As an interim instance between the subjects and me as the researcher, it conducted the experiments, collected the data, and separated personal from analysis data. I only received a truncated data set for my research purposes. Thereby, the personal data of study participants did not leave the University of Hamburg. Thus, all handing of sensible information was outsourced to the WiSo-Research Laboratory and lay with the University of Hamburg. As a service provider, the WiSo-Research Laboratory solely acts as a "data-collecting body" and not as a project partner.

The data collection process at the University of Hamburg is performed at high standards. The data is managed in a protected area within an isolated network and in partitioned folders. A procedural description confirming that the WiSo-Research Laboratory Hamburg acts as a data-collection provider and is solely responsible for delivering the truncated data is provided in Appendix [B.4.1](#).

In summary, the truncated data I received from the WiSo-Research Laboratory does not permit to match participants' identities and private information (i.e., name, email address, study program, etc.). Therefore, the laboratory setting described earlier, together with the particular data-collection procedure outlined above, grants complete anonymity and privacy to the study participants, ruling out any potential risks to both.

B.1.3 Information Sheet for Participants and Consent Form

Teilnehmerinformation

Titel der Studie: "Experimentelle Studie von individuellem Einsatz in Aufgaben mit unterschiedlichem Aufwand"

Ziele der Studie

Das Hauptziel dieser Studie ist die Analyse von individuellem Einsatz unter verschiedenen Verträgen. Abhängig von Ihren Entscheidungen und Ihrem Einsatz können Sie Geld verdienen.

Untersuchungsmethode und Ablauf

Diese Studie besteht aus mehreren Teilen: Sie treffen einige Entscheidungen, bearbeiten eine Aufgabe und füllen mehrere Fragebögen aus. Die gesamte Studie wird von Ihnen am Computer durchgeführt. Instruktionen zu den einzelnen Teilen werden Ihnen am Computerbildschirm genau erläutert. Nachdem Sie die zuvor genannten Teile absolviert haben, möchten wir Sie abschließend bitten noch einen kurzen Fragebogen, ebenfalls am Computer, auszufüllen. Die gesamte Studie nimmt ungefähr 1 Stunde und 30 Minuten in Anspruch.

Teilnahmebedingungen

Mit Ihrer Anmeldung zur Studie akzeptieren Sie die folgenden Teilnahmebedingungen: (1) Sie sind mindestens 18 Jahre alt, (2) die Nicht-Beachtung der Instruktionen während der Studie kann zum Ausschluss aus der Studie führen.

Risiken

Die Studie ist mit keinerlei Risiken für Sie verbunden. Abhängig von Ihren Entscheidungen und Ihrem Einsatz können Sie Geld verdienen. In jedem Fall erhalten Sie für Ihre Teilnahme ein entsprechendes Entgelt.

Vergütung

Für Ihren Zeitaufwand erhalten Sie ein fixes Entgelt von 5 EUR. Darüber hinaus können Sie im Verlauf des Experimentes Geld hinzuverdienen.

Rücktrittsrecht

Sie haben als Teilnehmerin/ Teilnehmer jederzeit das Recht ohne Angabe von Gründen aus der Studie auszutreten. Es entstehen Ihnen dadurch keine Nachteile. Sie erhalten jedoch, wenn Sie vorzeitig aus der Studie austreten, nur die feste Vergütung von 5 EUR. Mögliche monetäre Ansprüche aus der Studie erlöschen damit. Auf Ihren Wunsch können alle bis zu diesem Zeitpunkt innerhalb der Studie erhobenen Daten vernichtet werden.

Datenschutz

Die erhaltenen Daten werden vertraulich behandelt und irreversibel anonymisiert. Sie sind sicher auf Servern der ETH Zürich verwahrt und werden nur für Forschungszwecke durch die ETH Zürich und die Universität Basel verwendet. Die Mitglieder der Ethikkommission der ETH Zürich können diese Daten zu Prüf- und Kontrollzwecken einsehen, jedoch unter strikter Einhaltung der Vertraulichkeit. Ihre Privatsphäre wird während des gesamten Versuchsablaufes und bei der Datenverwaltung durch die Universität Hamburg sowie bei der Datenverarbeitung durch die ETH Zürich und die Universität Basel gewahrt.

Versicherungsschutz

Allfällige Gesundheitsschäden, die in direktem Zusammenhang mit der Studie entstehen und auf nachweisliches Verschulden der ETH Zürich zurückzuführen sind, sind durch die Betriebs-Haftpflichtversicherung der ETH Zürich (Police Nr. 30/4.078.362, Basler Versicherung AG) gedeckt. Darüber hinaus liegt die Unfall-/Krankenversicherung (z.B. für die Hin- und Rückreise) in der Verantwortung der Versuchsteilnehmerin/des Versuchsteilnehmers.

Kontakt, Projektfinanzierung und -genehmigung

Christian Waloszek, Professur für Öffentliche Finanzen, ETH Zürich, Leonhardstrasse 21, 8092 Zürich, Schweiz
Dieses Forschungsprojekt wird über durch die MTEC Foundation der ETH Zürich zur Verfügung gestellte Mittel finanziert und wurde durch die Ethikkommission der ETH Zürich bewilligt (EK-2018-N-08).

Figure B.3: Information sheet for the participants (Teilnehmerinformation)

Einverständniserklärung

- ⇒ Bitte lesen Sie dieses Formular sorgfältig durch.
- ⇒ Bitte kontaktieren Sie den/die Untersucher/in oder Ihre Kontaktperson, wenn Sie etwas nicht verstehen oder etwas wissen möchten.

Studientitel: Experimentelle Studie von individuellem Einsatz in Aufgaben mit unterschiedlichem Aufwand

Durchführungsort: WISO-Forschungslabor, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Deutschland

Untersucher: Christian Waloszek, Professur für Öffentliche Finanzen, ETH Zürich, Leonhardstrasse 21, 8092 Zürich, Schweiz

Dateneigner (Name und Vorname):

- ⇒ Ich nehme an dieser Studie freiwillig teil und kann jederzeit ohne Angabe von Gründen meine Zustimmung zur Teilnahme widerrufen, ohne dass mir deswegen Nachteile entstehen.
- ⇒ Ich wurde schriftlich über die Ziele, den Ablauf der Studie, über die zu erwartenden Wirkungen, über mögliche Vor- und Nachteile sowie über eventuelle Risiken informiert.
- ⇒ Ich habe die zur oben genannten Studie die Teilnehmerinformation für die Studienteilnehmer gelesen. Meine Fragen im Zusammenhang mit der Teilnahme an dieser Studie sind mir zufriedenstellend beantwortet worden.
- ⇒ Ich hatte genügend Zeit, um meine Entscheidung zu treffen.
- ⇒ Ich bestätige mit meiner Unterschrift, dass ich die im Informationsblatt genannten Bedingungen für die Studienteilnahme erfülle.
- ⇒ Ich bin darüber informiert, dass die allgemeine Haftpflichtversicherung der ETH Zürich (Police Nr. 30/4.078.362, Basler Versicherung AG) nur Gesundheitsschäden deckt, die in direktem Zusammenhang mit der Studie entstehen und auf nachweisliches Verschulden der ETH Zürich zurückzuführen sind. Darüber hinaus liegt die Unfall-/Kranken-versicherung (z.B. für die Hin- und Rückreise) in meiner Verantwortung.
- ⇒ Ich bin einverstanden, dass die zuständigen Untersuchenden und/oder Mitglieder der Ethikkommission zu Prüf- und Kontrollzwecken meine Originaldaten einsehen dürfen, jedoch unter strikter Einhaltung der Vertraulichkeit.
- ⇒ Ich bin mir bewusst, dass während der Studie die in der Teilnehmerinformation für Studienteilnehmer genannten Anforderungen und Einschränkungen einzuhalten sind. Im Interesse meiner Gesundheit kann mich die untersuchende Person auch ohne gegenseitiges Einverständnis von der Studie ausschließen.

Ort, Datum Unterschrift Studienteilnehmer/in

Ort, Datum Unterschrift Untersucher/in

Figure B.4: Consent form for the participants (Einverständniserklärung)

B.1.4 Instructions and Screenshots of the Experiment

The procedure of the experiment is illustrated in Figure B.1.

B.1.4.1 Experiment Content and Author Contributions

The experiment was programmed in Python using the experimental laboratory software oTree. For each element in the configuration a reference to the associated literature (if available) and information about the respective designer and programmer is provided in Table B.1.

B.1.4.2 Admission and Introduction

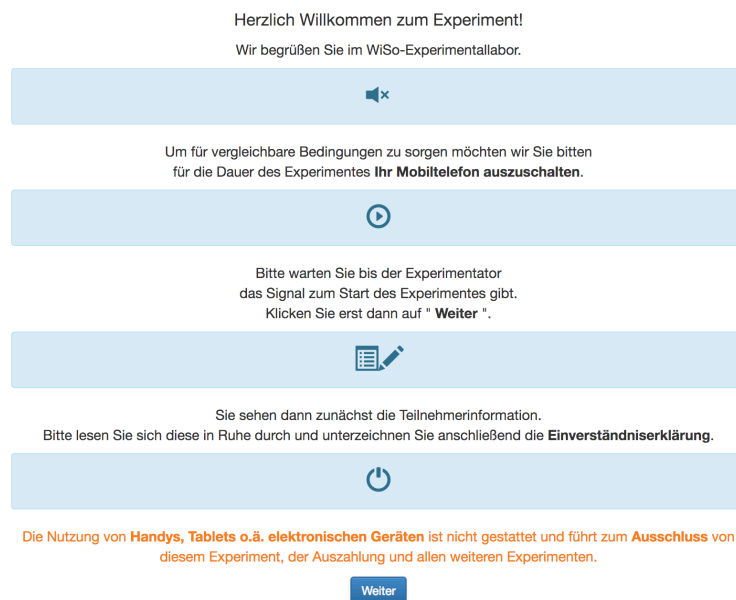


Figure B.5: Welcome screen

Participant information ("Teilnehmerinformation")

Studententitel

"Experimentelle Studie von individuellem Einsatz in Aufgaben mit unterschiedlichem Aufwand"

Ziele der Studie

Das Hauptziel dieser Studie ist die Analyse von individuellem Einsatz unter verschiedenen Verträgen. Abhängig von Ihren Entscheidungen und Ihrem Einsatz können Sie Geld verdienen.

Untersuchungsmethode und Ablauf

Diese Studie besteht aus mehreren Teilen: Sie treffen einige Entscheidungen, bearbeiten eine Aufgabe und füllen mehrere Fragebögen aus. Die gesamte Studie wird von Ihnen am Computer durchgeführt. Instruktionen zu den einzelnen Teilen werden Ihnen am Computerbildschirm genau erläutert. Nachdem Sie die zuvor genannten Teile absolviert haben, möchten wir Sie abschließend bitten noch einen kurzen Fragebogen, ebenfalls am Computer, auszufüllen. Die gesamte Studie nimmt ungefähr 2 Stunde in Anspruch.

Teilnahmebedingungen

Mit Ihrer Anmeldung zur Studie akzeptieren Sie die folgenden Teilnahmebedingungen: (1) Sie sind mindestens 18 Jahre alt, (2) die Nicht-Beachtung der Instruktionen während der Studie kann zum Ausschluss aus der Studie führen.

Risiken

Die Studie ist mit keinerlei Risiken für Sie verbunden. Abhängig von Ihren Entscheidungen und Ihrem Einsatz können Sie Geld verdienen. In jedem Fall erhalten Sie für Ihre Teilnahme ein entsprechendes Entgelt.

Vergütung

Sie erhalten für Ihren Zeitaufwand ein fixes Entgelt von 5 €. Darüber hinaus können Sie im Verlauf des Experimentes Geld hinzuverdienen.

Rücktrittsrecht

Sie haben als Teilnehmerin/Teilnehmer jederzeit das Recht ohne Angabe von Gründen aus der Studie auszutreten. Es entstehen Ihnen dadurch keine Nachteile. Sie erhalten jedoch, wenn Sie vorzeitig aus der Studie austreten, nur die feste Vergütung von 5 €. Mögliche monetäre Ansprüche aus der Studie erlöschen damit. Auf Ihren Wunsch können alle bis zu diesem Zeitpunkt innerhalb der Studie erhobenen Daten vernichtet werden.

Datenschutz

Die erhaltenen Daten werden vertraulich behandelt und irreversibel anonymisiert. Sie sind sicher auf Servern der ETH Zürich verwahrt und werden nur für Forschungszwecke durch die ETH Zürich und die Universität Basel verwendet. Die Mitglieder der Ethikkommission der ETH Zürich können diese Daten zu Prüf- und Kontrollzwecken einsehen, jedoch unter strikter Einhaltung der Vertraulichkeit. Ihre Privatsphäre wird während des gesamten Versuchsablaufes und bei der Datenverwaltung durch die Universität Hamburg sowie bei der Datenverarbeitung durch die ETH Zürich und die Universität Basel gewahrt.

Versicherungsschutz

Allfällige Gesundheitsschäden, die in direktem Zusammenhang mit der Studie entstehen und auf nachweisliches Verschulden der ETH Zürich zurückzuführen sind, sind durch die Betriebs-Haftpflichtversicherung der ETH Zürich (Police Nr. 30/4.078.362, Basler Versicherung AG) gedeckt. Darüber hinaus liegt die Unfall-/Krankenversicherung (z.B. für die Hin- und Rückreise) in der Verantwortung der Versuchsteilnehmerin/des Versuchsteilnehmers.

Kontakt, Projektfinanzierung und -genehmigung

Christian Waloszek, Professur für Öffentliche Finanzen, ETH Zürich, Leonhardstrasse 21, 8092 Zürich, Schweiz. Dieses Forschungsprojekt wird über durch die MTEC Foundation der ETH Zürich zur Verfügung gestellte Mittel finanziert.

Bitte unterzeichnen Sie nun die Einverständniserklärung.

Klicken Sie bitte erst auf "Weiter," wenn Sie die Einverständniserklärung unterzeichnet haben.

Mit dem Klick auf " Weiter " gelangen Sie zur Studie.

Weiter

General instructions ("Allgemeine Informationen")

Herzlich willkommen!

Sie werden jetzt an einem wissenschaftlichen Experiment teilnehmen, bei dem Sie Geld verdienen können. Der Betrag, den Sie verdienen, hängt von Ihren persönlichen Entscheidungen während des Experiments ab. Daher ist es wichtig, dass Sie die Anweisungen sorgfältig lesen.

Alles was Sie wissen müssen, um an diesem Experiment teilzunehmen, wird im Folgenden erklärt. Sollten Sie Schwierigkeiten haben, diese Anweisungen zu verstehen, heben Sie bitte Ihre Hand und warten Sie, bis einer der Experimentatoren zu Ihnen kommt.

Bitte beachten Sie, dass es während des Versuchs nicht gestattet ist, mit anderen Teilnehmern zu kommunizieren. Wenn Sie vorsätzlich gegen diese Regel verstoßen, werden Sie aufgefordert, zu gehen. In diesem Fall können Sie für Ihre Teilnahme **nicht** ausbezahlt werden.

Alle Teilnehmer erhalten eine **Teilnahmeentschädigung von 5 €**. Im Verlauf des Experiments können Sie Punkte sammeln. Alle Punkte, die Sie generieren, werden am Ende des Experiments in Euro umgerechnet und dann zu Ihrer Teilnahmeentschädigung hinzugefügt. Der Wechselkurs ist:

1000 Punkte = 1.00 €

1 Punkt = 0.001 €

Nach Abschluss des Versuchs werden die verdienten Einnahmen (zzgl. 5 € Teilnahmeentschädigung) bar und privat an Sie ausgezahlt: *Kein anderer Teilnehmer kann sehen, wie viel Sie erhalten.*

Anonymität: Alles ist anonym. Sie werden nicht den Namen und die Identität der Teilnehmer erfahren, mit denen Sie zusammen interagiert haben. Diese werden auch niemals Ihre Identität erfahren. Sie werden nicht erfahren, welche Entscheidungen von einem bestimmten Teilnehmer getroffen wurden und kein anderer Teilnehmer kann Ihre Entscheidungen einsehen.

Durchführung des Experimentes: Bitte lassen Sie sich nicht durch die Tastengeräusche anderer Teilnehmer irritieren. Falls Sie diese zu sehr stören verwenden Sie bitte das Ihnen gratis zur Verfügung gestellte Paar Ohrstöpsel. Gerne können Sie dieses nach dem Experiment behalten.

Die Dauer des Experiments beträgt ungefähr 120 Minuten. Manche Teilnehmer sind in der Bearbeitung schneller als andere und daher gegebenenfalls früher fertig. Eine vorzeitige Auszahlung ist **nicht** möglich.

Sollten Sie früher mit dem Experiment fertig sein, greifen Sie bitte auf die von Ihnen mitgebrachte Lektüre zurück, um die Zeit zu überbrücken. **Die Verwendung von Mobiltelefonen und anderen elektronischen Geräten ist während des gesamten Experimentes nicht gestattet.**

Auf Ihrem Arbeitsplatz finden Sie ein Quittungsformular. Bitte füllen Sie dieses erst am Ende des Experimentes, wenn Sie dazu aufgefordert werden, aus.

Weiter

Table B.1: Experiment content and author contributions

Item	Reference	Design development	oTree implementation
I. General			
Instructions	-	CW	CW
PANAVAKS (Introduction)	Schallberger (2005)	CW	CW
Socio-demographics	-	CW	CW
Further explanatory variables	-	CW	CW
II. Characterization			
<i>II.a Questionnaires</i>			
AFS, AKSU, AMS, BF10, BISBAS, BISS8, BNT, HAKEMP, IE4, KKI, ATs, PFS, RANK (short), SCM, SNS, SOP2, STS, UMS (stats), UMS (goals)	See C.1.1 (ref:AmmonsA1959-citation) Ammons & Ammons (1959) (ref:JaeggiThe2010-citation) Jaeggi et al. (2010) (ref:DohmenPerformance2011-citation) Dohmen & Falk (2011) (ref:KephartoTree2017-citation) Kephart (2017) (ref:ErkalRelative2011-citation) Erkal et al. (2011) (ref:BergerCan2011-citation) Berger & Pope (2011) (ref:GchterCombining2016-citation) Gächter et al. (2016)	CW	CW
<i>II.b Characterization tasks</i>			
Anagram Task	Ammons & Ammons (1959)	CW	AS
Dual-2-back Task	Jaeggi et al. (2010)	CW	CW
PANAVAKS (Characterization)	Schallberger (2005)	CW	CW
III. Real-effort tasks			
Introduction to Real-Effort Tasks	-	CW	CW
<i>III.a Pages that are identical for all tasks:</i>			
Confirm task-understanding	-	CW	CW
Count-down page (before each task begins)	-	CW	CW
Outside Option: "Snake"	-	AS	AS
Relaxation page (after each task)	-	CW	CW
<i>III.b Tasks:</i>			
A) Multiplication task	Dohmen & Falk (2011)	FW	FW
B) Word-transcription task	Waloszek modified from Kephart (2017), unpublished	CW	CW modified from CK
C) Code-transcription task	Kephart (2017)	CK	CK
D) Word-encryption task	Erkal et al. (2011)	FW	FW
E) ab-Typing task	Berger & Pope (2011)	CW	AS
F) Single-slider task	Mimeo	CW	AS
G) Ball-catching task	Gächter et al. (2016)	FW	FW
PANAVAKS (Real-effort tasks)	Schallberger (2005)	CW	CW
IV. Final questionnaires			
Personal-hit list	Rheinberg (1989)	CW	CW
Final motivational survey	-	CW	CW
V. End			
Study participation	-	CW	CW
Payment	-	CW	CW

Note:

For the sequence of experimental contents, see Figure B.1. For the tasks designed and implemented in oTree by Fabian Winter (FW), see Winter (2017). Contributions by Alexander Sandukovskiy are abbreviated with AS, by the author of this thesis with CW.

B.1.4.3 Characterization

PANAVA-Kurzskala (PANAVAKS)

Fragebogen

Wie fühlten Sie sich *unmittelbar* nach der Aufgabe?

	sehr 1	2	3	unent- schieden 4	5	6	sehr 7	
zufrieden	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	unzufrieden
energiegeladen	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	energieelos
gestresst	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	entspannt
müde	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	hellwach
friedlich	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	verärgert
unglücklich	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	glücklich
lustlos	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	hoch motiviert
ruhig	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	nervös
begeistert	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	gelangweilt
besorgt	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	sorgenfrei

Weiter

Figure B.6: **PANAVAKS scale:** The scale was implemented following [Schallberger \(2005\)](#). It was used throughout the experiment to determine the state of mind of the study participants: after the introduction, after the characterization (prior to the tasks) and after completion of all tasks (before the final questionnaires).

Socio-demographics

Anleitung

Bitte machen Sie im Folgenden einige Angaben zu Ihrer Person.

Geben Sie bitte Ihr Geschlecht an

- männlich
- weiblich

Die Ergebnisse dieser Befragung werden auch für unterschiedliche Altersgruppen ausgewertet. Bitte nennen Sie dazu Ihr Alter.

Bitte geben Sie Ihr Körpergröße in cm an.

Bitte geben Sie Ihr Körpergewicht in kg an.

Welche Staatsangehörigkeit haben Sie? (list provided with EU and further common countries; may chose "other" if home country not contained in list)

Wenn Sie bei der letzten Frage "andere" gewählt haben, bitte tragen Sie im folgenden Feld Ihr Herkunftsland ein. _____

Haben Sie derzeit eine feste Partnerschaft?

- Nein
- Ja

Wie setzt sich Ihr Haushalt zusammen? Ich wohne...

- bei meinen Eltern
- alleine
- zusammen mit Partner/Partnerin
- in einer Wohngemeinschaft

Bitte geben Sie an wie viele Kinder Sie haben: _____

Welche der folgenden Angaben beschreibt am besten das Fachgebiet, in dem Sie Ihren höchsten Abschluss erhalten haben? Wählen Sie bitte "kein Studium," falls Sie kein Studium absolvieren/absolviert haben.

- Lehramt
- Gesellschafts- und Sozialwissenschaften
- Rechts- und Wirtschaftswissenschaften
- Geistes- und Kulturwissenschaften
- Kunst und Gestaltung
- Medizin und Gesundheitswesen
- Agrar- und Forstwissenschaften
- Mathematik und Naturwissenschaften
- Ingenieurwissenschaften
- Sonstiges
- – kein Studium –

Wie hoch ist Ihr eigenes durchschnittliches monatliches Nettoeinkommen? Bei dieser Frage geht es darum, Gruppen in der Bevölkerung mit z. B. hohem, mittlerem oder niedrigem Einkommen auswerten zu können. Daher benötigen wir eine Einkommensangabe. Sie können sicher sein, dass Ihre Antwort nicht in Verbindung mit Ihrem Namen ausgewertet wird. Mit durchschnittlichem monatlichem Nettoeinkommen meinen wir die Summe, die sich aus Lohn, Gehalt, Einkommen aus selbstständiger Tätigkeit, Rente oder Pension ergibt. Rechnen Sie bitte auch die Einkünfte aus öffentlichen Beihilfen, Einkommen aus Vermietung und Verpachtung, Vermögen, Wohngeld, Kindergeld und sonstige Einkünfte hinzu und ziehen Sie dann Steuern und Sozialversicherungsbeiträge ab. Bitte wählen Sie die für Sie zutreffende Einkommensgruppe.

- 250 unter 250 €
- 500 251 bis 500 €
- 750 501 bis 750 €
- 1000 751 bis 1000 €
- 1250 1001 bis 1250 €

- 1500 1251 bis 1500 €
- 1750 1501 bis 1750 €
- 2000 1751 bis 2000 €
- 2500 2001 bis 2500 €
- 3000 2501 bis 3000 €
- 3001 über 3001 €

Wie gut haben Sie in der letzten Nacht geschlafen? - sehr gut - gut - ok - nicht so gut - schlecht

Und die Nacht zuvor? - sehr gut - gut - ok - nicht so gut - schlecht

Meditieren Sie manchmal? - wöchentlich mehrfach - wöchentlich einmal - gelegentlich - selten - nie

Im Allgemeinen, wie ist Ihr Energieniveau ...

am Morgen?

- sehr hoch
- hoch
- mittel
- niedrig
- sehr niedrig

am Abend?

- sehr hoch
- hoch
- mittel
- niedrig
- sehr niedrig

Nun geht es um Ihre allgemeine Lebenszufriedenheit. Wie zufrieden sind Sie gegenwärtig, alles in allem, mit Ihrem Leben? Bitte kreuzen Sie ein Kästchen auf der Skala an, wobei der Wert 0 bedeutet:

“überhaupt nicht zufrieden,” und der Wert 10: “völlig zufrieden.” Mit den Werten dazwischen können Sie Ihre Einschätzung abstufen.

überhaupt nicht zufrieden											völlig zufrieden
0	1	2	3	4	5	6	7	8	9	10	

[Weiter](#)

Further explanatory variables

Anleitung

Bitte machen Sie im Folgenden einige Angaben zu Ihren persönlichen Umständen und Ihren Einstellungen.

Anzahl meiner guten Freunde (die mir z.B. bei Problemen helfen würden): ca. _____ Personen

Anzahl meiner Bekannten (die ich z.B. zu einer Party einladen würde): ca. _____ Personen

Wieviele Stunden verbringen Sie in der Woche mit

- Arbeit gegen Entlohnung? _____ Stunden
- Engagement für Allgemeinwohl, z.B. durch Mitarbeit in gemeinnützigem Verein (z.B. Sport-, Musik- oder Schützenverein) oder auch z.B. als Schülerhilfe? _____ Stunden
- Betreuung von Angehörigen (Kinder, Eltern, anderes Familienmitglied)? _____ Stunden
- Ausübung von Hobbies (Sport, Musikinstrument, ...)? _____ Stunden
- Wenn Sie studieren: Wie viele Stunden wenden Sie normalerweise pro Woche für Ihr Studium auf? _____ Stunden
- Wie hoch ist ihr durchschnittlicher Stundenlohn während Ihrer Arbeitszeit (gerundet in €)? _____ €

Wie viel Stunden nutzen Sie schätzungsweise im Durchschnitt einen Computer oder ein Tablett pro Tag?

- mehr als 7 Stunden
- 5 bis 7 Stunden
- 3 bis 5 Stunden
- 1 bis 3 Stunden

- weniger als 1 Stunde

Beherrschen Sie das Zehnfingertippssystem ("10-Finger-Tastschreiben")?

- Sehr gut
- Gut
- Ein bisschen
- Überhaupt nicht

Haben Sie Schwierigkeiten Farben zu unterscheiden?

- Ich bin farbenblind und habe Tritanopia (Unempfindlichkeit gegenüber rotem Licht)
- Ich bin farbenblind und habe Deuteranopia (Unempfindlichkeit gegenüber grünem Licht)
- Ich bin farbenblind und habe Protanopia (Unempfindlichkeit gegenüber blauem Licht)
- Ein bisschen
- Überhaupt nicht

Weiter

Questionnaires

See Appendix C.1.1 for details on the utilized characterization questionnaires, i.e., AFS, AKSU, AMS, BFI10, BISBAS, BISS8, BNT, HAKEMP, IE4, KKI, PATs, PFS, RANK (short), SCM, SNS, SOP2, STS, UMS (goals and statements).

Anagram task *In this language proficiency task, subjects had to think up with anagrams for a given word. The instructions contained examples (for the two German words "Knaben" and "suchen") to introduce the basic idea of the task.*

The task included a trial round so that the study participants could familiarize themselves with the task. After the trial round, the subjects were asked whether they had fully understood the task. If not, they could return to the instructions page and re-read the directions for completing the task. A countdown page was

displayed before the actual task began. The subjects had 180 seconds to complete 6 anagrams (30 seconds per anagram).

The anagrams presented in the task all contained six letters (original German word and number of feasible anagrams in brackets):

- seven (“sieben,” 34),
- pigeons (“Tauben,” 38),
- to rejoice (“freuen,” 28),
- to love (“lieben,” 32),
- to do gymnastics (“turnen,” 28),
- winter (“Winter,” 26).

The page offering to re-read the instructions and the countdown page were very similar to the pages displayed in Appendix B.1.4.4 below and are, therefore, not included here.

In dieser Aufgabe geht es darum, aus einem vorgegebenen Wort neue Wörter, sogenannte Anagramme, zu bilden. Dabei dürfen **lediglich die Buchstaben** verwendet werden, die in dem **vorgegebenen Wort selbst enthalten** sind. Es müssen aber nicht alle Buchstaben des Wortes verwendet werden. Die Reihenfolge der Buchstaben im Wort kann selbstverständlich geändert werden.

Hier als Beispiel einige Anagramme für die Wörter “Knaben” und “suchen”:

<p>Knaben</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Banken </div> <div style="text-align: right; margin-right: 5px; border: 1px solid #ccc; padding: 2px;">Hinzufügen!</div> <ol style="list-style-type: none"> 1. Nabe 2. kann 3. Bank 4. Bann 5. Knabe 	<p>suchen</p> <div style="border: 1px solid #ccc; padding: 2px; margin-bottom: 5px;">Suche </div> <div style="text-align: right; margin-right: 5px; border: 1px solid #ccc; padding: 2px;">Hinzufügen!</div> <ol style="list-style-type: none"> 1. uns 2. Heu 3. neu 4. seh 5. euch 6. scheu
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Figure B.7: Exemplary anagrams

Ziel der Aufgabe ist es, in einer vorgegebenen Zeit **so viele Anagramme wie möglich pro Wort** zu finden. Als Anagramme sind dabei folgende Wörter erlaubt: **alle Wörter der deutschen Sprache, die im Duden erwähnt sind. Nicht erlaubt** sind:

- Personennamen
- Abkürzungen
- Wörter aus anderen Sprachen (ebenso englische Wörter, die im Sprachgebrauch verwendet werden, wie z.B. „run“, „bit“ oder „byte“)
- das Ausgangswort selbst, wie „Knaben“ und „suchen“ im obigen Beispiel

Weiterhin ist zu beachten:

- Es werden Ihnen **6 Wörter mit jeweils 6 Buchstaben** vorgegeben.
- **Pro Wort haben Sie 30 Sekunden** Zeit Anagramme zu finden.
- Bitte tragen Sie diese Anagramme in das vorgegebene Textfeld ein und bestätigen Sie Ihre Eingabe durch klicken des Knopfes «Hinzufügen»
– oder einfach durch das **Drücken der «Enter»-Taste** auf Ihrer Tastatur.
- Wenn Sie **Wörter mit Umlauten** eingeben wollen, verwenden Sie bitte die einzelnen Vokale, also **“ae”** statt “ä,” **“oe”** statt “ö” sowie **“ue”** statt “ü.”
- Für diese Aufgabe haben Sie **3 Minuten Zeit**.

suchen

Suche	Hinzufügen!
<ol style="list-style-type: none"> 1. uns 2. Heu 3. neu 4. seh 5. euch 6. scheu 	

Figure B.8: Trial round screen

Anagrammaufgabe: Übungsrunde.

Anagrammaufgabe: Ergebnisse Übungsrunde. *Anzahl erfolgreich gefundener Anagramme: 13*

Runde	Vorgegebenes Wort	Ihre Eingaben	richtige Anagramme
1	Knaben	Nabe, kann, Bank, Bann, Knabe, Banken	6
2	suchen	Uns, Heu, neu, seh, euch, scheu, Suche, Schnee	7
			Summe: 13

Falls Sie "False" und 0 Punkte für Ihre richtiges, letztes Anagramm sehen, dann haben Sie dieses leider erst nachdem der Timer abgelaufen ist abgesendet.

In dieser Übungsrunde haben Sie somit 13 Anagramme gefunden. Klicken Sie bitte auf " Weiter ".

Weiter

⇒ *The task itself mirrored the design of the trial round.*

Dual-2-back task *The dual-2-back task allows to test the short-term memory capacity of the subjects and is implemented according to Jaeggi et al. (2010). Colored squares appear one after the other for a short moment and at different positions on the screen. The subjects had to remember their color and position and indicate by pressing a corresponding key on the keyboard whether either or both were the same as two steps before. Not to be neglected is that the task also contains certain elements of a visual search and requires the ability to differentiate colors. Therefore, for the color selection, a color palette was used which can be distinguished by people with color difficulties.*

In dieser Aufgabe wird Ihnen nacheinander jeweils eine **quadratische Figur** angezeigt. Diese Figur erscheint jeweils nur für kurze Zeit auf dem Bildschirm und enthält ein **farbiges Kästchen**. Charakteristisch für das Kästchen sind zum einen seine **Farbe** und zum anderen seine **Position** (Feld 1-8) innerhalb der grossen quadratischen Figur.

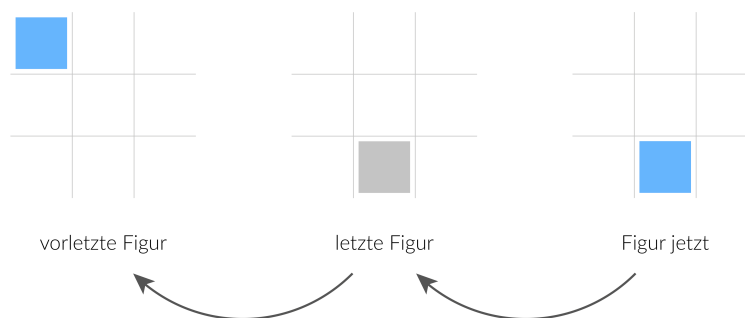


Figure B.9: **Possible colors and positions of squares:** Colors were chosen to meet requirements for color blind study participants.

Die auf dem Bildschirm angezeigten aufeinander folgende Figuren können sich also in zweierlei Hinsicht ähneln:

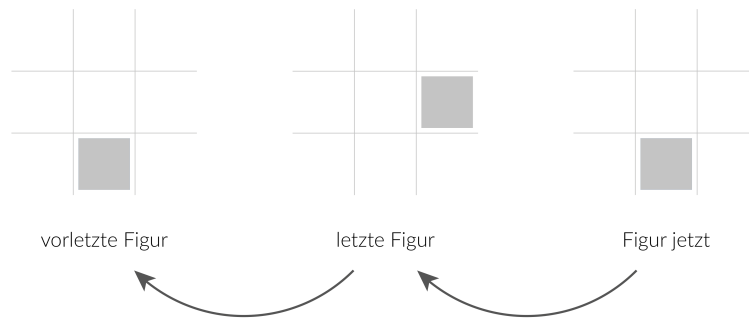
- die **Farbe** des Kästchens ist genau gleich,
- die **Position** des Kästchens innerhalb der Figur ist genau gleich.

Ihre Aufgabe ist es **schnellstmöglich durch Tastendruck** zu bestätigen, wenn die **aktuell angezeigte Figur** (also Kästchenfarbe und/oder Kästchenposition) der **zwei Bildschirme zurückliegenden Figur** ähnelt:



Drücken Sie "F" wenn die Farbe der vorletzten und der jetzigen Figur die gleiche ist

Figure B.10: **Instructions for permissible color sequence**



Drücken Sie "P" wenn die Position der vorletzten und der jetzigen Figur die gleiche ist

Figure B.11: Instructions for permissible position sequence

Das Kästchen in der aktuell angezeigten Figur kann somit dem Kästchen in der **zwei Bildschirme zurückliegenden Figur** entweder in **Farbe und Position** oder nur **in einem von beidem** entsprechen. Ziel ist es, dies so schnell wie möglich durch Tastendruck zu bestätigen (durch die Tasten **"f"** wie **"Farbe"** bzw. **"p"** wie **"Position"** auf der Tastatur). Versuchen Sie keine Fehler zu machen.

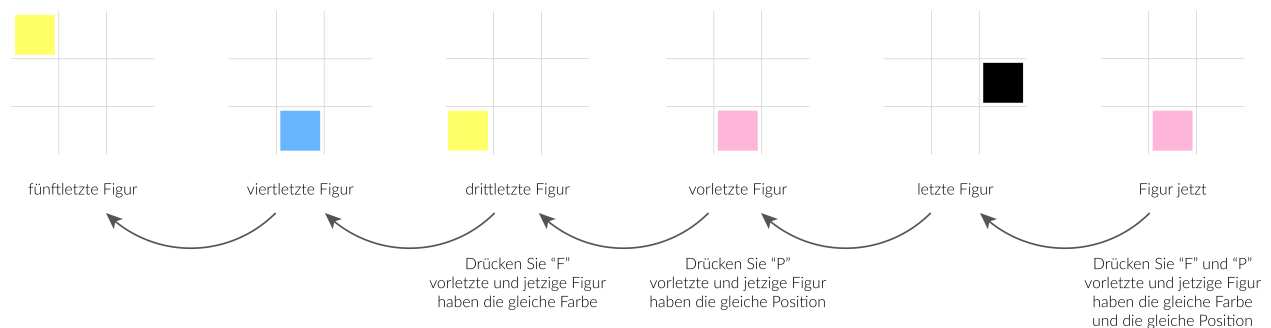
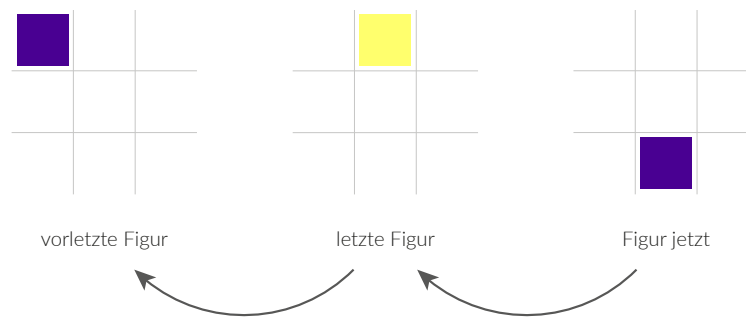


Figure B.12: Instructions for permissible color and position sequence

Gedächtnisaufgabe: Verständnisfragen. Bitte beantworten Sie die folgenden Fragen. (*correct answers in petrol*)

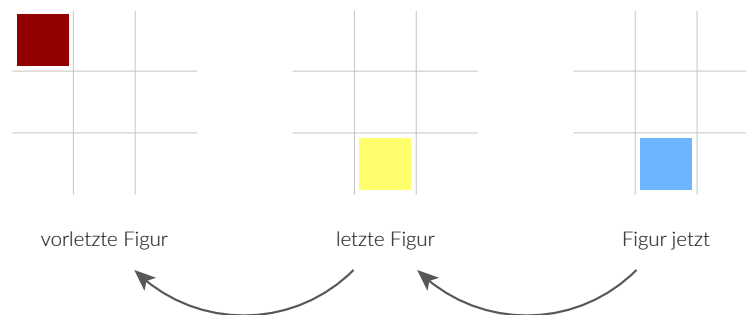
Frage 1: Welche Tasten müssen Sie zur Bewältigung der Aufgabe drücken?

