PHD

The time-course and quality of ideation: singular tasks, sequential tasks and support for well-formed ideas

Keating, Christina

Award date:
2019

Awarding institution:
University of Bath

Link to publication

The time-course and quality of ideation: singular tasks, sequential tasks and support for well-formed ideas

submitted by

Christina Keating

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Computer Science

March 2019

COPYRIGHT

Summary

Ideation is described as the process of generating useful ideas to help reach a specific goal or outcome. It is one of the task types with the longest standing roles in experimental studies of creativity. Despite being a well studied concept, many aspects of the cognitive processes underpinning ideation are poorly understood. A main model of thinking that is referred to when looking at ideation studies is the concept of Search for Ideas in Associative Memory (SIAM), which suggests that ideas are generated in semantically related clusters, due to the way we synthesise these ideas from memory. The SIAM model of thinking is likened to that of information foraging theories in this thesis, which forms the basis for our studies. In this thesis we investigate the time-course pattern of idea generation in an ideation task and ask whether we can find any influences on this. In particular, we look to classic stopping heuristics to try to explain the time-course of ideas across a continuous single question ideation task and the giving-up strategies used in a multiple sequential question ideation task. Throughout the work presented in this thesis, we ran into methodological problems with inter-rater disagreement between coders rating the sets of ideas on commonly used metrics such as novelty and value, as well as classifying ideas in semantic categories. The second area we investigate in this thesis is whether we can support ideation in such a way that judgement of ideas becomes more reliable.

We ran three ideation studies in order to develop theory in support of applications to assist in the ideation process. In our first study, we looked at continuous ideation across a single ideation task. Key findings from this study showed a weak verification of semantic clustering of ideas generated, however, we were able to replicate main effects others had found using this method. In the second study, we looked at ideation across multiple questions. Our key finding directly mirrors that found in other (non-ideation) studies on discretionary task switching: people combine rate of return and sub-goal completion as strategies for abandoning a task. In the third study, we addressed the issues of low inter-rater agreement by performing simple instructional manipulations for ideators. Key findings show a that the use of thematic roles have a positive effect on judged value of ideas. Novelty judgement agreement was also positively affected, although this was seen across conditions.

Although a lot more research is required in order to fully understand the cognitive processes in ideation, we hope that this work shows that by detailed experimental analysis of ideation we might learn some possible interface interventions through which ideators might be supported.

Acknowledgements

First and foremost, I would like to thank my supervisors, Professors Stephen J. Payne and Eamonn O'Neill. In the 6 months before submitting this thesis, Steve put up with my terrible questions and countless impromptu meeting requests. This thesis would not have been submitted was it not for his help and support throughout.

Many thanks to everyone in the Department of Computer Science for always being there with words of wisdom and helping where-ever you could, especially those of you who willingly jumped in to help me make sense of my data or debug experiments, even last minute.

Thank you to my family away from home, many of you in the Department of Computer Science, some outside, two I live with. Thank you for understanding me and always being there whenever I need a friend (or cake).

I especially want to thank my family for always being there for me and being supportive no matter what I choose to do in life. I wouldn't have made it through this without endless FaceTime support from mor, far, Maria, Momme, Moffe, Henrik and Rachel. And of course Richard, for being the best team-mate and co-pilot and for carrying me through these last few months. You make my life happier than I could ever have wished for.

Contents

1 Introduction 1

1.1 Research questions 5

1.2 Research methodology 7

1.2.1 Research ethics 7

1.3 Contributions 8

1.4 Thesis structure 9

2 Literature review 11

2.1 Ideation 12

2.1.1 What is ideation? 12

2.1.2 The process of ideation 15

2.2 Ideation Support Systems 17

2.2.1 Ideation studies in Human Computer Interaction 18

2.2.2 Flow versus impasse 20

2.3 Relations of ideation with Foraging Theory 23

2.3.1 The Search for Ideas in Associative Memory 23

2.3.2 Foraging Theory 25

2.3.3 Internal memory search and giving-up rules 30

2.4 Metrics and Measurements in Ideation Support Systems 32

2.5 Chapter Summary 34

3 Time-course of single ideation tasks 36

3.1 Chapter background 36

3.2 Study 1a: Ideation cued by an environment question 40

3.2.1 Method 40

3.2.2 General analytic approach 45

3.2.3 Results and discussion 47

3.2.4 Summary and conclusions 57

3.3 Study 1b: Ideation cued by a health question . . . 60
3.3.1 Method . . . 61
3.3.2 Results and discussion . . . 64
3.3.3 Summary and conclusions . . . 73
3.4 Semantic distance between adjacent ideas . . . 74
3.4.1 Method . . . 75
3.4.2 Results and discussion . . . 76
3.4.3 Summary and conclusions . . . 84
3.5 Chapter Summary . . . 85

4 Time-course of ideation across multiple ideation tasks 87
4.1 Chapter background . . . 88
4.2 Study 2: Ideation strategy across multiple consecutive tasks . . . 91
4.2.1 Method . . . 92
4.2.2 Results and discussion . . . 98
4.3 Chapter Summary . . . 107

5 Judging the quality of ideas: novelty and value 110
5.1 Method . . . 111
5.2 Results and discussion . . . 112
5.2.1 Coder and rating characteristics . . . 112
5.2.2 Quality coding . . . 115
5.2.3 The properties of agreed-good v agreed-bad v no-agreement ideas . 116
5.3 Chapter Summary . . . 117

6 Supporting the well-formedness of ideas 119
6.1 Chapter background . . . 121
6.2 Study 3: Prompting increased specificity of ideas in an online ideation task 122
6.2.1 Method . . . 122
6.2.2 Results and discussion . . . 131
6.3 Chapter summary and conclusions . . . 142

7 Discussion, conclusions and implications 144
7.1 General discussion . . . 145
7.1.1 The time-course of ideation and support for well-formed ideas . . . 147
7.1.2 Limitations and considerations for future work . . . 151
7.2 Implications for the design of ideation tools . . . 155

7.2.1 Design of ideation tools for the extension of methodology in support of experimenters . . . 156
7.2.2 Design guidelines for the development of ideation support tools . . . 159
7.3 Concluding remarks . . . 161

Bibliography **162**

Appendix **171**

A Appendices for studies 1a and 1b **172**
A.1 Participant information sheet S1 . . . 173
A.2 Participant consent form S1 . . . 174
A.3 Post-experiment questionnaire S1 . . . 175
A.4 Semantic difference coding instructions . . . 176

B Appendices for study 2 **177**
B.1 Questions for study 2 . . . 177
B.2 Participant information sheet S2 . . . 179
B.3 Participant consent form S2 . . . 180
B.4 Post-experiment questionnaire S2 . . . 181
B.5 Quality Coding Instructions . . . 183

C Appendices for study 3: screenshots of application built in Gorilla **184**

D General appendices **191**
D.1 Ethics . . . 191

List of Figures

1-1 The creative problem solving process, made up of stages of ideation-evaluation. Ideation is facilitated by creative thinking, evaluation facilitated by critical thinking. 2

2-1 Marginal value theorem (adapted from Charnov et al. (1976)) (Wilke et al., 2009) 26

2-2 Pattern of simplistic patch leaving behaviour theories. (a) Fixed-Number rule (b) Fixed-Time rule. A vertical jump on both graphs indicate an item found. The patch is left at the threshold time or threshold number of items found (denoted by the red line), regardless of patch potential. 27

2-3 Time pattern of patch leaving behaviour theories using items as an indicator of tendency to stay on patch. (a) Decremental rule (b) incremental rule. A vertical jump (upwards in incremental rule, downwards in decremental rule) indicate an item found. 28

2-4 Time pattern of patch leaving behaviour theories (a) giving-up time rule and (b) Green's rule. A vertical jump on both graphs indicate an item found. 29

3-1 Developed app for environment question study 1a. (a) Takes in participant number, (b) is the main ideation screen, showing an input area, a list of previously entered ideas and the question itself, (c) end of experiment screen appears automatically after 15 minutes, ending the ideation task. 44

3-2 Cumulative productivity (number of ideas generated) over time for each participant in environment question study 1a. 51

3-3 Cumulative productivity (number of ideas generated) over time for each participant in environment question study 1a, each vertical jump represents submitting an idea; horizontal lines represent response latencies. 51

3-4 Distribution of between category thinking times environment question study 1a showing a cluster of short thinking times and a long tail of longer thinking times, compared to mean within category thinking time (orange line). Note the > 20 overflow bin. 58

3-5 Cumulative productivity (number of ideas generated) over time for each participant in health question study 1b. 66

3-6 Cumulative productivity (number of ideas generated) over time for each participant in health question study 1b, each vertical jump represents submitting an idea; horizontal lines represent response latencies. 66

3-7 Distribution of between category thinking times health question study 1b, compared to mean within category thinking time (orange line). Note the > 20 overflow bin. 73

3-8 Summary correlation of difference ratings for study 1a and 1b: second coder mean difference ratings against each of the main coder difference ratings on the 10-point scale. 77

3-9 Distribution of ratings for main coder (438 datapoints) and second coder (152 datapoints) for studies 1a and 1b. 78

3-10 Correlation between mean difference scores and the fluency of an idea set; semantic difference scores calculated in data from studies 1a and 1b. 80

3-11 Thinking time as a function of semantic difference scores between adjacent items; semantic difference scores calculated in data from studies 1a and 1b. 81

4-1 Developed app for study 2: multiple sequential ideation. 95

4-2 The trajectories of number against time on task of four randomly chosen participants from study 2, in the 10 second delay condition. Vertical jumps represent finding an idea, horizontal lines represent thinking times. On the right, the thinking times (blue) and giving-up times (orange) plotted at the time during a task they occur for those same four participants. Note the variation in scales between participants. All scales are linear. 102

4-3 The trajectories of number against time on task of four randomly chosen participants from study 2, in the 25 second delay condition. Vertical jumps represent finding an idea, horizontal lines represent thinking times. On the right, the thinking times (blue) and giving-up times (orange) plotted at the time during a task they occur for those same four participants. Note the variation in scales between participants. All scales are linear. 103

4-4 Comparison of giving-up times, longest between-item time and mean times between items (thinking times) for each of the conditions in study 2. 104

4-5 The distribution of giving-up times, compared with the mean between-item thinking times (M=10.28, SD=5.23) in study 2. 105

4-6 Comparison of giving-up times of subjectively easy and hard tasks in study 2, against the median split on number of ideas generated, for each participant individually. 106

5-1 Distribution of novelty and value ratings as given by each individual coder; data coded is from study 2. 114

6-1 Study flow in Gorilla. Orange nodes represent randomisers that split participants first into prompt type then into question. Blue nodes represent the study screens themselves. Green nodes are questionnaire elements. Grey nodes control experiment logic (e.g. checkpoints verify that participants have not left the study before completion). 126

6-2 Ideation page in study 3, variation in prompts between the three conditions across health and phone question (note the question in these screenshots is that from the practice session.) 127

6-3 Estimated marginal means of *mean number of words*, averaged across ideas then participants, per prompt condition in each of the two questions in study 3. 134

6-4 Estimated marginal means of *fluency* per prompt condition in each of the two questions in study 3. 134

6-5 Estimated marginal means of log transformed thinking times per prompt condition in each of the two questions in study 3. A main effect for question is found, indicating people found the health question easier to generate ideas for in comparison to the phone question. 135

6-6 Cumulative productivity for each participant in the 6 conditions in study 3. All 6 conditions show a diminishing returns curve. Graphs on the left represent the health question in the three conditions: (a) no prompt, (c) implicit prompt, (e) explicit prompt. Graphs on the right represent the gain curves for the phone question in those same corresponding conditions. . . . 136

6-7 Summary of coder 4 difference ratings against each of coder 3's difference ratings on the 10-point scale for sample data in study 3. 139

7-1 Suggested design of a free-form interface that supports categorical ideation where categories are predetermined by the experimenter. 157

7-2 Suggested design of free-form interface that supports categorical ideation. The number and name of categories is determined by ideators, by selecting 'new category'. Ideas can be dragged and dropped between categories. Category boxes should resize dependent on number of ideas entered. 158

7-3 Suggested design alternative for visually cued length constraints on ideas, offering a multi-line box affording longer ideas than single line boxes. Colour-cued feedback on idea length, with red meaning too short, yellow meaning ok, green meaning well formulated. Alternatives to the standard traffic light colours advised in design for accessibility (e.g. colour blindedness). 160

7-4 Suggested design of template ideation tool that requires input into thematic role fields to inspire well-formulated ideas. 161

C-1 Study 3 consent form 185

C-2 Study 3 demographics form 186

C-3 Study 3 post-experiment questionnaire form 187

C-4 Instructions for ideators in the three different conditions in Gorilla 188

C-5 Ideation screens for no prompt and explicit conditions in Gorilla 189

C-6 Ideation screens for implicit example prompt (health and phone questions) in Gorilla 190

List of Tables

3.1 Category matrix for environment question consisting of 10 means and 5 goals. 46

3.2 The overall fluency, mean times and questionnaire responses for participants in environment question study 1a. Mean of means are used for fluency and timing values. Standard deviations in brackets. 49

3.3 Total number of categories (*diversity*) per participant according to two independent coders (IC) in environment question study 1a. Standard deviations shown in brackets. 52

3.4 Unclassified ideas and comments per independent coder in environment question study 1a. 53

3.5 The computation of ARC scores for each participant in environment question study 1a, using the codings of IC1. 54

3.6 The computation of ARC scores for each participant in environment question study 1a, using the codings of IC2 (variation between submitted ideas and N are due to coder leaving some ideas blank). 55

3.7 Category matrix for health question consisting of 11 means and 9 goals. . . 63

3.8 The overall fluency, mean times and questionnaire responses for participants in health question study 1b. Mean of means are used for fluency and timing values. Standard deviations shown in brackets. 65

3.9 Total number of categories (*diversity*) per participant according to two independent coders (IC) in health question study 1b. Standard deviations shown in brackets. 68

3.10 Unclassified ideas and comments per independent coder in health question study 1b. 68

3.11 The computation of ARC scores for each participant in health question study 1b, using the codings of IC3. 70

3.12 The computation of ARC scores for each participant in health question study 1b, using the codings of IC4. 70

3.13 Mean difference scores per participant compared to their overall *fluency* for data from studies 1a and 1b. Standard deviations shown in brackets. 80
3.14 Correlation between mean semantic difference scores and mean ideation timing metrics per participant for data from studies 1a and 1b. 82

4.1 Mean number of questions and ideas (SD) across all participants, split by delay condition and overall. Questions unattempted are questions seen and abandoned without generation of any ideas. Questions abandoned are questions that were abandoned by the ideator and not cut short by the experiment automatically ending. 99
4.2 Self-reported switching strategies (N=30) from study 2 participant questionnaire. Note that the numbers don't add up to 30, this is due to 6 people reporting multiple strategies (e.g. using an aggregate strategy involving time since last idea and a *failure* (N=5) or number of ideas generated (N=1)). . 107

5.1 Table showing mean ratings for novelty and value as given by each individual coder; data coded is from study 2. 113
5.2 Correlations of coder scores for novelty and value; data coded is from study 2.113
5.3 Correlations between timing measurements for each $participant \times question$ pair. No significant correlations were found between any of these metrics and the novelty and value scores given by any of the coders. 115

6.1 Table showing uneven split of participants between conditions due to issues with the integration between the Gorilla-Prolific platforms. 129
6.2 Table showing split of participants between conditions after removal of 13 datasets in which participants had not followed instruction. 130
6.3 Summary data from each of the 6 conditions in experiment 3. Number of words per idea, thinking times, typing times and response latencies are presented as the mean of means across participants, then condition. Standard deviations shown in brackets. 132
6.4 Correlations between coder 3 and coder 4 on each of the 6 conditions in study 3. All correlations are significant at the 0.01 level (2-tailed). 140
6.5 Correlations between the two quality coders on novelty and value. Novelty correlations have increased in comparison to previous study data, however, this has increased across all 3 conditions in the current study. *denotes correlations significant at the $P < .01$ level. 141

Chapter 1

Introduction

Creative problem solving plays an important part in our lives. It is not a process that is isolated to a particular setting. We encounter creative problem solving tasks in all facets of life: generation of new and competitive strategies in work settings, planning the perfect holiday in personal settings, or developing plans to make schools safer in community settings. Indeed, the need for creative thinking has repeatedly been identified as vital to maintaining a competitive advantage in organisations (Woodman, Sawyer and Griffin, 1993; Amabile, 1996; Basadur and Hausdorf, 1996).

Creative problem solving is often thought to be split into distinct stages or steps (Basadur, Graen and Green, 1982; Allen and Thomas, 2011; Sowden, Pringle and Gabora, 2015). Whilst the suggestions for stages that make up the creative process vary, they often include similar stages, e.g. problem clarification, problem solving, and implementation (Ochse, 1990). Each of these steps is considered to be an iteration of creative thinking (ideation) and critical thinking (evaluation) (Basadur, 1995; Dennett, 2017). That is, for each stage of the creative problem solving process, creative thinking facilitates the generation of new ideas (ideation) freely and without judgement. Ideation is often presented as a time-boxed activity, in which problem solvers widen the number of possible solutions to a problem. Once the ideation phase is complete, critical thinking is applied, facilitating the judgement (evaluation) of the relevance and value of the ideas generated in the ideation phase (see figure 1-1) to narrow down the number of viable solutions to the problem.

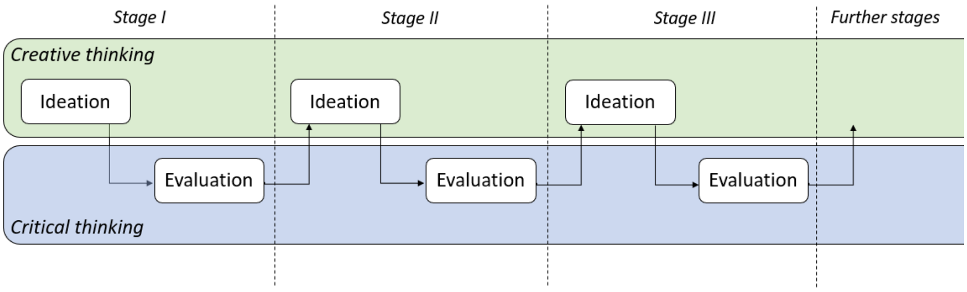


Figure 1-1: The creative problem solving process, made up of stages of ideation-evaluation. Ideation is facilitated by creative thinking, evaluation facilitated by critical thinking.

According to Shneiderman (2009), one of the founding fathers of Human Computer Interaction (HCI) research, supporting creativity is one of the field's grand challenges: "Understanding creative processes, especially when mediated by user-interfaces, will remain a continuing challenge." The significance of support for creativity lies in the process of ideation in each stage of creative problem solving.

Recognising the need to support creative problem solving, a large amount of psychological literature has focused on the ideation step of this process (e.g. Diedrich et al., 2015; Diehl, Munkes and Ziegler, 2002; Dennis and Valacich, 1993; Moss, Kotovsky and Cagan, 2007; Dugosh et al., 2000). In fact, ideation is one of the task types with the longest standing roles in experimental studies of creativity. The ideation task itself refers to tasks in which participants are asked to generate as many ideas as possible in response to a problem most often stated in the form of an open-ended question. Ideation tasks are often described as ill-defined problems lacking any specific goal or set of steps to reach this goal (Shah, Smith and Vargas-Hernandez, 2003; Chrysikou, 2006). It is therefore not uncommon when investigating support for ideation to look at how to support both divergent and convergent thinking, that is, thinking styles in which we might make associations to seemingly unrelated artefacts (e.g. method of analogy) to generate a breadth of novel ideas, and thinking styles in which we might focus on a specific topic or semantic space in order to generate, combine and hone in on specific features by exploring the depth of that specific space (e.g. Gabora, 2002; Sowden, Pringle and Gabora, 2015).

Ideation has been studied across a variety of fields: HCI, psychology, engineering and organisational literature. The most popular form of ideation has been identified as brainstorming (Osborn, 1957). Although this is not the only method of ideation, brainstorming is a well-studied phenomenon, both in groups as well as individual brainstorming. Nominal groups, in which people brainstorm separately and ideas are analysed as a pooled group, treated as if the individuals generated the ideas together, have often been shown to be more productive than in-situ group brainstorming (e.g. Taylor, Berry and Block, 1958; Diehl and Stroebe, 1987; Goldenberg, Larson Jr and Wiley, 2013). This focus on individual brainstorming has brought about studies on a range of technological tools to assist in brainstorming electronically: ideation support systems.

The type of factors tested in ideation support systems spans a wide range. There has been research on the concept of priming people to think divergently (Dennis, Minas and Bhagwatwar, 2013), as well as priming people with mood (Sowden and Dawson, 2011) as well as feelings of social belonging or separation (Ye and Robert Jr, 2017; Perry-Smith, 2006). Support for the creative process is not a new area of study in HCI, and from an HCI point of view, studies have been done that explore the effects of different tools on the productivity of ideation. These include, but are not limited to, interface designs that offer examples and keywords to prompt ideation (e.g. Diehl, Munkes and Ziegler, 2002; Chan et al., 2017), the effects of different styles of input, such as keyboards versus graphical pen and paper interfaces (Oviatt et al., 2012), on the fluency and appropriateness of ideas generated, as well as physical props such as inspiration cards (Golembewski and Selby, 2010).

Unfortunately, despite its long history as an object of scientific study, many aspects of the cognitive processes underpinning ideation are poorly understood. In more recent years, an emerging area in HCI focuses on the understanding of the concepts of *flow* (the free generation of ideas (Csikszentmihalyi, 1997)) and *impasse* (feelings of being stuck and unable to generate more ideas (MacGregor, Ormerod and Chronicle, 2001)), with a particular interest in understanding how the state of impasse arises and what can be done to mitigate the effects of it (e.g. Chan et al., 2017; Chan, Dang and Dow, 2016; Siangliulue et al., 2015). Indeed, as ideation is often presented with time-constraints, it makes sense that we would want to minimise the number of impasses in order increase productivity in the ideation phase. Experimenters in this research area mostly focus on the idea of giving prompts in the form of examples to ideators either when they are idle (not generating ideas), at set intervals, or simply when ideators request inspiration. Researchers in this area furthermore focus on the type of examples given (e.g. from expert or novice facilitators; prompts that are semantically far from the recently submitted ideas or semantically close) and the timing of when these examples should be offered.

A main model of thinking that is referred to in many of these studies is the concept of ideation being similar to Search in Associative Memory – a paradigm in recall studies in which people have been shown to strongly structure their recall in terms of categories (Raaijmakers and Shiffrin, 1981). Nijstad and Stroebe (2006) have developed a model of thinking, called Search for Ideas in Associative Memory (SIAM), in which they state that we generate clusters of semantically similar ideas in rapid succession, due to the way we synthesise these ideas from images in (faster) short-term memory. Ideas that are semantically different take longer to generate due to the nature of searching for other images in (slower) long-term memory. In our review of the literature, we briefly discuss this model and the evidence to suggest its validity. The theory itself closely aligns with the theories of information foraging (Pirolli and Card, 1999): when searching for information in a physical space, we move from patch to patch (e.g. books, internet, other people's knowledge) looking for information. The decision to give up on a patch in favour of another is dependent on a range of factors (such as time since the last successful retrieval of a useful item, number of useful items retrieved) and is widely studied in psychology literatures (e.g. Wilke et al., 2009; Payne, Duggan and Neth, 2007). If the search for ideas in memory involves moving from image to image in semantic memory, we would expect that these same rules that apply to information foraging would apply to ideation[1]. It is with this link in mind that we developed our first study, an ideation task in which individual participants generate a list of ideas in response to a problem. The study replicates that of Nijstad and Stroebe (2006) as we hope to further study the concept of categorical structuring of ideas, whether we see evidence that people cluster their ideas into categories, and the time-course within and between these clusters of ideas.

Our original motivation for the work presented in this thesis was to explore the time-course of ideas in an ideation task and how people allocate their time between clusters of semantically similar ideas. We were interested in these concepts with the aim of ultimately developing a stronger theoretical platform on which to develop and test approaches to ideation support - with specific focus on mitigating the concept of impasse. However, in the studies that followed this, we encountered a series of methodological problems. These seemed not to be particular to our studies but rather to suggest methodological difficulties in the literature at large. It seemed that fundamental issues in the empirical testing and measurement of ideation would need to be overcome before any theory based design could be attempted.

First and foremost, we were unable to replicate the findings in Nijstad and Stroebe (2006), in part due to the difficulty with which two independent coders found categorising the set of ideas generated in the study. Our analysis of the data showed that people generally do structure their ideas in semantic categories, however, the low inter-rater agreements coupled with the noisy nature of the category switches observed (not many sequential ideas falling into the same category; the data often followed an ABAB structure rather than

[1] This link has been alluded to in the above-mentioned psychological literatures in relation to accomplishment tasks that have a finite number of solutions. To the best of our knowledge, the link between foraging theories and ideation has not been explicitly stated yet in other research.

AABB) presented us with methodological problems. In a second study, in which ideators were asked to generate ideas in response to a series of sequential questions, we encountered a similar issue with inter-rater agreements. In this study, coders rated the novelty and value (two of the measures of quality) of the ideas. The lack of inter-rater agreement in both categorical coding in study 1 and quality coding in study 2 steered the work into additional investigations of methodology, somewhat removed from the initial motivations of this work.

1.1 Research questions

What started out as the intention to understand methods of supporting the cognitive state transitions in ideation and the categorisation of ideas generated ended up being a study on methodology. To report the research questions as they were formulated when the work for this thesis first took form would be misleading in terms of content and outcomes of the work presented. We were interested in understanding the categorical nature of idea generation, whether information foraging patch-leaving strategies could account for behaviour within a single and multiple ideation tasks. Our ultimate research question was whether we could inform interface styles that would support ideators in generation of more, higher quality ideas. After the failure to replicate results became apparent already in the first study, we decided to explore the topic from the point of view of the methodology. The research questions formed throughout the lifespan of this thesis were therefore:

RQ1: What influences the time-course pattern of ideas generated in ideation tasks?

This question relates to the overall time-course of ideas, the concepts of flow and impasse and whether we are able to see any patterns in the data itself with regards to these concepts. We analyse data from three different studies with the timings, number of ideas generated and properties of ideas themselves in mind. This question is further split into four measurable research objectives:

RO1: Assess whether semantic clustering of ideas takes place in continuous individual ideation on a single task.

RO2: Analyse whether any semantic clustering of ideas is reflected in the time-course of idea generation, and where possible relate these time-course effects to foraging theory heuristics for patch-leaving.

RO3: Test whether the simple foraging theory stopping rules, fixed items, fixed time, giving-up time or Green's rule can predict task-switching behaviour in multiple ideation tasks.

RO4: Investigate the quality as well as the quantity of ideas, by asking judges to rate ideas' novelty and value.

RO1 and **RO2** are addressed Chapter 3, in which we attempt to replicate the results of Nijstad and Stroebe (2006) to show that ideation is indeed semantically clustered. Finding longer times between clusters would suggest a patch like structure in the way we retrieve ideas from memory. **RO3** and **RO4** are addressed in Chapters 4 and 5, in which we perform an ideation study where participants generate ideas in response to multiple sequential questions. We analyse the times between ideas and compare these to the *giving-up times* (the time between last submitted idea and the decision to move on to another question) in order to respond to **RO3**. In addition to finding evidence of any foraging strategies people may use when deciding to switch, we investigate in Chapter 5 whether the quality of ideas might have an effect on this decision. The inherent problem with low inter-rater agreements in both Chapters 3 and 5 have led us to develop the following research question on support ideation. We hypothesise, based on evidence from Chapter 5, that better formulated ideas may in fact result in a higher judgement correlation between coders.

RQ2: Can we support ideation in such a way that judgement of ideas becomes more reliable?

This question relates to the general issues we encountered with disagreements between coders for the novelty and value judgement metrics of ideas, and also of the disagreement between coders for the semantic category judgements. We speculate whether specific qualities of individual ideas, such as length and specificity (detail), could have an impact on the idea judgements. In order to measure this, we have generated two research objectives:

RO5: Test whether priming ideators with examples or thematic roles improves the length of the ideas generated.

RO6: Observe whether the specificity of ideas affects the subjectivity of coding ideas or whether increased specificity improves inter-rater agreements.

Both of these are addressed in Chapter 6, in which we perform a simple prompting manipulation as a way of improving the well-formedness of ideas (**RO5**). Although we do not perform any further analyses on the specificity of ideas, the positive effect of the manipulation on the length of ideas allowed us to further explore the data in order to inform **RO6**.

1.2 Research methodology

In this thesis, we empirically explore the time-course of the process of coming up with ideas with the goal of gaining a stronger understanding of how we can support complex tasks such as ideation tasks. We furthermore attempt to answer the questions surrounding the methodological issues encountered. We therefore take a mixed-methods approach drawing mostly on quantitative techniques but also qualitative techniques.

We performed three ideation experiments (Lab-based: Chapter 3, Chapter 4; Online crowdsourcing: Chapter 6) in which we gathered quantitative data for timings and number of ideas. In our first and last study (Chapters 3 and 6) independent coders qualitatively analysed the data using a pre-determined set of categories, in order to code the data into categories for further analysis.

Additionally, in our first and last studies, we looked to independent coders to make judgements on the difference between adjacent ideas. In Chapters 5 and 6, we hired coders to make judgements on the data from the point of view of novelty and value – again rating qualitative data in order to facilitate further quantitative analyses. These results were further qualitatively coded by the experimenter to look for patterns in the language structure of individual ideas.

To understand further the concepts of specificity and generalisation of ideas (Chapter 6), we performed an online crowd-source experiment in which the same metrics (timings and number of ideas) were measured quantitatively. A sample of the data was coded qualitatively by 5 different coders across 3 coding schemes (categories, semantic difference scores between adjacent ideas, quality of the ideas). For each of the three ideation experiments we furthermore gathered responses on ideation strategies from participants using likert-scale questionnaires, reporting their level of feeling stuck.

1.2.1 Research ethics

The work done in this thesis involves asking human participants to engage in an ideation task, as well as asking human participants to independently code or rate sets of ideas generated. In order to ensure our studies were within the ethical bounds of the Department of Computer Science at the University of Bath, we followed the 13-point ethics checklist (see Appendix D.1) for each study we ran. In every study participants gave full informed consent and their data was stored safely and separately from consent forms so as to maintain the anonymity of participants.

1.3 Contributions

Current literature in HCI focuses on how to support ideation in a variety of ways. This includes provision of semantically near and far example ideas to support associative divergent thinking. Additionally, the timing of delivery of these examples is a topic studied in HCI as a method of understanding how to overcome the problem of impasse states. In this thesis we present three studies relating to ideation and the understanding of the time-course of ideation in relation to foraging theory (**RQ1**). We make these associations in order to observe whether classic stopping rules can account for the task-switching behaviour of ideators when presented with multiple ideation tasks. A stronger theoretical understanding of the cognitive processes that take place during ideation might highlight potential issues for the design of support tools. Our contributions to the theoretical understanding of cognitive processes taking place during ideation are:

- Replication and somewhat weak verification that generated ideas are clustered into semantic categories, and that the thinking time between category clusters is longer than thinking times of ideas that fall within category clusters.
- Distribution of *giving-up times* (the time since last submitted idea and the choice to move to a new question) (Chapter 4) displays a peak at low times and a long tail of very high giving-up times. This finding confirms that the decision to give up on an ideation task (accumulation style task) follows a similar pattern of sensitivity to the rate of generation plus the probability of switching upon sub-goal completion, as is found in other types of problem solving tasks in which there is a limited set of responses (accomplishment style task).

Our studies further found methodological issues in well-established methods of categorising ideas as well as judging the quality of ideas (through novelty and value). In response to these methodological issues we performed a series of investigations into the qualities of the ideas themselves and whether the variation in generalisation and specificity of ideas was to blame for these problems (**RQ2**). Our methodological contributions based on this area are:

- A simple instructional manipulation specifying the response format and thematic roles to include significantly increases the length of the ideas generated.
- An explicit instruction on the steps needed (specification of thematic roles) might provide higher value ideas, as is found in Chapter 6, as well as lower variation in value.
- Although no major difference was found in inter-rater agreements on categorisation

(low agreements), value (low agreements) or novelty (high agreements) between conditions with no manipulation and those with the thematic role manipulation, our findings open up the question of whether the visibility of a set of well-formulated ideas increased the novelty rating agreement levels between coders. This finding requires additional research.

In addition to the above contributions, we found further contributions resulting from our first and last studies:

- A more reliable method of measuring thought processes during ideation might be through analysis of the semantic difference between each idea. In study 1, we show that the thinking time between semantically close ideas is significantly shorter than the thinking times between semantically far ideas, loosely underpinning the SIAM model of thinking by showing that ideas with stronger mental associations are retrieved faster than those with weaker associations.

1.4 Thesis structure

The rest of this document is structured according to the work done chronologically.

Chapter 2 sets the scene for the rest of the work done in the thesis. We start by defining ideation, the main concept in this thesis, as a highly studied step in creative problem solving tasks. Key literatures on ideation support are reviewed leading us to the topics of flow and impasse. Through the Search for Ideas in Associative Memory (SIAM) model of thinking (Chapter 3), we link ideation behaviour to behaviour found in foraging theories, in which a set of variables are used to inform the decision to leave the present task. We then review the metrics and methods of evaluating ideas in ideation tasks in order to inform the methods used in Chapters 5 and 6.

Chapter 3 presents an experiment built on the theory covered in Chapter 2, in which we ask participants to continuously generate ideas in response to a single question (study performed for two separate questions). The data are analysed for patterns in the time-course of ideas, specifically informed by the categories that ideas fall into (**RO1** and **RO2**). We report a low inter-rater agreement on the independent category scores. Despite this, our findings indicate a categorical structure to thinking when generating ideas. In addition to the replication study, we offer a novel method of evaluating the relation between ideas and show that the semantic difference between ideas affects the time-course between ideas.

Chapter 4 presents an empirical experiment addressing the issues found in the noisy category switches in Chapter 3. We do this by asking participants to generate ideas in response to a sequential series of questions. In this study, we look for evidence of the use of foraging heuristics as a cue to moving to a new question (**RO3**). We rule out the use of the simplest foraging heuristics and show that switching might occur due to a combination of sensitivity to diminishing returns and the concept of sub-goal completion.

Chapter 5 reports follow-up analyses of the data gathered in Chapter 4, in order to assess the quality of ideas, either the novelty or value of an idea, and whether the variation in quality might play a part in the decision to switch (**RO4**). In this study, three coders rated the novelty and value on 10-point scales. Inter-rater agreement for both metrics was low. This result combined with the low agreement on categorisation in Chapter 3 guided the subsequent analysis, in which we performed a thematic analysis of highest agreed good ideas, highest agreed bad ideas, and highest disagreement ideas. The thematic analysis provided the foundations for the study following in Chapter 6.

Chapter 6 presents an empirical experiment in which we manipulate prompt type. We do so in order to affect the specificity with which participants formulate their ideas (**RO5**). We report a number of evaluations done on a sample of the data in order to verify whether the manipulation had an effect on the agreements between coders for all three coding types (categorisation, quality, and semantic differences) (**RO6**).

In **Chapter 7** we conclude the thesis, summarising the work done and the key findings from the studies. We identify key limitations in the work done in this thesis and offer directions for future work this might further our understanding of the cognitive processes taking place during an ideation task.

Chapter 2

Literature review

In the previous section we outlined our motivation as well as our initial overall goal of understanding the time-course of ideation (the creation of new ideas) in order to better inform ideation support tools, aimed at increasing the number and quality of ideas generated by diminishing the effects of an impasse occurring. In order to understand the time-course of ideation, we first look to the literature on creative problem solving in order to fully define ideation as a problem solving task. We then turn to literature on supporting ideation, both in the field of Human Computer Interaction as well as, more briefly, in other disciplines such as Psychology, Engineering and Organisational Management. In this chapter we will explore the literature that addresses the concept of ideation and current methods of supporting creativity and increased productivity. We explore the link between problem solving and theories of information foraging and propose that these theories might be a stronger method of understanding decision making in ideation. Furthermore, we review different methods of evaluating ideas.

In section 2.1 we discuss the concept of ideation, as a step in creative problem solving as well as the process with which we ideate. We review different types of problem solving task that vary in constraint and level of divergent and convergent thinking needed to solve the problem. We show that whilst ideation lends itself to the more creative divergent thinking type of task, the best solutions in ideation tasks are those that draw upon a combination of convergent and divergent thinking. Several methods of ideation are highlighted, with brainstorming seemingly being the most popular of these. We discuss electronic brainstorming as a way of supporting individual ideation; this is the method we will be using for the studies in this thesis.

Section 2.2 provides a brief overview of literatures that have studied the support of ideation through a variety of different methods such as tangible tools and novel interfaces. Studies in HCI have focused mainly on priming associative thinking and offering examples as a method of increasing productivity in ideation. In more recent years, through the aforementioned studies, the concept of impasse has emerged as being a key factor in ideation productivity. A few studies have since looked into how to predict an impasse and how to guide ideators back into a state of flow.

In section 2.3 we discuss the links made between foraging and search in memory and why this is relevant to the process of ideation. We review the key foraging theories and the studies in which modified versions of these foraging rules have been shown to apply to internal search tasks. We draw parallels between these internal search tasks and the notion that ideation is essentially a search for ideas in associative memory (SIAM model), suggesting that foraging rules may offer an understanding of the cognitive processes taking place during an ideation task.

In section 2.4 we review methods of judging the quality of ideas, highlighting the difference in metrics used, weights these metrics carry as well as what the term quality means in different types of task such as tests of creativity versus product design.

2.1 Ideation

2.1.1 What is ideation?

Ideation is simply put: the process of coming up with ideas. However, it is not easy to move toward a more precise definition. Ideation is sometimes described as the process of coming up with useful ideas that help the ideator reach a specific goal or outcome (Briggs and Reinig, 2010). Ideation tasks themselves can be described as ill-defined problems, in which the goal and the steps necessary to solve the problem are ambiguous (Chrysikou, 2006). Shah, Smith and Vargas-Hernandez (2003) highlight the fact that the specification of ideation tasks varies from task to task. In the famous 'unusual uses' or 'alternative uses' tests (Torrance, 1962; Guilford, 1967), the problem is typically expressed as 'think of as many alternative uses for a shoe', without any further constraints on the goal or method. In contrast to this, in design ideation the steps to achieve the goal may well be ill-defined but the problem and goal itself must adhere to a set of physical constraints and functional requirements in order to produce usable solutions that are fit for implementation.

This ambiguity in the definition of the goals of an ideation task has resulted in some disagreement over the years as to whether evaluation is and should be included in the ideation process (e.g. Peterson and Nemeth, 1996). For the research presented in this thesis, we share the views of Basadur (1995) and Dennett (2017): that ideation precedes evaluation and ideators are therefore advised to withhold any judgment or criticism of their own ideas and write them down no matter how outrageous they might feel the ideas are.

Ideation and problem solving

Creative problem solving can be divided into two steps: ideation and evaluation (Osborn, 1957; Basadur, Graen and Green, 1982). Ideation is the initial step taken towards solving a problem, which is performed without the imposition of judgement. The ideation process can engage either the individual, teams or a whole organisation in the process of coming up with new ideas. In organisational management, ideation follows the scoping phase of a project, but precedes any evaluation of the ideas generated (Briggs and Reinig, 2010). Creativity is the key factor at this stage and evaluation will not occur until initial ideation has completed. This is often achieved by setting a timer and allowing no criticism until the time is up. The evaluation stage looks at the ideas generated in the ideation stage of the problem solving task and judge their suitability as a solution to the problem and its constraints.

Problem solving tasks differ in scope, definition and solution requirements. On one end of the scale, we find well-defined tasks that tend towards a specific goal, single solution or set of solutions (Chrysikou, 2006). These tasks are usually more precisely scoped tasks that require either a single correct answer or set of answers. Intelligence tests are an example of these, as these require critical (convergent) thinking rather than creative (divergent) thinking. Another example of convergent thinking tasks are 'remote associates tasks' (Mednick, 1962), in which problem solvers are shown three words and asked what the fourth linking word is (e.g. for 'dream, break, and light' the linking word is 'day'). In contrast to single-solution tasks, other convergent thinking tasks include accumulation tasks such as finding words from a set of letters (anagram task) (Wilke et al., 2009; Payne, Duggan and Neth, 2007). Answers are easily judged as correct or incorrect and the problem solver can continue to enter words until either the list of possible solutions is exhausted or they fail to generate more words (cognitive failure). We will revisit these tasks in the Foraging Theories section, as accumulation tasks such as these, with large pre-determined sets of solutions, have been relatively widely studied in relation to linking foraging theories with the internal search for solutions (e.g. Harbison et al., 2009; Wilke et al., 2009; Payne, Duggan and Neth, 2007). At the other end of the scale we find open-ended, ill-defined tasks with few or no constraints on the solution space. An example of these tasks is Guilford's (1967) alternative uses task. In this task, problem solvers are asked to come up with a list of as many possible uses as they can for an everyday common item such as a brick or a shoe. These tasks are classified as divergent tasks, as the person generating these uses has to think broadly in order to generate a range of creative ideas, with barely any constraints.

Ideation tasks are often discussed as a key phase in creative problem solving, benefiting from creative, divergent thinking. Validity and correctness of answers in an ideation process is not a question of a simple yes or no answer but a matter of degree. However, it has to be noted that creative thinking, which facilitates ideation, does not always fall in the fully divergent thinking end of the scale of problem solving (Sowden, Pringle and Gabora, 2015). Yes, divergent thinking plays a big part in ideation tasks as the process to reach goals is often unconstrained. However, many ideation tasks will be presented together with some form of constraint that requires convergent thinking such as honing in on one feature in order form associations that could generate novel solutions within a specified design space (Shah, Smith and Vargas-Hernandez, 2003). Among these tasks we find design reasoning tasks, in which problem solvers are shown images of a specific design and asked to infer characteristics about that item. In Oviatt et al. (2012) an image of a chimpanzee foot and a human foot are shown and problem solvers are asked to think about the biological structure and evolution of these.

As a problem solving task, it is no surprise then that ideation is mentioned in conjunction with creative thinking but calls for the use of the combination of convergent and divergent thinking (Brophy, 2001; Runco and Acar, 2012; Kaufman and Sternberg, 2010; Cropley, 2006). Based on the fact that ideation is best solved using a combination of convergent and divergent thinking, we liken ideation to the general category of knowledge work, in which workers are not simply performing rote tasks with a well-defined solution space, nor are they necessarily applying creative thinking at all times. Creative ideation has been identified to be an important area of research as it is identified as being key to ensuring work-place effectiveness as well as competitive advantage over other organisations (Woodman, Sawyer and Griffin, 1993; Amabile, 1996; Basadur and Hausdorf, 1996).

To summarise, ideation tasks present themselves in different ways, such as open-ended questions, inferential reasoning tasks and unusual uses / alternative uses tasks. In the work presented in this thesis, we use open-ended questions as a method of representing ideation, specifically due to the inclusive nature of the types of questions, requiring little to no subject knowledge. Although ideation might often usefully be considered as the first or an early stage of a longer problem solving process, it is true to say that in many studies of ideation this stage is isolated, so that the task for ideators, as studied, is to generate ideas that are not further developed. This is the method adopted in this thesis. It might be criticised in that it takes ideation away from its broader context; but in addition to the standard defences of such an approach (i.e. that it allows careful measurement and analysis of ideation), it is worth noting that organisational structures quite often create a similar divide between ideation and the other problem solving processes that might follow it, for example in brainstorming meetings of project teams.

2.1.2 The process of ideation

The process of ideation itself, and of recording ideas, can be done in many different ways. One of the most popular methods of ideating is brainstorming, in which groups of people engage in the generative part of ideation, leaving the evaluative part until the end of the brainstorming session (Osborn, 1957).

Other methods include brain-writing (Geschka, Schaude and Schlicksupp, 1976), SCAMPER idea prompting method (Eberle, 2008), and mind-mapping (Buzan and Harrison, 2010). Brain-writing offers a group technique in which ideators write down their ideas individually to avoid the effects of production blocking and evaluation apprehension (see section 2.1.2 Electronic brainstorming as a method of ideation). Once an ideator feels stuck, they swap notes with another in order to keep elaborating and synthesising ideas. The SCAMPER idea prompting method makes use of 7 prompts (**S**ubstitute, **C**ombine, **A**dapt, **M**odify, **P**ut to another use, **E**liminate, **R**everse) in the form of action verbs in order to come up with new ideas, in particular in relation to product modification or new product design. Mind-mapping consists of a graphical representation of the connections between ideas generated by an individual, although this method can also be used in group brainstorming sessions to capture ideas according to different categories. This list of ideation methods is of course not exhaustive, Silverstein, Samuel and DeCarlo (2013) outline more than 50 methods that can be used for ideation, including analogy to existing systems, categorisation of already generated ideas as a method of synthesising solutions and De Bono's (2017) six thinking hats (a method of approaching a problem by thinking as different stakeholders would).

In the scientific and management literatures, ideation is most often discussed in relation to brainstorming, as brainstorming is one of the most popular methods of ideating. Traditional brainstorming takes place in groups, who verbally express ideas. The term brainstorming was coined by Osborn (1957), as a method of generating ideas over a fixed period of time without stopping to evaluate the ideas generated. Associated with this method he developed 4 constraints on the process itself:

1. Set a timer and do not apply criticism or evaluation on any idea until the time is up.
2. Generate as many ideas as possible as many ideas are conducive to combination and therefore more good ideas will be generated.

3. Generate 'wild' ideas, even if they seem irrelevant to the problem space.

4. Expand and elaborate on individual ideas, don't just see each idea as a separate item.

The concept of rules on brainstorming has at times been criticised for being too constraining on the creative process. Briggs and Reinig (2010) criticise the 'generate many ideas' rule as over-estimating the value of productivity (or *fluency*), arguing that there is a threshold at which good ideas are just no longer presented and the quality of ideas decreases. Moreover, Peterson and Nemeth (1996) argue that these rules may theoretically even prime creative thinkers into a thought pattern of 'conventionality', in particular the rule stating 'don't apply criticism', generating ideas that conform rather than offer novelty and value. Nonetheless these rules are employed in many successful brainstorming studies.

Electronic brainstorming as a method of ideation

Ideation in groups has been under a lot of scrutiny over the years as the concepts of evaluation apprehension, production blocking and free riding are seen as negative symptoms of verbal group ideation (Nijstad, Stroebe and Lodewijkx, 2003; Diehl and Stroebe, 1991, 1987). Evaluation apprehension is defined as the fear of being judged by others in the group and therefore refraining from voicing an idea, even though it might be a great idea. Production blocking occurs on the basis of a few scenarios; (1) if a group member feels their ideas are not as creative or valuable as others being generated, they will withhold those ideas; (2) the turn taking nature of group ideation means that at times ideators may come up with a few ideas before it is their turn and either forget or don't get the chance to fully formulate all of their ideas. The final problem with group ideation is free riding, a phenomenon seen in all types of group work, in which ideators decide to sit back and let others generate ideas while they do nothing.

A range of studies have shown that nominal group brainstorming produces more ideas (e.g. Taylor, Berry and Block, 1958; Diehl and Stroebe, 1987; Goldenberg, Larson Jr and Wiley, 2013). Note that these are measured on productivity of the ideators and not the quality of the ideas themselves. Nominal brainstorming groups are either individuals brainstorming without seeing anyone else in the group, or ideators generating ideas individually before discussing the ideas - the term "nominal group" refers to the pooling of such individuals' ideas. These findings were also found true for more constrained tasks such as anagram tasks (Kanekar and Rosenbaum, 1972). Girotra, Terwiesch and Ulrich (2010) found that *hybrid* group ideation was superior both in the productivity and quality of ideas. Hybrid group ideation is a method in which ideators first go off to generate their own ideas and then work together with the rest of the group once they've developed their own ideas.

As a method of supporting individual ideation, computer-mediated electronic brainstorming has been introduced. Dennis and Valacich (1993) showed that larger groups of 12 members that communicated electronically performed better than nominal groups of 12 members. For smaller groups of 6, the electronic communication condition and nominal group condition did not differ though. They propose that an electronic method of communication offered group process gains such as 'synergy and avoidance of redundant ideas' whilst avoiding the problems of regular face to face ideation (e.g. production blocking).

In summary, electronic brainstorming may be a method of supporting teams and avoiding the problems of face to face sessions. Kudrowitz and Wallace (2013) used individual brainstorming as a method of easier evaluation of individual ideators, to avoid the influence of other people's ideas on the assessment of quality of ideas of individual ideators.

2.2 Ideation Support Systems

As may be clear from the synthesis of ideation studies from sources such as Management, Psychology and Engineering, individual ideation is a fairly understudied phenomenon in Human Computer Interaction. Studies relating to ideation are mostly brainstorming studies that focus on group work and collaborative work, but little on the individual and computer mediated ideation (except in that nominal groups are made up of individuals - the emphasis in these studies is less on the individual's process than on the total productivity of a group of individuals). The work that looks at ideation and brainstorming for the most part focus on what factors help or hinder the ideation process.

Among the factors studied are the effects of social networks and social connection on ideation. Studies on how ideators see themselves in relation to the other members in a group ideation task have shown that feeling a part of a group (collectivism) increases overall fluency in comparison to groups that feel distanced from each other (e.g. Ye and Robert Jr, 2017; Perry-Smith, 2006). Similarly, positive mood has been shown to have an effect on ideation fluency. Sowden and Dawson (2011) performed a mood induction study in which participants watched funny, sad or neutral video clips prior to an ideation task as a method of priming their mood. They found that positive mood had a positive influence on the fluency but the mood conditions had no effect on the quality of ideas.

Efforts have been made to support ideation tasks in a variety of different ways. Recall that ideation and generating ideas comes from a synthesis of a person's own knowledge and forming associations in different ways. In Dennis, Minas and Bhagwatwar (2013) they offer a game to participants prior to the ideation task. The game asks them to link words presented on the screen together to make funny newspaper headlines. The participants who engaged in the fun game prior to the ideation study showed a higher fluency in ideation. The experimenters suggested this was due to the game priming the participants to form associations which in turn made ideation easier.

Other types of priming studied are input methods as a prime for the style of communication of ideas (Oviatt et al., 2012; Oviatt and Cohen, 2010). In these studies, they explored the effect of interface affordances on ideation and inheritance problem solving. They showed that, in a comparison of pencil and paper, digital pen and paper interface, pen tablet interface and graphical tablet interface that the digital pen and paper condition yielded the fastest problem solving. They furthermore compared ideation in pen interfaces and keyboard interfaces and found a higher level of linguistic content in keyboard interfaces, however the responses were not appropriate hypotheses, suffering from over-generalisations of science inferences. The digital pen condition afforded diagrammatic representation of solutions and were found to be overall more 'accurate' than the keyboard-based solutions for these types of task. The concept of free-form interfaces remains an understudied area in ideation, with few studies on the benefit of graphical versus textual inputs.

2.2.1 Ideation studies in Human Computer Interaction

Ideation in HCI is, similarly, often focused on brainstorming tasks in groups, such as participatory design tasks, and how to mediate these group sessions. Additionally, quite a lot of the work in ideation has looked at the development of new tools as a method of supporting creativity (Frich, Mose Biskjaer and Dalsgaard, 2018).

Quite a few of these tools are tangible tools, such as method cards (Wölfel and Merritt, 2013), designed to be used as a method of providing support to ideators and problem solvers. Golembewski and Selby's (2010) *Ideation decks* are a card-based tool that may help ideators reflect on the concept that is on the card and let it inspire their ideation process. A freeform spatial interface tool called *DataBoard* was developed by Kandogan et al. (2011), as a collaborative freeform tool for capturing ideas. The tool was found to have a positive influence on the ideators ability to map out a problem solving task, move elements around, as well as track and execute strategies, however, the board required quite a lot of effort and involvement and was found to detract from the tasks it was meant to assist. More pervasive support methods for group ideation in Human Computer Interaction include *Idea Expander* (Wang, Cosley and Fussell, 2010), a method which shows images to ideators based on their currently activated thinking patterns and concepts. This study is a proof of concept of this type of study where images are selected through wizard of oz method rather than selected through a developed system. The image is meant to prompt cognitive inspirations. The study showed higher productivity and quality of ideas. Andolina et al.'s

(2015) *InspirationWall* has a similar goal, in cognitively stimulating ideas to avoid slowing down of production of ideas. *InspirationWall* processes recognised expressions through speech recognition in order to present relevant key words on a screen, in the same semantic domain as the words being spoken. It does this in the background and brainstorming groups do not have to actively interact with the screen but can choose to look at the screen should they need inspiration for ideation. Unfortunately, although the *InspirationWall* was shown to have an effect on the overall pattern of ideation, it did not in fact increase productivity in the ideation tasks performed. Perhaps showing that although offering ideas, these might not be salient enough to offer any help to ideators in a group ideation setting.

As in other fields, the concept of priming has been studied in HCI. Bao et al.'s (2010) *Momentum* offers a method of group ideation in which members are first asked questions related to the words in the ideation task (e.g. "How to recruit adventurous tourists" will start the ideators off with the following prompts: "What is something that reminds you of tourists?", "With whom do you have an adventurous relationship?"). Ideators are asked to enter their responses by text and group them on a screen, visible to all ideators. All this is done prior to ideation and acts similarly to that of a prime, in that they are asked unexpected questions to encourage associations they would not normally make during the brainstorming session. Whilst the quality and productivity of the brainstorming session was not improved by the *Momentum* tool, they did find that ideators expressed a stronger awareness of the strategies they used in order to generate ideas.