1. n und m
2. f und p
3. q und p
4. 1 und 2
5. Linkspfeil und Rechtspfeil

Figure B.13: **Control question 2**

Frage 2: Welche Taste(n) sollten Sie jetzt drücken?

1. f für Farbe
2. p für Position
3. f und p für Farbe und Position
4. keine von beiden
5. Sie wissen es nicht

Figure B.14: **Control question 3**

Frage 3: Welche Taste(n) sollten Sie jetzt drücken?

1. f für Farbe
2. p für Position
3. f und p für Farbe und Position
4. keine von beiden
5. Sie wissen es nicht

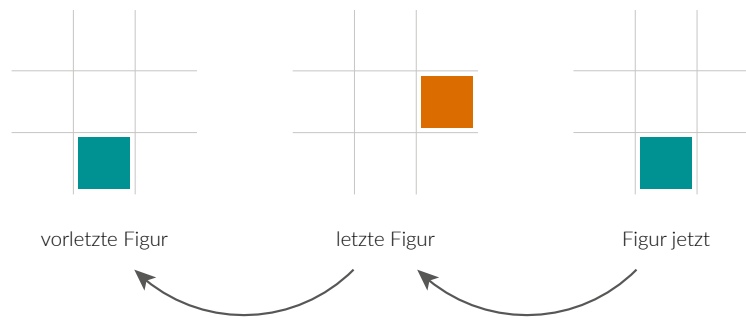


Figure B.15: **Control question 4**

Frage 4: Welche Taste(n) sollten Sie jetzt drücken?

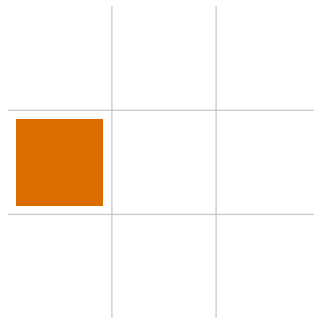
1. f für Farbe
2. p für Position
3. f und p für Farbe und Position
4. keine von beiden
5. Sie wissen es nicht

Mit dem Klick auf " **Weiter** " gelangen Sie zur Übungsrunde für die Aufgabe.

Weiter

Gedächtnisaufgabe: Übungsrunde

Drücken Sie die Taste "F,"	Drücken Sie die Taste "P,"
wenn die Farbe der vorletzten und der jetzigen Figur gleich ist.	wenn die Position der vorletzten und der jetzigen Figur gleich ist.



<i>Ihre Eingabe:</i>	Farbe	Position
		X
<i>Richtige Antworten:</i>	Farbe von Kästchen	Position von Kästchen
	1 / 4	2 / 4

Gedächtnisaufgabe: Ergebnisse der Übungsrunde. Ihre erbrachte Leistung:

Richtige Farbe der Kästchen	Richtige Position der Kästchen
9 / 20	12 / 20

Klicken Sie bitte auf "Weiter".

Weiter

⇒ The task itself mirrored the design of the trial round.

B.1.4.4 Real-Effort Tasks

Content:

- Introduction to the real-effort tasks
- Pages which are the same for all tasks
 - Confirm task-understanding
 - Count-down page (before each task begins)
 - Outside option: “Snake”
 - Relaxation page (after each task)
- Multiplication task
- Word-transcription task
- Code-transcription task
- Word-encryption task
- ab-Typing task
- Single-slider task
- Ball-catching task

The task duration for each task was five minutes. However, due to a technical limitation in the creation of the screenshots, the task instructions show a task duration of only zero minutes.

To prevent sabotage, access to the computer’s input devices has been restricted for certain tasks (e.g., the keyboard and scroll wheel have been disabled for the single-slider task; in other tasks subjects were prevented from using the arrow keys).

Introduction to the real-effort tasks

The trial round for the game “Snake” was only available to subjects in the treatment which included the game as an outside option.

Aufgabenteil - allgemeine Anweisungen

Im folgenden Teil des Experimentes werden Sie mehrere Aufgaben bearbeiten. Diese dauern **jeweils 5 Minuten**.

Für die erfolgreiche Bewältigung der vor Ihnen liegenden Aufgaben erhalten Sie unmittelbar Punkte gut geschrieben. Diese Punkte werden am Ende des Experiments mit dem Umrechnungskurs in Euro konvertiert und Ihrer anfänglichen Teilnahmeentschädigung von 5,00 € hinzugerechnet.

Ihre Auszahlung hängt somit von Ihnen ab - je nachdem wieviel Punkte Sie erzielen konnten!

Wenn Sie an Stelle einer der Aufgaben lieber eine Runde *Snake* spielen möchten, drücken Sie einfach auf den dafür vorgesehenen Knopf. Im Spiel *Snake* können Sie jedoch *keine* Punkte sammeln. Haben Sie innerhalb einer der Aufgaben zu *Snake* gewechselt können Sie **nicht** mehr zur Aufgabe zurückkehren, um diese fortzusetzen. Erst wenn die Zeit für diese Aufgabe abgelaufen ist kehren Sie wieder zum Aufgabenteil zurück, und können durch die Bewältigung der nächsten Aufgabe wieder Punkte sammeln.

Klicken Sie auf " **Weiter** ", um auf der nächsten Seite Verständnisfragen zum Aufgabenteil zu beantworten.

Weiter

Figure B.16: General instructions for the tasks

Aufgabenteil - Verständnisfragen

Bitte beantworten Sie die folgenden Fragen.

Frage 1: Wie lange haben Sie zur Bearbeitung der einzelnen Aufgaben Zeit?

5 Minuten

Frage 2: Dürfen Sie so viele Punkte erzielen, wie Sie möchten?

Ja

Frage 3: Hängt Ihre Auszahlung am Ende des Experiments von den in dieser Aufgabe erzielten Punkten ab?

Ja. Die Punkte werden mit dem Umrechnungskurs in Euro |

Frage 4: Sie haben sich entschieden eine der Aufgaben abzubrechen, um eine Runde *Snake* zu spielen. Können Sie anschließend wieder zu dieser Aufgabe zurückkehren, um sie fortzusetzen?

Nein, nicht mehr während dieser Aufgabe. Erst wenn die

Klicken Sie auf "Weiter" um auf der nächsten Seite zu einem Probelauf von *Snake* zu gelangen. Dort können Sie sich mit der Alternative zum Aufgab lösen für kurze Zeit vertraut machen, bevor der eigentliche Teil beginnt.

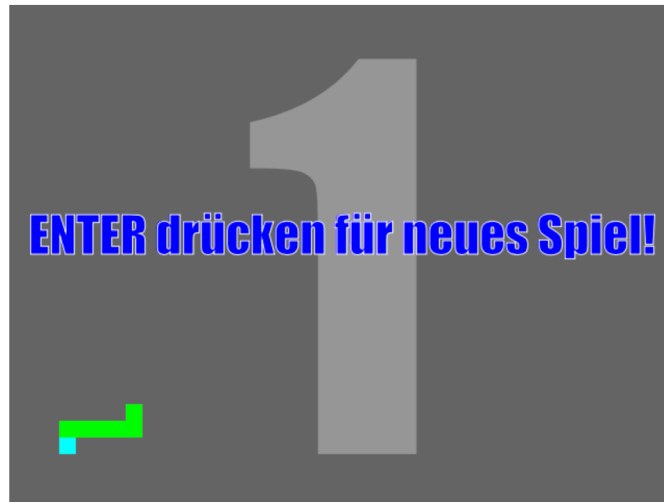
Falls Sie das Gefühl haben, dass Ihre Tastatur nicht funktioniert, prüfen Sie bitte die "Caps Lock" Taste: ist sie gedrückt (aktiviert) drücken Sie die Taste bitte erneut, um diese zu deaktivieren. Sollten weiterhin Probleme auftreten melden Sie sich bitte beim Experimentator mittels Handzeichen.

Weiter

Figure B.17: Control questions

Snake - Übungsrunde

Verbleibende Zeit für diese Seite. ⌚ 0:01



Zum Steuern der Schlange können Sie die **Pfeiltasten** oder die Tastenkombinationen **WASD** sowie **HJKL** verwenden.

Sammeln Sie den eingeblendeten Futternapf ein, damit die Schlange wächst und sich schneller bewegt.

Ist der nächste Futternapf einmal nicht sichtbar, so sind ist er unter der Schlange versteckt. Bewegen Sie diese zunächst um ihn freizulegen.

Figure B.18: **Snake trial round**: only available to the subjects in the treatment with outside option.

Aufgabenteil - Jetzt geht's los!

Jetzt beginnt der Aufgabenteil – **fühlen Sie sich frei, so viele Punkte wie möglich in den Aufgaben zu erzielen!**

Klicken Sie auf "Weiter" um zur ersten Aufgabe zu gelangen.

Weiter

Figure B.19: **Ready for the tasks**

Pages which are the same for all tasks



Figure B.20: **Confirm task-understanding:** Before beginning to complete a task, subjects had to confirm that they understood the task.

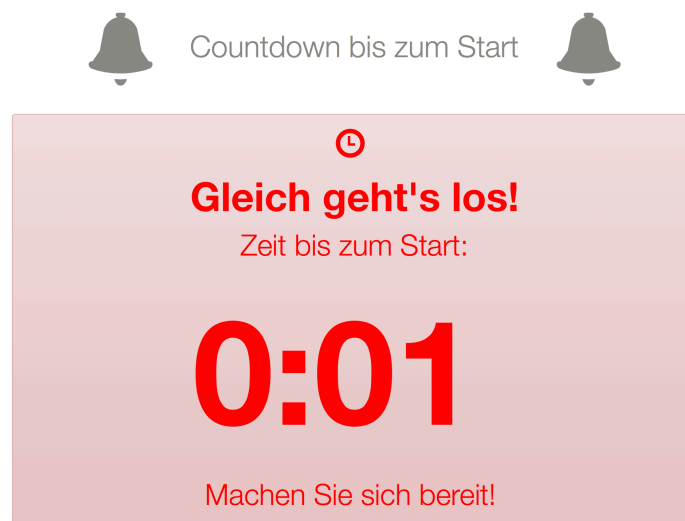


Figure B.21: **Count-down timer:** To make sure that subjects were focused and did not miss the start of a task, a five-second countdown timer was displayed on their screen before the start of each task. This was instrumental in ensuring that the effort allocation actually covered the full five minutes.

SNAKE - hier klicken um Snake zu spielen

Figure B.22: **Button to switch to outside option “Snake”**: Subjects in the treatment “with outside option” could choose to play the well-known computer game “Snake” as an alternative activity. However, once they switched to the outside option, they could not return to the current task. When the duration of effort provision for the current task (five minutes) had expired, subjects were informed about the number of points they had collected in the task. Afterwards they had the opportunity to collect points once again in the next task. The button to switch to Snake was displayed below each task, allowing subjects to abort the task and switch to the game at any time during the effort provisioning period.



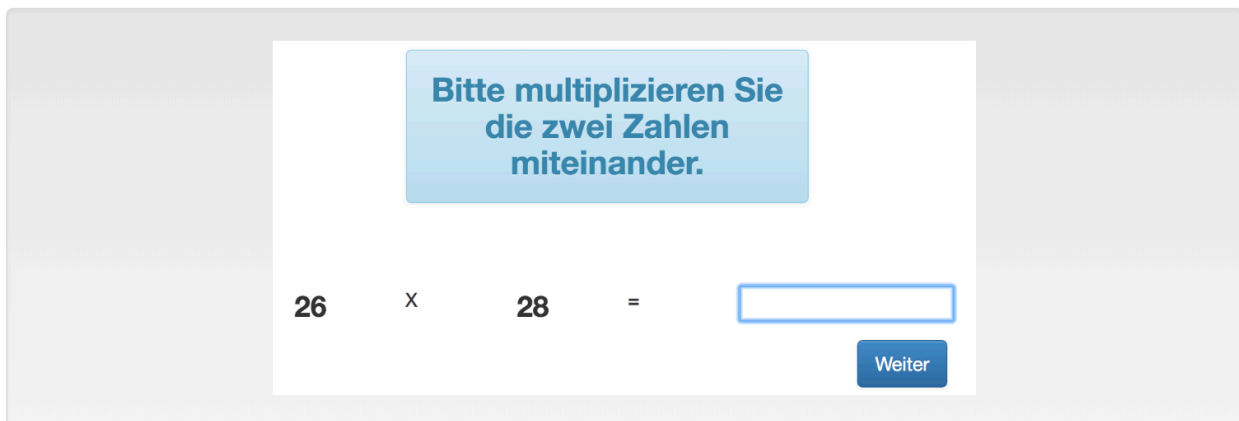
Figure B.23: **Break-page**: After each task, subjects were encouraged to take a short break to stretch, relax and recover in order to recharge their batteries and regain their focus and concentration.

Multiplication task (Dohmen & Falk, 2011)

Multiplikationsaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Lösen von Rechenaufgaben Geld verdienen. Es werden Ihnen zwei zweistellige Zahlen vorgegeben, die Sie miteinander multiplizieren müssen. Für **richtige Lösungen erhalten Sie Punkte**, für **falsche Lösungen** werden Ihnen nach **dreimaliger Eingabe Punkte abgezogen**. Sie können das bereitgelegte Papier und den Kugelschreiber bei der Rechnung zu Hilfe nehmen – **nicht aber Ihr Smartphone, Ihren Taschenrechner oder andere Hilfsmittel!**

Beispielhafte Rechenaufgabe (Das Ergebnis "728" muss noch eingetragen und bestätigt werden):



Bitte multiplizieren Sie die zwei Zahlen miteinander.

26 x 28 =

Weiter

Ziel der Aufgabe ist es, in einer vorgegebenen Zeit **so viele Multiplikationsrechnungen wie möglich korrekt zu lösen**.

Dabei ist zu beachten:

- Sie haben **drei Versuche das Ergebnis einer Rechnung einzutragen**.
- Bitte tragen Sie Ihre Lösung für die Rechnung in das vorgegebene Feld ein und bestätigen Sie Ihre Eingabe durch klicken des Bestätigungsknopfes «Absenden!»
- Zur Rechnung können Sie das zur Verfügung bereitgestellte Papier verwenden, **nicht aber Ihr Smartphone, Ihren Taschenrechner oder andere Hilfsmittel (führt zum Ausschluss aus dem Experiment und allen Zahlungen)!**
- Sie erhalten **225 Punkte** für jede **korrekt gelöste Multiplikationsrechnung**.
- nach drei missglückten Versuchen/**falschen Lösungen** werden Ihnen **225 Punkte** für diese Rechnung abgezogen.
- Sollten Sie am Ende dieser Multiplikationsaufgabe mehr falsche als richtige Lösungen eingegeben und somit einen Verlust erzielt haben, so wird Ihr Verdienst in dieser Aufgabe auf 0 gesetzt.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Figure B.24: Instructions for the task

Beispiel:

Nehmen Sie an, dass Sie bei drei Rechenaufgaben das richtige Ergebnis eintragen, wobei Sie bei einer drei Rechenaufgaben zwei Versuche brauchen, bis Sie das richtige Ergebnis eingeben. Bei einer vierten Aufgabe kommen Sie jedoch nicht zum richtigen Ergebnis und verrechnen sich dreimal.

korrekt gelöste Rechenaufgaben	fehlerhaft gelöste Rechenaufgaben	Anzahl an gesammelten Punkten
3 Rechenaufgaben	1 Rechenaufgabe	
+ 3 x 225 Punkte	- 1 x 225 Punkte	= 2 x 225 Punkte

Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.
In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure B.25: Instructions for the task

Multiplikationsaufgabe

Verbleibende Zeit für diese Seite. ⌚ 0:03

Bitte multiplizieren Sie
die zwei Zahlen
miteinander.

$$32 \times 80 = \input{text}$$

Absenden!

SNAKE - hier klicken um Snake zu spielen

Figure B.26: Multiplication task

Multiplikationsaufgabe: Ergebnisse

Anzahl effektiv, erfolgreich gelöster Rechenaufgaben: 0

korrekt gelöst	fehlerhaft gelöst	effektiv gelöst	Anzahl an gesammelten Punkten
0	0	0	0 Punkte

In dieser Aufgabe haben Sie somit **0 Punkte** gesammelt (225 Punkte pro korrekt gelöster Aufgabe abzüglich 225 für jede dreimal falsche eingegebene Lösung). Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

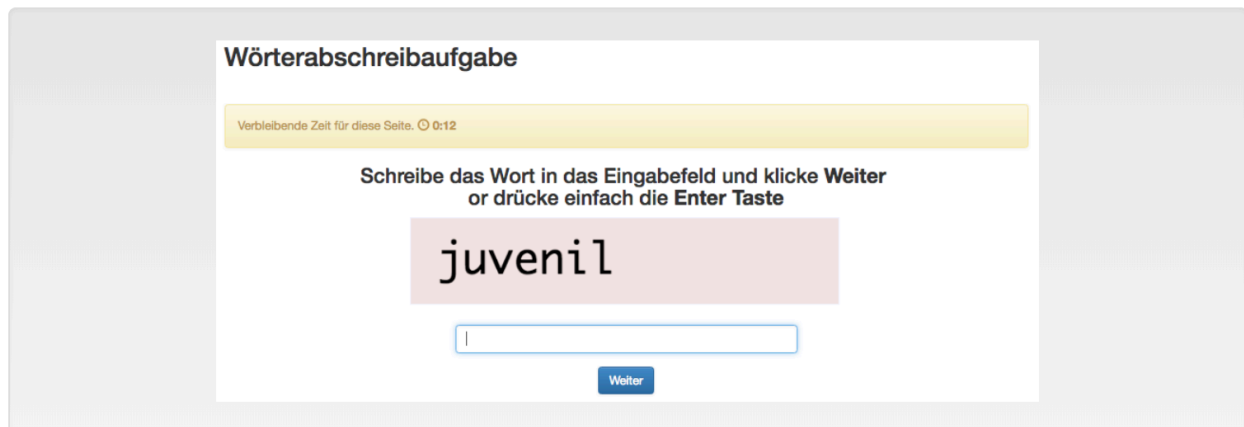
Figure B.27: Results page

Word-transcription task (modified from Kephart, 2017)

Wörterabschreibeaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Abschreiben von Wörtern Geld verdienen. Es wird Ihnen ein Bild mit einem Wort angezeigt. Ihre Aufgabe ist es nun **das Wort korrekt** in das **Eingabefeld einzutragen**.

Hier als Beispiel für das Wort "juvenil":



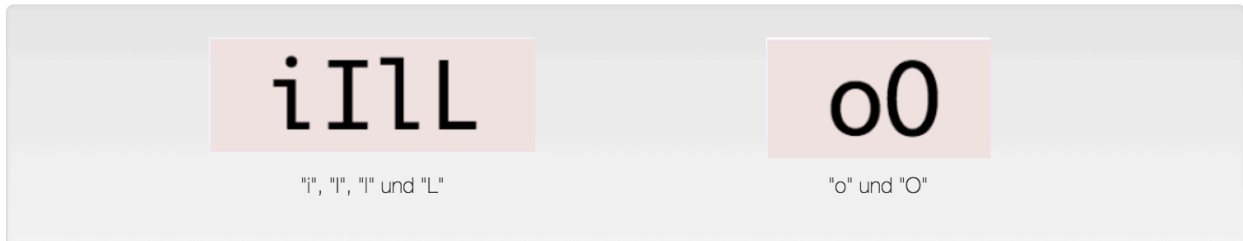
Das auf dem Bild angezeigte Wort muss exakt in das Eingabefeld abgeschrieben werden, um als korrekt anerkannt zu werden (**Achtung Groß- und Kleinschreibung!**).

Ziel der Aufgabe ist es, in einer gegebenen Zeit **so viele Buchstaben- und Zifferkombinationen wie möglich** einzutragen. Dabei ist zu beachten:

- Sie erhalten **32 Punkte** für jedes **korrekt eingetragene Wort**.
- Für **fehlerhaft eingetragene Wörter** erhalten Sie **keine Punkte**.
- Bitte tragen Sie die Wörter in das vorgegebene Feld ein und bestätigen Sie Ihre Eingabe durch klicken des Bestätigungsknopfes «Weiter» – oder einfach durch das **Drücken der «Enter»-Taste** auf Ihrer Tastatur.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Figure B.28: Instructions for the task

Die Schriftart der Wörter erlaubt Ihnen die Buchstaben leichter von einander zu unterscheiden, wie z.B.:



Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.
In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure B.29: Instructions for the task

Wörterabschreibeaufgabe

Verbleibende Zeit für diese Seite. 🕒 0:14

Schreibe das Wort in das Eingabefeld und klicke **Weiter**
oder drücke einfach die **Enter Taste**

abgeschieden

Weiter

SNAKE - hier klicken um Snake zu spielen

Figure B.30: Word-transcription task

Wörterabschreibeaufgabe: Ergebnisse

Anzahl korrekt eingetragener Wörter: 3

Runde	Richtiges Wort	Eingetragenes Wort	Wahr?	Anzahl Richtige
1	abgeschieden	abgeschieden	True	1
2	abgründig	abgründig	True	1
3	achtsam	achtsam	True	1
4	angelegentlich	angelegen	False	0

Falls Sie "**False**" und 0 Punkte für Ihre richtige, letzte Buchstaben- und Zifferkombinationen sehen, dann haben Sie diese leider erst nachdem der Timer abgelaufen ist abgesendet.

In dieser Übungsrunde haben Sie somit **96 Punkte** gesammelt (32 Punkte pro korrekt eingetragem Wort). Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure B.31: Results page

Code-transcription task (Kephart, 2017)

Buchstaben- und Zahlenabschreibeaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Abschreiben von Buchstaben- und Zifferkombinationen Geld verdienen. Es wird Ihnen ein Bild mit Buchstaben und Ziffern, die zusammen ein "Code" ergeben, angezeigt. Ihre Aufgabe ist es nun **den Code korrekt** in das **Eingabefeld einzutragen**.

Hier als Beispiel für den Code "Fa 8 L":

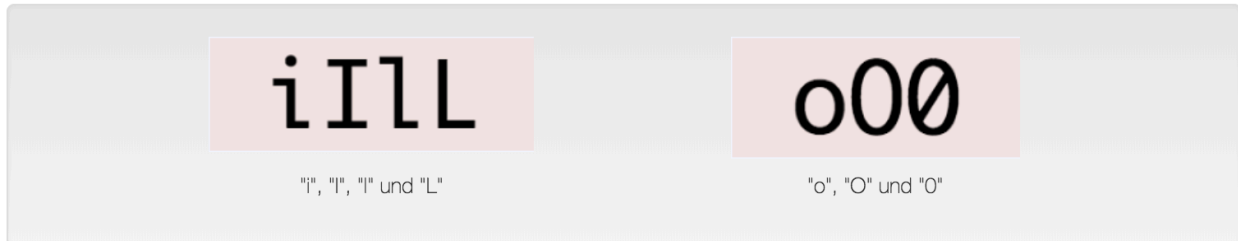
Die auf dem Bild angezeigte Buchstaben- und Zifferkombinationen muss exakt in das Eingabefeld abgeschrieben werden, um als korrekt anerkannt zu werden (**Achtung Groß- und Kleinschreibung!**).

Ziel der Aufgabe ist es, in einer gegebenen Zeit **so viele Buchstaben- und Zifferkombinationen wie möglich** einzutragen. Dabei ist zu beachten:

- Sie erhalten **30 Punkte** für **korrekt eingetragene Codes**.
- Für **fehlerhaft eingetragene Codes** erhalten Sie **keine Punkte**.
- Bitte tragen Sie die Buchstaben- und Zifferkombinationen in das vorgegebene Feld ein und bestätigen Sie Ihre Eingabe durch klicken des Bestätigungsknopfes «Weiter» – oder einfach durch das **Drücken der «Enter»-Taste** auf Ihrer Tastatur.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Figure B.32: Instructions for the task

Die Schriftart der Codes erlaubt Ihnen Buchstaben und Zahlen leichter von einander zu unterscheiden, wie z.B.:



Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.
In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure B.33: Instructions for the task

Buchstaben- und Zahlenabschreibeaufgabe

Verbleibende Zeit für diese Seite. ⌚ 0:09

Schreibe den Code in das Eingabefeld und klicke **Weiter**
oder drücke einfach die **Enter Taste**

uIzR

Weiter

SNAKE - hier klicken um Snake zu spielen

Figure B.34: Code-transcription task

Buchstaben- und Zahlenabschreibaufgabe: Ergebnisse

Anzahl korrekt eingetragener Buchstaben- und Zifferkombinationen: 0

Runde	Richtiger Code	Eingetragener Code	Wahr?	Anzahl Richtige
1	ulzR		False	0

Falls Sie "**False**" und 0 Punkte für Ihre richtige, letzte Buchstaben- und Zifferkombinationen sehen, dann haben Sie diese leider erst nachdem der Timer abgelaufen ist abgesendet.

In dieser Übungsrunde haben Sie somit **0 Punkte** gesammelt (30 Punkte pro korrekt eingetragem Code). Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure B.35: Results page

Word-encryption task (Erkal et al., 2011)

Buchstabenverschlüsselungsaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Verschlüsseln von Buchstaben Geld verdienen. Es werden Ihnen drei Buchstaben, die zusammen eine "Buchstabenabfolge" ergeben, sowie eine Verschlüsselungstafel vorgegeben. Ihre Aufgabe ist es nun, **aus der Verschlüsselungstafel die passenden Codes herauszusuchen** und diese **beim jeweiligen Buchstaben einzutragen**.

Hier als Beispiel für die Buchstabenabfolge "W M T":

Wort:	W	M	T
Code:	<input type="text" value="876"/>	<input type="text" value="574"/>	<input type="text"/>

hier „911“ eintragen

W	G	X	P	E	F	R	B	Q	O	N	C	U	A	T	L	H	K	M	S	J	V	I	Z	Y	D
876	691	791	329	846	270	738	415	187	417	460	186	917	876	911	318	443	218	574	445	143	333	352	780	940	308

Die Codes für die ersten beiden Buchstaben wurden bereits eingetragen. Der dritte Code (911) für den dritten Buchstaben T der Buchstabenabfolge muss noch eingetragen werden.

Ziel der Aufgabe ist es, in einer vorgegebenen Zeit **so viele Wortverschlüsselungen wie möglich** durchzuführen.

Dabei ist zu beachten:

- Bitte tragen Sie die Codes in die vorgegebenen Felder ein und bestätigen Sie Ihre Eingabe durch klicken des Knopfes «Codes absenden»
- Mit der **Tabulatortaste** auf Ihrer Tastatur können Sie schneller zwischen den Eingabefeldern wechseln.
- Sie erhalten **120 Punkte** für **korrekt verschlüsselte Buchstabenabfolgen**.
- für **fehlerhaft verschlüsselte Buchstabenabfolgen** erhalten Sie **keine Punkte**.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.

In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure B.36: Instructions for the task

Buchstabenverschlüsselungsaufgabe

Verbleibende Zeit für diese Seite. ⌚ 0:09

Wort Nummer: 1
Anzahl richtiger Verschlüsselungen: 0

Wort: C V E
Code:

T Z H D X Y W P G B E K M U S I A O N L C V R J Q F
311 448 669 928 580 559 451 865 729 426 913 330 450 714 706 185 106 125 581 458 983 723 692 882 192 986

Codes absenden

SNAKE - hier klicken um Snake zu spielen

Figure B.37: Word-encryption task

Buchstabenverschlüsselungsaufgabe: Ergebnisse

Anzahl korrekt verschlüsselte Buchstabenabfolgen: 0

korrekt verschlüsselte Buchstabenabfolgen	fehlerhaft verschlüsselte Buchstabenabfolgen	Anzahl an gesammelten Punkten
0	0	0 Punkte

In dieser Aufgabe haben Sie somit **0 Punkte** gesammelt (Punkte pro korrekt gelöster Aufgabe). Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure B.38: Results page

ab-Typing task (Berger & Pope, 2011)

"A oder B" Tastendruckaufgabe: Anleitung

Auf der nächsten Seite können Sie in einer einfachen Tastendruck-Aufgabe Geld verdienen. Ziel dieser Aufgabe ist es, **abwechselnd für 0 Minuten die Tasten "a" und "b"** auf Ihrer Tastatur so schnell wie möglich zu drücken. Jedes Mal erhalten Sie einen Punkt, wenn Sie erfolgreich die Taste "a" und dann alternierend die Taste "b" betätigen. Beachten Sie, dass Ihre Leistung nur dann mittels Punkte belohnt wird, wenn Sie die Taste abwechselnd betätigen: D.h. es werden **keine** Punkte erzielt, wenn Sie einfach nur die Taste "a" oder nur die Taste "b" drücken.

Das folgende Beispiel zeigt, dass als nächstes die Taste "b" gedrückt werden muss, um den nächsten Punkt zu erhalten:

a	b
42	41

Ziel der Aufgabe ist, es in einer vorgegebenen Zeit **so oft wie möglich abwechselnd die Tasten a und b** zu drücken. Dabei ist zu beachten:

- Sie erhalten jedes Mal **1,8 Punkte** für **einmal alternierend a und b drücken**.
- Für **fehlerhafte Eingaben** (wie "immer a" oder andere Tasten drücken) erhalten Sie **keine Punkte**.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe. In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

Weiter

Figure B.39: Instructions for the task

"A oder B" Tastendruckaufgabe

Verbleibende Zeit für diese Seite. ⌚ 0:02

a	b
13	12

SNAKE - hier klicken um Snake zu spielen

Figure B.40: ab-Typing task

"A oder B" Tastendruckaufgabe: Ergebnisse

Ihre erreichte Gesamtpunktzahl: 25 (Anzahl abwechselnder Tastendrucke)

In dieser Aufgabe haben Sie somit **45 Punkte** gesammelt (bei 1,8 pro abwechselnder Tastendrucke).
Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure B.41: Results page

Single-slider task

See Appendix A.1 for a detailed description of the task.

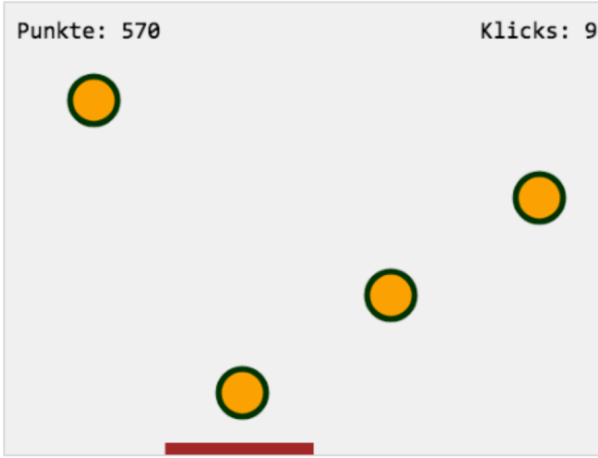
Ball-catching task (Gächter et al., 2016)

Balleinfangaufgabe: Anleitung

In der folgenden Aufgabe können Sie durch das Einsammeln von Bällen Geld verdienen. Von oben fallen kontinuierlich Bälle herunter. Ihre Aufgabe ist es nun **mit Hilfe eines verschiebbaren Korbes** die **Bälle einzufangen** wobei jede Bewegung des Korbes **mit Kosten verbunden ist**.

Durch klicken der Knöpfe "LINKS" und "RECHTS" kann der Korb (**brauner Balken**) verschoben werden, um die Bälle einzufangen:

Eingesammelte Bälle: 19	bisherige Kosten durch Klicks: 135	
-----------------------------------	----------------------------------------------	--

<p>Punkte: 570 Klicks: 9</p>  <p style="text-align: center;">Korb bewegen nach</p> <div style="display: flex; justify-content: center; gap: 20px;"> <= LINKS RECHTS => </div>	<p>Gewinn pro Ball: 30 Punkte</p> <p>Kosten pro Klick: 15 Punkte</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------

Ziel der Aufgabe ist es, in einer vorgegebenen Zeit **so viele Bälle wie möglich** einzufangen, wobei die **Kosten für jede Bewegung des Korbes zu berücksichtigen sind**.

Figure B.42: Instructions for the task

Bitte beachten Sie:

- Sie erhalten **30 Punkte** für **jeden eingesammelten Ball**;
- gleichzeitig **kostet Sie jede Bewegung des Korbes (nach rechts oder links) 15 Punkte pro Klick**;
- Ihr Verdienst entspricht der Summe von gesammelten Punkten für eingefangene Bälle abzüglich der Kosten für das Bewegen des Korbes.
- Sollten Sie am Ende dieser Balleinfangaufgabe mehr Kosten durch Korbbewegungen als Punkte durch eingefangene Bälle erreicht und somit einen Verlust erzielt haben, so wird Ihr Verdienst in dieser Aufgabe auf 0 gesetzt.
- Für diese Aufgabe haben Sie **0 Minuten Zeit**.

Mit dem Klick auf "**Weiter**" gelangen Sie zur Übungsrunde für die Aufgabe.
In der Übungsrunde gesammelte Punkte zählen *nicht* zu Ihrer späteren Auszahlung.

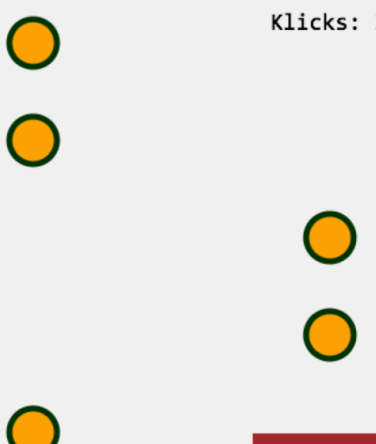
Weiter

Figure B.43: Instructions for the task

Balleinfangaufgabe

Verbleibende Zeit für diese Seite. ⌚ 0:04

Eingesammelte Bälle: 3	bisherige Kosten durch Klicks: 45	
----------------------------------	---------------------------------------------	--

Punkte: 90		Klicks: 3
-------------------	--------------------------------------------------------------------------------------	------------------

Korb bewegen nach	
<div style="display: flex; justify-content: space-around;"> <= LINKS RECHTS => </div>	
<div style="border: 1px solid gray; padding: 5px; display: inline-block;">SNAKE - hier klicken um Snake zu spielen</div>	

Gewinn pro Ball:
30 Punkte

Kosten pro Klick:
15 Punkte

Figure B.44: Ball-catching task

Balleinfangaufgabe

Effektiv gesammelte Punkte: 75 Punkte				
Anzahl eingefangene Bälle	Gesamtzahl Punkte durch Bälle	Anzahl getätigte Klicks	Gesamtkosten durch Klicks	Punktstand
4	120	3	45	75

In dieser Aufgabe haben Sie somit **75 Punkte** gesammelt (30 Punkte pro eingesammeltem Ball abzüglich 15 Punkte für jeden getätigten Klick).

Diese werden am Ende des Experimentes mit dem Umrechnungskurs in Euro konvertiert und an Sie ausbezahlt.

Weiter

Figure B.45: Results page

Real-effort task survey See Appendix B.2.

B.1.4.5 Final Questionnaires

Personal hit-list

Persönliches Ranking

Bitte denken Sie zunächst allgemein an die Tätigkeit, die Sie **zur Zeit am liebsten von allen machen** und von der Sie gar nicht genug bekommen können. Schreiben Sie diese bitte in das Feld mit dem lachenden Smiley 😄 ein.

Denken Sie jetzt bitte an die Tätigkeit, die Sie **zur Zeit am allerwenigsten von allen** mögen. Am liebsten würden Sie alles andere lieber tun als ausgerechnet diese Tätigkeit - **aber Sie müssen sie dennoch machen**. Tragen Sie diese bitte in das Feld mit dem traurigen Smiley 😞 ein.

Schreiben Sie nun bitte noch in das Feld in der Mitte eine Tätigkeit, die Sie weder besonders gern noch besonders ungern machen.

An diesen **geliebten, neutralen** und **ungeliebten** Tätigkeiten (visualisiert durch die Smileys) orientieren Sie sich beim Ankreuzen der folgenden Tabelle.

Arbeitsanweisung:

Im bisherigen Teil der Studie haben Sie mehrere Aufgaben erledigt. Bewerten Sie diese im Folgenden anhand **Ihres persönlichen Rankings**.

Mit den Kästchen zwischen den Endpunkten können Sie Ihre Antwort abstufen

	😊		😐		😞
	Fahrrad fahren		abwaschen		Fahrrad reparieren
Multiplikationsaufgabe	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Wortabschreibeaufgabe	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Buchstaben- und Zahlenabschreibeaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Wortverschlüsselungsaufgabe	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
ab-Tastendruckaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
"Links- & Rechts" Schiebaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Ballfangaufgabe	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>

Weiter

Figure B.46: Personal hit-list

Table B.7: **Final motivational questionnaire:** The self-assessment survey explicitly asks the study participants about their motivations for putting effort into the tasks. The survey items include the incentive effect i) of the subjects' enjoyment of the tasks in general, ii) of monetary incentives on them, iii) of wanting to comply with (any) expectations of the experimenter, and iv) induced by the subjects' self-image (to meet peer- and self-demand). The response scale has seven levels and ranges from (1) "do not agree at all" to (7) "agree fully." Since the items of the scale address different motivations, they cannot be aggregated into a single measure and have to be considered separately. The first item focuses specifically on the tasks performed in the study. The implication is that the subjects' responses to this item, i.e., their degree of "joy in completing tasks," cannot necessarily be transferred or generalized to other tasks such as washing dishes or repairing bicycles. The subjects' answers to the remaining items are of a more general nature and, therefore, likely valid in different experimental situations.

ID	Motive	Item
1	Joy in task-fulfilment	The fulfillment/completion of the task itself was already fun.
2	Earn a lot of money	I wanted to make/earn as much money as possible.
3	Meet experimenter expectations	I wanted to meet the expectations of the experimenter, who certainly expected my commitment.
4	Diligent attitude	I am a hardworking person and therefore wanted to show commitment accordingly.
5	Do not like Snake	I do not really like the game Snake.
6	Snake more fun than tasks	I had more fun playing Snake than completing the tasks.

Final motivational questionnaire

Fragebogen zu getätigten Aufgaben

Sie haben im vergangenen Teil der Studie mehrere Aufgaben erledigt, was mit einem gewissem Einsatz verbunden war.
Was war für Sie bei der Erledigung der Aufgaben Ihre Hauptmotivation?

*Mit den Kästchen zwischen den Endpunkten
können Sie Ihre Antwort abstimmen*

	Stimme überhaupt nicht zu						Stimme voll zu
Die Erfüllung der Aufgabe an sich hat mir bereits Spaß gemacht hat.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich wollte so viel Geld wie möglich verdienen.	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich wollte die Erwartungen des Experimentators erfüllen, der sicherlich Einsatz von mir erwartet hat.	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich bin ein fleißiger Mensch und wollte dementsprechend auch gerne Einsatz zeigen.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ich mag das Spiel Snake nicht wirklich gerne.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Mir hat es mehr Spass gemacht Snake zu spielen als die Aufgaben zu erledigen.	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Weiter

Figure B.47: **Final motivational questionnaire:** Screenshot of the German version implemented in the study.