The concept of offering examples and category labels for ideation as a way of inspiring ideas is not new (Diehl, Munkes and Ziegler, 2002; Dugosh et al., 2000; Nijstad, Stroebe and Lodewijkx, 2002; Baruah and Paulus, 2011). In other studies, it has been shown that ideas generated by others led to ideators generating ideas in a broader set of categories, showing more breadth of thinking. In electronic ideation tasks in which ideators are asked to generate as many responses as they can that could solve an open-ended, ill-defined question, examples have been used as prompts to make new associations and therefore continue to generate novel ideas. *Cheatstorm* is a phrase used by Faste et al. (2013) as a way of describing the inclusion of data from previous brainstorming sessions as input to assist in brainstorming. The ideas generated in an ideation task were found to be related to the ones in the *Cheatstorm* data that cued the brainstorming activity responses. The use of text stimuli during design activities has been shown to help novice ideators generate more novel ideas, although not more practical ideas (Goldschmidt and Sever, 2011). This ties in well with the concepts presented by Sosa and Dong (2013) who propose that an idea from a previous brainstorming session should not be dismissed, but rather used as a prompt in future brainstorming sessions to be elaborated and refined. Ideas that may have originally be seen as bad ideas, may in fact be good in the sense that access to them could inspire better ideas, making these 'bad' ideas facilitators of creativity.

In Chan, Dang and Dow's (2016) study they monitored ideation and had expert facilitators offer inspiration as a method of guiding productivity. Online ideators were offered the possibility of pressing the 'inspire me' button should they feel stuck. This study found that in conditions where experienced facilitators offered inspirations, ideator fluency and creativity increased. In conditions where novice facilitators offered inspirations to online ideators, they decreased the overall productivity and creativity. The investigators hypothesised that this is due to experts knowing how to analyse the train of thought that ideators are currently following and therefore offering ideas that are relevant and genuinely helpful to overcome a state of being stuck.

As can be seen from the previous examples, many studies on the methods of supporting ideation focus on the support of productivity (number of ideas generated) and quality of ideas, in particular in group ideation. These studies have a strong focus on tangible artefacts, methods of priming problem solvers before ideation and a subtle use of hints as a method of offering inspiration and examples during ideation. The following studies presented are from the more recent HCI literature. Although they focus on increasing productivity and quality of ideas, they do so by analysing what is hindering the creative process in an ideation task; in particular on the concept of stuckness, or impasse. These studies show a move towards a more empirical analysis of how problem solvers think and whether there are any specific cues that help or hinder ideators.

2.2.2 Flow versus impasse

Much of the research on ideation in Human Computer Interaction looks at methods for supporting both group and individual ideation. In the literature specifically on idea generation support methods, the idea of stuckness (or impasse) has emerged. Inspired by Csikszentmihalyi's (1997) *flow* state, the state in which ideas appear rapidly and freely to the ideators, a state of *impasse* (e.g. MacGregor, Ormerod and Chronicle, 2001) is synonymous with ideators losing their train of thought or having a moment of inability to come up with more new ideas[1]. Forming an understanding of what brings on this state of impasse may be beneficial to aid in the development of ideation support systems. In order to overcome this phenomenon, a few studies have estimated the occurrence of impasse states in a pragmatic way, by relying on ideators to self-report when they feel stuck (Chan et al., 2017; Siangliulue et al., 2015).

[1] Also referred to as a *failure* in research by Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003), and Nijstad and Stroebe (2006).

In Siangliulue et al. (2015), the authors focus on a key aspect of supporting ideation, namely *when* to intervene. Participants were asked to generate ideas for a period of 15 minutes, in which they would enter their ideas electronically into an online form, with a list of their own ideas populating a window down the side of the screen. The ideation task was an open-ended task that required no subject knowledge, just imagination: generating product ideas for a touch sensitive fabric that could render high resolution images and videos. Participants were assigned to one of four conditions: (1) on-idle condition, in which the application assumes that a 30 second idle time means the ideator is experiencing an impasse and offers example solutions; (2) on-demand condition in which the ideator had the option of pressing a button saying 'inspire me' in order to receive example solutions; (3) on-interval condition in which participants saw a new set of examples every 3 minutes; (4) no examples condition. The investigators found that in the on-demand condition, ideators generated more novel ideas, but not more ideas overall. Additionally, they found that the on-idle condition benefited by generating more ideas than in the other conditions. An interesting speculation is that although on-idle clearly prompted the continued production of ideas, the ideator was perhaps not stuck but busy formulating new ideas. Indeed, the threshold of 30 seconds for predicting an impasse might not have been a good threshold. They compare this to the longest idle times in the on-demand condition, showing that the first idle times before requesting examples was shorter at the start of the study but exceeded 1 minute later in the study.

In the above study, examples were offered to the ideator as a method of getting past a cognitive impasse. It has been shown that the use of examples in ideation tasks might have a negative effect on the ideation process as it might result in fixation on the topic of the examples and an inability to generate novel ideas (e.g. Agogué et al., 2014; Jansson and Smith, 1991). Chan et al. (2011) studied this phenomenon in relation to the type of examples shown and concluded that the problems of fixation could be overcome by presenting examples of ideas that are less common. Semantically distant ideas were also shown to help avoid fixation and promote ideation by analogy, in which ideators use unrelated objects as a method of forming novel and unusual connections. This study was followed up in order to find out whether, despite supporting the generation of novel ideas, semantically distant ideas would hinder the cognitive state in ideation, specifically in relation to impasse states (Chan et al., 2017). This study takes a step into cognitive theories of thinking, in particular it looks at that of Search for Ideas in Associative Memory (SIAM), a concept we will review in more detail in the next section. The SIAM model predicts that we generate ideas within categories, in such a way that we will continue generating ideas until we can no longer think of any more within that category. That stage is likened to an impasse, in which we search for a new category to generate ideas within.

In Chan et al.'s (2017) study, they asked participants to generate ideas electronically for 8 minutes in response to an open-ended question about themed weddings. The task itself was semi structured, asking them to write a theme and a prop before typing out their idea. The theme and prop fields were used by their system to perform simple real-time analyses of the semantic category in which the participant was ideating. Participants were split into 5 different conditions. For the purpose of this thesis we are mostly interested in the two 'match'-'mismatch' conditions. In both of these conditions, participants received examples at regular intervals (although, it is not specified what these intervals are in the study). Additionally, participants were offered the button 'give me inspiration', as was used in their earlier study (Siangliulue et al., 2015), in order to request more inspiration on demand. The experimenters used this method as a way of assuming ideators were in an impasse state. This impasse state would remain until the participant submitted another idea, in which the participant was assumed to have re-entered flow state. They report confidence in this method of estimating impasse states as they had found in their previous study that participants were able to notice when they were stuck and through self-reporting, stated that their motivation for pressing the button really was due to feeling stuck. As an additional measure, the experimenters verified this by running a short study in which they used brain sensors (functional near-infrared spectroscopy) which uses changes in blood oxygen concentration to infer brain activity changes.

In the match condition, participants were shown semantically close ideas at intervals (ideas that were very related to the current ideas being generated) within *flow* states and semantically far ideas whenever the ideator pressed 'give me inspiration'. It was hypothesised that this scheme supported ideation in the category the participant was already ideating within on *flow* states, and offering new categories to pursue upon *impasse* states. In the mismatch condition, participants were shown semantically far ideas when in the *flow* state and semantically close ideas when in the *impasse* state. It was hypothesised that the mismatch state would block productivity in such a way that participants would generate fewer ideas. The results were consistent with this hypotheses, showing that the mismatch condition blocked productivity and left ideators asking for inspiration more often. The results however did not show that providing ideas such as in the match condition has any benefit over a control condition where no examples were shown.

In this section we reviewed the literature on ideation support systems, in particular in relation to methods of supporting ideators in both groups and individually. We found that many ideation support studies focus on the invention of tangible tools. In more recent years, the concept of ideation and how to support it from a theoretical standpoint have come into focus. The concept of impasse has emerged from these studies as a hindrance to the creative problem solving process. In the next section, we look to the theory of Search for Ideas in Associative Memory and its link to foraging theories in order to answer the question of whether ideation might be better informed by theory.

2.3 Relations of ideation with Foraging Theory

In the previous section we looked at ideation support systems and how experimenters in Engineering, Management and Human Computer Interaction (HCI) have been building tools to support ideation. We encountered the concepts of flow and impasse as cognitive states of ideation. A few recent studies in HCI have studied these states; one of these studies linked the concept of impasses to the Search for Ideas in Associative Memory (SIAM). In this section, we cover this theory in more detail, linking it to foraging theories, and in particular the rules for patch leaving that have been identified by researchers in Optimal Foraging Theory and Information Foraging.

There has been some Psychology research in the field of time allocation in problem solving for a number of years, although this remains quite an understudied phenomenon. The studies performed look in depth at ways to understand why people choose to switch away from a particular problem they are working on. Some strategies and theoretical models are used to explain switching decisions made by foragers in nature. These are often compared to human external (physical) or internal (cognitive) switching. Reasons for giving up on a task are explored in the Psychology literature, in particular in terms of highly constrained tasks where problem solvers work through sub-tasks or steps toward a clearly defined goal. Many of these models have been tested in laboratory settings and are often validated with the use of highly specified problem sets and goals. This section aims to look at these models for abandoning tasks in more detail.

2.3.1 The Search for Ideas in Associative Memory

In associative theory (Mednick, 1962), there are three main ways of coming up with creative solutions: serendipity, in which mental images (frames) that help form novel ideas appear almost by chance; similarity, in which seemingly different elements can be associated in order to formulate new ideas; mediation, in which common elements are used in order to form new associations for new ideas. It is in the interest of the creative problem solver in this case to make far associations, for example in the way of analogies, in order to come up with increasingly more creative solutions.

The concept of similarity as an idea-generator is considerably elaborated in a cognitive model of idea generation developed by Nijstad and Stroebe (2006): the Search for Ideas in Associative Memory (SIAM). This model suggests that ideation tasks are a result of repeated searches in memory (knowledge activation) in order to generate ideas. SIAM is used to explain the process followed when generating ideas, similar to, and derived from, the Search of Associative Memory model (SAM) (Raaijmakers and Shiffrin, 1981) used for memory retrieval in free recall tasks. In free recall tasks you might expect a person to be asked to recall as many animals as possible. In a task like this, SAM proposes that people will group their responses in specific categories, e.g. animals with fur, ears, or 4 legs. Indeed, empirically, people do have a strong tendency to categorise their responses. The SIAM model proposes that the search for ideas works in a similar way by following a 2-step process:

1. Knowledge activation - *images*[2] are retrieved from long term memory.
2. The *features* of the retrieved image are used to make associations and generate ideas.

These steps make the assumption that we have two memory systems: long term memory (in which images are retrieved) and short-term memory (in which images are temporarily stored and used to make associations). Images are seen as 'knowledge structures' which have a central theme, such as 'hotel', around which associations are made, such as 'lobby', 'check-in', etc.

Through a series of studies focusing on individual and group brainstorming, Nijstad and Stroebe (2006) showed that people really do tend to generate semantically similar ideas in clusters. In these studies, participants were allowed 20 minutes to generate as many ideas to a single open ended question (e.g. "What can people do to preserve the environment?"). The resulting ideas were coded into multiple categories (for this particular question, a predetermined set of categories developed by Diehl (1991)) and performed an Adjusted Ratio of Clustering (ARC) analysis on the results. Their findings indicate a high, beyond chance, level of clustering, suggesting a categorical structure to thinking when generating ideas. Interestingly, this model supports quite a lot of theories across the creativity literature. Sternberg (1998) states that people do not blindly activate and deactivate frames in their minds when problem solving in the domain of creativity. Rather, they apply formal (domain knowledge) and informal (social knowledge) to the frames they activate in such a way that one idea will lead to another in a 'contagious' way, where one idea may lead to another by association. The process of generating ideas in a creative space is likened to bringing up existing knowledge and trying to form new knowledge structures within that conceptual space (Koestler, 1964; Boden, 2004). Rietzschel, Nijstad and Stroebe

[2] Not to be confused with visuals, images are simply themes or topics that can subsequently possess a number of features or associations.

(2007) showed through priming participants to specific categories that ideators generate more ideas within the categories they have been primed to and that a strong within-category fluency, that is generating ideas within a single category, led to more novel and unusual ideas. Recall that in associative theory (Mednick, 1962), the three main ways of generating creative solutions involve forming associations, either by luck, through analogy with common elements or analogy with similar elements. Although often represented as being at odds with each other (associative theory states that generating ideas across several categories yields a greater number of novel ideas, whereas SIAM states that generating ideas within the same categories leads to stronger formulation of many good ideas) the concept of making associations with images is common to both theories.

The findings in Nijstad, Stroebe and Lodewijkx's (2002) studies indicate that ideas generated are semantically organised into categories, and that these are temporally clustered in such a way that the time between ideas within a semantic category is shorter than the time between categories. They explain this through the concept of failures. As an ideator reaches a failure, or impasse, the perceived exhaustion of ideas generated from an image leads the ideator to give up and activate another image from long term memory, in order to generate more ideas. In a series of papers by the aforementioned authors, a failure is defined as either the inability to generate *any* ideas or the inability to generate *new* ideas (i.e. thinking of the same ideas repeatedly). The differences in the timings of between category items and within category items is hypothesised to be the time it takes to activate a new image. It is therefore not surprising that their findings confirm a higher number of ideas generated by ideators whose ideas show high levels of clustering.

Although not explicitly discussed in Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003), or Nijstad and Stroebe (2006), the SIAM model and time-course across ideas when applying this model resembles the concepts of patches and patch leaving rules described in foraging theory. In the next section, we are going to be covering the literature on foraging theory, the basic giving up time rules and how these apply to the SIAM model, as well as how they might be used to better inform the concepts of flow and impasse.

2.3.2 Foraging Theory

Most of the theoretical work on why agents give up on tasks or switch between tasks is informed by foraging theory, where patch leaving has been one of the very central constructs. This section begins by looking at the patch leaving literature, to highlight how foraging theory and theories surrounding human information foraging have come about.

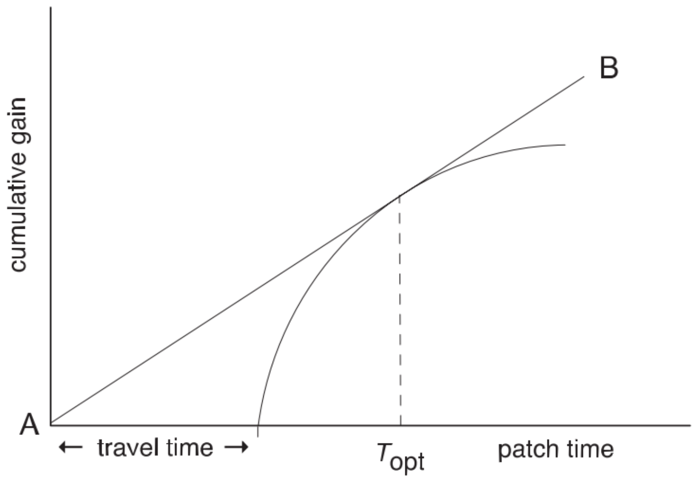
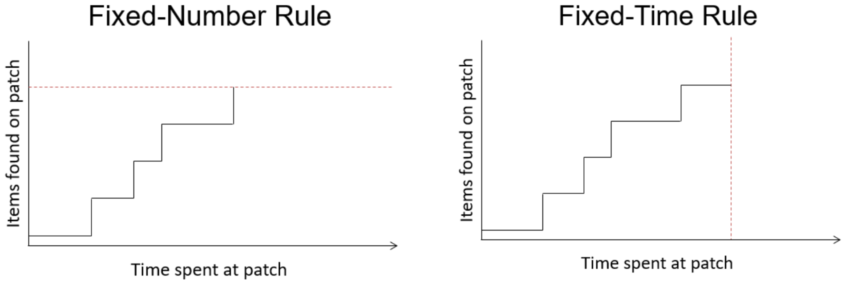


Figure 2-1: Marginal value theorem (adapted from Charnov et al. (1976)) (Wilke et al., 2009)

Information foraging literature from Behavioural Sciences makes an analogy between human and animal foraging behaviour (Wilke et al., 2009; Hills, Jones and Todd, 2012). Patches in nature have limited, and varied, resources. As foraging animals search for resources in a patch, they must decide on the optimal time to leave for another patch, in order to maintain a high overall level of gain and rate of gain. There is a trade-off in this decision: at what time does it become more valuable to switch to a new source (patch) rather than stay and continue foraging in the current patch, keeping in mind the fact that there will be a travel time (a cost) to the next patch. To make an analogy to human (external) information foraging: when we look for information (the resource), we will look in a range of patches, such as physical files, electronic filing systems, search online, ask someone, etc. Variable amounts of information can be found in each of these patches and switching between them can take time. The information seeker, as with the foraging animal, must therefore determine when the value of switching to a new source is higher than trying to find more information at the current source (Pirolli and Card, 1999; Sandstrom, 1994).

When determining what strategy animal foragers use, an often-mentioned strategy is Charnov et al.'s (1976) *Marginal Value Theorem (MVT)*. This optimal foraging theory states that when the momentary (or *marginal*) rate of return of a resource at the current patch is equal to the mean return rate for all foraging patches met so far, the patch should be left in search of another unexplored patch. The rate of return and the time it takes to travel to the next patch dictates the optimal time to leave the current patch (see figure 2-1).



(a)

(b)

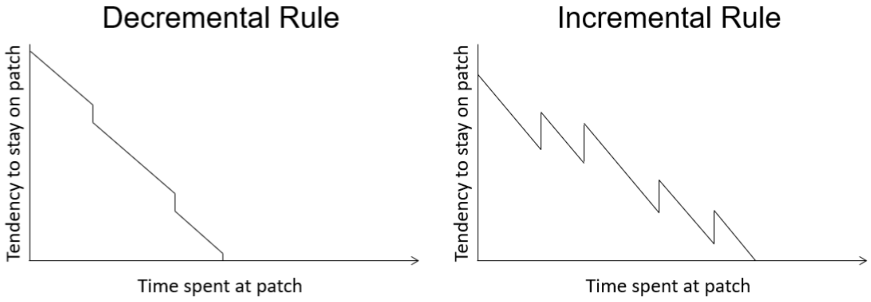
Figure 2-2: Pattern of simplistic patch leaving behaviour theories. (a) Fixed-Number rule (b) Fixed-Time rule. A vertical jump on both graphs indicate an item found. The patch is left at the threshold time or threshold number of items found (denoted by the red line), regardless of patch potential.

Unfortunately, Charnov et al.'s (1976) *MVT* presents a limitation: current, instantaneous rate of return must be tracked in addition to the average rate of return so far across all patches. The theorem is successful empirically but is better thought of as a theory of what animals are computing than of how they are computing it. It seems implausible that animals employ this level of cognitive capacities during a simple foraging task. To this effect, behavioural biologists have come up with a range of simplistic heuristic theories about how animals behave, which might approximate the marginal value theorem, given certain assumptions about the nature and distribution of patches. These heuristics have influenced work on human giving-up rules and are therefore briefly reviewed here.

Simple foraging heuristics

The simplest giving-up heuristics include fixed-number and fixed-time strategies (Iwasa, Higashi and Yamamura, 1981; Stephens and Krebs, 1986) (see figure 2-2). In the fixed-number strategy, animals will leave a patch regardless of its quality (number of resources available) after a predetermined number of resources have been found. In the fixed-time strategy, animals will leave a patch after a predetermined time has passed. These rules have very low cognitive demands but suffer from being insensitive to the actual number of resources on a patch; nevertheless they might be effective in certain environments, for example when all patches have very similar diminishing returns gain curves that an animal knows in advance.

More complex rules, using number of items encountered as the information cue are the incremental or decremental rules (Waage, 1979). In the incremental rule (see figure 2-3b), the tendency to stay on a patch is a function of the time taken to successfully find a



(a)

(b)

Figure 2-3: Time pattern of patch leaving behaviour theories using items as an indicator of tendency to stay on patch. (a) Decremental rule (b) incremental rule. A vertical jump (upwards in incremental rule, downwards in decremental rule) indicate an item found.

resource. As a resource is found, the tendency to stay increases a small amount. As the time between successful finds increases, theoretically due to the number of resources left on the patch, the animal eventually gives up and moves to another patch. Similarly, the decremental rule is dependent on the time taken to successfully find a resource (see figure 2-3a), however inversely dependent on resources found. Every resource found decreases the tendency to stay on the patch; whilst a logical rule, this does not take into account the variation of patches. In the decremental rule, a highly abundant patch might be left before it has been fully foraged.

Adaptive foraging heuristics

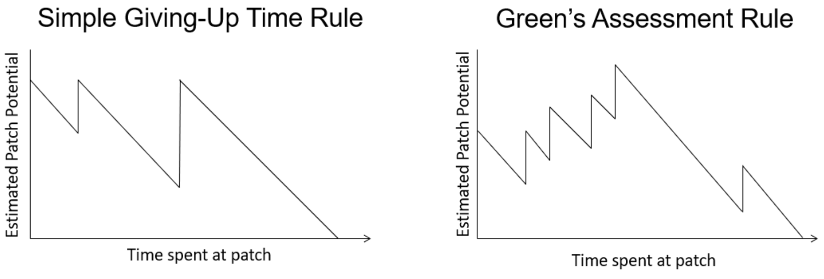
The aforementioned rules suffer the problem of being too simplistic, especially when thinking about patches that may vary in potential gain. In both human and animal foraging, the potential (or quality) of a patch is often unknown and variable, leaving the simple heuristics to fall short as suitable patch leaving rules.

More adaptive heuristics, best used in situations where the quality of a patch is unknown or variable, are *Green's rule* (Green, 1984) and *giving-up time* rules (Stephens and Krebs, 1986). Green's rule states that animals may be estimating the potential of a patch by increasing their perception of the potential every time a resource is successfully found (see figure 2-4b). The potential decreases as a function of time and successfully finding an item results in a vertical jump in potential estimation. If no item is found for a while, the patch potential estimation will continue to drop as a function of time until it drops below a pre-set threshold, after which the patch is abandoned. As such, Green's rule predicts that foragers will have longer visit times at richer patches. In Green's rule the length of a visit is determined by $V = T + IG$ where:

$V =$ visit time at a patch,
$T =$ minimum visit time (e.g. if no items are found),
$I\ =$ number of items found during the visit,
$G =$ gain in visit time for each item found.

Conversely, giving-up time is the time between the last resource found and the decision to give up. A prior threshold of patience in terms of time is set for the patch. The number of successful items found does not have an effect on this, nor does the total time on the patch. This heuristic relies solely on the time taken since the last successful item found. If this time increases beyond a set threshold, then the patch is abandoned.

As animal foraging has been likened to external information foraging by humans, external human foraging has often been likened to that of human internal foraging (e.g. Wilke et al., 2009; Hills, Jones and Todd, 2012; Payne, Duggan and Neth, 2007). The distinction made here is that human external foraging follows the same rules as animal foraging, where patches of resources are external to the forager and they physically have to move between the patches. Internal foraging relates to free recall, where a problem solver searches their working memory or long-term memory for information stored. Again, it is interesting to find that there is no explicit comparison of SIAM to the theories of foraging. However, looking at the categorical generation of ideas from patch-like images, it is easy to assume that foraging theories might be applied as a method of understanding ideation tasks. In particular, we suggest here that flow can be seen as successful patch foraging whereas impasse can be likened to running out of resources and moving to another patch.



(a) (b)

Figure 2-4: Time pattern of patch leaving behaviour theories (a) giving-up time rule and (b) Green's rule. A vertical jump on both graphs indicate an item found.

The next section gives an overview of a few studies comparing internal and external foraging rules. These studies introduce slight modifications to the rules reviewed above as a result of comparing these rules to the behaviour of problem solvers in tasks such as anagram tasks, showing that these theories are as valid for internal information foraging as they are for external foraging.

2.3.3 Internal memory search and giving-up rules

Much of the research looking at internal memory search and giving-up times (e.g. Dougherty and Harbison, 2007; Harbison et al., 2009; Payne, Duggan and Neth, 2007), refers to foraging theory to try to explain the same phenomenon: when given the choice, what makes people stop trying to complete a task or sub-tasks or retrieve information from memory (internal search) and what factors affect problem solvers decision to abandon one task in favour of another.

Wilke et al. (2009) ran a series of studies to show the difference in task abandonment between external (physical) and internal (cognitive) searches. They specifically looked for the use of any of the foraging heuristics as indicators of the patch leaving rules people use when abandoning tasks, and whether these rules were different in the two different tasks. Their first task included an external (electronic) physical search; finding fish in a pond. This was set up in a virtual landscape with fish appearing at random intervals. When moving from one pond to the next, extra time was added to mimic the time taken to walk from one pond to the other, also known as a switch cost. Their second comparison study included an internal search task, presenting problem solvers with a basic anagram task, a task in which participants have to form as many words as possible from a string of letters. An important observation they made was that when performing the external search, there was a tendency to switch immediately after an item was found, thus the giving-up times were very low. This suggested that in the external task, the principle of 'just one more' was applied, a form of sub-goal completion if you will. Surprisingly, they did not find this in internal searches; these offered much longer giving-up times and rather than switching upon finding a word, problem solvers switched after long idle times. They speculate that this could be due to an internal finding potentially giving rise to more 'ideas' and therefore an aversion to make an immediate switch. In contrast, they speculate that an external forager may be aware that they are close to depleting the patch and therefore doesn't want to waste any more time.

Dougherty and Harbison (2007) looked at internal search in a free recall word memorisation task in detail to explain what they call *exit latency* i.e. the time the participant was willing to keep trying to retrieve words from memory after their last successful retrieval before giving up. In foraging theory this term is better known as *giving-up time*. The study showed that for more difficult lists of low-frequency words, the decisiveness scores (an individual difference measured by a questionnaire) and exit latency were correlated; that is, higher decisiveness meant a decrease in exit latency. They propose that exit latency is a product of the motivation of the person performing the task and the difficulty of the task. They relate their findings to a more classic stopping rule theory by Anderson and Milson (1989) called the *PG-C* model. The *PG-C* model incorporates *P*, the probability that another item will be retrieved; *G*, the gain or value associated with retrieving this new item and *C*, the cost of finding this item. Following this model, a person is able to make a prediction on whether it is worth continuing to search for more items; as *PG-C* approaches zero, that is, the cost of finding a new item increases and the probability and gain of finding a new item decrease, this may act as a mental cue for the person to abandon their search and thus the task. In a follow up study by Harbison et al. (2009), they showed that the number of items retrieved in an internal memory search had an increasing effect on time spent on a task and a decreasing effect on the exit latency. Whilst it seems self-evident that more items result in increased time spent on a problem, it is interesting to see that more items found leads to the perception that the probability of finding more items is lowered, thus the patience threshold after the final item retrieved is lowered too.

An interesting observation in task abandonment is this concept of sub-goal completion. In a study by Payne, Duggan and Neth (2007), participants were given two anagram tasks varying in difficulty (easy and hard), and asked to make their own judgement on when to switch between these tasks. The experimenters were interested in understanding the underlying strategies that guide the decision to switch away from a task, and whether these could be attributed to any or multiple of the foraging giving-up heuristics. The experimenters found that more time was allocated to the easy task than to the hard task. Whilst this finding shows adaptive allocation of time across the tasks, it also rules out the two simplest foraging rules: a simple fixed-time rule would have resulted in the same amount of time being spent on both tasks; the simple fixed-number rule would have resulted in more time being spent on the hard, less productive task, in order to be able to generate the same amount of items as in the easy task. The experimenters further observed that giving-up times were longer for the more difficult task and that some between-item times were longer than the giving-up times. These findings rule out the use of a simple giving-up time heuristic. A much better fit to participants' patch-leaving decisions was offered by Green's rule, which essentially computes (in a way that makes few demands on working memory) the rate of gain since a patch was entered (since a task was started) and quits when this rate falls below a threshold.

Analysing the giving-up times more closely, Payne, Duggan and Neth (2007) found that quite a high number of task switches took place immediately after a word was found, behaviour which Green's rule allows, but not to the extent that it was observed in the data. They attributed this behaviour to sub-goal completion, modelling task abandonment behaviour on this task by extending Green's rule to include the probability that a switch would be made upon sub-goal completion. In summary, they found that giving-up decisions were best thought of as having two probabilistic causes acting in concert: a rate-based decision, in which quitting happened when rate of gain fell below some threshold ($V = T+IG$), and a success-based decision, in which quitting might happen immediately (though with low probability) after any local success.

Recall from section 2.1.1 that problem solving tasks such as anagram tasks differ from ideation. Indeed, you would expect these to follow the same pattern as an ideation task as they draw on a combination of image activation and association, as well as offering an accumulative solution space. However, they differ from ideation as both the process and goal is more specified, e.g. generate words from these letters in an anagram task will have a distinctly correct set of responses, immediately perceived by the problem solver.

As a summary, our interest in the topics of foraging theories lie in trying to understand the underlying processes that take place during an ideation task, specifically in relation to the decision to switch to a new patch in semantic memory in a single ideation task, as well as the decision to switch from question to question in sequential ideation tasks.

2.4 Metrics and Measurements in Ideation Support Systems

The measurement of quality varies and researchers in this area often come up with their own metrics or suggest new ways of interpreting the metrics (Dean et al., 2006). Historically, in creativity tasks, the quality of ideas is thought of as the quality of the entire set of ideas generated within an ideation task, rather than individual ideas, in such a way that the quality assesses the creativity of the person who generated the ideas. In Guilford's (1967) alternative uses tasks, the entire idea-set is scored on four components to judge the level of creative divergent thinking of the ideator: originality, fluency, flexibility and elaboration. Each component has a strong associated scale, such that there is little ambiguity when it comes to rating these. Originality is rated by comparison to all people who gave the test, and responses given by only 5% or 1% of people are scored points. Fluency is simply the total number of responses. Flexibility depends on categorisation of the ideas, such that, if alternative uses for a newspaper were to 'wear it as a top' and 'wear it as a dress', these would fall in the same category. The total number of categories is counted and used in the measurement of creativity. Finally, elaboration is scored, giving 'clothing' 0 points and 'a hat you can wear on your head when it is raining' 2 points for elaboration of type of clothing and detail about the circumstances.

In both Torrance (1962) and Guilford's (1967) tests, the metrics are used as a method of judging creativity only. This method is however not applicable in all settings. In design studies and in ideas generated in order to reach a certain goal state in an organisational setting, the quality of ideas takes on a different meaning. Novelty is desirable but feasibility and usefulness play a big part in the overall quality and suitability of the idea (e.g. Briggs and Reinig, 2010; Reinig, Briggs and Nunamaker, 2007; Howard-Jones and Murray, 2003). Additionally, in organisations specifically, the number of ideas generated is often irrelevant, only the quality of the ideas themselves. This is due to the fact that the purpose of ideation in these settings is to generate ideas towards the inherent goals of the organisation and therefore management are not interested in sub-par ideas but ideas that are actually going to benefit the company and put them in a strategically competitive position. Bounded ideation theory challenges the notion set forward by Osborn (1957), that the quantity of ideas increases the quality of ideas. Instead, bounded ideation theory proposes a model in which the ratio of good ideas to total ideas depends on a variety of factors such as understanding of the task, goal congruence, ability of the ideator and openness of the solution space, as well as the cognitive state of the ideator (mental exhaustion and attentional resources). Reinig, Briggs and Nunamaker (2007) fully change the metrics in order to judge ideas individually rather than as a set. They consider instead the economical, technical and political feasibility to the responses for the question 'What can be done to resolve the problems of the school of business?' on a scale of 1-4.

In the majority of the studies we previously reviewed in relation to ideation, experimenters hire two independent coders to rate ideas on a variety of scales (5, 7, and 10-point scales are common) (Ye and Robert Jr, 2017; Shah, Smith and Vargas-Hernandez, 2003; Dennis et al., 1999; De Dreu, Baas and Nijstad, 2008; Diehl and Stroebe, 1987). They do this in order to generate two sets of independent scores, that can be compared to verify that the coders agree on these scores. The independent coders will usually rate novelty and value metrics, whilst the experimenter can simply count the number of ideas to get the fluency metric. In each of these studies, they make use of one to three quality metrics only, from the following list of most common measures of quality of ideas:

- Novelty[3]: how original and surprising an idea is

[3] Note the split between P-novel (never been thought of by the person before) vs. H-novel (never been thought of by anyone before) - (Boden, 2004)

- Value: how useful and practical the idea is, whether it makes sense as a solution to the problem
- Fluency: how many ideas are generated, also called productivity in a few studies
- Elaboration/Specificity: how well formulated is the idea

Quality can further be used to estimate the quality of the whole set of ideas. Whilst mean quality (of all ideas generated by a single person) might seem a plausible way of measuring ideation quality, this does not work as it does not take into account the amount of ideas generated. If one person has come up with one great idea and another has 4 great but one bad, the first person will have a higher mean quality rating. The quality of good ideas paradigm tries to overcome this issue by adding up the quality of ideas that have a quality rating of (median point of their scale) 3 or more. This aims to reward all ideators for good ideas and not bad (Diehl and Stroebe, 1987; Dennis et al., 1999).

The concept of generating a composite quality score has been criticised by Shah, Smith and Vargas-Hernandez (2003) as being nonsensical. The novelty of an idea and the value of an idea are not measured on a similar scale - it might therefore be more advantageous to look at the relationship between novelty and variety in a set of ideas rather than build a composite score of those two metrics. For the work done in this thesis, we make use of the metrics novelty and value as methods of measuring ideation quality, due to their popularity as measurements of this in the ideation literature.

2.5 Chapter Summary

In this chapter we reviewed the literature on ideation, situating ideation tasks in the space of creative problem solving tasks that are usually ill-defined and loosely constrained (at least in the HCI literature on this topic). This topic has been identified as being an important challenge as we perform creative ideation tasks in our everyday lives. In particular, organisations have identified the need for creative problem solving as a way of gaining competitive advantage. Ideation makes use of convergent and divergent thinking, activating frames and making associations in order to generate novel ways of solving a problem. As a popular method of ideation, if not the most popular, we looked at brainstorming and studies involving individual electronic brainstorming.

In recent years, literature on understanding the cognitive processes that take place during ideation has started looking to the concept of impasse in order to develop novel mechanisms in which we can assist ideators, and support them through the feeling of being stuck. These studies relate the feeling of being stuck to the exhaustion of an image or category in semantic memory from which ideators generate their ideas. Surprisingly, none of the studies explicitly liken these cognitive processes to foraging theory. The notion that people search for ideas in semantic memory and choose to retrieve a new image in long term memory when they run out of ideas seems almost intuitively related to the concept of patch foraging and the decision to switch patches when potential decreases. These internal search patterns form the basis of our first study, where we try to replicate SIAM's main model of thinking: that people generate ideas within categories, and that production of ideas within categories occurs at a faster rate than between categories.

Finally, we reviewed the methods of judging the quality of ideas, including the methodology and metrics used in such an analysis. We make use of these methods in our second study, in which we attempt to explain the time-pattern of switching between tasks through the variation in quality of ideas generated. We make use of these methods again in our third study, in which we perform a simple manipulation on ideation prompting in order to see whether there are any effects on the agreement scores between coders for quality ratings and category scoring.

In the next chapter we present our first study, based on the study developed by Nijstad, Stroebe and Lodewijkx (2002). We perform this study as an exploratory approach to understanding the concept of impasse, as a function of switching between categories (SIAM model), and if there are any ways we can account for the occurrence of impasse states using any of the foraging theories.

Chapter 3

Time-course of single ideation tasks

This chapter focuses on the time it takes to come up with ideas on a single problem or task, and whether any patterns can be found in the time-course of coming up with ideas. The studies presented in this chapter attempt to replicate the findings of Nijstad and Stroebe (2006), suggesting that idea generation relies on associative semantic memory through showing a clear category structure of ideas, with the goal to find a link between the concepts of *flow* and *impasse*, Search for Ideas in Associative Memory (SIAM) and the notion that the search for ideas can be likened to that of external foraging for ideas. Looking for patterns in the time taken to generate ideas, and specifically the time *between* groups of related ideas compared with time *within* groups of related ideas, may help us understand whether people generating ideas are applying foraging theory-like stopping rules on a per-category basis.

In the following section we briefly revisit the relevant theories in literature that are reviewed in Chapter 2 and which motivate this study.

3.1 Chapter background

An important idea in the HCI literature on ideation support is the notion that we can support ideator productivity, especially in light of the fact that ideation is often a timebound process. A recognised hindrance to productivity is when ideators are *stuck* or suffering from an *impasse*. Impasses are commonly defined in contrast to Csikszentmihalyi's (1997) *flow* state, the state in which ideas appear rapidly and freely to the ideators. A state of impasse is synonymous with ideators losing their train of thought or having a moment of inability to come up with more new ideas[1]. Forming an understanding of what brings on this state of *impasse* may be beneficial to aid in the development of ideation support systems.

In the literature, we find studies have been run in which researchers attempt to measure or identify the occurrence of impasse states in a variety of ways, e.g. brain sensor data (fMRI), where changes in blood oxygen concentration indicate changes in brain activity (Chan et al., 2017); eye tracking, where long fixations with no saccades are taken to mean an impasse (Knoblich, Ohlsson and Raney, 2001). Whilst these methods have been shown to be somewhat accurate, they are not always practical to use in situations where ideation takes place outside a lab setting. Chan et al. (2017) (and Siangliulue et al. (2015)) use a pragmatic approach in which they rely on ideators to self-report an impasse by clicking a button saying "Give me other inspirations".

An alternative to such self-reporting is perhaps to simply use time between ideas as a continuous proxy for impasse states. Indeed, in Siangliulue et al. (2015), the offer of ideas after 30 seconds of idle time increases the number of ideas generated (in this case, their measure of success). They highlight though that an analysis of idle times showed that these were shorter at the start at the study and longer later in the study. The threshold for idle times in the on-demand condition increased towards the end of the study as well, an indicator that they were slowing down in the generation of ideas.

In the present study, we intend to measure these between idea times and verify whether they're a proxy for impasse states. In order to achieve this, we will be looking at the ideas generated by individuals over a fixed period of time. Throughout the studies in this thesis, we record the time-stamps of the *start* and *end* of each idea generated by the ideator participants. This is recorded in order to capture the time between the submission of one idea and the start of the next (*thinking times*) and the time it takes the ideator to type an idea (*typing times*). The sum of these two timings give us the overall formulation time of a single idea, also known as *response latency* (a term used by Nijstad, Stroebe and Lodewijkx (2002)). Instead of asking whether ideators are in a flow or impasse state, we look for aspects of ideation behaviour that are associated with longer or shorter thinking and typing times.

The work presented in Nijstad, Stroebe and Lodewijkx (2002) (and subsequently Nijstad, Stroebe and Lodewijkx (2003) and Nijstad and Stroebe (2006)) presents us with an organising principle which predicts the time-course of ideation, namely categorical structure. As reviewed in the literature, their theory of the cognitive model of idea generation called SIAM

[1] Also referred to as a failure in research by Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003), and Nijstad and Stroebe (2006).

(Search for Ideas in Associative Memory) resembles that of Search of Associative Memory (SAM) (Raaijmakers and Shiffrin, 1981) used to model memory retrieval in free recall tasks. It presents a cognitive model used to explain the process followed when generating ideas, suggesting that ideas are a result of repeated searches in memory (knowledge activation) in order to generate ideas. Much like SAM, SIAM makes use of the notion that search in memory consists of two stages: (1) Knowledge activation - *images* are retrieved from long-term memory; (2) The *features* of the retrieved image are used to make associations and generate ideas.

SIAM resembles information foraging in the sense that each image brought up in memory can be likened to a patch. The perceived exhaustion of ideas generated from an image leads the ideator to give up and activate another image from long-term memory, in order to generate more ideas. The findings from Nijstad, Stroebe and Lodewijkx (2002) show that ideas generated are semantically organised and that these are further temporally clustered, such that ideation times within a cluster are faster than between clusters. If a categorical pattern is to be observed during an ideation task, we might suppose that we are able to gain a stronger understanding of flow of ideation and understand the time management strategies associated with moving to another image in memory.

As already noted, this general approach is similar but somewhat distinct from the foraging theory approaches reviewed in Chapter 2. For example, if people choose to move on from a category of ideas then, according to the threshold in giving-up time models of thinking, this is clearly related to the concept of impasse-driven switching. However, if instead ideators' switch decisions are governed by a heuristic such as Green's rule, which is sensitive to the rate at which ideas are being generated, then the suggestion is that performance is more continuously monitored, and discrete states of flow and impasse might be a too simplistic way of modelling cognitive states during ideation. Whether either of these theories are correct, foraging theory offers a more refined set of stopping heuristics, and allows us to make connections and compare with time-management phenomena in very different tasks, such as accomplishment tasks. For this reason our analysis will include foraging theory's heuristics.

In this chapter, we will be replicating the study found in Nijstad, Stroebe and Lodewijkx (2002) (and subsequently Nijstad, Stroebe and Lodewijkx (2003) and Nijstad and Stroebe (2006)), hoping to replicate the results found in their individual[2] ideator condition - that is, temporally and semantically clustered ideas generated by an individual. In the aforementioned studies, participants are asked to generate a list of written ideas to respond

[2] Note that the research in these papers is mostly concerned with group performance. That said, their studies compare to an individual ideator control condition.

to an open-ended question. From the point of view of the research in this thesis, this is an exploratory study, designed to provide empirical data that is aimed to help further our understanding of the time-course of idea generation in a single question task. The study is grounded in the literature on SIAM with specific interest in the semantic and temporal clustering of ideas. Finding a categorical structure in a set of generated ideas would suggest that ideation tasks do indeed follow the patch-sequence structure of information foraging theories, often applied to more rote tasks in the literature (**RO1**). Furthermore, temporal clustering of idea categories may help us understand the cognitive state of a person as they are generating ideas, in particular as they slow down and eventually switch to another patch in memory (**RO2**).

It is conceivable that as an ideator exhausts the number of ideas they're able to generate based on an image in memory, they will employ a switching strategy to switch to another image in memory. Likely strategies have been covered in detail in the literature survey and include adaptive rules such as Green's rule (Green, 1984) or simply fixed-time or fixed-number rules (Iwasa, Higashi and Yamamura, 1981). These will be studied in more detail in Chapter 4.

Looking at the timestamps of submitted ideas in a single question ideation task allows us to ask questions about whether any of these patch-leaving heuristics are being used, but only in a limited way. One particular difficulty is that the time interval at category switches cannot be divided: the giving-up time can't be separated from the time to generate the next item. A further problem (to anticipate some of our results) is that category boundaries are likely to be much less distinct to the analyst.

As an addition to the original study by Nijstad, Stroebe and Lodewijkx (2002), we measure thinking and typing times. In the original study, the experimenters measured solely the time taken for each idea (response latencies). The typing time for an idea should not vary according to the categorical structure of idea-generation (or indeed according to any other cognitive process of ideation), unless ideators continue to formulate an idea after having begun to type it, indeed we perform this split in the knowledge that this might be the case, and understand that this might be a fairly crude estimation of thinking time. However, should we find that typing time does not vary according to categorical structure, we might also find that separating response latencies into thinking times and typing times may increase the sensitivity of our analyses. We predict that longer thinking times will arise when giving up on an image in memory and moving to another.

The rest of this chapter is divided into four parts. The initial study was run with ideators individually performing a single question ideation task by typing out as many solutions as possible to answer the question "What can the individual do to preserve the environment?" for 15 minutes.

The ideas generated in response to the initial (environment) question in this study were analysed and coded by independent coders to find semantic and temporal clustering of ideas. Analysis of the data for the initial study proved difficult, therefore in the second part of this chapter, the study-design was repeated using a different question, "What should a person do in order to maintain or improve their health?" in order to rule out the possibility of question-specific problems. Analysis of the data for the second (health) question study revealed similar issues, indicating the method itself may have limitations in its applicability. These issues seem to be in relation to the accepted sets of categories used, despite the fact that these sets have been used successfully in other studies. This will be discussed further in the chapter summary.

In the third part we combine the results of both studies and perform a follow-up analysis of the semantic difference between adjacent ideas in the idea sets. This analysis removes the need for pre-determined categories, and supports a shift from the assumed discrete states of flow and stuckness, in keeping with the continuous-monitoring aspects of some foraging theory stopping rules.

The fourth part is a summary of the chapter. We highlight the issues in coding with semantic categories and show how a focus on the semantic differences between adjacent pairs of ideas might allow the semantic structuring of ideation performance to more reliably be analysed.

3.2 Study 1a: Ideation cued by an environment question

In this section, a study is presented in which we attempt to replicate the findings of Nijstad, Stroebe and Lodewijkx (2002). The task used is an open-ended question designed to elicit a range of ideas from participants in a variety of areas. The task does not require any specific subject knowledge, making it appropriate to use with participants from different backgrounds.

3.2.1 Method

Design

A laboratory experiment was conducted. This followed a similar paradigm to that used by Nijstad, Stroebe and Lodewijkx (2002, 2003); Nijstad and Stroebe (2006): each participant was given a single ideation task to complete within 15 minutes[3] and asked to list as many ideas as they could that would serve to solve the particular problem given. Participants were given the same question as in one of the tasks listed in the Nijstad, Stroebe and Lodewijkx (2002) paper:

Ideation task: "What can the individual do to preserve the environment?"

This question was chosen as it offers a reference point for comparing our data to, as well as pre-determined list of possible categories, developed by Diehl (1991), consisting of a matrix of goals and means that can be combined to form categories in which to code ideas. For the environment question, the pre-determined category system consists of 10 goals (e.g. reduce water pollution, protect climate and atmosphere) and 5 means (e.g. consumption, production). In order to determine whether semantic clustering of ideas is occurring, we will be calculating the Adjusted Ratio of Clustering (ARC) scores for each of the datasets. This will be discussed further in the analysis section. The study itself aligns closely with Nijstad, Stroebe and Lodewijkx (2002) as does the initial part of the analysis. Whilst Nijstad, Stroebe and Lodewijkx (2002) did indeed measure timings of consecutive ideas that fell within a category, and the timings between category groups, they did so with the intention of showing that categorisation of ideas had an effect solely on response latencies. A variation we apply to this study is to further analyse whether thinking times (prior to actually typing - or formulating - the idea) is affected by any possible category switches. We explore this to see if there is any evidence to suggest impasse states happen during long thinking (*idle*) times and whether these correspond with category switches, as per the SIAM theory as well as foraging theories.

As we are looking to replicate the results from Nijstad, Stroebe and Lodewijkx (2002), we will test the same main hypotheses as stated in their studies. Note that Hypothesis 2 is adapted to reflect our split of response latency into thinking and typing times:

H1: Semantic clustering is occurring at a higher than chance level.

H2: Thinking times between semantically related ideas is shorter than that between semantically unrelated ideas.

H3: Because categorical structure is an effective generative device, clustering is positively correlated with the number of ideas generated (*overall fluency*).

[3] Nijstad, Stroebe and Lodewijkx (2002) gave participants 20 minutes. Note that this is a recognised design flaw in the present study. Diehl and Stroebe (1991) found that in both group and individual ideation sessions, the length of the session had no effect on productivity levels. That said, we will cover this topic further in our final Chapter: Discussion, conclusions and implications.

Participants

Twelve participants (4 female), age range from 24 to 31 (M=27.42, SD=2.40), were recruited from around the University of Bath, consisting of undergraduate and postgraduate students as well as university staff. Participants were recruited from a variety of disciplines although predominantly from the department of Computer Science. All participants were inexperienced ideators, that is, none of the participants performed ideation for a living, nor were they previously trained in ideation methods. Whilst the majority of these participants work in Human Computer Interaction, an often design-oriented field, none of the participants in this study were considered professional designers. No further profiling constraints were applied to recruitment for this study, due to the difficulty in finding volunteer participants. Participants were recruited by word of mouth and mailing lists. No reward was given for their participation.

Two participants' data were excluded as they failed to follow instruction. One participant did not write a list of ideas but a long stream of consciousness; the other generated a total of 4 ideas and showed no sign of activity after the 9-minute mark. Whilst these types of data are interesting in themselves, they did not adhere to the instructions and we were therefore not able to apply the same types of analysis as performed on the remaining 10 participants' data.

Materials

An application was specifically designed for this study using Visual Studio 2015. The application interface is illustrated in Figure 3-1. It consists of three screens. The first screen (figure 3-1a) takes the participant number as an input and uses this to set up a data file that captures user input. The second screen (figure 3-1b) consists of a simple input form with a submit button. To submit an idea, participants can either click submit or press enter. Both are included to allow participants to select the method they are most familiar with - this should not be a distraction from the ideation task. Above the input form is a large empty text box. Each time the user submits an idea, this is displayed in the large text box. As many ideas are expected, the text box is set to scroll to allow users to scroll up to see all the ideas they've submitted. Once ideas are submitted, they cannot be edited.

The input form automatically shuts down after 15 minutes, taking the participant to the third screen (figure 3-1c), which thanks the participant for their time and ends the ideation part of the study. The participant also has the possibility of ending the experiment early themselves by pressing the "end ideation" button in the bottom right hand corner of the screen. Data is captured by the environment automatically. Ideas entered as well as timestamps for the start of an idea and timestamps for when users press enter are captured in an excel file. This allows the calculation of the time participants take to think between ideas entered.

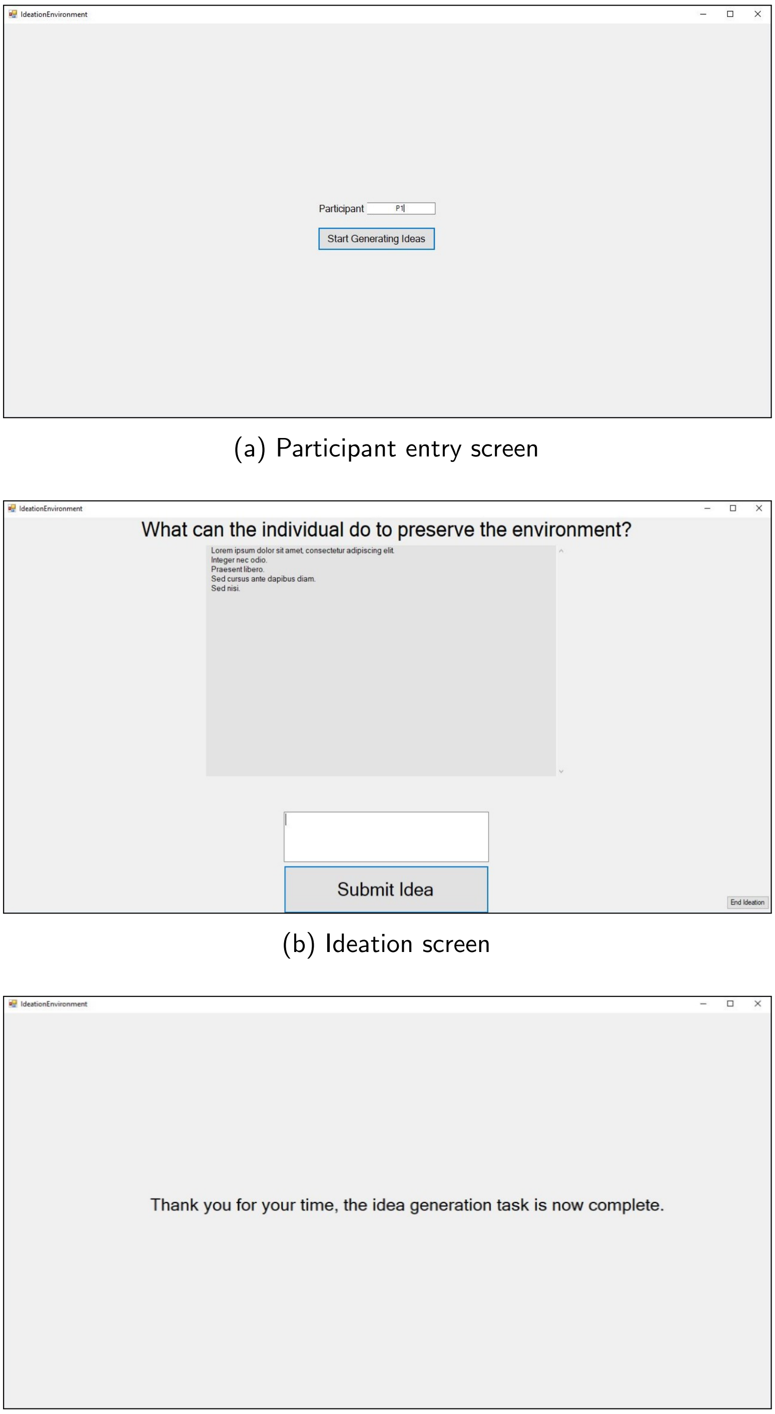
The software was run on a Lenovo Yoga 710-14ISK laptop running Windows 10. To ease input, a 22" screen, wired mouse and keyboard were attached to the laptop.

Procedure

Each participant attended the study individually. They were led into a quiet room situated in the University of Bath Computer Science Department. The room itself had no obvious distractions or décor that could act as a primer for ideas. Participants were handed an information sheet to read and could ask any question of the experimenter before starting the study (see appendix A.1 and A.2 for exact wording on information sheet and instructions). They were then explicitly instructed to generate ideas for 15 minutes according to a modified version of Osborn's (1957) brainstorming rules:

1. To generate many ideas
2. To avoid judging their own ideas
3. To generate 'wild' ideas
4. And to formulate their ideas beyond just a single word

Participants were asked to expand their ideas to incorporate specific behaviours; e.g. instead of saying "manage electricity", they were asked to indicate how they propose you could do so, e.g. "don't overfill a kettle if you are only making a single cup of tea". Additionally, they were welcomed to submit the same idea multiple times if they felt they had a better way of expressing it after submitting it. These instructions were included on the information sheet, which was placed on the table next to the participant should they wish to review these instructions. Once consent forms had been signed, participants were situated in front of the computer and asked to respond to a 3-minute practice task ("How can the number of tourists visiting the city of Bath be increased?") in order to familiarise themselves with the ideation environment. Once participants felt comfortable with the environment, they were provided with the main ideation task and given 15 minutes to type as many ideas as they could think of. Whilst participants were aware of the 15-minute time constraint, no clock was visible, so as to avoid distraction.



(a) Participant entry screen

What can the individual do to preserve the environment?

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Integer nec odio.
Praesent libero.
Sed cursus ante dapibus diam.
Sed nisi.

(b) Ideation screen

Thank you for your time, the idea generation task is now complete.

(c) End of experiment screen

Figure 3-1: Developed app for environment question study 1a. (a) Takes in participant number, (b) is the main ideation screen, showing an input area, a list of previously entered ideas and the question itself, (c) end of experiment screen appears automatically after 15 minutes, ending the ideation task.