B.1.4.6 End of the Experiment

Final questionnaire on study participation

Abschlussfragebogen. Abschließend möchten wir Sie bitten noch einige Angaben zu Ihrer Studienteilnahme zu machen.

Bitte beantworten Sie die folgenden Fragen.

Haben Sie die Instruktionen in diesem Experiment immer verstanden?

- überhaupt nicht
- in sehr geringem Maße
- in gewissem Maße
- in hohem Maße
- in sehr hohem Maße

Aus welchem Grund haben Sie an dieser Studie teilgenommen?

- Persönliche Interesse an der experimentellen Wirtschaftsforschung
- Neugier (allgemein)
- Verdienst/Entlohnung
- Sonstiges

Sonstige Gründe zur Teilnahme bitte hier eintragen: _____

Waren Ihnen bereits Teile dieses Experimentes bekannt?

- Nein, noch nicht
- Ja

Bitte tragen Sie die bekannten Teile stichwortartig ein: _____

In den nächsten beiden Fragen möchten wir Sie bitten uns zu helfen die Qualität der erhobenen Daten besser einzustufen zu können. Bitte machen Sie daher eine ehrliche Selbsteinschätzung zu Ihrer Studienteilnahme. Ihre Auszahlung hängt nicht von Ihren Antworten ab.

Wie konzentriert und fokussiert waren Sie während dieser Studie?

- überhaupt nicht
- in sehr geringem Maße
- in gewissem Maße
- in hohem Maße
- in sehr hohem Maße

Wie ehrlich waren Sie mit Ihren Antworten während dieser Studie?

- überhaupt nicht
- in sehr geringem Maße
- in gewissem Maße
- in hohem Maße
- in sehr hohem Maße

Wenn Sie den Experimentatoren noch etwas mitteilen möchten, können Sie dies im folgenden Feld eintragen.

Payments

Auszahlung

Umrechnung Ihrer gesammelten Punkte

Sie sind am Ende des Experimentes angekommen. Ihre gesammelten Punkte wurden mit dem folgenden, angekündigten Kurs umgerechnet:

100 Punkte = 0,10 €

1 Punkt = 0,001 €

Ihre persönliche Auszahlung

Zunächst erhalten Sie Ihre Teilnahmeentschädigung von 5,00 €.

Aufgabe	gesammelte Punkte
Multiplikationsaufgabe	0 Punkte
Wortabschreibaufgabe	96 Punkte
Buchstaben- und Zahlenabschreibaufgabe	0 Punkte
Wortverschlüsselungsaufgabe	0 Punkte
ab-Tastendruckaufgabe	45 Punkte
Schieberaufgabe	20 Punkte
Ballfangaufgabe	60 Punkte

Als Dankeschön für Ihre Teilnahme und Ihre Bemühungen beim Ausfüllen der Fragebögen erhalten Sie zusätzlich eine Gutschrift von 1000 Punkte. Insgesamt haben Sie somit 1221 Punkte akkumuliert, was umgerechnet 1,22 € entspricht.

Aufgerundet auf 10 Cent beträgt **Ihre persönliche, finale Auszahlung** daher **6,3 €**.

Bitte tragen Sie die genannten Werte nun in das **Quittungsformular** ein.

Weiter

Figure B.48: **Payment screen:** The accumulated score was converted into Euros according to the conversion rate announced at the beginning of the experiment. Participants were kindly asked to complete and sign the provided payment slip.

Vielen Dank!

Wir danken Ihnen für Ihre Teilnahme am Experiment.

Sie können nun auf die von Ihnen mitgebrachte Lektüre zurückgreifen, um die Zeit zu überbrücken bis alle Teilnehmer das Experiment abgeschlossen haben. **Zur Erinnerung: Die Verwendung von Mobiltelefonen und anderen elektronischen Geräten ist während des gesamten Experimentes nicht gestattet.**

Alle Teilnehmer des Experimentes werden gemäss ihrer Tischnummer nacheinander zur Auszahlungsstelle gebeten. Wir bitten Sie daher noch um ein wenig Geduld, bis Sie an der Reihe sind.

Bitte zur Auszahlungsstelle mitbringen:

- das ausgefüllte **Quittungsformular**
- Ihre **Platzkarte**
- den **Kugelschreiber**
- Ihren **Studienausweis**.

Die zur Verfügung gestellte Packung Ohropax können Sie gerne behalten.

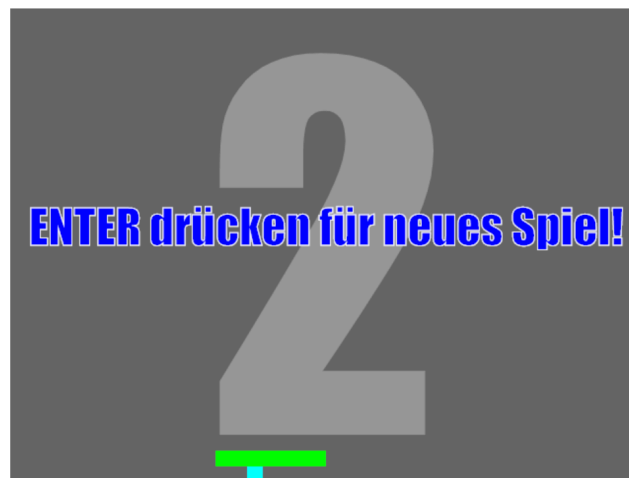
Wir wünschen Ihnen noch einen schönen Tag und würden uns freuen, wenn Sie wieder einmal an einer unserer experimentellen Studien teilnehmen würden.

Falls Sie zum Überbrücken der Zeit noch eine Runde Snake spielen möchten, klicken Sie einfach "**Weiter**".

Weiter

Figure B.49: “Thank you for your participation”-screen: Express thanks for taking part in the study. The participants were kindly asked to remain seated until all participants had completed the experiment.

Vielen Dank für Ihre Studienteilnahme!



Zum Steuern der Schlange können Sie die **Pfeiltasten** oder die Tastenkombinationen **WASD** sowie **HJKL** verwenden.

Sammeln Sie den eingeblendeten Futternapf ein, damit die Schlange wächst und sich schneller bewegt.

Ist der nächste Futternapf einmal nicht sichtbar, so sind ist er unter der Schlange versteckt. Bewegen Sie diese zunächst um ihn freizulegen.

Figure B.50: **Waiting screen with option to play Snake:** The experiment ended when all subjects had finished. To bridge the time until (individual) payment, the study participants had the opportunity to play the game “Snake” again.

B.2 Real-Effort Task Survey: Survey Structure and German Version

The composition of items of the real-effort task survey is not the result of successive item reduction based on a large pool of candidate items. Instead, the survey development followed an exploratory approach based on the [design criteria and practices](#) presented in Chapter 2, which synthesize the literature on real effort. These propose adjustments in the design and implementation of tasks to curb effort provision motivated by activity-related and undesirable purpose-related incentives. The design practices that address the former are intended to counter curiosity and enjoyment in task performance; those addressing the latter seek to reduce or mitigate the influence of various effort-inducing consequences. The motivational items in the survey capture the fulfillment of these design practices. Of the variety of proposed design practices, the survey focuses on the task-dependent practices because the independent ones can be implemented in any task. In addition, the survey contains two items to assess how physically and mentally demanding a task is. Therefore, the survey covers four dimensions: Design practices that predominantly address 1) activity-related incentives or 2) purpose-related incentives, as well as 3) physical effort and 4) mental effort. Figure [B.51](#) depicts the allocation of survey items to these dimensions on the far right. A factor analysis largely confirms this grouping structure underlying the questionnaire (see Appendix [B.3.3.1](#)).

Figure [B.52](#) depicts the German version of the real-effort task survey as implemented in the laboratory experiment. The survey contains eight items, which are arranged in a vertical grid with the poles differing for each item. The seven radio buttons of each response scale are labeled with integer anchors and are distributed horizontally with equal spacing. The resulting grid structure of the survey ensures the subject's perception of an equidistant scale. Subject's ratings can, therefore, be treated as interval data ([S. Uebersax, 2000](#)). The user-friendly survey design facilitates its completion, making it easier to obtain truthful responses and reducing any measurement errors. Overall, the survey is short and concise, and not too elaborate to be filled out several times in a row in repeated-measures designs.

Item	Survey Item	Activity-related incentive	Purpose-related incentive	Effort type	Survey dimension
1	Not fun	Task enjoyment in general			ARI
2	No feedback	No challenge	No performance feedback		PRI
3	Boring	Monotone & unexciting task			ARI
4	Tedious & tiring	Tedious, toilsome & tiring task	Annoying task		ARI
5	Meaningless		Meaningless task		PRI
6	Physically demanding			Physical effort	Physical effort
7	Mentally demanding			Mental effort	Mental effort
8	No outcome		Task without outcome		PRI

Figure B.51: **Grouping of the items of the real-effort task survey:** On the left side are abbreviated survey items, in the middle are the design practices addressed by the survey item, and on the right side is the survey dimension to which the item belongs.

Fragebogen

Bitte machen Sie im Folgenden Angaben zu der soeben getätigten Aufgabe.

Die Aufgabe ...

	sehr 1	2	3	unent- schieden 4	5	6	sehr 7	
... hat mir Spaß gemacht	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... hat mir keine Freude bereitet
... gab mir eine Ziel-/ Leistungsmessung, die mich anspornte	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... gab mir keine Rückmeldung
... hat meine Neugierde geweckt/ war unterhaltsam	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... war sehr uninteressant/ langweilig
... war ansprechend/ mühelos bewältigbar	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	... war sehr mühselig/ nervig/ ermüdend
... erschien als sinnhaft	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... erschien als sinnlos
... war physisch einfach	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	... war physisch anstrengend
... war mental einfach	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... war mental anstrengend
... bringt etwas hervor/ erreicht ein Ziel	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	... produziert nichts/ hat kein messbares Ergebnis

Sind Sie mit Ihrer Leistung in der Aufgabe zufrieden?

eher schon

Weiter

Figure B.52: **Survey implementation in the laboratory software oTree:** Wording of items in German as employed in the laboratory study presented in Section 3.3. Each of the eight survey items contains two statements that express opposing perceptions of a task. The subjects assess the tasks on a seven-level response scale between both statements, which allows a differentiated measurement. Study participants complete the survey after each task.

B.3 Additional Figures and Tables

B.3.1 Descriptive Statistics

B.3.1.1 Study Participants

Descriptive statistics for the study participants are provided in Table C.2 in the Appendix to Chapter 4, as they are set directly in relation to subjects' performance in the tasks.

B.3.1.2 Real-Effort Task Survey Data

The motivational items of the real-effort task survey can be grouped according to whether the design practices they capture address more activity-related incentives or purpose-related incentives. Together with the effort items, the survey thus covers four dimensions (see also Figure B.51). The dimensions assessed by the survey are also reflected in the structure of the distribution of responses obtained (see Figure B.53 and Figure B.54). These response patterns are described below.

- The first, third and fourth item cover design practices that address *activity-related incentives*. The response patterns for item 1 (“no fun”) and item 3 (“boring”) are quite similar but gently differ from the pattern for item 4 (“tedious & tiring”), especially for the multiplication task.
- The second, fifth, and eighth survey items capture to what extent the design practices to attenuate *purpose incentives* are met. Subjects' responses to item 5 (“meaningless”) and item 8 (“no outcome”) are distributed very similarly, but not necessarily in the same way as item 2 (“no feedback”), which can be seen especially for the ab-typing task and single-slider task. Thus, aggregating subjects' responses for all purpose-related items would conceal these subtle differences. The ab-typing task and the single-slider task are perceived as particularly meaningless and as producing nothing (items 5 and 8). However, unlike the single-slider task, the design of the ab-typing task gave subjects much more of a goal that spurred them on (item 2).

- The greater variance among the responses for the multiplication task suggests that the task is perceived very unevenly – some subjects have strong feelings in favor of the task, others against it (see also the examination of subjects' task preferences in Appendix B.3.4.2).
- The neither mentally nor physically demanding ball-catching task is perceived as rather motivating.

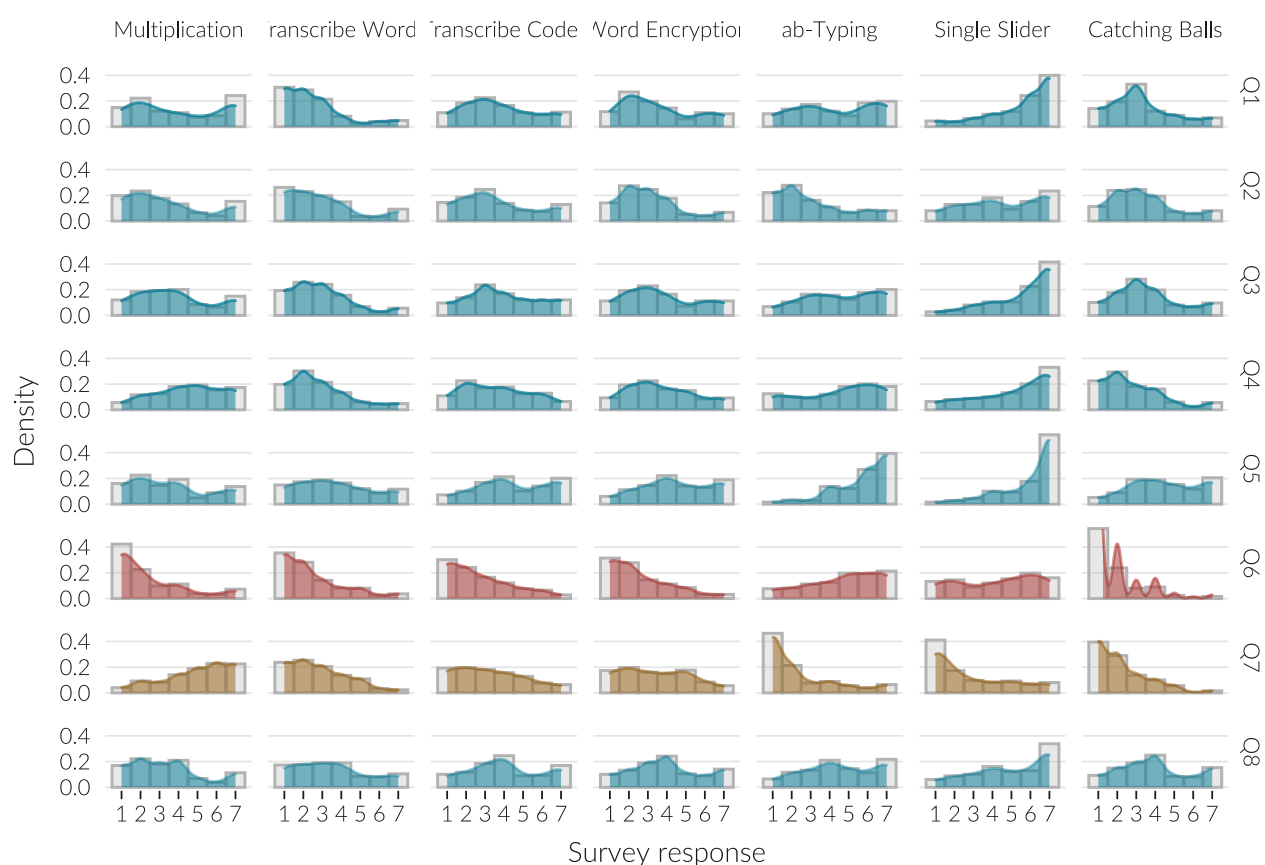


Figure B.54: **Distribution of survey responses for all tasks and survey items:** The densities for the motivational items are colored in petrol blue, for the physical effort item in red and for the mental effort item in brown.

B.3.2 Comparing Real-Effort Tasks Along Survey Dimensions

B.3.2.1 Orthogonal Contrast-Coding Scheme

Table B.8: **Orthogonal contrast-coding scheme:** Despite their differences, the [tasks in the selection](#) share certain similarities and can be grouped accordingly (see Figure 3.5). Each contrast compares two such commonality-based subsets of the task selection. Every pair of contrasts is orthogonal, such that the properties of orthogonal contrasts are fulfilled. The contrasts can be combined in a matrix that was applied to the factor variable *task* when preparing the data for the regression analysis.

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls	Total
1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{3}$	$-\frac{1}{3}$	$-\frac{1}{3}$	0
2	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	0
3	1	-1	0	0	0	0	0	0
4	0	0	1	-1	0	0	0	0
5	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$	-1	0
6	0	0	0	0	1	-1	0	0

B.3.2.2 Fixed Effects ANOVA

Table B.9: **ANOVAs for the fixed effect “task” for each motivational survey item:** For each item, strong evidence is found against the null hypothesis that the tasks are perceived to be the same by the study participants. *p*-values are adjusted neither for multiple comparisons nor for sphericity (see explanation in Section 3.3.2.3).

	Sum Sq	Mean Sq	NumDF	DenDF	F-value	Pr(>F)
Not fun	1294.2	215.71	6	1482	93.1	2.4e-99***
No feedback	387.9	64.64	6	1482	29.6	1.3e-33***
Boring	1030.5	171.76	6	1482	88.6	4.5e-95***
Tedious, tiring	1084.0	180.67	6	1482	80.7	1.7e-87***
Meaningless	1247.4	207.91	6	1482	112.7	3.1e-117***
No outcome	440.4	73.40	6	1482	41.6	4.2e-47***

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.10: **ANOVAs for the fixed effect “task” for each effort-related survey item:** Again, strong evidence the found in support of the factor *task* for both survey items for physical and mental effort. As before, *p*-values are not adjusted for multiple comparisons or for sphericity.

	Sum Sq	Mean Sq	NumDF	DenDF	F-value	Pr(>F)
Physically demanding	1613.9	268.99	6	1482	124.3	2.0e-127***
Mentally demanding	1230.4	205.07	6	1482	94.2	2.5e-100***

Note: *p*-values are neither adjusted for multiple comparisons nor for sphericity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.3.2.3 Adjusting for Repeated-Measures Design and Multiple Comparisons

Adjusting for the repeated-measures design. The observations for each subject are *dependent* due to the within-subject design of the study. The *Greenhouse-Geisser estimate* $\hat{\epsilon}$ provides a conservative measure to correct the degrees of freedom (and consequently the *p*-values) for the resulting effect of sphericity. For the simple model described in Section 3.3.2.2 one finds rather high *Greenhouse-Geisser’s* estimates $\hat{\epsilon}$. The values are relatively close to 1 and are summarized in Table B.11. This means that the Greenhouse-Geisser adjusted *p*-values do not differ considerably from the original *p*-values.

Table B.11: **Greenhouse-Geisser estimates:** The values are approaching 1, which suggests that the Greenhouse-Geisser adjusted *p*-values are not much larger than the original *p*-values. Adjustments of *p*-values due to the within-subject design are, therefore, not necessary.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	Physically demanding	Mentally demanding	No outcome
$\hat{\epsilon}$	0.85	0.88	0.87	0.88	0.86	0.75	0.84	0.89

Adjusting for multiple comparisons. In the following tables, the *p*-values are corrected to account for multiple hypothesis testing using the *Bonferroni-Holm* approach (Table B.12 for the motivational items and Table B.13 for the effort-related items). Multiplicity adjustments are performed separately for each multiple regression, i.e., for each survey item across the six contrasts. Even after adjusting for multiple hypothesis testing, *p*-values remain at similar significance levels. For the motivational items, solely the coefficient for the second contrast is no longer significant for the model for the second survey item (“no feedback”). In terms of the effort-related items, the coefficient for the second

contrast is no longer significant for the model for the sixth survey item ("physical demand"), as well as the coefficient for the sixth contrast for the seventh survey item ("mental demand").

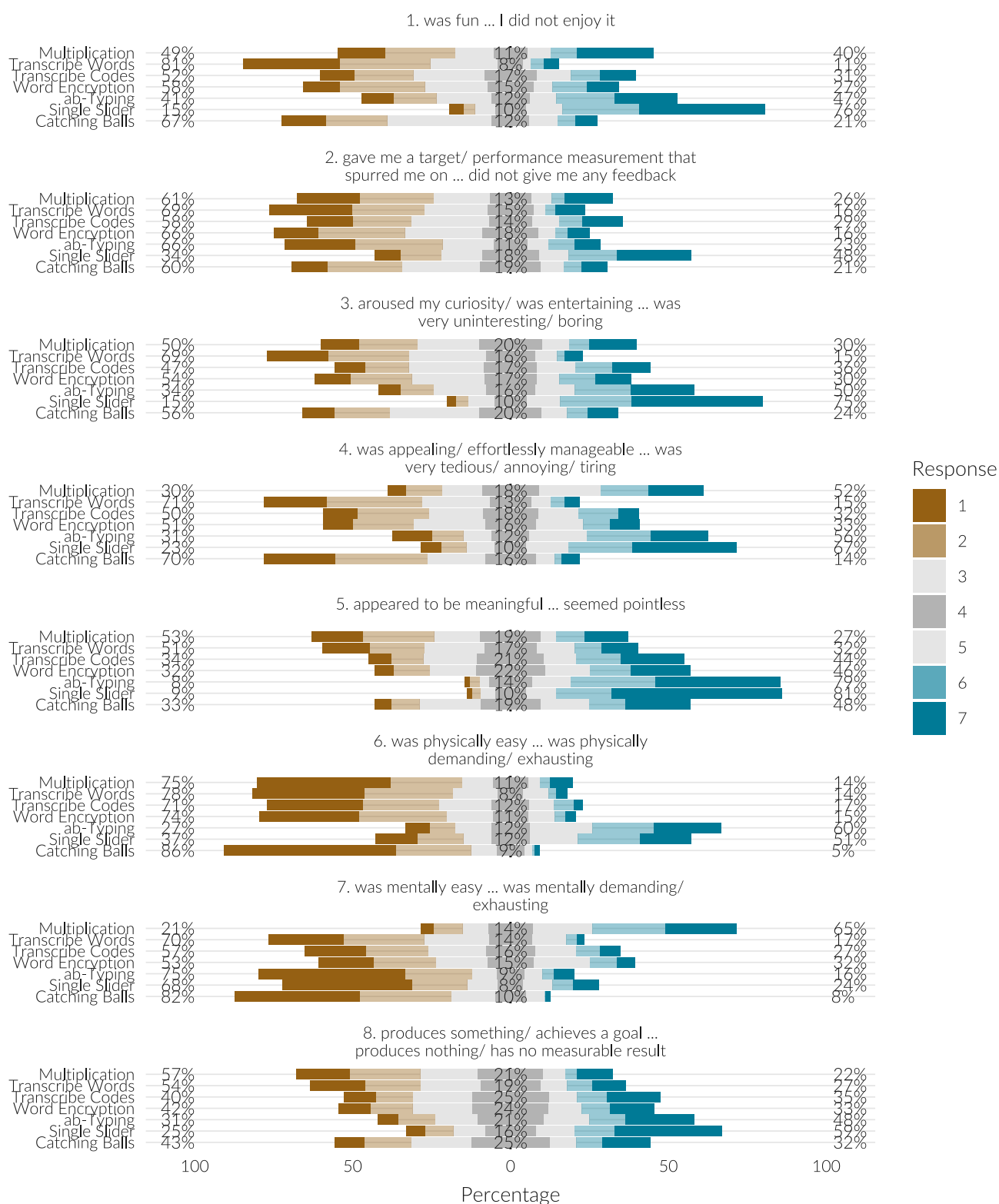


Figure B.53: **Percentage responses for all items of the real-effort task survey:** Subjects evaluate the tasks on a response scale with seven levels with (1) indicating agreement with the left statement and (7) corresponding to agreement with the statement on the right.

Table B.12: Regression estimates for all motivational survey items with adjusted p-values (simple model)

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	No outcome
(Intercept)	3.828*** (0.000)	3.417*** (0.000)	3.986*** (0.000)	3.869*** (0.000)	4.590*** (0.000)	4.027*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.609*** (0.000)	0.680*** (0.000)	1.664*** (0.000)	0.765*** (0.000)	2.331*** (0.000)	1.257*** (0.000)
C2: Cognitive → Memory	0.351** (0.002)	0.192 (0.256)	0.458*** (0.000)	-0.020 (1.000)	0.800*** (0.000)	0.583*** (0.000)
C3: Cognitive: Math → Words	-0.706*** (0.000)	-0.216** (0.008)	-0.393*** (0.000)	-0.804*** (0.000)	0.050 (0.975)	0.075 (0.809)
C4: Memory: Codes → Encryption	-0.109 (0.562)	-0.222** (0.006)	-0.117 (0.358)	0.038 (0.997)	0.008 (1.000)	-0.056 (0.948)
C5: Mechanical → Fun	-1.098*** (0.000)	-0.277** (0.002)	-0.980*** (0.000)	-1.301*** (0.000)	-0.891*** (0.000)	-0.507*** (0.000)
C6: Mechanical: ab-Typing → Single Slider	0.615*** (0.000)	0.690*** (0.000)	0.526*** (0.000)	0.310*** (0.000)	0.107 (0.441)	0.236*** (0.001)
N (obs.)	1736	1736	1736	1736	1736	1736
N (subjects)	248	248	248	248	248	248
R2 (fixed)	0.18	0.06	0.16	0.16	0.18	0.07
R2 (total)	0.45	0.41	0.49	0.42	0.53	0.53
AIC	6792.09	6720.14	6535.99	6707.38	6470.81	6460.93

Note: p-values are adjusted for multiple-hypothesis testing using the Bonferroni–Holm method.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.13: Regression estimates for all effort-related survey items with adjusted p-values (simple model)

	Physically demanding	Mentally demanding
(Intercept)	3.025*** (0.000)	3.134*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.799*** (0.000)	-1.979*** (0.000)
C2: Cognitive → Memory	0.183 (0.299)	-0.500*** (0.000)
C3: Cognitive: Math → Words	-0.016 (1.000)	-1.060*** (0.000)
C4: Memory: Codes → Encryption	-0.067 (0.928)	0.060 (0.957)
C5: Mechanical → Fun	-1.726*** (0.000)	-0.265** (0.004)
C6: Mechanical: ab-Typing → Single Slider	-0.234** (0.003)	0.179* (0.046)
N (obs.)	1736	1736
N (subjects)	248	248
R2 (fixed)	0.24	0.19
R2 (total)	0.45	0.42
AIC	6624.53	6645.75

Note: p-values are adjusted for multiple-hypothesis testing using the Bonferroni–Holm method.

* p < 0.05, ** p < 0.01, *** p < 0.001

B.3.2.4 Multiple Comparison with the Best: Exemplary Calculation

The *multiple comparisons with the best* method of Hsu (1996) analyzes the differences between level means. It identifies the factor levels with the largest mean (the “best”), those that are indistinguishable from these “best,” and those that are significantly different from the best. Multiple comparisons tests are typically performed on treatment means, but can also be applied to other statistics like variances (in the present case the analysis was performed both for means and variances). The following description of the multiple comparisons procedure with the *largest mean* is largely based on Kuehl (2000).

To compare a set of tasks based on subjects assessments, the difference D_i between the mean response for each task \bar{y}_i , and the largest mean response of the remaining task $\max(\bar{y}_j)$ with $j \neq i$ has to be calculated.

$$D_i = \bar{y}_i - \max_{j \neq i}(\bar{y}_j), \text{ for } i = 1, 2, 3, \dots, t$$

Moreover, one has to derive M for which the tabled statistic for one-sided comparisons $d_{\alpha, k, \nu}$ has to be looked up (see Appendix Table VI, p. 597 in Kuehl, 2000).

$$M = d_{\alpha,k,\nu} \sqrt{\frac{2s^2}{r}}$$

To derive M , the following values allow to obtain $d_{\alpha,k,\nu}$:

- the family-wise error rate is fixed at $\alpha_E = 0.05$;
- the number of comparisons is $k = 7 - 1 = 6$ for seven tasks;
- the degrees of freedom for the experimental variance is $\nu = 1729$;

In the mentioned table one finds $d_{\alpha,k,\nu} = 2.29$ (with $\nu = \infty$, since $1729 \gg 120$). To calculate M , one further needs:

- the experimental variance: $s^2 = MSE = 3$ (estimate of experimental error variance for the experiment);
- the number of replicates $r = 248$ (the number of observations, which in this case corresponds to the total sample size);

and finds $M = 2.29 * \sqrt{\frac{2*3}{248}} = 0.38$.

With all this, the $100(1 - \alpha)\%$ simultaneous constrained confidence intervals can be derived:

- lower confidence interval bound for $\mu_i - \max(\bar{y}_j)$

$$L = \begin{cases} D_i - M, & \text{if } (D_i - M) < 0 \\ 0, & \text{otherwise} \end{cases}$$

- upper confidence interval bound for $\mu_i - \max(\bar{y}_j)$

$$U = \begin{cases} D_i + M, & \text{if } (D_i + M) > 0 \\ 0, & \text{otherwise} \end{cases}$$

Table B.14 provides the results of the calculations for the last item of the real-effort task survey. The confidence intervals (CIs) indicate whether a task is “best” (*lower bound* of CI is zero), whether tasks are “insignificantly different from best” (CI contains zero), and whether tasks are “significantly different from best” (*upper bound* of CI is zero). The last column concludes whether the task is selected as “best.”

Table B.14: **Calculations for Hsu’s multiple comparisons with the best (for means):** for the last item of the real-effort task survey (“no outcome”)

Task	\bar{y}_i	$\max(\bar{y}_j)$	D_i	$D_i - M$	$D_i + M$	$\frac{95\%SCI}{(L,U)}$	Select?
Multiplication	3.35	4.94	-1.59	-1.97	-1.21	(-1.97,0)	No
Transcribe Words	3.50	4.94	-1.44	-1.82	-1.06	(-1.82,0)	No
Transcribe Codes	4.06	4.94	-0.88	-1.26	-0.49	(-1.26,0)	No
Word Encryption	3.95	4.94	-0.99	-1.37	-0.60	(-1.37,0)	No
ab-Typing	4.46	4.94	-0.47	-0.86	-0.09	(-0.86,0)	No
Single Slider	4.94	4.46	0.47	0.09	0.86	(0,0.86)	Yes (Best)
Catching Balls	3.94	4.94	-1.00	-1.38	-0.61	(-1.38,0)	No

B.3.2.5 Robustness of Results

Two hundred forty-eight study participants provided their assessments of seven tasks. In total, the survey was completed 1736 times. In 6.8% of these cases, subjects provided the same response to all survey items. 22.2% of subjects engaged in this undesired subject behavior at least once, which steadily increases with time (8 incidents for the first task compared to 24 incidents for the last one in the task sequence). In light of this, it seems unlikely that the observed pattern of responses is due to unclear instructions or a poorly designed survey. Instead, the behavior may rather be attributed to the long duration and the burdensome content of the experiment.⁵ Meanwhile, filtering out these potentially fraudulent observations leaves the results of the subsequent analysis unchanged. For this

⁵Desselle (2005) note that “a respondent who provides a neutral response consistently for much of the items may be sending a message they did not care to participate in the study.” In roughly a third of the cases where subjects made the same choice for all survey items they chose the neutral response.

reason, these observations are not discarded and the results based on the full data set are reported.

As described in Section 3.3.1.2, each experimental session contained two groups, each with their own unique sequence of experimental components, including a distinct task sequence. The twelve sessions yielded 24 groups of, on average, ten subjects who completed the tasks and the subsequent survey in the same order. This approach represents a non-optimal randomization procedure. Ideally, each subject would have been assigned an individual task sequence. However, this approach was technically not feasible with the laboratory software. The following will examine whether the chosen randomization of the task order could have led to spill-over effects.

Figure B.55 shows the position index of each task across all experimental sessions. Figure B.56 depicts the responses to the real-effort task survey conditional on the position of a task in the task sequence. The visual results suggest that the assumption, that a randomization of the task sequence across all subject groups is sufficient to mitigate order effects, may hold. To substantiate this presumption the next section presents an extension to the parsimonious model presented in Section 3.3.2.

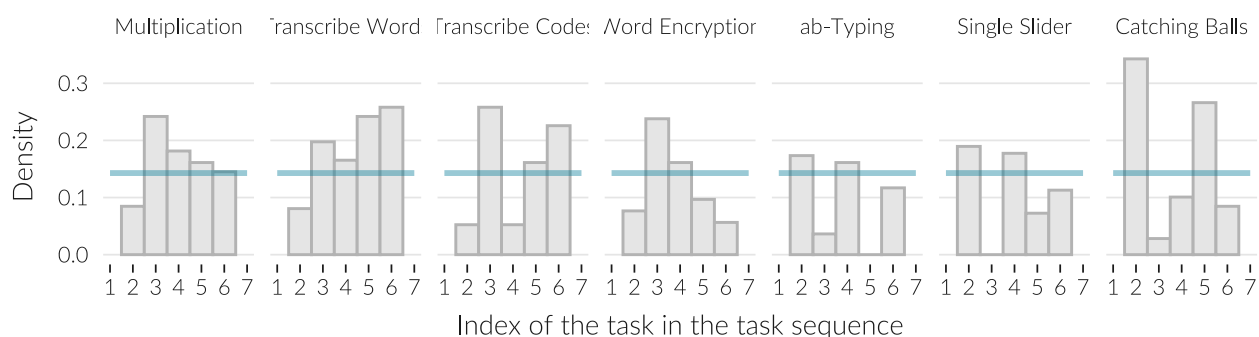


Figure B.55: **Distribution of task indices for all tasks:** The order of tasks was randomized across groups of subjects. Each of the twelve experimental session was split into two groups, yielding a total of 24 groups with a unique task sequence (for comparison, the petrol blue line indicates if all indices would occur equally frequently). The distributions are quite different from each other, and coincidentally, certain tasks did not occur at particular positions in the sequence (e.g., the ball-catching task was never the last). Nevertheless, none of the distributions deviates sharply from a uniform distribution in average, i.e., none of them is strongly skewed to the left or right or exhibits a spike.



Figure B.56: **Survey responses conditional on task order:** The grey bars resemble the mean responses of the subjects to each survey item and for each task conditional on the index of the task in the task sequence (CI: normal). The each task and possible index in the task sequence ($i \in [1, 7]$), the number of subjects who completed the task in a particular position is stated (summed across all sessions). For example, 14 subjects completed the word-encryption task as their sixth task.

Extended regression model. Section 3.3.2 presented a simple model to investigate the main questions of interest. For each survey item q it is assessed i) whether subjects perceive the design of tasks to be different (i.e., a *task effect*), and ii) whether subjects perceive a given tasks differently (i.e., the variability between different subjects with regards to the general task liking level). The complex model extends the simple model to address the following supplementary questions regarding the controlled variables:

- III) Is there an *order effect* w.r.t. the *position of task in the task sequence*?
- IV) How large is the *variability between different groups* with regards to the *general task liking level*?

These questions are addressed through a fixed effect for *task index* with seven levels (i.e., one for each possible position of a task in the task sequence) and a random effect for *group* (see [definition of groups](#) above). Thereby, the *group* variable may capture potential order effects w.r.t. i) the *position of task in the task sequence*, ii) the *succession of tasks in the task sequence*, and iii) the *succession of experimental components within the experiment*. Moreover, it may reflect the general experimental circumstances, including the timing of experiment, within-session incidents or the experimenter supervising the session. Here *group* is mainly added as a control to assess the second point ii), i.e., order effects w.r.t. the *succession of tasks in the task sequence*. The simple linear mixed-effects model introduced in Section 3.3.2 is extended by adding a fixed effect for *task index* and a random effect for *group*:

$$Y_{tisg}^q = \mu^q + \alpha_t^q + \beta_i^q + \delta_s^q + \gamma_g^q + \varepsilon_{tisg}^q \quad (\text{B.1})$$

In the complex model, observation Y_{tisg}^q is the response of subject s to survey item q for task t . The model contains all variables contained in the simple model, i.e., for each survey item q

- μ^q global mean
- α_t^q fixed effect of task t (deviation from global mean due to task t)
- δ_s^q random effect of subject (“general task liking level of subject s ”)

and in addition

- β_i^q fixed effect of *task index* (position of task in the task sequence),
- γ_g^q random effect of *group* (“general task liking level of group g ”)
- random error term: ε_{tisg}^q .

For each survey item q it is assumed that the error terms are normally distributed and i.i.d.. Furthermore, it is assumed that the random effects are normally distributed.

As in the previous analysis, the *motivational items* and the *effort-type items* of the survey are examined separately. In both cases, the influence of the fixed effects *task* and *task position in the task sequence* (task index) is examined first, which are treated as factors in the entire analysis. The following tables [B.15](#) and [B.16](#) summarize the results of the ANOVAs for each survey item. Significant differences

in means are found in terms of p -values for each survey item. Therefore, the analysis with planned comparisons can be performed as intended.

Table B.15: **ANOVAs for the fixed effects “task” and “index” for each motivational survey item:** To control for potential order effects, a fixed effect for the *task index* as well as a random effect for the *group* are additionally included in the more complex regression model. For each survey item a significant effect of task is observed, and, for half of the items, a significant effect of task index.

	Sum Sq		Mean Sq		NumDF	DenDF	F-value		Pr(>F)	
	Task	Index	Task	Index			Task	Index	Task	Index
Not fun	1175.7	20.48	195.9	3.41	6	1476	84.7	1.476	2.5e-91***	1.8e-01
No feedback	406.6	84.17	67.8	14.03	6	1476	31.7	6.566	4.9e-36***	7.4e-07***
Boring	946.0	26.44	157.7	4.41	6	1476	81.8	2.285	1.9e-88***	3.4e-02*
Tedious, tiring	1045.9	45.54	174.3	7.59	6	1476	78.7	3.425	2.0e-85***	2.3e-03**
Meaningless	1003.3	9.29	167.2	1.55	6	1476	90.6	0.839	6.8e-97***	5.4e-01
No outcome	403.1	15.28	67.2	2.55	6	1476	38.2	1.448	2.9e-43***	1.9e-01

Note: p -values are neither adjusted for multiple comparisons nor for sphericity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.16: **ANOVAs for the fixed effects “task” and “index” for each effort-related survey item:** For the survey items for “physical effort” and “mental effort” a significant effect of *task* is found; for position of the task in the task sequence (*index*) is only significant for “mental effort.”

	Sum Sq		Mean Sq		NumDF	DenDF	F-value		Pr(>F)	
	Task	Index	Task	Index			Task	Index	Task	Index
Physically demanding	1455	41.1	243	6.84	6	1476	113.1	3.19	1.7e-117***	0.0041**
Mentally demanding	1101	27.1	184	4.51	6	1476	84.6	2.08	3.1e-91***	0.0528

Note: p -values are neither adjusted for multiple comparisons nor for sphericity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For the apriori defined comparisons the same orthogonal contrasts as detailed in Section 3.3.2.1 are employed. The linear mixed-effect model is fit for each survey item by restricted maximum likelihood. Tables B.17 and B.18 summarize the results of the regression analysis for the complex model for the *motivational* items and for the *effort type* items respectively.