Nijstad, Stroebe and Lodewijkx (2002) developed a short questionnaire, in order to gauge the ideators self-awareness of failures (impasses). They found that the responses correlated with overall fluency (number of ideas generated). In line with our goal to replicate their results, once the ideation task was complete, participants were given this short questionnaire asking them to rate the following on a scale from 1-10:

- "How difficult was it to keep on generating ideas?"
- "How often were you unable to generate ideas?"
- "How often did an idea you previously generated occur to you again?"

At the end of the study, the participant was debriefed and thanked for their time.

3.2.2 General analytic approach

264 ideas were generated in total by ten participants. No duplicates were found; however, 4 ideas were determined as being afterthoughts to the prior idea rather than ideas in themselves. These compound ideas were disaggregated into single unique ideas, resulting in a total of 260 ideas generated by the ten participants (M=26, SD=7.58). Note that combining these also meant the first key stroke of the original idea was noted as the start of typing time and the final enter-press of the afterthought was noted as the end of the typing time.

All 260 ideas were coded by two independent coders[4]. The independent coders were selected as they were external to the process and were therefore not familiar with the goals of the research. The coders were both male, aged 28 and 33 and were graduate students at the University of Bath in Mechanical Engineering and Computer Science. The coders were volunteers and received no payment for their participation. Although both coders were familiar with other methods of analysis, neither coder had experience in coding this type of data before. Consequently, prior to coding the full datasets, each of the coders received the following instructions:

They were instructed to classify each idea in a category using the means by goal matrix developed by Diehl (1991). For the environment question, this consisted of 10 goals and 5 means: 50 categories in total (see table 3.1). These goals and means function as a matrix of categories, such that when classifying an idea, the coder will match one goal and one means to create a composite category consisting of a 3-digit code. The idea "campaign for an organisation that saves whales" would be Goal 10 and Mean 5 = category 105. Coders

[4] Note that this coding was originally trialled by the experimenters - however, concern was raised about possible bias due to familiarity with study hypotheses. It was therefore decided to recruit external coders, blind to the study hypotheses. This will be discussed further in the conclusion of this chapter.

were asked to enter the codes into three separate columns in a spreadsheet of ideas; the first being the goal, the second column the mean and the third column any notes or additional goals/means they felt an idea could fall into.

Each independent coder was sent the data and the means by goal matrix. They individually coded 15 ideas each across 2-3 datasets of their choice and then met with the experimenter to talk through each of the scores. This gave the coder an opportunity to discuss their reasoning and raise any concerns they might have about the coding. Once the coders had shown full understanding of the meaning of each of the goals and means in the matrix, they were asked to categorise the remaining ideas themselves. Coders were not informed of the study hypotheses upon completion of the coding. This was to ensure that they would remain suitable candidates for coding possible future datasets.

	Means:		Goals:
01	Reduce waste	1	Consumption
02	Reduce chemical or toxic substances	2	Production
03	Reduce water pollution	3	Treatment of Waste
04	Reduce air pollution	4	Information
05	Reduce pollution of soil	5	Organization and action
06	Protect climate and atmosphere		
07	Reduce use of natural resources		
08	Reduce energy use and promote green energy		
09	Protect landscape		
10	Protect animals and plants		

Table 3.1: Category matrix for environment question consisting of 10 means and 5 goals.

The terms *fluency* and *productivity* will be used interchangeably to refer to the total number of ideas generated by a participant. The *diversity* (number of categories a participant covered in their responses) and *Adjusted Ratio of Clustering* (ARC) were calculated in line with Nijstad, Stroebe and Lodewijkx (2002). The Adjusted Ratio of Clustering (Roenker, Thompson and Brown, 1971) is an index used to show the occurrence of clustering in a data set. In ARC scores, chance clustering is shown by zero and perfect clustering is 1. Note that negative scores are possible and mean a 'less than chance' level of clustering. We use this calculation in order to find out if clustering is indeed taking place and to be able to use the information to verify if time allocation between and within clusters varies. ARC scores are calculated as follows:

$$ARC = \frac{R - E(R)}{\max R - E(R)} \tag{3.1}$$

where:

R = total number of observed category repetitions (i.e. the number of times an idea in one category follows an idea from the same category),
$\max R$ = maximum possible number of category repetitions, and
$E(R)$ = expected (chance) number of category repetitions.

Note that $\max R = N - k$ where *N* = total number of ideas generated, and *k* = number of categories represented in the ideation protocol. *E(R)*, the chance number of category repetitions, is calculated as:

$$E(R) = \frac{\sum\limits_{i} {n_i}^2}{N} - 1 \tag{3.2}$$

where n_i = total number of ideas generated in category *i*, and *N* the total of all ideas generated, as before.

When thinking times and typing times were analysed by ANOVA or t-tests they were log-transformed to normalise the distribution (as is common practice in the analysis of response times). Statistical analysis of correlations, t-tests and Cohen's *k* were done (as throughout this thesis) using IBM SPSS Statistics version 23. Tables and graphs were generated in Microsoft Excel for Office 365 Pro.

3.2.3 Results and discussion

We present the data in four parts. First, an overview of the data is presented, and a summary of the sets of ideas per participant including mean timings of ideas, overall fluency per participant as well as correlations between these and the responses to the post-study questionnaire. The second aspect of results covers the category coding, the agreement scores between our two independent coders and subsequently, results that highlight issues with the category system and coding method itself. Third, we compute the ARC scores based on the category coding for each participant (**H1**). The ARC scores are presented per participant and are correlated with *fluency* to see whether these are positively correlated (**H3**) Finally, despite the low ARC scores and low coder agreement scores, we present a tentative analysis of the effect of category switches on *thinking times* (**H2**) and *typing times*. We repeat this analysis using half category switches (in which adjacent ideas have either goal or means in common, but not both). Where possible we will draw lessons about strategy by comparing category switches with patch-leaving decision strategies.

Participants produced a mean of 26 ideas (SD=7.58) in their 15 minutes ideating (i.e. 1.73 ideas per minute; this compares with a mean productivity of 32.4 (SD=10.54) in 20 minutes, ie. 1.62 ideas per minute on the same task by Nijstad, Stroebe and Lodewijkx (2002)). The high variance in fluency can be explained by individual differences in the ideators. That said, these variances are often found in ideation studies (see e.g. Nijstad, Stroebe and Lodewijkx (2002)). This might incorporate a range of factors, such as ideators differing levels in motivation, threshold for what they feel is an idea worth writing down (self-evaluation), threshold for what is an acceptable number of ideas to generate per minute, etc. Note that individual differences might also account for the high variance in thinking times, typing times and response latencies.

Data for all 10 participants are shown in table 3.2. Means of thinking times, typing times, and response latencies have been calculated. These are presented along with the overall fluency per participant and their questionnaire responses to the three questions: Q1: "How difficult was it to keep on generating ideas?", Q2: "How often were you unable to generate ideas?", and Q3: "How often did an idea you previously generated occur to you again?". All three questionnaire responses were given on a 10-point scale, with higher numbers indicating "very difficult" or "often", representative of the participant perceiving a *failure* or *impasse*.

As would be expected, *fluency* and *response latencies* were highly negatively correlated ($r(8) = -.963$ at 0.01 level significance (2-tailed)), as well as *fluency* and *typing times* ($r(8) = -.849$ at 0.01 level significance (2-tailed)). No correlation was found between *fluency* and *thinking times*. Comparing *fluency* to post-experiment questionnaire responses, tests of correlation showed no significant correlations between Q1-fluency and Q3-fluency, but did show for Q2("How often were you unable to generate ideas?")-fluency (Q1-fluency: $r(8) = -.41, p = .24$. Q2-fluency: $r(8) = .85, p = .002$. Q3-fluency: $r(8) = .04, p = .92$)

No significant internal correlations were found between questionnaire responses, despite the fact that Q1 and Q2 seem to be asking very similar questions (Q1-Q2: $r(8) = .48, p = .16$; Q2-Q3: $r(8) = -.24, p = .50$; Q1-Q3: $r(8) = -.03, p = .93$)

These results are interesting as they do not replicate that found in Nijstad, Stroebe and Lodewijkx (2002), in which they show significant internal correlations between the questionnaire responses. Following these analyses, and due to the odd correlation encountered, we are not going to pay further attention to the questionnaire answers in this study. The questionnaire was used as a subjective measure of failures (impasse). We would expect

		Mean of Idea Timings			Questionnaire Responses		
Participant	Overall Fluency	Thinking Time	Typing Time	Response Latency	Q1	Q2	Q3
1	30	9.44 (10.30)	20.35 (12.97)	29.79 (16.02)	7	7	3.5
2	20	6.53 (4.12)	36.58 (31.36)	43.11 (32.21)	4	3	7
3	22	17.25 (11.91)	20.37 (13.63)	37.62 (18.22)	6	5	1.5
4	21	2.16 (1.79)	39.83 (21.07)	41.99 (21.18)	2	2	1.5
5	27	11.01 (13.65)	21.80 (14.44)	32.81 (18.12)	3.5	4	4
6	41	10.28 (14.42)	9.94 (7.14)	20.22 (16.68)	3	9	5
7	15	12.78 (13.27)	42.94 (38.67)	55.72 (38.92)	2	4	2
8	24	3.09 (2.35)	34.18 (16.06)	37.27 (16.48)	5	2	7
9	37	7.57 (8.88)	16.44 (9.78)	24.02 (13.62)	8	9	2
10	23	16.11 (15.83)	22.30 (16.70)	38.41 (21.91)	5	4	6
M (10)	26	9.62	26.47	36.09			
SD (10)	*7.58*	*4.74*	*10.48*	*9.63*			

Table 3.2: The overall fluency, mean times and questionnaire responses for participants in environment question study 1a. Mean of means are used for fluency and timing values. Standard deviations in brackets.

that those who rate ”How often were you unable to generate ideas?” highly to have a lower fluency. This would show that the subjective and objective measures match up, however, we have the opposite result. The positive Q2-*fluency* correlation seems counter-intuitive, although it is conceivable that people who are generating more ideas perceive themselves to be stuck more readily than others who generate fewer ideas, in light of the fact that participants have no comparison other than their own performance.

Cumulative productivity over time per participant is shown in figure 3-2. Notice the shapes of the curves seem to be much more linear than would be expected if a participant was running out of ideas. Participant on average generated fewer ideas in the final 3 minutes (M=4.4, SD=1.78) than they did in the first 3 minutes (M=6.2, SD=4.16), but this difference was not significant, $(t(9) = 1.247, p = .244, d = .39)$. Overlaying the generally linear relationship between number of ideas and time spent, there are clear spurts and times of slowing down, suggesting somewhat differing levels of flow and stuckness.

Figure 3-3 shows the step-wise cumulative productivity per participant. Each vertical hop indicates a submitted idea and horizontal lines represent response latencies. Although there does not seem to be a pattern to be seen in all the idea sets, some (e.g. P6, P7 and P9) show some evidence of clusters of quick successive ideas generated with some longer thinking times between these spurts - a promising, if informal, finding as it suggests the category structure of ideation proposed in SIAM may be occurring in these data sets.

It seems possible that the initial rate of generation of ideas will be slowed by reading the question, in which case the first idea thinking time will tend to be slow. However, a paired-samples t-test between first idea *thinking time* (M=13.84, SD=11.75) and the mean of subsequent *thinking times* (M=9.46, SD=4.73) showed no significant difference in means $(t(9) = 1.217, p < .225, d = .385)$. Mean first *typing times* (M=23.22, SD=13.53) and mean of subsequent *typing times* (M=26.69, SD=11.01) similarly showed no significant difference in means $(t(9) = -.622, p < .550, d = -.197)$. These analyses justify including first-idea times in the tables above and in subsequent analyses.

Coding ideas by category

In order to determine whether ideas are generated in semantic category clusters, each idea had to be coded into a category. The following section contains a summary of the results of this coding exercise, including how many of the categories each coder used. The two independent coders (IC) were asked to look at each of the 260 ideas and judge the goals and means of the ideas according to the category matrix shown in table 3.1 (in the analysis section of this chapter). IC1 used 33 of the possible 50 categories to code this dataset (Mean number of ideas per category = 7.88, SD=12.06). IC2 used 27 of the 50 categories (Mean number of ideas per category = 9.37, SD=10.66).

IC2 left 7 ideas uncoded (blank), with the comment "not relevant to the schema; not an environmental topic". The number of categories coded by each coder per participant is shown in table 3.3. The category coding gave us the *diversity* for each participant, that is, the total number of categories used by a participant in the 15-minute study. *Within-category fluency* is calculated by dividing *fluency* (*N*) by *diversity*, indicating how many ideas fall into the same category. Number of clusters (*category repetitions*) is calculated by looking at the number of times an item follows an item from the same category. *Cluster length* is calculated by $N/(N-R)$ where R is the number of category repetitions and N is the *fluency*, total number of ideas generated by a participant.

Coder agreement

The preliminary results in the previous section show that the overall numbers of categories used by each coder was fairly similar. Cohen's *k* was computed to determine if there was agreement between the two independent coders in classifying ideas given to the environment question. Cohen's *k* was selected as a suitable test due to the data being categorical (nominal). The same test was used in Nijstad and Stroebe (2006), giving us a good reference point for accepted values for this type of data. There was only a "fair"[5] agreement

[5] Fair according to (Landis and Koch, 1977, p.165).

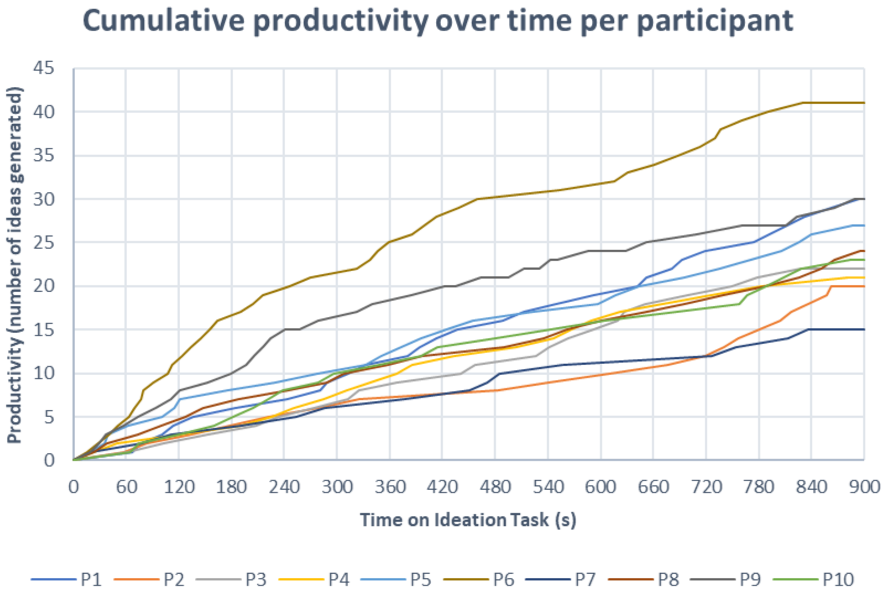


Figure 3-2: Cumulative productivity (number of ideas generated) over time for each participant in environment question study 1a.

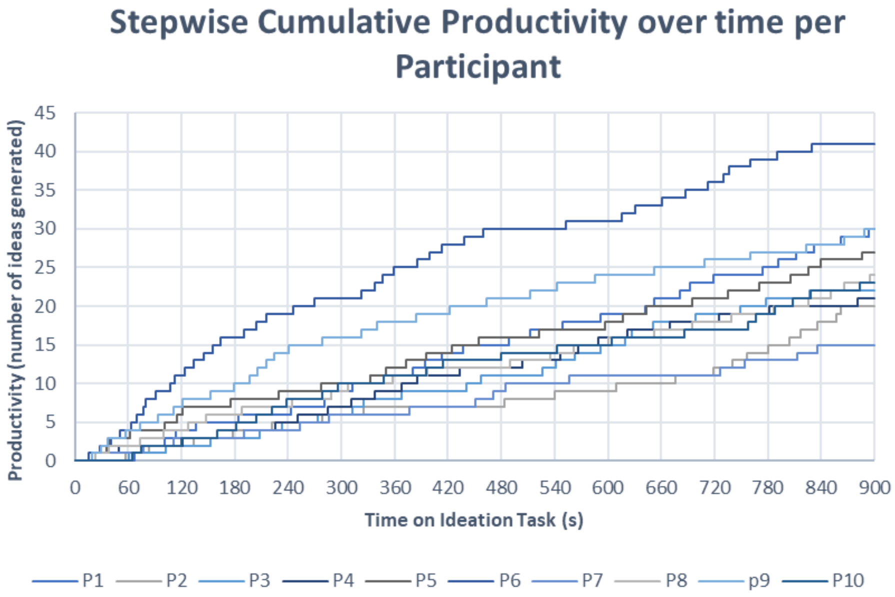


Figure 3-3: Cumulative productivity (number of ideas generated) over time for each participant in environment question study 1a, each vertical jump represents submitting an idea; horizontal lines represent response latencies.

Participant	Fluency	IC1 Category Count	IC2 Category Count
1A	30	10	12
2A	20	7	9
3A	22	10	12
4A	21	11	11
5A	27	11	11
6A	41	12	13
7A	15	10	10
8A	24	13	12
9A	37	14	15
10A	23	12	13
Mean Diversity (SD)		11.00 (1.84)	11.80 (1.60)
Mean Within-Category Fluency (SD)		2.37 (.57)	2.18 (.45)
Mean Cluster Length (SD)		1.29 (.10)	1.24 (.14)
Mean Category repetitions (SD)		5.80 (2.82)	5.10 (3.39)

Table 3.3: Total number of categories (*diversity*) per participant according to two independent coders (IC) in environment question study 1a. Standard deviations shown in brackets.

between the two coders, Cohen's $k = .36, p < .0001$[6].

Due to only finding fair agreement, Cohen's k was calculated for agreement solely on Means and for agreement solely on Goals, to find whether there are interesting findings based on those. There was moderate agreement between the two coders, Cohen's k $Means = .477, p < .0001$ and Cohen's k $Goals = .462, p < .0001$.

During the coding exercise, coders were encouraged to leave a comment on exceptional cases indicating if they found the idea too vague and therefore could be classified under several different categories. Coders left more comments than expected, see table 3.4 which shows total number of comments left per coder. Almost half of the comments indicated that the idea commented on could be categorised into more than just one additional category due to the idea being too vague or non-specific.

IC1 made a series of comments on the overall structure of the matrix coding scheme, reporting that it felt outdated. It was noted that most goals had some overlap and it was up to the discretion of the coder to decide which one they felt the idea aligned to the best. IC2 made further comments on the coding system itself but also on the ideas, indicating that many assumptions had to be made as to what the ideators meant, and stating that

[6] Comparison to the values found by Nijstad, Stroebe and Lodewijkx (2002); Cohen's $k = .87$, Cohen's $k = .89$, and by Nijstad, Stroebe and Lodewijkx (2003): Cohen's $k = .88$

many ideas in themselves were not formulated well enough to ascertain intended meaning with much confidence. IC2 commented on the duality of the ideas they were coding but additionally on the ”age” of the coding scheme – e.g. action taken through social media, would this come under means 4. Information or means 5. Organisation and action? Social media as well as behavioural change are topics that have rapidly developed, yet these were not covered fully in the environment topic category system. IC1 noted that although not wrong, the focus on reduction of use of materials and the treatment of waste materials relies heavily on the current population thinking about the ozone layer etc, whereas currently the population thinks more about action they can take through different types of dissemination of information on multiple media platforms. IC1 indicated that the goals of ”reduce energy use and promote green energy”, ”to protect climate and atmosphere” were usually tied together and it was difficult to distinguish whether an idea was more focused on the action of reduction to protect or protection by reducing.

Overall, verbal feedback from IC1 indicated that they felt they had to make assumptions about what the ideator meant in order to code the idea into the most appropriate category. Verbal feedback from IC2 made it clear that they, from the start, found a lot of ambiguity and room for interpretation in the coding system itself. IC1 and IC2 both indicated that coding was found to be very subjective and they both felt the need to recheck their coding to ensure consistency.

Number of unfilled		Number of comments	Number of comments with more than 1 addl. category
IC1	0	99	48
IC2	7	115	61

Table 3.4: Unclassified ideas and comments per independent coder in environment question study 1a.

Evidence of idea generation in categorical clusters

Here we present the results of the *Adjusted Ratio of Clustering* calculation (see equation 3.1) performed on the ideation data submitted for the environment question. We do this to seek evidence of a higher than chance occurrence of categorical clustering of ideas. Due to the lack of agreement between the two coders, we have been unable to use a compound score to analyse data, or to solely rely on one of the coding protocols to have been correct. We therefore treat the category coding from IC1 and IC2 separately and perform ARC score analysis on both sets. It is worth noting, before presenting the formal calculations, that the number of categories for all participants is rather high, given their overall fluency.

Participant	Submitted ideas	N	k	maxR	R	Ni	E(R)	ARC
1a	30	30	10	20	13	146	3.87	**0.566**
2a	20	20	7	13	7	88	3.40	**0.375**
3a	22	22	10	12	7	76	2.45	**0.476**
4a	21	21	11	10	2	73	2.48	*-0.063*
5a	27	27	11	16	8	103	2.81	**0.393**
6a	41	41	12	29	11	313	6.63	**0.195**
7a	15	15	10	5	3	45	2.00	**0.333**
8a	24	24	13	11	4	70	1.92	**0.229**
9a	37	37	14	23	11	173	3.68	**0.379**
10a	23	23	12	11	6	57	1.49	**0.474**

Table 3.5: The computation of ARC scores for each participant in environment question study 1a, using the codings of IC1.

In table 3.5 IC1's categories were used to test the existence of a categorical structure in the data. Parameters for the ARC score calculation were calculated for each participant: *N* is the total number of ideas included, *k* is the total number of categories represented in the set, R is the total number of observed category repetitions, maxR the maximum possible number of category repetitions (*N*-*k*), and n_i is the total number of ideas generated in each category. E(R), the chance number of category repetitions, was calculated using equation 3.2. ARC scores calculated from IC1's coding ranged from -0.06 to 0.57. Overall, the average of these (M=0.336 , SD=0.179)[7] was positive and significantly different from zero $(t(9) = 5.927, p < .001, d = 1.877)$. This is evidence that according to IC1'a category judgements, the participants ideas follow some (rather weak) systematic categorical structure, weakly accepting **H1**.

There was no significant correlation between submitted ideas and the ARC score $(r(8) = .03, p = .94)$, showing that we have no evidence in the data that using categories as a generative device is an effective strategy that affects productivity (**H3**).

We performed the same analyses using the categorisation done by IC2. The parameters and ARC scores were again calculated and are presented in table 3.6. Note here that 'submitted ideas' and 'coded ideas (*N*)' differ due to IC2 leaving some entries blank. The results of this followed somewhat the same pattern as with IC1's coding. This analysis returned a much lower mean ARC score although this was still significantly different from zero. Results ranged from -0.207 to 0.407. The average of these (M=0.180, SD=0.175)[8] was positive and significantly different from zero $(t(9) = 3.250, p = .010, d = 1.03)$. This

[7] Comparable to Nijstad and Stroebe (2006) who had mean ARC scores ranging from 0.20 - 0.41.

[8] Comparably lower than the 0.20 - 0.41 ARC scores in Nijstad and Stroebe (2006).

Participant	Submitted ideas	N	k	maxR	R	Ni	E(R)	ARC
1a	30	29	11	18	9	111	2.83	**0.407**
2a	20	20	9	11	4	64	2.20	**0.205**
3a	22	19	11	8	3	39	1.05	**0.280**
4a	21	21	11	10	4	67	2.19	**0.232**
5a	27	26	10	16	8	126	3.85	**0.342**
6a	41	41	13	28	12	317	6.73	**0.248**
7a	15	14	9	5	0	26	0.86	*-0.207*
8a	24	24	12	12	4	84	2.50	**0.158**
9a	37	37	15	22	5	143	2.86	**0.112**
10a	23	22	12	10	2	62	1.82	**0.022**

Table 3.6: The computation of ARC scores for each participant in environment question study 1a, using the codings of IC2 (variation between submitted ideas and N are due to coder leaving some ideas blank).

is evidence that according to this coder's categories, the participants' ideation has some systematic categorical structure and again lets us weakly accept **H1**. As with the data from IC1, there was no significant correlation between submitted ideas (fluency) and the ARC scores calculated from IC2's category scores $(r(8) = .33, p = .23)$, again refuting **H3**.

Effect of category switches on thinking and typing times

Having gained an overall understanding at the data, we can now turn to our primary question, which is whether the categorical structure of responses is reflected in the time-course of responses, as one would expect if category switches are analogous to patch leaving decisions. The following results are based solely on the codings of IC1[9]. All ideas were grouped into two groups, according to whether they were in the same category as the preceding idea or in a different category, i.e. after a category switch.

As noted above, the number of ideas per category was low, which might compromise these analyses. Nonetheless, we analysed the *thinking times* and *typing times* of ideas that were *within* a category and those that were *after* a category switch in order to understand whether categorisation has an effect on time to generate and formulate ideas. The means of the log transformed *within* category and *between* category timings were calculated for each participant. A paired samples t-test showed that the *thinking time* between (M=1.75, SD=.62) categories was significantly longer than within (M=1.08, SD=.61)

[9] In Nijstad, Stroebe and Lodewijkx (2002), all ideas were categorised by a main coder and a 10% sample was categorised by a second coder. Further analyses were performed solely using their main coder's data. In our study, the selection of the codings from IC1 is solely based on this coder being the first coder. No further criteria were used in selection of the data from IC1.

$(t(9) = 3.297, p = .009, d = 1.043)$. This of course meant that the overall *response latency* between categories (M=3.45, SD=.26) was also significantly longer than the overall *response latency* within categories (M=3.17, SD=.45) $(t(9) = 3.361, p = .008, d = 1.06)$, comparable with the findings in Nijstad and Stroebe (2006). This result shows support for **H2** that *thinking times* are shorter within categories, indeed suggesting that moving from one patch to another in semantic memory has an associated time-cost. Not surprisingly, the *typing time* within (M=2.86, SD=.49) and between (M=3.01, SD=.46) categories did not differ significantly. In the studies by Nijstad and Stroebe (2006), they found that response latency between categories was longer than within categories. These results support our decision to split thinking and typing times. We extend this finding by showing that it is the thinking aspect, not the typing aspect, of response latency that is affect by category switches. This supports the SIAM theory that a switch to a new category requires time to retrieve a new image from memory.

The effect of half category switches on thinking and typing times

It is plausible that people pivot around central themes when ideating. The concept of full category switches, combined with the problems we have encountered in the coding process itself, may be flawed in the sense that it relies too much on discrete states of flow and impasse. Looking away from the full category switches, we were interested in knowing if perhaps ideators perform half category switches, letting the goal of one idea inspire another idea of the same goal but different means. Likewise, the means of one idea could inspire another of the same means but different goal.

For these analyses, we divided the category coding done by IC1 into three different groups. Ideas with the same goal and mean as the preceding idea remained in the within category group as per the previous analysis. Ideas that shared either goal or mean, but not both, with the preceding idea was marked as a half category switch. Ideas that shared neither goal nor mean with the preceding idea were marked as full category switches. Log transformed timings were used again for this analysis, although the same patterns were seen in analyses on non-transformed data.

We first performed paired samples t-tests to see whether there was a significant difference between full and half switches. The means of *thinking times* in full switches (M=1.76, SD=.63) and *thinking times* for half category switches (M=1.77, SD=.69) were not found to be significantly different $(t(9) = -.096, p = .925, d = -.03)$. As expected, the means of *typing times* in full switches (M=2.98, SD=.55) and *typing times* for half category switches (M=2.99, SD=.42) were also not found to be significantly different $(t(9) = -.103, p = .920, d = -.03)$.

Secondly, we performed paired samples t-tests to compare the difference in means of half switches and within category timings. The means of *thinking times* in half switches (M=1.77, SD=.69) were significantly higher than *thinking times* for within category (M=1.08, SD=.61) $(t(9) = 3.241, p = .010, d = 1.02)$. No significant difference was found in the means of *typing times* in half switches (M=2.99, SD=.42) and *typing times* for within categories (M=2.86, SD=.49) $(t(9) = .905, p = .389, d = .29)$.

The typing time results are not surprising as these were not affected in the regular between and within time analyses. The significant difference between half switches and within category timings, but not full switches, may be due to the fact that the half switches form a large subset of the original full switches data. As half-switches were not significantly different from full switches, it does not seem as if people are pivoting around a central theme, or the types of central themes stipulated by our goals and means category matrix.

In terms of both full and half-category switches, then, there is some evidence in the time-course data that these index strategic choices, in that switches take longer than within-category thinking times. It is not possible to separate giving-up times from next item times at category breaks, but we might still draw some tentative lessons about heuristics that underpin the category-switch decisions.

Figure 3-4 shows the untransformed distribution of thinking times identified to be between categories (note the > 20 overflow bin). The mean of thinking times identified to be within categories (M=6.43, SD=8.28) is highlighted on the graph. The distribution of thinking times shows one of the characteristic findings that led Payne, Duggan and Neth (2007) to propose a two-factor giving up rule. First, most between-category thinking times were considerably longer than the mean within-category thinking times. This is suggestive of a patch-leaving decision influenced by the failure to generate an idea, and the decline in idea-generation rate. Secondly, there are some between-category thinking times that are considerably quicker than the mean within-category time. This is suggestive of a rarer success-based patch-leaving decision, leaving a patch immediately after an idea has been encountered, a kind of sub-goal completion.

3.2.4 Summary and conclusions

Ideation performance and time-course

There was some evidence in the main data that participants found ideas harder to come by as time progressed, but certainly 15 minutes did not completely exhaust ideators stock of ideas. Nevertheless, all participants might be described as being "stuck" sometimes, as can be seen by the ragged cumulations of ideas over time. Some degree of stuckness is

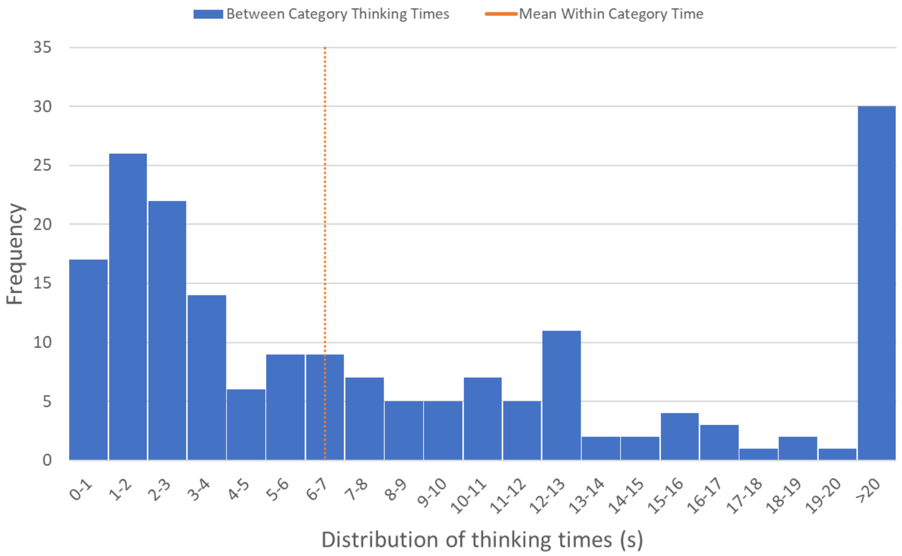


Figure 3-4: Distribution of between category thinking times environment question study 1a showing a cluster of short thinking times and a long tail of longer thinking times, compared to mean within category thinking time (orange line). Note the > 20 overflow bin.

also apparent in the high standard deviation of thinking times that shows that some values are more than twice the size of the mean. In figures 3-2 and 3-3 the time-course of each participant's idea-set, steeper areas of a curve represent a faster ideation rate. These could be interpreted as representing some kind of flow state. The shallower parts of the curve represent a slower ideation rate and may be interpreted to represent impasse, although they could alternatively be interpreted as mini breaks or periods of distraction.

Post-experiment questionnaire responses

The attempt to gain insight into ideation through direct questions was not very successful. The questions used were shown to successfully correlate subjective perception of failures and objective measures of failures in Nijstad, Stroebe and Lodewijkx (2002) but responses to the questions in our study did not show these correlations. This could possibly be due to the small number of participants, although the control condition in Nijstad, Stroebe and Lodewijkx (2002) was comparable in size. We reported an anomalous and surprising correlation between the question "How often were you unable to generate ideas?" and fluency, showing that those who reported they were unable to generate ideas more often, were also those who generated the most ideas. This is likely a chance effect, but may be due to the ideators not having anything to compare against. It is conceivable that someone coming up with many ideas is looking to achieve a high productivity and so may be more aware of when they are stuck, in comparison to someone who generates ideas at a more leisurely pace and takes their time formulating each idea. Individual differences in ideators like this might be one of the factors leading to the noisiness in the data.

Evidence of clustering in idea generation

We encountered some issues with the category coding itself, which will be covered later in this section. Due to this, we performed the ARC score calculations on the data twice, based on both coders' categorisations of the data. Our mean ARC scores according to IC1 only, are comparable with those of Nijstad and Stroebe (2006), whose ARC scores all fell between .20 and .41. The ARC scores according to IC2's coding were somewhat lower. The mean ARC scores were significantly different from zero, which is the value for a chance level of clustering, offering support for the hypothesis that ideas are generated in semantic clusters. However, these results must be treated with caution given fairly low inter-rater agreements as well as their comments on the category structure.

According to the SIAM model, ARC should be correlated with fluency, because clustering itself supports the search for ideas (Nijstad and Stroebe, 2006). However, in this study there was no evidence for the predicted correlation between ARC and fluency. The limitation here could be due to a range of variables, especially the inherent issue with the categorical coding system itself, and the low number if ideas per category in all participants' data according to either coder's categories.

Allocation of time within and between clusters

Interestingly, and despite the evident noisiness of our category codings, looking at the timings between and within categories, we observed a significant difference in the *thinking time* (and therefore the overall *response latency*). This offers some support for the suggestion that there is, if only small, a pause between categories in which ideators will stop to retrieve another image.

There is no similar effect on *typing time*, which further suggests that although the split of thinking and typing time might have seemed crude (in that participants could continue to think having initiated typing), in fact the split has been helpful in identifying and locating ideation impasses that happen between categories. This is a small additional methodological contribution to the literature on ideation.

Despite the general influence of category boundaries on thinking times, for each of the participants in this study, the maximum within category thinking time is higher than their minimum between category thinking time. This perhaps argues against a simple relation between thinking time and experienced impasse states. Similarly, Chan et al. (2017) discuss how idle times might indicate actively formulating the next idea or being distracted by external stimuli.

Finally, we noted that some between-category switches had very short thinking times. This finding relates to the theory of task switching rules by Payne, Duggan and Neth (2007), who suggested that task switches might usually occur when production was difficult, but occasionally probabilistically occur immediately after a sub-goal success. We explore this further in study 2 in Chapter 4, in which we analyse the timestamps of explicit switches between tasks.

Limitations of this study: issues with the category structure and coder agreement

The low agreement between coders' categorisations was a major nuisance for our initial research objectives. The agreement between the two coders, Cohen's $k = .365, p < .0001$ was far from that of the values found in Nijstad, Stroebe and Lodewijkx (2002), Cohen's $k = .87$, and Nijstad, Stroebe and Lodewijkx (2003), Cohen's $k = .88$. Krippendorff (2018) states that drawing conclusions from any data that shows a less than 0.67 k value should not be done.

For these reasons, and in the light of the coders' comments on their difficulties, a second version of our first study was run using a different problem. Refinement of the environment coding scheme was initially considered, however, with the availability of a second question and associated coding scheme, as well as the possibility of a different approach to analysis (i.e. semantic distances, see section 3.4), it was decided that this was out of scope of the research goals. In the next section, we present the results of a similar 15-minute ideation study cued by a question on personal health rather than the environment. This study was run in the hope of being better able to replicate the findings of Nijstad, Stroebe and Lodewijkx (2003), in particular, with a focus on coder agreement scores for the categorisation of ideas.

3.3 Study 1b: Ideation cued by a health question

In the study presented in the previous section, we attempted to replicate the findings of Nijstad, Stroebe and Lodewijkx (2002). Our goal was to gain a stronger understanding of the time-course of coming up with ideas in a single ideation task by studying the link between Search for Ideas in Associative Memory, the concepts of flow and impasse, and possible foraging theories that could be applied to explain the patch-like pattern followed when generating ideas. The basis for our theory that ideation follows a foraging like paradigm lay in the clustering of ideas in semantic categories, a phenomenon shown to occur in ideation (Nijstad, Stroebe and Lodewijkx, 2002). Our study included an environment question that had a verified category system. Unfortunately, although there was fair agreement between category coders, this was far off that of Nijstad, Stroebe and Lodewijkx (2002). The category structure itself was highlighted as having issues by both coders who complained that multiple categories could be assigned to a single idea and that some of the categories were ambiguous. The comment from one coder on the datedness of some of the categories and the lack of means relating to social media and persuasion brings the environment category system into question. We found promising results based on the individual coder's protocols, showing that there is evidence of clustering of ideas and that there is a significant difference to be found in the thinking times of between cluster and within cluster ideas. We therefore repeat this study with a different question, in order to rule out question, and associated category structure, specific issues.

In this section, we present a study similar to that in section 3.2, in which we attempt to replicate the findings of Nijstad, Stroebe and Lodewijkx (2002). A different task was chosen from those studied in Nijstad, Stroebe and Lodewijkx (2003), to explore whether the problems discovered in the analysis of the previous study are question-specific or based on methodological issues.

3.3.1 Method

Design

This study was identical to the first in every respect except for those noted below. The method, approach and hypotheses were unaltered. Whilst the 15-minute timespan for the ideation task has been suggested to be too short to exhaust ideators stock of ideas, this study was intended to follow the same design as the environment question study, this timespan therefore remains the same for this study. In study 2 in Chapter 4 and study 3 in Chapter 6 we increase the timespan and discuss the impact this has on the ideation tasks. The most important change was the ideation task question itself.

> *Ideation task 2: "What should a person do in order to maintain or improve their health?"*[10]

Again, this question was chosen due the existence of a pre-determined category system, developed by Diehl (1991) and the ability to compare data to that in Nijstad, Stroebe and Lodewijkx (2003). The category system follows the same matrix of goals and means that

[10] The wording on this task was changed from "What can everybody do to improve or maintain one's own health?" as this was found to be an awkward translation. Original wording "

can be combined. For the health question, the pre-determined category system consists of 11 goals (e.g. avoid physical injuries, maintain healthy teeth) and 9 means (e.g. medication and treatment, social factors). These will be discussed further in the analysis section.

Participants

Ten participants (3 female), age range from 22 to 31 (M=25.1, SD=3.36), were recruited from around the University of Bath. The participant pool mostly consisted of undergraduate and postgraduate students as well as university staff. Participants were recruited from a variety of disciplines although predominantly from the department of Computer Science. All participants were inexperienced ideators, that is, none of the participants performed ideation for a living, nor were they trained in ideation methods. Whilst the majority of these participants work in Human Computer Interaction, an often design-oriented field, none of the participants in this study were considered professional designers. No further profiling constraints were applied to recruitment for this study, due to the difficulty in finding volunteer participants. Participants were recruited by word of mouth and mailing lists. No reward was offered for this study.

Analysis by category

The ideas were categorised into the 99 possible categories by two coders and ARC scores computed in order to see if ideas follow a categorical structure. Timings within and between category clusters were analysed to see if foraging theory applies to this type of data. Recall that timestamps were recorded by the software, providing us with thinking times, the time between submitting an idea and typing the first letter of the next, typing times, the time between typing the first letter of an idea and pressing enter to submit it, and response latencies the sum of the two former timings.

199 ideas were generated in total by the ten participants. No duplicates were found, however, 1 idea was judged as being an afterthought to the prior idea. This was therefore combined with the prior idea, resulting in a total of 198 ideas generated by the ten participants (M(10)=19.8, SD=6.06). Note that combining these also meant the first key stroke of the original idea was noted as the start of typing time and the final enter-press of the afterthought was noted as the end of the typing time.

In Nijstad, Stroebe and Lodewijkx (2003), a coder classified all ideas submitted by all participants. A random subset of about 10% was then classified by a second coder. Cohen's $k = .85(p < .001)$. In our study, all 198 were coded by two new independent coders (different to the environment question). The independent coders were selected as they

	Means:		Goals:
01	Maintain or improve physical fitness	1	Food and substance intake
02	Maintain or improve perception (sight, hearing, etc)	2	Medication and treatment
03	Optimize metabolism and avoid deficits	3	Clothing
04	Avoid strain on skeleton, improve posture	4	Bodily hygiene
05	Avoid weight problems	5	Environmental factors
06	Avoid physical injuries	6	Physical exercise
07	Maintain or improve psychological health	7	Information
08	Avoid physical strain	8	Life style (health-relevant behaviors/attitudes)
09	Avoid poisonous substances and radiation	9	Social factors
10	Practice health prevention		
11	Maintain healthy teeth		

Table 3.7: Category matrix for health question consisting of 11 means and 9 goals.

were external to the process and therefore not familiar with the goals of the research. Coder 3 was male, aged 30 and a graduate student at the University of Bath in Computer Science. Coder 4 was female, aged 28 and a graduate student at the University of Bath in the School of Management. The coders were volunteers and received no payment for their participation. Although both coders were familiar with other methods of analysis, neither coder had experience in coding this type of data before. Consequently, prior to coding the full datasets, each of the coders received the same instructions as coders 1 and 2:

They were instructed to classify each idea in a category using the means by goal matrix developed by Diehl (1991). For the health question, this consisted of 11 goals and 9 means: 99 categories in total (see table 3.7). Note that not all categories will be used by ideators.

These goals and means function as a matrix of categories, such that when classifying an idea, the coder matched one goal and one mean to create a composite category consisting of a 3-digit code. The idea "eat lots of fish to avoid an omega-3 deficit" would be Goal 03 and Mean 1 = category 031. As done in the environment question, coders were asked to enter the codes into three separate columns: the first being the goal, the second being the means and the third column being any notes or additional goals/means they felt an idea could fall into. Each independent coder was sent the data and the means by goal matrix, coded a subset individually and met with the experimenter to discuss. Once the coders had shown full understanding of the meaning of the goals and means in the matrix, they were asked to categorise the remaining ideas themselves.

3.3.2 Results and discussion

In the following section we present our data from the follow-up study in four parts. The first is an overview of the data and a summary of the sets of ideas per participant including mean timings of ideas, overall fluency per participant as well as correlations between these and the responses to the post-study questionnaire. We analysed the post-study questionnaire and show inter-correlation scores as all the questions cover the concept of *failure*. The second part of this section covers the category coding for the health question along with the agreement scores between our two independent coders.

Third, we computed the ARC scores based on the category coding for each participant (**H1**). The ARC scores are presented per participant and is correlated with *fluency* to see whether these are positively correlated (**H3**). Finally, despite finding low ARC scores and low coder agreement scores again in the health question, we present a tentative analysis of the effect of category switches on *thinking times* (**H2**).

Participants produced a mean of 19.8 (SD=6.06) ideas in 15 minutes, equivalent to 1.32 ideas per minute, which is markedly less fluent than Nijstad, Stroebe and Lodewijkx's (2002) participants' mean fluency of 45.87 (SD=20.03) in 20 minutes (2.29 ideas per minute). In an ideation study where participants are asked to generate as many ideas as possible, generating an average of this few ideas seems odd. We suspect this could perhaps be related to the lack of financial incentive on the motivation of ideators, however, if this were to be the case, this result is still surprising given the close match in fluencies in the first study (which likewise lacked financial incentive) and the comparison study. It seems likely a matter of chance, given the high variance in fluency in all studies.

Data for all 10 participants are shown in the table 3.8. Means of *thinking times*, *typing times*, and *response latencies* have been calculated. These are presented along with *fluency* and their questionnaire responses to the questions: Q1: "How difficult was it to keep on generating ideas?", Q2: "How often were you unable to generate ideas?", and Q3: "How often did an idea you previously generated occur to you again?".

As would be expected, *fluency* and *response latencies* are highly negatively correlated ($r(8) = -.973$ at 0.01 level significance (2-tailed)), as well as *fluency* and *typing times* ($r(8) = -.834$ at 0.01 level significance (2-tailed)). No correlation was found between *fluency* and *thinking times*. These results replicate our results for the environment question. This is an intuitive result as this would be expected in a timebound task – longer response latencies result in less productivity across the 15-minute time-course.

		Mean of Idea Timings			Questionnaire Responses		
Participant	Overall Fluency	Thinking Time	Typing Time	Response Latency	Q1	Q2	Q3
1	30	12.55 (10.78)	16.29 (11.91)	28.84 (14.99)	4	7	9
2	19	13.20 (11.09)	33.67 (10.98)	46.87 (14.60)	4	3	2
3	14	25.19 (28.25)	35.22 (18.54)	60.41 (39.60)	5	6	3
4	15	21.28 (20.23)	38.38 (12.21)	59.66 (24.06)	5	6	6
5	20	10.36 (15.89)	34.05 (19.32)	44.41 (26.16)	6	7	9
6	17	2.85 (2.50)	49.63 (17.29)	52.48 (17.57)	2	3	4
7	16	20.53 (19.66)	33.53 (15.21)	54.05 (28.39)	7	5	7
8	21	10.41 (8.92)	31.43 (18.53)	41.84 (23.14)	8	2	7
9	32	6.29 (11.16)	21.27 (9.64)	27.56 (13.81)	2	2	3
10	14	11.76 (7.27)	41.62 (23.09)	53.38 (27.19)	1	2	1
M (10)	19.80	13.44	33.51	46.95			
SD (10)	*6.06*	*6.60*	*8.96*	*10.96*			

Table 3.8: The overall fluency, mean times and questionnaire responses for participants in health question study 1b. Mean of means are used for fluency and timing values. Standard deviations shown in brackets.

Comparing *fluency* to post-experiment questionnaire responses, no significant correlations were found between fluency and any question responses $(Q1 - fluency : r(8) = -.14, p = .70.$ $Q2 - fluency : r(8) = -.04, p = .90.$ $Q3 - fluency : r(8) = .28, p = .43)$

Interestingly, significant internal correlations were found between questionnaire responses, although not between Q1 and Q2, that were essentially asking the same question. Q1-Q2: $r(8) = .37, p = .30$; Q2-Q3: $r(8) = .65, p = .04$; Q1-Q3: $r(8) = .66, p = .04$.

Cumulative productivity over time, figure 3-5, looks similar to the curve found for the environment question, if not even more linear. This appears to be a more linear relationship than you might expect in a situation of diminishing returns. However, participants generated fewer ideas on average in the final 3 minutes (M=2.9, SD=1.6) than in the first 3 minutes (M=4.8, SD=1.93) $(t(9) = 3.943, p. = 003, d = 1.25)$.

Figure 3-6 shows the cumulative productivity per participant. Each vertical hop indicates a submitted idea and the horizontal lines represent response latencies. Quick successive ideas generated with some longer thinking times were not easily discerned from this dataset, these were more prominent in the environment question.

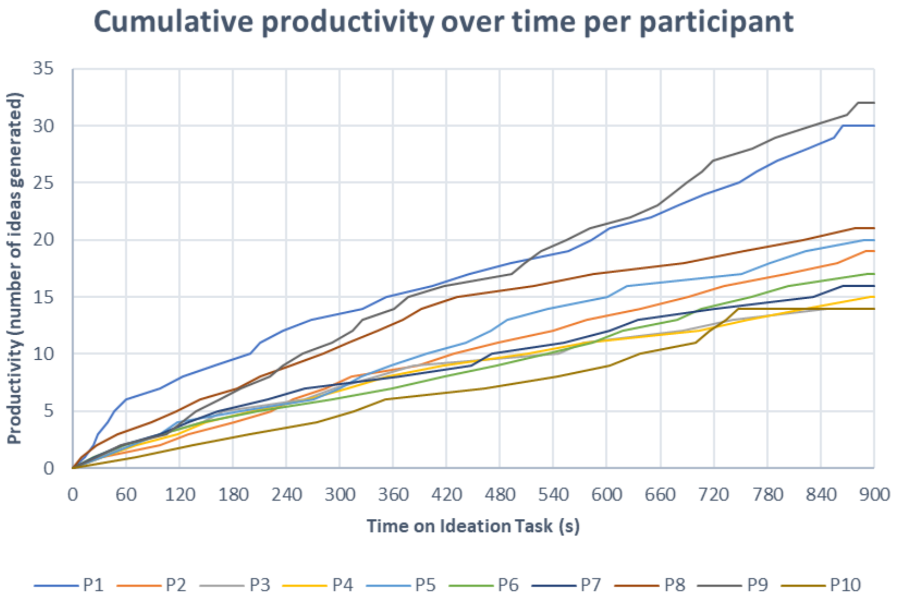


Figure 3-5: Cumulative productivity (number of ideas generated) over time for each participant in health question study 1b.

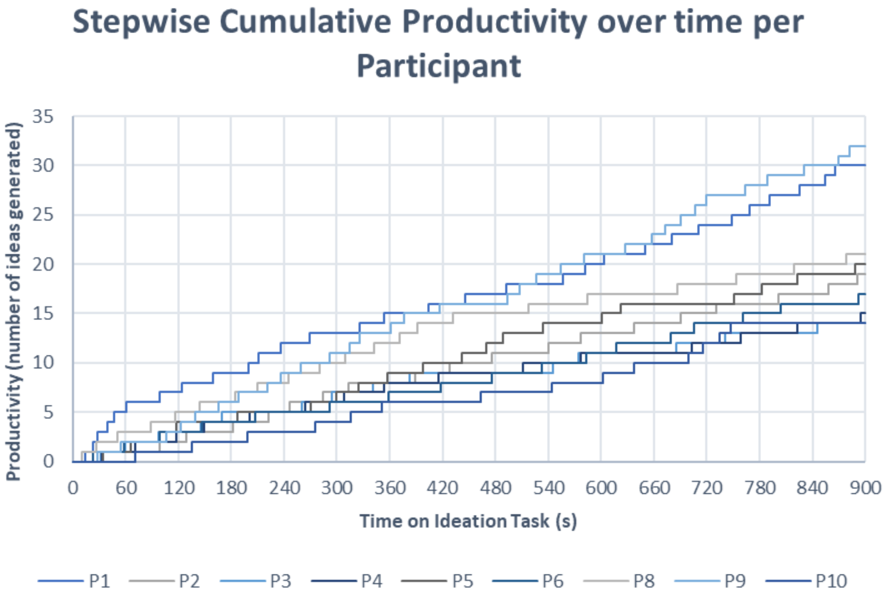


Figure 3-6: Cumulative productivity (number of ideas generated) over time for each participant in health question study 1b, each vertical jump represents submitting an idea; horizontal lines represent response latencies.

As in study 1a, the time to generate the first idea was examined. This analysis showed that first idea *thinking times* (M=5.71, SD=1.96) were significantly smaller than remaining *thinking times* (M=12.82, SD=15.95), $(t(9) = -3.557, p = .006, d = -1.12)$; a result opposite to what was expected if you take reading of the question into account.

Coding ideas by category

In order to determine whether ideas are generated in semantic category clusters, we needed to code the data qualitatively. In the environment question, this step proved to be problematic due to only a fair agreement between coders, as well as an overwhelming amount of comments on both the vagueness of the pre-set categories themselves and the difficulty of having to make assumptions about what the participant meant in badly specified ideas. In the health question for the current study, we hoped to overcome this issue and show that this disagreement was a question or category specific observation. Two new coders, coders 3 and 4, were asked to look at each of the 198 ideas and judge the goals and means if the ideas according to the category matrix shown in table 3.7.

IC3 coded this dataset into 35 of the possible 99 categories (Mean number of ideas per category = 5.66 SD=7.74). IC4 coded this dataset into 28 of the 99 possible categories (Mean number of ideas per category = 7.07, SD=8.45). No ideas were left blank by either coder.

As in the environment question, category coding gives us the following variables: *diversity* for each participant, the total number of categories used by a participant in the 15-minute study; *within-category fluency* calculated by dividing *fluency* (*N*) by *diversity*; number of clusters (*category repetitions*), calculated by looking at the number of times an item follows an item from the same category; and finally *cluster length*, calculated by $N/(N-R)$ where R is the number of category repetitions and N is the *fluency*, total number of ideas generated by a participant (see table 3.9).

Coder agreement

Comparable to the environment question, the summary results show that the overall number of categories used by each coder was fairly similar, with agreement on the cluster length and number of clusters. Cohen's *k* was run to determine if there was agreement between the two independent coders in classifying ideas given to the health question. There was a good[11] agreement between the two coders, Cohen's $k = .61, p < .0001$, much higher than the agreement found in the environment question.

[11] Good according to (Landis and Koch, 1977, p.165)

Participant	Fluency	IC3 Category Count	IC4 Category Count
1B	30	17	12
2B	19	10	8
3B	14	11	10
4B	15	7	10
5B	20	7	9
6B	17	12	12
7B	16	8	9
8B	21	11	11
9B	32	13	11
10B	14	11	10
Mean Diversity (SD)		10.70 (2.87)	10.20 (1.25)
Mean Within-Category Fluency (SD)		1.90 (.48)	1.94 (.51)
Mean Cluster Length (SD)		1.19 (.15)	1.19 (.15)
Mean Category repetitions (SD)		3.20 (2.56)	3.20 (2.56)

Table 3.9: Total number of categories (*diversity*) per participant according to two independent coders (IC) in health question study 1b. Standard deviations shown in brackets.

As in the environment question, and unsurprisingly given there are fewer 'half categories', agreement was higher when measured independently for Goals and Means: Cohen's *k* for Goals was good, Cohen's $k = .77, p < .0001$. Cohen's *k* for Means was good, Cohen's $k = .74, p < .0001$.

During the coding exercise, coders were encouraged to leave a comment on exceptional cases indicating if they found ideas or parts of the classification system too vague. For the health question, coders also left more comments than expected, see table 3.10 which shows the total number of comments left per coder. In the environment question, almost half of the comments related to ideas that could be coded into several categories. This number was much lower in the health question.

Number of unfilled		Number of comments	Number of comments with more than 1 addl. category
IC3	0	56	22
IC4	0	71	11

Table 3.10: Unclassified ideas and comments per independent coder in health question study 1b.

In discussions with the independent coders after coding, IC3 said "some seem to clearly focus around good and bad things to consume, these MUST be in the same category, but in the category structure you are missing a consume mean. Also consumption hops from things that are bad for you to healthy foods to medicine. There is a goal for each one of these." "If the schema is not specific it is hard to code but being specific means that many single ideas can fit into multiple categories, there is an overlap". IC4 struggled to select the correct goal for some as there were clearly multiple goals. Just as IC1 and IC2 in the environment question, IC3 and IC4 indicated that coding was found to be very subjective.

Evidence of idea generation in categorical clusters

In this section we present the results of the *Adjusted Ratio of Clustering* calculation (see equation 3.1) performed on the ideation data submitted for the health question. We did this analysis to find evidence of a higher than chance occurrence of categorical clustering of ideas. Although a much higher agreement was found between the two coders for the health question in relation to the environment question, this still does not compare to the Cohen's $k = .88$ found in Nijstad, Stroebe and Lodewijkx (2003) for this question. As a result of this, we treated the category coding from IC3 and IC4 separately and performed the ARC score analysis on both sets to see if they yielded similar results.

In table 3.11 IC3's categories were used to calculated the parameters for ARC score calculations for each participant. ARC scores calculated from IC3's coding ranged from -0.24 to 0.53. Overall, the average of these (M=0.225, SD=0.209)[12] was positive and significantly different from zero $(t(9) = 3.402, p = .008, d = 1.076)$. This is evidence that according to IC3'a category judgements, participants' ideas follow some systematic categorical structure, accepting **H1** for the health question as well as the environment question. We should note however that the ARC scores are generally low, even if usually positive.

In line with our findings for the environment question, there was no significant correlation between fluency and the ARC score $(r(8) = .38, p = .27)$, showing that we had no evidence in the data that using categories as a generative device is an effective strategy that affects productivity (**H3**).

We performed the same analyses using the categorisation done by IC4. The parameters and ARC scores were again calculated and presented in table 3.12. We performed the analyses in order to see if we could replicate the results of IC1, IC2 and IC3, however for IC4 we were not able to find an ARC score mean significantly different from zero. Results ranged from -0.217 to 0.412. Overall, the average of these (M=0.096, SD=0.184) was not significantly different from zero $(t(9) = 1.650, p = .133, d = .52)$. This is evidence that according to IC4's category coding, participants' ideation do not follow a systematic categorical structure. As with the data from IC3, there was no significant correlation between submitted ideas

[12] Comparable to Nijstad and Stroebe (2006) who had mean ARC scores ranging from 0.20 - 0.41.

(fluency) and the ARC scores calculated from IC4's category scores ($r(8) = .45, p = .07$), again refuting **H3**.

IC3 submitted feedback commenting on a couple of issues with the coding scheme. In particular, they found sleep to be a difficult one to code as the ideator was not specifying why they were recommending more sleep. Additionally, they found overlap in a few of the goals, e.g. is optimisation of metabolism and avoiding deficits not a method of practising health prevention?

Participant	Submitted ideas N	k	maxR	R	Ni	E(R)	ARC
1b	30	17	13	6	114	2.8	**0.31**
2b	19	10	9	5	69	2.63	**0.37**
3b	14	11	3	0	22	0.57	*-0.24*
4b	15	7	8	3	49	2.27	**0.13**
5b	20	7	13	7	96	3.80	**0.35**
6b	17	12	5	3	29	0.71	**0.53**
7b	16	8	8	4	74	3.63	**0.09**
8b	21	11	10	3	55	1.62	**0.16**
9b	32	13	19	8	128	3.00	**0.31**
10b	14	11	3	1	20	0.43	**0.22**

Table 3.11: The computation of ARC scores for each participant in health question study 1b, using the codings of IC3.

Participant	Submitted ideas N	k	maxR	R	Ni	E(R)	ARC
1b	30	12	18	7	148	3.93	**0.22**
2b	19	8	11	6	89	3.68	**0.32**
3b	14	10	4	0	24	0.71	*-0.22*
4b	15	10	5	1	33	1.20	*-0.05*
5b	20	9	11	6	70	2.50	**0.41**
6b	17	12	5	1	29	0.71	**0.07**
7b	16	9	7	2	50	2.13	*-0.03*
8b	21	11	10	2	57	1.71	**0.03**
9b	32	11	21	6	160	4.00	**0.12**
10b	14	10	4	1	24	0.71	**0.09**

Table 3.12: The computation of ARC scores for each participant in health question study 1b, using the codings of IC4.

Effect of category switches on thinking and typing times

Having gained an understanding of this data and the categorical structure found by IC3 (but not IC4) allowed us to respond to our primary question for the health task: whether the categorical structure of responses was reflected in the time-course of responses. The following results are based on analyses made solely by IC3 as those by IC4 did not show evidence of an overall categorical structure to the ideas generated. All ideas were again coded into two groups, whether they fell in the same category as the preceding idea or in a different category, i.e. a category switch. These groupings allow us to compare thinking times and typing times of ideas that fall within the same category, and ideas that are generated at a category switch. Comparing these gives us a stronger understanding of whether categorisation influences the time it takes to generate ideas. The means of log transformed within category and between category timings were calculated for each participant. A paired samples t-test showed that the *thinking times* between categories (M=1.96, SD=.69) were significantly larger than the *thinking times* within categories (M=1.29, SD=.82) $(t(8) = 3.11, p = .014, d = 1.04)$.

Performing the same analyses on *typing times* between (M=3.33, SD=.36) and within (M=3.27, SD=.63) categories showed no significant differences in means.

Both of these findings are in line with our findings in the environment question in study 1a, providing stronger support for accepting **H2**, that thinking times are shorter within categories and confirm our suspicion that typing times remain unaffected by generating ideas within or between categories.

The effect of half category switches on thinking and typing times

In the previous study we discussed the plausibility of ideators pivoting around central themes when ideating. Full category switches rely on discrete states of flow and impasse when in fact we may be pivoting around topics, such that ideators are letting the goal of one idea inspire another idea of the same goal but different means. Likewise, the means of one idea could inspire another of the same means but different goal. The analyses in study 1a showed that mean timings of half category (either mean or goal) switches were not significantly different from full category switches, suggesting that either ideators are not letting a central theme inspire other ideas or this was simply the effect of this specific pre-set category system. For this reason, we perform the same analysis on the data from the current health question study.

For these analyses, we divided the category coding done by IC3 into three different groups. Ideas with the same goal and mean as the preceding idea remained in the within category group as per the previous analysis. Ideas that shared either goal or mean, but not both, with the preceding idea was marked as a half category switch. Ideas that shared neither goal nor mean with the preceding idea were marked as full category switches. Log transformed timings were used again for this analysis, although the same patterns were seen in analyses on non-transformed data.

First, paired samples t-tests between full and half switches were performed on thinking times and typing times. The means of *thinking times* in full switches (M=2.01, SD=.70) and *thinking times* for half category switches (M=1.98, SD=.92) were not found to be significantly different $(t(8) = .033, p = .975, d = .01)$. As expected, the means of *typing times* in full switches (M=3.34, SD=.39) and *typing times* for half category switches (M=3.24, SD=.43) were also not found to be significantly different $(t(8) = .718, p = .493, d = .24)$.

Secondly, we performed paired samples t-tests to compare the difference in means of half switches and within category timings. The means of *thinking times* in half switches (M=1.98, SD=.92) were significantly higher than *thinking times* for within category (M=1.29, SD=.82) $(t(8) = 3.256, p = .012, d = 1.09)$. No significant difference was found in the means of *typing times* in half switches (M=3.24, SD=.43) and *typing times* for within categories (M=3.24, SD=.66) $(t(7) = -.058, p = .956, d = -.02)$. These results replicate our findings in study 1a. The typing time results were not surprising as these were not affected in the regular between and within time analyses. The significant difference between half switches and within category timings, but not full switches, may have been due to the fact that the half switches form a large subset of the original full switches data. As half-switches were not significantly different from full switches, half switches may well be treated as a transition to another patch in semantic memory, just as full switches. Again, we look at the *distribution* of between-category thinking times for some tentative insights into participants' strategies.

Figure 3-7 shows the untransformed distribution of *thinking times* identified to be between categories (again, note the > 20 overflow bin). The mean of thinking times identified to be within categories (M=9.37, SD=13.52) is highlighted on the graph. Again, the distribution showed that most between-category thinking times were considerably higher than mean within-category thinking time, but that there was a number of very fast between-category thinking times. This pattern mirrors that found by Payne, Duggan and Neth (2007) in a completely different task (a scrabble task) and supports their argument for a two-factor (failure versus success) based patch-leaving decision.

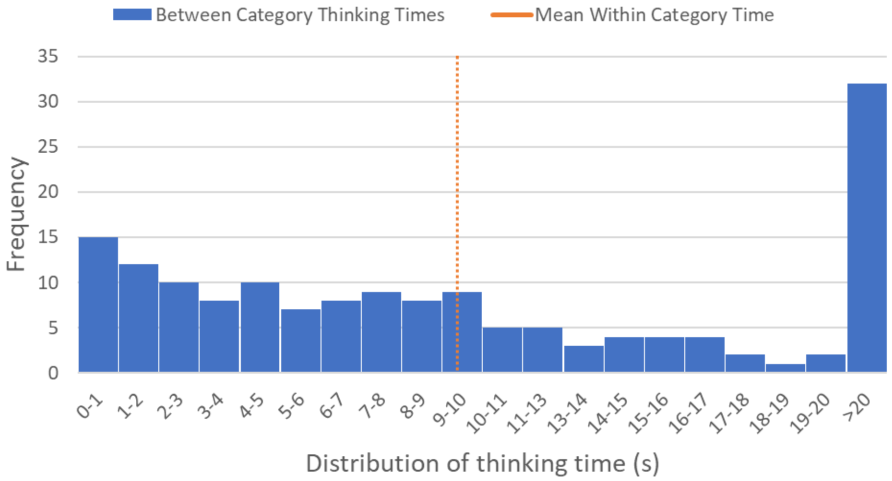


Figure 3-7: Distribution of between category thinking times health question study 1b, compared to mean within category thinking time (orange line). Note the > 20 overflow bin.

3.3.3 Summary and conclusions

Ideation performance

Participants' fluency was lower in this study, and there was stronger statistical evidence of ideas being harder to come by after 15 minutes of ideation. Again, the cumulative ideas curves (figures 3-5 and 3-6) show some longer thinking times that might indicate stuckness and short bursts that could be taken to indicate flow, but such mappings are highly speculative. This further highlights the effects of individual differences on ideation performance in relation to, for example, the threshold of what they might feel is an idea worth writing down. This is supported by the fact that some participants generated 14 ideas within a 15-minute timespan - suggesting a more leisurely pace of ideation than would be expected from a task in which ideators are asked to generate as many ideas as possible. Similarly, the lack of financial incentive might have had an impact on the productivity of ideators. This will be addressed further in study 2, Chapter 4, in which we provide financial incentives as a means of increasing productivity.

Post-experiment questionnaire responses

In this task, some of the responses to the post-experiment questionnaire were intercorrelated. Nevertheless, we were again unable to replicate the fluency-questionnaire correlation.

Evidence of clustering in idea generation

Agreement between coders was better in this study, as hoped. Despite this, coders raised problems with the scheme they were asked to use and the ideas they were asked to judge. Furthermore, and unfortunately, we were only able to replicate the finding of reliably positive ARC scores using IC3's coding (not IC4's). This could be due to the small number of participants, exacerbated by the fact that for one participant there were no inter-category ideas at all.

As in study 1a, analysis done on clusters identified by IC3 showed overall means of shorter thinking times within clusters than between clusters. Results like this suggest that if the identified clusters are to be likened to patches in memory, switching to another patch does indeed have an associated time-cost. This might be attributed to a person running out of ideas within an image and searching for a new image in memory from which further ideas can be generated. Again, there was no significant difference in thinking times between half and full category switches.

Limitations of this study: issues with the category structure and coder agreement

The two empirical studies in this chapter were a modest attempt to replicate the study and analyses of Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003) and Nijstad and Stroebe (2006). However, despite some encouraging reliable effects in the data, there were several failures to replicate the cleanliness of their data across two different questions. In particular coders didn't agree on category membership of ideas, and our participants didn't show as strong ARC scores. The lack of coder agreement and low ARC scores may not be unrelated, but they point to methodological issues for ideation research that we will address in later chapters of this thesis. In the next section of this chapter we attempt a different approach to the semantics of ideas that might allow more reliable coding and better exhibit participants' search strategies. Instead of categorising ideas according to pre-established categories, we inspect the semantic similarity between adjacent ideas. This approach can still capture a category or patch structure, and still allows questions about the time-course of idea generation within and between patches, but it replaces a binary flow-impasse analysis with one based on a continuous notion of semantic difference.

3.4 Semantic distance between adjacent ideas

In the two versions of the study we've just presented, we came across problems with the method of analysis chosen, despite this being a well-established and verified method. What started out being a simple replication study revealed that we were unable to fully replicate the effects that were expected from the study. We speculate that this could be attributed to a variety of reasons, such as fundamental problems with the pre-set categories themselves, the lack of well-formedness of the ideas generated and ultimately a lack of support for the SIAM model of searching for ideas in memory.

We try to move away from the concept of ideation states as being discrete such as flow and impasse, generating ideas in categories or switching categories. A continuous marginal rate of gain may be more in line with foraging theories. To that effect, this section is dedicated to how semantically related ideas are and what effect this has on the time-course of the ideation task. We explore this by calculating the *difference scores* between adjacent ideas. This was done to understand how closely related ideas were and whether this was a stronger method of mapping an idea set rather than expecting it to follow a category structure. This type of analysis was done by Chan et al. (2017), however, this was automated using Global Vectors for Word Representation (GloVe) (Pennington, Socher and Manning, 2014). They performed this analysis as verification that within-category ideas are semantically more closely related with between categories. They did not compare response latencies or thinking times of semantically similar and semantically different ideas as coded by GloVe.

The mapping in this section was done to try to understand whether the time-course of an ideation task could be explained by the semantic differences between ideas. Despite being primarily an exploratory analysis, we propose the following hypotheses:

H4: *Fluency* is negatively correlated with average difference scores, such that higher average difference scores are seen in smaller datasets

H5: Higher difference scores result in higher thinking times (and therefore response latencies) between subsequent ideas.

3.4.1 Method

Two independent coders were asked to rate the difference between adjacent ideas (e.g. ideas that follow each other) for each idea generated by each participant on a 10-point scale. This scale was chosen due to the lack of an obvious midpoint, acting as a method to make the coders think about their choice. One coder was male in his early thirties, one was female in her late twenties, both graduate students in computer science at the University of Bath, unaware of any of the research hypotheses. External coders were selected due to their lack of familiarity with the research goals. The coders were volunteers and received no payment for their participation. Although both coders were familiar with other methods of analysis, neither coder had experience using this approach before.

The first coder was given all 458 responses and the second coder was given a subset of 159 ideas (7 participants, approximately $1/3$ of the participant set). Coders were first explained what the participants had done during the study, then shown the set of ideas from the health question and the environment question separately. They were then given the following instructions, including examples, indicating how to apply the ratings (for exact wording of instructions see appendix A.4).

The following would score a 1 or 2, as they are essentially touching on the same concepts: recycling and plastic materials.

- Idea 1: "recycle plastic bottles"
- Idea 2: "recycle plastic bags"

These would score a 5-6 as they have something in common but aren't quite the same.

- Idea 1: "recycle plastic bottles"
- Idea 2: "teach people why recycling plastic bottles saves the planet"

The following would score a 10 as they have nothing in common:

- Idea 1: "recycle plastic bottles"
- Idea 2: "support an organisation that protects pandas"

Coders were asked to individually talk through the first 10 semantic difference ratings they performed, with the experimenter. Once the coders showed full understanding of the process, they were asked to complete the remainder of the difference ratings themselves. As difference score rating of the ideas was done on a scale, not categorically as in the previous study, inter-rater agreement was calculated using Pearson's correlation between the responses of the two coders.

3.4.2 Results and discussion

In the following section we present our results from our follow-up analysis of the data from both parts of our first study. We present this in four parts. The first presents the agreement between coders and the characteristics of their coding protocol. In the second part we look at the mean difference ratings for each idea set and correlate this with fluency to see whether generating semantically related ideas affects the amount of ideas an ideator can generate in a timebound task. In the third part we compare the difference scores and overall timings of ideas, *thinking time*, *typing time*, and *response latencies* in order to see if there is a link between difference between two adjacent ideas and the time it takes to generate ideas that are semantically near or far from the previous idea. Finally, we compare the timings of top different ideas and top similar ideas using t-tests see whether there is a significant difference in the means of these.

Coder agreements on difference ratings

Pearson correlation performed on the 152 difference scores revealed a high positive correlation between the ratings of the two coders for difference scores, $r(150) = .715, p < .01$. This is a much more promising result than in our category analyses, as it means we are more confident in the conclusions we draw from these analyses. Figure 3-8 shows the mean difference ratings submitted by the second coder compared to the ratings submitted by the main coder. This was calculated by grouping all the ratings of the second coder according to the main coder's ratings.

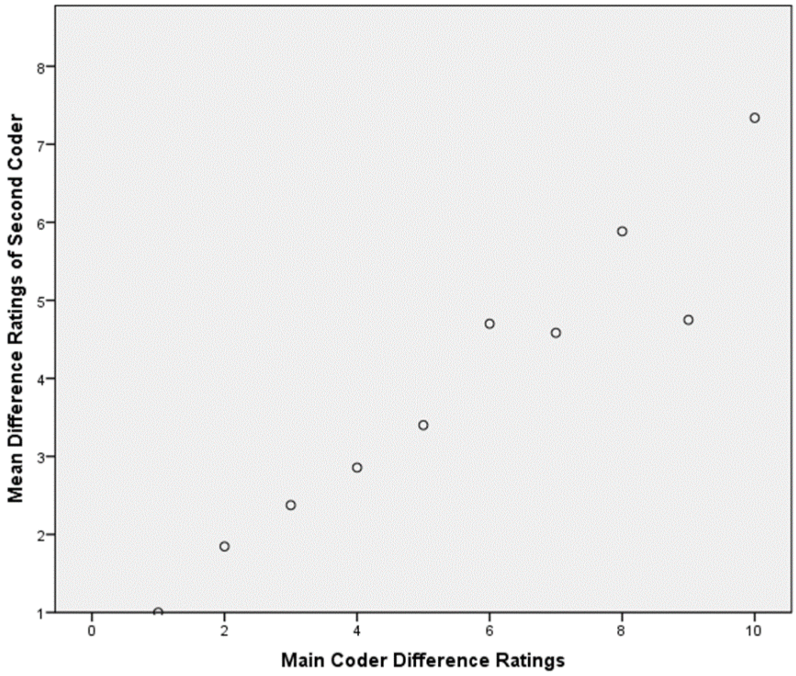


Figure 3-8: Summary correlation of difference ratings for study 1a and 1b: second coder mean difference ratings against each of the main coder difference ratings on the 10-point scale.

The coders seemed to use the 10-point scale differently. The main coder seemed very inclined to use a rating of 10 frequently (see distribution in figure 3-9). Although not

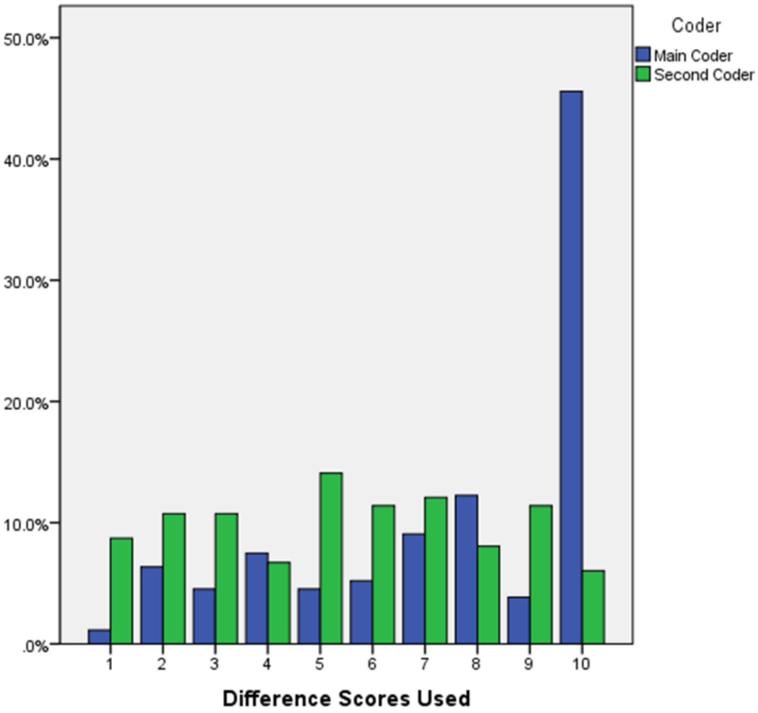


Figure 3-9: Distribution of ratings for main coder (438 datapoints) and second coder (152 datapoints) for studies 1a and 1b.

ideal, this supports the finding in the first two parts of this chapter, in which a very high number of category switches were identified in comparison to a very small number of within category ideas. The second coder used mostly the full 10-point scale creating a fairly even distribution between the 10 points. Note that this was not attributed to the sample itself, as scores from the main coder from the same sample gave a similar distribution as the main coder had for the whole set.

Descriptive data

In the following section, we present analyses of $N - 1$ datapoints per participant, where N is the number of ideas in the participant's dataset, or *fluency*. Difference scores were given between two adjacent ideas, and not for every idea. The first idea of a dataset did therefore not have an associated difference score.

Table 3.13 shows the mean difference scores compared to mean fluency for each participant in the two parts of the study. Pearson correlation between mean difference score and fluency is $r(18) = -.69$. Figure 3-10 shows this negative correlation between mean difference scores of an idea-set and the *fluency* of that idea-set. This result evidences that there is a negative correlation between difference scores and the number of ideas generated, allowing us to accept **H4**. We also computed the correlation between fluency and variance of the semantic difference score. We suppose that variance is a kind of proxy for semantic structure, as it will arise when some adjacent ideas are semantically close (within patch) and others are distant (between patch). Indeed there was a positive correlation between variance of semantic difference and fluency, $r(18) = .69, p < .01$.

Difference score correlations with time-course patterns

In the next section we compare the difference scores to the time span across individual ideas. In total, we had 438 difference scores between ideas developed by 20 participants across the two parts of the study. We compared thinking times, which consist of the time between the two ideas that have been given a similarity score, the typing time of the second idea and the overall response latency of the second idea in the difference score set.

When looking at the timings of ideas, i.e. thinking times, typing times and response latencies, we are only looking at ideas following the first idea. As the first idea cannot be compared to anything before it, it is not possible to include this in the analyses of course.

We calculated the correlation between difference scores and the three timing metrics for each participant (see table 3.14). We used these to see if there was evidence of a positive correlation between the time taken to generate an idea and its difference from the previous

Study 1a			Study 2a		
Participant	Fluency	Mean Difference Score	Participant	Fluency	Mean Difference Score
1A	30	6.55 (3.16)	1B	30	7.38 (3.00)
2A	20	7.11 (3.14)	2B	19	8.44 (2.11)
3A	22	6.52 (2.89)	3B	14	9.38 (1.15)
4A	21	8.85 (1.42)	4B	15	9.00 (1.69)
5A	27	6.54 (2.95)	5B	20	8.21 (2.07)
6A	41	7.33 (2.89)	6B	17	8.31 (2.26)
7A	15	8.07 (1.98)	7B	16	8.53 (2.16)
8A	24	8.17 (2.20)	8B	21	8.70 (1.55)
9A	37	6.56 (3.28)	9B	32	7.61 (2.36)
10A	23	7.27 (2.88)	10B	14	9.46 (1.39)
M (10)	*26*	*7.3*		*19.8*	*8.5*
SD (10)	*(7.58)*	*(0.78)*		*(6.06)*	*(0.65)*

Table 3.13: Mean difference scores per participant compared to their overall *fluency* for data from studies 1a and 1b. Standard deviations shown in brackets.

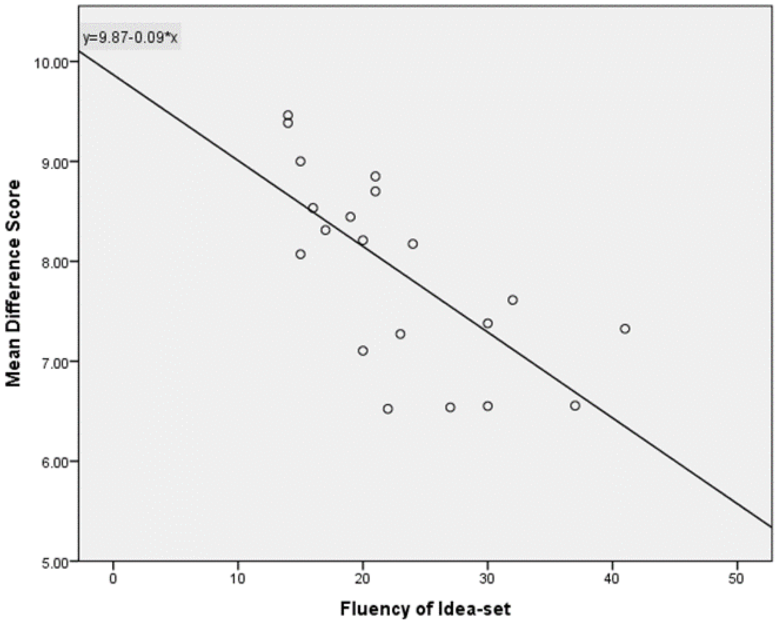


Figure 3-10: Correlation between mean difference scores and the fluency of an idea set; semantic difference scores calculated in data from studies 1a and 1b.

idea, in order to answer **H5**. *Thinking time* correlations ranged from -0.26 to 0.48. Overall, the average of these (M=0.19, SD=0.17) was positive and significantly different from zero $(t(19) = 4.933, p < .001, d = 1.12)$.

Typing time correlations ranged from -0.31 to 0.61. Overall, the average of these (M=0.04, SD=0.23) was positive but not significantly different from zero $(t(19) = .768, p = .452, d = 0.17)$. *Response latency* correlations ranged from -0.28 to 0.61. Overall, the average of these (M=0.17, SD=0.19) was positive and significantly different from zero $(t(19) = 3.746, p = .001, d = 0.89)$.

These positive correlations give us evidence that producing semantically different ideas will take longer than semantically similar ideas, allowing us to accept **H4** and **H5**. Correlations between typing times and difference scores is again not surprising and re-enforces the result found in the category coding sections as well. The positive correlations between thinking times and difference scores suggest that generating ideas does in fact follow a patch like structure in memory.

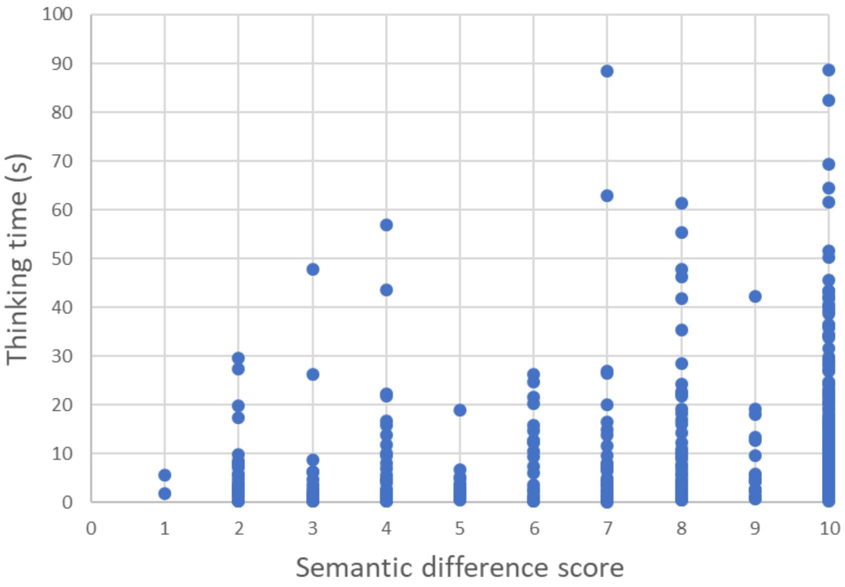


Figure 3-11: Thinking time as a function of semantic difference scores between adjacent items; semantic difference scores calculated in data from studies 1a and 1b.

Figure 3-11 shows all data across all participants. The overall scatter plot on which every thinking time is plotted against its semantic difference score is of particular interest. It shows that the increase in thinking times with semantic difference is due to some elongated thinking times, and a greater spread in the distribution at greater semantic differences.

	Correlation between difference score and timing		
Participant	Thinking Time	Typing Time	Response Latency
1A	0.36	-0.18	0.06
2A	0.23	0.26	0.28
3A	0.18	-0.17	-0.01
4A	0.11	0.61	0.61
5A	0.16	0.06	0.17
6A	0.39	-0.31	0.21
7A	0.48	0.19	0.36
8A	0.19	0.31	0.32
9A	0.19	0.11	0.20
10A	0.26	-0.18	0.06
1B	0.17	0.19	0.28
2B	0.27	0.17	0.34
3B	-0.24	-0.22	-0.28
4B	0.21	-0.26	0.05
5B	0.30	-0.24	0.00
6B	0.30	0.05	0.09
7B	0.24	0.06	0.20
8B	-0.26	-0.05	-0.15
9B	0.14	0.17	0.23
10B	0.21	0.25	0.27
Mean (20)	*0.19*	*0.04*	*0.17*
SD (20)	*0.17*	*0.23*	*0.19*

Table 3.14: Correlation between mean semantic difference scores and mean ideation timing metrics per participant for data from studies 1a and 1b.

Even at the most extreme semantic difference there are some very quick thinking times. Informally, this again suggests multiple processes at work – allowing ideas to come to mind independently of semantic similarity, and/or allowing participants to deliberately shift focus very soon after an idea has been generated, rather than only in the case of a long idle interval.

Differences in means timings of highly similar and highly different ideas

In addition to the correlations between difference scores and thinking times, typing times, and response latencies, we compared the timings of ideas that were rated with the highest similarity and ideas rated with the highest difference. Data was calculated as averages per participant and then compared using paired samples t-tests. The data was split into the following groups:

1. Average thinking times, typing times and response latencies of any idea coded as having a difference score of 1-2 (high similarity)
2. Average thinking times, typing times and response latencies of any idea coded as having a difference score of 9-10 (high difference)

Paired samples t-tests conducted on *thinking times* in the high similarity (M=5.79, SD=5.18) and high difference (M=12.09, SD=5.25) conditions showed a significant difference in the two groups $t(10) = 3.72, p = .004, d = 1.12$.

Paired samples t-tests conducted on *typing times* in the high similarity (M=19.63, SD=10.53) and high difference (M=23.68, SD=12.14) conditions unsurprisingly showed no significant difference in the two groups $t(10) = .997, p = .342, d = .30$.

Paired samples t-tests conducted on *response latencies* in the high similarity (M=23.30, SD=12.95) and high difference (M=35.74, SD=9.58) conditions showed a significant difference in the two groups $t(11) = 3.47, p = .005, d = 1.00$.

The same analyses were performed on an even split down the middle of the data based on difference scores, resulting in the following two groups:

1. Average thinking times, typing times and response latencies of any idea coded as having a difference score of 1-5, bottom half difference.
2. Average thinking times, typing times and response latencies of any idea coded as having a difference score of 6-10, top half difference.

Paired samples t-tests conducted on *thinking times* in the bottom half difference (M=5.11, SD=3.12) and top half difference (M=11.93, SD=6.24) conditions showed a significant difference in the two groups $t(17) = 4.86, p = .000, d = 1.15.$

Paired samples t-tests conducted on *typing times* in the bottom half difference (M=24.08, SD=11.79) and top half difference (M=30.96, SD=13.04) conditions showed no significant difference in the two groups $t(17) = 1.91, p = .073, d = .45.$

Paired samples t-tests conducted on *response latencies* in the bottom half difference (M=29.19, SD=11.49) and top half difference (M=42.89, SD=15.03) conditions showed a significant difference in the two groups $t(17) = 3.28, p = .004, d = .77.$

The results further support our hypotheses that *thinking times* are affect by the difference between adjacent ideas (**H5**) and that *typing times* remain unaffected.

3.4.3 Summary and conclusions

For semantic difference scores, the inter-rater agreement was within an acceptable range. Coding difference ratings seems to be a more viable way of approaching data than coding into pre-set categories. In particular, with data such as ours where there is a varied level of generalisation and specificity across the ideas, making it difficult at times to discern what it is the ideator meant specifically. Looking at the distribution of scores, it is clear to see that the main coder felt willing to view ideas as semantically completely unrelated, seen in the high number of 10-scores used. The lack of certain confidence in the scale might be assumed from looking at the low number of 1-scores given by the coder. The overall average difference scores rated by our coders are in general quite high, showing that ideators are indeed jumping between fairly different ideas. Despite this, we have still seen interesting patterns in the data. The negative correlation between fluency and the average difference ratings between ideas is a promising result as it shows that ideas are thought of as distinctly different rather than following a train of thought in order to develop better ideas. Similarly, it seems that smaller datasets do not conform to SIAM. This was difficult to see in the initial analysis as calculating ARC scores of a small dataset proved problematic. It can be assumed here that high difference scores are representative of distinct category switches in which ideators try to think of ideas that are distinctly different from previous ideas. This behaviour is not in line with the instructions which stated that the ideator must come up with many wild ideas, and not judge them as they go along. Despite instructions, participants show that they're still more interested in generating distinct "good" ideas rather than type out their stream of consciousness. Again, this highlights how individual differences might affect ideation performance; instead of "simply" ideating, some people may impose thresholds on what they feel is worth writing down and employ not only creative but critical thinking.

Indeed, we might expect that all participants in these studies display varying tendencies and abilities to shift between creative and critical thinking. In order to overcome issues of varying motivation to generate more ideas rather than self-evaluating ideas, we offer a financial incentive to generation of many ideas in our next study.

The positive, albeit small, reliable correlations shown between difference scores and all timings suggest that it does indeed take longer to think of an idea that has a higher difference from the previous idea than not.

Splitting the data into highest similarity idea timings and highest difference idea timings showed that more similar ideas have significantly lower thinking times and response latencies, suggesting a faster development of ideas that follow on from similar ideas. This confirms a prediction of the SIAM theory, but suggests, perhaps, that rather than ideas being structured according to a series of discrete images in memory, they are generated by associations with already active ideas.

3.5 Chapter Summary

In this chapter we presented the results of two studies, with two separate ideation problems, developed to replicate the findings of Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003), and Nijstad and Stroebe (2006). We performed these studies in order to verify that ideas are generated in a categorical way, and to test the distribution of typing times and thinking times within and between categories. This was done with the goal of building a foundation for the understanding of impasse and flow states in terms of foraging theory, and the eventual plan of supporting ideators to escape from impasses. We have reported problems with the methodology of categorising ideas into these pre-set semantic categories developed by Diehl (1991). There were two separate sources for these problems, presumably, as indicated in the coders' comments. One source was the categorisation schemes – these were those used successfully by Nijstad and Stroebe (2006), but might no longer be optimal. Nijstad and Stroebe (2006) note that (even) their range of levels of clustering (.20-.41) was fairly low in comparison to clustering found in free recall tasks (.60-.70 in Basden et al. (1997)). They attributed this to the category system itself. The second source of difficulty for category coding was the ideas themselves, as expressed by the participants. Ideation tasks as usually studied, and as studied in this thesis, until the final study, are very unconstrained, so that ideas can be expressed at various levels of generalisation/specificity. This issue was also raised in the coders' comments. We will return to this in Chapter 5. A key limitation to highlight here is the use of non-professional coders. The main researchers originally performed the coding, however, concerns were raised about a possible bias towards categorising vague ideas based on the previous idea, due to the knowledge of the research hypotheses. The decision was therefore made to bring in external coders. The lack of access to professional coders and financial means meant that the coders involved were volunteer coders, albeit all familiar with data analysis. It was ensured that these received instructions and training by going through a part of the dataset with the main researcher.

Despite these difficulties with category coding, we have reported reliable indices of category structure, and reliable effects on ideation times. We have found that, overall, within and between category times vary significantly. Nevertheless, we do not place enough confidence in these data to go beyond some speculations regarding specific foraging stopping rules.

Arguably, the main success of this chapter and the main novel contribution is an alternative approach to semantically coding ideas, not in terms of pre-existing categories but in terms of semantic differences between adjacent ideas. Whilst semantic differences have been calculated automatically in other studies, it has been done so in order to verify that ideas within groups are closely related (Chan et al., 2017). We tested the effect of the difference between two ideas on the thinking times between those ideas and reported a reliable positive correlation between thinking times and difference scores. This result could, speculatively, be couched in relation to the SIAM model, which postulates that ideas are generated as associations made to a particular image retrieved from long-term memory. It seems possible that generating associations with a retrieved image might not produce well-structured, perfectly clustered ideas in categories but nonetheless have a semantic relation to the current image being looked at. As a result of this finding, it was decided not to update the category scheme. The possibility of an alternative approach, namely semantic distance between ideas, led to the decision to move on from the use of categorisation schemes as a way of analysing the time-course of a single question ideation task. Note that we do revisit the health category scheme in Chapter 5, after performing an instructional manipulation for ideators, in order to assess whether our manipulation impacts coder agreement.

Our original motivation for these studies was to look at the boundaries between category switches in order to see if we could explain idea generation using information foraging theories. The weak categorisation results inhibited this approach. In the next study we therefore explore a different type of ideation task, in which ideators are given a practically unlimited number of questions to respond to in a time-bound setting. This approach allows us to treat the objective sequence of ideation problems as a sequence of patches, and to investigate the rules that ideators use to switch from one problem to the next.

Chapter 4

Time-course of ideation across multiple ideation tasks

In the previous chapter, we looked at the time span of ideation across a single ideation task. The category scheme available for each of the questions presented to participants, was identified as problematic, resulting in difficulty categorising ideas, yielding a fairly low to moderate agreement between coders. Comparing this to the study being replicated (Nijstad and Stroebe, 2006), which had Cohen's k agreement scores of 0.7-0.9, we questioned the usefulness of this method for our purposes. Coders indicated they felt they had to assume what the ideator meant as some ideas were vague and difficult to assess in terms of the given category structure. The ARC scores and within- and between-category switches did show interesting patterns in the environment question, however, this was not so apparent in the health question.

More positively, we found evidence of a small correlation between the judged semantic difference between two adjacent ideas (an idea pair) and the time it takes to formulate the second idea in the given idea pair (response latency). Likewise, t-tests showed that our split of thinking and typing times was valuable as there was a significant difference in thinking times for highly different and highly similar ideas, but not for typing time, indicative of the ideator stopping to formulate an idea in their head before writing it down. The higher time taken to formulate highly different ideas additionally seems to support SIAM, in that it can be explained by the concept of moving to a different image in long term memory. Nevertheless, continuously variable semantic difference ratings do not allow us to test the heuristics that derive from foraging theory concerning how an ideator might choose to abandon one semantic category in favour of another.

Therefore, we developed a new study that looks at how ideas are generated when the user has the freedom to abandon a question for another at their own will. Category switches were originally considered to represent the switch from one patch of information to another in semantic memory, likened to information foraging. The switch from one question to another (unknown before the switch) was in the current study used as a stronger indicator of giving up and starting information foraging in a new patch. This method does not require classification based on a pre-determined set of categories as switches are explicitly seen when the ideator chooses to stop one task and begin another.

This task-sequence design had similar goals to the first study: we were looking for patterns in the data that could suggest the strategies ideators use and whether they are switching to new tasks in such a way that they maximise productivity. In the following section, we revisit the theories discussed in Chapter 3 as well as further theories that relate to this study. The main aim of this study was to use a discretionary task-switching paradigm to investigate the time-course of ideation and the decision making of ideators. The study can also be seen as a contribution to the work on discretionary giving-up, and to the understanding of multi-tasking in general, because it extends an experimental paradigm that has most typically been used on simple tasks or highly constrained problem solving tasks to a highly creative task.

4.1 Chapter background

In our first study, we were particularly interested in the concepts of *flow* (Csikszentmihalyi, 1997) and *impasse* (e.g. Nijstad and Stroebe, 2006; Chan et al., 2017; Siangliulue et al., 2015). Impasse was defined as a failure to produce new ideas and is described in the literature as being a cause of longer idle times causing ideators to request inspiration. We posed singular questions in our first study and categorised the responses according to a pre-determined set of categories. The timings between and within category clusters were an attempt to address whether ideators might switch topics if stuck for ideas.

In the current study, we wanted to observe more explicit switches which could be analysed to better understand strategies employed to give up one patch so as to switch to another. As we were looking at switches made by the ideator themselves, we suppose that they might be more aware of the strategies they're using in this study than in our previous single task study. Rather than simply looking at impasse we refer further to classic stopping rules and whether these can be used to explain the behaviour of ideators when given the freedom to switch (**RO3**).

The classic switch decisions can be found in the foraging theory sub-chapter of the literature review. Charnov's Marginal Value Theorem (MVT) (Charnov et al., 1976) is a proposed optimal foraging theory in which the forager must use the momentary (or *marginal*) rate of return of the current patch and mean rate of return for all patches visited as a method of estimating if it is worth remaining on the current patch. Calculating the optimal time to leave the current patch involves taking these rates of return into account as well as the time it takes to travel to a new patch, Charnov's Marginal Value Theorem imposes a very a high cognitive load on the forager. Ware et al. (2016) have shown that during performance of tasks, strategies employed by problem solvers become less optimal the more demanding a task is. One simple prediction of Charnov's Marginal Value Theorem is that patches will be left sooner if the travel time to the patch (when no successes are achieved by definition) is shorter. To test this prediction, a delay was imposed every time a participant quit one ideation task to move to another (see Wilke et al. (2009) for a similar manipulation). In a between-groups manipulation, this delay was set to 10 seconds for half the participants and 25 seconds for the remainder.

More simplistic strategies include fixed-number and fixed-time rules (Iwasa, Higashi and Yamamura, 1981), in which a patch is left based on a predetermined set time or number of items have been proposed, because they might approximate the optimal strategy under certain environmental conditions. For ideation, this would mean choosing to give up after a set amount of time, or after generating a set number of ideas. Another quite simplistic strategy is the rule which uses a threshold of *giving-up time*. The giving-up time rule solely takes into account the time since the last item found, regardless of how many items have been found before. This rule is based on a giving-up time threshold set by the forager, and if the time since the last item found increases above this threshold, the patch is abandoned. This rule leads to foragers spending longer in patches where successes are more frequent. Another heuristic that adapts to the experienced yield of a patch is Green's rule (Green, 1984). Green's rule states that a forager may be estimating the potential of the patch they are in, using a combination of the number of items found and the time taken between them. The potential decreases as a function of time until it drops below a threshold at which the patch is abandoned for another. Finding an item is represented by a positive increase in the perceived potential. Green's rule allows a forager to monitor the current rate of gain of a patch, simply by keeping track of the number of successes. Looking at strategies for making the decision to switch to a new question, we revisit the studies on free recall and word search. Multiple tasks in which a person can abandon one for another has been done in more rote tasks such as recall and anagram tasks (Payne, Duggan and Neth, 2007; Wilke et al., 2009) but, to our knowledge, not in ideation tasks. These tasks are different to ideation in the sense that they align more with accomplishment rather than accumulation tasks - the responses will be either correct or incorrect and there is a finite number of responses. Note that in this section we use 'decision to switch' and 'giving-up' interchangeable, as a decision to switch is essentially a decision to give up on the current question. In Harbison et al. (2009), they explore 4 different rules for terminating search in memory for word lists: (1) fixed-time, (2) time since last successful retrieval of a word (giving-up time), (3) number of successive failures to retrieve a word and (4) when the rate of retrieval has reached the participant's own accepted threshold (Green's rule). Time on task and time since last idea generated are two plausible rules for ending an ideation task. Presumably so would the rate of return threshold rule be as it might well be a signal to the ideator that their productivity on the current task is dropping. Translating rule (3), number of successive failures, to an ideation task may be difficult, as the word *failures* needs to be redefined. In achievement tasks, measuring a failure is straight forward as this consists of failing to give a correct response or failing to give any response when prompted to remember an item. In ideation tasks, classifying a response as a failure is less straight forward as there are no correct or incorrect ideas. Recall that in Nijstad and Stroebe (2006) failures are defined as the inability to continue to come up with new ideas, the same as impasse. A method in which an ideator could be aware of multiple failures would be a situation in which they fail to generate ideas, or generate new ideas.

In Payne, Duggan and Neth (2007), an anagram task is used in order to see what type of strategies people use to maximise productivity. The tasks given to participants are in this study split into easy and hard tasks. When given the chance to switch between tasks freely, participants were found to spend longer on the harder tasks with lower rates of return, rather than easier tasks with a higher rate of return and therefore a better score. Giving-up times were longer in the harder tasks. In addition to this, they found that switching behaviour could not simply be explained by any one of the above-mentioned rules, but by a composite of Green's rule and a new parameter: the probability of switching immediately after finding an item (*sub-goal completion*). In the anagram study, the experimenters allowed the use of task interleaving, letting participants switch back and forth between an easy and a hard set of letters. In the present study, as we are recording the thinking time and typing time of each idea, we have to use a different method in order to assess whether participants are spending their time optimally. If participants would be allowed to switch back and forth between tasks, this would be a different type of study entirely as they may let the question sit in the back of their mind whilst performing another task, and then returned once they could think of more ideas[1]. In our study we were mostly interested in the task abandonment

[1] A concept known as incubation. Interestingly, incubation is sometimes considered a key part of creative problem solving. For the purpose of this study, we avoid the introduction of an incubation period as this would interfere with our goal of observing flow-impasse states and the use of giving-up strategies.

paradigm in which, when a participant moves on from a task, they are essentially giving up.

In Wilke et al. (2009), a similar anagram study was performed in which they liken the word search task to foraging in the physical world. They introduced a switching cost to their anagram task: it would cost 25 seconds to switch from one task to another. They did this in order to compare the results to an electronic 'fishing' task in which switching to another pond took 25 seconds. They did not compare a 25 second switch cost to any other delay time. In their study, they allowed only giving up on a task, not switching back and forth between tasks as in Payne, Duggan and Neth (2007). They similarly found that the more simplistic giving-up rules could not explain the participant foraging behaviour in their study, however, they did find evidence that the time since last item found (and time since the item found before that) seemed to be the main indicator involved in the decision to switch. They explain that in most cases, there did not seem to be a threshold time after which participants would switch, and therefore a simple giving-up time rule would be too simplistic to explain the decision to give up.

The above strategies for giving up and maximising productivity provide us with a basis of possible patterns to look for in our current study, in order to determine how people allocate their time in a multiple problem ideation task. People don't know the gain curve of a task prior to starting it and so will need to start the task before they can form an understanding of the gain curve. Taking into consideration the constantly changing gain curve, we were interested in finding out how people solve the time allocation problem and how well they do this. The ideation study in this chapter was run with ideators individually, performing a multiple question ideation task by typing out as many solutions as possible to a set of 17 questions. No participant managed to complete all 17 questions before time was up. They could freely skip to another question, however, would either have a 10 second delay or a 25 second delay. The analysis of the data looks at timings of ideas and responses to a post-experiment questionnaire in the same way as in the first study. We additionally look at giving-up times as well as propose a method of estimating how difficult a participant found each individual question.

4.2 Study 2: Ideation strategy across multiple consecutive tasks

In this section, we present our second experimental study, aimed at understanding switch strategies in ideation and whether people use simple switching strategies such as fixed-time or fixed-number rules, or more complex switching strategies such as Green's rule (Green, 1984). This study looks at the span of ideas across multiple ideation tasks and allows the participant to switch to another question should they feel stuck on the current one. The study consists of 17 open-ended questions designed to elicit a range of ideas from participants in a variety of topics (see appendix B.1). A high number of questions were developed in order to provide a seemingly infinite pool of tasks, allowing participants to switch as often as they felt. No participant managed to get through all 17 ideas in 30 minutes.

4.2.1 Method

Design

A laboratory experiment was conducted following a broadly similar paradigm to that used in the previous study in this thesis. However, each participant was given 30 minutes[2] to generate as many ideas as possible across a series of ideation tasks. They were given the discretion to stay ideating on a single task or to jump to another (unknown) task. To test the elementary foraging theory prediction that travel times between patches will influence patch leaving decisions, participants were assigned to one of two conditions, the switch cost conditions: (1) 10 second delay between tasks or (2) 25 second delay between tasks. To avoid order effects, the tasks were randomised at the start of each session and participants were presented the tasks in a random order. Tasks were shown in serial and participants were not allowed to interleave between them – once a task had been abandoned, participants moved on to the next one and would not be able to go back to a previously abandoned task.

The questions were chosen as appropriate tasks for this study as they were open-ended questions that allowed ideators to think around the topic without requiring any real subject knowledge. Whilst we did show that there is a difference in the time it takes between and within groups of semantically related ideas, having multiple questions presents us with a stronger split, determined by participants' move from one question to another rather than being based on the categorisation done by an independent coder, which might suffer from being subjective rather than objective.

We were interested in the patterns in response latencies and whether these could predict giving up the current question and switching to the next. In particular, we were looking to see if we could classify this pattern as a simple fixed-time or fixed-number pattern or whether there was any evidence of running out of ideas (slower response latencies) before

[2] The 30 minute time limit was a fairly arbitrary decision as there was no similar ideation study to compare to. For other types of accomplishment task we find varying lengths, such as Payne, Duggan and Neth's (2007) 15-minute Scrabble task, or Wilke et al.'s (2009) 60-minute anagram task.

a switch is made, in which case a giving-up time heuristic or a rate-sensitive heuristic like Green's rule might be a better model of participants' switch decisions. We compare *thinking times* and *giving-up times* to see if there are any time threshold parameters involved and look for any other unexpected patterns in the data. We propose the following hypotheses:

H1: When the time-cost of switching tasks is higher, participants will spend longer on each task before switching.

H2: We do not expect participants' switch decisions to be well explained by time on task, number of items generated, or giving-up times.

H3: We predict that switch decisions will exhibit dual processes, with some switches almost immediately after an idea has been generated, but most after a considerable giving-up time longer than the mean between-item time.

Additionally, we look at whether there is any evidence to suggest that people spend longer on more difficult questions rather than those in which they are more productive (i.e. generating more ideas). In an ideation task like this, despite the use of questions that should not require specific subject knowledge, it is to be expected that participants' individual interests might have an effect on their motivation and ability to generate ideas for a given question. We therefore expect that objective difficulty of a question cannot be determined, but subjective difficulty can. We estimate subjective difficulty using a range of different metrics in order to assess whether participants indeed do spend longer on more difficult tasks.

Participants

30 participants (10 female), age range 19 to 53 (M=23.93, SD=6.51), were recruited from around the University of Bath, consisting of undergraduate and postgraduate students, as well as academic and administrative staff. Participants were recruited from a variety of disciplines. They were recruited by mailing lists and posters presented around campus. All participants were inexperienced ideators, that is, none of the participants performed ideation for a living, nor were they previously trained in ideation methods. No further profiling constraints were applied to recruitment for this study, due to the difficulty in sourcing participants, despite being a paid study. In study 1, we discussed the possibility that the low number of ideas generated by some participants could be due the lack of financial incentive as a way of motivating the generation of many ideas rather than self-evaluation and restraint in writing down ideas. We address this in the current study by offering a reward dependent on productivity. A reward was given for participation – 20p per coherent idea formed during the experiment (capped at £10) paid within a week of

the study. This was included as an incentive to maximise productivity, as more ideas would result in a higher reward. Additionally, participants were entered in a prize draw for two amazon.co.uk vouchers of £25 each. This study was within the ethical bounds of the Department of Computer Science at the University of Bath and every participant gave full informed consent.

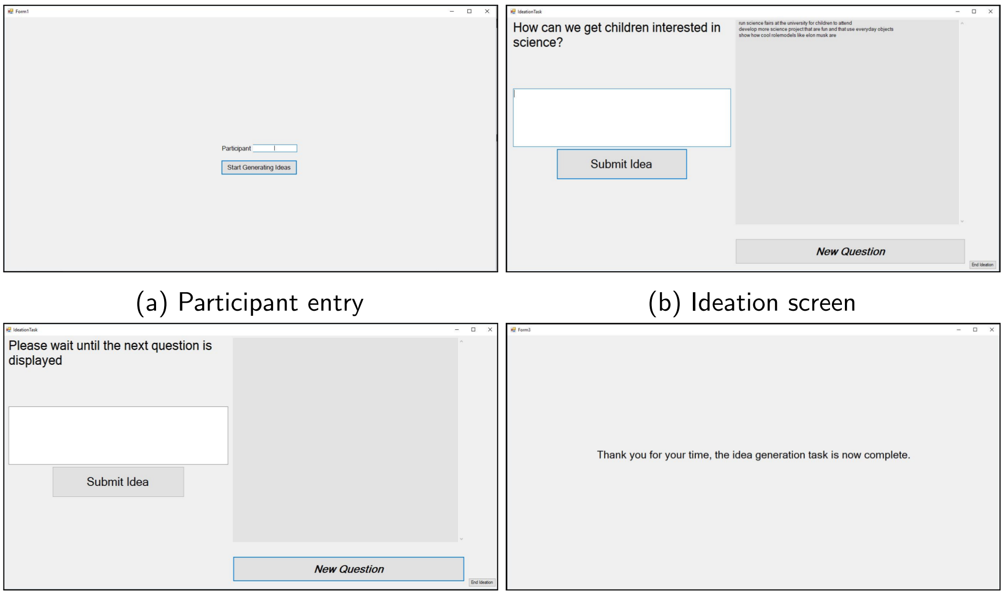
Materials

The application from our first study was adapted for this study using Visual Studio 2015. The application was edited to consist of 4 screens (see figure 4-1). The first screen (figure 4-1a) takes in the participant number as input and uses this to set up the data file that captures the participants input. When the "Start Generating Ideas" button is pressed, the application randomises the 17 questions in the array using the Fisher-Yates shuffle and saves these back into the array in their new order.