Table B.17: **Regression estimates for all motivational survey items (complex model):** The complex model contains all variables of the simple model (a fixed effect for the task and a random effect for *subject*). In addition, it contains a fixed effect for the position of the task in the task sequence (*index*) and a random effect to control for the imperfect randomization of the task order (*group*). The results of the regression analysis are discussed in comparison to those of the simpler model in Appendix B.3.2.6.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	No outcome
(Intercept)	3.690*** (0.000)	2.974*** (0.000)	3.753*** (0.000)	3.529*** (0.000)	4.747*** (0.000)	3.849*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.661*** (0.000)	0.843*** (0.000)	1.776*** (0.000)	0.906*** (0.000)	2.241*** (0.000)	1.267*** (0.000)
C2: Cognitive → Memory	0.372** (0.002)	0.260 (0.068)	0.480*** (0.000)	0.026 (1.000)	0.787*** (0.000)	0.610*** (0.000)
C3: Cognitive: Math → Words	-0.712*** (0.000)	-0.243** (0.003)	-0.400*** (0.000)	-0.813*** (0.000)	0.054 (0.995)	0.069 (0.956)
C4: Memory: Codes → Encryption	-0.098 (0.837)	-0.169 (0.122)	-0.090 (0.834)	0.060 (0.995)	-0.008 (1.000)	-0.047 (0.998)
C5: Mechanical → Fun	-1.084*** (0.000)	-0.286** (0.004)	-0.960*** (0.000)	-1.331*** (0.000)	-0.906*** (0.000)	-0.547*** (0.000)
C6: Mechanical: ab-Typing → Single Slider	0.615*** (0.000)	0.690*** (0.000)	0.533*** (0.000)	0.308*** (0.000)	0.102 (0.665)	0.233** (0.001)
N (obs.)	1736	1736	1736	1736	1736	1736
N (subjects)	248	248	248	248	248	248
N (groups)	24	24	24	24	24	24
R2 (fixed)	0.18	0.07	0.16	0.17	0.19	0.07
R2 (total)	0.45	0.42	0.49	0.43	0.53	0.54
AIC	6809.07	6706.47	6549.03	6713.34	6492.92	6475.28

Note: *p*-values are neither adjusted for multiple hypothesis testing nor for sphericity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.18: **Regression estimates for all effort-related survey items (complex model):** A discussion of the regression results is provided in the following Appendix B.3.2.6.

	Physically demanding	Mentally demanding
(Intercept)	2.715*** (0.000)	2.854*** (0.000)
C1: Cognitive & Memory → Mechanical & Fun	1.866*** (0.000)	-1.908*** (0.000)
C2: Cognitive → Memory	0.226 (0.175)	-0.460*** (0.000)
C3: Cognitive: Math → Words	-0.020 (1.000)	-1.070*** (0.000)
C4: Memory: Codes → Encryption	-0.057 (0.996)	0.079 (0.946)
C5: Mechanical → Fun	-1.734*** (0.000)	-0.292** (0.003)
C6: Mechanical: ab-Typing → Single Slider	-0.235** (0.005)	0.177 (0.084)
N (obs.)	1736	1736
N (subjects)	248	248
N (groups)	24	24
R ² (fixed)	0.24	0.19
R ² (total)	0.45	0.43
AIC	6634.27	6662.31

Note: p -values are neither adjusted for multiple hypothesis testing nor for sphericity.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In this extended model framework, the random deviation of an individual study participant consists of a general shift δ , (“subject specific general task perception” or “general task liking level of a subject”) and a general shift γ_g (“group specific general task perception” or “general task liking level of a group”) due to the subject belonging to a specific experimental group. Moreover, there is an “average (over the whole population) task perception” (or “preference profile”) with regards to the seven different positions a task can have in the task sequence, given by the fixed effect β_i . Both stated fixed effects (α_i^g for the task and β_i for the task index) have to be interpreted as population averages.

B.3.2.6 Model Comparison

According to Table B.15 and B.16 the *index of the task in the task sequence* yields only a minor effect on the response variable, subject’s task perception. A first view of the full regression results seems to confirm this also for the *group* variable. Therefore, the additional explanatory value of the more complex model appears to be limited. This supposition is examined in two steps: First, the AIC-criterion allows for a relative comparison of models and rewards goodness of model fit (Akaike, 1974); a subsequent ANOVA ensures that any differences found are significant. By combining the two ap-

proaches, one can determine whether the addition of another predictor adds substantial explanatory value (in terms of a smaller AIC value indicating a better model fit) and does significantly differ from a simpler model (according to an ANOVA). A more complex model that has only a slightly lower AIC value than a simpler model and does not differ from the latter based on an ANOVA implies that the additional predictor is not necessarily worthwhile. As before, the analysis is performed separately for the motivational and effort-type items of the survey.

Compare models with the AIC-criterion. The AIC-criterion serves as a measure of goodness of model fit to assess the impact of additional regressors accounting for the dependencies in the experiment. The following tables [B.19](#) and [B.20](#) summarize the AIC values for the motivational items and the effort-related items of the survey. According to [Akaike \(1974\)](#), the model with the lowest AIC value is preferred among the compared set of models.

Covariates are gradually added to a baseline model, which only contains a fixed effect for the *task*. The simple model extends the baseline model with a random effect for the *subject*. Adding a random effect for the *group* allows for a slight improvement of the AIC values for the motivational items, except for item 2 (“no feedback”), for which the complex model with the *index* of the task in the sequence has the lowest AIC value. Whereas the addition of the random effect for the *subject* yields a large improvement in AIC values for all survey items, the addition of further predictors does not lead to greater changes and is thus not particularly worthwhile.⁶

For the effort-related survey items, the simple model turns out to have the lowest AIC values. The difference between the models is further substantiated by the model comparison with ANOVAs provided below.

Compare models with ANOVAs. According to the analysis with the AIC criterion above, more regressors yield no improvement over the simple model for most of the survey items. Lower AIC values are found only for item 2 (“no feedback”), if including an additional fixed effect for *index* and a random effect for the *group*, and item 8 (“no outcome), if including an additional random effect for

⁶Solely adding the *group* as a random effect in addition to the simple model specification could be considered. Yet, the AIC values for this model are not very different to the ones from the simple model, which only contains *task* as a fixed effect and *subject* as a random effect. This is confirmed by calculations of the relative likelihood as estimate of the probability that the model minimizes information loss (results available upon request).

Table B.19: **Assess the impact of additional regressors for the motivational items with the AIC-criterion:** The table presents the AIC values for a number of models differing in complexity (across columns) for all motivational survey items (rows): 1) the *base model* solely contains a fixed effect for the *tasks*; 2) in the *simple model*, a random effect is added to provide for the within-subject design (all subjects completed all tasks); 3) to take into account the position of the task in the task sequence, a fixed effect (*index*) is added to the simple model; 4) to control for the imperfect randomization of the task order, a random effect is added to the simple model (*group*: 24 groups of on average 10 subjects completed the tasks and the following survey in the identical order); 5) the *complex model* contains both the fixed effect for the position index of the task and the group variable.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	No outcome
Base model: task (FE)	7099.91	7090.34	6957.06	6967.05	6941.46	7123.43
Simple model: BM + subject (RE)	6792.09	6720.14	6535.99	6707.38	6470.81	6460.93
Simple model + index (FE)	6809.92	6708.28	6550.07	6713.85	6493.82	6480.58
Simple model + group (RE)	6791.24	6718.33	6534.94	6706.87	6469.91	6455.64
Complex model: SM + index + group	6809.07	6706.47	6549.03	6713.34	6492.92	6475.28

Table B.20: **Assess the impact of additional regressors for the effort-related items with the AIC-criterion**

	Physically demanding	Mentally demanding
Base model: task (FE)	6840.70	6879.27
Simple model: BM + subject (RE)	6624.53	6645.75
Simple model + index (FE)	6632.59	6660.36
Simple model + group (RE)	6626.21	6647.70
Complex model: SM + index + group	6634.27	6662.31

the *group*. A set of ANOVAs is conducted to assess whether any observed differences between the candidate models are significant. Separately for each survey item, an ANOVA examines, the impact of adding to the simple model i) the “index of task in the task sequence” as fixed effect, and ii) the “group” as random effect, and iii) adding both terms at once. In terms of adding regressors to the simple model, significant differences are found only for the items mentioned. In conjunction with the AIC values it can be concluded that for these items the models with the additional regressors (item 2: *index* and *group*; item 8: *group*) perform better than the simple model. Additionally, significant differences are found for the items 3 (“boring”) and 4 (“tedious, tiring”). According to the AIC values, the simple model performs significantly better than the models which contain *index* as an additional predictor for these items. For the remaining motivational items (1: “not fun”; 8: “meaningless”), no

statistically significant difference is observed between the models.

Likewise, ANOVAs were performed for the effort-related survey items. Significant differences are only found for item 6 (“physically demanding”); according to the AIC values, the simple model outperforms the model containing an additional fixed effect for *index* as well as the complex model.⁷

Table B.21: Assess the impact of additional regressors for the motivational items with an ANOVA: According to the AIC values, the observed significant differences between the simple model and its extension containing “task index” as a fixed effect are in support of the simple model on the third and fourth items, and in favor of the model containing the task index on the second item. No difference is found between the simple model and its extension containing “group” as a random effect, except for the last survey item. According to the AIC values, the model with “group” performs better here. The observed differences between the simple and the complex model result from a composite of these observations.

	Not fun	No feedback	Boring	Tedious, tiring	Meaningless	No outcome
Simple model ⇒ SM + index (FE)	0.179	0***	0.032*	0.002**	0.536	0.189
Simple model ⇒ SM + group (RE)	0.118	0.069	0.106	0.144	0.118	0.01*
Simple model ⇒ Complex model	0.124	0***	0.022*	0.002**	0.378	0.032*

Note: *p*-values are not adjusted for multiple hypothesis testing.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B.22: Assess the impact of additional regressors for the effort-related items with an ANOVA

	Physically demanding	Mentally demanding
Simple model ⇒ SM + index (FE)	0.004**	0.051
Simple model ⇒ SM + group (RE)	0.67	0.927
Simple model ⇒ Complex model	0.007**	0.084

Note: *p*-values are not adjusted for multiple hypothesis testing.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

⁷One could further consider to add an interaction term between both fixed effects (*task* and *task index*). This was examined in an additional regression and no significant influence was found according to an ANOVA. Considering the respective AIC value, the model does not improve by adding the interaction term, it even deteriorates in comparison to the simpler models.

B.3.2.7 Inter-rater Agreement

One property of a task that is beneficial for [task applications one and two](#) is when the task has the same motivational influences on the subjects and is equally effortful for all of them. The [real-effort task survey](#) can be used to identify tasks that are viewed similarly by study participants. The more consistent the study participants' task ratings are, the more it can be expected that motivational influences and effort demands are the same for them.

A large body of literature examines the consistency or homogeneity of evaluations by different raters using the same scale to assess a given object. Exemplary applications of *inter-rater reliability* or *inter-rater agreement measures* include comparing doctors' ratings of patients' symptoms or judges' interpretations of legal disputes.⁸ The strand of the literature usually refers to "raters," so the term is used synonymously with *study participants* in the following. According to this literature, there can be several reasons why subjects may disagree in rating a task. [S. Uebersax \(2000, p. 2\)](#) distinguishes three components of disagreement of interval-level ratings: "effects on the *correlation or association of raters' ratings, rater bias, and rater differences in the distribution of ratings*" [emphasis added]. Each of them is closer detailed in the following, including potential impact on study outcomes and ways to mitigate these.

Rater association. The subjects interpret the survey items differently. In their assessment, they take a variety of factors into account; which factors they consider may vary from subject to subject and also how much weight they attach to them in their rating. To reduce any random noise caused by this, special care was taken in the preparation of the instructions for the tasks and the survey. Attention was paid to compiling the bipolar statements of the survey as unambiguous and antithetical as possible.⁹ This is why some of the items may seem wordy at first sight.

Moreover, the rating process always contains some degree of random error, which means that even the same rater will form different judgments about the selfsame item over time. This constitutes an

⁸Further applications include content analysis and computational linguistics (labeling content or text as having certain features), pharmaceutical research (evaluating drug effects), survey research, and programming (agreement among programmers on how they would solve a particular software problem).

⁹As noted in Section 3.2, feedback from participants of a pilot study confirmed that the survey contains a clear and precise wording.

inevitable part of any such study and a reason for choosing such a large sample size ($n = 248$). The experiment included a variety of tasks to help keep the subjects engaged. The experiment included a diverse selection of tasks, both to reflect the variety of tasks available and to provide subjects with variation to keep them engaged. They were also instructed to take a short break after each task to relax and recover in order to regain focus and attention.

Rater bias. Certain subjects generally tend to rate tasks higher or lower because they generally like or dislike real-effort tasks or because they interpret the rating scale calibration in a different way. Rater bias resembles a “subject specific general task perception” or a “general task liking level of a subject” and is acknowledged in the regression analysis by addition of a random effect for subject.

Rating distribution. Each study participants has a subjective view of a task – and these task perceptions may differ largely. Put differently, subjects inevitably will disagree and that is totally acceptable in this study. Instructions were clear that raters should truthfully report their view of the tasks and that they should not follow any sort of norm “how they should perceive a (type of) task” (see also Section 3.2 regarding measures taken to prevent socially desirable responses).

Among these three reasons why subjects may disagree on the rating of a task, the third *difference in rating distributions* is of greater interest here. In the literature on inter-rater agreement, several measures have been proposed to assess the degree of similarity of raters’ ratings distributions, i.e., to which degree two raters assign some object the same score. Applied to the presented study, the strength of inter-rater agreement would reflect the extent to which subjects hold the same perception of a particular task. However, the study does not present a typical inter-rater agreement study (cf. the examples provided above). The number of raters (248 study participants) is very large and there are only few units (eight survey responses per task), which are at best examined separately. Several measures were compared to determine the most appropriate.

Pearson correlation expresses the strength of the relation between two sets of values. While it assess correlations it can not capture the degree of agreement between raters (Bland & Altman, 2003; Kottner et al., 2011). In other words, the Pearson correlation may catch when the evaluations of two raters have the same melody, but not that they disagree in the tonality, i.e., the pitch of the melody. Also,

it does not deliver any reasonable results if there is no variance among the ratings, and unfortunately some raters gave the same response to all survey items (see Appendix B.3.2.5).

Overall percent agreement is defined as the number of observations where two subjects agree divided by the total number of observations considered. *Cohen's kappa* corrects this value for any agreement between the subjects that occurred by chance. However, it allows to compare the valuations of two raters only. *Lights kappa* and *Fleiss kappa* are adaptations of Cohen's kappa for more than two raters. Just like the *Intra-class correlation coefficient* they do not permit to account for the data type through weighting. Conversely, *Weighted kappa* allows for quadratic weighting to match the ordinal data structure, but can only cope with two raters.

Krippendorff's Alpha Reliability Coefficient represents a generalized version of Cohen's Alpha for a large number of raters and allows for weighting. It assesses whether the raters judgments are reliable, i.e., how much of the overlapping in ratings can be attributed to chance. Krippendorff's Alpha seems to be a promising measure. However, it appreciates if raters have to assess very different units (such that the complete range of the response scale is utilized. This means that all diagonal elements of the *coincidence matrix*, which has to be calculated in the process of deriving Krippendorff's Alpha, have high values. At the same time Alpha penalizes if the opposite is the case, i.e., if only a fraction of the response scale is used (even if raters show great agreement and thus only the corner elements of the coincidence matrix have high values).¹⁰ This is exactly the case in the current setting, wherefore Krippendorff's Alpha is not ideal.

Personal communication with Professor Krippendorff to find a weighted measure for multiple raters that does not require variance has proven fruitful to derive a suitable measure.¹¹ However, a detailed presentation and discussion would go beyond the scope of this paper. Interested readers are referred to a recent publication by Professor Krippendorff, who took up my use case with great interest and dedicated a section to it (see [Krippendorff, 2021](#)).

For now, heat maps of coincidence matrices allow a visual comparison of differences in rater distributions, and are sufficient to illustrate congruence in subjects' task perceptions. The coincidence matrices are also called concordance matrices and represent an intermediate step in the calculation

¹⁰A detailed description of how the coincidence matrix is derived can be found below.

¹¹Personal communication between December 2019 and February 2020.

of Krippendorff's alpha. They can be derived as follows. The coincidence matrices are also called concordance matrices and can be constructed as follows. In a first step, contingency tables are created for each possible pairing of raters i and j to determine the congruence in their ratings. Then, all contingency tables are superimposed to obtain a contingency table with the total congruency in the raters' assessments. To obtain the coincidence matrix, this table of superimposed contingency tables is added to its transpose, making the coincidence matrix symmetric along the diagonal and losing any connection to the raters. The derived coincidence matrices can then be plotted as a heat map to illustrate the consistency of ratings across the levels of the response scale.

In the present case, study participants rated the tasks along the eight dimensions of the real-effort task survey on a seven-level response scale. The elements in a contingency table correspond to the frequency with which rater i chose response $r_i \in [1, 7]$ and rater j chose response $r_j \in [1, 7]$. Thus, each contingency table consists of a 7×7 matrix. Consequently, the coincidence matrix, as a superposition of the contingency tables for all possible rater pairs plus their transpose, has the same format.

The described process can be either be applied to compare ratings *across (selected) tasks* or *across (selected) survey items*. Figure B.57 follows the latter approach and depicts concordance matrices for four selections of survey items separately for each task: all items, the motivational items, the mental effort item, and the physical effort item respectively.¹² Since the raters assessed the survey on a response scale with seven levels, the values on both axis range from one to seven (higher values correspond to a less motivating and more effortful task). The more frequently a choice combination between a pair of raters occurred, the more **petrol blue** the respective cell is colored.¹³ In short, tasks that are darkly colored in the lower left area are perceived as pleasant by the subjects, while tasks that are darkly colored in the upper right area are perceived as unpleasant by the subjects. Moreover, the softer and more dispersed the coloring of a graph, the more differently the task was assessed by the subjects. It is quickly apparent that the subjects agreed in their views regarding the physical demands for the multiplication task and ball-catching task, and concerning the mental demands for the ab-typing task, singles-slider task and ball-catching task. In addition, the subjects strongly concurred in

¹²Illustrations of concordance matrices across different groupings of tasks are available upon request.

¹³The heat maps on the far left include all survey items and consequently contain eight times the data. The same applies to the heat maps for the motivational dimension, which covers six survey items and comprises six times the data. The informative value of these respective heat maps is correspondingly higher than those for the effort items.

their assessment of the motivational dimensions of the questionnaire for the singles-slider task, and to a slightly lesser extent for the ab-typing task.

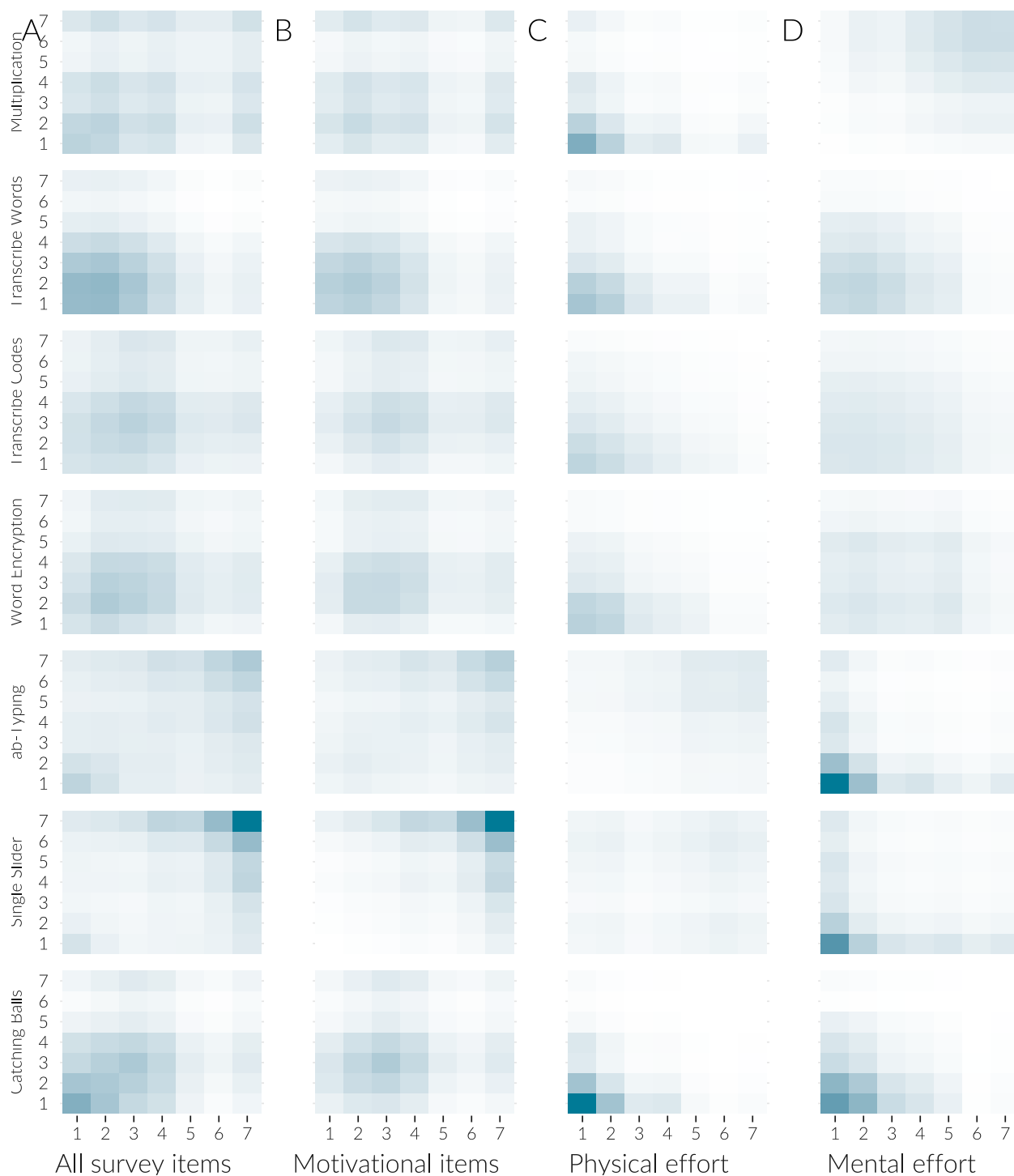


Figure B.57: **Visualization of the consistency of the subjects' perception of the tasks:** Each heat map is based on a concordance matrix, which in turn is a superposition of the contingency tables for all possible pairings of raters to which the transpose of the superposition is added. Tasks colored darkly in the bottom left were considered as pleasant by the subjects, while tasks colored darkly at the top right were found as unpleasant by the subjects. The more dampened and distributed the coloration of a graph, the more differently the task was perceived.

B.3.3 Survey Validation

B.3.3.1 Exploratory Factor Analysis for the Real-Effort Task Survey

An exploratory factor analysis for the real-effort task survey examines in what ways and to what extent its items are correlated. This can reveal their underlying structure to assess whether or not the survey is capturing the constructs as intended. In this approach the entire collected data is considered, i.e., subjects responses to all items and all tasks.¹⁴ The factor analysis revealed that the survey contained four factors; all survey items were part of a factor and had loadings (see Figure B.58 for an illustration of the derived factor structure and Table B.23 for the precise loadings of the identified factors). The derived factor structure resembles the intended survey summarized very closely (see Figure B.51). The factor *MR1* predominantly covers activity-related incentives while factor *MR3* captures purpose-related incentives. The factors *MR2* and *MR4* mainly record the mental and physical demands of a task.

The fourth survey item examines whether subjects perceive a task as *tedious, annoying, or tiring*. One could consider dropping the item since none of the factor loadings is above 0.4, and also loads on several of the factors. If an abbreviated version of the survey is desired, the survey item may be omitted. However, scale purification through item elimination can affect construct validity, such that the measure no longer assesses what it was originally intended to measure.

¹⁴In a pre-step, the data is split into two random samples for training and verification. Performing a factor analysis on both sub-samples yields the same factor loadings, giving support to the approach. The reported results are calculated on the complete data set.

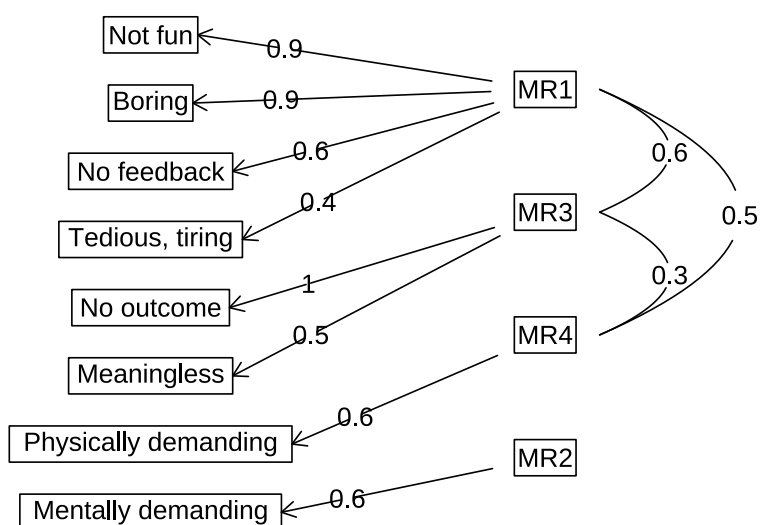


Figure B.58: **Factor analysis for the real-effort task survey:** Parallel analysis suggests that the number of factors is four. The root mean square of residuals (RMSR) is 0.003 (< 0.05 is good), the root mean square error of approximation index (RMSEA) is 0.03 (< 0.05 is good); the Tucker-Lewis Index (TLI) of factoring reliability is 0.994 (anything above 0.9 is acceptable). Only items with loadings above 0.4 are included in the factors depicted. These correspond fairly closely to the intended survey structure (see also Figure 3.3).

Table B.23: Detailed factor loadings and factor correlations

Below the factor loadings, the correlations of the four factors are given.

Variable	MR1	MR3	MR4	MR2	b^2	u^2	com
Not fun	0.93	-0.05	0.01	0.05	0.84	0.16	1.01
No feedback	0.65	0.20	-0.17	0.13	0.54	0.46	1.44
Boring	0.88	0.03	0.07	-0.11	0.85	0.15	1.05
Tedious, tiring	0.39	0.07	0.37	0.33	0.71	0.29	3.01
Meaningless	0.30	0.46	0.25	-0.23	0.68	0.32	2.88
Physically demanding	-0.02	0.05	0.60	0.10	0.42	0.58	1.07
Mentally demanding	-0.01	0.02	0.12	0.59	0.40	0.60	1.08
No outcome	-0.02	1.01	-0.01	0.02	1.00	0.00	1.00
SS loadings	2.55	1.48	0.79	0.61			
MR1	1.00	0.61	0.55	0.16			
MR3	0.61	1.00	0.35	0.03			
MR4	0.55	0.35	1.00	0.26			
MR2	0.16	0.03	0.26	1.00			

B.3.3.2 Usage of Outside Option “Snake” by Study Participants

209 of the 248 subjects had the option to abort the current task and to instead perform an alternative activity. However, they could not return to the task to earn money once they had quitted the task. After the timer for the current task expired, they continued with the next task like the other study participants. The usage of the outside option varies across tasks. Overall, relatively few subjects decided to abandon the task and to stop earning money as Table B.24 shows. Striking is the number of subjects who switched to the outside option in the multiplication task. Nearly 10% of the participants decided to perform a more pleasurable activity instead. This substantiates that this task is strongly disapproved by a share of the study participants.¹⁵

Table B.24: **Outside option usage pattern:** The number and share of subjects that switched to the outside option *Snake* are reported for each task. Besides, the average number of seconds spent on the game is given per subject in general and for those subjects who switched.

	Multipli- cation	Transcribe Words	Transcribe Codes	Word En- ryption	ab-Typing	Single Slider	Catching Balls
Subjects using outside option Snake	20	1	2	2	2	4	0
Share of subjects switching to Snake	0.096	0.005	0.01	0.01	0.01	0.019	0
Use of Snake in sec per subject	13.88	1.052	1.04	2.008	0.9718	2.202	0
Use of Snake in sec per subject who switched	172.1	261	129	249	120.5	136.5	

¹⁵As noted in Appendix B.3.4.2, the average valuation of subjects who strongly disapproved the multiplication task in the Personal Hit-List was 7.15 compared to 4.89 for the remainder of the subjects (the scale ranged from 1 for a strongly liked activity to 9 for a strongly disliked activity).

B.3.3.3 Personal hit-list

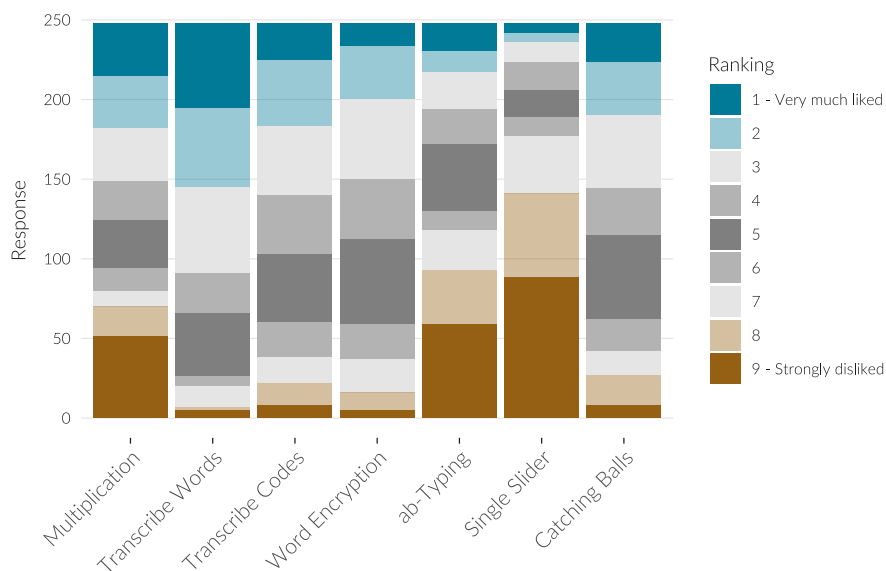


Figure B.59: **Distribution of the responses to the Personal Hit-List for all tasks:** The results align very well with the findings from the real-effort task survey.

To create the individual ranking scale for the Personal Hit-List, subjects were asked to name one activity that they currently very much like to do and one that they strongly dislike to get involved in (see Section 3.4.5 for a more detailed description and Figure B.46 for an illustration of the survey implementation). The majority of the activities stated by the subjects are summarized in Table B.25. Although it was intended that the subjects list activities from their everyday life, some of them decided to use the tasks they had just completed as a frame of reference. Rather strikingly, some tasks were primarily assigned to the “positive activity” to be named (the transcription tasks as well as the ball-catching task), while others were mainly ascribed to the “negative activity” to be named (ab-typing task as well as the single-slider task). In addition, the multiplication task is encountered with mixed feelings: some subjects used it as a reference point for the “positive activity,” but even more employed it for the “negative activity.” These observations also closely mirror those in Appendix B.3.4.2. As expected, the spectrum of other activities specified varies greatly for both the positive and negative directions. Of central importance for the construction of the individual scale and in order to enable further intra- and inter-subject comparisons is above all that the frame of reference was formulated in the same way in which the subjective ranking can then take place.

Table B.25: **Activities subjects stated as strongly disliked activity and strongly liked activity**

	Strongly disliked activity	Strongly liked activity
<i>Tasks</i>		
Multiplication	19	23
Transcription tasks	6	32
Word Encryption	3	4
ab-Typing	22	5
Single Slider	32	1
Catching Balls	1	25
<i>Working and studying</i>		
Working	14	0
University (general)	30	0
University (exam)	15	0
<i>Housework and other</i>		
Housework (general)	39	0
Housework (dishes)	9	0
Non-sense work	21	0
<i>Hobbies</i>		
Hobbies (general)	1	35
Sports	0	46
Video gaming	0	16
<i>Social life and needs</i>		
Socializing	0	28
Basic needs	0	15
<i>Share of subjects</i>	0.85	0.93

To assess the correlation between the real-effort task survey and the Personal Hit-List, subjects' responses to all items of the real-effort task survey were added for each task. For every study participant, this single score provides a composite measure of the individual perception of a task along a multitude of dimensions. The overall impression of a task is evaluated by the personnel-hit list, which represents a subjective ranking scale prepared by each study participant. Figure B.60 presents the Pearson correlation between the two measures, with the aggregate score of the real-effort task survey is tracked on the vertical axis and the Personal Hit-List ranking recorded on the horizontal axis. Correlations are color-coded according to whether they are **positive** or **negative**; non-significant correlations have been omitted. In the upper right triangle, the probability values are corrected for multiple testing. The similarities between the tasks, which were discussed several times and on which the design of the contrasts defined in Section 3.3.2.1 was also based, are again clearly evident (com-

pare also Figure 3.5).

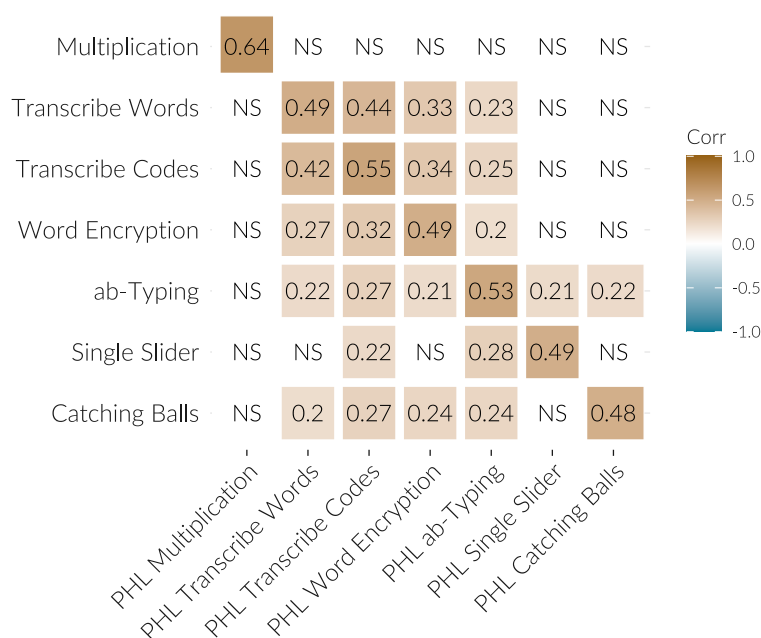


Figure B.60: **Correlation of the real-effort task survey and the Personal Hit-List:** The real-effort task survey is considered as a summated rating scale, aggregating subjects' responses to all survey items for each task. Both measures are coded in the same way, such that lower values resemble a more preferred task.

B.3.4 Clustering Subjects According to Their Preferences

B.3.4.1 Identifying Subject Types With the Real-Effort Task Survey

The following scatter plots depict the results of clustering analysis for three different tasks: the *multiplication task* (Figure B.61), the *single-slider task* (Figure B.62) and the *ball-catching task* (Figure B.63).¹⁶ For each of these tasks, subjects were assigned to one of two clusters (*petrol blue* or *brown*) according to their ratings. The graphs show the average responses for the subjects in each cluster for the respective task, whereby the number of subjects per cluster is indicated on the right. The composite graph further contains plots for the remaining tasks. These reveal how the two groups of subjects identified for task i rate the other tasks $j \neq i$ respectively (the clusters are colored in the same shade but slightly lighter).

¹⁶Clustering was performed using the k -means algorithm across all items of the real-effort task survey. Silhouette-width analysis as well as elbow plots (with k -means clustering for different values of k) suggested $k = 2$ as number of clusters.

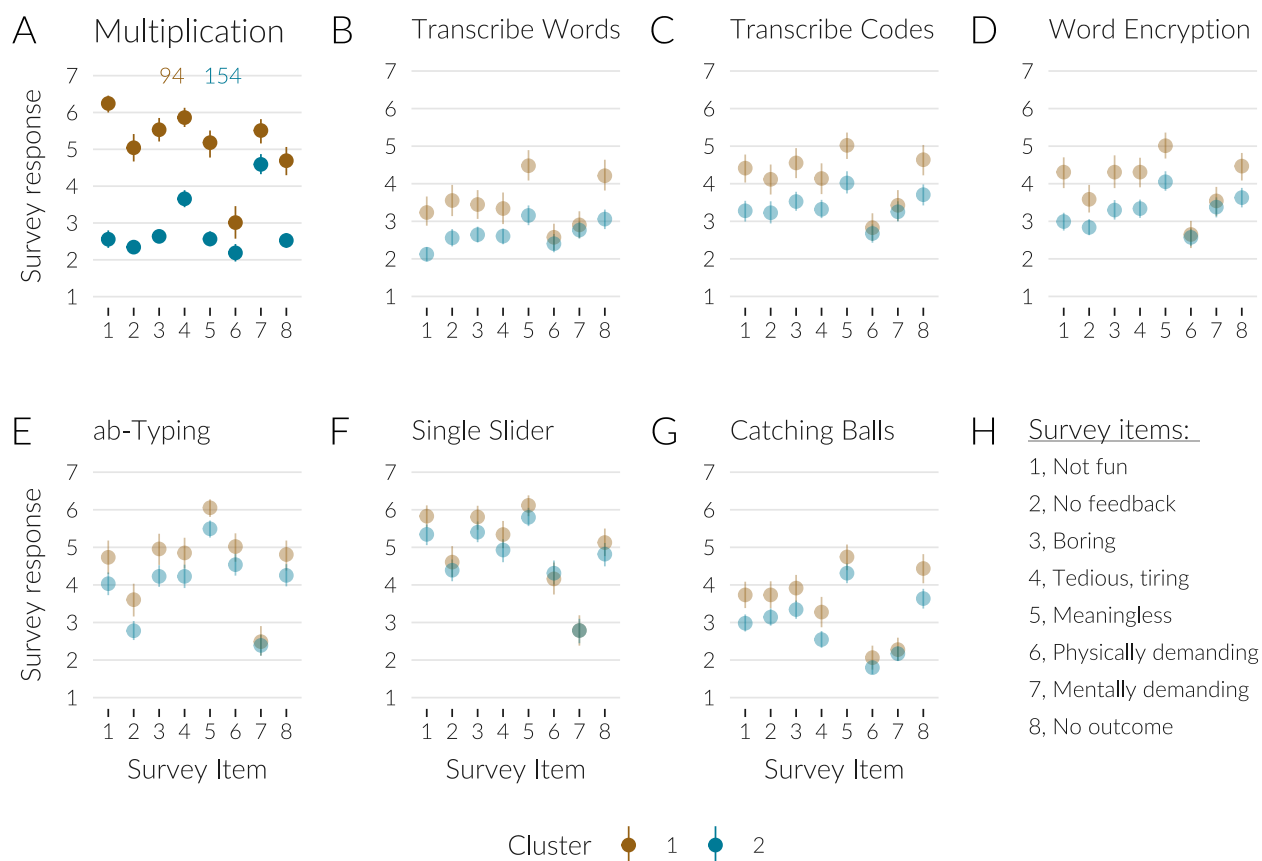


Figure B.61: **Mean task ratings for clusters based on subjects' rating of the multiplication task:** Using *k-means clustering*, two clusters of subjects are formed based on their responses to all items of the real-effort task survey for the multiplication task (petrol blue and brown, with the number of subjects per cluster indicated). Based on the clustering for this task, the mean responses for all other tasks and survey items are depicted in separate plots (in lighter shades of petrol blue and brown). The clustering can nicely split apart two groups of subjects for the multiplication task. The petrol-blue cluster of subjects who rate the multiplication task as not so enjoyable also find the other cognitive tasks not so attractive (subjects gave high responses to most survey items). Interestingly, both clusters of subjects do not differ much in their views of the single-slider task.

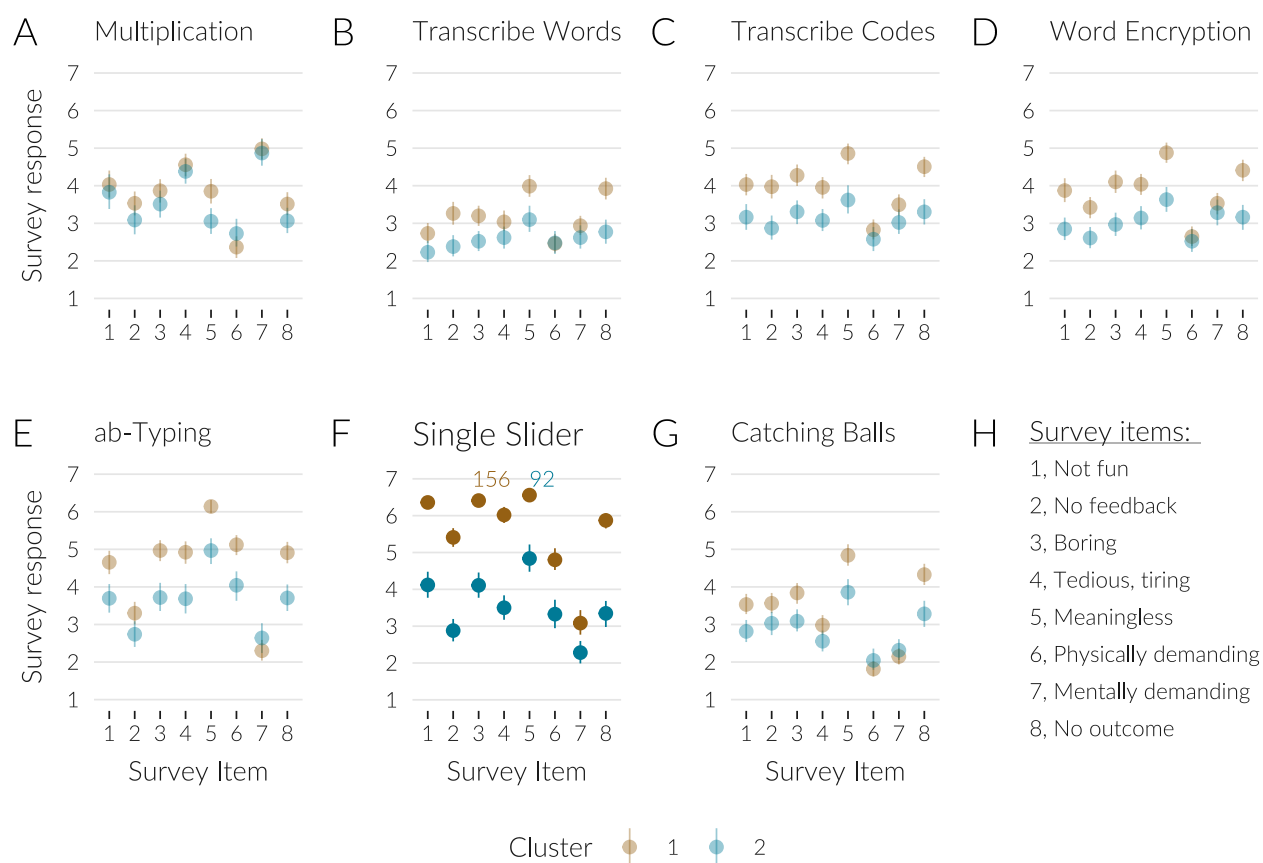


Figure B.62: **Mean task ratings for clusters based on subjects' rating of the single-slider task:** The k-means clustering algorithm again identifies two clusters for the single-slider task. The **brown cluster** consist of a very large number of subjects, all of whom perceived the task as less enjoyable. The identified clusters also show different attitudes towards the ab-typing task and partially towards the memory tasks. For the remaining tasks, the differences between the two clusters are not particularly large, especially for the multiplication task.

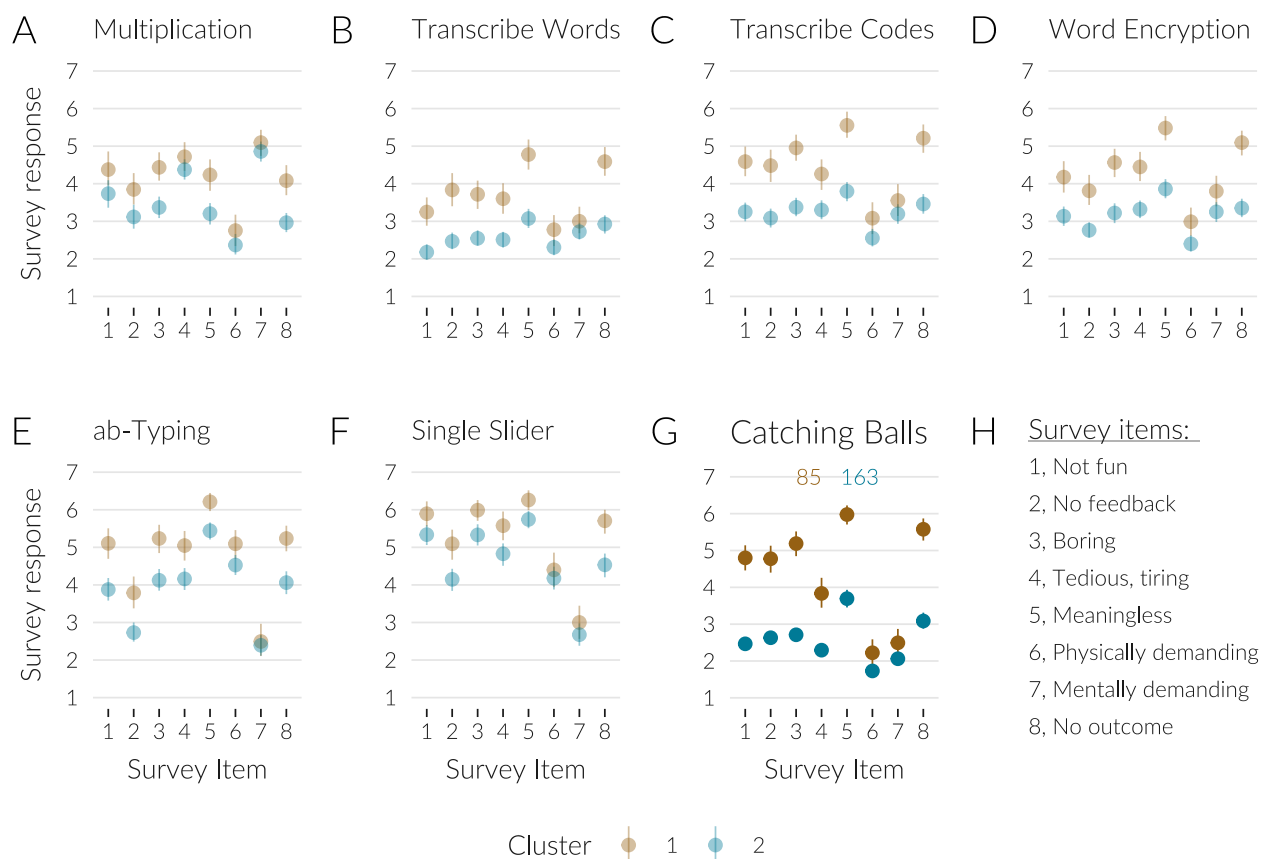


Figure B.63: **Mean task ratings for clusters based on subjects' rating of the ball-catching task:** The two clusters found for the ball-catching task have very different views of the task, with the exception of the two effort items: All subjects seem to have virtually the same view of the task's requirement for physical and mental effort. The two clusters identified for this task also diverge quite considerably in their task perceptions for the other tasks. Subjects who do not find the ball-catching task as appealing, thus giving **high responses** to the survey, think similarly about the memory tasks in particular.

B.3.4.2 Identifying Subject Types With the Personal Hit-List

The results of the Personal Hit-List further allow to identify groups of subjects with similar preferences. In a first step, two groups of subjects are defined *for each task*. The “lovers of task” rate the task ≤ 2 ; the “haters of task” assign it a rating ≥ 8 . Table B.26 lists the number of lovers and haters determined for each task according to this definition. The following Table B.27 shows for the lovers of a particular task (rows) the average rating of all tasks (columns), with their mean rating subtracted. Similarly Table B.28 depicts the same for the haters of a specific task. If these subjects have a tendency to like another task, the average rating is colored in **brown**, whereas if they tend to dislike the task

it is colorized in petrol blue.¹⁷

Table B.26: **“Lovers” and “haters” of tasks according to the Personal Hit-List:** Subjects who indicated that they had a greater preference for a task (task rating in the Personal Hit-List ≤ 2) are defined as “task lovers”; in contrast, those who more or less equated the task with an activity that they currently highly dislike to perform are referred to as “task haters” (rating ≥ 8).

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
Task lovers	65	102	64	47	31	11	58
Task haters	70	7	22	15	92	141	27

Table B.27: **Average rating of all tasks for the “lovers of a particular task”:** For the lovers of a given task (task rating in the Personal Hit-List is ≤ 2) the average valuation of all tasks is calculated. From the obtained values, the mean rating of all subjects of the respective task is subtracted and presented in the table. Subjects who love the multiplication task are not particularly fond of any other task. Those who like any of the transcription or memory tasks enjoy the other tasks in this category of tasks as-well. Interestingly, subjects who favor the word-encryption may also like the ball-catching task, possibly because both tasks present a certain visual challenge. Subjects who strongly enjoy one of the two mechanical tasks very much also appreciate the other task. In addition, they are fond of the code-transcription task but dislike the multiplication task. This holds particularly strongly for the single-slider task. However, there are only eleven subjects who love this task. Subjects who prefer the ball-catching tasks are also inclined to favor the ab-typing task.

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
Multiplication	-3.39	-0.31	-0.26	0.13	0.42	0.19	0.33
Transcribe Words	-0.25	-1.75	-1.09	-0.64	-0.18	0.21	0.01
Transcribe Codes	-0.44	-1.28	-2.49	-1.11	-0.82	-0.41	0.02
Word Encryption	-0.12	-0.88	-1.37	-2.57	-0.35	-0.22	-0.71
ab-Typing	0.56	-0.4	-1.43	-1.02	-4.44	-1.38	-0.47
Single Slider	2.11	-0.4	-1.64	-0.61	-2.9	-5.56	-0.11
Catching Balls	0.47	-0.12	-0.33	-0.55	-0.88	-0.33	-2.69

¹⁷Employing petrol blue for negative values would have been more intuitive. However, the Personal Hit-List is originally coded such that activities with lower values are preferred and those with higher values disliked. Preferred activities are thus colored in the positive color.

Table B.28: **Average rating of all tasks for the “haters of a particular task”**: Same procedure as in Table B.27 was applied for all subjects who strongly dislike a given task (rating in the Personal Hit-List is ≥ 8). Subjects who detest the multiplication task favor the single-slider task instead. Those who have an aversion to one of the transcription or memory tasks also disregard all the other tasks in that category as well as the multiplication task. Interestingly, subjects who dislike memory tasks are to some degree averse to the ab-typing task. This suggests that even simple typing tasks require a certain amount of memory to remember the correct keys on the keyboard. Subjects who loathe one of the two mechanical tasks are equally not very enthusiastic about the other. There are 92 subjects who strongly disapprove of the ab-typing tasks, whereas 141 detest the single-slider task. Subjects who strongly dislike the ball-catching tasks also have an aversion to the mechanical tasks.

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
Multiplication	3.84	0.16	0.32	0.15	-0.42	-0.61	-0.27
Transcribe Words	1.82	5.48	2.43	2.01	0.82	-0.06	0.87
Transcribe Codes	0.88	2.22	4.22	1.68	0.92	0.17	0.45
Word Encryption	0.61	2.14	2.3	4.03	0.42	0.13	-0.02
ab-Typing	-0.33	0.06	0.29	0.19	2.74	0.85	0.63
Single Slider	-0.41	-0.18	0.13	0.17	0.74	1.57	0.3
Catching Balls	-0.19	0.39	0.01	-0.17	0.7	0.31	4.02

B.4 Supporting Documents

B.4.1 Official Approval to Conduct the Experiments



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ETH Zürich
Herr Christian Waloszek
Professur für Öffentliche Finanzen
LEE G 112
Leonhardstrasse 21
8092 Zürich

**Vizepräsident für Forschung und
Wirtschaftsbeziehungen**

ETH Zurich
Prof. Dr. Detlef Günther
HG F 57
Rämistrasse 101
8092 Zürich

Kontakt:
Stab Forschung
iturizaga@sl.ethz.ch

10. Juli 2018 ri
EK 2018-N-38

Sehr geehrter Herr Waloszek

Ihr Gesuch

Experimental study of individual effort across different real-effort tasks (03.04.2018)
wurde online durch die folgenden Mitglieder der Ethikkommission beurteilt:

Name	Institut	am Beschluss beteiligt		
		ja	nein (Grund)	
			abwesend	im Ausstand
Prof. Dr. Lutz Wingert, Präsident	Professur für Philosophie		X	
Prof. Dr. Jörg Goldhahn	Institut für Translationale Medizin	X		
Prof. Dr. Otmar Hilliges	Institut für Pervasive Computing	X		
Prof. Dr. Christoph Hölscher	Professur Kognitionswissenschaften		X	
Prof. Dr. Julian Mausbach	Rechtswissenschaftliches Institut	X		
Dr. Marino Menozzi	Innovations- und Technologiemanagement		X	
Dr. Kai-Uwe Schmitt	AGU Zürich	X		
Prof. Dr. Michael Siegrist	Institut für Umweltentscheidungen	X		
Prof. Dr. William R. Taylor	Institute for Biomechanics		X	
Prof. Dr. Effy Vayena	Institut für Translationale Medizin	X		
Dr. Peter Wolf	Institut für Robotik und intelligente Systeme	X		

Aufgrund der Empfehlung der Ethikkommission der ETH Zürich ist die Schulleitung zu folgendem Beschluss gekommen:

- Bewilligung
- Bewilligung mit Vorbehalt (schriftliche Mitteilung an Ethikkommission ausreichend)
- Rückweisung zur Überarbeitung mit Auflage
 - Schriftliche Mitteilung an Ethikkommission ausreichend
 - Neubegutachtung durch Ethikkommission notwendig
- Negativ (mit Begründung und Erläuterung für Neubeurteilung)
- Nicht-Eintreten (mit Begründung)

Vorbehalte

Stellungnahmen der Mitglieder der Ethikkommission zu Ihrem Gesuch:

o) Vorbehalte Generell

- a) Es erschliesst sich nicht, warum die Studie in Hamburg durchgeführt werden soll. Begründen Sie bitte inwiefern die dortigen Bedingungen eine Durchführung an diesem Ort als günstiger erscheinen lassen.
- b) The paragraph on insurance coverage via ETH should be sanity checked. Since the experiment is entirely conducted in Hamburg - the insurance situation surely is complex and should be treated carefully.
- c) Gemäss Gesuch ist die Universität Basel kein Forschungspartner. Warum darf sie gemäss Info-Blatt die Daten verwenden? Bitte Rolle der Uni Basel erklären.

1) Abstract

-

2) Projekt

- a) Der Vorgang der Anonymisierung sollte nachvollziehbarer dargestellt werden. Was wird exakt von den Daten entnommen oder abgeschnitten, damit die Anonymität gegeben ist. Behält die Universität Hamburg einen Datensatz, der diesen Anonymisierungsprozess nicht durchläuft? Die angegebenen Guidelines sind diesbezüglich nicht ausreichend bzw. konkret genug betreffend das vorliegende Projekt.
- b) Bitte Termine für die Projektlaufzeit anpassen.
- c) Fragebogen/ Teil 8: bei sozio-demographischen Fragen bitte ein Feld mit „keine Angabe“ oder die Möglichkeit des Überspringens der Frage einfügen - insbesondere bei Fragen, bei denen nur wenige, sehr traditionelle Antwortmöglichkeiten angeboten werden. Zudem das Mindestalter anpassen; es muss 18 Jahre betragen (nicht 17).

3) Zu erwartende Risiken und entsprechende Vorsichtsmassnahmen

-

4) Projektleiter/-in

-

5) Probanden/-innen

-

6) Informationsblatt für Probanden/-innen

-

7) Einverständniserklärung

-

Fachliche Kommentare

Die fachlichen Kommentare sind nur als Hinweise zu verstehen und müssen im Falle einer Überarbeitung nicht berücksichtigt werden.

-

Wir machen Sie darauf aufmerksam, dass gegenüber der Ethikkommission der ETH Zürich in folgenden Situationen eine Meldepflicht besteht:

- a) Unverzüglich bei Auftreten von unerwarteten Ereignissen, welche die Sicherheit der Versuchspersonen und/oder die Weiterführung des Versuches beeinflussen können
- b) Bei Änderungen am Forschungsprotokoll und bei Versuchspersonen
- c) Bei Abbruch der Studie

Freundliche Grüsse

Prof. Detlef Günther
Vizepräsident für Forschung &
Wirtschaftsbeziehungen

Prof. Lutz Wingert
Vorsitzender der Ethikkommission

cc: Departementsvorsteher MTEC



UHH – Datenschutzbeauftragter – Mittelweg 177 – 20148 Hamburg

Universität Hamburg
Fakultät für Wirtschafts- und
Sozialwissenschaften
WISO-Forschungslabor
Herrn Olaf Bock

per E-Mail

Bernd Uderstadt

Datenschutzbeauftragter
Stabsstelle Recht, R16 / DSB
Mittelweg 177
Raum N0051
20148 Hamburg
Tel. +49 (0) 40 - 42838 -2957
datenschutz@verw.uni-hamburg.de
www.uni-hamburg.de

UHH-R16/DSB - 920.9410-0010/002/058:0008

Hamburg, 13.04.2018

Forschungsprojekt WALO2 + WALO-Schrittstudien

Dortiges Geschäftszeichen: 2018_03_29_WALO_2, 2018_03_29_WALO_Schrittstudien

Sehr geehrter Herr Bock,

die anliegenden, abgestimmten Verfahrensbeschreibungen habe ich in die Verfahrensübersicht der Universität Hamburg aufgenommen. Sie werden bei mir unter dem Geschäftszeichen ‚UHH/DSB 920.9410-0010/002/058:0008‘ (wie bereits für WALO_1 verwendet) geführt.

Bei technischen, organisatorischen oder rechtlichen Änderungen im Fachverfahren bitte ich um entsprechende Überarbeitung und Zusendung einer aktualisierten Version.

Nach § 19 Abs. 6 HmbDSG (Hamburgisches Datenschutzgesetz) sind die gespeicherten Daten regelmäßig alle vier Jahre von der (fach-)verantwortlichen Stelle auf ihre Erforderlichkeit hin zu überprüfen.

Im Hinblick auf die aktuelle Novellierung des Datenschutzrechts (am 25. Mai 2018 wird die Europäische Datenschutzgrundverordnung - [EU-DSGVO](#) - unmittelbar wirksam), bleibt abzuwarten, inwieweit diese Vorschrift Bestand haben wird. Die Neufassung des Hamburgischen Datenschutzgesetzes (HmbDSG) wird derzeit im Bürgerschaftsausschuss ‚Justiz und Datenschutz‘ strittig erörtert..

Den öffentlichen Teil der Verfahrensbeschreibung (Teil A = Ziffern 1-7) halte ich auf Antrag durch jede Person bis zum 24. Mai 2018 zur Einsichtnahme vor. Nicht einsehbar ist gemäß Ihrer Vorgabe der Teil B der Verfahrensbeschreibung (Ziffern 8 und 9).

Mit Wirksamkeit der EU Datenschutzgrundverordnung (EU-DSGVO) am 25.05.2018 entfällt diese vom Datenschutzbeauftragten wahrzunehmende Veröffentlichungspflicht der verantwortlichen Stelle - künftig ‚Verantwortlicher‘ genannt - ersatzlos.

Bis zu diesem Zeitpunkt sind die entsprechenden Informationen der Verfahrensbeschreibung in das künftige ‚Verzeichnis von Verarbeitungstätigkeiten‘ zu überführen; vgl. [Art 30 EU-DSGVO](#). Zum gegenwärtigen Zeitpunkt ist mir noch nicht bekannt, von welcher Stelle der Universität Hamburg das Verzeichnis künftig verwaltet und gepflegt wird.

Mit freundlichem Gruß

Bernd Uderstadt

B.4.2 Financial Support: Grant From the MTEC Foundation

Subject: Application for funding through the MTEC Foundation
Date: Friday, 23 December 2016 at 13:12:16 Central European Standard Time
From: Felix Tobler
To: Waloszek Christian
CC: Christopher Klenk

Sehr geehrter Herr Waloszek

Sie haben für das Projekt „Active through Incentive Mechanisms“ ein Forschungsförderungsgesuch an die MTEC Foundation gerichtet. Dafür danken wir Ihnen und Ihrem Projektpartner.

Der Stiftungsrat hat sich an seiner Sitzung vom 30. November 2016 mit den eingereichten Gesuchen befasst und über die Vergabe der zur Verfügung stehenden Fördermittel entschieden. Von den eingereichten 15 Anträgen konnten 4 berücksichtigt werden.

Ich freue mich, Ihnen mitteilen zu können, dass der Stiftungsrat Ihr Gesuch im beantragten Umfang von **CHF 52'750** gutgeheissen hat. Der Betrag steht Ihnen ab sofort zur Verfügung.

Um den Förderbetrag abzurufen, wollen Sie sich bitte an mich wenden. Damit ich die Auszahlung veranlassen kann, benötige ich einen **Auszahlungsantrag** (schriftlich oder via E-Mail); ein Satz genügt. Der Antrag muss die genaue **Zahlungsverbindung** (in der Regel das Konto der ETH Zürich bei der Schweizerischen Nationalbank) und den **Zahlungsvermerk** enthalten.

Wir bitten Sie, dem Stiftungsrat nach Abschluss des Projekts unaufgefordert einen kurzen schriftlichen **Abschlussbericht** zu erstatten (an meine Adresse). Trägt das unterstützte Projekt ursächlich oder mitursächlich zum Entstehen einer **Publikation** bei, bitten wir Sie, uns ein Belegexemplar zur Verfügung zu stellen.

Im Namen des Stiftungsrates wünsche ich Ihnen einen erfolgreichen und befriedigenden Projektverlauf. Für Rückfragen stehe ich Ihnen gerne zur Verfügung.

Herzliche Grüsse

Für den Stiftungsrat der MTEC Foundation:

Felix Tobler, Geschäftsführer

Rämistrasse 3
 8024 Zürich
 Telefon 044 251 50 90
tobler@tobler-law.ch
www.tobler-law.ch

Figure B.69: **Financial support for the laboratory experiments:** The project was funded by the *MTEC Foundation* through a research grant awarded to Christopher Klenk and Christian Waloszek in December 2016.

C

Appendix to Chapter 4

C.1 Experimental Design

C.1.1 Characterization: List of Psychological Questionnaires

	Construct	Measure	Abbrev.	Reference	Year	Items
1 Directly Task-related Ability Dimensions						
1.1 Physical Skills and Endurance						
1	Computer usage	Frequency of computer usage	PC_freq	Own measure	2018	1
2	Computer skills	Ability to use 10-finger typing	PC_skill	Own measure	2018	1
3	Colorblindedness	Assess difficulty to differentiate colors on computer screen	colorblind	Own measure	2018	1
4	Real-effort task difficulty	Personal perception of task difficulty of the real-effort task	Exp2_RET survey	Own measure	2018	2
1.2 Concentration: Ability to Focus						
1	Self-control	Self-control Measure	SCM	Ameriks, J., Caplin, A., Leahy, J., Tyler, T., 2007. Measuring Self-Control Problems. <i>American Economic Review</i> 97, 966–972.	2007	4
2	Temptation and Impulsiveness	Susceptibility to Temptation Scale	STS	Steel, 2002. The measurement and nature of procrastination. ProQuest Information & Learning.	2002	11
3	Self-Regulation: Action- versus state-orientation	HAKEMP - ACS-24 German Version	HAKEMP	Kuhl, J. (1990). Kurzanweisung zum Fragebogen HAKEMP 90. Manuskript. Fachbereich Psychologie, Universität Osnabrück.	1990	24
4	Grit	Grit Scale German Version	BISS8	Schmidt, F. T., Fleckenstein, J., Retelsdorf, J., Eskreis-Winkler, L., & Möller, J. (2017). Measuring Grit. <i>European Journal of Psychological Assessment</i> .	2017	8
1.3 Vigilance/Attention (Wachsamkeit/ Aufmerksamkeit)						
1	Affectivity: State during task execution	Positive Aktivierung, Negative Aktivierung, Valenz	PANAVA KS	Schallberger, U. (2005). Kurzskalen zur Erfassung der Positiven Aktivierung, Negativen Aktivierung und Valenz in Experience Sampling Studien (PANAVA-KS). Theoretische und methodische Grundlagen, Konstruktvalidität und psychometrische Eigenschaften bei der Beschreibung intra- und interindividueller Unterschiede.	2005	10
2	Attention	Psychomotor Vigilance Task	PVT	Dinges, D., Pack, F., Williams, K., Gillen, K., Powell, J., Ott, G., Aptowicz, C., Pack, A., 1997. Cumulative Sleepiness, Mood Disturbance, and Psychomotor Vigilance Performance Decrements During a Week of Sleep Restricted to 4–5 Hours per Night. <i>Sleep</i> 20, 267–277.	1997	2 min
1.4 Quantitative & Analytical Reasoning (number, math)						
1	Cognitive Ability	Berlin Numeracy Test	BNT	Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. <i>Judgment and Decision Making</i> , 7(1), 25.	2012	4
2	Cognitive Ability	Subjective Numeracy Scale	SNS	Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the Subjective Numeracy Scale. <i>Medical Decision Making</i> , 27(5), 672–680.	2007	7
3	Cognitive Ability	Kurzskala kristalline Intelligenz	KKI	Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C., Rammstedt, B., 2014. Kurzskala kristalline Intelligenz (BEFKI GC-K). Zusammenstellung sozialwissenschaftlicher Items und Skalen. doi 10.	2014	12

Figure C.1: **List of characterization questionnaires:** Psychological measures employed to assess subject characteristics (a more detailed version is available upon request).

1.5 Working Memory

1	Working Memory	2-back-task	RET_dual_2_back_task	Jaeggi, S., Buschkuhl, M., Perrig, W., Meier, B., 2010. The concurrent validity of the N-back task as a working memory measure. <i>Memory</i> 18, 394–412 Instructions: CW	2010	2 min
---	----------------	-------------	----------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	-------

1.6 Reading and Writing Ability

1	Language skills	Standard Anagram Task	RET_anagram	Ammons, Ammons, 1959. A Standard Anagram Task. <i>Psychol Rep</i> 5, 654–656. Instructions: CW	1959	6
---	-----------------	-----------------------	-------------	---------------------------------------------------------------------------------------------------	------	---

2 Decision-Making Measures

1	Decision approach: Tendency to postpone decision	Prokrastinationsfragebogen für Studierende	PFS	Glöckner-Rist, A., Engberding, M., Höcker, A., & Rist, F. (2009). Prokrastinationsfragebogen für Studierende (PFS). In <i>Zusammenstellung sozialwissenschaftlicher Items und Skalen</i> , hg. von Angelika Glöckner-Rist, Bonn.	2014	7
---	--------------------------------------------------	--------------------------------------------	-----	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	---

3 Motivation Measures

3.1 Social Desirability

1	Social Desirability	Questions based on „learning motivation scale“	Q_MOT	Wild, K. P., Krapp, A., Schiefele, U., Lewalter, D., & Schreyer, I. (1995). Dokumentation und Analyse der Fragebogenverfahren und Tests. <i>Berichte aus dem DFG-Projekt „Bedingungen und Auswirkungen berufsspezifischer Lernmotivation</i> , (2).	1995	6
---	---------------------	------------------------------------------------	-------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	---

3.2 Self-Regulation

1	Self-Regulation: Incentive responsiveness	BIS/BAS German version	BISBAS	Strobel, A., Beauducel, A., Debener, S., & Brocke, B. (2001). Eine deutschsprachige Version des BIS/BAS-Fragebogens von Carver und White. <i>Zeitschrift für Differentielle und Diagnostische Psychologie</i> .	2001	20
2	Interest/enjoyment	KIM: Interesse (interest)	IMIs_I	Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzsкала intrinsischer Motivation (KIM). <i>Zeitschrift für Didaktik der Naturwissenschaften</i> , 15.	2009	3
3	Competence	KIM: Wahrgenommene Kompetenz (perceived competence)	IMIs_C	Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzsкала intrinsischer Motivation (KIM). <i>Zeitschrift für Didaktik der Naturwissenschaften</i> , 15.	2009	3
4	Pressure	KIM: Druck (perceived tension)	IMIs_P	Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzsкала intrinsischer Motivation (KIM). <i>Zeitschrift für Didaktik der Naturwissenschaften</i> , 15.	2009	3
5	Autonomy	KIM: Wahrgenommene Wahlfreiheit (perceived choice)	IMIs_A	Wilde, M., Bätz, K., Kovaleva, A., & Urhahne, D. (2009). Überprüfung einer Kurzsкала intrinsischer Motivation (KIM). <i>Zeitschrift für Didaktik der Naturwissenschaften</i> , 15.	2009	3

2

Figure C.2: **List of characterization questionnaires (cont.):** page 2.

4 Personality Measures

4.1 Personality Inventories

1	Personality	Big Five Inventory (10-item short version)	BFI10	Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. <i>Journal of research in Personality</i> , 41(1), 203-212.	2007	10
---	-------------	--------------------------------------------	-------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	----

4.2 Personality-construct Measures

4.2.1 Control Measures						
1	Locus of control	Internale-Externale-Kontrollüberzeugung-4	IE4	Kovaleva, A., Beierlein, C., Kemper, C., & Rammstedt, B. (2014). Internale-Externale-Kontrollüberzeugung-4 (IE-4). In D. Danner, & A. Glöckner-Rist, <i>Zusammenstellung sozialwissenschaftlicher Items und Skalen</i> .	2014	4
4.2.2 Time-Orientation Measures						
1	Patience	Validating an Ultra Short Survey measure of patience	PATs	Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., & Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. <i>Economics Letters</i> , 120(2), 142-145.	2013	1
2	Short- vs. long-term consequences	Consideration of Future Consequences	CFC	Strathman, A., Gleicher, F., Boninger, D. S., & Edwards, C. S. (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. <i>Journal of Personality and Social Psychology</i> , 66(4), 742-752.	1994	12

5 Diagnosis of Incentive Qualities

1	Task preference: Relative task attractiveness	Persönliche Hitliste	PHL	Rheinberg, F. (2004). <i>Motivationsdiagnostik: Kompendien. Psychologische Diagnostik</i> . Göttingen:Hogrefe	1989/ 2004	1
2	Task experience: qualitative assessment of task attractiveness	Perception of and joy in performing the real-effort task	Exp2_RET survey	Own measure	2018	7
3	Task vs. purpose motivation	Anreiz-Fokus Skala	AFS	Rheinberg, F., Iser, I., & Pfauter, S. (1997). Freude am Tun und/oder zweckorientiertes Schaffen? Zur transsituativen Konsistenz und konvergenten Validität der Anreizfokus-Skala. <i>DIAGNOSTICA-GOTTINGEN-</i> , 43, 174-191.	1997	20

6 Assessing the Attractiveness of Results of Action

1	Performance motive	Achievement Motives Scale	AMS	Engeser, S. (2005). Messung des expliziten Leistungsmotivs: Kurzform der Achievement Motives Scale. Retrieved on 02/10/2017 from https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PGA/bilder/Engeser__2005__Kurzform_der_AMS.pdf .	2005	10
---	--------------------	---------------------------	-----	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	----

3

Figure C.3: **List of characterization questionnaires (cont.):** page 3.

2	Motives: Affiliation	Unified Motive Scale: Affiliation	UMS	Schönbrodt, F. D., & Gerstenberg, F. X. (2012). An IRT analysis of motive questionnaires: The unified motive scales. <i>Journal of Research in Personality</i> , 46(6), 725-742.	2012	6
3	Motives: Power	Unified Motive Scale: Power	UMS	Schönbrodt, F. D., & Gerstenberg, F. X. (2012). An IRT analysis of motive questionnaires: The unified motive scales. <i>Journal of Research in Personality</i> , 46(6), 725-742.	2012	3
4	Motives: Achievement	Unified Motive Scale: Achievement	UMS	Schönbrodt, F. D., & Gerstenberg, F. X. (2012). An IRT analysis of motive questionnaires: The unified motive scales. <i>Journal of Research in Personality</i> , 46(6), 725-742.	2012	10
5	Motives: Fear	Unified Motive Scale: Fear	UMS	Schönbrodt, F. D., & Gerstenberg, F. X. (2012). An IRT analysis of motive questionnaires: The unified motive scales. <i>Journal of Research in Personality</i> , 46(6), 725-742.	2012	12

7 Self-efficacy Expectations

1	Self-Efficacy	Allgemeine Selbstwirksamkeit Kurzsкала	AKSU	Beierlein, C., Kemper, C., Kovaleva, A., & Rammstedt, B. (2013). Kurzsкала zur Erfassung allgemeiner Selbstwirksamkeitserwartungen (ASKU). <i>Methoden, Daten, Analysen (mda)</i> , 7(2), 251-278.	2014	3
---	---------------	----------------------------------------	------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------	---

8 Controls

8.1 Socio-demographics

1	Gender	Demographische Standards 1	gender	Hoffmeyer-Zlotnik, J. H., Glemser, A., Heckel, C., von der Heyde, C., Quitt, H., Hanefeld, U., ... & Mohr, S. (2010). <i>Statistik und Wissenschaft: Demographische Standards Ausgabe 2010 (Band 17)</i> .	2010	1
2	Age	Demographische Standards 2	age	Hoffmeyer-Zlotnik, J. H., Glemser, A., Heckel, C., von der Heyde, C., Quitt, H., Hanefeld, U., ... & Mohr, S. (2010). <i>Statistik und Wissenschaft: Demographische Standards Ausgabe 2010 (Band 17)</i> .	2010	1
3	Nationality	Based on Demographische Standards 3	nationality	Hoffmeyer-Zlotnik, J. H., Glemser, A., Heckel, C., von der Heyde, C., Quitt, H., Hanefeld, U., ... & Mohr, S. (2010). <i>Statistik und Wissenschaft: Demographische Standards Ausgabe 2010 (Band 17)</i> .	2010	1
4	Relationship status	SOEP question 154a	relationship	Berlin, D. I. W. (2016). <i>SOEP 2016-Erhebungsinstrumente 2016 (Welle 33) des Sozio-oekonomischen Panels: Personenfragebogen, Stichproben A-L3 (No. 345)</i> . SOEP Survey Papers.	2010	1
5	Household	Household composition	household	Own measure	2017	1
6	Children	Number of children	children	Own measure	2017	1
7	Education level	Based on SOEP question 21	degree	Berlin, D. I. W. (2016). <i>SOEP 2016-Erhebungsinstrumente 2016 (Welle 33) des Sozio-oekonomischen Panels: Personenfragebogen, Stichproben A-L3 (No. 345)</i> . SOEP Survey Papers.	2010	1

4

Figure C.4: **List of characterization questionnaires (cont.):** page 4.

8	Life Satisfaction	Kurzskala zur Erfassung der Allgemeinen Lebenszufriedenheit	L-1	Beierlein, C., Kovaleva, A., László, Z., Kemper, C. J., Rammstedt, B. (2015). Kurzskala zur Erfassung der Allgemeinen Lebenszufriedenheit (L-1). Zusammenstellung sozialwissenschaftlicher Items und Skalen.	2015	1
9	Optimism & Hope	Skala Optimismus-Pessimismus-2	SOP2	Kemper, C., Beierlein, C., Kovaleva, A., & Rammstedt, B. (2012). Eine Kurzskala zur Messung von Optimismus-Pessimismus: Die Skala Optimismus-Pessimismus-2 (SOP2). GESIS.	2014	2
8.2 Study Participation						
1	Chronotype	Chronotype measurement	CIRENS	Ottoni, G. L., Antonioli, E., & Lara, D. R. (2011). The Circadian Energy Scale (CIRENS): two simple questions for a reliable chronotype measurement based on energy. <i>Chronobiology international</i> , 28(3), 229-237.	2011	2
2	Alertness	Assessment of attentiveness and quality of sleep in previous 2 nights	ALERT1-3	Own measure	2017	3
3	Literacy	Understanding of experimental instructions	literacy	Own measure	2017	1
4	Study participation	Reason/purpose of study participation	study_participation	Own measure	2017	1
5	Experiment content	Inquire if participants were familiar with content	exp_content	Own measure	2017	1
6	Honesty and concentration	Inquire if participant was concentrated & honest while answering surveys	concentration, honesty	Own measure	2018	2
					Total:	255

Figure C.5: List of characterization questionnaires (cont.): page 5.

C.2 Additional Figures and Tables

C.2.1 Descriptive Analysis

Table C.1 summarizes the main descriptive statistics for the set of real-effort tasks (Figure C.6 reiterates the score distributions to allow for a simultaneous graphical examination). The variance of the distributions differs widely from task to task. Especially for the multiplication task, but also for the ab-typing task, the scores are spread further away from the mean. The score distributions vary extensively in midspread (IQR), with the multiplication task having the most, the code-transcription task, the single-slider task and the ball-catching task the least. The multiplication task has positive skew such that the tail of its distribution is longer on the right. The remaining tasks have negative skew, although this is so small for the word-transcription task, word-encryption task, and the ab-typing task that their score distributions do not differ significantly from a normal distribution in terms of their skewness. Regarding kurtosis, only the code-transcription task is comparable to the normal distribution. The single-slider task and the ball-catching task possess heavy-tailed distributions, the remaining tasks have (very) light tails. To illustrate the skew and kurtosis of the distributions, the first and the fourth quartiles are highlighted in Figure C.6. For completeness, their numerical values are provided in Table C.1.