The second screen (figure 4-1b) is the main ideation input screen, split into two main parts. On the left it displays the first ideation question at the top and contains a simple input field and a submit button saying: "Submit Idea". On the right is a large empty text box in which submitted ideas are displayed. As in the single task study, the text box is set to scroll when more ideas are submitted than can be displayed. Under the text box is the "New Question" button, which participants were instructed to press should they want to abandon the current task for another question.

If the "New Question" button is pressed, the third screen is shown (figure 4-1c). This is the delay screen - it is presented exactly as the second screen, however, it is not possible to enter any ideas. All ideas submitted for the previous question are saved to the output file and on-screen text box is cleared, ready for the next question. The ideation question is replaced with "Please wait until the next question is displayed". Participants were made aware of which delay condition they were in, 10 seconds or 25 seconds, prior to starting the study. The third screen will be displayed for the specified delay time. Once the delay time has passed, the input screen is shown again, displaying the next question in the array.

The input form automatically shuts down after 30 minutes, taking the participant to the fourth screen (figure 4-1d), which thanks them for their time and ends the ideation part of the study. All data are captured by the environment automatically. Ideas entered, the question they are associated with, as well as the timestamps for the start of an idea and the time enter is pressed (i.e idea submitted) are captured.



How can we get children interested in science?

run science fairs at the university for children to attend
develop more science project that are fun and that use everyday objects
show how cool rolemodels like elon musk are

Please wait until the next question is displayed

Thank you for your time, the idea generation task is now complete.

(a) Participant entry

(b) Ideation screen

(c) Question switching delay screen

(d) End of experiment screen

Figure 4-1: Developed app for study 2: multiple sequential ideation.

The delay time condition was selected for each participant according their order of participation, and had to be adjusted manually. The software was run on a Lenovo Yoga 710-14ISK laptop running windows 10. To ease input, a 22" screen, wired mouse and keyboard were attached to the laptop.

Procedure

Each participant attended the study individually in a quiet room situated in the Computer Science department at the University of Bath. This was done to avoid any distractions or any items that could act as a prime for certain types of idea. The participants were handed an information sheet to read and had the chance to ask questions before the study started (see appendix B.2). They were explicitly instructed to come up with ideas using the same modified version of Osborn's (1957) brainstorming rules used in the first study. As in the first study, they were asked to expand their ideas to incorporate specific behaviours; instead of saying "manage electricity", they were asked to indicate how they propose you could do so, e.g. "don't overfill a kettle if you are only making a single cup of tea". They were explained that they could submit the same idea twice should they wish to expand and elaborate.

Once consent forms had been signed (appendix B.3), participants were situated in front of the computer and talked through the environment. The three-minute practice task was started and participants were asked to come up with as many ideas as they could in response to "How can the number of tourists visiting the city of Bath be increased?"

Once participants felt comfortable with the environment, the real experiment application was started, and participants were left alone in the room to generate as many ideas across as many or few questions as they decided suitable in the 30 minutes. Participants were aware of both the 30-minute time constraint and of the delay condition to which they were assigned. No clock was visible so as to avoid distraction. Once the ideation task was complete, participants were given a short questionnaire developed specifically for this study (for full questionnaire, see appendix B.4). The first question asked them to comment on their strategy and what cues they used to decide to give up on a question. This first open-ended question was asked before the likert-scale questions, in order to avoid any priming effects from the wording of the scale questions, allowing the participant to formulate their own strategy wording before seeing other possible strategies. They were then asked to respond on a scale of 1-10, how often they used the following cues to switch: number of ideas, time spent on a question, time interval since previous idea generated and overall feeling of stuckness. They were then asked to rate, on a scale from 1-10, their level of agreement to the following questions:

- "I only pressed 'next question' when I felt stuck",
- "During a question, I continuously switched between the feelings of being stuck and un-stuck",
- "It was very difficult to think of new ideas" and
- "The same idea occurred to me several times."

At the end of the study, the participant was debriefed and thanked for their time. Within a week of participating, the number of unique ideas were counted, and the participant was compensated (up to a value of £10) for the number of ideas generated.

General analytic approach

With a goal of understanding giving-up decisions and how people allocate their time when ideating, we analysed the data with a focus on which strategies were used to maximise productivity. Participants had an extrinsic incentive to generate more ideas, as this would result in receiving a higher financial reward. As in our first study, the timings were automatically recorded by the experiment application. Note that the last ideation task in some of the datasets was not ended by the participant but by the application itself as the 30 minutes had elapsed and the participant ran out of time. These final tasks were removed from the data before analysis as they did not have a giving-up time associated with them and therefore could not be included in the analysis comparing giving-up times to between idea thinking times. This left us with 1264 ideas generated by 30 participants (M=42.10, SD=15.34) across all 17 questions.

In this chapter we use the same timings as in our previous study, but we introduce a new dependent variable, the *giving-up time*. Timings of the ideas generated are defined as follows:

- *First keystroke*: the first time a key is pressed, indicating that a letter has been entered in the answer area.
- *Thinking times*: the time between submitting the previous idea and the first keystroke of the subsequent idea. Note that for the first idea, thinking time is the time from when the question appears until the first keystroke. Thinking times are also referred to as *between-item times* for the sake of giving-up time analysis, as seeing any *between-item times* larger than *giving-up times* would rule out a simple giving-up time rule.
- *Typing times*: the time between the first keystroke of each idea and pressing the submit button.
- *Response latency*: as in study 1, this is the time between two submit presses. Response latency represents the full time it takes to formulate an idea; it is the sum of the thinking and typing time of an idea.
- *Giving-up time*: the time between when the submit button has been pressed on the last idea and the "New Question" button has been pressed.

The above acted as our dependent measures. In addition to the time dependent variables, we also looked at the *rate of return* (ideas per minute) per task and *fluency* (total number of ideas) per task. Analyses to find overall patterns in the time span of the task regardless of delay condition were done to estimate strategies used by participants to switch to new questions.

The easy and hard task split in Payne, Duggan and Neth (2007) is interesting as it may give further insight into time allocation and switching strategies. In these ideation tasks, determining what was considered a hard and an easy task was less straight forward than in an anagram task as the difficulty of a question is subject to interest in the topic and varies from person to person. In exploring median splits on a range of metrics, we found that looking at the *number of items* generated for a question yielded results, which we report. Questions that had a higher number of items generated were placed in the "easy" questions group and those with a lower number of items generated in the "hard" questions group.

Finally, the post-experiment questionnaires were prepared for analysis. The responses to open-ended question about switching strategy were coded by the experimenter into one of the following categories: *failure (F)*, *time since last idea (t)*, *time on question (Tq)*, *number of ideas (N)*, or *other (0)*. This was done simply on analysis of the wording used in the text. The first four are based on the well-known strategies found in the literature, where *failure* is what is also known as an *impasse*, feeling stuck or generating the same idea; time since last idea suggests a *giving-up time* rule; *number of ideas* and *time on question* suggest simple fixed-time or fixed-number rules. The final, *other*, was used for any strategies that did not easily fall into the other categories. Correlations were run in order to see if the self-reported strategies were in line with actual behaviour as well as whether they were correlated with the likert scale responses on the remainder of the questions on the questionnaire.

Statistical analyses of correlations and t-tests were done in IBM SPSS Statistics version 23. Tables and graphs were generated in Microsoft Excel for Office 365 Pro.

4.2.2 Results and discussion

The results are presented in three parts. First, we present an overview of the data including overall fluency per participant, fluency per question and number of questions attempted. Average *thinking times*, *typing times*, *response latencies*, and *giving-up times* are reported. We focus on the effect of the delay-time conditions on the decision to switch by looking at the number of questions attempted as well as the effect of the differing delay times on thinking times, typing times, response latencies, and *giving-up times*. The second part looks for evidence of the use of particular giving-up or switch strategies. The third covers self-reported strategies taken from the questionnaire responses and compares the written responses to those from the likert scales filled in. Additionally, we look at whether the self-reported strategies are reflected in individuals' behavioural data.

In the following section we present descriptive statistics of the number of questions attempted per participant, the number of ideas generated per participant and the rates of return (see table 4.1).

Effect of delay times on switch decisions

Independent t-tests showed that a delay of 10 seconds resulted in a significantly higher number of questions attempted, in comparison to the 25 second delay. Delay time had a significant effect on both the number of questions attempted $(t(28) = 2.210, p = .035, d = .807)$ and the number of questions seen, i.e. attempted and skipped questions $(t(28) = 2.288, p = .03, d = .836)$.

	delay condition 10	delay condition 25	Overall
Questions attempted	8.00 (3.31)	5.73 (1.95)	6.87 (2.94)
Questions unattempted	1.00 (1.26)	0.47 (0.88)	0.73 (1.12)
Questions abandoned	9.00 (4.05)	7.07 (1.98)	7.60 (3.53)
Ideas per attempted question	5.66 (3.41)	6.14 (2.48)	5.90 (2.99)
Mean ideas generated	43.80 (13.17)	40.40 (17.08)	42.10 (15.34)
Rate of return (ideas per minute)	1.88 (0.57)	1.63 (0.7)	1.75 (0.65)

Table 4.1: Mean number of questions and ideas (SD) across all participants, split by delay condition and overall. Questions unattempted are questions seen and abandoned without generation of any ideas. Questions abandoned are questions that were abandoned by the ideator and not cut short by the experiment automatically ending.

It was not possible to see a significant effect of delay time on the decision to skip a question without attempting to respond to it $(t(28) = 1.293, p = .207, d = .472)$. There was no significant difference in the mean number of ideas generated per question $(t(28) = -.431, p = .670, d = -.163)$ as well as total number of ideas generated $(t(28) = .590, p = .560, d = .223)$ for each of the two delay conditions. No significant difference was found for average time spent on each task between the two conditions. No significant difference was found for rate of return between the two conditions[3].

The timing data were first natural log transformed in order to normalise the data for t-tests. Log transformed data was then compared to see if the delay times affected any of the timings between participants. Independent t-tests comparing means of *thinking times* for the first idea of each question showed that in the 10 second delay condition (M=2.12, SD=.34), the first idea *thinking time* was significantly lower than that in the 25 second delay condition (M=2.43, SD=.34) $(t(28) = -2.463, p = .020, d = 0.899)$. No other timings showed a significant difference between the two conditions.

Giving-up decision strategies: overall patterns in time-course regardless of delay condition

Paired t-tests were conducted on log transformed data to compare thinking times, typing times and overall response latencies across both conditions. All comparisons of first idea

[3] Despite following a slightly different paradigm, that is, allowing participants to abandon a question for another, it is worth noting here that the mean rate of return of 1.75 ideas per minute in this study is comparable to that of 1.73 ideas per minute in study 1a. More importantly, this is comparable to other paid ideation studies, (e.g. Nijstad and Stroebe, 2006), although these are also single question ideation tasks.

to overall ideas exclude first idea times in the overall idea number. All comparisons of last idea timings to overall ideas exclude the last idea timings in the overall idea number. All patterns of significance are the same without log transform, with the exception of those marked by a star*.

There was a significant difference in the *thinking time* of the *first idea* (M=2.28, SD=.37) and the thinking time of the remainder of the ideas (M=1.61, SD=.56) $(t(29) = 6.69, p = .000, d = 1.22)$*. An opposite significant difference was observed for *typing time* of first idea (M=2.96, SD=.57) versus typing time of remainder of ideas (M=3.19, SD=.54) $(t(29) = -4.79, p = .000, d = -.87)$. No significant difference was found for overall response latency compared to first idea response latency.

There was a significant difference in the *thinking time* of the *last idea* (M=2.00, SD=.87) and mean thinking time for the remainder of the ideas (M=1.71, SD=.48) $(t(29) = 2.141, p = .040, d = 0.39)$. This was also found in the *typing time* of the last idea (M=3.36, SD=.61) compared to overall typing time (M=3.11, SD=.52) $(t(29) = 4.45, p = .000, d = .81)$ and the *response latency* of the last idea (M=3.75, SD=.55) compared to the overall response latency of all ideas (M=3.46, SD=.44) $(t(29) = 4.79, p = .000, d = .87)$.

The rate of return (ideas per minute) was calculated for the first 2 and last 2 ideas. In order to do this, only questions with four or more response ideas were included. This left us with a sample of 137 questions answered by 29 participants. Analysis of the rate of return for the *first two ideas* (M=2.57, SD=1.00) and *last two ideas* (M=1.77, SD=.84) showed a significantly higher rate of return in the first two ideas $(t(28) = 6.186, p = .000, d = 1.15)$.

In general these comparisons fit with the idea that participants typically find ideas harder to generate before choosing to quit a task. We cover this further in the analysis of giving-up times.

Simple heuristics for switch decisions

The very simplest heuristics suggested in the foraging literature (see Chapter 2) are that foragers might choose to quit a patch (a task in our case) after a fixed time, or after finding a fixed number of items (ideas). If tasks are roughly equivalent in difficulty, or if rate of production is very irregular, so that later performance on a task is not well predicted by earlier performance, these very simple heuristics will be reasonably efficient. Nevertheless, they seem, on the basis of the foraging literature, unlikely. The simplest evidence against them in this study was to inspect the time-course of each participants foraging across tasks and to look at how variable *time* and *number of items* per task are. A visualisation used by Wilke et al. (2009) for this purpose is shown in figure 4-2 and figure 4-3 (left panel - we will discuss the right panel in the next section), for 8 randomly selected participants, 4 in each delay condition. Inspecting these figures strongly confirms that these simple heuristics are not operational. All participants quit some tasks much quicker than others, and generate many fewer ideas for some tasks than others.

Giving-up times and switch decisions

The next simplest foraging heuristic for time allocation across patches (tasks) is a giving-up time rule, in which foragers quit a patch after a fixed threshold of time without a success. This is a more flexible and adaptive rule in many situations, as it results in more time being allocated to more productive tasks.

There are many ways of inspecting giving-up time data to test the explanatory power of the giving-up time rule in our data. First consider the scatter plots in the right-hand panels of figures 4-2 and 4-3. A simple giving-up time rule would predict that no thinking times would be longer than giving-up times. From the scatter plots from 8 randomly chosen participants (4 in each delay condition), it is clear to see that this was not the case, and that although giving-up times were generally longer, quite a few between-item thinking times were found to be longer.

Paired samples t-tests were performed on log transformed data for mean giving-up times and mean between times per participant. Mean giving-up times (M=2.10, SD=.80) were found to be significantly larger than mean thinking times (also called between times for the purpose of our study) (M=1.75, SD=.49) ($t(29) = 2.261, p = .031, d = 0.41$)[4].

Giving-up times and the longest between-item times are shown in figure 4-4. Tasks in which no ideas had been generated were excluded from this analysis, leaving us with 206 data points across 30 participants.

Out of the 206 sets of ideas submitted (i.e. $participant \times task$), 32% had longer between times (thinking times) than giving-up times (36% in the 10 second delay condition, 27% in the 25 second delay condition). This is simple and strong evidence against the use of a simple giving-up time heuristic. Figure 4-5 shows the overall distribution of giving-up times, with the mean between-item thinking time shown for comparison.

[4] Non-transformed data: Mean giving-up times (M=17.64, SD=11.53) was significantly larger than mean thinking times for all ideas (M=10.28, SD=5.23) ($t(29) = 3.727, p = .001, d = .68$). Note these findings were equally reflected if questions with less than 3 and 4 ideas were removed.

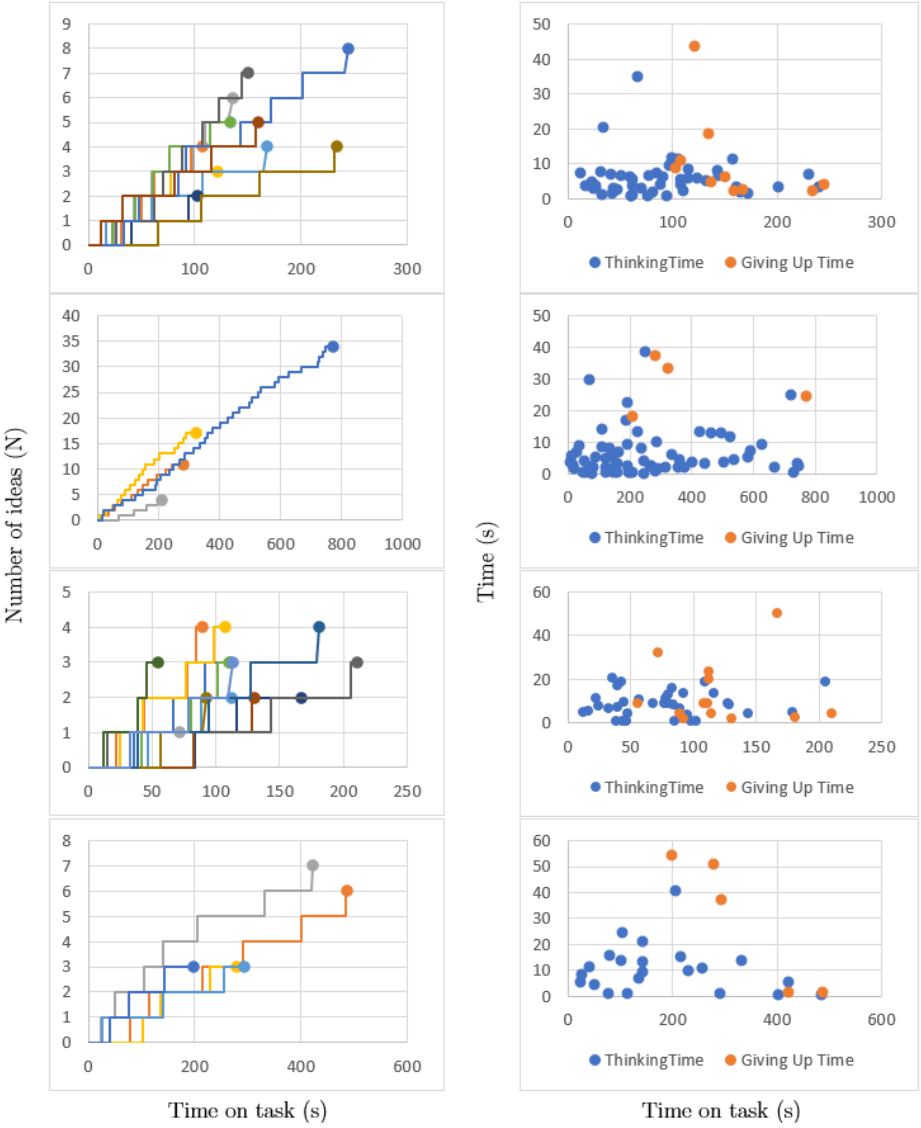


Figure 4-2: The trajectories of number against time on task of four randomly chosen participants from study 2, in the 10 second delay condition. Vertical jumps represent finding an idea, horizontal lines represent thinking times. On the right, the thinking times (blue) and giving-up times (orange) plotted at the time during a task they occur for those same four participants. Note the variation in scales between participants. All scales are linear.

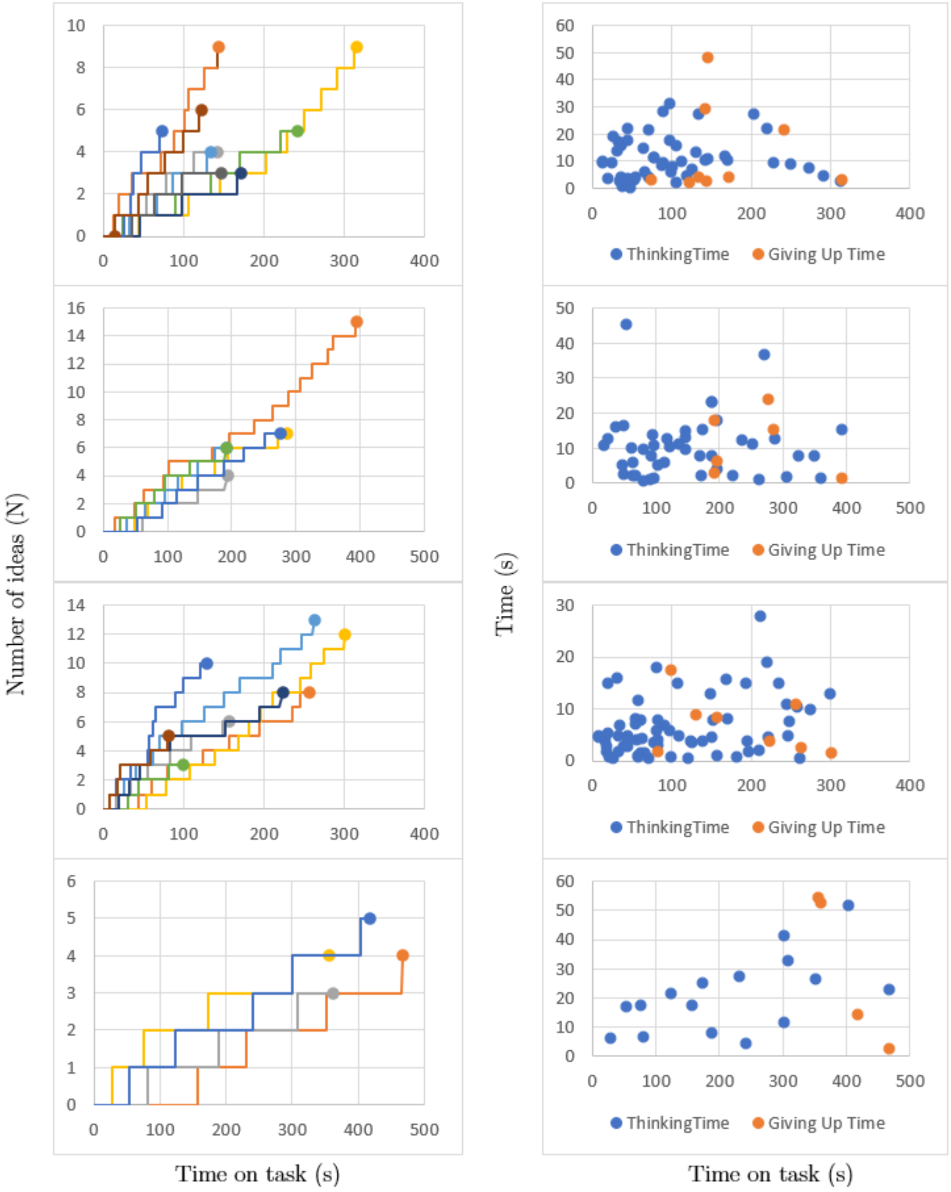


Figure 4-3: The trajectories of number against time on task of four randomly chosen participants from study 2, in the 25 second delay condition. Vertical jumps represent finding an idea, horizontal lines represent thinking times. On the right, the thinking times (blue) and giving-up times (orange) plotted at the time during a task they occur for those same four participants. Note the variation in scales between participants. All scales are linear.

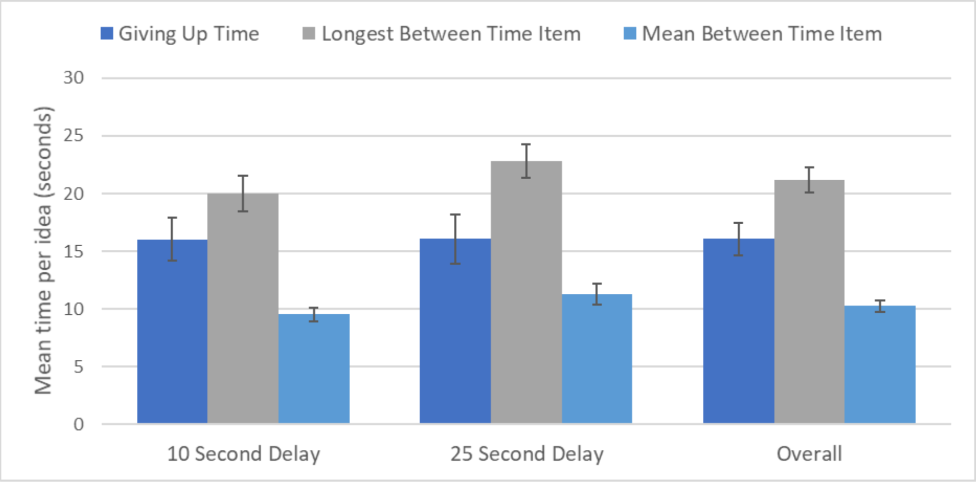


Figure 4-4: Comparison of giving-up times, longest between-item time and mean times between items (thinking times) for each of the conditions in study 2.

Overall, the giving-up time data supports **H3**. Too many tasks had between-item times that were longer than giving-up times to support a simple giving-up time heuristic. However, giving-up times were typically longer than between-item times, as are times to generate the last two items. This supports the operation of a rate-based heuristic such as Green's rule, where quit decisions are typically made when ideas are becoming harder to find. At the same time, some giving-up times were brief, which supports the idea of a two-process strategy: occasionally participants quit a task on sub-goal completion.

The effects of task difficulty

Due to participants' varying interest in, and motivation to respond to, each of the questions, we tested a range of measures to assess subjective difficulty. It is problematic to answer questions about the effects of task difficulty in this experiment, as most indices of difficulty for a participant are to some extent under strategic control. In earlier work (e.g. Payne, Duggan and Neth, 2007), tasks were purposefully designed to manipulate difficulty, and its effects could therefore be investigated. That is perhaps not impossible to do for ideation tasks, but the high variance of fluency in the studies in the literature is suggestive that difficulty of different tasks will vary among participants.

In this experiment, it is tempting to assume that if a participant generates ideas at a higher rate in some tasks than in others then that task is easier for that participant - but because participants can choose to persist with some tasks for longer than others that is not necessarily a reasonable approach. Nevertheless, in Payne, Duggan and Neth (2007), the

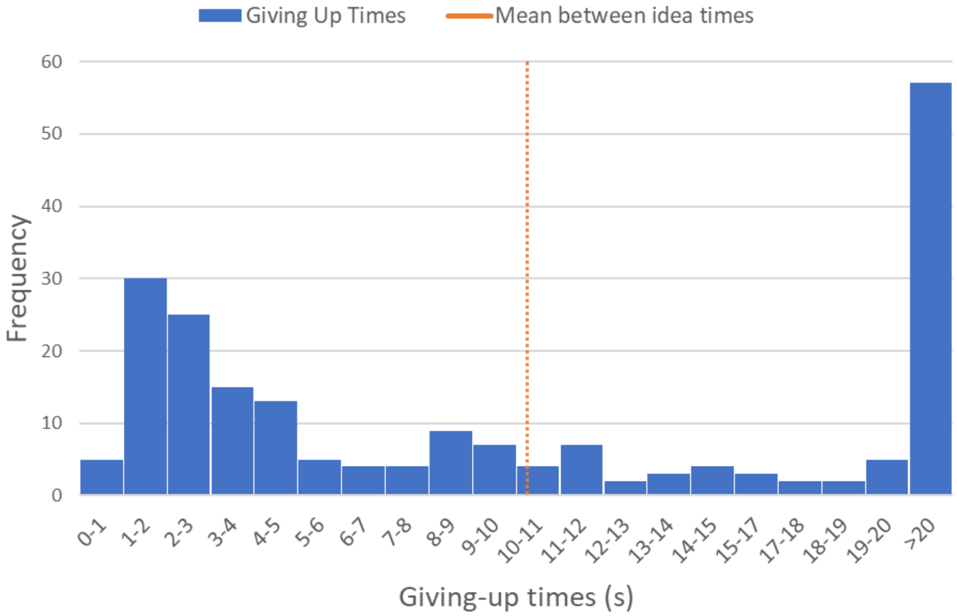


Figure 4-5: The distribution of giving-up times, compared with the mean between-item thinking times (M=10.28, SD=5.23) in study 2.

reliable finding that people spent longer in more rewarding tasks, and that giving-up times were higher in the more difficult task were both informative about participants strategies. It does not seem possible to test the first of these relationships, for the reasons mentioned immediately above, however, this first finding supports the suggestion that foragers are adaptive. The second reliable effect supports a particular heuristic, i.e. Green's rule. As (Payne, Duggan and Neth, 2007, p.377) argue:

> '[According to Green's rule,] In richer patches, the visits are longer, but items occur more densely within that time period. Therefore, the chances of an item occurring shortly before the leave decision (producing a short giving-up time) are increased. Shorter giving-up times in richer patches (easier tasks) thus occur as a probabilistic effect.'

To test whether giving-up times are shorter in richer patches in the current experiment required a model of richness that is completely independent of giving-up times. Clearly rate of return was not such an index, but fluency was: participants generated more ideas on tasks that were easier for them.

Note that all tasks with 0 ideas submitted were removed for the purpose of this analysis, leaving us with 206 tasks performed by 30 participants. A median split was performed on

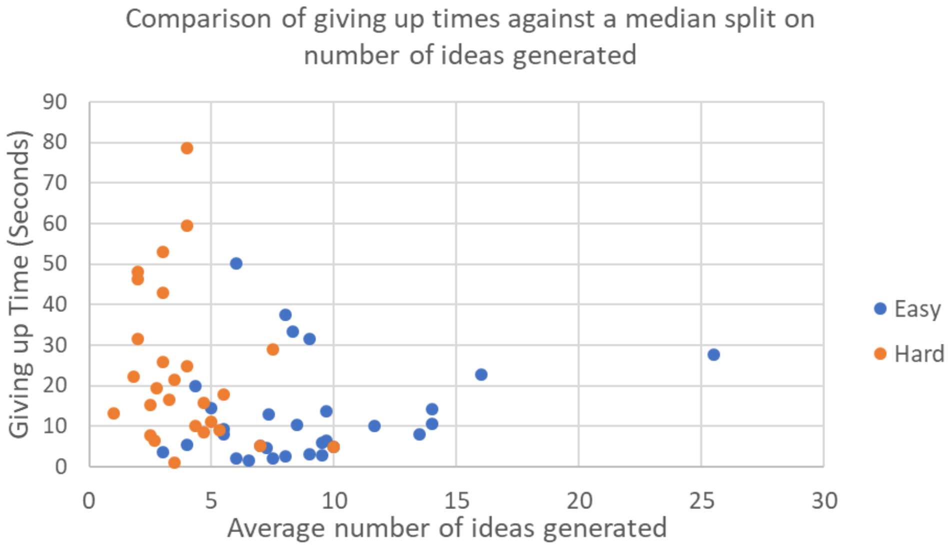


Figure 4-6: Comparison of giving-up times of subjectively easy and hard tasks in study 2, against the median split on number of ideas generated, for each participant individually.

the number of ideas per task for each participant: providing a set of difficult versus easy tasks for each individual participant, and the means of giving-up time were then calculated for the two halves of the data. Some participants' datapoints could not be calculated as performing a median split on number of ideas on datasets of 2,2,2,3 yielded no clear split. Figure 4-6 shows two data points per participant, one for mean giving-up time in the *easy* questions and one for mean giving-up time in *hard* questions.

Giving-up time in the tasks identified to be *easy* (M=13.55, SD=12.29) by the median split was found to be significantly smaller than giving-up time in the tasks identified as *hard* (M=23.17, SD=19.31) $(t(27) = -2.466, p = .020, d = -.47)$. This confirms the relation between giving-up time and patch richness and is supportive of the hypothesis that participants are making giving-up decisions according to Green's rule (**H2**).

Self-reported strategies

Questionnaire responses showed that most participants were using a switch strategy based on failure (feeling stuck, generating duplicate ideas, being unable to think of new ideas). We read through the qualitative responses on the questionnaire and analysed this for key words or strategies. These were coded into 5 different strategies, shown in table 4.2.

No significant differences in the means of any scale question in the post-experimental questionnaire was found in comparison with the self-reported switch strategy - an interesting result in itself, as it shows that despite a person being aware they are using a specific strategy, they still might not recognise the questions as being representative of their strategies.

Strategy Summary	
Failure (F)	20
Time since last idea (t)	8
Time on question (Tq)	0
Number of ideas (N)	2
Other (O)	6

Table 4.2: Self-reported switching strategies (N=30) from study 2 participant questionnaire. Note that the numbers don't add up to 30, this is due to 6 people reporting multiple strategies (e.g. using an aggregate strategy involving time since last idea and a *failure* (N=5) or number of ideas generated (N=1)).

Intercorrelation analysis between questionnaire responses showed little evidence of correlations. There was a positive moderate correlation between the responses to number of ideas generated being a switch cue and time spent on current question being a switch cue $(r = .576, p < .01)$ and low negative correlation between "It was very difficult to think of new ideas" and number of ideas as a cue $(r = -.494, p < .01)$, as well as time spent on question as a cue $(r = -.499, p < .01)$. "I only pressed next question when I felt stuck" and feeling stuck on current question as a cue to switch had a low positive correlation $(r = .493, p < .01)$.

Low positive correlation was found for the questions "I continuously switched between the feelings of being stuck and un-stuck" and "I only pressed 'new question' when I felt stuck" $(r = .362, p < .05)$. A low positive correlation was also found between the former question and "the same idea occurred to me several times" $(r = .388, p < .05)$.

No correlation was found between "Number of ideas generated on current question" as a cue to switch, and the variance in number of ideas generated per task $(r = -.265, p > .1)$

4.3 Chapter Summary

In this chapter, we presented a study in which ideators were given a practically unlimited number of questions to respond to in a time-bound setting. This approach allowed us to treat the sequence of ideation problems as a sequence of patches, in order to investigate what rules might govern ideation behaviour and the decision to switch from one problem to the next. As in the previous study, which was a replication of the studies presented in Nijstad and Stroebe (2006), we explored the behaviour of non-expert ideators. In contrast to the previous study, for this study we provided a financial incentive to the generation of many ideas. Whilst it is not wholly possible to measure the impact of this due to the different structure of this study, the rate of generation of ideas is comparable to other paid single question ideation studies. Interestingly, the rate of idea generation is also comparable to that found in the environment question in study 1a, but markedly higher than that found in the health question study 1b. The high rate of idea generation in the environment question could potentially show an overall higher interest from the participant group in that question, or it could simply be coincidence, affected by the small number of participants.

By presenting ideators with a sequence of ideation tasks and allowing them to choose when to abandon any task to move on to the next, we have been able to test the extent to which patch-leaving rules from foraging theory are able to explain ideators' decisions to switch tasks.

First, by manipulating a time-cost invoked at every change of task (comparable to travel time between patches), we have found support for generally adaptive behaviours as predicted by Charnov et al.'s (1976) Marginal Value Theorem: ideators will attempt less tasks if the time cost of accessing the next task is larger. This is a simple but fundamental support for our proposal that foraging theory might speak to time management in ideation tasks. We then turned to investigate simple patch-leaving heuristics as potential explanations of this adaptive behaviour. Analysis of the variation in rates of idea generation, quite clearly appear to rule out very simple rules such as switching after a fixed time, or after a fixed number of ideas. This is not surprising: these heuristics have not been shown to operate in any study of human foraging, whether externally, or internally (Payne, Duggan and Neth, 2007; Wilke et al., 2009; Hills, Jones and Todd, 2012).

More interestingly, analyses of giving-up times show that a simple giving-up time rule is also a poor candidate for explaining ideators' switch behaviour. It is very common for participants to have thinking times between items within a task that are longer than giving-up times. Furthermore, giving-up times are reliably longer on less productive tasks. Again, this replicates the findings in the information foraging literature, and extends them to this more creative foraging task using creative thinking as a tool for foraging for ideas.

Finally, the distribution of thinking times is quite strikingly similar, in certain ways, to that found by Payne, Duggan and Neth (2007) in their study of a Scrabble-like task. Thinking times are longer in general for ideas occurring later in the task than at the start of the task. Giving-up times in general are longer than between-item times, and (as above) longer for more difficult tasks, but sometimes are very short, shorter than the average thinking time. In all these respects the data suggest that ideators are sensitive to the within-task rate of generation of ideas, and quit when this is poor, but also might sometimes prefer to quit immediately after completing a subgoal – i.e. generating an idea.

As noted by Payne, Duggan and Neth (2007), this two-factor account of task switching is quite compatible with observations of multi-tasking in complex office environments (e.g. González and Mark, 2004).

Our original motivation for this study was to investigate what rules ideators might apply when deciding to switch from one question to another. We found a two factor account of task switching as has been shown in other types of task such as a Scrabble-task. Considering the iterative ideation-evaluation process of creative problem solving, and our speculations in Chapter 3 (study 1) that ideators evaluate their own ideas as they write them down - could it be the case that ideators use some self-judged quality of their ideas as a prompt to give up on a task? Indeed if people use subjective quality of ideas as a measure of giving-up, this might give further insight into the decision to switch immediately after submitting an idea versus after a longer period of idle time. In the next chapter, we present a short analysis of the data from the current study, originally setting out to explore this phenomenon further. Unfortunately, methodological issues with the quality coding exercise itself became clear early in the analysis, changing the focus of the analysis away from the quality of the ideas themselves, towards an investigation of the properties of the ideas themselves; properties that might be having a negative impact on independent coder agreements.

Chapter 5

Judging the quality of ideas: novelty and value

In the previous chapter, we presented a study in which ideators were given a seemingly unlimited number of questions to respond to within a 30-minute time limit. We looked at the time-span of ideation across multiple ideation tasks and found evidence of a two factor account of task switching, a sensitivity to rate of return and the tendency to switch immediately upon submission of an idea, suggesting that we might be able to apply classic foraging theories to our understanding of ideation behaviour.

The analyses of patch leaving strategies in the study presented in Chapter 4 assume that participants are sensitive to the number of ideas they are generating, and the rate of generation. As a results of the ideation-evaluation nature of creative problem solving, it seems possible, additionally or alternatively, that participants will care about idea quality, and will prefer to stay in patches that are yielding the best rate of high-quality ideas rather than ideas per se (**RO4**). The quality of ideas in ideation tasks like the ones used in these studies is quite a challenging concept, with some discrepant suggestions in the literature. Having reviewed the different methods of assessing quality in Chapter 2, it was decided to focus on two independent qualities, namely novelty and value. We asked independent coders to rate these properties, using definitions and instructions often used in the literature (e.g. Ye and Robert Jr, 2017; Dennis et al., 1999; Dennis, Minas and Bhagwatwar, 2013; Diehl and Stroebe, 1987).

5.1 Method

Three coders were hired through word of mouth. As in study 1, it was decided to hire external coders unfamiliar with the research goals to avoid bias. All three coders, 1 male and 2 female, aged 20-21, were native English speaking third year undergraduate students in the Department of Computer Science at the University of Bath. All 3 coders individually rated all 1264 ideas given in response to the 17 questions by all 30 participants. The coders were each paid £60 for the entire coding exercise. Due to a lack of access to professional coders, we hired coders with no prior experience in coding this type of data. Consequently, prior to coding the full datasets, each of the coders received thorough instructions and were asked to talk through their coding to show understanding.

The work itself took place in 3 stages. First the independent coders had an introductory session in which they met in person with the experimenter individually. They were explained the purpose of the research, but not the goals of the analysis, trained on how to perform the ratings and asked to talk through how they would rate a set of ideas in order to verify they had understood the brief.

The coders were asked to individually rate every idea on two metrics: *value* and *novelty*. They were asked not to meet or discuss their ratings as this would invalidate the ratings being independent. These were to be rated on a scale of 1-10, 1 being 'no value/not novel' and 10 being 'high value/very novel' (for full written instructions given to coders, see appendix B.5). These were for the purpose of this study defined as:

- Novelty: how original and surprising is the idea?
- Value: how useful and practical is this idea and does it make sense as a solution to the problem?

Training in the use of the novelty and value scales consisted of verbally going through each point on the scale, assigning this to an idea in a sample of the data, with each coder. Once each point on the scale had been explained, the coder was asked to rate a sample of the data and talk through their thinking, in order to show they had understood how to use the scales. Separate samples were used for each coder to avoid influencing the agreement scores. They were given an estimate of three hours to complete the task, although were allowed to spend more time should they need to do so. They were allowed to complete this in their own time. They were told that although the work was quite simple they should ensure they focus on it rather than have distractions in the background. Each question's response set was listed on individual tabs, resulting in 17 tabs in total. Coders were instructed to work on one question at a time by initially reading through all ideas for that single question, without making any judgements. They were then asked to go back to the top and give "novelty" and "value" scores to each individual idea. Coders were specifically reminded that novelty and value were meant to be treated as two separate scores and they should not use one idea's novelty score as a cue for the value score. Coders were asked to use the notes field in case they found some difficult to rate, such that these could be discussed in the final stage, the debrief.

After completing the quality coding, coders individually attended a 30-minute debrief session with the experimenter, in which they talked through some of the ideas they found difficult to code. Once the debrief session was over, coders were thanked for their time and paid the £60 fee for completing the work.

5.2 Results and discussion

In the following section we present the results of the quality coding performed by three coders. In the first part, we discuss the characteristics of the coders and the ratings of novelty and value that they have submitted. In the second part we present the correlation between coder ratings per metric and per participant and question.

Due to low inter-rater agreement, we present a quick overview of analyses done on each of the set of ratings. These can only be seen as tentative conclusions as the low agreement rules out certainty in the conclusions we make. The following results highlight the analyses we would be performing had we had a higher agreement level between coders.

5.2.1 Coder and rating characteristics

Table 5.1 shows a summary of the means and standard deviations of ratings done by each coder, as well as the range of the 10-point scale used. In figure 5-1 we show the distribution of the rating given by all three coders (individually).

Correlation between *novelty* and *value* for each individual coder revealed no correlation for coder 1 ($r(1257) = -.047, p = .096$). A weak negative correlation was found between novelty and value scores for coder 2 ($r(1262) = -.205, p = .000$) as well as a weak negative correlation in the same scores for coder 3 ($r(1253) = -.233, p = .000$). Note that the differing degrees of freedom are due to coders 1 and 3 leaving some responses blank.

Coders were asked to leave comments in case they wished to elaborate on their response or highlight the reason for a specific response. Only one comment was left by coder 1, whereas there were 95 comments left by coder 2. Among the comments left by coder 2, comments such as "not practical/possible", "socially unacceptable" and "not enough

		Mean	SD	Min	Max	Left Blank
Coder 1						
	Novelty	5.11	2.71	1	10	5
	Value	6.64	2.20	1	10	5
Coder 2						
	Novelty	3.18	1.64	1	9	0
	Value	6.32	1.75	1	10	0
Coder 3						
	Novelty	4.13	1.71	1	10	11
	Value	5.54	1.30	1	9	11

Table 5.1: Table showing mean ratings for novelty and value as given by each individual coder; data coded is from study 2.

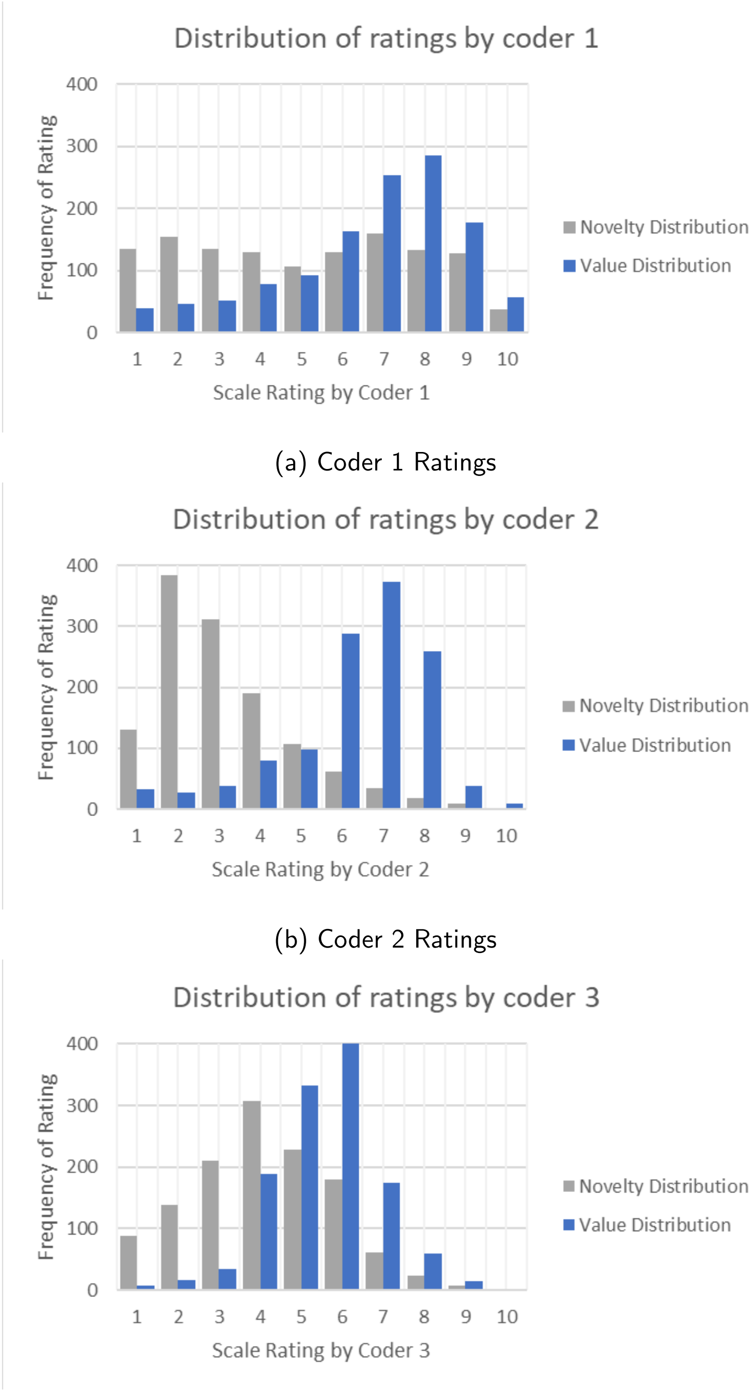
	Novelty			**Value**		
	Coder 1	**Coder 2**	**Coder 3**	**Coder 1**	**Coder 2**	**Coder 3**
Coder 1	-	.511	.384	-	.360	.228
Coder 2	-	-	.470	-	-	.282

Table 5.2: Correlations of coder scores for novelty and value; data coded is from study 2.

detail" appeared multiple times. Coder 3 left 19 comments. The majority of these were about the idea not being applicable (and subsequently, "not an idea to [solve the key problem in the question]").

Pearson correlations between the two coders were all significant, however, the correlations showed mostly weak and moderate agreement between ratings (see table 5.2). Running Krippendorff's alpha[1] (Hayes and Krippendorff, 2007) (ordinal) analysis on novelty agreements between the 3 coders resulted in a disagreement rate of $a = 0.316$ (3776 decisions across 1264 cases). Running Krippendorff's alpha (ordinal) analysis on value agreements between the 3 coders resulted in a disagreement rate of $a = 0.163$ (3776 decisions across 1264 cases).

[1] Krippendorff's alpha compares 'observed' disagreement with the 'expected' disagreement. Expected disagreement is strongly influenced by the ratio of values, 1-10.



(a) Coder 1 Ratings

(b) Coder 2 Ratings

(c) Coder 3 Ratings

Figure 5-1: Distribution of novelty and value ratings as given by each individual coder; data coded is from study 2.

	Coder 1		Coder 2		Coder 3	
	Novelty	**Value**	**Novelty**	**Value**	**Novelty**	**Value**
N of ideas per question /participant	0.023	0.101	-0.054	0.04	-0.063	-0.057
Giving- up times (s)	0.121	0.126	0.056	0.055	0.075	-0.017
LN of Giving-up times	0.133	0.133	0.06	0.01	0.062	0.011
Time spent ideating (s)	0.173	0.04	0.093	-0.018	0.084	-0.109
LN of time spent ideating (s)	0.152	0.033	0.089	-0.009	0.137	-0.097
Time on task (s)	0.179	0.075	0.093	0.011	0.146	-0.108
LN of time on task (s)	0.189	0.058	0.101	-0.01	0.094	-0.11

Table 5.3: Correlations between timing measurements for each $participant \times question$ pair. No significant correlations were found between any of these metrics and the novelty and value scores given by any of the coders.

5.2.2 Quality coding

Due to the low agreement level between the coders, analyses on quality per participant have not been calculated. Although stated that it is not fair to calculate quality as mean quality (due to 1 great idea and 3 bad ones scoring lower than a mediocre idea), we are interested here in looking at the mean quality per idea set and seeing if this had an influence on the ideators choice to move on to another question. For this, tasks that had 0 ideas (e.g. abandoned tasks) were removed leaving us with 206 sets of ideas across 17 tasks by 30 participants. We calculated the correlations between novelty and value ratings, given by all 3 coders separately, against a number of metrics in order to see if a measure of quality was used by participants in study 2 to decide when to give up, how much time to spend on the task and whether number of ideas correlated with the value/novelty of the set. None of these correlations were significant (see table 5.3). Interpreting this as being evidence that quality is not a part of the strategy to give up on a question would be imprudent considering the disagreement between our coders. Instead we look to the data in order to see whether these disagreements (also found in study 1 category codings) could be attributed to the properties of the ideas themselves.

5.2.3 The properties of agreed-good v agreed-bad v no-agreement ideas

Faced with the problem of very poor inter-rater agreement of judged value and novelty, we wondered if this might be an effect of the variable specificity with which ideas are expressed. We conducted two exploratory analyses to investigate this speculation. One analysis used a very coarse measure of specificity, namely the word length of ideas. The second analysis instead considered the thematic roles that were present in ideas.

The themes used are from classic case grammar. The ideas are generally speaking sentences and therefore typically exhibit the classic sentence structure. It seemed plausible to wonder whether better formed ideas have more of these thematic categories in their expression and whether the worst formed ideas are reduced sentences that only hold an action or theme.

A subset of the judged ideas was inspected to find out whether ideas rated as overall agreed good (between our three coders) stated explicit roles, themes or goals of the ideas, in comparison to overall agree bad ideas.

We analysed the top 20 agreed good ideas (aggregate found by calculating the means of all three novelty and value ratings) as well as the lowest 20 agreed ratings, looking at the number of words in the ideas. The following are examples of the low and high agreed quality ideas:

- Low: "Have less things"
- Low: "Do the things you want to"
- High: "Have a partnership with the city council of Bath for building houses which would serve not only for students but can be for example temporary houses for tourists during the summer."
- High: "Your phone could warn you of known areas with poor cell data coverage, to help you avoid such areas if you are streaming music for example"

The number of words in agreed low ideas (M=11.00, SD=6.00) was significantly lower than that in agreed high (M=18.30, SD=8.72) $(t(38) = -3.083, p = .004)$. For reference, the 20 ideas that had the highest disagreement between coders, a mean word length of 13.35 (SD=9.34) was found.

We performed a thematic analysis, using a variation of the thematic roles set out by Berk (1999), to see whether we could account for the difference in idea lengths by observing differences in specificity of the ideas. A sample of the 20 agreed highest scores, 20 agreed lowest scores and 20 highest disagreement sets were coded with the set of thematic roles listed. We found the following thematic roles explicitly or implicitly specified in the agreed good ideas:

- Agent: who is performing this action?
- Action: what are they doing?
- Instrument: what are they using (e.g. phone, a map)
- Recipient: who is benefiting from this action?[2]
- Purpose: why is this being done, what is the goal of this?[3]

In the agreed-highest scores set (M=6.15, SD=0.96), the number of thematic roles occurring was higher than in the overall agreed-lowest scores set (M=3.75, SD=1.22). T-tests revealed that this was a significant effect $(t(38) = 6.73, p < .001)$. For reference the average number of thematic categories in the highest disagreement set (M=3.70, SD=1.76).

We used the concept of these thematic roles in order to inform our third and final study, in which we used prompts as a method of improving the overall length (and presumably specificity) of ideas. We furthermore applied this manipulation in order to understand if longer ideas might have an effect on the agreement between coders in the different coding methods we have used (categorisation, semantic differences and novelty/value ratings) through the previous 2 studies in this thesis.

5.3 Chapter Summary

Unfortunately, despite training and testing that coders understood the task, the inter-rater agreements for *value* and *novelty* ranged between 23%-51%, much lower than we would have expected to see. According to Krippendorff (2018, p.241), a values higher than .800 are the standard agreed acceptable value. When a tentative conclusion is acceptable, a values of higher than .667 is the lowest limit. Unfortunately, our scores were closer to 0 and we could therefore not accept these ratings as being reliable as moderate agreement had not been found.

Inter-rater disagreement may have been due to inadequate instructions – despite our care with this aspect of the procedure and it being modelled on published studies. Showing coders good and bad ideas, with accompanying rationale, might be another instruction

[2] Recipient and beneficiary were two separate roles, however, for the purpose of this analysis, these were combined into simply one role.

[3] Similarly, purpose, benefit, and goal were separate roles, however, these were found to be similar as well and therefore combined into one role for the purpose of our study.

method to avoid such high disagreement. Further coder training seems to be needed on the basis of these results.

Alternatively, or additionally, we might consider again the issue of well-formedness of ideas that was raised in Chapter 3. Our attempts to find indices of ideas that separate agreed-good from agreed-bad and from ideas where coders disagreed have shown two simple, related and promising patterns: well-formed ideas are longer (in terms of number of words) and contain more thematic roles. In the next study, this finding is used as the basis for an experimental manipulation of prompts for ideators, and the effects of these prompts on timing and number of ideas as well as judged idea quality are tested.

Chapter 6

Supporting the well-formedness of ideas

In the studies presented in our previous two chapters, we attempted to understand the time-course of ideation across a single task and across multiple tasks. We are interested in both single and multiple task paradigms as both are encountered in everyday lives. Performing a single ideation task is comparable to being given a problem at work and needing to brainstorm ideas to solve this particular problem, whereas multiple tasks might be encountered on a day to day basis when performing knowledge work. We likened ideation tasks to knowledge work, as opposed to simple rote tasks such as puzzle tasks. Presumably, people do in fact switch between multiple knowledge tasks in work settings. Both of these paradigms were therefore found worthwhile studying in order to gain more evidence for appropriate methods of supporting ideation task performance.

Our findings for a single ideation tasks showed promising results in relation to ideators generating ideas in a categorical structure in two different questions, a result we were attempting to replicate from the studies in Nijstad, Stroebe and Lodewijkx (2002), Nijstad, Stroebe and Lodewijkx (2003), and Nijstad and Stroebe (2006). We found some evidence of a categorical structure to thinking, however, despite generating ideas within similar categories, ideators showed evidence of switching often between categories. Fluency, presumably affected by categorisation of ideas, was not correlated with the *Adjusted Ratio of Clustering* scores, indicating that clustering was not having an effect on productivity. A variation we performed on the original study was to split thinking and typing times. We were interested mostly in seeing the effect of category switches on idle times between ideas rather than the full response latency. Our variation was successful and showed that typing times remained unaffected and therefore thinking times were an appropriate measure of idle times. Thinking times were successfully shown to be significantly longer when an idea followed one in another category (between category timings) than when an idea followed ideas in the same category (within category timings).

We found some inherent problems in the study design itself. Initially, we did not observe a decline in productivity as might be expected in a longer ideation task, which may be attributed to the short length of the task itself. A major limitation in the study was encountered when we received the results from independent coders of the data. We had access to a pre-determined set of categories, developed by Diehl (1991) and verified by Nijstad and Stroebe (2006), however, the independent coders expressed concern with both the vagueness of the category system itself as well as with the ideas being categorised. Performing inter-rater agreement scores reaffirmed these concerns. It is therefore important that we take the results evidencing a categorical structure as tentative results as our hypothesis was not possible to accept with certainty based on the low coder agreement scores. In our second study on multiple ideation tasks, we moved away from relying on subtle category switches as a method of understanding decision making in ideation tasks and towards more explicit task switching in order to understand decision making. This study was likened with foraging studies such as Payne, Duggan and Neth (2007) and Wilke et al. (2009) in order to show that the decision to give up on a patch in ideation can be likened to that of giving up on patches in more rote tasks. An average of 32% of the participant-question idea-sets submitted had longer between idea thinking times than giving-up times, ruling out the possibility of a simple giving-up time strategy for leaving a question. Similarly, more simple strategies were shown not to be used as no correlations were found between number of items and giving-up times, as well as time on task and giving-up times. We were able to verify that giving-up times follow a fairly bi-modal distribution, that there are a high number of higher giving-up times, however, a spike in giving-up times is found in the 1-3 second span. This is indicative of ideation being subject to what Payne, Duggan and Neth (2007) calls sub-goal completion.

We furthermore found that longer delay times resulted in a tendency to stay on questions rather than attempt more questions, indicative that an awareness of long switching costs is a deterrent to moving on to another question. Rate of return for a task was found to be strongly related to the time taken to generate the two first ideas in that task. Despite this, median split analyses on time taken to generate first ideas did not show significant differences in giving-up times and time spent on task. This indicates that ideators were not using initial generation rate as a cue to allocating time efficiently.

With a stronger understanding of the time-course of sequential ideation tasks, we were further interested in analysing the effect of quality of ideas on the decision to give up. We encountered similar problems in this study design to that of the design in study 1, despite training of coders and verifying they had understood the task. On both metrics such as value and novelty, coders had a fairly low agreement, leaving us again with the ability to only make tentative conclusions about the data. The comments from coders remained the same as in the individual ideation task, that the ideas were vague and at times were open to interpretation.

In this chapter we address the problems encountered in our first two studies with a particular focus on increasing the well-formedness of ideas generated. We aim to discover whether the problems encountered of agreements between coders may be attributed to the fact that the ideas generated were too non-specific, making it difficult to understand what the ideator actually meant.

6.1 Chapter background

In an attempt to support higher agreement between coders in both quality as well as categorisation of ideas, we look to the concepts of generalisation and specificity of the ideas themselves. It is conceivable that the vagueness of the ideas, and therefore, the need for subjective interpretation has a negative effect on the ability to judge the set of ideas somewhat consistently.

Very little research in ideation focuses on the specificity of ideas. In fact, in a review by Dean et al. (2006), they found that out of 51 studies looking at the quality of ideas, only 10% spoke about the concept of specificity, described in a variety of terms such as generality (Taylor, Berry and Block, 1958), detail and clarity (Durand and VanHuss, 1992), abstraction (Reinig, Briggs and Nunamaker, 2007) and thoroughness of description (MacCrimmon and Wagner, 1994; Cady and Valentine, 1999).

The rest of this chapter describes the ideation study run with ideators individually, performing a singular ideation task by typing out as many solutions as possible to one of the following questions: "What can the individual do to increase their general level of health?" and "Your phone can tell you exactly where you are, come up with functions or apps that can use this." This was a between participants study looking at the effect of different levels of prompt on the length of ideas, quality of ideas and ultimately, the ease at which ideas were rated.

6.2 Study 3: Prompting increased specificity of ideas in an online ideation task

In this section, we present our third study in which we explored the effect of two different methods of prompting on the specificity of ideas and whether we could help ideators produce more well-formed ideas in comparison to a no prompt condition (**RO5**). Our main motivation was to look at whether we were able to improve coder agreements across quality ratings as well as categorisation of ideas (**RO6**)[1]. This between participants study compared three conditions: no prompt, implicit prompt by example, and explicit prompt by providing a list of thematic roles to include in the idea. Each condition was run over two questions in order to rule out question specific results.

6.2.1 Method

Design

An online experiment run on the Gorilla platform (Anwyl-Irvine et al., 2018) was conducted. This followed a similar paradigm to that in study 1 in that each participant was given a singular question, however, the time was increased to 25 minutes to generate as many ideas as they could that would serve to solve the given problem. Online ideation studies are often run in 15-minute sessions. However, in study 1, we found that a 15-minute timespan was not enough to exhaust ideators stock of ideas. We therefore decided to increase the ideation time substantially in order to see if we could observe a diminishing returns curve. The study followed a 3x2 between participants design. The independent variables consisted of prompt type (no prompt, implicit prompt and explicit prompt) and question (health question, phone question). The questions were chosen as appropriate tasks for this as they were successfully used before in the previous two studies. Question one was adapted from our first two studies: "What can the individual do to increase their general level of health?". It was selected as it had the highest average number of responses in study 2, with a range of responses between 4 and 34 ideas, e.g. a difference of 30 where most other ideas had a range of 5-17. This had the added benefit of being comparable the category coding we performed in the first study, in which the agreement between independent coders for category coding was good but unfortunately not as high as expected. Question two was taken from study 2: "Your phone can tell you exactly where you are, come up with functions or apps that can use this." This question was selected due to the perceived narrower technical context in comparison to the broad context of personal health.

[1] Note that we are only performing categorisation of ideas in the health question. This is partly due to the lack of a pre-set category system for the phone question.

Participants were randomly assigned to one of the 6 groups by the Gorilla application, answering only one question in one of the prompt conditions. In the no prompt condition, participants were electronically given the same instructions as in the previous two studies: to generate ideas according to Osborn (1957)'s brainstorming rules. On the main ideation screen they simply saw the question, the text entry box and a list being populated as they submit their ideas.

In addition to being given Osborn's (1957) brainstorming rules, participants in the example prompt condition were further offered a list of 3 examples on the main ideation entry page (but not in their instructions). The examples chosen were previously submitted ideas to the relevant questions from study 2. These ideas were chosen based on the following criteria: they feature in the top 10 highest overall scores given by all 3 coders in study 2 and include at least 4 of the identified thematic roles (such as "Use apps or fitness trackers which can help keep you on track and motivate you, e.g. you can compare number of steps walked on each day."). The concept of presenting example ideas has been reviewed in the literature. Type and timing of ideas as well as semantic similarity have been shown to cause fixation (e.g. Agogué et al., 2014; Jansson and Smith, 1991). It is with an awareness of these findings that we have chosen simply to present a few ideas of high specificity throughout the study and rather than the ideas themselves presenting inspiration for the ideator, the structure of the ideas themselves might inspire the ideator to structure their own ideas accordingly. To this effect we had originally designed this condition to present good and bad ideas to the ideator, however, this suffers from logic drawbacks. We specifically instruct ideators not to evaluate or judge the ideas they are writing down. The use of the words "examples of good ideas" and "examples of bad ideas" goes against the brainstorming rules by Osborn (1957) that state that you should not evaluate or judge the ideas you are writing down.

In the thematic roles (explicit) condition, ideators were given the same brainstorming rule instructions as well as instructions to generate ideas by incorporating all or many of the thematic roles. These instructions were given to the participants on both the instruction page as well as on the main ideation page, such that they could easily refer to these. They were given the instruction to generate as many well-formed ideas using the following structure as a method of producing well-formulated ideas:

- Who is performing...
- What action...
- With what instrument or object and...

- Whom is receiving or benefiting from this action and...
- Why is this being done.

They were told to incorporate all or most of the above elements into their idea in order to construct well-formed ideas. We were interested a few different findings. Firstly, whether the manipulation of prompts has worked. We did so by looking at the length of ideas in each of the three conditions to see whether we could find significantly longer ideas. We therefore performed this study with the following hypotheses in mind:

H1: Prompting an ideation task explicitly with thematic roles increases the well-formedness of the ideas generated, measured in the number of words in the idea.

H2: Prompting an ideation task implicitly with examples increases the well-formedness of the ideas generated, measured in the number of words in the idea in comparison to no-prompts, but not in comparison to explicit prompts.

H3: Prompting an ideation task explicitly with thematic roles decreases fluency in a set of ideas in comparison to no prompt conditions.

H4: Prompting an ideation task implicitly with examples decreases fluency in a set of ideas in comparison to no prompt conditions, but not in comparison to explicit prompts.

Granted the manipulation worked and the above hypotheses would be accepted, we hoped to show that the well-formedness of an idea has a significant impact on quality ratings for both the phone question and the health question, and a significant impact on our category coding scores for the health question. For these goals, we had the following hypotheses:

H5: Inter-rater agreement on novelty and value will be higher in the explicit prompt condition than in the no prompt condition.

H6: Inter-rater agreement on novelty and value will be higher in the implicit prompt condition than in the no prompt condition.

H7: Independent coder agreements of category classification will be higher in the explicit prompt condition than in no prompt conditions.

H8: Independent coder agreements of category classification will be higher in the implicit prompt condition than in no prompt conditions.

We did not make hypotheses about semantic difference score coding as this has already been shown to have a high inter-rater agreement. We performed this analysis nonetheless to explore whether this was still true for our conditions.

Participants

107 participants (58 female), age range 18 to 60 (M=29.44, SD=10.00[2]), were recruited from the online crowd sourcing website Prolific Academic[3]. Participants were pre-screened to ensure they were aged over 18 and fluent in English. Pre-screening further ensured that only participants using a laptop or desktop could access the study, in order to avoid any effect a mobile on-screen keyboard might have on thinking and typing times. 93% of participants were educated at A-level (or equivalent) or above. 56% of participants were educated at University undergraduate level or above. No further profiling constraints were applied to recruitment for this study; including whether or not the participants were experienced ideators, as this was not a pre-screening option. Participants were paid £2.81 for their time (approximately £5.00/hr, given average completion times of 33-34 minutes).

Materials

The study was designed in Gorilla (Anwyl-Irvine et al., 2018), an online cognitive psychology experiment environment. The input to the study consisted of a consent form, demographics questionnaire, main ideation screen and a three question post-experiment questionnaire (see all screenshots in appendix C).

Figure 6-1 shows the logic flow of the study. All participants started at the same node and completed the consent form and demographics questionnaire. The experiment then split all participants into three groups: no prompt, implicit prompt and explicit prompt. Once a participant was allocated, they were further allocated to one of the two questions - health or phone. Participants were given different instructions depending on the condition as well as different ideation screens (see figure 6-2). Depending on the condition they were assigned to they would either receive no additional information on the ideation screen (figure 6-2a), example ideas permanently staying on the screen (figure 6-2b), or a list of thematic roles to include in the idea (figure 6-2c). Once complete, all participants were directed to the same post-experiment questionnaire and thanked for their time. The experiment in Gorilla automatically redirected participants back to Prolific Academic for completion of study and payment.

[2] Estimated mean and standard deviation of ages based on the mid-point of ranges in which participants disclosed age.

[3] https://www.prolific.ac/

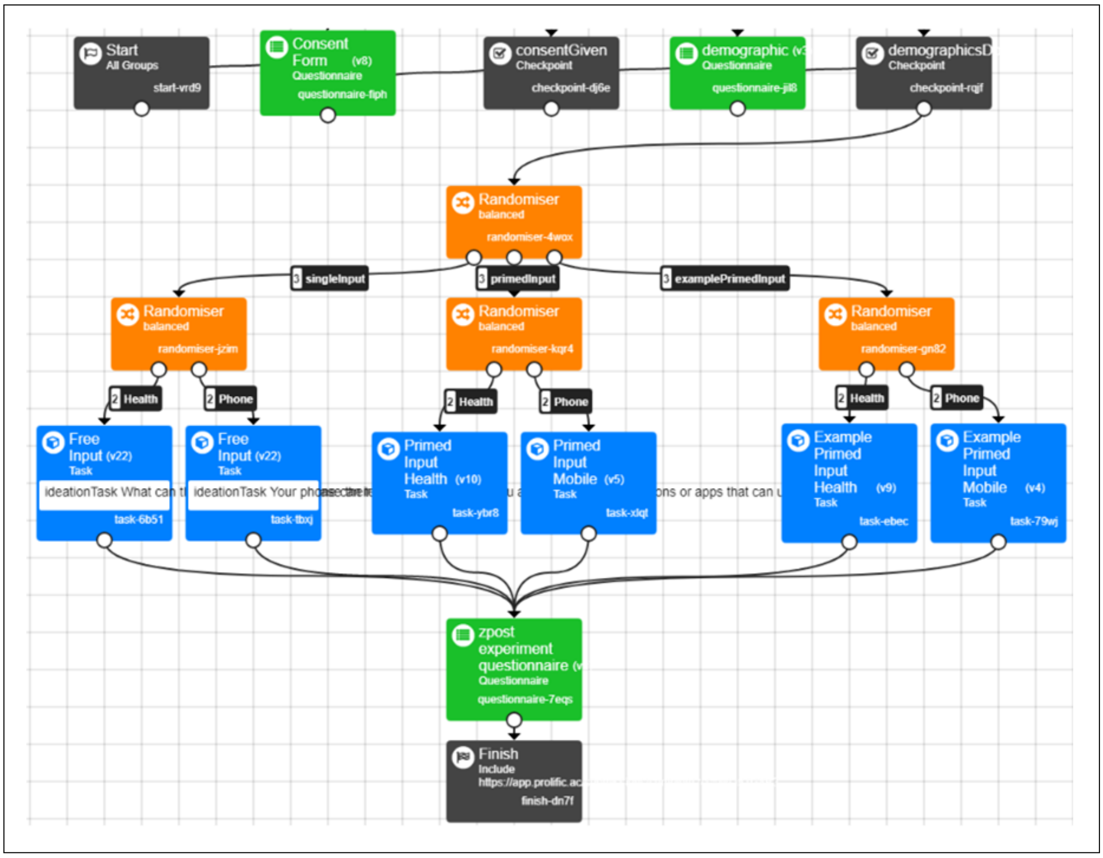
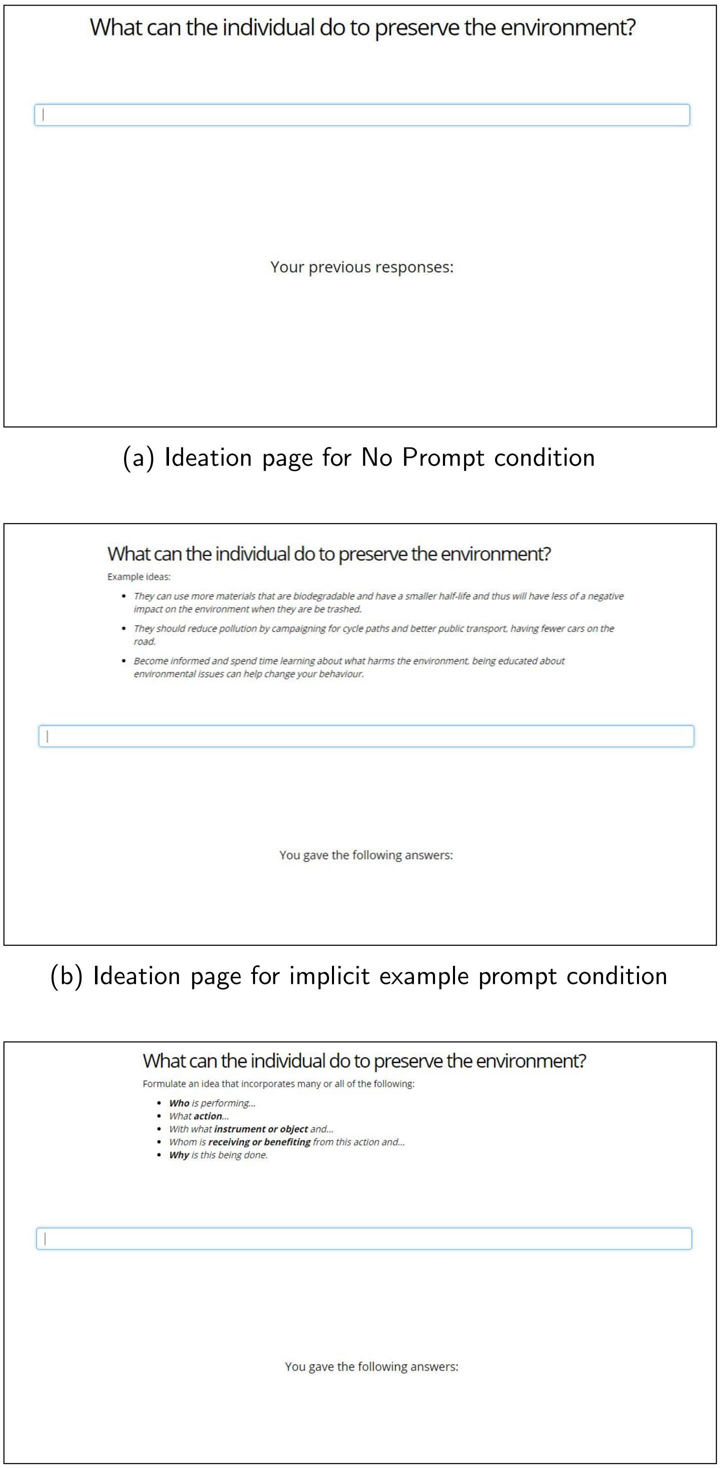


Figure 6-1: Study flow in Gorilla. Orange nodes represent randomisers that split participants first into prompt type then into question. Blue nodes represent the study screens themselves. Green nodes are questionnaire elements. Grey nodes control experiment logic (e.g. checkpoints verify that participants have not left the study before completion).



What can the individual do to preserve the environment?

Your previous responses:

(a) Ideation page for No Prompt condition

What can the individual do to preserve the environment?

Example ideas:

- *They can use more materials that are biodegradable and have a smaller half-life and thus will have less of a negative impact on the environment when they are be trashed.*
- *They should reduce pollution by campaigning for cycle paths and better public transport, having fewer cars on the road.*
- *Become informed and spend time learning about what harms the environment, being educated about environmental issues can help change your behaviour.*

You gave the following answers:

(b) Ideation page for implicit example prompt condition

What can the individual do to preserve the environment?

Formulate an idea that incorporates many or all of the following:

- ***Who** is performing...*
- *What **action**...*
- *With what **instrument or object** and...*
- *Whom is **receiving or benefiting** from this action and...*
- ***Why** is this being done.*

You gave the following answers:

(c) Ideation page for explicit thematic role condition

Figure 6-2: Ideation page in study 3, variation in prompts between the three conditions across health and phone question (note the question in these screenshots is that from the practice session.)

Procedure

Each participant was recruited through Prolific Academic and was redirected to the study hosted by Gorilla. Participants first saw a brief introduction to the study, what was going to happen, and the time commitment expected from them. Participants could then choose to consent to take part or not consent and exit the study. Once consent was given, the participant was asked for demographics: age range, gender and level of education.

Once demographics were submitted, they were shown full instructions for the study. These varied depending on prompt condition. Participants in the no prompt and implicit prompt were shown the same instructions (see figure C-4a in appendix C). These instructions matched how we instructed participants in the previous 2 studies. Participants were explicitly asked to generate ideas for 25 minutes according to Osborn's (1957) brainstorming rules. For the third condition, explicit prompts, participants were given the same instruction but elaborated with the 5 thematic roles we identified in Chapter 5 (see figure C-4b in appendix C); for each idea they were asked to consider "Who is performing...", "What action...", "With what instrument or object and...", "Whom is receiving or benefiting from this action and...", "Why is this being done?".

Once read, they were taken to a 2 minute practice task asking: "What can the individual do to preserve the environment?". Once they had completed the 2 minute practice task, they were shown the instructions again and asked to press enter when ready.

Each participant had 25 minutes to come up with as many ideas as they could in relation to the question they were given. The question (health, phone) was fully randomised by the application itself, as well as the condition they were in (no prompt, implicit example prompt, explicit thematic roles prompt). Once the 25 minutes were over, participants were asked to rate three questions (from Nijstad and Stroebe (2006)) on a scale of 1-10 about their ideation experience. These were the questions used in our first study, in order to gage the ideators self-awareness of failures (*impasses*). Despite not being able to replicate the results from Nijstad and Stroebe (2006), we included this questionnaire in the current study for consistency. The study questions were:

- "How difficult was it to keep on generating ideas?"
- "How often were you unable to generate ideas?"
- "How often did an idea you previously generated occur to you again?"

	Health Question	Phone Question
No Prompt	22	17
Implicit Example	19	20
Explicit Example	15	14

Table 6.1: Table showing uneven split of participants between conditions due to issues with the integration between the Gorilla-Prolific platforms.

Once the study was over, they were thanked for their time. Participants had a maximum completion time of 50 minutes, so as to not have them spend too long on any of the questionnaire pages. All results were screened, and participants were paid through Prolific Academic within 48 hours.

General analytic approach

The random assignment mechanism on Gorilla was designed to balance participants evenly across the 6 conditions, however it failed to do so. This was due to the integration between Prolific and Gorilla. Referring to the study flow in figure 6-1, Gorilla allows for checkpoints at different stages of the study. Should a participant not give consent or not pass the demographic checkpoint, Gorilla will immediately send back their information to Prolific, stating that no data was collected. In this instance, no money is taken from the experimenter account on either platform.

Should a participant successfully submit consent and demographics, they are allocated to a condition in Gorilla. After being allocated to a condition, if participants abandon the experiment before completion, the 50 minute experiment timer[4] will have to run out before Prolific is informed to recruit more participants. No money is taken from the Prolific experimenter account. Unfortunately, this is not the case for Gorilla, in which the participant has taken up a space in one of the 6 conditions, and money is taken from the experimenter account.

Due to the speed at which participants are acquired through Prolific, all 6 conditions were filled before the 50-minute experiment timer. As it was not possible to see which participants had abandoned the task before this, we ended up with an uneven split across conditions (see table 6.1). Due to financial- and time-constraints it was not possible to collect more data to rectify this issue.

The datasets were screened to see if all participants had followed instructions. In total, 13

[4] The amount of 50 minutes was set to allow participants to complete the ideation task and the questionnaire without being interrupted.

	Health Question	Phone Question
No Prompt	20	15
Implicit Example	18	17
Explicit Example	12	12

Table 6.2: Table showing split of participants between conditions after removal of 13 datasets in which participants had not followed instruction.

participants were removed from the final dataset. Some submitted ideas in other languages[5]. Others were not answering the question and solely submitted text irrelevant to the question or failed to press enter between ideas, therefore submitting a long string of ideas with no timestamps. Whilst these types of data are interesting in themselves, they did not adhere to the instructions and we were therefore not able to apply the same types of analysis as performed on the remaining participants data. The final total number of participants was 94 (see table 6.2).

2316 ideas were generated by 94 participants across the 6 conditions. 32 ideas were removed as they were either not ideas or they were duplicates of ideas[6]. Exclusion of these ideas left us with a final number of 2284 ideas. The entire dataset was summarised by *fluency*, *mean number of words* per idea per participant, *thinking time*, *typing time*, and *response latency* for each condition. The effect of the prompt manipulation was then tested by implementing two-way ANOVAs with significance level set at $p < .05$. The difference in productivity in the last five minutes versus the first five minutes was calculated per participant and then compared using paired-samples t-tests.

To verify the manipulation had an effect on coder agreement, we performed the three different coding tasks used in studies 1 and 2: category coding (of the health question only), semantic difference coding of both health and phone questions, and quality coding of both health and phone question. All independent coders for this study were selected as they were external to the process, in line with our previous two studies. The coders were volunteers and received no payment for their participation. Although all of the coders for this were familiar with other methods of analysis, all but one had no experience in coding this type of data before. The one with experience had coded ideas into categories in our first study. Prior to coding the full datasets, each coder, including the experienced coder, received full instructions. The coding was done on a sample of the data due to time

[5] This was flagged to Prolific as this goes against the profiling criteria

[6] Examples of submissions that were not ideas include clarifications on previous ideas such as self-correcting spelling, single letter ideas (one idea entered was simply "t") and submissions like "done" and "finished". Repeated ideas were adjacent ideas that were listed multiple times (e.g. one participant wrote "robots" 3 times in their trial).

constraints. It was not deemed necessary to code all the data in order to see whether there was agreement between coders.

For the category coding, a sample of 15 participants were selected, 5 from each condition in the health question. The mean number of words per idea was calculated for each of the three conditions, the 5 participants then chosen for being the closest to the mean in their condition. This resulted in a set of 480 ideas across 15 participants in the three prompt conditions. For the semantic difference coding and the quality (novelty / value) coding, the same sample of 15 participants from the health question was selected, as well as 15 participants from the phone question, selected on the same criteria. This resulted in a set of 480 ideas for the health question and 227 ideas for the phone question, or 707 ideas in total from 30 participants in the three prompt conditions. In the next section we present the results of these analyses.

6.2.2 Results and discussion

In the following section we present data from our study in six parts. The first is an overview of the data gathered with a summary of number of words per idea, number of ideas, thinking times, typing times, and response latencies for all ideas across each of the six conditions in our study. We furthermore present correlations of the post-experiment questionnaire responses with fluency. In the second part, we analyse our data using factorial ANOVAs, in order to see whether our manipulation of the prompt type and question had an effect on the fluency, length of ideas, and log transformed thinking times. The threshold for significance was set at $(p < .05)$. The third part presents the productivity over time per participant. We show a visual representation of the time-course of the ideation tasks in all 6 conditions. In the fourth part, we are interested in the categorisation scores performed by two independent coders for the health question only, and whether agreement scores (Cohen's k) have been positively improved as a result of the prompt manipulation. In the fifth part we show the inter-rater agreement scores between semantic difference score coders, to show that this method of rating ideas is reliable independent of the specificity of ideas. Finally, we show the agreement between two coders on the novelty and value scores for sampled data from all 6 conditions. Our goal is to show that whilst novelty and value are accepted methods of judging the quality of ideas, the subjectivity of the method may decrease with the length of the ideas themselves.

In table 6.3 we present an overview of the results for all 3x2 conditions. Mean values of *fluency*, number of words per idea, *thinking times*, *typing times*, and *response latencies* are calculated as the mean of means across participants in each condition. As in our first study, the high variance in fluency might be explained by the individual differences in ideators.

	No Prompt		Implicit Example Prompt		Explicit Thematic Role Prompt	
	Health	Phone	Health	Phone	Health	Phone
Number of participants	20	15	18	17	12	12
Mean fluency	41.5 (21.39)	23.53 (11.67)	26.28 (15.21)	13.47 (4.92)	19.5 (8.38)	13.75 (5.63)
Mean number of words per idea	6.1 (3.65)	18.25 (20.62)	15.06 (10.38)	20.02 (8.18)	17.89 (5.37)	21.39 (10.52)
Mean thinking times	35.2 (22.39)	58.1 (43.97)	63.07 (61.75)	67.38 (29.4)	53.84 (41.3)	88.23 (78.76)
Mean typing times	14.48 (19.64)	40.15 (42.91)	31.13 (29.72)	72.28 (43.98)	52.35 (20.5)	71.66 (43.63)
Mean response latencies	49.68 (33.76)	98.25 (82.13)	94.2 (65.26)	139.66 (56.85)	106.19 (47.58)	159.88 (96.24)

Table 6.3: Summary data from each of the 6 conditions in experiment 3. Number of words per idea, thinking times, typing times and response latencies are presented as the mean of means across participants, then condition. Standard deviations shown in brackets.

Looking at the mean fluency in the first 15 minutes of the task (M=30.52, SD=14.14), giving a rate of return of 2.04 ideas per minute, we might suppose that the financial incentive given to ideators in the current study has had an effect on the motivation to generate ideas faster. It would not be fair to compare to the full 25 minutes in this condition due to the likelihood of diminishing returns. Indeed, data from the full 25-minute set show a mean fluency of M=41.5 (SD=23.53), resulting in a 1.66 ideas per minute rate of return.

Comparing *fluency* to post-experiment questionnaire responses, no significant correlations were found between fluency and any question responses (Q1-fluency: $r(92) = -.05, p = .65$. Q2-fluency: $r(92) = -.12, p = .24$. Q3-fluency: $r(92) = -.10, p = .32$). This replicates our finding from both questions in study 1 and rules out the finding being due to small participant numbers. Interestingly, low positive but significant internal correlations were found: Q1-Q2: $r(92) = .28, p = .006$; Q1-Q3: $r(92) = .28, p = .006$; Q2-Q3: $r(92) = .18, p = .075$. Although Q1 was found to have a low correlation to Q2 and Q3, these were unfortunately not very highly correlated and we can therefore not report that we have been able to replicate the finding from Nijstad and Stroebe (2006) in any of our studies in relation to these questionnaire questions.

Effects of prompting ideation on fluency and specificity

A main interest and motivation for the manipulation done in this study was to see whether these firstly have an effect on the behaviour of ideators. The no prompt condition was comparable to studies 1 and 2. We hoped to find an improvement in the well-formedness of ideas in both the more implicit example driven prompt condition and the explicitly stated thematic roles driven prompt condition. In addition to the well-formedness of ideas, we were interested in seeing whether the formulation of these ideas (*thinking times*) varied with question or prompt conditions.

Conducting a two-way ANOVA comparing the effects of $PromptType \times Question$ on *mean number of words* per idea, per participant yielded a main significant effect for Prompt Type ($F_{2,900.959} = 3.478, p = .035$). Tukey LSD post hoc tests showed that *mean number of words* was lower in the no prompt condition (M=11.31, SD=15.25) than in the explicit prompt condition (M=19.64, SD=8.71; p=.010) but not in the implicit prompt condition (M=17.47, SD=9.84; p=.066). This evidence allows us to accept **H1**, as our manipulation had a positive significant effect on the well-formedness of ideas, but not **H2** as the implicit condition did not differ from the no prompt condition. The implicit and explicit prompt conditions did not differ. A main effect was similarly found for Question ($F_{1,1066.657} = 8.236, p = .005$). No interaction effect was found for $PromptType \times Question$ ($F_{2,336.470} = 1.299, p = .278$). Figure 6-3 shows the estimated marginal means of mean number of words in each of the 3 by 2 conditions.

Given that ideas were longer in the prompted conditions, we might assume that our manipulation had a negative influence on the fluency, as participants may have been spending longer formulating individual ideas in the implicit and explicit prompt conditions. A two-way ANOVA was conducted comparing the effects of $PromptType \times Question$ on overall *fluency* per participant. A significant main effect was found for Prompt Type ($F_{2,4378.431} = 11.338, p < .001$). Tukey LSD post hoc tests showed that *fluency* was higher in the no prompt condition (M=33.80, SD=20.26) than in the implicit prompt condition (M=20.06, SD=13.30; $p < .001$) and the explicit prompt condition (M=16.63, SD=7.86; $p < .001$), whereas the implicit and explicit prompt conditions did not differ. Again, we accept hypotheses **H3** and **H4** as this evidences a significant effect of the prompt conditions on *fluency*. A main effect was similarly found for Question ($F_{1,3354.162} = 17.371, p < .001$). No interaction effect was found for $PromptType \times Question$ ($F_{2,526.845} = 1.364, p = .261$). Figure 6-4 shows the estimated marginal means of fluency in each of the 3 by 2 conditions.

Conducting a two-way ANOVA comparing the effects of $PromptType \times Question$ on log transformed *thinking time* per idea, per participant yielded no significant effect for Prompt

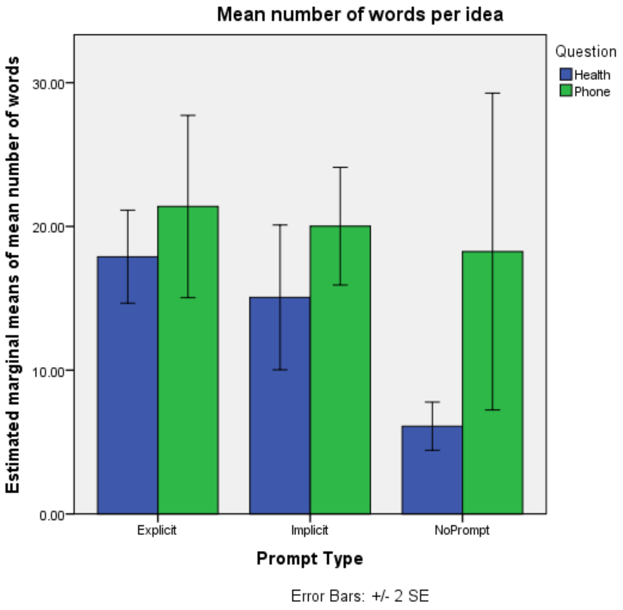


Figure 6-3: Estimated marginal means of *mean number of words*, averaged across ideas then participants, per prompt condition in each of the two questions in study 3.

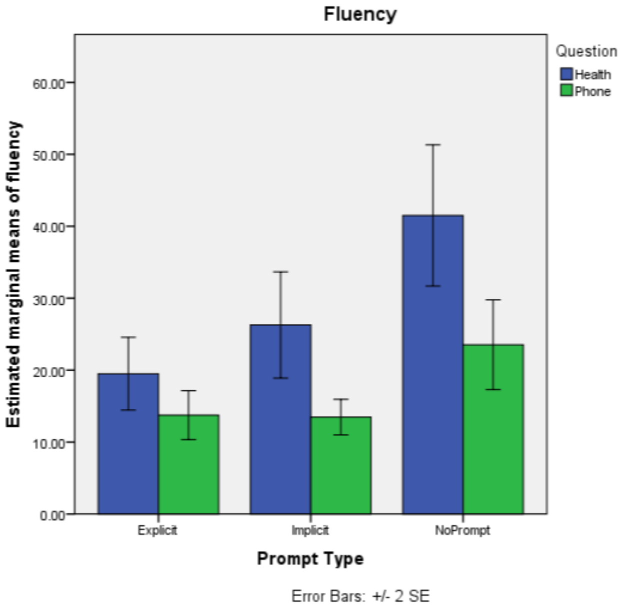


Figure 6-4: Estimated marginal means of *fluency* per prompt condition in each of the two questions in study 3.

Type $(F_{2,3.986} = 2.797, p = .066)$. A main effect was found for Question $(F_{1,11.360} = 15.946, p < .001)$ suggesting participants may have found the phone question more difficult than the health question, a result also reflected in the difference in *fluency* across the two questions. No interaction effect was found for $PromptType \times Question$ $(F_{2,.049} = .035, p = .966)$. Figure 6-5 shows the estimated marginal means of average thinking time per idea in each of the 3 by 2 conditions.

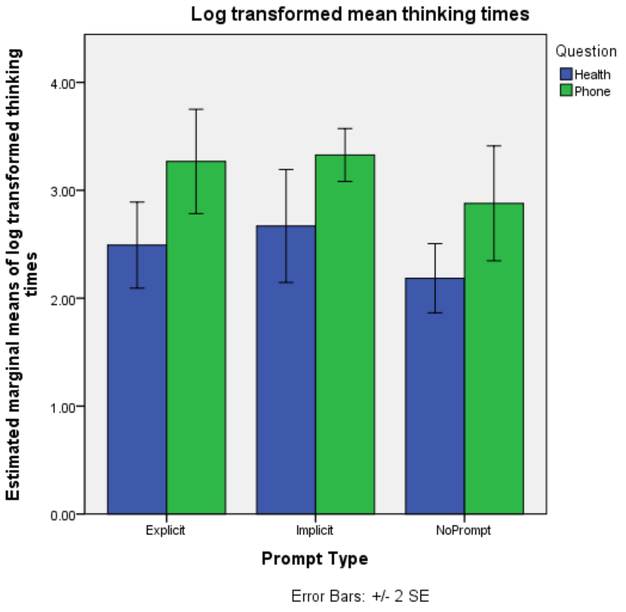


Figure 6-5: Estimated marginal means of log transformed thinking times per prompt condition in each of the two questions in study 3. A main effect for question is found, indicating people found the health question easier to generate ideas for in comparison to the phone question.

Time-course of productivity of ideation task

The cumulative productivity over time for each condition is shown in figure 6-6. The shapes of the curves suggest a slight diminishing returns as expected. These are seen more evenly in the health question cumulative productivity (left panel) in comparison to the phone question (right panel). In the latter, the overall returns appear to be decreasing over time, however, there are quite a few sudden jumps, possibly attributed to the lower number of ideas generated in these conditions.

The overall productivity per participant in the first five minutes of the task (M=8.28, SD=6.19) was significantly higher than that in the last five minutes of the task (M=3.85, SD=3.28) $(t(93) = 7.677, p < .001, d = .79)$.

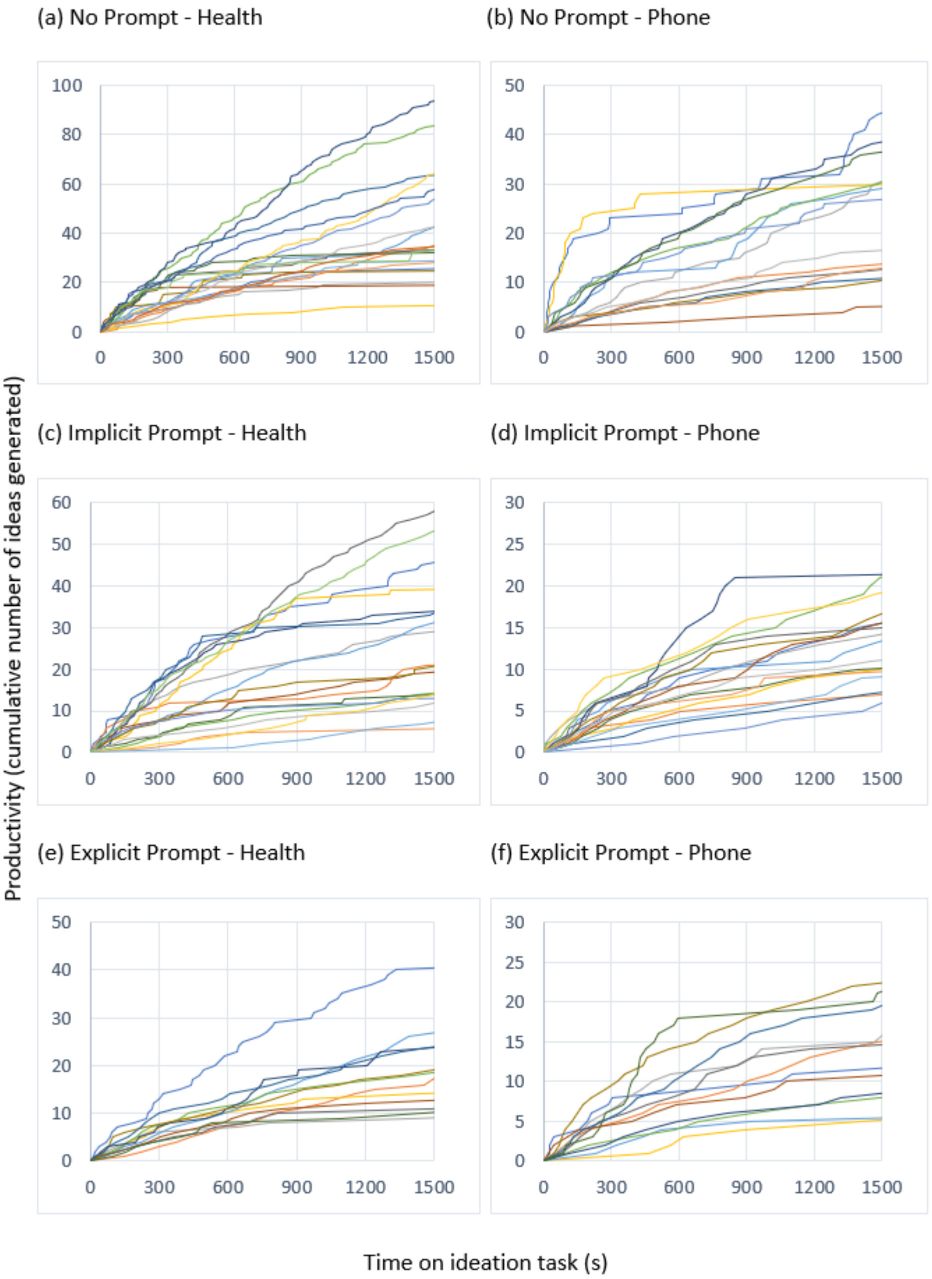


Figure 6-6: Cumulative productivity for each participant in the 6 conditions in study 3. All 6 conditions show a diminishing returns curve. Graphs on the left represent the health question in the three conditions: (a) no prompt, (c) implicit prompt, (e) explicit prompt. Graphs on the right represent the gain curves for the phone question in those same corresponding conditions.

Coding ideas by category: agreement between coders

Two independent coders were asked to categorise all 480 ideas, submitted for the health question by 15 participants across all three prompt conditions, 5 in each condition. Coder 1 is male aged 28, postgraduate student in the Department of Mechanical Engineering. This coder had performed this type of analysis earlier in our work. Coder 2 is female aged 26, postgraduate student in the Department of Computer Science, both coders from the University of Bath. This coder had no prior experience coding this type of data before. Prior to coding the full datasets, both coders received the following instructions (the same that were used in study 1):

They were instructed to classify each idea to a category using the means by goal matrix developed by Diehl (1991), also used in study 1 (see table 3.7 for full category matrix). For the health question, this consisted of a matrix of 11 goals and 9 means, resulting in 99 categories. Note that not all categories will be used by ideators. Coders were asked to enter the codes into three separate columns in a spreadsheet: the first being the goal, the second being the means and the third being any notes or additional goals/means they felt an idea could fall into. Each independent coder was sent the data and the means by goal matrix, coded a subset individually and met with the experimenter to discuss. Once the coders had shown full understanding of the meaning of the goals and means in the matrix, they were asked to categorise all 480 ideas.

Cohen's k was run to determine if there was agreement between the two independent coders. Overall across all three prompt conditions there was a fair agreement, Cohen's $k = .45, p < .001$, although lower than that in our first experiment for the health question.

Unfortunately, the prompt manipulation did not increase the kappa value for any of the three conditions (refuting **H7** and **H8**). Cohen's k of the no prompt condition (N=217) was $.45$, in the implicit prompt condition (N=152) was $.47$, and in the explicit prompt condition (N=110) it was $.42$. We can therefore not say it is the length of ideas that affect the inter-rater agreement scores for the categorisation system.

Interestingly, comparing the coding protocol of coder 1 versus coder 2, it was possible to pick out apparent flaws in the coding system itself, the vagueness and overlap in the schema. Short ideas, such as "Donate blood" was assumed as mental health (feeling good about yourself) by one and health prevention by another. "Eat more fruit" was coded differently as well, where the same mean applied (food intake) but no explicit goal is expressed, leaving it up to the coder to assess whether the ideator meant "avoid weight problems", "optimize metabolism", or "practice health prevention". A longer idea "[replace] fizzy drinks with water to help improve her health as well as improve her teeth and general lifestyle" was coded as generic health prevention by one but maintain healthy teeth by another. The idea was specific enough to be clear about the goals, but did not follow the schema in such a way that it could be easily coded into a single category. Finally, "Do not overwork yourself" was seen as a mental health goal by coder 1 whereas coder 2 saw this as an avoidance of physical strain goal; both goals seem plausible.

These findings replicate our findings from study 1b and underpin the issue that whilst people might ideate within categories, these are not well defined. Indeed, the definition of these categories might vary from person to person, depending on what they consider as being connected. This does not negate the SIAM model, it simply reinforces the notion that the connections made from an image in semantic memory might differ dependent on experience.

Coding ideas by semantic distance: agreement between coders

Two independent coders were asked to rate the difference between adjacent ideas (e.g. ideas that follow each other) on a 10-point scale. They were asked to rate the same subset of ideas as rated in the categorisation section. Coder 3 male aged 62, professor in the Department of Computer Science at the University of Bath, coder 4 was female aged 30, postgraduate student in the Department of Psychology at the University of Bath. Whilst both coders were familiar with data analysis methods, neither coder had prior experience in coding this type of data. Both coders were given all 707 sample responses, from 30 participants, across the 6 conditions. This resulted in $N-1$ datapoints per participant, where N is the number of ideas in the participant's dataset, due to the difference score being between two adjacent ideas and not for every idea. Both coders were blind to prompt condition but informed of that there were 2 questions, as they were shown the responses for the two questions separately. They were given the same instructions, including examples, as were given in study 1 (see section 3.4.1):

The following would score a 1 or 2, as they are essentially touching on the same concepts: recycling and plastic materials.

- Idea 1: "recycle plastic bottles"
- Idea 2: "recycle plastic bags"

These would score a 5-6 as they have something in common but aren't quite the same.

- Idea 1: "recycle plastic bottles"
- Idea 2: "teach people why recycling plastic bottles saves the planet"

The following would score a 10 as they have nothing in common:

- Idea 1: "recycle plastic bottles"
- Idea 2: "support an organisation that protects pandas"

Coders were asked to individually talk through the first 10 semantic difference ratings they performed with the experimenter. Once the coders showed full understanding of the process, they were asked to complete the remainder of the difference ratings themselves. As difference score rating of the ideas was done on a scale, not categorically as in the previous study, inter-rater agreement was calculated using Pearson's correlation between the responses of the two coders.

Overall agreement between the two coders was positive and fairly high $r(675) = .645, p < .001$. Figure 6-7 shows the mean difference ratings submitted by coder 4 compared to the ratings submitted by coder 3.

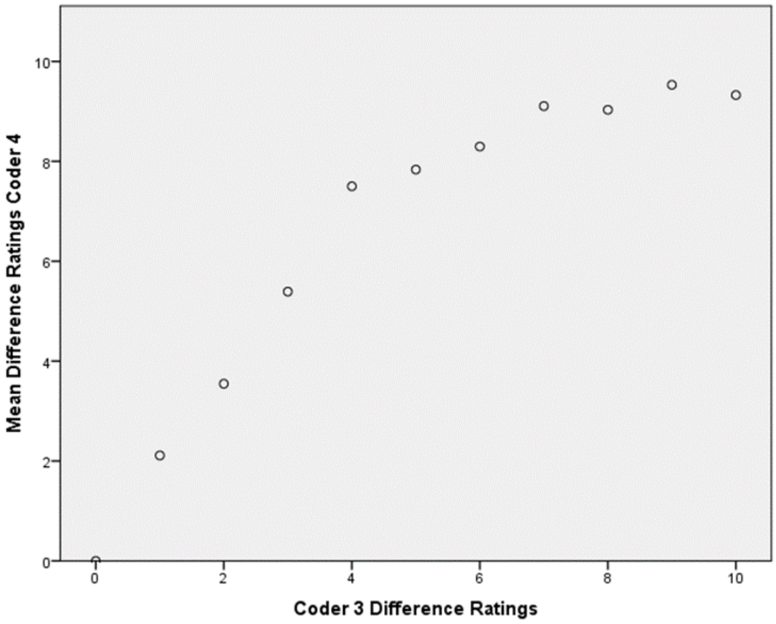


Figure 6-7: Summary of coder 4 difference ratings against each of coder 3's difference ratings on the 10-point scale for sample data in study 3.

Table 6.4 shows the correlations between coder scores for each of the conditions. This shows a decreased correlation in the implicit example prompt. Overall, correlation remains distinctly high and comparable in all 3 prompt conditions, showing that semantic difference scores remain unaffected by the well-formedness of ideas.

	Health	Phone	Total
No prompt	.736	.695	*.686*
Implicit example prompt	.602	.570	*.593*
Explicit thematic role prompt	.771	.586	*.669*
Total	*.669*	*.636*	

Table 6.4: Correlations between coder 3 and coder 4 on each of the 6 conditions in study 3. All correlations are significant at the 0.01 level (2-tailed).

Coding ideas by novelty and value: agreement between coders

In the following section we present the results of quality coding performed by 2 coders, in order to observe whether the manipulation intended to affect specificity has had an effect on inter-rater agreement. Two independent coders were sent the same sample of data (707 ideas across 30 participants in the 6 conditions) used in the difference scores section and asked to rate the novelty and value on a 10-point scale. Coder 5 was male aged 28, postgraduate student in the Department of Computer Science. Coder 5 had no prior experience in coding this type of data. The other coder for this task was coder 1, male aged 28 and postgraduate student in the Department of Mechanical Engineering, who had previously coded the set of the data into categories.

The coders were asked to individually rate every idea on two metrics: *value* and *novelty*. They were asked not to meet or discuss their ratings as this would invalidate the ratings being independent. These were to be rated on a scale of 1-10, 1 being "no value/not novel" and 10 being "high value/very novel" (full written instructions were the same as in study 2, see appendix B.5). These were for the purpose of this study defined as:

- Novelty: how original and surprising is the idea?
- Value: how useful and practical is this idea and does it make sense as a solution to the problem?

Prior to coding the full datasets, the two coders received thorough instructions and were asked to talk through their coding to show understanding. Training in the use of the novelty and value scales consisted of verbally going through each point on the scale, just as was done for study 2 data in Chapter 5. Each point was assigned to an idea in the data sample. Training was done with each coder individually. Once each point on the scale had been explained, the coder was asked to rate a sample of the data and talk through their choices, in order to show that they had understood how to use the scales.

In table 6.5 we show the overall correlations between the two coders across all 30 sampled participants for novelty and value. These correlations are further broken down by condition. Interestingly, novelty correlations were considerably higher in all 3 conditions, and in the free input condition, these have particularly improved, in comparison to the novelty agreements found in study 2 (r values ranged from .38-.51). No difference in ratings between prompt conditions was seen, refuting **H5** and **H6**. Value ratings did not increase in comparison to the ratings in study 2 (r values ranged from .23-.36). We theorise that this lack of variation in agreements across conditions could be due to a priming effect on coders from having access to the full set of well-formed and abstract ideas. Although no consideration was given to order effects in coding these, coders were instructed to read through all ideas first to get a sense of the domain before judging novelty.

We were interested in observing an increase in the quality of ideas in the conditions in which we offer prompts. We observed novelty and value separately and that well-formed ideas will have higher mean novelty and value scores. We also predicted a higher variance of qualities due to ideas being more thought through than lower quality ideas, which we would presume may have a lower range of quality scores due to the abstract nature of the idea.

Conducting a two-way ANOVA comparing the effects of $PromptType \times Question$ on mean value per idea, per participant yielded a main effect for Prompt Type ($F_{2,5.656} = 3.533, p = .045$) where mean value for explicit prompts = 7.23 (SD=.66), implicit prompts = 6.22 (SD=.89), and no prompt = 6.41 (SD=1.02). No main effect was found for Question ($F < 1$). No interaction effect was found for $PromptType \times Question$ ($F_{2,.313} = .121, p = .886$).

	No Prompt	**Implicit Prompt**	**Explicit Prompt**	**Overall**
	N = 315	**N = 222**	**N = 168**	**N = 705**
Novelty	.744*	.592*	.583*	**.665***
Value	.355*	.402*	.148	**.329***

Table 6.5: Correlations between the two quality coders on novelty and value. Novelty correlations have increased in comparison to previous study data, however, this has increased across all 3 conditions in the current study. *denotes correlations significant at the $P < .01$ level.

Conducting a two-way ANOVA comparing the effects of $PromptType \times Question$ on the standard deviation of value per idea, per participant yielded a main effect for Prompt Type ($F_{2,.900} = 4.979, p = .016$) where SD value for explicit prompts = 1.17 (SD=.22), implicit prompts = 1.57 (SD=.23), and no prompt = 1.53 (SD=.42). No main effect was found for Question ($F < 1$). No interaction effect was found for $PromptType \times Question$ ($F_{2,.215} = 1.08, p = .355$).

Whilst we recognise this is a small sample, this evidence of higher mean value and lower variation in value in the explicit prompt condition is indeed a promising finding, as it shows prompting ideators with thematic roles as a small device for supporting ideation.

Conducting a two-way ANOVA comparing the effects of $PromptType \times Question$ on mean novelty per idea, per participant, we do not find the same effects. No effect for Prompt Type ($F_{2,4.978} = 1.931, p = .167$) where mean novelty for explicit prompts = 4.35 (SD=1.02), implicit prompts = 5.20 (SD=.89), and no prompt = 4.32 (SD=1.41). No main effect was found for Question ($F_{1,3.127} = 2.426, p = .132$). No interaction effect was found for $PromptType \times Question$ ($F < 1$).

A similar lack of effects was found of $PromptType \times Question$ on variance in novelty per idea, per participant. No effect for Prompt Type ($F_{2,.467} = 1.077, p = .357$) where mean novelty for explicit prompts = 4.35 (SD=1.02), implicit prompts = 5.20 (SD=.89), and no prompt = 4.32 (SD=1.41). No main effect was found for Question ($F_{1,.032} = 1.479, p = .236$). No interaction effect was found for $PromptType \times Question$ ($F < 1$).