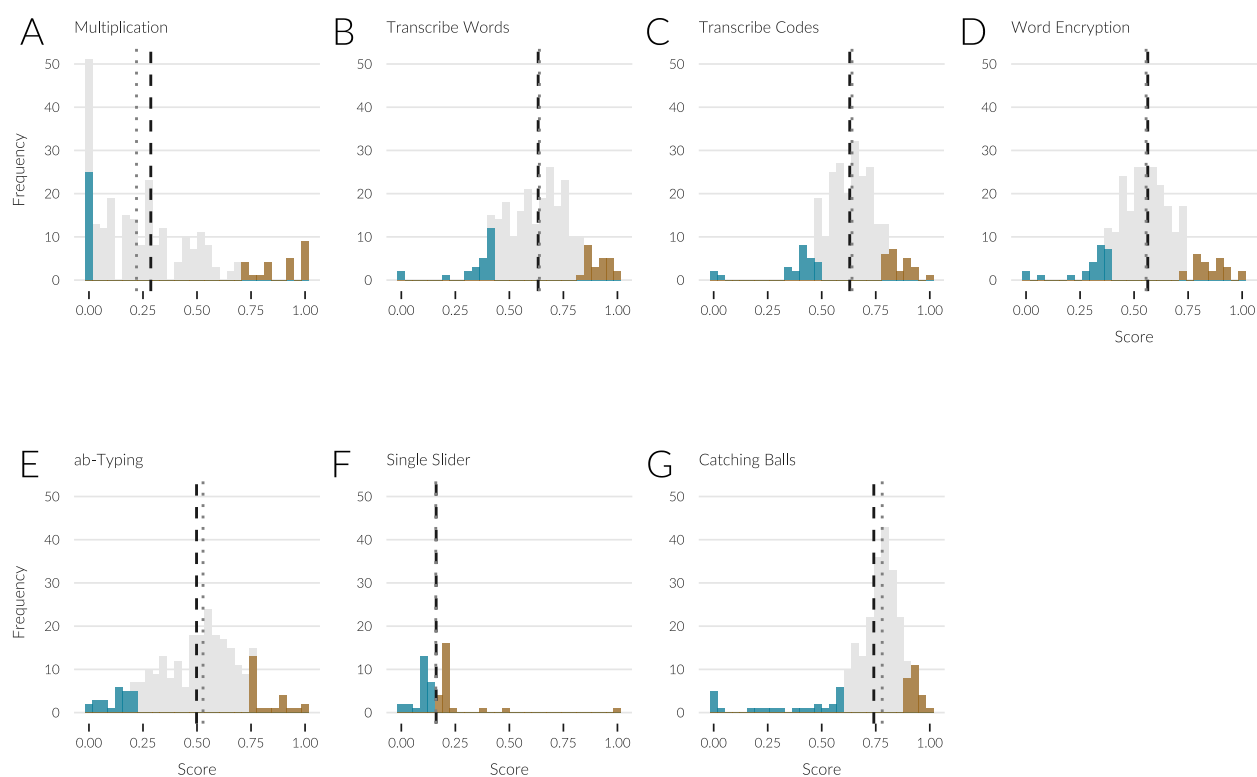


Figure C.6: **Score distributions for the selection of real-effort tasks:** To better compare the shape of the distributions, the scores are min-max normalized and plotted on the same y-axis. For each task, the median values (dotted) and mean values (dashed), as well as the scores of the best and worst-performing study participants (1st quartile in petrol blue, 4th quartile in brown), are indicated.

Table C.1: **Descriptive statistics of all real-effort tasks (normalized scores)**

	mean	sd	se(mean)	IQR	skewness	kurtosis	25%	75%
<i>Mentally demanding tasks</i>								
Multiplication	0.29	0.27	0.02	0.40	1.00	0.29	0.04	0.44
Transcribe Words	0.63	0.17	0.01	0.23	-0.33	0.74	0.52	0.75
Transcribe Codes	0.63	0.14	0.01	0.16	-0.86	3.57	0.55	0.71
Word Encryption	0.56	0.16	0.01	0.22	-0.12	1.03	0.44	0.67
<i>Physically demanding tasks</i>								
ab-Typing	0.50	0.20	0.01	0.28	-0.24	-0.23	0.35	0.64
Single Slider	0.16	0.07	0.00	0.03	8.94	111.11	0.14	0.17
<i>Neither physically nor mentally demanding tasks</i>								
Catching Balls	0.74	0.17	0.01	0.13	-2.45	7.60	0.70	0.84

Prior to conducting the study, the piece rates were adjusted according to a pilot trial to ensure that a similar payoff could be achieved for each task at an average performance. Yet, average payoffs

slightly varied across tasks, and the total payoffs do not provide a good measure of the overall effort. To obtain a better estimate of the “total effort provided” by a subject in the experiment, the min-max-normalized scores for each task were summed up. This improved aggregate measure was then min-max-normalized again to be comparable with the normalized scores the subject achieved in the tasks. In Figure C.7, the subjects are ranked according to this “normalized total score.” For each subject, their min-max normalized scores for the remaining tasks are displayed to the right. The figure exemplifies that subjects perform very differently across the selection of tasks, which differ considerably in the skills required to excel. Accordingly, the subjects’ scores are only slightly correlated between the tasks (see also Figure 4.7). A stronger correlation between the tasks is only observed for the two transcription tasks. When additionally adjusting for multiple tests, some of the weak and very weak correlations are no longer significant (e.g., multiplication task and word-encryption task).

Besides, Figure C.8 lists the correlations between the subjects’ performance in each task and the overall performance measure derived from them. A greater correlation is observed for the ab-typing task, the two transcription tasks (words and codes), and the word-encryption task. These tasks might, therefore, serve (better) as a proxy for the *overall provided level of effort*. If, however, these tasks turn out to be very skill-dependent in the further course of the study, this “measure of overall performance” is more likely to resemble the “overall skill level” of the subjects.

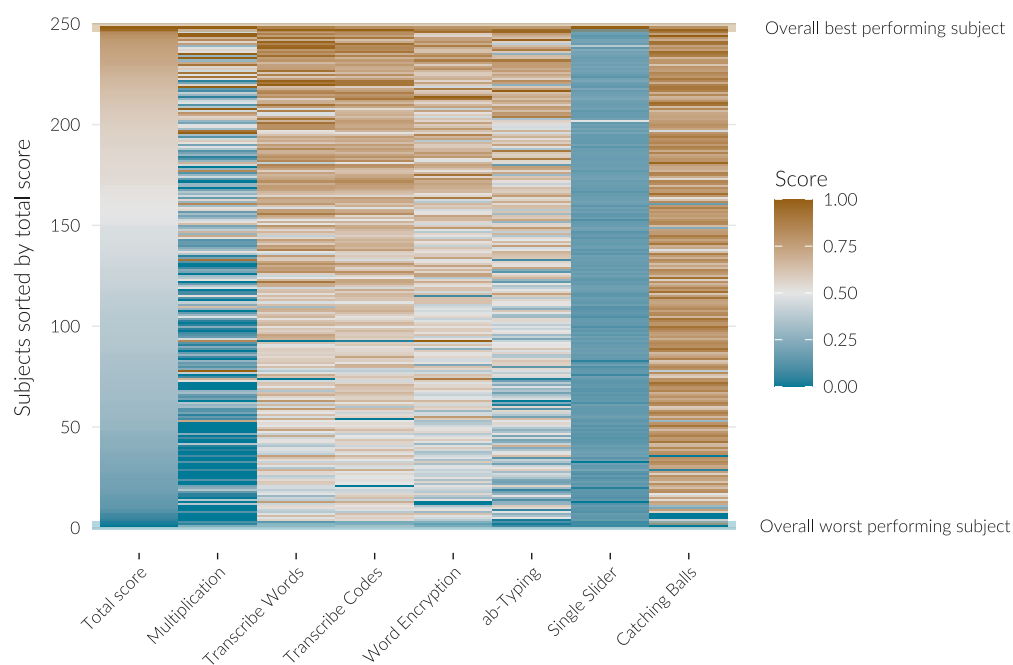


Figure C.7: **Ranking of subjects according to their “normalized total score”**: The graph ranks the subjects according to their overall performance (left-most bar) to illustrate that they performed very differently across the tasks. The ranking was performed on the sum of the min-max normalized scores from all tasks instead of on the de facto earned final payoffs, to account for differences in piece rates. Next to their overall performance, subjects’ normalized scores are displayed for each of the tasks. These do not have very much in common, which illustrates the limited predictive power of scores across tasks.

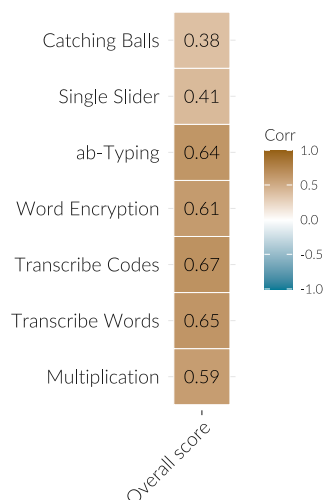


Figure C.8: **Correlation between subjects’ “normalized total score” and their score in each task**

C.2.2 Regression Analysis

C.2.2.1 Correlation Analysis

In the following, several correlograms are presented, reporting *correlations between the subject personality traits and skills* (Figure C.9), *correlations between the motivational variables* (Figure C.10 and Figure C.11), and *correlations between the motivational variables and subjects' performance* (Figure C.12). Each correlogram shows Pearson correlations (above the diagonal), variable distributions (on the diagonal), and scatter plots with smoothed lines of best fit (below the diagonal).

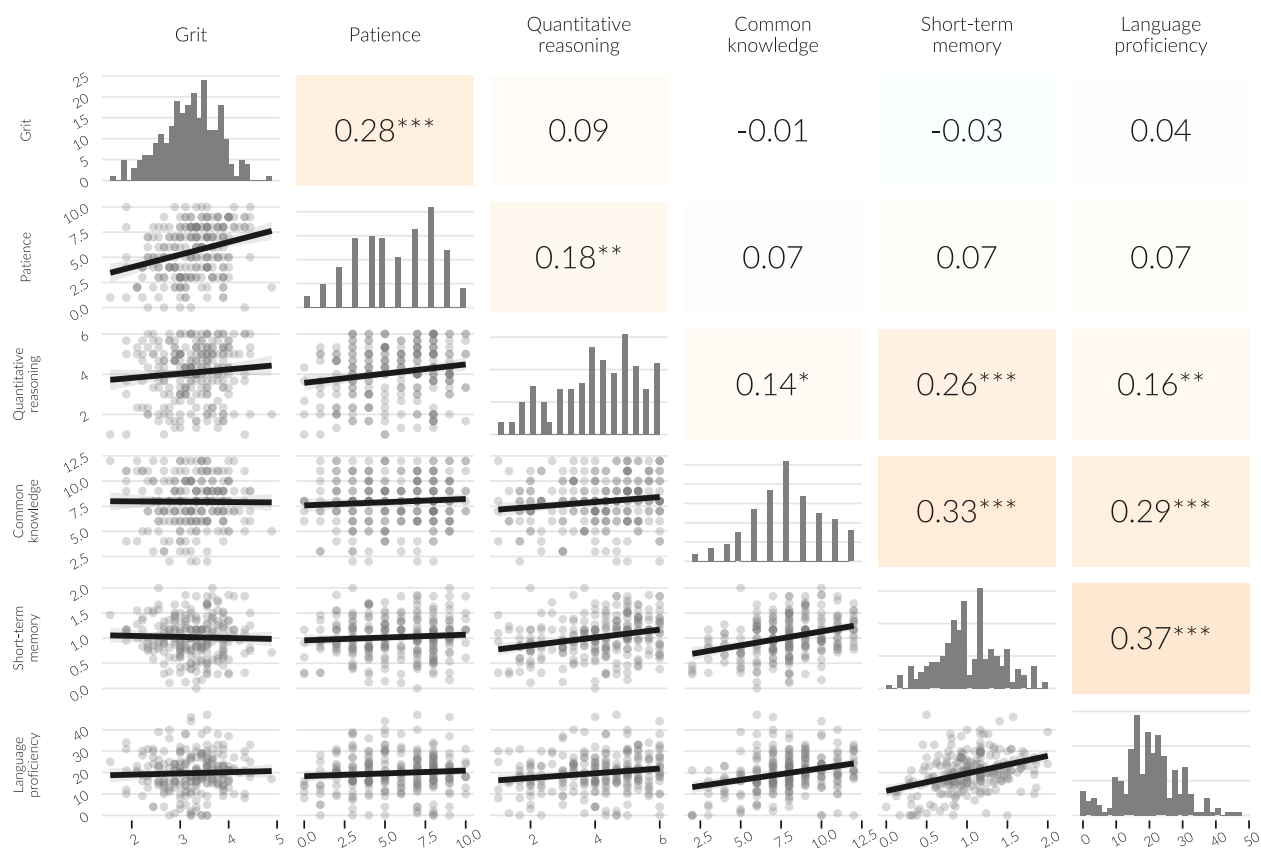


Figure C.9: **Correlation between subjects' qualities:** Correlation between the skills and personality traits obtained from the surveys listed in Table 4.1. Their frequency distributions are presented along the diagonal. For pairwise combinations of subjects' qualities, correlations are reported as values with significance stars (upper right triangle) and illustrated by a scatter plot, including a linear approximation (lower left triangle). Only complete observations are considered to calculate the correlation matrix. The observations of four subjects who did not finish certain surveys in time are, therefore, excluded from the analysis. Moreover, only numeric variables are included (*touch-typing* is assessed on a response scale with four levels, thus constitutes an ordered factor and is consequently not considered in the correlation analysis). The correlations between the subjects' qualities range from very weak to weak.

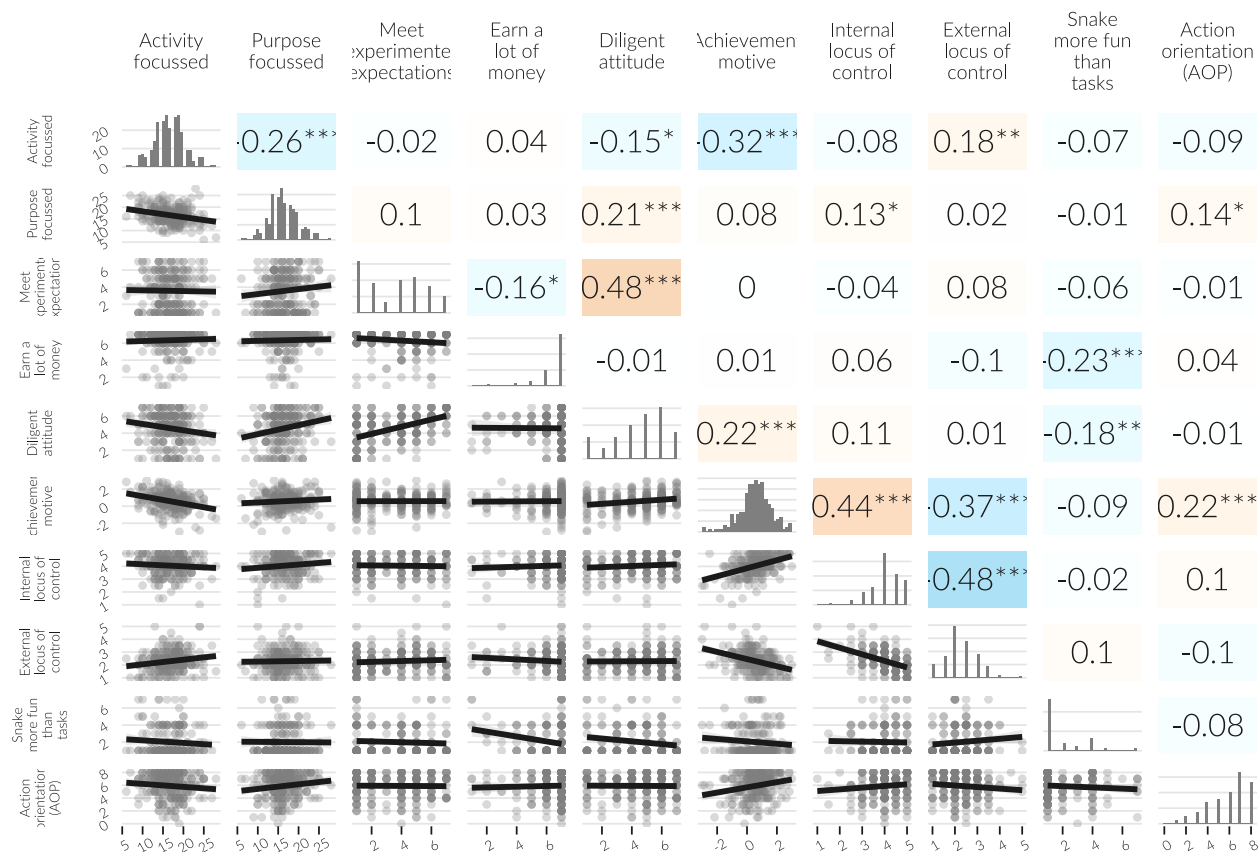


Figure C.10: **Correlation between subjects' motivations:** including the frequency distributions and significance values. The correlation table contains all combinations of the evaluated motivation characteristics (see Table 4.2). Low correlation is observed between the elements of the motivation diagnostics approach: *activity centring* is (as expected) negatively correlated with *purpose centring* as well as with the *performance motive*. Large effects are observed for the global motivations, particularly between having a *diligent attitude* and 1) *meeting experimenter expectations*, 2) *having joy in task fulfillment*, and 3) *earning a lot of money*.

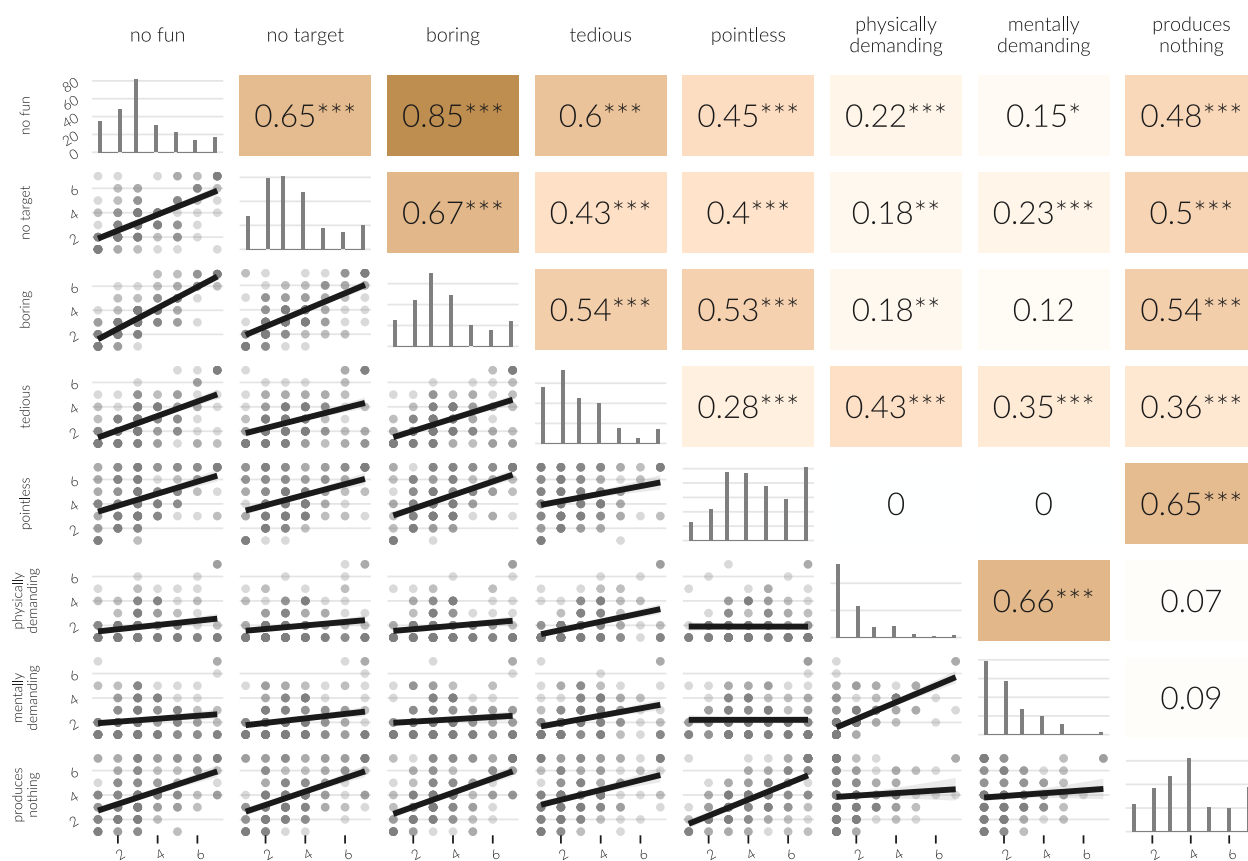


Figure C.11: **Correlation between the items of the real-effort task survey:** exemplarily for the ball-catching task of Gächter et al. (2016). The survey represents a semantic differential scale with two bipolar verbal anchors. Its eight items are evaluated on a 7-point response scale, with 1 representing the positively connoted statement anchor, e.g., (1) “the task was fun” or (6) “was physically easy,” and 7 the negatively connoted statement anchor, e.g., (1) “I did not enjoy it” or (6) the task “was physically demanding/exhausting.” The item correlations observed for the other tasks are, in essence, quite similar.

After completing each task, subjects filled in the real-effort task survey introduced in Chapter 3 to elicit subjects’ perceptions of each task. Figure C.11 presents the Pearson correlations between all dimensions of the survey as an example for the ball-catching task of Gächter et al. (2016) (see Appendix B.3.3.1 for a general treatment of the correlation between the survey items). The correlations between the items found for the remaining tasks are essentially very similar.

Figure C.12 reports Pearson-correlations for subjects’ performance in each task and their individual responses to the survey. Negative correlation values indicate that subjects who tended in their response towards the negatively connoted anchor (7) scored worse on the task, while subjects who

perceived the task more positively and chose a response towards the positively connoted anchor (1) scored better (see also [C.13](#) for a graphical illustration). Subjects' performance in a task hardly affects their perception of the task. This holds for all tasks except for the multiplication task and, to a smaller degree, for the ball-catching task. Subjects who perceive the ball-catching task more physically or mentally demanding score lower in the task (since their responses to the respective survey items are weakly negatively correlated with their performance in the task). For the multiplication task, however, weak to moderate correlation is observed for *all* survey items (except for the item that asks subjects to evaluate the task's physical demand). Thus, subjects' assessment of the multiplication task is greatly influenced by their performance in the task (this finding was already anticipated in [Section 3.5](#):¹ The task is highly cherished by some subjects, but strongly disapproved by others, which results in a pronounced motivational bias).

In the last column, the table further reports Pearson-correlations for subjects' task scores and their subjective performance assessments (see also discussion in [Section 4.3.3](#)).

¹Appendix [B.3.4](#) attempts to determine subjects' preferences for tasks using the real-effort task survey and the Personal Hit-List.

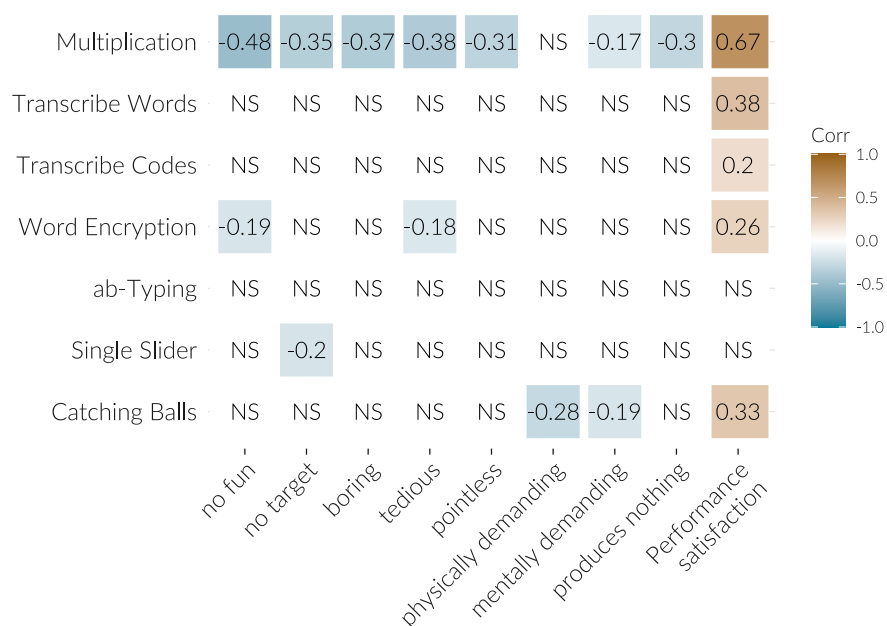


Figure C.12: **Pearson-correlations for task scores and responses to the real-effort task survey:** Correlations are assessed between subjects' scores in a task and i) their individual survey responses (column 1 to 8) and ii) their subjective performance assessment (column 9) both for this particular task. i) For most tasks, subjects' perception of the tasks is uncorrelated with their performance in the respective task. Only for the Multiplication task, weak to moderate correlation is observed. ii) Correlations for subjects' task scores and their subjective performance assessments are reported in the last column: Except for the physical tasks, which involve rather unusual working procedures, the subjects can assess their performance quite accurately (see also the discussion in 4.3.3 on subjects' subjective performance assessment). The table solely includes significant correlations ($p < 0.05$; results basically do not change for a significance level of $p < 0.1$).

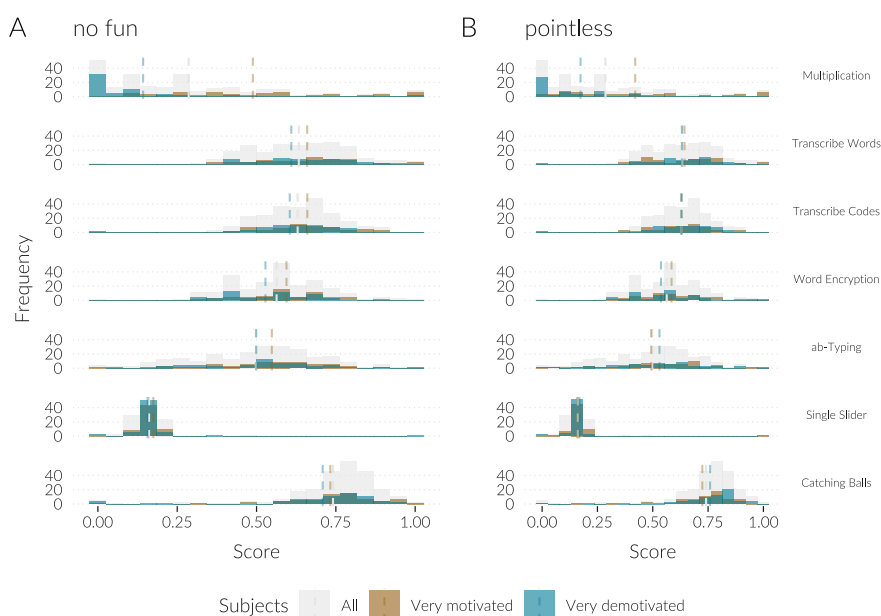


Figure C.13: **Superimposed scores conditional on subject's task perception:** (A) for each task, the score distribution for all subjects is shown in grey. In addition, the point distributions of subjects who indicated that they found the task to be highly “fun” in the first item of the real-effort task survey are colored in brown, while those who found the task to be genuinely “not fun” are colored in petrol-blue (first and fourth quartiles of the distribution, respectively). (B) The same procedure is applied to the fifth survey item, which assesses whether subjects perceive the task as meaningful or instead as meaningless.

C.2.2.2 Results of OLS Regressions: Tables Including Control Variables

The regression analysis included several control variables: gender (binary), age (cts), German nationality (binary)², relationship status (binary), familiarity with the content of the experiment (binary; comment box available). Table C.2 summarizes descriptive statistics for the control variables. The following Tables C.4, C.5 and C.6 extend the Tables 4.3, 4.4 and 4.5 presented in the results section of Chapter 4 by including the coefficients for the controls. Various findings concerning the influence of the controls on subjects' performance in the tasks are discussed below. In summary, all control parameters included in the regression had a significant effect on one or the other task, whereby the coefficient sizes were not negligibly small (apart from *age*). Strong gender effects can be observed in

²This control variable aims to capture differences in familiarity with the German language, but also cultural differences. However, it cannot be ruled out that some of the 49 non-German students were nevertheless very well versed in the German language, e.g., students from Austria (no Swiss students participated in the study). Therefore, the analysis was additionally carried out differentiating between Germans, Europeans, and non-Europeans. The results remain qualitatively the same; any differences observed are explicitly stated.

Table C.2: **Descriptive statistics for the control variables:** Subjects performance split by gender and descriptives for all controls.

		Female (N=141)		Male (N=107)		Diff. in Means	Std. Error
		Mean	Std. Dev.	Mean	Std. Dev.		
Multiplication		0.26	0.26	0.32	0.28	0.07	0.04
Transcribe Words		0.62	0.16	0.65	0.17	0.04	0.02
Transcribe Codes		0.62	0.13	0.64	0.16	0.02	0.02
Word Encryption		0.56	0.16	0.57	0.16	0.01	0.02
ab-Typing		0.43	0.19	0.59	0.18	0.16	0.02
Single Slider		0.15	0.03	0.17	0.09	0.02	0.01
Catching Balls		0.73	0.18	0.75	0.16	0.02	0.02
Age		24.93	5.36	26.37	5.03	1.44	0.66
		N	%	N	%		
Nationality: non-German/German	German	110	78.0	89	83.2		
	non-German	31	22.0	18	16.8		
Relationship status	In relationship	70	49.6	54	50.5		
	Single	71	50.4	53	49.5		
Exp. content	No	94	66.7	68	63.6		
	Yes	47	33.3	39	36.4		

the ab-typing task, probably due to the greater competitiveness of men.³ A less distinct but yet significant effect is found for the single-slider task. While men tend to use computers more frequently, the effect cannot be explained by a greater level of typing skills.⁴ It seems as if another crucial dimension is not queried, which is, therefore, picked up by the dummy variable (e.g., competitive thinking). Notably, the gender effects observed by [Benndorf et al. \(2014\)](#) for the word-encryption task could not be reproduced.⁵

A significant but negligibly small effect of age on subject's performance was observed in most of the tasks.

Furthermore, subjects from outside of Germany perform significantly better in the multiplication task, when only the motivational variables are included in the regression (see [Table C.5](#)). However, the coefficient is no longer significant when the subject's personality and skills are considered in the

³Researchers who plan to use the task in their experiments are advised to control for gender in any subsequent analysis (e.g., endowments generated with the task are likely to show pronounced gender bias).

⁴A chi-square test for independence showed that there was no significant association between gender and the ability to use touch-typing, which serves as a proxy for typing skills and physical dexterity in the use of computers ($\chi^2(3,248) = 5.85, p = 0.12$). However, men are likely to use the computer more frequently than women ($\chi^2(4,248) = 7.97, p = 0.09$) with a medium effect size according to Cramer's V as a measure of association.

⁵Since the [Benndorf et al. \(2014\)](#) used the task in a repeated-measures design, the effects may occur only when the task is performed repeatedly. However, the authors do not find any difference in learning behavior in the task.

analysis (see both Table C.4 and Table C.5). This observation indicates that the dummy variable likely captures some of the (additional) dimensions covered in these regressions.⁶

As expected, participants with German nationality scored higher on the word-transcription task, which can be attributed to their more extensive vocabulary. This difference persists even if all considered personality traits and skills of a subject are taken into account (see Table C.4).⁷

However, non-German subjects also perform worse in the structurally very similar code-transcription task. This is much less to be expected as the codes, which the subjects had to transcribe, are randomly generated, and no language skills are required to perform the tasks. There are two possible explanations for this: *Firstly*, the instructions for the transcription tasks, which are understandably quite similar, are not formulated well enough to be understood by non-native speakers. Participants had the opportunity to review the instructions for each task after the trial round. Since only a tiny proportion of subjects made use of this option in the transcription tasks, this explanation seems less plausible (see Table C.3 below).⁸ *Secondly*, and all the more likely, the provision of German keyboards may have been a decisive factor in the experiment. Native speakers are accustomed to the German keyboard layout (regardless of whether they can use touch-typing).⁹ For non-Germans, however, this keyboard layout might take a while to get used to.¹⁰ Subjects' relationship status is highly significant for the multiplication task. Since it is uncorrelated with other variables such as quantitative reasoning abilities or task liking, it appears to capture other effects of greater importance.

At the end of the experiment, subjects could indicate whether they were familiar with contents of the

⁶Note also that no difference was observed in the self-reported mathematical abilities ($t(70.65) = -0.76, p > 0.05$).

⁷An anagram task served as a measure to assess language and verbalizing skills. German subjects were able to generate 21.63 (SD = 8.1) anagrams, whereas non-German study participants came up with an average of 12.22 (SD = 7.87) anagrams. A Welch two-samples t-test revealed that the difference was statistically significant, $t(75.06) = -7.45, p < 0.001$.

⁸Conversely, this explanation appears appropriate for the ball-catching task: If one (further) differentiates between Germans, Europeans, and non-Europeans, the last group scored significantly worse. Considering that non-Europeans re-read the instructions of the task twice as often as the others, it seems very likely that a large proportion of the subjects had difficulties understanding the task and thus fulfilling it (see also Table C.3).

⁹However, non-German subjects do have greater knowledge of touch-typing ($\chi^2(3,248) = 11.81, p = 0.01$).

¹⁰The fact that non-German subjects also perform significantly worse in the ab-typing task provides some support for the second explanation: Being rather unfamiliar with the keyboard layout, non-German subjects were bound to perform below their abilities (see Table C.5). Note that the coefficient of the dummy variable for German descent is no longer significant for the word-encryption task when the subject's personality and skills are considered. The influence is absorbed by other variables e.g., language skills (see both Table C.4 and Table C.5).

Table C.3: **Share of subjects that re-read the instructions for each task**

Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
10.48	1.21	2.82	13.31	0.81	4.44	10.48

experiment.¹¹ Familiarity with the experiment’s content appears decisive for both transcription tasks, the word-encryption task, as well as the single-slider task. In terms of the first three mentioned tasks, some of the subjects likely encountered the same or similar tasks in other experimental researchers’ studies. More surprising and interesting is the significant effect found for the single-slider task, which was first implemented in this study.¹² This indicates that prior knowledge – *no matter of which task* – can have substantial effects on experimental outcomes (i.e., these subjects “know the game”). Therefore, the dummy variable serves somewhat as an indicator for frequent participation in laboratory studies, i.e., as a proxy for “being a lab rat.”

¹¹See Section 4.2.2 for a more detailed discussion of the share of subjects who had previous encounters with specific tasks.

¹²More precisely, a pilot experiment was conducted approximately two months before the study was conducted. However, participants of the pilot were not admitted for participation.

Table C.4: Task performance conditional on all subjects' qualities, including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.191 (0.221)	0.459*** (0.000)	0.604*** (0.000)	0.449*** (0.000)	0.472*** (0.000)	0.126*** (0.001)	0.374*** (0.000)
<i>Physical abilities</i>							
Touch typing	0.048 (0.246)	0.112*** (0.000)	0.116*** (0.000)	0.099*** (0.000)	0.022 (0.457)	0.006 (0.564)	-0.023 (0.390)
<i>Prerequisites to focus and to maintain concentration</i>							
Grit	-0.043 (0.150)	-0.010 (0.489)	-0.019 (0.152)	0.016 (0.337)	0.014 (0.520)	-0.006 (0.419)	0.021 (0.275)
Patience	-0.005 (0.446)	-0.006* (0.058)	-0.002 (0.464)	-0.004 (0.315)	0.000 (1.000)	-0.002 (0.376)	-0.003 (0.493)
<i>Cognitive abilities</i>							
Quantitative reasoning	0.063*** (0.000)	0.004 (0.497)	0.003 (0.569)	0.007 (0.382)	0.011 (0.222)	0.002 (0.556)	0.016* (0.060)
Common knowledge	-0.008 (0.315)	0.004 (0.278)	0.000 (0.959)	-0.007 (0.119)	-0.010* (0.092)	0.002 (0.308)	0.006 (0.272)
Short-term memory	0.023 (0.627)	0.073*** (0.002)	0.035 (0.103)	0.057** (0.039)	0.045 (0.190)	0.008 (0.504)	0.040 (0.201)
Language proficiency	0.000 (0.846)	0.006*** (0.000)	0.004*** (0.000)	0.003** (0.017)	0.004*** (0.005)	0.001** (0.011)	0.002 (0.160)
<i>Control variables</i>							
Gender: female	-0.029 (0.404)	-0.025 (0.138)	-0.020 (0.213)	-0.016 (0.416)	-0.159*** (0.000)	-0.021** (0.018)	0.007 (0.754)
Age	0.001 (0.837)	-0.003* (0.070)	-0.004** (0.015)	-0.001 (0.701)	-0.003 (0.151)	0.000 (0.589)	0.003 (0.193)
Nationality: German	-0.097** (0.050)	0.085*** (0.000)	0.049** (0.029)	-0.026 (0.362)	0.028 (0.425)	-0.006 (0.633)	0.030 (0.356)
In relationship	0.101*** (0.002)	0.026 (0.110)	0.008 (0.589)	0.012 (0.510)	0.013 (0.581)	-0.008 (0.312)	-0.015 (0.485)
Familiarity exp. content	0.035 (0.328)	0.043** (0.015)	0.035** (0.031)	0.077*** (0.000)	0.032 (0.210)	0.001 (0.878)	0.019 (0.428)
Num.Obs.	248	248	248	248	248	248	248
R2	0.205	0.473	0.364	0.207	0.271	0.119	0.114
R2 Adj.	0.154	0.438	0.323	0.156	0.224	0.062	0.057

Note: p-values are not adjusted for multiple hypothesis testing.

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

Table C.5: Task performance conditional on all subjects' motivations, including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.387* (0.083)	0.644*** (0.000)	0.761*** (0.000)	0.582*** (0.000)	0.792*** (0.000)	0.246*** (0.001)	0.998*** (0.000)
<i>Q1. Capturing activity-related incentives</i>							
PHL	-0.043*** (0.000)	-0.028*** (0.000)	-0.012** (0.021)	-0.016** (0.011)	-0.001 (0.919)	-0.006** (0.014)	-0.013** (0.018)
Activity focussed	-0.006 (0.188)	0.002 (0.482)	-0.002 (0.400)	-0.004 (0.190)	-0.007* (0.072)	-0.001 (0.501)	-0.006* (0.075)
Purpose focussed	-0.007 (0.120)	0.001 (0.783)	-0.002 (0.558)	0.001 (0.791)	0.002 (0.637)	-0.002 (0.290)	-0.008** (0.018)
<i>Q2. Capturing externally controlled incentives</i>							
Meet experimenter expectations	0.001 (0.942)	-0.020*** (0.001)	-0.011** (0.048)	-0.007 (0.254)	0.001 (0.938)	0.000 (0.983)	-0.002 (0.730)
<i>Q3. Diagnosis of desired outcomes</i>							
no target	-0.020** (0.022)	0.016** (0.015)	-0.005 (0.369)	-0.003 (0.657)	-0.015** (0.048)	-0.006** (0.016)	0.005 (0.518)
<i>Q4. Capturing outcome dependent incentives</i>							
Earn a lot of money	0.027** (0.040)	-0.001 (0.870)	-0.009 (0.273)	0.011 (0.240)	0.001 (0.897)	0.006 (0.165)	0.009 (0.311)
Diligent attitude	0.005 (0.661)	0.013* (0.084)	-0.001 (0.829)	0.009 (0.277)	0.005 (0.554)	0.000 (0.938)	-0.001 (0.858)
Achievement motive	-0.025 (0.207)	-0.018 (0.167)	-0.024** (0.049)	-0.005 (0.704)	-0.002 (0.901)	-0.001 (0.864)	0.023* (0.098)
<i>Q5. Self-efficacy expectations</i>							
Internal locus of control	0.074*** (0.007)	0.021 (0.235)	0.036** (0.030)	0.012 (0.535)	0.006 (0.776)	-0.002 (0.823)	-0.018 (0.336)
External locus of control	0.014 (0.573)	-0.014 (0.378)	-0.011 (0.483)	0.008 (0.642)	-0.011 (0.618)	-0.012 (0.138)	-0.003 (0.855)
<i>Q6. Aversiveness of the activity</i>							
Snake more fun than tasks	0.006 (0.552)	0.001 (0.916)	-0.003 (0.656)	-0.003 (0.693)	-0.007 (0.410)	0.000 (0.886)	-0.022*** (0.002)
<i>Q7. Capturing self-regulation and volition</i>							
Action orientation (AOP)	-0.020** (0.025)	0.005 (0.403)	0.007 (0.181)	0.009 (0.150)	-0.003 (0.646)	0.001 (0.721)	0.005 (0.451)
<i>Control variables</i>							
Gender: female	-0.059* (0.081)	-0.050** (0.025)	-0.033 (0.114)	-0.030 (0.206)	-0.174*** (0.000)	-0.025** (0.025)	0.004 (0.862)
Age	0.000 (0.936)	-0.004** (0.034)	-0.006*** (0.003)	-0.004* (0.053)	-0.005** (0.031)	0.000 (0.795)	-0.001 (0.597)
Nationality: German	-0.038 (0.386)	0.099*** (0.001)	0.086*** (0.001)	0.005 (0.865)	0.051 (0.160)	0.009 (0.521)	0.051* (0.092)
In relationship	0.067** (0.033)	0.016 (0.438)	0.006 (0.748)	0.019 (0.392)	0.021 (0.422)	-0.008 (0.428)	0.003 (0.884)
Familiarity exp. content	0.001 (0.968)	0.041* (0.057)	0.042** (0.037)	0.078*** (0.001)	0.019 (0.497)	0.003 (0.748)	0.056** (0.018)
Num.Obs.	205	205	205	205	205	205	205
R2	0.416	0.317	0.251	0.197	0.274	0.169	0.232
R2 Adj.	0.359	0.251	0.179	0.120	0.204	0.089	0.158

Note: p-values are not adjusted for multiple hypothesis testing.

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

Table C.6: Task performance conditional on all subject characteristics (qualities and motivations), including controls

	Multiplication	Transcribe Words	Transcribe Codes	Word Encryption	ab-Typing	Single Slider	Catching Balls
(Intercept)	0.398 (0.133)	0.481*** (0.001)	0.686*** (0.000)	0.524*** (0.004)	0.786*** (0.000)	0.221*** (0.005)	0.891*** (0.000)
<i>Physical abilities</i>							
Touch typing	0.082** (0.035)	0.095*** (0.000)	0.111*** (0.000)	0.089*** (0.001)	0.023 (0.467)	0.001 (0.928)	-0.014 (0.606)
<i>Prerequisites to focus and to maintain concentration</i>							
Grit	-0.026 (0.488)	-0.036* (0.079)	-0.022 (0.273)	-0.011 (0.653)	-0.029 (0.341)	-0.011 (0.324)	0.008 (0.746)
Patience	-0.012* (0.079)	-0.004 (0.266)	-0.001 (0.729)	-0.006 (0.172)	-0.001 (0.876)	-0.003 (0.140)	-0.007 (0.118)
<i>Cognitive abilities</i>							
Quantitative reasoning	0.029** (0.049)	0.007 (0.327)	0.006 (0.365)	0.009 (0.296)	0.008 (0.479)	0.006 (0.162)	0.012 (0.212)
Common knowledge	0.004 (0.588)	0.003 (0.528)	0.003 (0.402)	-0.006 (0.240)	-0.010 (0.110)	0.004 (0.112)	0.006 (0.290)
Short-term memory	-0.034 (0.460)	0.062** (0.015)	0.028 (0.263)	0.046 (0.133)	0.057 (0.139)	0.010 (0.472)	0.019 (0.573)
Language proficiency	0.001 (0.562)	0.007*** (0.000)	0.005*** (0.000)	0.004*** (0.002)	0.005*** (0.006)	0.002*** (0.006)	0.002 (0.153)
<i>Q1. Capturing activity-related incentives</i>							
PHL	-0.037*** (0.000)	-0.013** (0.012)	-0.012*** (0.007)	-0.013** (0.034)	-0.005 (0.418)	-0.008*** (0.001)	-0.014** (0.016)
Activity focussed	-0.008* (0.098)	0.001 (0.639)	-0.003 (0.186)	-0.005 (0.136)	-0.007* (0.090)	-0.001 (0.356)	-0.006 (0.104)
Purpose focussed	-0.009* (0.073)	0.002 (0.428)	-0.002 (0.412)	0.001 (0.810)	0.002 (0.539)	-0.002 (0.250)	-0.008** (0.015)
<i>Q2. Capturing externally controlled incentives</i>							
Meet experimenter expectations	0.006 (0.525)	-0.012** (0.021)	-0.004 (0.375)	-0.002 (0.711)	0.002 (0.772)	0.002 (0.451)	0.000 (0.939)
<i>Q3. Diagnosis of desired outcomes</i>							
no target	-0.021** (0.020)	0.009* (0.099)	-0.003 (0.496)	-0.004 (0.541)	-0.013* (0.084)	-0.006** (0.018)	0.003 (0.636)
<i>Q4. Capturing outcome dependent incentives</i>							
Earn a lot of money	0.024* (0.063)	-0.006 (0.369)	-0.011 (0.120)	0.008 (0.340)	-0.002 (0.821)	0.003 (0.501)	0.004 (0.656)
Diligent attitude	0.005 (0.678)	0.016*** (0.009)	0.000 (0.955)	0.008 (0.282)	0.005 (0.556)	0.000 (0.967)	-0.003 (0.699)
Achievement motive	-0.027 (0.200)	-0.016 (0.176)	-0.024** (0.033)	-0.001 (0.922)	0.009 (0.613)	0.001 (0.843)	0.020 (0.174)
<i>Q5. Self-efficacy expectations</i>							
Internal locus of control	0.080*** (0.004)	0.014 (0.344)	0.031** (0.034)	0.005 (0.776)	-0.001 (0.968)	0.000 (0.964)	-0.016 (0.423)
External locus of control	0.006 (0.820)	-0.011 (0.407)	-0.010 (0.455)	0.009 (0.604)	-0.011 (0.620)	-0.008 (0.289)	0.002 (0.919)
<i>Q6. Aversiveness of the activity</i>							
Snake more fun than tasks	0.007 (0.544)	0.002 (0.676)	0.000 (0.935)	-0.003 (0.700)	-0.006 (0.465)	0.002 (0.642)	-0.021*** (0.006)
<i>Q7. Capturing self-regulation and volition</i>							
Action orientation (AOP)	-0.020** (0.029)	0.005 (0.355)	0.006 (0.232)	0.007 (0.246)	-0.005 (0.507)	0.001 (0.770)	0.003 (0.631)
<i>Control variables</i>							
Gender: female	-0.049 (0.154)	-0.029 (0.131)	-0.023 (0.211)	-0.022 (0.328)	-0.161*** (0.000)	-0.017 (0.112)	0.010 (0.688)
Age	0.001 (0.737)	-0.002 (0.139)	-0.004*** (0.007)	-0.002 (0.289)	-0.003 (0.222)	0.000 (0.628)	-0.001 (0.793)
Nationality: German	-0.051 (0.301)	0.056** (0.036)	0.055** (0.034)	-0.016 (0.628)	0.027 (0.497)	-0.012 (0.415)	0.016 (0.649)
In relationship	0.078** (0.013)	0.026 (0.130)	0.017 (0.303)	0.026 (0.212)	0.029 (0.261)	-0.006 (0.561)	0.002 (0.925)
Familiarity exp. content	0.013 (0.697)	0.030 (0.108)	0.036** (0.043)	0.074*** (0.001)	0.011 (0.714)	-0.002 (0.872)	0.049** (0.042)
Num.Obs.	205	205	205	205	205	205	205
R2	0.462	0.562	0.483	0.343	0.347	0.275	0.278
R2 Adj.	0.380	0.496	0.405	0.242	0.247	0.165	0.168

Note: p-values are not adjusted for multiple hypothesis testing.

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

C.2.2.3 Results of OLS Regressions With Priors

The following tables report the results of the regression analysis based on the prior considerations presented in Section B.1.1 for a subset of the task selection. The selection includes two heavily skill-demanding tasks (multiplication task and word-transcription task) and one physically demanding task (single-slider task). For each of the tasks, the tables contain the following model specifications: *With priors and without control variables* (1), *without priors and without control variables* (2) and *with control variables* (3). Priors in terms of skills and personality traits, as reported in Table 4.1 are assessed in Table C.7. The pre-considerations on subjects' motivations, as summarized in Table 4.2, could again form the basis for priors. However, these only contain an additional motivation evaluation for the multiplication task. Nevertheless, to enable a direct comparison of the regression results with and without control variables, the following Table ?? presents the results for the same selection of three tasks. Finally, Table ?? includes the prior-considerations for both dimensions.

Several observations can be drawn from Table C.7. (1) For all tasks, the adjusted R^2 increases considerably by adding additional explanatory variables (i.e., moving from left to right, i.e., from “with priors and without controls” to “without priors and with controls”).

(2) The coefficient for *performance-related action orientation* is no longer significant in the multiplication task when adding the control variables into the model. Instead, subjects' *relationship status* appears to matter above all. As discussed above, this is probably less a spurious correlation than the fact that the study has so far neglected an essential component, which is now instead taken into account to a certain extent by the relationship status.

(3) For the word-transcription task, the size of the coefficients increases for *touch-typing* and decreases for *short-term memory*, as the number of covariates is increased (i.e., moving from the left to the right from “with priors and without controls” to “without priors and with controls”). (4) Analogously for the single-slider task, the coefficients' size slightly increases for *touch-typing* with greater model complexity. (5) It turns out that the control variables gradually added in the model comparison play a non-negligible role, particularly the familiarity with experimental content.

Table C.7: Task performance conditional on subjects' characteristics for selected tasks

	Multiplication			Transcribe Words			Single Slider		
	with prior	without prior		with prior	without prior		with prior	without prior	
	no controls	no controls	controls	no controls	no controls	controls	no controls	no controls	controls
(Intercept)	0.260** (0.024)	0.246** (0.037)	0.191 (0.221)	0.378*** (0.000)	0.388*** (0.000)	0.459*** (0.000)	0.193*** (0.000)	0.129*** (0.000)	0.126*** (0.001)
<i>Physical abilities</i>									
Touch typing		0.052 (0.192)	0.048 (0.246)	0.083*** (0.000)	0.080*** (0.000)	0.112*** (0.000)		0.006 (0.536)	0.006 (0.564)
<i>Prerequisites to focus and to maintain concentration</i>									
Grit	-0.040 (0.183)	-0.037 (0.209)	-0.043 (0.150)		-0.010 (0.513)	-0.010 (0.489)	-0.009 (0.223)	-0.009 (0.220)	-0.006 (0.419)
Patience	-0.006 (0.411)	-0.006 (0.341)	-0.005 (0.446)		-0.007* (0.053)	-0.006* (0.058)	0.000 (0.858)	-0.001 (0.527)	-0.002 (0.376)
<i>Cognitive abilities</i>									
Quantitative reasoning	0.070*** (0.000)	0.068*** (0.000)	0.063*** (0.000)		0.004 (0.516)	0.004 (0.497)		0.004 (0.179)	0.002 (0.556)
Common knowledge	-0.012* (0.084)	-0.011 (0.143)	-0.008 (0.315)	0.012*** (0.003)	0.008** (0.038)	0.004 (0.278)		0.002 (0.239)	0.002 (0.308)
Short-term memory		0.031 (0.514)	0.023 (0.627)		0.090*** (0.000)	0.073*** (0.002)		0.008 (0.519)	0.008 (0.504)
Language proficiency		-0.001 (0.638)	0.000 (0.846)	0.009*** (0.000)	0.008*** (0.000)	0.006*** (0.000)		0.001** (0.019)	0.001** (0.011)
<i>Control variables</i>									
Gender: female			-0.029 (0.404)			-0.025 (0.138)			-0.021** (0.018)
Age			0.001 (0.837)			-0.003* (0.070)			0.000 (0.589)
Nationality: German			-0.097** (0.050)			0.085*** (0.000)			-0.006 (0.633)
In relationship			0.101*** (0.002)			0.026 (0.110)			-0.008 (0.312)
Familiarity exp. content			0.035 (0.328)			0.043** (0.015)			0.001 (0.878)
Num.Obs.	248	248	248	248	248	248	248	248	248
R2	0.119	0.152	0.205	0.354	0.407	0.473	0.007	0.082	0.119
R2 Adj.	0.104	0.120	0.154	0.340	0.384	0.438	-0.001	0.048	0.062

Note: Three different OLS-regressions are performed for each of the selected tasks, ranging from *with priors and without controls* to *without priors with controls*.

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

C.2.3 Additional Findings

Further additional findings with potential influence on subjects' performance in real-effort tasks are reported below.

C.2.3.1 The Role of an Outside Option

In Section 2.4.2.2, it was argued that the availability of an alternative activity could alter the outcome of an experiment. More precisely, offering study participants an outside option may impact their motivation to complete a given task, e.g., by countering active participation. As noted earlier, this concern has been confirmed by several studies (Corgnet, Hernán-González, & Schniter, 2015; Erkal et

al., 2017). However, previous studies focused on a single task at a time. To provide further evidence of the influence of outside options on subjects' behavior and compare their impact for different tasks, *Snake* was offered as an alternative activity to 209 of the 248 study participants. These subjects could abort the current task at any time to play the well-known computer game. However, once they had switched, they could not return to the task in order to continue earning money. This outside option was not available to the remaining 39 subjects.

The availability of the outside option was discussed in Appendix B.3.3.2 in the context of the real-effort task survey to substantiate the survey's validity. To recall the crucial findings: The usage of the outside option varies widely between the different tasks. In general, only few study participants made use of it (see also Table B.24 which summarizes subjects' usage pattern of the outside option). However, the number of subjects who decided to quit and no longer earn money in the multiplication task is remarkable.

Table C.8 reports the average scores for all examined tasks for both treatment groups, i.e., with and without outside option *Snake*. If subjects decide to switch to the outside option, there is a tendency that they earn less than their fellow study participants. For the two tasks with the largest number of switching subjects, Table C.9 takes a closer look at the scoring in the task in relation to the availability and use of the outside option. In a next step, the hypothesis is tested that the respondents' performance is negatively affected when they are offered an outside option. That is, when subjects are offered an alternative activity, the task becomes less attractive and they put less effort into completing the task.

Table C.8: **Task performance conditional on the availability of an outside option**

	No (N=39)		Yes (N=209)		Diff. in Means	Std. Error
	Mean	Std. Dev.	Mean	Std. Dev.		
Multiplication	0.291	0.302	0.285	0.268	-0.006	0.052
Transcribe Words	0.611	0.174	0.637	0.164	0.026	0.030
Transcribe Codes	0.628	0.100	0.630	0.146	0.002	0.019
Word Encryption	0.551	0.148	0.565	0.160	0.015	0.026
ab-Typing	0.511	0.214	0.495	0.202	-0.016	0.037
Single Slider	0.163	0.022	0.161	0.071	-0.002	0.006
Catching Balls	0.730	0.203	0.743	0.164	0.013	0.035

Figure C.14 displays the score distributions for all tasks separately for subjects who had the outside

Table C.9: **Outside option usage for specific tasks:** The majority of the subjects could switch to an outside option during the completion of each task. Only a fraction of them eventually used it and played the offered alternative, the game Snake. For the two tasks with the highest use of the outside option, the table summarizes the frequencies of availability and usage as well as the average score (with standard deviation).

Outside option availability	Use of outside option	N	mean score	sd
Multiplication task				
Not available	did not play Snake	39	0.291	0.302
Available	did not play Snake	189	0.312	0.267
Available	did play Snake	20	0.036	0.093
Single-slider task				
Not available	did not play Snake	39	0.163	0.022
Available	did not play Snake	205	0.162	0.070
Available	did play Snake	4	0.079	0.070

option available and those who did not. Below the label of each task, the results of Welch t-tests comparing both groups of subjects for that task are reported. In the single-slider task, the subjects who were able to switch to the game *Snake* achieved a mean score of 0.16 (SD = 0.07), whereas those who did not have an outside option available accumulated a score of on average 0.16 (SD = 0.02). A Welch two-samples t-test showed that the difference was statistically significant, $t(197.09) = 0.3$, $p < 0.05$.

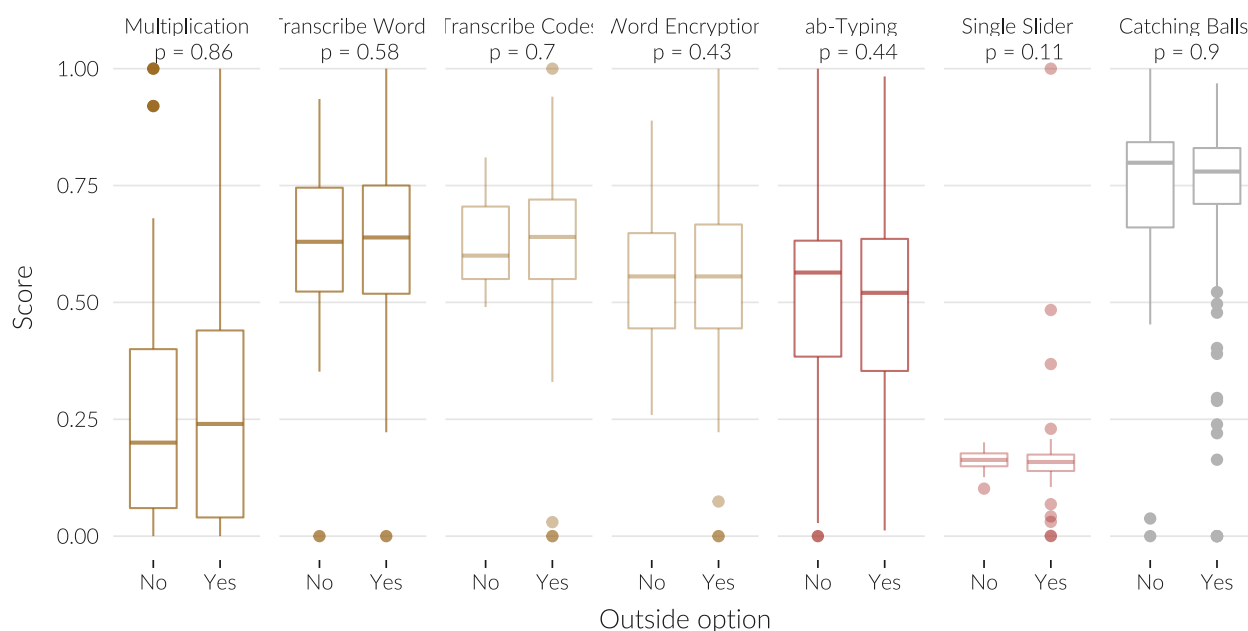


Figure C.14: **Compare the performance of subjects with and without outside option available:** The results of separate Welch t-tests comparing the scores of subjects with and without the option to switch to Snake are provided. Only for the single-slider task is a significant difference observed between the score distributions.

In the next step, the subjects who were able to switch to Snake but actually did not do so were compared to those who did not have the option to switch to the outside option in the first place. Figure C.15 presents the distributions for each task, again including the results of the Welch t-tests. In the single-slider task, the subjects who were able to switch but did not do so achieved a mean score of 0.16 (SD = 0.07), whereas those who did not have the option available accumulated a score of on average 0.16 (SD = 0.02). A Welch two-samples t-test showed that the difference was statistically significant, $t(195.29) = 0.03$, $p < 0.1$.

In summary, for some tasks, the mere availability of an outside option already makes a difference to subjects' performance. However, the effect varies from task to task. As expected, the effect is more pronounced, i) for tasks that can have a discouraging effect upon mere sight (e.g., the multiplication task, as discussed in Section 4.3.3), and ii) for tasks that are tremendously monotonous and mind-numbing (e.g., the single-slider task).

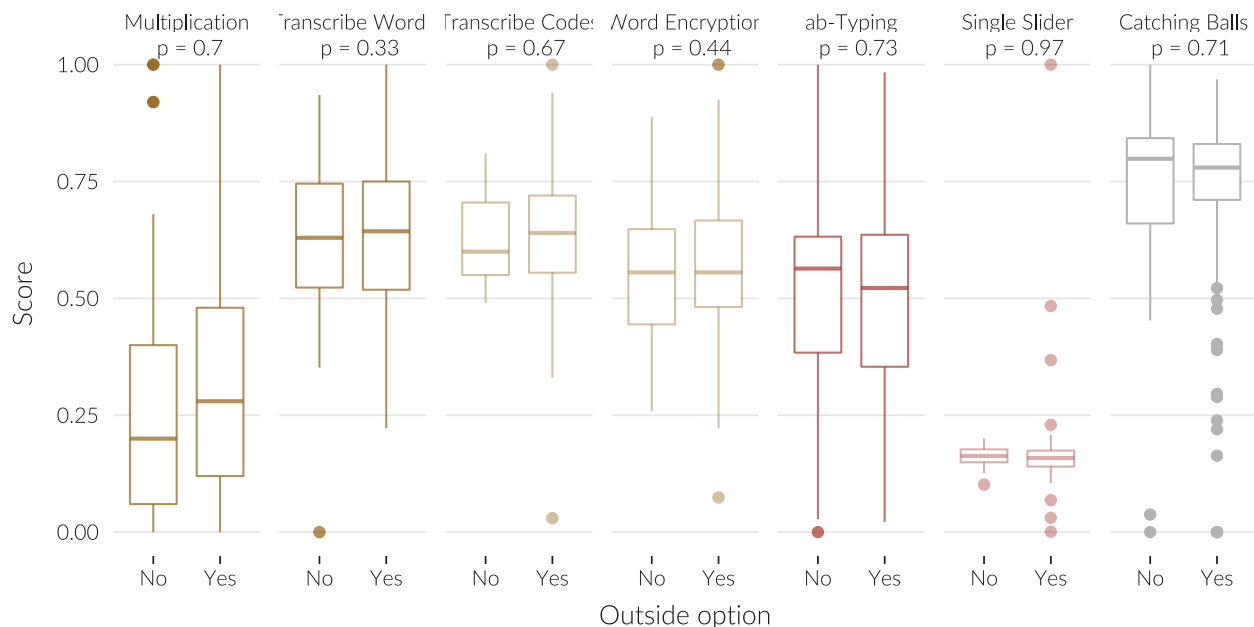


Figure C.15: **Compare the performance of subjects with outside option available, but did not use it, and without:** The figure reports the results of Welch t-tests between subjects who did not make use of an available outside option and those who did not have one available from the outset (separately for each of the tasks). For the single-slider task, a significant difference between the point value distributions is observed only at a significance level of 0.1.

C.2.4 Machine Learning Approach

The “hold out” technique was used to train and test the model:

1. A *train-test split* was performed, splitting the data into two disjoint sets, i.e., the *training data* and *test data*.
2. A regression model was built on the training data (using cross-validation folds); Since sampling the *training data* into a *train set* and a *validate set* can affect the performance measures, *cross-validation* was employed (sampling a *train set* and *validate set* multiple times with different separations). Aggregating the results for all partitions provides a robust measure.
3. Compare the root-mean-squared error (RMSE) for different parameter constellations to find an optimal set of parameters, i.e., obtain a “final model.”
4. Train the final model on complete training data and calculate RMSE within the training data
5. Predict the outcome of the test data; calculate the RMSE within test data

6. Compare *test RMSE* and *training RMSE* to determine model performance/accuracy Repeat steps 2.-5. for different modeling approaches (OLS, lasso, ridge, random forest, SVM, GBM).

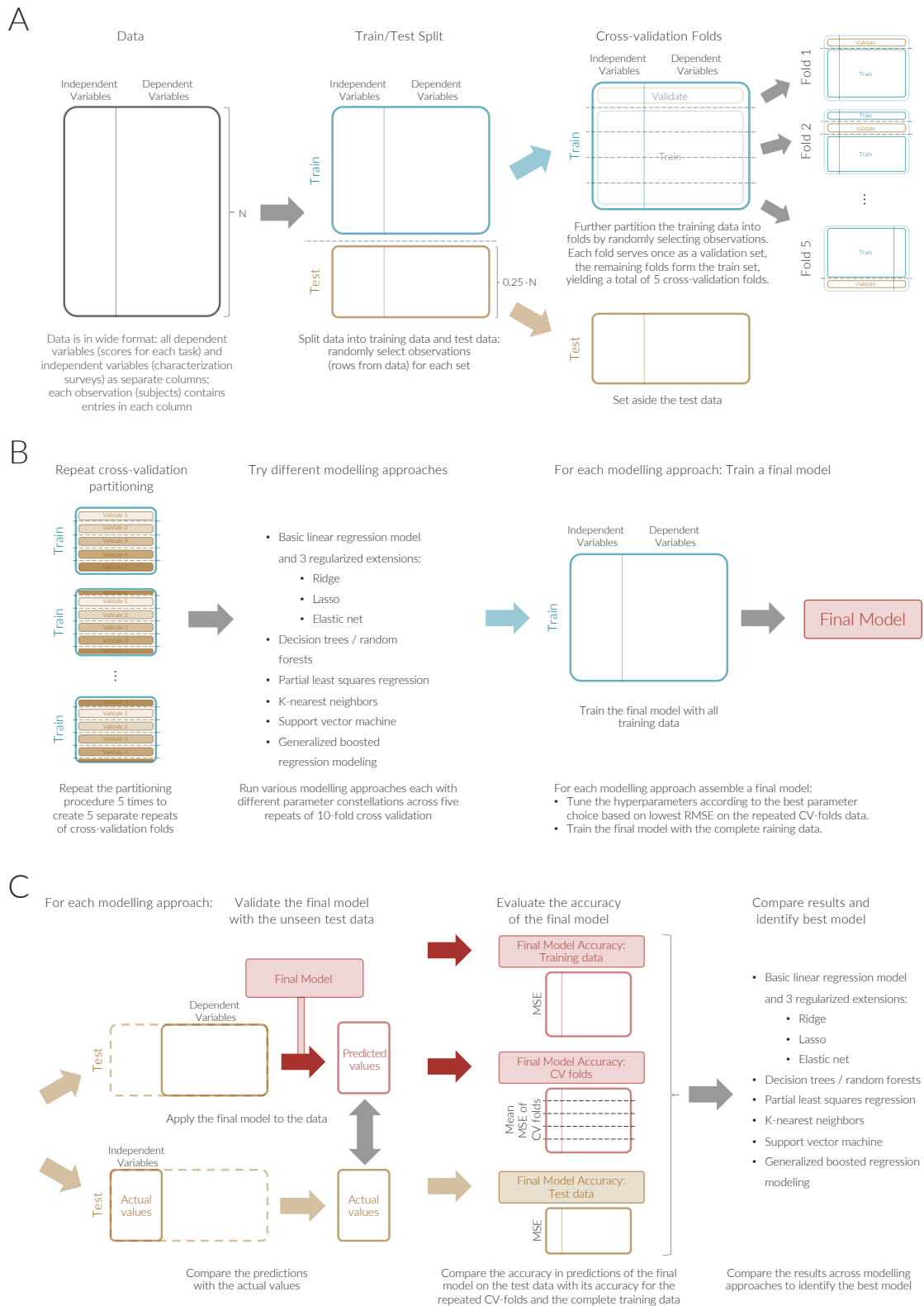


Figure C.16: **Schematic representation of the machine learning approach:** A) train/test split and cross-validation folds; B) model development, selection and final-model training; C) final-model testing and evaluation of accuracy.

The coefficients in the regression analysis only provide an idea of what is going on “on average.” More specifically, they do not allow to differentiate whether all subjects suffer from a minor motivation problem or only a few subjects suffer from a major one. The estimated coefficients are very sensitive to the model specification, i.e., which motivational characteristics are included as predictors in the model or not. This indicates that, despite eliciting different constructs, there are larger connections between the surveys than observable at first (the non-negligible degree of collinearity among the predictors points into this direction; existence of a further mediating variable is not excluded) Or similarly likely, certain characteristics “come in pairs,” i.e., if someone is tall she is very likely to wear larger shoes. For this reason, clustering the subjects provides a more appropriate approach. Performing clustering analysis prior to the regression analysis allows to disentangle the influence of a given motivational characteristics on different subsets of the study participants and thereby indirectly allows to better cope with the collinearity among the predictors.

To further come closer to the structure of the diagnostic scheme by [Rheinberg \(2004\)](#), the regression analysis was, therefore, additionally performed separately for each stage of the diagnosis scheme. However, more detailed insights could only be gained by performing a cluster analysis for the different stages. For each stage, subjects were grouped into clusters based on the questionnaires surveyed in the stage using the k-means clustering algorithm. The identified cluster were then labeled based on their mean values for the respective motivational characteristics. The cluster assignments obtained for the different stages were then added to the regression analysis. However, the analysis becomes very elaborate, and a detailed account of this approach goes beyond the scope of this work. Only this much may be said: Clustering at the stage level in combination with a subsequent regression analysis allows to gain more insight into which of the identified groups are motivated in which way or which motivation problem they suffer from. This is achieved in particular by means of the interaction terms between the clustering variable of a given stage and the respective motivational characteristics that are collected in the stage.

As a result, the overall explanatory power of the model continues to increase, further reducing the proportion that cannot be explained by the characteristics or motivation of the subjects. If this were only due to the motivation through monetary incentives, then it would not be a problem but rather desirable. However, as the results indicate, this is unfortunately only very partially the case. Other incentives likewise motivate the subjects to make an effort in the tasks.

Colophon

This thesis was written in R Markdown and \LaTeX , and rendered into PDF document using *ETHdown*, an adaptation of [huskydown](#) to the thesis guidelines and corporate design of [ETH Zurich](#), and [bookdown](#).

This document was typeset using the XeTeX typesetting system, and the [University of Washington Thesis class](#) class by Jim Fox. It is based on the [University of Washington Thesis LaTeX template](#) with slight modifications to ensure that the document conforms precisely to ETH Zurich submission standards. Other elements of the source code for formatting the document were taken from [Latex](#), [Knitr](#), and [RMarkdown templates for UC Berkeley's graduate thesis](#), and [Dissertate: a LaTeX dissertation template to support the production and typesetting of a PhD dissertation at Harvard, Princeton, and NYU](#).

The thesis is typeset with the *Lato* font. In some PDF readers, the keyword search in the PDF document is only available if the font is installed on the respective device (the font can be [downloaded for free](#)).

The source files for this thesis, along with all data files, have been organized into an R package that is available from the author upon request.

This version of the thesis was generated on 2021-05-14 05:13:35. The repository is currently at this commit:

Commit: 577160c36fb74637e2fdb0a81beb3183f3c409e

Author: Chris Waloszek (WORK) <CWaloszec@ethz.ch>

When: 2021-05-14 03:04:31 GMT

Thesis ready for publishing

1 file changed, 2 insertions, 0 deletions

Rmds/Ch3_Comparing-Real-Effort-Tasks/3_Results.Rmd | -0 +2 in 1 hunk

Reproducibility of Results: Information on R Environment and Packages

The version of R listed in Table C.10 and the packages listed in Table C.11 were used to generate this thesis in R markdown. The R markdown file containing the code of the thesis can be used to reproduce this PDF document and the analyses it contains at any time. To compile the code used, the versions of the packages listed in Table C.11 are required. The corresponding versions can be installed from snapshots on the [Checkpoint Server for Reproducibility](#).

Table C.10: **Session info on R environment**

Setting	Value
version	R version 4.0.4 (2021-02-15)
os	macOS Catalina 10.15.6
system	x86_64, darwin17.0
ui	X11
language	(EN)
collate	en_US.UTF-8
ctype	en_US.UTF-8
tz	Europe/Zurich
date	2021-05-14

Table C.11: Required packages

	Package	Loaded version	Date		Package	Loaded version	Date
1	broom	0.7.0	2020-07-09	31	lme4	1.1-23	2020-04-07
2	car	3.0-9	2020-08-11	32	lmerTest	3.1-2	2020-04-08
3	carData	3.0-4	2020-05-22	33	magrittr	1.5	2014-11-22
4	cluster	2.1.0	2019-06-19	34	markdown	1.1	2019-08-07
5	cowplot	1.1.0	2020-09-08	35	MASS	7.3-53	2020-09-09
6	data.table	1.13.0	2020-07-24	36	Matrix	1.3-2	2021-01-06
7	dendextend	1.14.0	2020-08-26	37	modelsummary	0.6.5	2021-01-16
8	devtools	2.3.1	2020-07-21	38	multcomp	1.4-13	2020-04-08
9	DiagrammeR	1.0.6.1	2020-05-08	39	mvtnorm	1.1-1	2020-06-09
10	dplyr	1.0.2	2020-08-18	40	plyr	1.8.6	2020-03-03
11	estimatr	0.26.0	2020-09-07	41	psych	2.0.8	2020-09-04
12	ez	4.4-0	2016-11-02	42	psychTools	2.0.8	2020-08-12
13	forcats	0.5.0	2020-03-01	43	purrr	0.3.4	2020-04-17
14	Formula	1.2-3	2018-05-03	44	RcmdrMisc	2.7-1	2020-08-13
15	GGally	2.0.0	2020-06-06	45	readr	1.3.1	2018-12-21
16	ggcorrplot	0.1.3	2019-05-19	46	readxl	1.3.1	2019-03-13
17	ggplot2	3.3.2	2020-06-19	47	reshape2	1.4.4	2020-04-09
18	ggpubr	0.4.0	2020-06-27	48	rmarkdown	2.7	2021-02-19
19	ggthemes	4.2.0	2019-05-13	49	sandwich	2.5-1	2019-04-06
20	git2r	0.27.1	2020-05-03	50	showtext	0.9	2020-08-13
21	gplots	3.0.4	2020-07-05	51	showtextdb	3.0	2020-06-04
22	Hmisc	4.4-1	2020-08-10	52	stringr	1.4.0	2019-02-10
23	huskydown	0.0.5	2020-09-10	53	survival	3.2-3	2020-06-13
24	huxtable	5.0.0	2020-06-15	54	sysfonts	0.8.1	2020-05-08
25	jtools	2.1.2	2021-01-07	55	TH.data	1.0-10	2019-01-21
26	kableExtra	1.3.4	2021-02-20	56	tibble	3.0.3	2020-07-10
27	knitr	1.31	2021-01-27	57	tidyr	1.1.2	2020-08-27
28	latex2exp	0.4.0	2015-11-30	58	tidyverse	1.3.0	2019-11-21
29	lattice	0.20-41	2020-04-02	59	usethis	1.6.1	2020-04-29
30	likert	1.3.5	2016-12-31	60	vcd	1.4-7	2020-04-02
				61	xtable	1.8-4	2019-04-21

The Single-Button Task

There seems to be a tradition in the family to develop mind-numbing real-effort tasks. I became aware of this when I asked my father to draw the sketches for the introduction to the work and Chapter 1 from my drafts (he is simply the better painter). He had developed the “one-button task” some forty years ago, well before [Swenson \(1988\)](#) invented the *pressing-keys task* and [Berger & Pope \(2011\)](#) reworked the concept with the *ab-typing task*.



Figure C.17: **The single-button task** (of D. Waloßek, personal communication).

References

- 10 Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference Points and Effort Provision. *American Economic Review*, 101(2), 470–492. <http://www.jstor.org/stable/29783680>
- Aellig, S. (2004). *Über den Sinn des Unsinn: Flow-Erleben und Wohlbefinden als Anreize für autotelische Tätigkeiten* [PhD thesis].
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Ammons, R., & Ammons, C. (1959). A Standard Anagram Task. *Psychological Reports*, 5(3), 654–656. <https://doi.org/10.2466/pr0.1959.5.3.654>
- Araujo, F. A., Carbone, E., Conell-Price, L., Dunietz, M. W., Jaroszewicz, A., Landsman, R., Lamé, D., Vesterlund, L., Wang, S. W., & Wilson, A. J. (2016). The slider task: an example of restricted inference on incentive effects. *Journal of the Economic Science Association*, 2(1), 1–12. <https://doi.org/10.1007/s40881-016-0025-7>
- Ariely, D., Bracha, A., & Meier, S. (2009). Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99(1), 544–555. <https://doi.org/10.1257/aer.99.1.544>
- Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, 76.
- Augenblick, N., Niederle, M., & Sprenger, C. (2015). Working over Time: Dynamic Inconsistency in

- Real Effort Tasks. *The Quarterly Journal of Economics*, 130(3), 1067–1115. <https://doi.org/10.1093/qje/qjv020>
- Augenblick, N., & Rabin, M. (2015). An experiment on time preference and misprediction in unpleasant tasks. *The Review of Economic Studies*. <https://doi.org/10.1093/restud/rdy019/4996235>
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215. <https://doi.org/10.1037/0033-295x.84.2.191>
- Bartling, B., Fehr, E., Maréchal, M., & Schunk, D. (2009). Egalitarianism and Competitiveness. *American Economic Review*, 99(2), 93–98. <https://doi.org/10.1257/aer.99.2.93>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Baumeister, R. F. (1984). Choking under pressure: Self-consciousness and paradoxical effects of incentives on skillful performance. *Journal of Personality and Social Psychology*, 46(3), 610. <https://doi.org/10.1037/0022-3514.46.3.610>
- Benabou, R., & Tirole, J. (2003). Intrinsic and Extrinsic Motivation. *Review of Economic Studies*, 489–520. <https://doi.org/10.1111/1467-937X.00253>
- Benndorf, V., Rau, H., & Sölch, C. (2014). Minimizing learning behavior in experiments with repeated real-effort tasks. Available at SSRN. <https://ssrn.com/abstract=2503029>
- Berger, J., & Pope, D. (2011). Can Losing Lead to Winning? *Management Science*, 57(5), 817–827. <https://doi.org/10.1287/mnsc.1110.1328>
- Bland, J. M., & Altman, D. G. (2003). Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology*, 22(1), 85–93. <https://doi.org/10.1002/uog.122>
- Blumkin, T., Ruffle, B. J., & Ganun, Y. (2012). Are income and consumption taxes ever really equivalent? Evidence from a real-effort experiment with real goods. *European Economic Review*, 56(6), 1200–1219. <https://doi.org/10.1016/j.euroecorev.2012.06.001>
- Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: Hamburg registration and organization online tool.