6.3 Chapter summary and conclusions

In this chapter an experiment was reported that manipulated the way ideators were prompted on the screen where they enter and list ideas generated as solutions to the specified single problem. To test the idea that unprompted ideas tend to vary in level of specificity and that this causes problems of quality and the judgement of quality, we compared these with ideas prompted by (1) examples of well-formed ideas and (2) a list of the thematic roles that our earlier work has shown were more present in good and agreed high value ideas. This manipulation was successful in showing that the prompts led to longer thinking times and thus fewer ideas, but longer ideas. Unfortunately, the effects of the manipulations on the judgements of the properties of the ideas was harder to understand.

First, assigning ideas to categories was done even more weakly than in the earlier studies (on the same question). This result was hard to interpret but might suggest that the problem lies with the categories themselves rather than the varying length of ideas. Second, judgements of semantic difference were, if anything, slightly reduced although remained at an acceptable level. Finally, judgements of novelty did show the hoped for improvement in agreement between coders. However, this improvement was not limited to the prompted groups. We speculate, perhaps optimistically, that the presence of longer and better formed ideas facilitate the judgement of all ideas. There were no such effects on value judgements. The value of ideas as measured as the average of the two coders ratings was significantly positively affected by the prompts, the highest value ideas were generated in the thematic role prompted condition. Given the size of the sample analysed by coders, this is a limited finding but promising nonetheless.

Chapter 7

Discussion, conclusions and implications

In this chapter we summarise the progress made during the thesis as well as noting the limitations of the work done and possibilities for future work. The investigations of this thesis were originally aimed toward developing theory in support of applications that could assist in the ideation process. By investigating the time-course of an ideation task, whether continuous ideation in response to a single question or ideation in response to multiple sequential questions, and how people allocate their time, we hoped to develop a stronger understanding of the concept of impasse and whether people in general use reasonable strategies to overcome this cognitive state (e.g. switching to a different semantic space or category, in the case of a single task, or a more productive question in the sequential ideation task). In the studies performed to inform this theory, we encountered methodological difficulties inherent in the measurement of ideation, essentially thwarting the original intentions of this research. Addressing these methodological issues became an end in itself. The results of the studies in this thesis weakly support the notion of Search for Ideas in Associative Memory (SIAM) and show promising results towards our hypotheses that ideation follows a two factor information foraging style heuristic, involving giving-up times and the tendency to switch upon sub-goal completion. We were, however, unable to fully inform ideation theory, due to the methodological problems encountered. Through this research, the concept of impasse still remains a question to be answered. The method of ideation research has come into question as it does not seem as straightforward to analyse data that does not conform to a specific format. This methodological worry has itself led to suggestions for a design intervention primarily for methodological reasons, but which may also have practical relevance. Although these have not been developed, results from our analysis of study 2's data and the manipulations done in study 3 offer intermediate-level knowledge, knowledge that sits between theory and specific instance (Höök and Löwgren, 2012). These manifest in the form of design insights into the development of ideation tools that we may draw upon to support the increased value in ideas generated, as well as methodological insights involving asking ideators to self-rate and reformulate their ideas.

This chapter is split into three sections. In the first section, we discuss the studies presented in this thesis, the most prominent results and how these contribute to the overall literature on creative problem solving and ideation. We present the overall limitations of our studies and highlight how some design decisions made throughout these studies might have impacted the ideation and coding processes themselves. In the second section, we highlight the practical relevance of our key findings, and present the implications for the design of ideation tools. In the third section, we offer a summary of the thesis and concluding remarks.

7.1 General discussion

In this section, we discuss the work performed in this thesis as a whole and the contribution of each chapter. Throughout, we reason for the design choices in the studies and explicitly highlight the wider limitations of this thesis and how future work might address these.

This thesis began with a literature review, which expanded the definition of ideation tasks with respect to various empirical literatures in Psychology, before considering recent work in Human Computer Interaction research with concerns very much in line with those of this thesis. Ideation tasks, defined as the process of coming up with ideas in response to a loosely constrained and ill-defined problem, are identified as being an important part of everyday life. The significance of supporting such tasks is seen in a range of literature in Psychology, Engineering, Management and Human Computer Interaction. Ideation is recognised as a key step in creative problem solving, in which we perform several iterations of ideation-evaluation. We do this as a way of widening the number of possible solutions to a problem (ideation) and then narrowing these down to a few viable options (evaluation) (e.g. Basadur, 1995; Dennett, 2017; Ochse, 1990).

Among the main lessons from this literature review is that support for ideation may be informed by better understanding the time-course of ideation – in order to better understand cognitive processes underlying creative problem solving. The concepts of impasse and failures have been studied in an attempt to understand how we might better support ideators increase their productivity. Whilst not explicitly stated in the literature reviewed, the ability to offer support when ideators run out of ideas is beneficial in time-bound ideation tasks.

Indeed, in more recent research on methods of supporting ideation, Chan et al. (2017) studied the concept of impasse, as a way of understanding what types of support we might offer. One of the primary theories of ideation, Search for Ideas in Associative Memory (SIAM) makes the claim that ideas generated in an ideation task will be organised in clusters by semantic categories. Additionally, that these will be temporally clustered in such a way that time to formulate ideas within a cluster will be shorter than the time to formulate ideas between clusters. We covered the close relation of this theory to information foraging theories in the literature review. *Information foraging theory*, derived from optimal foraging theory, is widely influential in Psychology and Human Computer Interaction research, but has not, to our knowledge, been applied to any work in ideation, including those in HCI. Its application to understanding the time-course of ideation is one of the primary novel aspects of this thesis. Having reasoned for an *information foraging* perspective (through its link with the SIAM model of idea generation) on the time-course of ideation, three main experimental studies were reported, in which individual ideators produce lists of ideas on a computer. An understanding of the cognitive processes that take place during ideation, and specifically when ideation becomes less productive, might help us understand the occurrence of impasses better, as well as how to better support ideators when this happens.

Prior to our discussion of the studies performed in this thesis, their contributions and overall limitations, it is worth pointing out the choice made for the design of these studies. Ideation and creativity research brings a certain level of complexity; there is a wide range of ideation styles, methods for structuring solutions, existing support tools, and settings in which ideation can be performed. The work in this thesis sits within the research on ideation that takes an experimental approach to this type of creative thinking task. Whilst we need to acknowledge that this approach to ideation may have a negative effect on the spontaneity of creative thought, detailed experimental analysis allows us a certain level of control over the variables involved in the complexity of creativity. As in much of the reviewed literature (e.g. Chan, Dang and Dow, 2016; Ye and Robert Jr, 2017; Oviatt and Cohen, 2010; Runco and Sakamoto, 1999), we treat ideation here as a time-constrained task in isolation. Whilst it might be argued that this takes ideation away from its broader context, some studies point out that this is not wholly void of validity as time-constraints often present themselves in the real world (Allen and Thomas, 2011). As a result of taking this approach, we are able to develop a *general* understanding of the thought processes that occur during ideation. The studies presented here focused on time-bound ideation tasks in which individuals electronically generate ideas by brainstorming in relation to an open ended question. The choice to study individual ideation was partly founded in the idea that we may be better able to control certain variables that affect ideation. Group ideation offers the benefits of multiple view-points and the ability for ideators to synthesise and build on each other's ideas. Indeed, under certain conditions, brainstorming in groups may result in increased productivity (see e.g. Dennis and Valacich (1993), in which computer-mediated electronic brainstorming in groups of 12 or more resulted in better performance than individual brainstorming). Group ideation may unfortunately also suffer from the effects of evaluation apprehension, production blocking and free riding (Nijstad and Stroebe, 2006); all factors that could affect the flow of idea generation. Controlling for such influences was vital in order to perform a more focused analysis of the occurrence of impasses. Whilst our work looks at individual brainstorming, we also inform the research on nominal group brainstorming, in which ideas generated separately by ideators are pooled together as though they were generated in a group (Taylor, Berry and Block, 1958; Goldenberg, Larson Jr and Wiley, 2013).

We further acknowledge that ideation itself can be done in many different ways, with brainstorming being one of the most popular methods, possibly due to experimenters strong control over variables in this method. We use brainstorming as a representative method of ideation, however, as covered in the literature, there exists a range ideation methods that can be used such as analogy to existing systems (Silverstein, Samuel and DeCarlo, 2013), SCAMPER (Eberle, 2008) and mind-mapping (Buzan and Harrison, 2010). Each of these methods themselves have benefits for specific use in design ideation, as each of them offer different tools to help guide thinking (e.g. mind-mapping affords a hierarchical way of structuring ideas; analogy brings in design ideas from an external source as inspiration). We will discuss this further in this chapter in the discussion of the results of the manipulation performed in study 3.

7.1.1 The time-course of ideation and support for well-formed ideas

In this section, we summarise the studies performed in this thesis and highlight key findings and contributions. Our first study, the environment and health questions in studies 1a and 1b (Chapter 3), was a replication of the studies run by Nijstad, Stroebe and Lodewijkx (2002) (and subsequently Nijstad, Stroebe and Lodewijkx (2003) and Nijstad and Stroebe (2006)) in which we attempted to replicate SIAM's fundamental prediction, i.e. that the list of ideas produced by an ideator in response to a problem will tend to be categorically organised, with adjacent ideas more likely to belong the same category than more distant ideas. In addition to the study replication, in which the original authors of this study recorded solely the overall response latency for each answer given, we provided a further depth to the measure of timing. Our split of response latency into thinking and typing times provide a novel (if simple) contribution to our knowledge. We analysed these in accordance with the categorical structures laid out by SIAM model in order to test the predictions that foraging theory could somewhat explain the time-course across ideas.

These studies exposed a methodological problem - the first of several in the thesis which forced additional investigations somewhat away from the original plan of work. The methodological problem presented itself in the disagreements between independent coders who did not agree on the category allocation of the expressed ideas. Comparing this finding to Nijstad and Stroebe (2006), this is surprising as they performed the same study with the same questions and identical category scoring. Our version of their study was not devoid of limitations. In the first instance, we are unable to make comparisons to the ideators in the original study as profiling constraints were not reported. The ideators recruited in our study were not professional designers; the noisiness in the data might possibly have been avoided by the inclusion of professional designers who are aware of the need to generate many ideas over evaluation and self-censoring. Additionally, our data was coded by non-expert coders. Finally, although this might not have had an effect on the noisiness of the data, we ran this study using a 15-minute time-limit, rather than the 20 minutes stipulated in the original study. We will address these general limitations later in this chapter, as they relate to all three studies.

Despite the disagreements between the coders, who categorised ideas into previously generated categories, if we analyse the category coding for all four coders individually, we do find weak evidence for categorical structure in participants' protocols (IC1 and IC3), a replication of the results in the original study by Nijstad and Stroebe (2006). In addition to this finding, we observed further that our novel split of thinking and typing time was successful in showing that the categorical structure affects participants' thinking times (but not typing times), with category switching taking longer on average than generating ideas from the same category (within category thinking times). Whilst response latency might be significantly different between and within categories, our results suggest that thinking time might be a more refined way of showing that we indeed do spend time switching to new images in our mind.

The category judgement problem prompted us to formulate a new and, as it transpired, more reliable approach to measuring the thought process during an ideation task: i.e. asking coders to rate the semantic distance between adjacent ideas. This method was performed much more reliably by independent coders, with agreement scores over 70%, offering a promising alternative for analysing semantic effects in ideation. In addition to this positive finding, we observe that semantic differences predict the temporal intervals between ideas well. This observation offers a novel contribution to the understanding of cognitive processes in ideation. Our finding is somewhat in support of the SIAM model, but makes a shift from the rather binary view of ideating within categories. Instead, it is suggestive of a more fluid associative ideation method, in which an idea within the same image might have associations to the next idea, but not be within the exact same semantic category.

Our second experiment was designed specifically to allow a more refined test of patch-leaving rules. The sequential-idea design of this study has not been formulated in ideation studies before, to our knowledge. This style of study is often used in Psychology literature for accomplishment-style tasks such as Scrabble tasks, in order to study foraging theories in relation to task-performance (e.g. Wilke et al., 2009). The timing effects in the data from study 1 allowed some limited, speculative insight into the foraging rules that may underpin category switching (under the analogy that a category is a patch that an ideator chooses to abandon). However, the noisy nature of category switching compromised our ability to perform reliable analyses of this. Participants were asked to generate ideas in response to multiple sequential distinct problems. The sequence and therefore potential of the problems was not known in advance; participants were free to decide when to abandon any problem and move on to the next. This design allows us to measure *giving-up times* and to test whether any of the foraging theory heuristics might underpin the ideators decisions to quit tasks.

Analyses of the time-course of ideas in this second experiment followed quite closely the analyses of anagram solutions in Payne, Duggan and Neth (2007) and conclude similarly that participants' switch decisions are not governed simply by the number of items or the time on a task. Neither are they governed solely by the giving-up time (i.e. a kind of patience threshold of time since the last idea came to mind). Rather, participants seem to be sensitive to the rate of generation of ideas on a task. Furthermore, participants occasionally switched to a new task immediately on generating an idea: a finding which directly mirrors Payne, Duggan and Neth (2007) model of discretionary task switching. This is an interesting finding in itself as, although these are two different types of task, the concept of sub-goal completion applies as a prompt for leaving a task in ideation tasks as well. Indeed some participants indicated that they did make a switch as a result of submission of what they believed was a really good idea. These are in striking contrast to those who said they eventually just switched question because they felt they couldn't think of anymore good ideas. The insights from the initial timing analysis of study 2 give us a novel contribution to ideation research, namely that we might be able to explain ideator behaviour, of switching from one task to another, using foraging theory. This might not be that surprising a finding, as these theories have been able to offer insight into other types of task, for example accomplishment-style tasks. A positive outcome of this finding is that it suggest we could further study the lessons from the research on other problem solving style tasks in order to understand the cognitive processes that take place during ideation.

All the analyses of experiment 2 depended simply on the number of ideas and their time-course. In addition to these simple measures, we considered other measures that might have an effect on switching strategies. Surely, quality of ideas might also matter, both for participants who monitor their own successes and for researchers who would wish to support ideation. In trying to analyse quality, a second major methodological awkwardness was uncovered. Three coders were asked to judge the *novelty* and the *value* of ideas - two aspects of quality that are widely thought to be important and independent. On neither aspect did the coders agree very closely (they did agree significantly, but the correlation between coders' scores was only around 0.3).

Rather than regard this as merely a methodological nuisance, we reasoned that it likely exposed a general difficulty with ideation tasks: they leave it up to participants to choose how specific or vague they choose to be in their description of the proposed solutions (generated ideas). Indeed, the style of ideation task often used in this type of question is an open-ended, fairly broad, question that requires little to no subject specific knowledge - the openness of these types of question do not afford a specific length or elaboration in response. If specificity is low, it is hardly surprising to find that judged quality is variable. Founded in comments from coders and from looking at the data, this hypothesis was speculative, but it suggested a design intervention that seemed important for the practical ambitions of the thesis - to support ideation - as well as for the theoretical aims - to understand it.

Some simple and limited exploratory analyses were conducted to test the speculation that specificity or well-formedness might explain the low agreement in the quality ratings. We analysed the 20 highest agreed, 20 lowest agreed and 20 most disagreed quality scores, discovering that ideas which coders agreed were good were significantly longer and contained significantly more thematic roles than low-quality ideas or those that coders disagreed about.

These findings suggest that ideation might be improved by a user interface which encourages ideators to construct better-formed ideas. The third main experiment tested a rather simple implementation of this idea in a between-groups experiment. This experiment, and the variations therein, was of our own formulation, specifically designed to test the hypotheses that 1) we could increase the specificity of the ideas and 2) that better formulated ideas increases the agreement between coders - essentially making the ideas less ambiguous and less open to personal interpretation. The general study involving a single open-ended task followed a similar structure to that of regular electronic brainstorming studies, however, the instruction was manipulated in order to test our hypotheses. One group of participants produced ideas under instructions identical with those of the first studies. A second group was presented with well-formed example ideas above the ideation window, visible throughout the ideation process. These well-formed ideas were selected as they contained 4 or more of the thematic roles found in agreed good ideas, however, they did not explicitly mention these anywhere during the experiment for this group. A third group were asked explicitly to construct ideas which included all or many of a list of thematic roles that remained visible throughout. The list of semantic roles was transformed from "agent, theme, instrument, etc" to easier, non-technical constructs ("who is performing", "what action", etc). Despite the simplicity of this manipulation, these prompts had significant effects on ideation: prompted participants produced fewer and longer ideas.

The quality of the ideas generated by a subset of participants (10 per prompt condition) was rated for quality (novelty and value) by two coders. Agreement scores for these coders showed higher agreement than in experiment 2, but perplexingly this improvement was not limited to the prompted conditions. The improvement might have been a chance effect, but it is also possible that the presence of longer, better-formed ideas somehow moderated or influenced the judgment of all ideas. Indeed, the coders were informed to read through the entire set of ideas before beginning the novelty and value ratings, an instruction often used in these judgement tasks.

The judged novelty of ideas did not vary across the three prompt conditions. However, the ideas of those participants who were prompted with thematic roles were judged to be of significantly more value (and significantly less variable value) than those of the other conditions. This effect should be treated as suggestive rather than definitive as it is a small study (due to the high time-cost of judging quality, we limited the sample of judged ideas to those from only 30 participants), but we hope that it illustrates the main ambition of the thesis: through experimental analysis we may gain insights into interface interventions through which ideators might be supported. Referring back to our discussion of different types of ideation, brainstorming need not solely be a tool to generate many ideas, regardless of quality. Indeed, we may be more interested in the value of an idea than its novelty (e.g. Reinig, Briggs and Nunamaker, 2007; Briggs and Reinig, 2010; Howard-Jones and Murray, 2003), and if this is the case, our work suggests that we may be able to offer methods of support for more valuable ideas to be generated.

7.1.2 Limitations and considerations for future work

The studies in this thesis are not without limitations. We recognise that the selected participants do not fully represent the average population. However, despite their affiliation with a university, they may be somewhat representative of the average inexperienced ideator. More experienced ideators, for example Engineers with training in methods like Triz and SCAMPER, professional designers and people who ideate for a living will likely perform differently on ideations tasks such as these. Whilst we have attempted to understand general cognitive processes with little attention to individual differences, it is worth mentioning here that these results might have looked very different had we used experienced ideators. We might have found ourselves with less noisy data consisting of well-formulated ideas. Considering this, it would be interesting to see these studies performed using expert ideators.

At the very start, we ran into methodological issues with the categorical schema and the noisiness of the data. Whilst we have discussed the possible issues with the pre-determined categorical schema, we continued to experience issues with coders when rating the novelty and value of ideas. We therefore need to bring attention to another limitation in our work: the coders recruited for this study. Coders were selected based on availability, as we unfortunately did not have access to experienced coders. Additionally some of the coders from early studies were not included in subsequent studies. This was due to a range of reasons: these studies were performed over a long period of time and some were simply no longer accessible; coding for all but the novelty/value ratings was done on good will and therefore not paid; some coders were unable to provide more time than coding a single set of data would take. We highlight this as a serious limitation as this may have affected a large part of the work performed in this thesis. As before, it would be interesting to see these codings performed using expert coders.

A further prominent limitation of these studies is the use of coding schema and novelty and value ratings. The categorical schema was possibly too vague and not appropriate for use in these tasks, despite having been used successfully in other studies. An obvious questions this raises might be: why do we not simply create a new schema? This would be interesting, and has been done in other studies (e.g. Chan et al., 2017). In our work, an attempt at a restructured schema was made, however, this took a turn from the main research goals and eventually it was decided to move towards an analysis of the methodology itself. The successful use of Diehl's (1991) original categorical structure in other studies raises interest in this area and we suggest a refinement of these for future researchers. Particularly, self-assessment might be the way forward for this, in which ideators categorise their own ideas, or a situation in which coders work together to formulate a coding schema for each set of ideation data.

The novelty and value ratings are often-used measurements of idea quality (e.g. Boden, 2004; Ye and Robert Jr, 2017; De Dreu, Baas and Nijstad, 2008). The low agreement between coders using these ratings suggests that the ambiguity of the definitions quality metrics might need to be refined. Take for example "value: how useful and practical is this idea and does it make sense as a solution to the problem?" Useful and practical are two very different things. Might this complexity have had an impact on the quality ratings performed by our coders? We suggest here that future research involving ratings like these consider what is truly meant by value, and whether this metric should be split into components such as feasibility, cost, impact, generalisability and scalability; or whether it should be clearly define which of these components is represented by *value*.

We have previously discussed the reasoning for the choice of brainstorming as our ideation task. We have also discussed the concept of brainstorming tasks in themselves as being a very unconstrained, ill-defined tasks in which the goal and the steps necessary to produce viable solutions is ambiguous (Chrysikou, 2006). We chose to use open-ended questions in this thesis due to the inclusive nature of the types of question, requiring little to no subject knowledge. The specific questions chosen for our first study were a replication of questions used in other ideation studies by Nijstad and Stroebe (2006). The questions subsequently used in our second study were developed with consideration for the initial questions - all open-ended questions designed to be possible for a general audience to answer. It is worth noting that some of our questions were designed with the participant population available to us in mind: students and staff associated with the University of Bath. These questions may have limited usage in other situations (e.g. "How can we improve student housing?"). Altering the *type* of ideation task might be a useful way of exploring the research questions in this thesis with fewer inter-rater reliability issues. Using other methods of ideation, such as SCAMPER or analogy to existing systems, which specify a structure or use exemplars to guide how a solution should be formulated, we might find more specific or elaborated ideas. Indeed, this brings up the question of whether ideas generated through a brainstorming activity might not lend themselves to be objectively evaluated.

The task time-limits used in these studies varied quite a bit throughout the research. At the outset, we were looking to eventually run our studies as online crowd-sourced ideation tasks. Other studies (e.g. Siangliulue et al., 2015) that have been run on online platforms offer ideators 15 minutes to generate as many ideas as possible to an open-ended question, similar to those we have used in our research. We therefore set our first study time-limit to 15 minutes. As our first study was an unpaid replication study, we ran this study in a lab setting (not online). Unfortunately, the study we were replicating was a 20-minute ideation task; this is an unfortunate design flaw, but did highlight the fact that an eventual move to online studies could prove less productive than hoped. The 15 minute time-limit did not yield as many ideas as anticipated. For our second study, the 30 minute time-limit was a fairly arbitrary decision as there was no similar ideation study to compare to.

For other accomplishment-style tasks we find varying lengths, such as Payne, Duggan and Neth's (2007) 15-minute scrabble task, or Wilke et al.'s (2009) 60-minute anagram task. The 30 minutes worked well for us in this study as we were not interested in seeing the overall generation of ideas across this time, but the generation of ideas within and between each of the tasks. The time-limit for the third study was 25 minutes. Online ideation studies are often run in 15-minute sessions, however, in our first study, we found that a 15-minute timespan was not enough to exhaust ideators stock of ideas. We therefore decided to increase the ideation time substantially in order to see if we were able to observe a diminishing returns curve. Whilst this decision was grounded in our experience from the first study, the limitation of this decision has to be acknowledged - particularly in relation to the nature of online studies. Although the crowd-sourced ideators were informed of the length and nature of the study in advance, we might attribute the high drop-out rate to the length of the study. Indeed, some participants showed signs of not generating any ideas in the last few minutes, suggesting that a time-limit of less than 25 minutes might be preferable for online studies, and might explain the (often standard) choice of 15 minutes made in other studies of this type.

We chose to run our third study on an online platform in order to make use of crowd-sourcing data collection methods. This is an increasingly popular method in the literature, due to quick response times, time and cost efficiency and access to a larger work-force than might be possible in a lab-based setting (Eickhoff, 2018). Recent research in design studies and creativity shows a strong interest in the use of online platforms as a way of generating ideas for inspiration (Goucher-Lambert and Cagan, 2019), involving users in the generation of ideas for new products (Schemmann et al., 2016), or even using crowd-sourced work to asses the quality of ideas (Kudrowitz and Wallace, 2013). Among the aforementioned benefits of crowd-sourcing participants, platforms such as Prolific Academic[1], the one used in our research, allow experimenters to perform specific profiling on participants. In terms of time and cost efficiency, we were able to recruit 107 participants within two hours in our third study. This is compared to the hour spent with each participant in the lab-based first and second studies we performed. It is worth highlighting why the first studies were not crowd-sourced. A trade-off when running online studies is that these require monetary incentive. The initial study we ran, our replication study, was an unpaid 15-minute study and it was therefore not possible to run this online. The second study was rewarded based on performance (based on the number of distinct ideas generated), a mechanic not yet supported by online platforms such as Prolific Academic. A further disadvantage of crowd-sourcing is the weak supervision of these studies. This may increase the likelihood of

[1] https://www.prolific.ac/

dropouts in which participants start a study but for unknown reasons choose to abandon the study. Online studies might also be subject to low quality submissions from either malicious participants, simply engaging in the study for monetary gain, or from unqualified participants who have somehow slipped through the profiling constraints (Allahbakhsh et al., 2013; Eickhoff and de Vries, 2013; Buchanan and Scofield, 2018). Indeed, a few of our datasets from the online study did contain responses in a foreign language, suggesting that the 'native English speaker' constraint may not be 100% reliable. Despite the issues inherent in the weak supervision of crowd-sourced data collection, there is no doubt that the ability to access a large pool of participants, in a short amount of time, opens up larger possibilities for ideation research, e.g. in relation not only to generation of ideas, but in relation to possible crowd-sourced coding of ideas as well. Although counter-indicated by the strong implication in our results that such coders require careful training, it is worth exploring further how we can use crowd-sourcing platforms to recruit judges whilst mitigating the effects of lack of careful training and no supervision.

Although the experiments reported in this thesis are fairly small in terms of participant numbers, the number of ideas these participants generated is substantial. This is good news for the statistical power of some of the within-subjects tests of effects, but also a challenge in terms of the considerable workload required from volunteer and paid coders, who are necessary for the analysis of semantics and quality. With respect to our original goals of informing the design of ideation support systems, it must be admitted that the final study in the thesis is too small-scale and exploratory to offer a major contribution to theory-based design. However, it does hopefully show the potential of rather lightweight but theoretically derived interventions at the user interface to shift ideators' behaviour and therefore, in principle, to improve that behaviour for practical purposes.

7.2 Implications for the design of ideation tools

The initial intention of the work in this thesis was to build a stronger understanding of the time-course of ideation, in order to inform methods for supporting cognitive state transitions. Throughout our studies we encountered methodological issues that steered our work away from our original goals and towards the study of the methodology itself. Despite this outcome, we might still make inferences from our data, offering insights into possible design guidelines for ideation tools. In this section, we discuss the implications of our results from two different viewpoints: (1) extensions to methodology in ideation studies in support of experimenters, with focus on the coding difficulties we have encountered in our own work; (2) general support for ideators and how we might influence the design of ideation support tools.

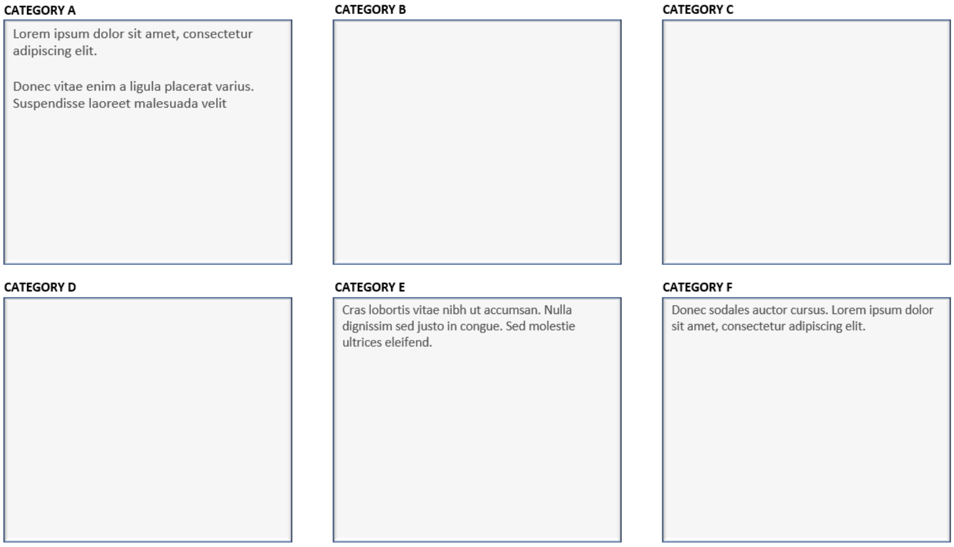
7.2.1 Design of ideation tools for the extension of methodology in support of experimenters

Our first research question "What influences the time-course pattern of ideas generated in ideation tasks?" prompted the replication of an existing single task study in our first study, and the development of our second (sequential task) study. Key learnings from these studies included a weak replication of the findings by Nijstad and Stroebe (2006): that ideas are generated in clusters of semantic similarity. As this finding provides a weak verification of a categorical structure to thinking during ideation, we might look to the methodological issues encountered for insights into how experimenters can better their understanding of categorical ideation. Findings from these studies highlighted another, fairly similar, methodological issue: difficulty in subjective novelty and value ratings of ideas. We propose three main methodological extensions that may support experimenters when designing these types of studies.

1. Self-categorisation of ideas generated by ideators themselves

Addressing the categorisation of ideas, we suggest the development of an interface that allows for self-categorisation of ideas. This might take the form of a free-form interface, rather than the often static list format used in simple brainstorming experiments. We suggest offering ideators a number of different named category boxes or spaces within which ideators might either enter ideas into a set of categories predetermined by the experimenter (see figure 7-1), or we might imagine them generating the categories, naming them according to their own categorical structure, then dragging and dropping ideas between these (see figure 7-2). The number of spaces representing categories and whether these are labelled might depend on the intended outcome of the ideation problem itself. If we are truly interested in observing ideators cognitive processes, we might leave the decision on number and name of categories up to the ideators themselves. However, it would likewise be interesting to see the effect pre-determined categories might have on the elaboration of ideas. Indeed, studies have been run on the concept of priming people to certain categories, showing that people are more likely to generate a depth of ideas within the few categories they are primed with, rather than a breadth of categories (Rietzschel, Nijstad and Stroebe, 2007). It would seem fair to say that varying constraints on number of categories and whether these can be self-selected is an interesting topic for future work in itself.

Not only may support for explicit self-categorisation further our understanding of the SIAM cognitive model, but also how ideators make semantic connections between the ideas they generate themselves. It may give a stronger insight into how ideas are categorised, without the need for post-experiment categorisation done by external coders, who may be unfamiliar with the thought-processes of the ideators themselves. The concept of self-categorisation might further be an interesting area of future study, to observe whether the explicit knowledge of the concept of categories helps ideators generate more ideas. Indeed, the concept of an explicit structure of thinking is an often-used method of prompting creative thinking. Consider for example mind-mapping, in which ideators are asked to structure ideas in a hierarchical manner (Buzan and Harrison, 2010). The mere existence of a set of physical categorical spaces to ideate within might support improved ideation flow. In addition to this, it would seem plausible that encouraging self-categorisation would prompt a strong awareness of strategies used in idea generation, which might be beneficial for ideators as well as experimenters.



CATEGORY A

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Donec vitae enim a ligula placerat varius. Suspendisse laoreet malesuada velit

CATEGORY B

CATEGORY C

CATEGORY D

CATEGORY E

Cras lobortis vitae nibh ut accumsan. Nulla dignissim sed justo in congue. Sed molestie ultrices eleifend.

CATEGORY F

Donec sodales auctor cursus. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Figure 7-1: Suggested design of a free-form interface that supports categorical ideation where categories are predetermined by the experimenter.

2. Self-evaluation of idea quality by ideators themselves

We can further extend the methodology by including an evaluation phase in electronic brainstorming research. This is, again, not a new concept and has been done in previous studies (e.g. Sowden, Pringle and Gabora, 2015). However, based on the ambiguity of the meanings of novelty (surprising and original) and value (useful and practical, not to mention feasibility, cost, impact, etc.) we propose an extension to the methodology in the form of ideators evaluating their own ideas, similar to a heuristic evaluation. A finding from our third study revealed a higher agreement on novelty ratings, possibly as a result of the visibility of the set of well-formulated ideas. Whilst this finding requires additional research, the use of exemplars might improve the ability of ideators themselves to perform ratings.

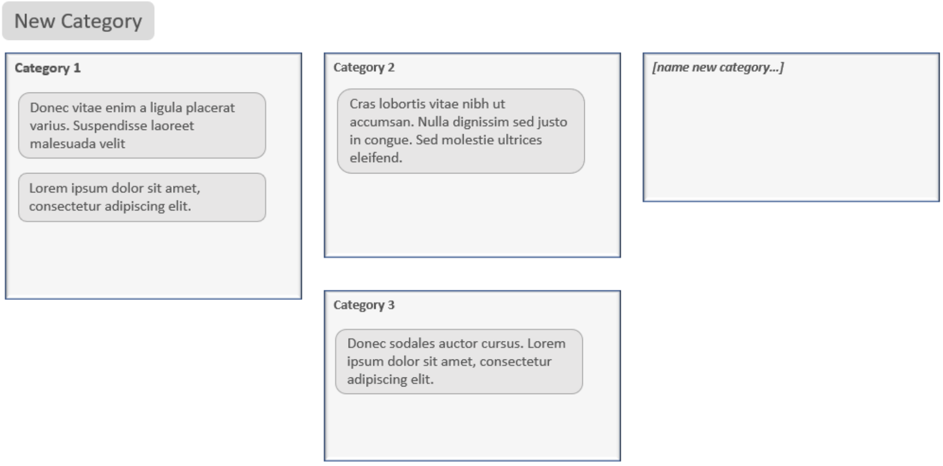

New Category
Category 1
Donec vitae enim a ligula placerat varius. Suspendisse laoreet malesuada velit
Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Category 2
Cras lobortis vitae nibh ut accumsan. Nulla dignissim sed justo in congue. Sed molestie ultrices eleifend.
[name new category...]
Category 3
Donec sodales auctor cursus. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Figure 7-2: Suggested design of free-form interface that supports categorical ideation. The number and name of categories is determined by ideators, by selecting 'new category'. Ideas can be dragged and dropped between categories. Category boxes should resize dependent on number of ideas entered.

3. Ideators reformulation of own ideas and revisiting earlier ideas

The reformulation of ideators own ideas and revisiting earlier ideas is not a new concept in ideation tools. The process of revisiting and reformulation is usually represented in conjunction with the evaluation stages of creative problem solving. In the IDEO Design Thinking[2] process for creative problem solving, ideators are encouraged to brainstorm potential solutions, then select and develop them to generate even better solutions to the problem, in an iterative manner. Again, we might imagine that observation and careful recording of this process can offer experimenters a stronger understanding of thought processes that take place during ideation.

It is worth mentioning here that these methodological extensions go somewhat against the original decision to treat ideation in isolation, separate to evaluation. Whilst studies of ideation in isolation provide insights into aspects of ideation such as time-course, the problems encountered in our studies suggest that it is likely that methodological extensions such as these may offer further insight into the cognitive processes that take place during ideation.

[2] https://www.ideou.com/pages/design-thinking

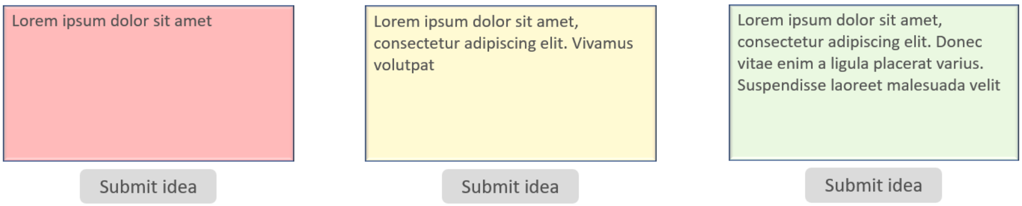
7.2.2 Design guidelines for the development of ideation support tools

Our second research question "Can we support ideation in such a way that judgement of ideas becomes more reliable?" was generated as a result of the methodological difficulties encountered in our first two studies, and prompted the development of a simple manipulation in our third and final single task study. Although a small manipulation, it allowed us to make some interesting observations and inferences about how ideation interface designs might impact different qualities of ideas generated. Overall, we found that an explicit instruction on what thematic roles to consider when generating solutions to a problem resulted in, on average, longer ideas. Whilst this also resulted in fewer ideas, these were interestingly judged as being of higher value, with a focus on usefulness and practicality as per our coding instructions. From these analyses we might offer two insights into the design of ideation support tools looking to support increased value of ideas.

Before discussing these, we want to highlight that the suggested extensions to the methodology of ideation studies, discussed in the previous section, might have further positive impacts on the quality of ideation itself. Perhaps not on ideation in isolation, but self-categorisation, self-evaluation and reformulation of ideas all seem to fit into the ideation-evaluation iterative model of creative problem solving. Indeed, a stronger self-awareness of ideation strategies combined with iterative refinement of ideas might in itself be considered a form of ideation support.

1. Constraints on the length of ideas

The result of our manipulations to instruction in the third study revealed significantly longer ideas generated in the explicit prompt condition, in which ideators were explicitly asked to include thematic roles in their solutions. Ideas generated in this condition were further judged as having significantly higher value. Although this was a small finding, we propose that a simple method supporting valuable ideation might involve setting constraints on the length of ideas. In the first instance, by simply making the input field larger than a single line input. Further designs might involve ways of guiding "the length of a good idea" by using simple visual cues such as a colour scales (see figure 7-3), or using smiley (frown/smile) feedback styles. In addition to support for more elaborated ideas, this offers the benefit of allowing ideators a visual way of tracking performance. Whilst the value of an individual idea might not always be correlated to its length, explicitly showing "expected" idea length might prompt ideators to consider how ideas might be elaborated.



Lorem ipsum dolor sit amet

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus volutpat

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec vitae enim a ligula placerat varius. Suspendisse laoreet malesuada velit

(a) Idea too short (b) Idea elaborated (c) Idea well-formulated

Figure 7-3: Suggested design alternative for visually cued length constraints on ideas, offering a multi-line box affording longer ideas than single line boxes. Colour-cued feedback on idea length, with red meaning too short, yellow meaning ok, green meaning well formulated. Alternatives to the standard traffic light colours advised in design for accessibility (e.g. colour blindedness).

2. Template input explicitly specifying thematic roles

The analysis of the data from our second study suggests that agreed good (high value and high novelty) ideas contained the following thematic roles:

- Agent: who is performing this action?
- Action: what are they doing?
- Instrument: what are they using (e.g. phone, a map)
- Recipient: who is benefiting from this action?
- Purpose: why is this being done, what is the goal of this?

Our simple manipulation further showed that prompting ideators to consider each of these thematic roles might increase the value of the ideas generated. We therefore propose a simplistic design guideline for ideation tools which offers a template input, offering several distinct boxes each labelled with the thematic roles that we discovered were included in the formulation of good ideas (see figure 7-4). Method cards[3] are an existing tool in the IDEO toolkit, which makes use of prompts such as learn, look, ask and try. A similar tool in the IDEO toolkit are nature cards[4], in which ideators are encouraged to take inspiration for their designs from nature; designing by analogy by looking at outside sources of inspiration. Our design suggestions contribute to existing tools like these by adding a structure to the way ideas are formulated.

[3] https://www.ideo.com/post/method-cards
[4] https://www.ideo.com/post/nature-cards

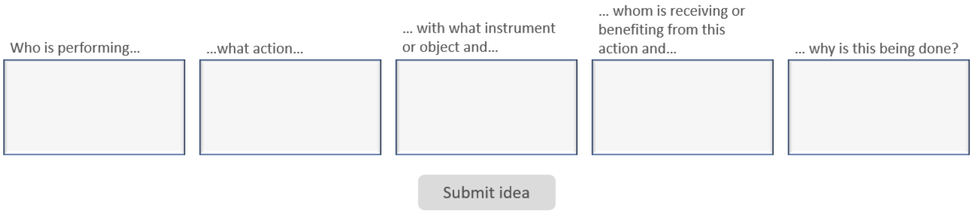

Who is performing...
...what action...
... with what instrument or object and...
... whom is receiving or benefiting from this action and...
... why is this being done?
Submit idea

Figure 7-4: Suggested design of template ideation tool that requires input into thematic role fields to inspire well-formulated ideas.

7.3 Concluding remarks

In this thesis, we set out to understand the cognitive processes that take place during ideation. Our original intent was to discover methods of supporting ideation through cognitive states such as impasses. The emergence of methodological difficulties encountered throughout the studies performed in this thesis steered the work towards a study of the methodology itself. We performed three studies, looking at both continuous ideation across a single ideation task, and continuous ideation across multiple sequential tasks. Whilst we found evidence of semantic clustering of ideas, we encountered low inter-rater agreements on category scores as well as novelty and value ratings of ideas. Our investigations of how to address these issues revealed that the use of thematic roles as prompts during ideation tasks have a positive effect on the length of ideas as well as on judged value of ideas.

As a result of the difficulties identified in our studies, we were able to identify insights into how methodologies in ideation studies can be extended to include self-categorisation, self-evaluation and reformulation of ideas. Although considered as the evaluative part of creative problem solving, inclusion of these elements might offer experimenters further insight into the cognitive processes that occur during ideation. We have additionally offered suggestions for simple design interventions that may affect the value and specificity of ideas generated: constraints on the length of ideas and explicit specification of thematic roles to consider when generating ideas. Although more research is required to fully understand the cognitive processes in ideation, we hope that this work shows that by detailed experimental analysis of ideation, we might learn some possible interface interventions through which ideators might be supported.

Bibliography

Agogué, M., Kazakçi, A., Hatchuel, A., Le Masson, P., Weil, B., Poirel, N. and Cassotti, M., 2014. The impact of type of examples on originality: Explaining fixation and stimulation effects. *The Journal of Creative Behavior*, 48(1), pp.1–12.

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H.R., Bertino, E. and Dustdar, S., 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2), pp.76–81.

Allen, A.P. and Thomas, K.E., 2011. A dual process account of creative thinking. *Creativity Research Journal*, 23(2), pp.109–118.

Amabile, T.M., 1996. *Creativity in context: Update to the social psychology of creativity*. Hachette UK.

Anderson, J.R. and Milson, R., 1989. Human memory: An adaptive perspective. *Psychological Review*, 96(4), p.703.

Andolina, S., Klouche, K., Cabral, D., Ruotsalo, T. and Jacucci, G., 2015. Inspirationwall: supporting idea generation through automatic information exploration. *Proceedings of the 2015 acm sigchi conference on creativity and cognition*. ACM, pp.103–106.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N. and Evershed, J., 2018. Gorilla in our midst: An online behavioral experiment builder. *bioRxiv*, p.438242.

Bao, P., Gerber, E., Gergle, D. and Hoffman, D., 2010. Momentum: getting and staying on topic during a brainstorm. *Proceedings of the sigchi conference on human factors in computing systems*. ACM, pp.1233–1236.

Baruah, J. and Paulus, P.B., 2011. Category assignment and relatedness in the group ideation process. *Journal of Experimental Social Psychology*, 47(6), pp.1070–1077.

Basadur, M., 1995. *The power of innovation: How to make innovation a way of life and put creative solutions to work*. Financial Times Management.

Basadur, M., Graen, G.B. and Green, S.G., 1982. Training in creative problem solving: Effects on ideation and problem finding and solving in an industrial research organization. *Organizational Behavior and Human Performance*, 30(1), pp.41–70.

Basadur, M. and Hausdorf, P.A., 1996. Measuring divergent thinking attitudes related to creative problem solving and innovation management. *Creativity Research Journal*, 9(1), pp.21–32.

Basden, B.H., Basden, D.R., Bryner, S. and Thomas III, R.L., 1997. A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(5), p.1176.

Berk, L.M., 1999. *English syntax: from word to discourse*. Oxford University Press, USA.

Boden, M.A., 2004. *The creative mind: Myths and mechanisms*. Routledge.

Briggs, R.O. and Reinig, B.A., 2010. Bounded ideation theory. *Journal of Management Information Systems*, 27(1), pp.123–144.

Brophy, D.R., 2001. Comparing the attributes, activities, and performance of divergent, convergent, and combination thinkers. *Creativity Research Journal*, 13(3-4), pp.439–455.

Buchanan, E.M. and Scofield, J.E., 2018. Methods to detect low quality data and its implication for psychological research. *Behavior research methods*, 50(6), pp.2586–2596.

Buzan, T. and Harrison, J., 2010. *Use your head: How to unleash the power of your mind*. Pearson.

Cady, S.H. and Valentine, J., 1999. Team innovation and perceptions of consideration: What difference does diversity make? *Small group research*, 30(6), pp.730–750.

Chan, J., Dang, S. and Dow, S.P., 2016. Improving crowd innovation with expert facilitation. *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*. ACM, pp.1223–1235.

Chan, J., Fu, K., Schunn, C., Cagan, J., Wood, K. and Kotovsky, K., 2011. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of mechanical design*, 133(8), p.081004.

Chan, J., Siangliulue, P., Qori McDonald, D., Liu, R., Moradinezhad, R., Aman, S., Solovey, E.T., Gajos, K.Z. and Dow, S.P., 2017. Semantically far inspirations considered harmful?: Accounting for cognitive states in collaborative ideation. *Proceedings of the 2017 acm sigchi conference on creativity and cognition*. ACM, pp.93–105.

Charnov, E.L. et al., 1976. Optimal foraging, the marginal value theorem.

Chrysikou, E.G., 2006. When shoes become hammers: Goal-derived categorization training enhances problem-solving performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(4), p.935.

Cropley, A., 2006. In praise of convergent thinking. *Creativity research journal*, 18(3), pp.391–404.

Csikszentmihalyi, M., 1997. Flow and the psychology of discovery and invention. *Harper-Perennial, New York*, 39.

De Bono, E., 2017. *Six thinking hats*. Penguin UK.

De Dreu, C.K., Baas, M. and Nijstad, B.A., 2008. Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model. *Journal of personality and social psychology*, 94(5), p.739.

Dean, D.L., Hender, J., Rodgers, T. and Santanen, E., 2006. Identifying good ideas: constructs and scales for idea evaluation.

Dennett, D.C., 2017. *Brainstorms: Philosophical essays on mind and psychology*. MIT press.

Dennis, A.R., Aronson, J.E., Heninger, W.G., Walker, I. and Edward, D., 1999. Structuring time and task in electronic brainstorming. *MIS Quarterly*, 23(1).

Dennis, A.R., Minas, R.K. and Bhagwatwar, A.P., 2013. Sparking creativity: Improving electronic brainstorming with individual cognitive priming. *Journal of Management Information Systems*, 29(4), pp.195–216.

Dennis, A.R. and Valacich, J.S., 1993. Computer brainstorms: More heads are better than one. *Journal of applied psychology*, 78(4), p.531.

Diedrich, J., Benedek, M., Jauk, E. and Neubauer, A.C., 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1), p.35.

Diehl, M., 1991. Kollektive kreativität: Zur quantität und qualität der ideenproduktion in kleingruppen (collective creativity: On quantity and quality of idea production in small groups). *Unpublished postdoctoral thesis, University of Tübingen*.

Diehl, M., Munkes, J. and Ziegler, R., 2002. Brainstorming and cognitive stimulation: When does being exposed to the ideas of others facilitate or inhibit one's own idea generation. *conference of the european association of experimental social psychology, san sebastian, spain*. vol. 200.

Diehl, M. and Stroebe, W., 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of personality and social psychology*, 53(3), p.497.

Diehl, M. and Stroebe, W., 1991. Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of personality and social psychology*, 61(3), p.392.

Dougherty, M.R. and Harbison, J., 2007. Motivated to retrieve: How often are you willing to go back to the well when the well is dry? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), p.1108.

Dugosh, K.L., Paulus, P.B., Roland, E.J. and Yang, H.C., 2000. Cognitive stimulation in brainstorming. *Journal of personality and social psychology*, 79(5), p.722.

Durand, D.E. and VanHuss, S.H., 1992. Creativity software and dss: Cautionary findings. *Information & management*, 23(1), pp.1–6.

Eberle, B., 2008. *Scamper: Creative games and activities for imagination development*. Prufrock Press.

Eickhoff, C., 2018. Cognitive biases in crowdsourcing. *Proceedings of the eleventh acm international conference on web search and data mining*. ACM, pp.162–170.

Eickhoff, C. and Vries, A.P. de, 2013. Increasing cheat robustness of crowdsourcing tasks. *Information retrieval*, 16(2), pp.121–137.

Faste, H., Rachmel, N., Essary, R. and Sheehan, E., 2013. Brainstorm, chainstorm, cheatstorm, tweetstorm: new ideation strategies for distributed hci design. *Proceedings of the sigchi conference on human factors in computing systems*. ACM, pp.1343–1352.

Frankel, F. and Cole, M., 1971. Measures of category clustering in free recall. *Psychological Bulletin*, 76(1), p.39.

Frich, J., Mose Biskjaer, M. and Dalsgaard, P., 2018. Twenty years of creativity research in human-computer interaction: Current state and future directions. *Proceedings of the 2018 designing interactive systems conference*. ACM, pp.1235–1257.

Gabora, L., 2002. Cognitive mechanisms underlying the creative process. *Proceedings of the 4th conference on creativity & cognition*. ACM, pp.126–133.

Geschka, H., Schaude, G.R. and Schlicksupp, H., 1976. Modern techniques for solving problems. *International Studies of Management & Organization*, 6(4), pp.45–63.

Girotra, K., Terwiesch, C. and Ulrich, K.T., 2010. Idea generation and the quality of the best idea. *Management science*, 56(4), pp.591–605.

Goldenberg, O., Larson Jr, J.R. and Wiley, J., 2013. Goal instructions, response format, and idea generation in groups. *Small Group Research*, 44(3), pp.227–256.

Goldschmidt, G. and Sever, A.L., 2011. Inspiring design ideas with texts. *Design Studies*, 32(2), pp.139–155.

Golembewski, M. and Selby, M., 2010. Ideation decks: a card-based design ideation tool. *Proceedings of the 8th acm conference on designing interactive systems*. ACM, pp.89–92.

González, V.M. and Mark, G., 2004. Constant, constant, multi-tasking craziness: managing multiple working spheres. *Proceedings of the sigchi conference on human factors in computing systems*. ACM, pp.113–120.

Goucher-Lambert, K. and Cagan, J., 2019. Crowdsourcing inspiration: Using crowd generated inspirational stimuli to support designer ideation. *Design Studies*, 61, pp.1–29.

Green, R.F., 1984. Stopping rules for optimal foragers. *The American Naturalist*, 123(1), pp.30–43.

Guilford, J., 1967. *The nature of human intelligence*, McGraw-Hill series in psychology. McGraw-Hill.

Harbison, J.I., Dougherty, M.R., Davelaar, E.J. and Fayyad, B., 2009. On the lawfulness of the decision to terminate memory search. *Cognition*, 111(3), pp.397–402.

Hayes, A.F. and Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1), pp.77–89.

Hills, T.T., Jones, M.N. and Todd, P.M., 2012. Optimal foraging in semantic memory. *Psychological review*, 119(2), p.431.

Höök, K. and Löwgren, J., 2012. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3), p.23.

Howard-Jones, P. and Murray, S., 2003. Ideational productivity, focus of attention, and context. *Creativity research journal*, 15(2-3), pp.153–166.

Iwasa, Y., Higashi, M. and Yamamura, N., 1981. Prey distribution as a factor determining the choice of optimal foraging strategy. *The American Naturalist*, 117(5), pp.710–723.

Jansson, D.G. and Smith, S.M., 1991. Design fixation. *Design studies*, 12(1), pp.3–11.

Kandogan, E., Kim, J., Moran, T.P. and Pedemonte, P., 2011. How a freeform spatial interface supports simple problem solving tasks. *Proceedings of the sigchi conference on human factors in computing systems*. ACM, pp.925–934.

Kanekar, S. and Rosenbaum, M.E., 1972. Group performance on a multiple-solution task as a function of available time. *Psychonomic Science*, 27(6), pp.331–332.

Kaufman, J.C. and Sternberg, R.J., 2010. *The cambridge handbook of creativity*. Cambridge University Press.

Knoblich, G., Ohlsson, S. and Raney, G.E., 2001. An eye movement study of insight problem solving. *Memory & cognition*, 29(7), pp.1000–1009.

Koestler, A., 1964. *The act of creation*. London Hutchinson.

Krippendorff, K., 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Kudrowitz, B.M. and Wallace, D., 2013. Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design*, 24(2), pp.120–139.

Landis, J.R. and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *biometrics*, pp.159–174.

MacCrimmon, K.R. and Wagner, C., 1994. Stimulating ideas through creative software. *Management science*, 40(11), pp.1514–1532.

MacGregor, J.N., Ormerod, T.C. and Chronicle, E.P., 2001. Information processing and insight: a process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(1), p.176.

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

Moss, J., Kotovsky, K. and Cagan, J., 2007. The influence of open goals on the acquisition of problem-relevant information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(5), p.876.

Nijstad, B.A. and Stroebe, W., 2006. How the group affects the mind: A cognitive model of idea generation in groups. *Personality and social psychology review*, 10(3), pp.186–213.

Nijstad, B.A., Stroebe, W. and Lodewijkx, H.F., 2002. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of experimental social psychology*, 38(6), pp.535–544.

Nijstad, B.A., Stroebe, W. and Lodewijkx, H.F., 2003. Production blocking and idea generation: Does blocking interfere with cognitive processes? *Journal of experimental social psychology*, 39(6), pp.531–548.

Ochse, R., 1990. *Before the gates of excellence: The determinants of creative genius*. CUP Archive.

Osborn, A.F.A.F., 1957. *Applied imagination : principles and procedures of creative thinking*. Rev. ed ed. New York : Charles Scribner's Sons.

Oviatt, S., Cohen, A., Miller, A., Hodge, K. and Mann, A., 2012. The impact of interface affordances on human ideation, problem solving, and inferential reasoning. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3), p.22.

Oviatt, S.L. and Cohen, A.O., 2010. Toward high-performance communications interfaces for science problem solving. *Journal of science education and technology*, 19(6), pp.515–531.