- European Economic Review*, 71(C), 117–120. <https://EconPapers.repec.org/RePEc:eee:eecrev:v:71:y:2014:i:c:p:117-120>
- Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: Hamburg registration and organization online tool. *European Economic Review*, 71(C), 117–120. <https://EconPapers.repec.org/RePEc:eee:eecrev:v:71:y:2014:i:c:p:117-120>
- Bonein, A., & Denant-Boèmont, L. (2015). Self-control, commitment and peer pressure: a laboratory experiment. *Experimental Economics*, 18(4), 543–568. <https://doi.org/10.1007/s10683-014-9419-7>
- Bortolotti, S., Devetag, G., & Ortmann, A. (2016). Group incentives or individual incentives? A real-effort weak-link experiment. *Journal of Economic Psychology*, 56, 60–73. <https://doi.org/10.1016/j.joep.2016.05.004>
- Bosman, R., & Winden, F. van. (2002). Emotional Hazard in a Power-to-Take Experiment. *Royal Economic Society, Wiley*, 112(476), 147. <http://www.jstor.org/stable/798435>
- Bowles, S., & Polanía-Reyes, S. (2012). Economic Incentives and Social Preferences: Substitutes or Complements? *Journal of Economic Literature*, 50(2). <https://doi.org/10.1257/jel.50.2.368>
- Breyer, B., & Danner, D. (2015). Grit Scale for Perseverance and Passion for Long-Term Goals. *GESIS - Zusammenstellung Sozialwissenschaftlicher Items Und Skalen*.
- Brüggen, A., & Strobel, M. (2007). Real effort versus chosen effort in experiments. *Economics Letters*, 96(2). <https://doi.org/10.1016/j.econlet.2007.01.008>
- Bühler, K. (1922). *Die geistige Entwicklung des Kindes* (3. Aufl.). Fischer.
- Calsamiglia, C., Franke, J., & Rey-Biel, P. (2013). The incentive effects of affirmative action in a real-effort tournament. *Journal of Public Economics*, 98, 15–31. <https://www.sciencedirect.com/science/article/pii/S0047272712001314>
- Carpenter, J., Holmes, J., & Matthews, P. (2014). “Bucket auctions” for charity. *Games and Economic Behavior*, 88, 260–276. <https://doi.org/10.1016/j.geb.2014.09.007>
- Carpenter, J., & Huet-Vaughn, E. (2017). *Real effort tasks* (A. Schram & A. Ule, Eds.; Handbooks

- of Research Methods and Applications series). Edward Elgar. <https://www.e-elgar.com/shop/handbook-of-research-methods-and-applications-in-experimental-economics>
- Chan, J. C. (1991). Response-order effects in Likert-type scales. *Sage Publications Sage CA: Thousand Oaks, CA*, 51(3), 531–540. journals.sagepub.com
- Charness, G., Cobo-Reyes, R., & Sánchez, Á. (2016). The effect of charitable giving on workers' performance: Experimental evidence. *Journal of Economic Behavior & Organization*, 131, 61–74. <https://doi.org/10.1016/j.jebo.2016.08.009>
- 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>
- Charness, G., & Grieco, D. (2014). Creativity and financial incentives. *University of California, Santa Barbara, Working Paper*.
- Charness, G., & Kuhn, P. (2011). *Handbook of Labor Economics*. 4, 229–330. [https://doi.org/10.1016/S0169-7218\(11\)00409-6](https://doi.org/10.1016/S0169-7218(11)00409-6)
- Charness, G., Masclet, D., & Villeval, M. (2013). The Dark Side of Competition for Status. *Management Science*, 131105054351001. <https://doi.org/10.1287/mnsc.2013.1747>
- Chen, D. L., 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>
- Cooper, D. J., & Kagel, J. H. (2016). Other regarding preferences: A selective survey of experimental results. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics, volume two* (Vol. 2, pp. 217–289). Princeton university press. <https://doi.org/10.1515/9781400883172>
- Cooper, D. J., & Kagel, J. H. (2016). Other regarding preferences: A selective survey of experimental results. In J. H. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics, volume two* (Vol. 2, pp. 217–289). Princeton university press. <https://doi.org/10.1515/9781400883172>
- Corgnet, B., Gómez-Miñambres, J., & Hernán-González, R. (2016). Goal Setting and Monetary Incen-

- tives: When Large Stakes Are Not Enough. *Management Science*, 61(12). <https://doi.org/10.1287/mnsc.2014.2068>
- Corgnet, B., Hernan-Gonzalez, R., & Rassenti, S. J. (2011). *Real effort, real leisure and real-time supervision: Incentives and peer pressure in virtual organizations*.
- Corgnet, B., Hernán-González, R., Kujal, P., & Porter, D. (2015). The Effect of Earned Versus House Money on Price Bubble Formation in Experimental Asset Markets. *Review of Finance*, 19(4), 1455–1488. <https://doi.org/10.1093/rof/rfu031>
- Corgnet, B., Hernán-González, R., & Schniter, E. (2015). Why real leisure really matters: incentive effects on real effort in the laboratory. *Experimental Economics*, 18(2). <https://doi.org/10.1007/s10683-014-9401-4>
- Cox, J. C., & Sadiraj, V. (2019). *Incentives* (A. Schram & A. Ule, Eds.; Handbooks of Research Methods and Applications series}). Edward Elgar. <https://www.e-elgar.com/shop/handbook-of-research-methods-and-applications-in-experimental-economics>
- Cronbach, L. J. (1949). *Essentials of psychological testing*. <https://doi.org/10.1002/sce.3730350432>
- Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. Jossey-Bass.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and Reliability of the Experience-Sampling Method. *The Journal of Nervous and Mental Disease*, 175(9), 526–536. <https://doi.org/10.1097/00005053-198709000-00004>
- DellaVigna, S., List, J. A., Malmendier, U., & Rao, G. (2016). *Estimating social preferences and gift exchange at work*.
- DellaVigna, S., & Pope, D. (2016). What motivates effort? Evidence and expert forecasts. *NBER WORKING PAPER SERIES*. www.nber.org
- Desselle, S. P. (2005). Construction, Implementation, and Analysis of Summated Rating Attitude Scales. *American Journal of Pharmaceutical Education*, 69(5), 97. <https://doi.org/10.5688/aj690597>
- Dickinson, D. L. (1999). An Experimental Examination of Labor Supply and Work Intensities. *Journal*

- of *Labor Economics*, 17(4). <https://doi.org/10.1086/209934>
- Dijk, F. van, Sonnemans, J., & Winden, F. van. (2001). Incentive systems in a real effort experiment. *European Economic Review*. [https://doi.org/10.1016/S0014-2921\(00\)00056-8](https://doi.org/10.1016/S0014-2921(00)00056-8)
- Dohmen, T., & Falk, A. (2011). Performance Pay and Multidimensional Sorting: Productivity, Preferences, and Gender. *American Economic Review*, 101(2), 556–590. <https://doi.org/10.1257/aer.101.2.556>
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and Passion for Long-Term Goals. *Journal of Personality and Social Psychology*, 92(6), 1087–1101. <https://doi.org/10.1037/0022-3514.92.6.1087>
- Duncker, K. (1941). On Pleasure, Emotion, and Striving. *Philosophy and Phenomenological Research*, 1(4), 391. <https://doi.org/10.2307/2103143>
- Dutcher, G., Salmon, T., & Saral, K. (2015). Is 'Real' Effort More Real? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2701793>
- Eckartz, K. (2014). Task enjoyment and opportunity costs in the lab: The effect of financial incentives on performance in real effort tasks. *Jena Economic Research Papers*.
- Eckartz, K., Kirchkamp, O., & Schunk, D. (2012). How do incentives affect creativity? *CESifo Working Paper Series*, 4049. https://ideas.repec.org/p/ces/ceswps/_4049.html
- Engeser, S. (2005). *Messung des expliziten Leistungsmotivs: Kurzform der Achievement Motives Scale*. https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PGA/bilder/Engeser__2005__Kurzform_der_AMS.pdf
- Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32, 158–172. <https://doi.org/10.1007/s11031-008-9102-4>
- Engeser, S., Rheinberg, F., Vollmeyer, R., & Bischoff, J. (2005). Motivation, Flow-Erleben und Lernleistung in universitären Lernsettings. *Zeitschrift für pädagogische Psychologie*, 19(3), 159–172. <https://doi.org/10.1024/1010-0652.19.3.159>

- Engeser, S., & Vollmeyer, R. (2005). Tätigkeitsanreize und Flow-Erleben. In R. Vollmeyer (Ed.), *Motivationspsychologie und ihre anwendung*. Kohlhammer.
- Eriksson, T., Mao, L., & Villeval, M. (2017). Saving face and group identity. *Experimental Economics*, 20(3), 622–647. <https://doi.org/10.1007/s10683-016-9502-3>
- Eriksson, T., Poulsen, A., & Villeval, M. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, 16. <https://doi.org/https://doi.org/10.1016/j.labeco.2009.08.006>
- Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative Earnings and Giving in a Real-Effort Experiment. *American Economic Review*, 101(7), 3330–3348. <https://doi.org/10.1257/aer.101.7.3330>
- Erkal, N., Gangadharan, L., & Xiao, K. (2017). Monetary and non-monetary incentives in real-effort tournaments. <https://doi.org/10.1016/j.eurocorev.2017.10.021>
- Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, 27(5). <https://doi.org/10.1177/0272989x07304449>
- Fahr, R., & Irlenbusch, B. (2000). Fairness as a constraint on trust in reciprocity: earned property rights in a reciprocal exchange experiment. *Economics Letters*, 66(3). [https://doi.org/10.1016/s0165-1765\(99\)00236-0](https://doi.org/10.1016/s0165-1765(99)00236-0)
- Falk, A., & Huffman, D. (2007). Studying labor market institutions in the lab: Minimum wages, employment protection, and workfare. *Journal of Institutional and Theoretical Economics JITE*, 163.
- Falk, A., & Ichino, A. (2006). Clean Evidence on Peer Effects. *Journal of Labor Economics*, 24(1), 39–57. <https://doi.org/10.1086/497818>
- Fleming, P., & Zizzo, D. (2015). A simple stress test of experimenter demand effects. *Theory and Decision*, 78(2), 219–231. <https://doi.org/10.1007/s11238-014-9419-2>
- Fochmann, M., Kiesewetter, D., & Sadrieh, A. (2012). Investment behavior and the biased perception of limited loss deduction in income taxation. *Journal of Economic Behavior & Organization*, 81(1), 230–242. <https://doi.org/10.1016/j.jebo.2011.10.014>

- Fochmann, M., Weimann, J., Blaufus, K., Hundsdoerfer, J., & Kiesewetter, D. (2013). Net Wage Illusion in a Real-Effort Experiment*. *The Scandinavian Journal of Economics*, 115(2). <https://doi.org/10.1111/sjoe.12007>
- Frank, B. (1998). Good news for experimenters: subjects do not care about your welfare. *Economics Letters*, 61(2), 171–174. [https://doi.org/10.1016/s0165-1765\(98\)00162-1](https://doi.org/10.1016/s0165-1765(98)00162-1)
- Friedman, H. H., ...P. J. of, & Simcha. (1994). *The biasing effects of scale-checking styles on response to a Likert scale*. 792. rangevoting.org
- Fu, Q., Ke, C., & Tan, F. (2015). “Success breeds success” or “Pride goes before a fall?” Teams and individuals in multi-contest tournaments. *Games and Economic Behavior*, 94, 57–79. <https://doi.org/10.1016/j.geb.2015.09.002>
- Gächter, S., Huang, L., & Sefton, M. (2016). Combining “real effort” with induced effort costs: the ball-catching task. *Experimental Economics*, 19(4), 687–712. link.springer.com
- Ghatak, M., & Mueller, H. (2011). *Thanks for nothing? Not-for-profits and motivated agents*.
- Gill, D., & Prowse, V. (2015). A novel computerized real effort task based on sliders. *Institute of Labor Economics*, 5801. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.710.7007&rep=rep1&type=pdf>
- Giusti, G., & Dopeso-Fernández, R. (2018). Incentive Magnitude, Reference Point Shifting and Intrinsic Motivation: A Laboratory Experiment. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3231265>
- Gneezy, U., & List, J. A. (2006). Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments. *Econometrica*, 74(5), 1365–1384. <https://doi.org/10.1111/j.1468-0262.2006.00707.x>
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118.
- Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3), pp. 791–810. <https://www.jstor.org/stable/2586896>

- Gravert, C. (2014). *Examples of real effort tasks*. ESA Experimental Methods Discussion. [https://groups.google.com/forum/#!searchin/esa-discuss/Comparable\\$20but\\$20Different\\$20Real\\$20Effort\\$20Tasks%7Csort:date/esa-discuss/XOV5YJgi-Gk/DidUdb9LxSQJ](https://groups.google.com/forum/#!searchin/esa-discuss/Comparable$20but$20Different$20Real$20Effort$20Tasks%7Csort:date/esa-discuss/XOV5YJgi-Gk/DidUdb9LxSQJ)
- Günther, C., Ekinçi, N. A., Schwierén, C., & Strobel, M. (2010). Women can't jump?—An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, 75(3), 395–401. <https://doi.org/10.1016/j.jebo.2010.05.003>
- Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching and Learning*, 7(6), 836–850. <https://doi.org/10.1016/j.cptl.2015.08.001>
- Hayashi, A. T., Nakamura, B. K., & Gamage, D. (2013). Experimental Evidence of Tax Salience and the Labor–Leisure Decision. *Public Finance Review*, 41(2), 203–226. <https://doi.org/10.1177/1091142112460726>
- Heckhausen, H., & Rheinberg, F. (1980). Lernmotivation im Unterricht, erneut betrachtet. *Unterrichtswissenschaft*, 8, 7–47.
- Heckhausen, J., & Heckhausen, H. (Eds.). (2018). *Motivation and Action* (3rd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-65094-4_1
- Hennig-Schmidt, H., & Sadrieh, A. (2010). In search of workers' real effort reciprocity – a field and a laboratory experiment. *Journal of the European Economic Association*. <https://doi.org/10.1111/j.1542-4774.2010.tb00541.x>
- Heyman, J., & Ariely, D. (2004). Effort for Payment. *Psychological Science*, 15(11), 787–793. <https://doi.org/10.1111/j.0956-7976.2004.00757.x>
- Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7.
- Hollard, G., Massoni, S., & Vergnaud, J.-C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, 80(3), 363–387. <https://doi.org/10.1007/s11238-015-9509-9>
- Houser, D., Schunk, D., Winter, J., & Xiao, E. (2017). Temptation and Commitment in the Laboratory.

- Games and Economic Behavior*. <https://doi.org/10.1016/j.geb.2017.10.025>
- Houy, N., Nicolai, J., & Villeval, M. (2016). Doing Your Best When Stakes are High? Theory and Experimental Evidence. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2735343>
- Hsu, J. (1996). *Multiple Comparisons: Theory and Methods*. Chapman; Hall/CRC.
- Huang, L., & Murad, Z. (2017). Impact of Relative Performance Feedback on Beliefs, Preferences and Performance across Dissimilar Tasks. *University of Surrey, Discussion Paper Series*, 02.
- Imas, A. (2014). Working for the “warm glow”: On the benefits and limits of prosocial incentives. *Journal of Public Economics*, 114, 14–18. <https://doi.org/10.1016/j.jpubeco.2013.11.006>
- Jaeggi, S. M., Buschkuhl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, 18(4), 394–412. <https://doi.org/10.1080/09658211003702171>
- Johnson, D., & Creech, J. C. (1983). Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error. *American Sociological Review*, 48(3), 398. <https://doi.org/10.2307/2095231>
- Jones, D., & Linardi, S. (2014). Wallflowers: Experimental Evidence of an Aversion to Standing Out. *Management Science*, 60(7), 1757–1771. <https://doi.org/10.1287/mnsc.2013.1837>
- 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>
- Kephart, C. (2017). *oTree project with a number of Real Effort Task (RET) example apps*. https://github.com/EconomiCurtis/otree_rets
- Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of Behavioral Research*. <http://lib.ugent.be/catalog/rug01:000845957>
- Kessler, J. B., & Norton, M. I. (2016). Tax aversion in labor supply. *Journal of Economic Behavior & Organization*, 124, 15–28. <https://doi.org/10.1016/j.jebo.2015.09.022>
- Konow, J. (2000). Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions. *Amer-*

- ican Economic Review*, 90(4). <https://doi.org/10.1257/aer.90.4.1072>
- Kosfeld, M., & Neckermann, S. (2011). Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics*, 3(3), 86–99. <https://doi.org/10.1257/mic.3.3.86>
- Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *International Journal of Nursing Studies*, 48(6), 661–671. <https://doi.org/10.1016/j.ijnurstu.2011.01.016>
- Kovaleva, A., Beierlein, C., Kemper, C., & Rammstedt, B. (2014). Internale-Externale-Kontrollüberzeugung-4 (IE-4). D. Danner, & A. Glöckner-Rist, *Zusammenstellung Sozialwissenschaftlicher Items Und Skalen*.
- Krippendorff, K. (2021). *The reliability of generating data (forthcoming)*. Taylor; Francis.
- Ku, H., & Salmon, T. C. (2012). The Incentive Effects of Inequality: An Experimental Investigation. *Southern Economic Journal*, 79(1), 46–70. <https://doi.org/10.4284/0038-4038-79.1.46>
- Kuehl, R. O. (2000). *Design of experiments: statistical principles of research design and analysis*. Duxbury.
- Kuhl, J. (1983). *Motivation, Konflikt und Handlungskontrolle*. Berlin : Springer-Verlag.
- Kuhl, J. (1990). Kurzanweisung zum Fragebogen HAKEMP 90. *Manuskript. Fachbereich Psychologie, Universität Osnabrück*.
- Kuhl, J., & Beckmann, J. (Eds.). (1985). *Action control: From cognition to behavior*. Springer. <https://www.springer.com/gp/book/9783642697487>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Laske, K., & Schröder, M. (2017). *Quantity, Quality and Originality: The Effects of Incentives on Creativity*.
- Lei, V., Noussair, C. N., & Plott, C. R. (2001). Nonspeculative Bubbles in Experimental Asset Markets: Lack of Common Knowledge of Rationality vs. Actual Irrationality. *Econometrica*, 69(4), 831–859. <https://doi.org/10.1111/1468-0262.00222>

- Lezzi, E., Fleming, P., & Zizzo, D. J. (2015). *Does it matter which effort task you use? a comparison of four effort tasks when agents compete for a prize.*
- Lévy-Garboua, L., Masclet, D., & Montmarquette, C. (2009). A behavioral Laffer curve: Emergence of a social norm of fairness in a real effort experiment. *Journal of Economic Psychology*, 30(2), 147–161. <https://doi.org/10.1016/j.joep.2008.09.002>
- Linardi, S., & McConnell, M. A. (2011). No excuses for good behavior: Volunteering and the social environment. *Journal of Public Economics*, 95(5-6), 445–454. <https://doi.org/10.1016/j.jpubeco.2010.06.020>
- List, J. A., & Momeni, F. (2017). When Corporate Social Responsibility Backfires: Theory and Evidence from a Natural Field Experiment. *NBER Working Papers*, No. 24169. <https://doi.org/10.3386/w24169>
- Masclet, D., Peterle, E., & Larribeau, S. (2015). Gender differences in tournament and flat-wage schemes: An experimental study. *Journal of Economic Psychology*, 47, 103–115. <https://doi.org/10.1016/j.joep.2015.01.003>
- McMahon, M. (2015). *Better Lucky Than Good: The Role of Information in Other-Regarding Preferences.* https://trace.tennessee.edu/utk_graddiss/3353
- Mohnen, A., Pokorny, K., & Sliwka, D. (2008). Transparency, Inequity Aversion, and the Dynamics of Peer Pressure in Teams: Theory and Evidence. *Journal of Labor Economics*, 26(4), 693–720. <https://doi.org/10.1086/591116>
- Montmarquette, C., Rullière, J.-L., Villeval, M.-C., & Zeiliger, R. (2004). Redesigning Teams and Incentives in a Merger: An Experiment with Managers and Students. *Management Science*, 50(10), 1379–1389. <https://doi.org/10.1287/mnsc.1040.0280>
- Nabanita, Datta Gupta, Poulsen, A., & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1), 816–835. <https://doi.org/10.1111/j.1465-7295.2011.00378.x>
- Neyse, L., Friedl, A., & Schmidt, U. (2014). Payment Scheme Changes and Effort Provision: The Effect of Digit Ratio. *MPRA Working Papers*. <https://mpra.ub.uni-muenchen.de/59549/>

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much? *The Quarterly Journal of Economics*, 122, 1067–1101. <https://www.jstor.org/stable/25098868>

Nikiforakis, N., Noussair, C. N., & Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, 96(9-10), 797–807. <https://doi.org/10.1016/j.jpubeco.2012.05.014>

Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15(5), 625–632. <https://doi.org/10.1007/s10459-010-9222-y>

Noussair, C. N., & Stoop, J. (2015). Time as a medium of reward in three social preference experiments. *Experimental Economics*, 18(3), 442–456. <https://doi.org/10.1007/s10683-014-9415-y>

Oehlert, G. W. (2010). *A first course in design and analysis of experiments*. <http://users.stat.umn.edu/~book/fcdae.pdf>

OfficeTeam. (2017). *Working hard or hardly working*. <http://rh-us.mediaroom.com/2017-07-19-WORKING-HARD-OR->

Pokorny, K. (2008). Pay – but do not pay too much. *Journal of Economic Behavior & Organization*, 66(2), 251–264. <https://doi.org/10.1016/j.jebo.2006.03.007>

Reinstein, D., & Riener, G. (2009). House money effects on charitable giving: an experiment. *Unpubl Work Pap*. https://www.researchgate.net/publication/255570050_House_Money_Eects_on_Charitable_Giving_An_Experiment

Rey-Biel, P., Sheremeta, R. M., & Uler, N. (2018). When income depends on performance and luck: The effects of culture and information on giving. *MPRA Paper*, 83940. <https://mpra.ub.uni-muenchen.de/83940/>

Rheinberg, F. (1989). *Zweck und Tätigkeit : motivationspsychologische Analysen zur Handlungsveranlassung: Vols. Band 11*. Göttingen : Zürich [etc.] : Verlag für Psychologie Hogrefe.

Rheinberg, F. (2004). *Motivationsdiagnostik: Vols. Band 5*. Hogrefe.

Rheinberg, F. (2006). Motivationsdiagnostik. In F. Petermann & M. Eid (Eds.), *Handbuch der psy-*

- chologischen diagnostik (pp. 511–521). Hogrefe. <https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/11609>
- Rheinberg, F. (2011). Motivation, Volition und Ziele. *Enzyklopädie Der Psychologie: Methodologie Und Methoden-Psychologische Diagnostik-Persönlichkeitsdiagnostik*, 4.
- Rheinberg, F., & Engeser, S. (2018). *Intrinsic Motivation and Flow*. 579. https://doi.org/10.1007/978-3-319-65094-4_14
- Rheinberg, F., Iser, I., & Pfauser, S. (1997). Freude am Tun und/oder zweckorientiertes Schaffen? Zur transsituativen Konsistenz und konvergenten Validität der Anreizfokus-Skala. *DIAGNOSTICA-GOTTINGEN-*, 43.
- Rheinberg, F., & Manig, Y. (2003). Was macht Spaß am Graffiti-Sprayen? *Report Psychologie*, 222–234. <https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/544%22>, <https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-6296%22%5D>
- Rheinberg, F., & Vollmeyer, R. (2003). Flow-Erleben in einem Computerspiel unter experimentellvariierten Bedingungen. *Zeitschrift für Psychologie*, 211, 161–170. <https://doi.org/10.1026//0044-3409.211.4.161>
- Rheinberg, F., & Vollmeyer, R. (2012). *Motivation* (8th ed.). Kohlhammer. <http://www.psych.uni-potsdam.de/people/rheinberg/personal/pubs-books-d.html>
- Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen (Langversion, 2001). *Diagnostica*, 2.
- Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). *Die erfassung des flow-erlebens*.
- Rosaz, J., Slonim, R., & Villeval, M. (2016). Quitting and peer effects at work. *Labour Economics*, 39, 55–67. <https://doi.org/10.1016/j.labeco.2016.02.002>
- Rosenberg, B. D., Navarro, M. A., & Frey, B. B. (2018). *Semantic Differential Scaling* (pp. 1504–1507). SAGE Publications, Inc. <http://dx.doi.org/10.4135/9781506326139.n624>
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement.

- Psychological Monographs: General and Applied*, 80(1). <https://doi.org/10.1037/h0092976>
- Rutström, E. E., & Williams, M. B. (2000). Entitlements and fairness: an experimental study of distributive preferences. *Journal of Economic Behavior & Organization*, 43(1), 75–89. [https://doi.org/10.1016/s0167-2681\(00\)00109-8](https://doi.org/10.1016/s0167-2681(00)00109-8)
- Ryan, R. M. (1982). Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of Personality and Social Psychology*, 43.
- Schallberger, U. (2005). Kurzskaalen zur Erfassung der Positiven Aktivierung, Negativen Aktivierung und Valenz in Experience Sampling Studien (PANAVA-KS). *Theoretische Und Methodische Grundlagen, Konstruktvalidität Und Psychometrische Eigenschaften Bei Der Beschreibung Intra-Und Interindividueller Unterschiede*.
- Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C., & Rammstedt, B. (2014). Kurzsкала kristalline Intelligenz (BEFKI GC-K). *GESIS - Zusammenstellung Sozialwissenschaftlicher Items Und Skalen*, 10. <https://doi.org/10.6102/zis220>
- Schneider, J. W., & McGrew, K. S. (2012). The Cattell-Horn-Carroll Model of Intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). The Guilford Press. <https://psycnet.apa.org/record/2012-09043-004>
- Schulze, C. (2020). *Testing corporate social responsibility in a meaningful work-environment: Its impact on cheating and its spillover effects on further prosocial behavior* [Master's Thesis]. ETH Zurich.
- Schwarz, N., Knäuper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *The Public Opinion Quarterly*, 55(4), 570–582. <https://www.jstor.org/stable/2749407>
- Simons, D. J., & Chabris, C. F. (1999). Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception*, 28(9), 1059–1074. <https://doi.org/10.1068/p281059>
- Smith, V. L. (1982). Microeconomic Systems as an Experimental Science. *The American Economic Review*, 72(5), 923. <http://www.jstor.org/stable/1812014>
- Smith, V. L. (2010). *Behavioural and Experimental Economics* (p. 120). <https://doi.org/10.1057/>

9780230280786_16

- Sullivan, G. M., & Jr, A. R. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, 5(4), 541–542. <https://doi.org/10.4300/jgme-5-4-18>
- Sutter, M., & Weck-Hannemann, H. (2003). Taxation and the Veil of Ignorance – A Real Effort Experiment on the Laffer Curve. *Public Choice*, 115(1/2). <https://doi.org/10.1023/a:1022873709156>
- Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3), 3–3. <https://doi.org/10.1145/1077246.1077253>
- Swenson, C. W. (1988). Taxpayer behavior in response to taxation An experimental analysis. *Journal of Accounting and Public Policy*, 7(1), 1–28. [https://doi.org/10.1016/0278-4254\(88\)90002-6](https://doi.org/10.1016/0278-4254(88)90002-6)
- Thaler, R. (1985). Mental Accounting and Consumer Choice. *Marketing Science*, 4(3), 199–214. <https://doi.org/10.1287/mksc.4.3.199>
- Thaler, R. H., & Johnson, E. J. (1990). Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice. *Management Science*, 36(6), 643–660. <https://doi.org/10.1287/mnsc.36.6.643>
- Torgler, B. (2002). Vertical and Exchange Equity in a Tax Morale Experiment. *WWZ Discussion Papers*.
- Trougakos, J. P., & Hideg, I. (2009). *Momentary work recovery: The role of within-day work breaks* (S. Sonnentag, "Pamela. L. Perrewé", & "Daniel. C. Ganster"), Eds.; Vol. 7, pp. 37–84). Emerald Group Publishing Limited. [https://doi.org/10.1108/s1479-3555\(2009\)0000007005](https://doi.org/10.1108/s1479-3555(2009)0000007005)
- Uebersax, J. S. (2006). *Likert scales: dispelling the confusion. Statistical methods for rater agreement.* <http://john-uebersax.com/stat/likert.htm>
- Uebersax, S. (2000). *Agreement on Interval-Level Ratings.* <http://john-uebersax.com/stat/cont.htm>
- Vandegrift, D., & Brown, P. (2003). Task difficulty, incentive effects, and the selection of high-variance strategies: an experimental examination of tournament behavior. *Labour Economics*, 10(4), 481–497. [https://doi.org/10.1016/s0927-5371\(03\)00033-2](https://doi.org/10.1016/s0927-5371(03)00033-2)
- Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., & Wagner, G. G. (2013). Validating

- an ultra-short survey measure of patience. *Economics Letters*, 120(2), 142–145. <https://doi.org/10.1016/j.econlet.2013.04.007>
- Vollmeyer, R., & Rheinberg, F. (2003). Task difficulty and flow. *EARLI-Conference at Padova*, 08–13.
- Wild, K.-P., Krapp, A., Schiefele, U., Lewalter, D., & Schreyer, I. (1995). Dokumentation und Analyse der Fragebogenverfahren und Tests. *Berichte Aus Dem DFG-Projekt „Bedingungen Und Auswirkungen Berufsspezifischer Lernmotivation*.
- Winter, F. (2017). A Library of Real Effort Tasks. *In Preparation*. https://www.coll.mpg.de/11577/Research-Group-_Mechanisms-of-Normative-Change
- Winter, F., Rauhut, H., & Helbing, D. (2012). How Norms Can Generate Conflict: An Experiment on the Failure of Cooperative Micro-motives on the Macro-level. *Social Forces*, 90(3). <https://doi.org/10.1093/sf/sor028>
- Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. *Journal of Labor Economics*, 32(1), 161–198. <https://doi.org/10.1086/673324>
- Zizzo, D. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1), 75–98. <https://doi.org/10.1007/s10683-009-9230-z>
- Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie Canadienne*, 34(4), 390. <https://doi.org/10.1037/h0078865>

Abstract (German)

In der experimentellen Forschung werden häufig so genannte "Aufgaben mit tatsächlichem Aufwand" verwendet, um das Erbringen von Leistung zu untersuchen. Dazu existiert eine grosse Vielfalt derartiger Aufgaben, von denen jede ihre eigenen Merkmale besitzt. Die Aufgaben werden in verschiedenen Anwendungsbereichen eingesetzt und weisen unterschiedliche Eigenschaften auf, die allerdings je nach Anwendungsbereich Vor- und Nachteile mit sich bringen. Diese Arbeit widmet sich der Wahl einer Aufgabe und deren Folgen.

Das *erste Kapitel* führt in die Thematik ein und liefert Hintergrundinformationen zur Messung des Aufwands in Experimenten. Da sich Aufgaben mit tatsächlichem Aufwand in ihrer Gestaltung stark unterscheiden, werden mehrere Möglichkeiten zu deren Klassifizierung vorgestellt. Die Klassifizierung der Aufgaben umfasst dabei ihren *Grad an Realitätsbezug*, das *Ausmass, in dem das generierte Resultat nützlich ist*, und die *Fähigkeiten und Charaktereigenschaften*, die erforderlich sind, um eine bestimmte Aufgabe gut auszuführen.

Das *zweite Kapitel* gibt einen Überblick über die Literatur, die sich kritisch mit dem Konzipieren und Implementieren von Aufgaben mit tatsächlichem Aufwand auseinandersetzt. Als Synthese dieser Literatur wird eine Reihe von *Gestaltungskriterien* vorgestellt. Sie zielen darauf ab, die *experimentelle Kontrolle* zu steigern und gleichzeitig die *grössere Realitätsnähe* von Aufwandsmessungen mit "tatsächlicher Anstrengung" im Vergleich zu jenen mit "deklarerter Anstrengung" zu erhalten. Um dies zu erreichen, werden *Gestaltungspraktiken* vorgestellt, die die experimentelle Kontrolle verbessern und zwar i) über die Aufwand-Kosten-Funktion, um sicherzustellen, dass die *freiwillige Bereitstellung von Anstrengung* auf ein Minimum beschränkt wird, sowie ii) über die Output-Produktions-Funktion, um sicherzustellen, dass *tatsächlich Aufwand erforderlich ist*, die Aufgabe zu erledigen.

Um Aufgaben im Hinblick auf diese Aspekte zu bewerten und zu vergleichen, wird im *dritten Kapitel* eine neue Methodik eingeführt, die *Umfrage bezüglich des tatsächlichen Aufwands einer Aufgabe*. Die Befragung wird von (potenziellen) Studienteilnehmern beantwortet und ermittelt deren subjektive Wahrnehmung der Aufgabengestaltung. Dies ist entscheidend, denn nur sie selbst können beurteilen, i) inwieweit eine Aufgabe sie zu freiwilliger Anstrengung motiviert hat und ii) wie anstrengend diese für sie tatsächlich war. Darüber hinaus werden die Ergebnisse einer ersten Erhebung, bei der sieben häufig verwendete Aufgabentypen miteinander verglichen wurden, vorgestellt.

Um die Auswirkung der Eigenschaften von Aufgaben auf die Anstrengungsmessung zu beleuchten, wird in *Kapitel vier* der Einfluss der Merkmale der Probanden auf ihre individuelle Leistung untersucht. Zu diesem Zweck enthält die in Kapitel drei vorgestellte Studie mehrere zusätzliche Elemente zur Charakterisierung der Studienteilnehmer. Mit Hilfe von Methoden der Motivationsdiagnostik und des Maschinellen Lernens lässt sich zeigen, dass Fähigkeiten, Persönlichkeit und Motivation einen grossen Teil der Variation in der beobachteten Anstrengung der Probanden erklären können.

Kapitel fünf bildet den Abschluss dieser Arbeit, fasst die Ergebnisse und Beiträge zusammen, setzt sie in Beziehung zueinander und gibt einen Ausblick auf zukünftige Forschungsmöglichkeiten.

Diese Arbeit zielt darauf ab, das Bewusstsein für die verschiedenen Eigenschaften von Aufgaben, ihre Unterschiede und ihre jeweilige Eignung für eine konkrete Anwendung zu schärfen. Dazu liefert die Arbeit mehrere konzeptionelle und methodische Beiträge und dient dem Praktiker zur Klassifizierung, Gestaltung, Auswahl und Implementierung von Aufgaben. Zusammenfassend lässt sich sagen, dass Aufgaben weder neutral, noch einfach austauschbar sind. Aus diesem Grund ist die Wahl der Aufgabe von entscheidender Bedeutung und muss stets auf die zu untersuchende Fragestellung abgestimmt sein.

Christian Waloszek

Ulrichstrasse 14, 8032 Zurich, Switzerland
cwaloszek@ethz.ch, 26.08.1985

Education

- 9/2014–3/2021 Doctoral candidate in Economics, Chair of Public Economics, ETH Zurich, Switzerland
- 1/2015–12/2015 Swiss Program for Beginning Doctoral Students in Economics, Study Center Gerzensee, Switzerland, supported by the Swiss National Bank
- 4/2010–11/2013 Master in Physics, Free University of Berlin, Berlin, Germany
with minor in economics
- 11/2012–11/2013 Preparation of Master Thesis, Chair of Sociology, in particular of Modeling and Simulation, ETH Zurich, Switzerland
- 10/2005–3/2010 Bachelor in Physics, University of Konstanz, Konstanz, Germany
with minors in chemistry and biology
- 7/2008–12/2008 Semester abroad, Uni. of the Western Cape, Cape Town, South Africa
 - 9/2007–12/2007 Semester abroad aboard *The Scholar Ship*
- 9/1995–7/2004 Humboldt-Gymnasium Ulm, Ulm, Germany
-

Practical experience

- 9/2014–3/2021 Doctoral researcher at the Chair of Public Economics, ETH Zurich, Switzerland
- 1/2014–4/2014 Research Associate at the Center for Law & Economics, ETH Zurich, Switzerland
- 12/2012–10/2013 Research assistant at the Chair of Sociology, in particular Modeling and Simulation, ETH Zurich, Switzerland
- 11/2009–10/2011 Founder and board member of *Recrear*, Berlin, Germany, an international nonprofit organization whose goal is to inspire and support young people to get involved in society
- 3/2010–8/2010 Organizer of two conferences at the BMW Foundation Herbert Quandt
 - 11/2009–9/2010 Main organizer of the summer academy *Recrear.beta*
-

Social engagement

- 6/2018–present Fitness coach, Rowing Club Zurich, Zurich, Switzerland
- 8/2016 Representative of ETH Zurich, *DGIST World-class University Rowing Festival*, South Korea
- 4/2012–7/2014 Member of the *Young Scholars Initiative*, Institute for New Economic Thinking
- 11/2011–5/2013 Member of the *Fortschrittsforum* of the Friedrich-Ebert-Stiftung, Berlin, Germany
- 3/2011–9/2012 Ambassador of the *Living Proof Campaign* of ONE and the Bill and Melinda Gates Foundation
- 8/2006–5/2007 Delegate at the *National Model United Nations 2007*, New York, USA
- 9/2006–10/2007 Head of the *Abstract Art Group*, University of Konstanz, Konstanz, Germany
- 9/2002–7/2004 Member of the Youth Parliament, Ulm, Germany
-

Skills

- Computer skills: R, R markdown, RStudio, Git, Adobe Creative Suite, \LaTeX , MS Office
- Language skills: German (native), English (professionally fluent), French (beginner)

PEC Dissertation Series

1. Waloszek, Christian: Real-Effort Tasks in Experiments: The Task Choice Matters, 2021

ETH Zürich
Chair of Public Economics
Leonhardstrasse 21
8092 Zurich
Switzerland

www.pec.ethz.ch

Publisher: Chair of Public Economics
Author: Christian Waloszek
Printing press: ETH Print + Publish

© ETH Zürich, 2021