Payne, S.J., Duggan, G.B. and Neth, H., 2007. Discretionary task interleaving: heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General*, 136(3), p.370.

Pennington, J., Socher, R. and Manning, C., 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*. pp.1532–1543.

Perry-Smith, J.E., 2006. Social yet creative: The role of social relationships in facilitating individual creativity. *Academy of Management journal*, 49(1), pp.85–101.

Peterson, R.S. and Nemeth, C.J., 1996. Focus versus flexibility majority and minority influence can both improve performance. *Personality and Social Psychology Bulletin*, 22(1), pp.14–23.

Pirolli, P. and Card, S., 1999. Information foraging. *Psychological review*, 106(4), p.643.

Raaijmakers, J.G. and Shiffrin, R.M., 1981. Search of associative memory. *Psychological review*, 88(2), p.93.

Reinig, B.A., Briggs, R.O. and Nunamaker, J.F., 2007. On the measurement of ideation quality. *Journal of Management Information Systems*, 23(4), pp.143–161.

Rietzschel, E.F., Nijstad, B.A. and Stroebe, W., 2007. Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of experimental social psychology*, 43(6), pp.933–946.

Roenker, D.L., Thompson, C.P. and Brown, S.C., 1971. Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, 76(1), p.45.

Runco, M.A. and Acar, S., 2012. Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, 24(1), pp.66–75.

Runco, M.A. and Sakamoto, S.O., 1999. Experimental studies of creativity. *Handbook of creativity*, pp.62–92.

Sandstrom, P.E., 1994. An optimal foraging approach to information seeking and use. *The library quarterly*, 64(4), pp.414–449.

Schemmann, B., Herrmann, A.M., Chappin, M.M. and Heimeriks, G.J., 2016. Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development. *Research Policy*, 45(6), pp.1145–1154.

Shah, J.J., Smith, S.M. and Vargas-Hernandez, N., 2003. Metrics for measuring ideation effectiveness. *Design studies*, 24(2), pp.111–134.

Shneiderman, B., 2009. Creativity support tools: A grand challenge for hci researchers. *Engineering the user interface*, Springer, pp.1–9.

Siangliulue, P., Chan, J., Gajos, K.Z. and Dow, S.P., 2015. Providing timely examples improves the quantity and quality of generated ideas. *Proceedings of the 2015 acm sigchi conference on creativity and cognition*. ACM, pp.83–92.

Silverstein, D., Samuel, P. and DeCarlo, N., 2013. *The innovator's toolkit: 50+ techniques for predictable and sustainable organic growth*. John Wiley & Sons.

Sosa, R. and Dong, A., 2013. The creative assessment of rich ideas. *Proceedings of the 9th acm conference on creativity & cognition*. ACM, pp.328–331.

Sowden, P.T. and Dawson, L., 2011. Creative feelings: the effect of mood on creative ideation and evaluation. *Proceedings of the 8th acm conference on creativity and cognition*. ACM, pp.393–394.

Sowden, P.T., Pringle, A. and Gabora, L., 2015. The shifting sands of creative thinking: Connections to dual-process theory. *Thinking & Reasoning*, 21(1), pp.40–60.

Stephens, D.W. and Krebs, J.R., 1986. *Foraging theory.* Princeton University Press.

Sternberg, R.J., 1998. Cognitive mechanisms in human creativity: Is variation blind or sighted? *The Journal of Creative Behavior*, 32(3), pp.159–176.

Taylor, D.W., Berry, P.C. and Block, C.H., 1958. Does group participation when using brainstorming facilitate or inhibit creative thinking? *Administrative Science Quarterly*, pp.23–47.

Torrance, E.P., 1962. *Guiding creative talent.* Prentice-Hall, INC.

Waage, J.K., 1979. Foraging for patchily-distributed hosts by the parasitoid, nemeritis canescens. *The Journal of Animal Ecology*, pp.353–371.

Wang, H.C., Cosley, D. and Fussell, S.R., 2010. Idea expander: supporting group brainstorming with conversationally triggered visual thinking stimuli. *Proceedings of the 2010 acm conference on computer supported cooperative work*. ACM, pp.103–106.

Ware, C., Rogers, D., Petersen, M., Ahrens, J. and Aygar, E., 2016. Optimizing for visual cognition in high performance scientific computing. *Electronic Imaging*, 2016(16), pp.1–9.

Wilke, A., Hutchinson, J.M., Todd, P.M. and Czienskowski, U., 2009. Fishing for the right words: Decision rules for human foraging behavior in internal search tasks. *Cognitive Science*, 33(3), pp.497–529.

Wölfel, C. and Merritt, T., 2013. Method card design dimensions: a survey of card-based design tools. *Ifip conference on human-computer interaction*. Springer, pp.479–486.

Woodman, R.W., Sawyer, J.E. and Griffin, R.W., 1993. Toward a theory of organizational creativity. *Academy of management review*, 18(2), pp.293–321.

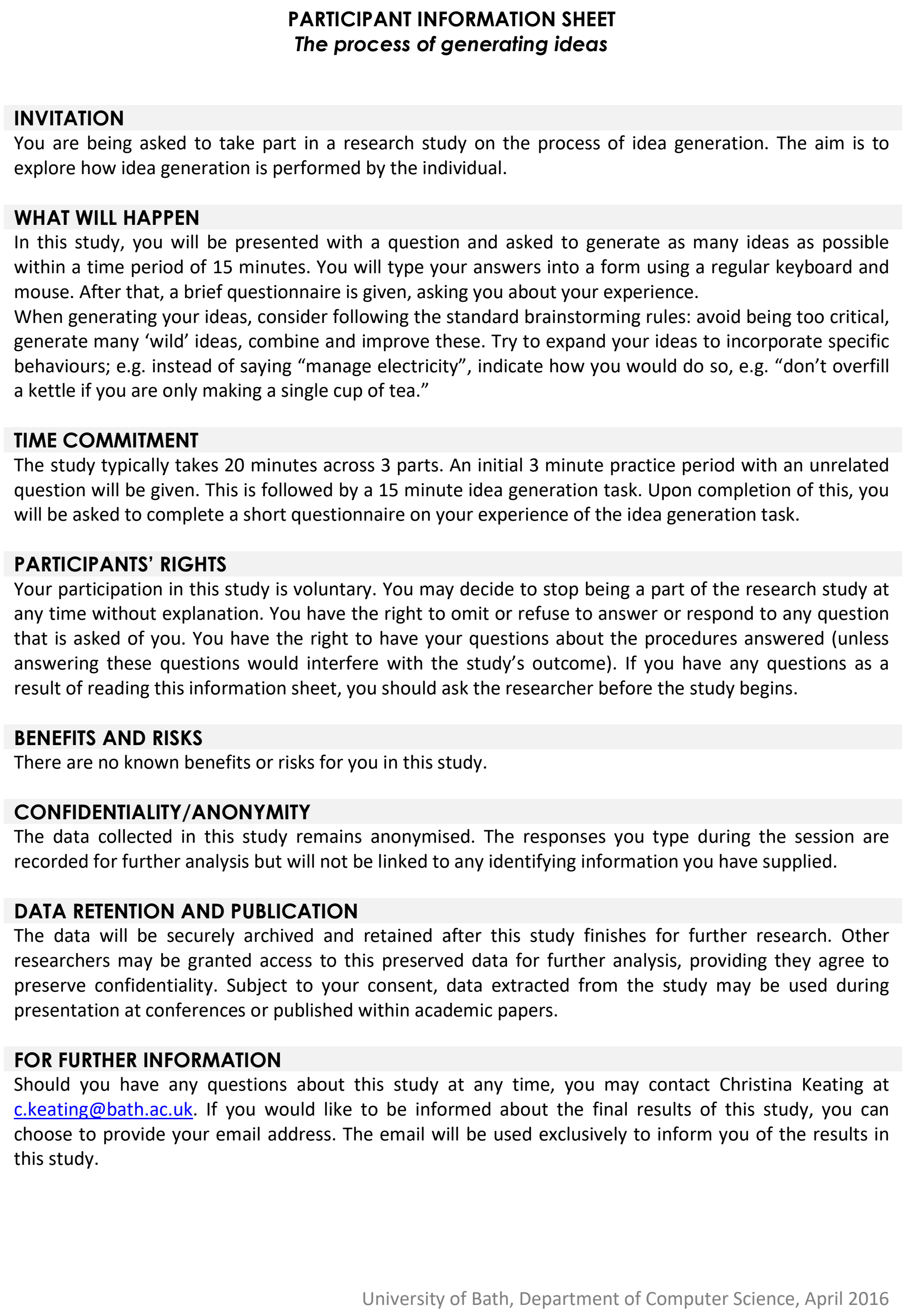
Ye, T. and Robert Jr, L.P., 2017. Does collectivism inhibit individual creativity?: The effects of collectivism and perceived diversity on individual creativity and satisfaction in virtual ideation teams. *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. ACM, pp.2344–2358.

Appendices

Appendix A

Appendices for studies 1a and 1b

A.1 Participant information sheet S1



PARTICIPANT INFORMATION SHEET
The process of generating ideas

INVITATION

You are being asked to take part in a research study on the process of idea generation. The aim is to explore how idea generation is performed by the individual.

WHAT WILL HAPPEN

In this study, you will be presented with a question and asked to generate as many ideas as possible within a time period of 15 minutes. You will type your answers into a form using a regular keyboard and mouse. After that, a brief questionnaire is given, asking you about your experience.

When generating your ideas, consider following the standard brainstorming rules: avoid being too critical, generate many 'wild' ideas, combine and improve these. Try to expand your ideas to incorporate specific behaviours; e.g. instead of saying "manage electricity", indicate how you would do so, e.g. "don't overfill a kettle if you are only making a single cup of tea."

TIME COMMITMENT

The study typically takes 20 minutes across 3 parts. An initial 3 minute practice period with an unrelated question will be given. This is followed by a 15 minute idea generation task. Upon completion of this, you will be asked to complete a short questionnaire on your experience of the idea generation task.

PARTICIPANTS' RIGHTS

Your participation in this study is voluntary. You may decide to stop being a part of the research study at any time without explanation. You have the right to omit or refuse to answer or respond to any question that is asked of you. You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study's outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

BENEFITS AND RISKS

There are no known benefits or risks for you in this study.

CONFIDENTIALITY/ANONYMITY

The data collected in this study remains anonymised. The responses you type during the session are recorded for further analysis but will not be linked to any identifying information you have supplied.

DATA RETENTION AND PUBLICATION

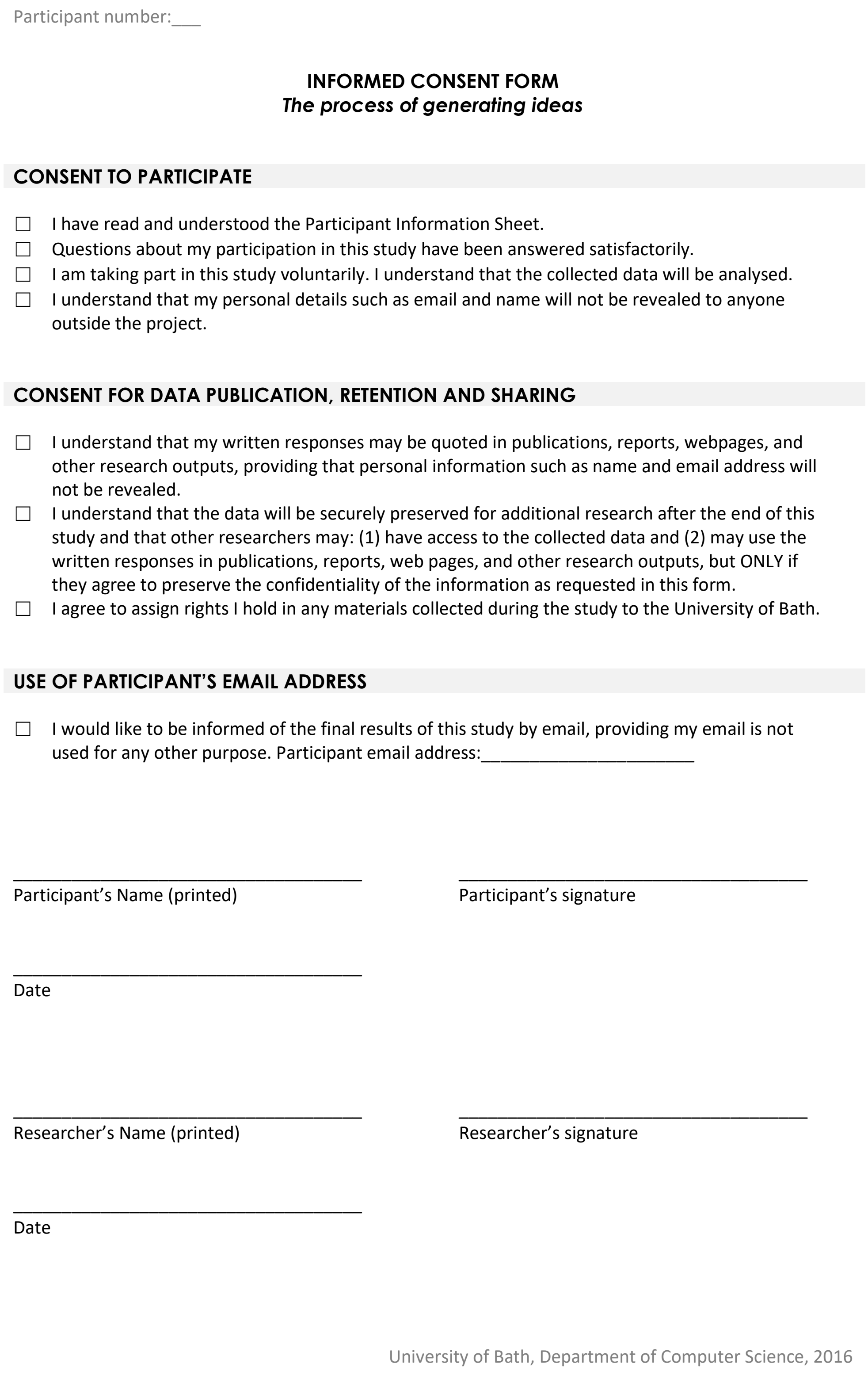
The data will be securely archived and retained after this study finishes for further research. Other researchers may be granted access to this preserved data for further analysis, providing they agree to preserve confidentiality. Subject to your consent, data extracted from the study may be used during presentation at conferences or published within academic papers.

FOR FURTHER INFORMATION

Should you have any questions about this study at any time, you may contact Christina Keating at c.keating@bath.ac.uk. If you would like to be informed about the final results of this study, you can choose to provide your email address. The email will be used exclusively to inform you of the results in this study.

University of Bath, Department of Computer Science, April 2016

A.2 Participant consent form S1



Participant number:___

INFORMED CONSENT FORM
The process of generating ideas

CONSENT TO PARTICIPATE

☐ I have read and understood the Participant Information Sheet.
☐ Questions about my participation in this study have been answered satisfactorily.
☐ I am taking part in this study voluntarily. I understand that the collected data will be analysed.
☐ I understand that my personal details such as email and name will not be revealed to anyone outside the project.

CONSENT FOR DATA PUBLICATION, RETENTION AND SHARING

☐ I understand that my written responses may be quoted in publications, reports, webpages, and other research outputs, providing that personal information such as name and email address will not be revealed.
☐ I understand that the data will be securely preserved for additional research after the end of this study and that other researchers may: (1) have access to the collected data and (2) may use the written responses in publications, reports, web pages, and other research outputs, but ONLY if they agree to preserve the confidentiality of the information as requested in this form.
☐ I agree to assign rights I hold in any materials collected during the study to the University of Bath.

USE OF PARTICIPANT'S EMAIL ADDRESS

☐ I would like to be informed of the final results of this study by email, providing my email is not used for any other purpose. Participant email address:______________________

____________________________________ ____________________________________
Participant's Name (printed) Participant's signature

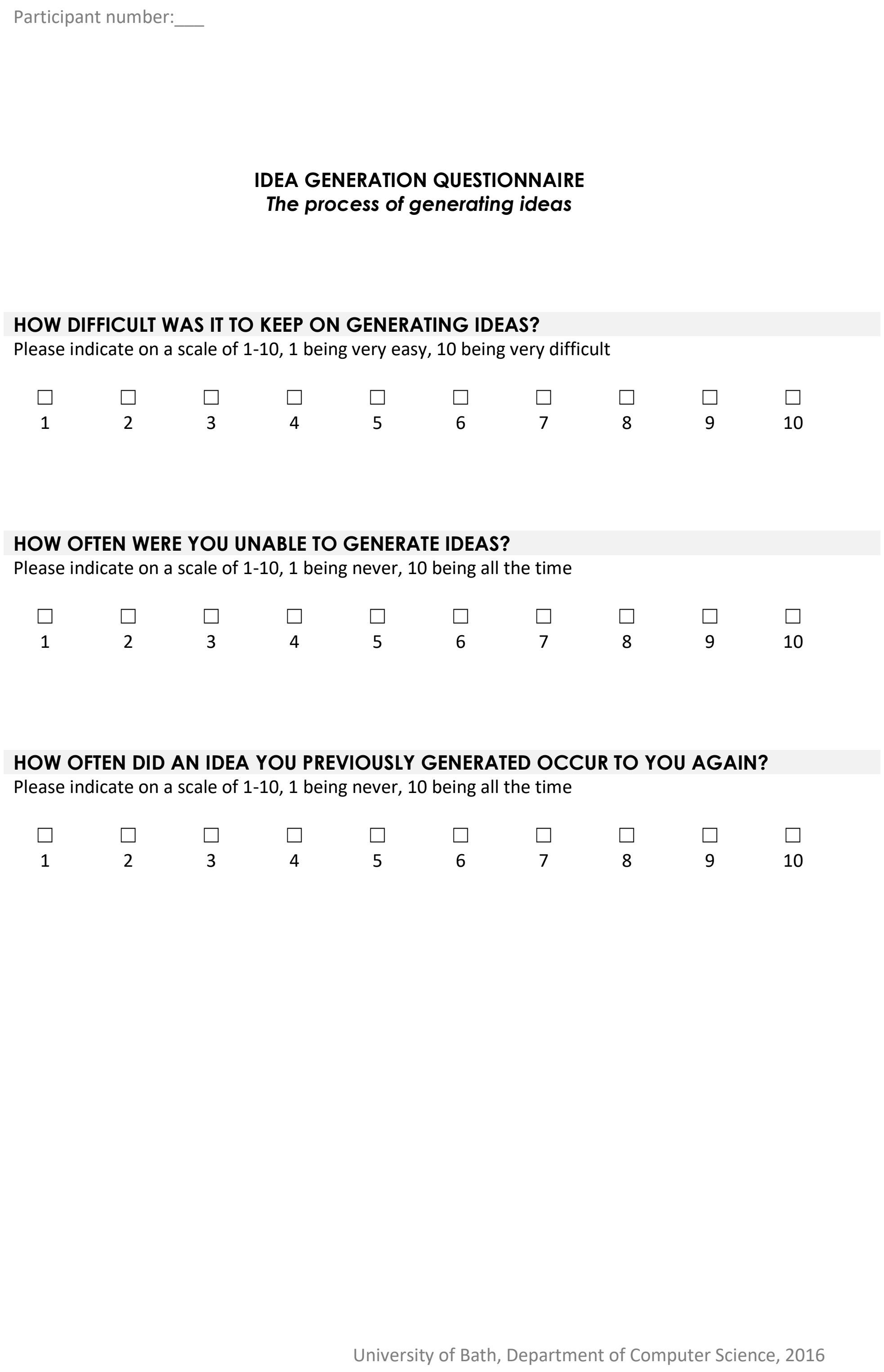
____
Date

____________________________________ ____________________________________
Researcher's Name (printed) Researcher's signature

____
Date

University of Bath, Department of Computer Science, 2016

A.3 Post-experiment questionnaire S1



Participant number:___

IDEA GENERATION QUESTIONNAIRE
The process of generating ideas

HOW DIFFICULT WAS IT TO KEEP ON GENERATING IDEAS?
Please indicate on a scale of 1-10, 1 being very easy, 10 being very difficult

□	□	□	□	□	□	□	□	□	□
1	2	3	4	5	6	7	8	9	10

HOW OFTEN WERE YOU UNABLE TO GENERATE IDEAS?
Please indicate on a scale of 1-10, 1 being never, 10 being all the time

□	□	□	□	□	□	□	□	□	□
1	2	3	4	5	6	7	8	9	10

HOW OFTEN DID AN IDEA YOU PREVIOUSLY GENERATED OCCUR TO YOU AGAIN?
Please indicate on a scale of 1-10, 1 being never, 10 being all the time

□	□	□	□	□	□	□	□	□	□
1	2	3	4	5	6	7	8	9	10

University of Bath, Department of Computer Science, 2016

A.4 Semantic difference coding instructions

Background:

My first exploratory study asked 10 participants the following question: "What can the individual do to preserve the environment?" and 10 participants "What can a person do to maintain or improve their health?". They were asked to come up with as many ideas as possible in the space of 15 minutes. The attached spreadsheet contains 20 tabs of raw data from this study.

As part of my research, I am looking at the semantic difference between subsequent ideas and whether people are grouping their ideas into categories or if each idea wildly varies from the previous one. I need independent raters for this (you)!

My task for you:

The attached spreadsheet contains 20 tabs of ideas. Could I ask you to score, on a scale from 1-10, how semantically different two adjacent ideas are?

Column B contains the ideas. Column A contains the area where you can input your score.

Just to give you an idea of how you can think about this:

I would score the following a **1 or 2**, as they are essentially touching on the same concepts: recycling and plastic materials.

Idea 1: "recycle plastic bottles"

Idea 2: "recycle plastic bags"

I would score this a **5-6** as they have something in common but aren't quite the same.

Idea 1: "recycle plastic bottles"

Idea 2: "teach people why recycling plastic bottles saves the planet"

I would score the following a **10** as they have nothing in common (in my personal opinion!).

Idea 1: "recycle plastic bottles"

Idea 2: "support an organisation that protects pandas"

As this is not a study I do not have the usual consent form, however, just to let you know: there is NO wrong way of doing this. This is purely based on your own best judgement and personal opinion.

Appendix B

Appendices for study 2

B.1 Questions for study 2

The following questions were developed for study 2 in order to avoid any one participant exhausting the number of questions available.

1. Your phone can tell you exactly where you are, come up with functions/apps that can use this.
2. What can one do the preserve the environment?
3. What can the individual do to increase their general level of health?
4. How could we attract more women to study technology?
5. What can the individual do to be more productive?
6. How can we improve student housing?
7. What can society do to make the world safer for wild animals?
8. How can we make transportation/travel/driving safer?
9. What can the individual do for the community in their spare time?
10. How can we get children interested in science?
11. How can we be safer online? (Mentally, privacy, financial safety, security, cyber bullying)
12. How can we facilitate world peace?

13. What methods can we use to make more friends in a new city?
14. How can we address the problems of an ageing population?
15. How can the individual break an addiction?
16. How can the individual improve their ability to remember things?
17. How can you save space if you live in a small apartment?

B.2 Participant information sheet S2

PARTICIPANT INFORMATION SHEET
The process of generating ideas

INVITATION

You are being asked to take part in a research study on the process of idea generation. The aim is to explore how idea generation is performed by the individual.

WHAT WILL HAPPEN

In this study, you will be presented with a question and asked to generate as many ideas as possible in 30 minutes. You will type your answers into a form using a regular keyboard and mouse. If you feel you have run out of ideas, click the “new question” button to receive a new question on another topic. Your goal is to produce as many good and well formulated ideas as possible in total, i.e. across all the questions you attempt. It doesn’t matter how many questions you attempt or how many ideas you generate for any particular question, all that matters is the total. After the 30 minutes are up, a brief questionnaire is given, asking you about your experience.

When generating your ideas, consider following the standard brainstorming rules: avoid being too critical, generate many ‘wild’ ideas, combine and improve these. Try to expand your ideas to incorporate specific behaviours; e.g. instead of saying “manage electricity”, indicate how you would do so, e.g. “don’t overfill a kettle if you are only making a single cup of tea.”

TIME COMMITMENT

The study typically takes 40 minutes across 3 parts. An initial 2-minute introduction to the form will be given. This is followed by a 30-minute idea generation task. Upon completion of this, you will be asked to complete a short questionnaire on your experience of the idea generation task.

PARTICIPANTS’ RIGHTS

Your participation in this study is voluntary. You may decide to stop being a part of the research study at any time without explanation. You have the right to omit or refuse to answer or respond to any question that is asked of you. You have the right to have your questions about the procedures answered (unless answering these questions would interfere with the study’s outcome). If you have any questions as a result of reading this information sheet, you should ask the researcher before the study begins.

COMPENSATION

As a thank you for your participation in this study, you will automatically be entered in a prize draw for two Amazon.co.uk voucher of £25 each. Additionally, you will receive 20p per coherent idea formed during the experiment (maximum £10). This will be paid within 1 week of completion of the study.

BENEFITS AND RISKS

There are no known benefits or risks for you in this study.

CONFIDENTIALITY/ANONYMITY

The data collected in this study remains anonymised. The responses you type during the session are recorded for further analysis but will not be linked to any identifying information you have supplied.

DATA RETENTION AND PUBLICATION

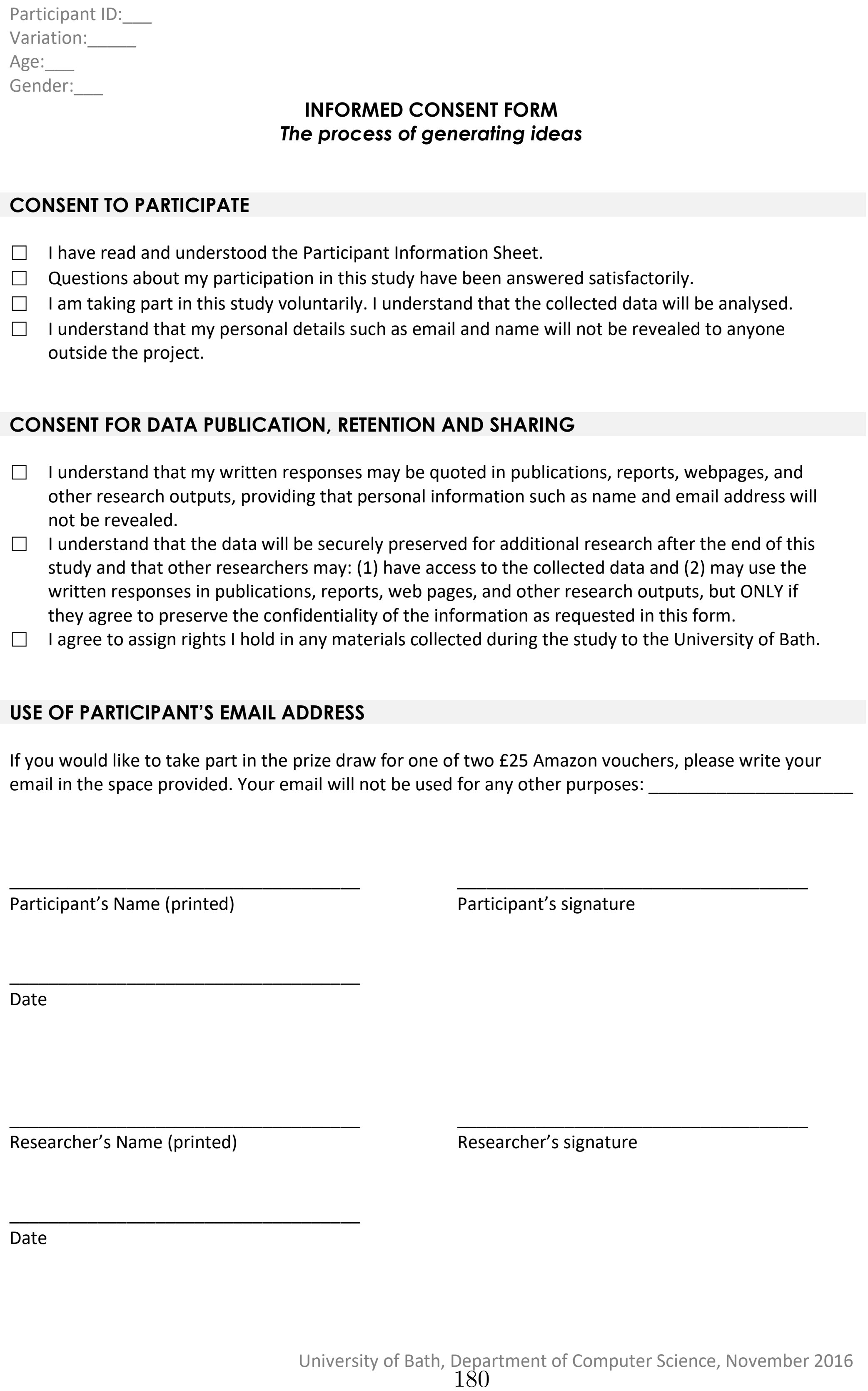
The data will be securely archived and retained after this study finishes for further research. Other researchers may be granted access to this preserved data for further analysis, providing they agree to preserve confidentiality. Subject to your consent, data extracted from the study may be used during presentation at conferences or published within academic papers.

FOR FURTHER INFORMATION

Should you have any questions about this study at any time, you may contact Christina Keating at c.keating@bath.ac.uk.

University of Bath, Department of Computer Science, November 2016

B.3 Participant consent form S2



Participant ID:___
Variation:_____
Age:___
Gender:___

INFORMED CONSENT FORM
The process of generating ideas

CONSENT TO PARTICIPATE

☐ I have read and understood the Participant Information Sheet.
☐ Questions about my participation in this study have been answered satisfactorily.
☐ I am taking part in this study voluntarily. I understand that the collected data will be analysed.
☐ I understand that my personal details such as email and name will not be revealed to anyone outside the project.

CONSENT FOR DATA PUBLICATION, RETENTION AND SHARING

☐ I understand that my written responses may be quoted in publications, reports, webpages, and other research outputs, providing that personal information such as name and email address will not be revealed.
☐ I understand that the data will be securely preserved for additional research after the end of this study and that other researchers may: (1) have access to the collected data and (2) may use the written responses in publications, reports, web pages, and other research outputs, but ONLY if they agree to preserve the confidentiality of the information as requested in this form.
☐ I agree to assign rights I hold in any materials collected during the study to the University of Bath.

USE OF PARTICIPANT'S EMAIL ADDRESS

If you would like to take part in the prize draw for one of two £25 Amazon vouchers, please write your email in the space provided. Your email will not be used for any other purposes: ______________________

___________________________________ Participant's Name (printed)

___________________________________ Participant's signature

___________________________________ Date

___________________________________ Researcher's Name (printed)

___________________________________ Researcher's signature

___________________________________ Date

University of Bath, Department of Computer Science, November 2016

B.4 Post-experiment questionnaire S2

Participant number:___

IDEA GENERATION QUESTIONNAIRE
The process of generating ideas

DID YOU COME UP WITH A PARTICULAR STRATEGY TO DETERMINE WHEN TO SWITCH QUESTIONS?

If YES, how would you describe this strategy, what cues did you use to decide when to give up on a question and move to the next?

____

____

____

____

____

HOW OFTEN DID YOU USE THE FOLLOWING CUES TO SWITCH TO A NEW QUESTION?

	Never									All the Time
Number of ideas generated on current question	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
Time spent on current question	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
Time interval since previous idea on current question	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
Feeling stuck on the current question	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10

Participant number:___

PLEASE INDICATE YOUR LEVEL OF AGREEMENT TO THE FOLLOWING STATEMENTS:

	Fully Disagree									Fully Agree
I only pressed “next question” when I felt stuck.	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
During a question, I continuously switched between the feelings of being stuck and un-stuck	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
It was very difficult to think of new ideas	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10
The same idea occurred to me several times	□ 1	□ 2	□ 3	□ 4	□ 5	□ 6	□ 7	□ 8	□ 9	□ 10

B.5 Quality Coding Instructions

Judgement of Ideas

Background

We have recently run a study looking at how people generate ideas. This was designed as an exploratory study – we are not just looking for pre-determined factors but also hoping that the data will tell us something we had not expected or anticipated.

The study involved 29 participants, each was given 30 minutes to generate ideas. They were given a list of 17 questions. Participants were free to spend the 30 minutes as they liked, e.g they could spend all 30 minutes on one question, or try to answer all 17. As part of the analysis, we are interested in the quality of ideas. Commonly used metrics for measuring this are:

- *Novelty: How original and surprising is the idea?*
- *Value: How useful and practical is this idea and does it make sense as a solution to the problem?*

The Work

You have been provided with a spreadsheet of 1264 ideas, given in response to 17 questions. As participants were given the choice of what questions to answer, some questions will have a longer list of ideas than others.

Each question is represented on its own tab. Please do one tab at a time, and try to take a break between each if you can in order to clear your mind. The overall payment for the work is £60 and will be paid upon completion of the debrief. The work should take approximately 3 hours.

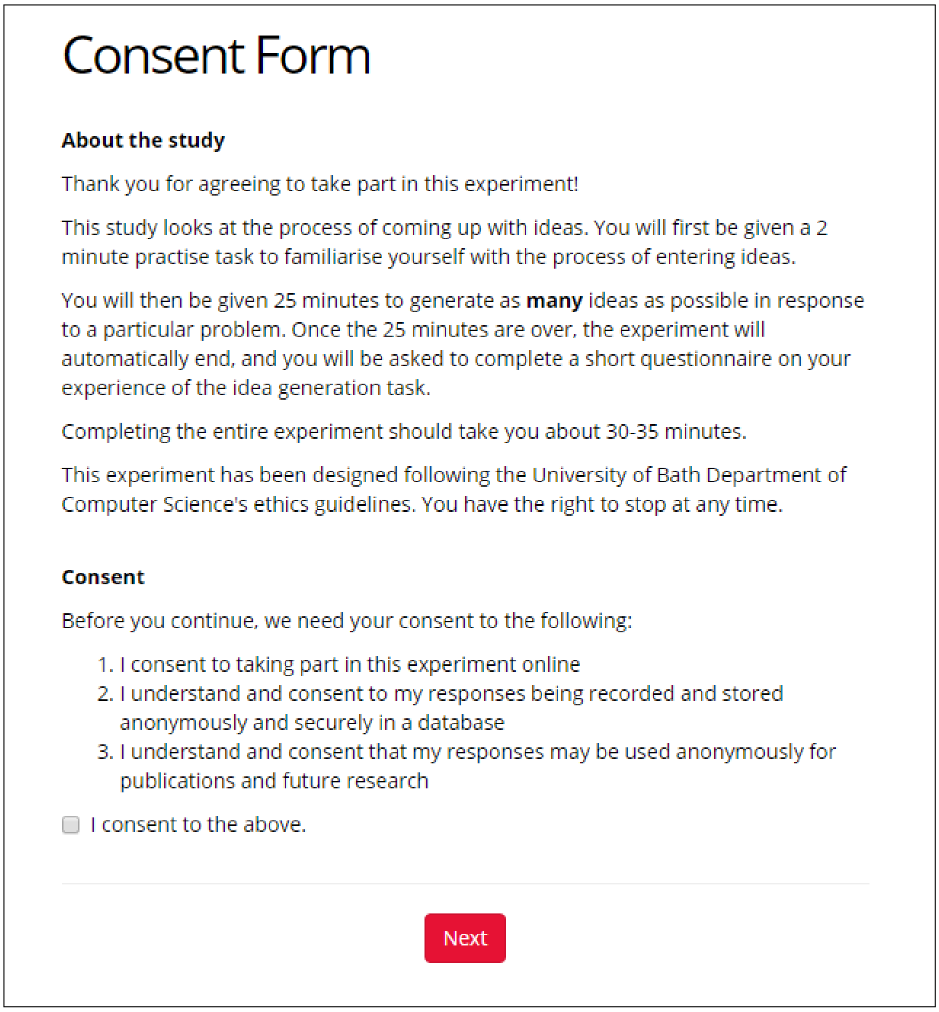
Instructions

1) Select a single question (tab).
2) Initially read through all the ideas for the question you have selected without making any judgements.
3) Go back to the top, and then give a "novelty" and "value" score to each individual idea. Novelty and Value do not necessarily go hand in hand so treat them as two separate scores. Please score these on the below scale.
4) You should not have the need to use the notes field, however if you do find some difficult, you are welcome to use this to enter your thoughts and we can discuss these in the debrief.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
No Value / Not Novel									High Value / Very Novel

Appendix C

Appendices for study 3: screenshots of application built in Gorilla



Consent Form

About the study

Thank you for agreeing to take part in this experiment!

This study looks at the process of coming up with ideas. You will first be given a 2 minute practise task to familiarise yourself with the process of entering ideas.

You will then be given 25 minutes to generate as **many** ideas as possible in response to a particular problem. Once the 25 minutes are over, the experiment will automatically end, and you will be asked to complete a short questionnaire on your experience of the idea generation task.

Completing the entire experiment should take you about 30-35 minutes.

This experiment has been designed following the University of Bath Department of Computer Science's ethics guidelines. You have the right to stop at any time.

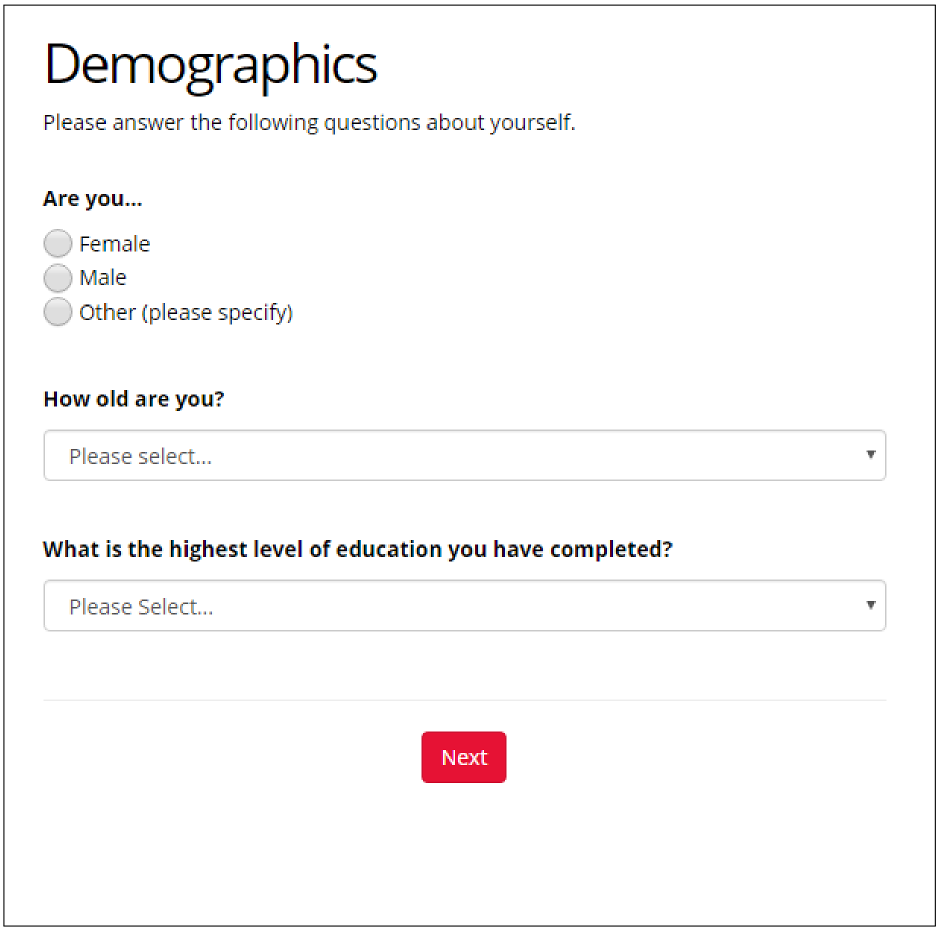
Consent

Before you continue, we need your consent to the following:

1. I consent to taking part in this experiment online
2. I understand and consent to my responses being recorded and stored anonymously and securely in a database
3. I understand and consent that my responses may be used anonymously for publications and future research

□ I consent to the above.

Figure C-1: Study 3 consent form



Demographics

Please answer the following questions about yourself.

Are you...

○ Female
○ Male
○ Other (please specify)

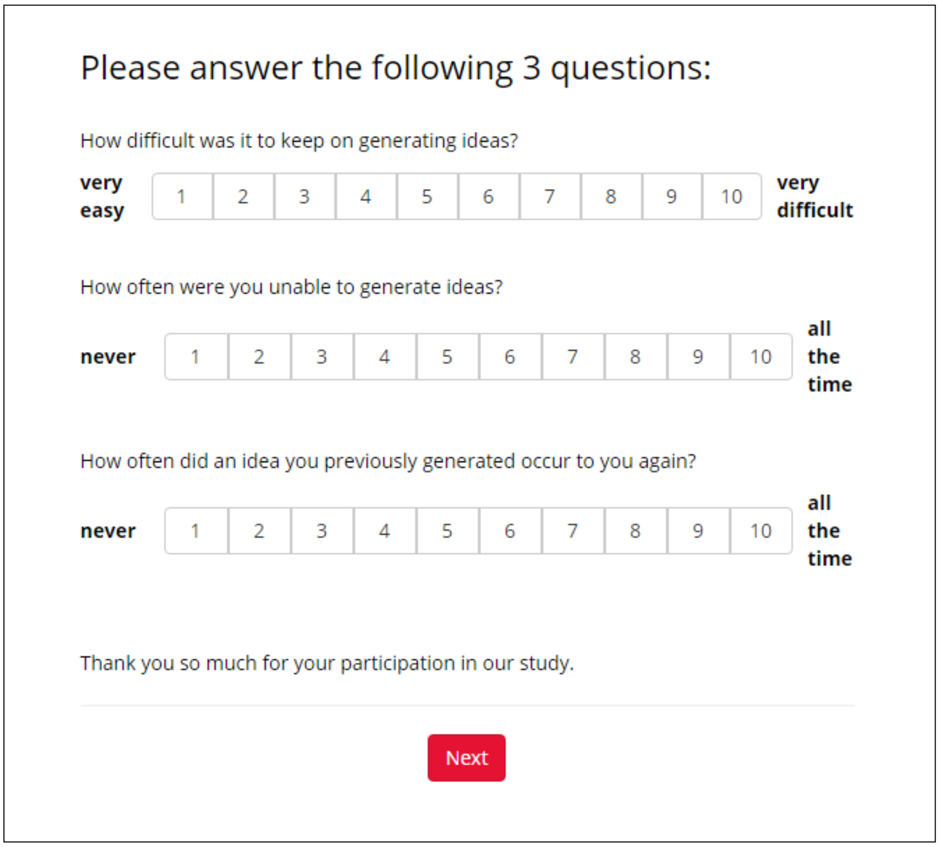
How old are you?

Please select... ▾

What is the highest level of education you have completed?

Please Select... ▾

Figure C-2: Study 3 demographics form



Please answer the following 3 questions:

How difficult was it to keep on generating ideas?

very easy	1	2	3	4	5	6	7	8	9	10	very difficult

How often were you unable to generate ideas?

never	1	2	3	4	5	6	7	8	9	10	all the time

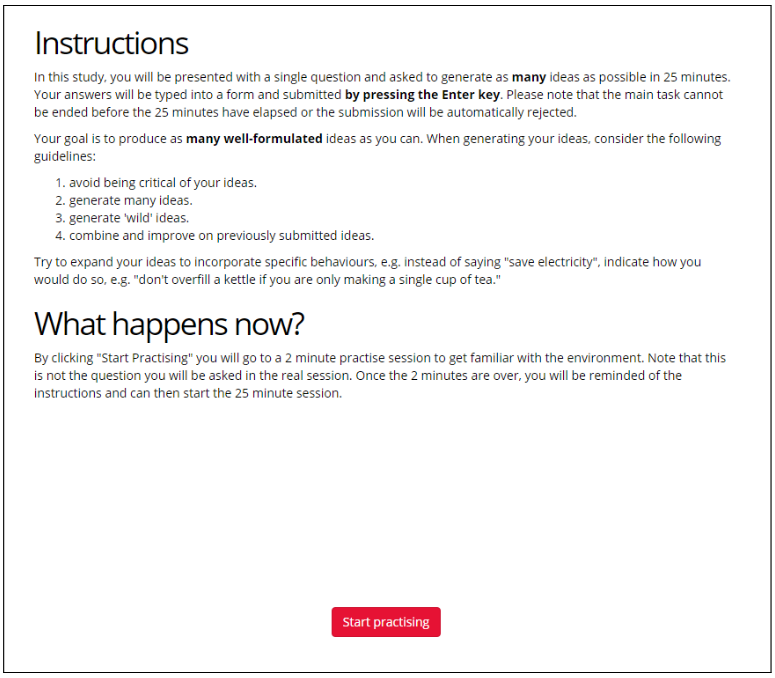
How often did an idea you previously generated occur to you again?

never	1	2	3	4	5	6	7	8	9	10	all the time

Thank you so much for your participation in our study.

Next

Figure C-3: Study 3 post-experiment questionnaire form



Instructions

In this study, you will be presented with a single question and asked to generate as **many** ideas as possible in 25 minutes. Your answers will be typed into a form and submitted **by pressing the Enter key**. Please note that the main task cannot be ended before the 25 minutes have elapsed or the submission will be automatically rejected.

Your goal is to produce as **many well-formulated** ideas as you can. When generating your ideas, consider the following guidelines:

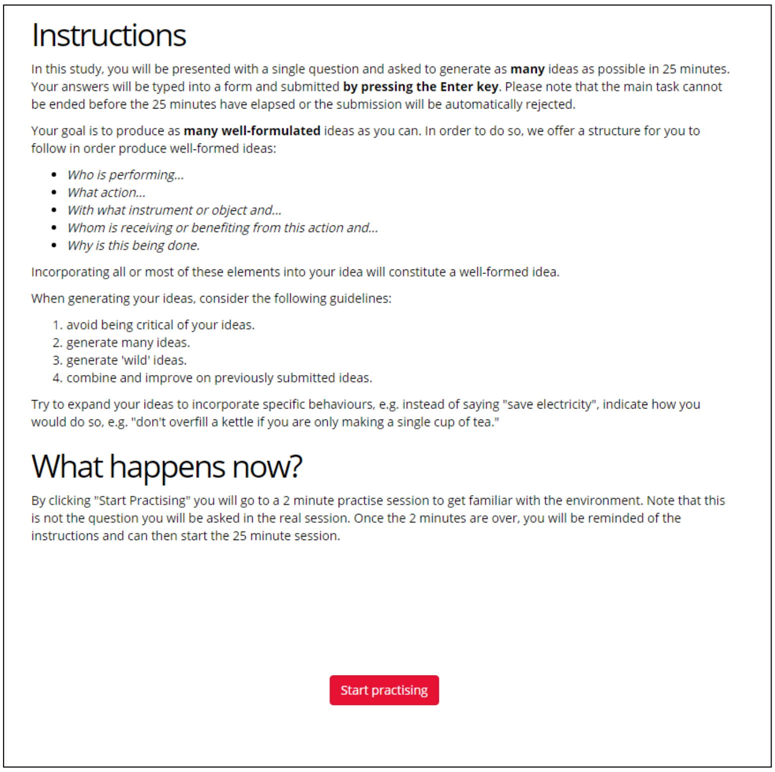
1. avoid being critical of your ideas.
2. generate many ideas.
3. generate 'wild' ideas.
4. combine and improve on previously submitted ideas.

Try to expand your ideas to incorporate specific behaviours, e.g. instead of saying "save electricity", indicate how you would do so, e.g. "don't overfill a kettle if you are only making a single cup of tea."

What happens now?

By clicking "Start Practising" you will go to a 2 minute practise session to get familiar with the environment. Note that this is not the question you will be asked in the real session. Once the 2 minutes are over, you will be reminded of the instructions and can then start the 25 minute session.

(a) Instructions to no prompt and implicit example prompt conditions



Instructions

In this study, you will be presented with a single question and asked to generate as **many** ideas as possible in 25 minutes. Your answers will be typed into a form and submitted **by pressing the Enter key**. Please note that the main task cannot be ended before the 25 minutes have elapsed or the submission will be automatically rejected.

Your goal is to produce as **many well-formulated** ideas as you can. In order to do so, we offer a structure for you to follow in order produce well-formed ideas:

- *Who is performing...*
- *What action...*
- *With what instrument or object and...*
- *Whom is receiving or benefiting from this action and...*
- *Why is this being done.*

Incorporating all or most of these elements into your idea will constitute a well-formed idea.

When generating your ideas, consider the following guidelines:

1. avoid being critical of your ideas.
2. generate many ideas.
3. generate 'wild' ideas.
4. combine and improve on previously submitted ideas.

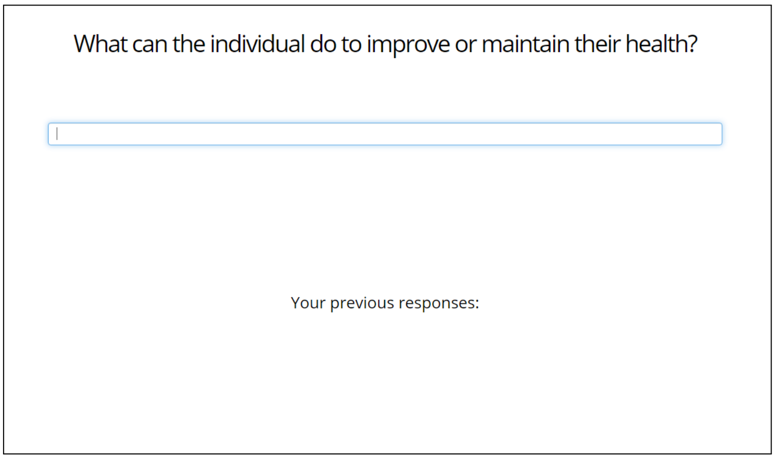
Try to expand your ideas to incorporate specific behaviours, e.g. instead of saying "save electricity", indicate how you would do so, e.g. "don't overfill a kettle if you are only making a single cup of tea."

What happens now?

By clicking "Start Practising" you will go to a 2 minute practise session to get familiar with the environment. Note that this is not the question you will be asked in the real session. Once the 2 minutes are over, you will be reminded of the instructions and can then start the 25 minute session.

(b) Instructions to explicit thematic role prompt condition

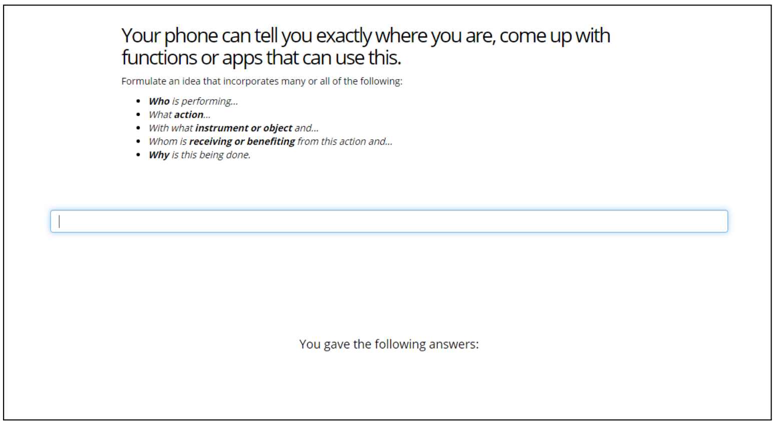
Figure C-4: Instructions for ideators in the three different conditions in Gorilla



What can the individual do to improve or maintain their health?

Your previous responses:

(a) Ideation screen for no prompt condition



Your phone can tell you exactly where you are, come up with functions or apps that can use this.

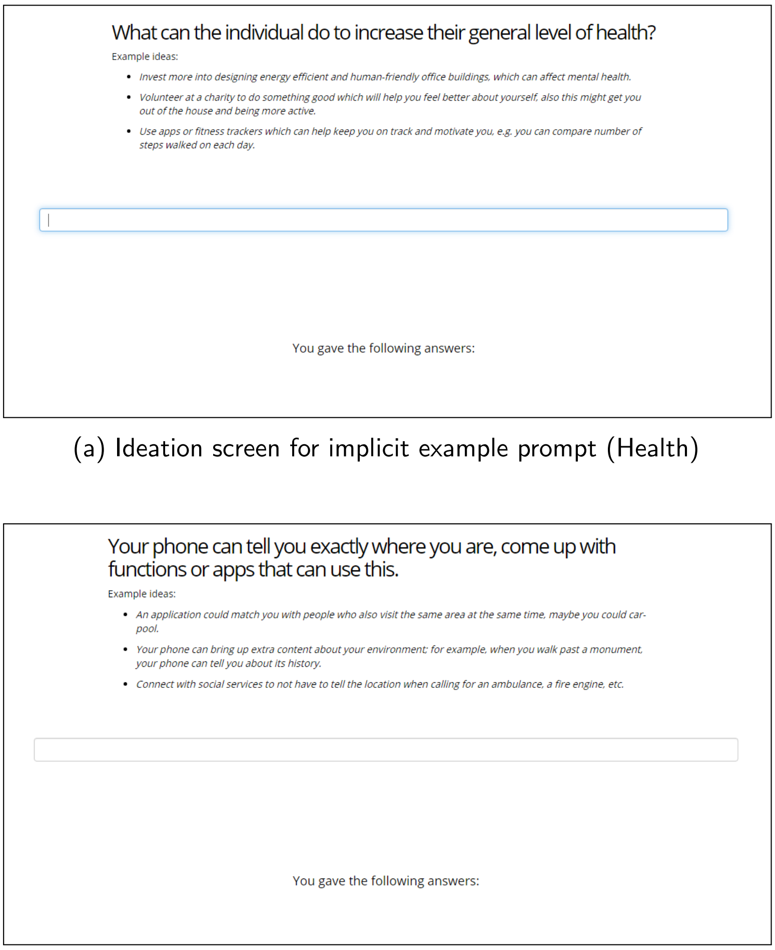
Formulate an idea that incorporates many or all of the following:

- ***Who** is performing...*
- *What **action**...*
- *With what **instrument or object** and...*
- *Whom is **receiving or benefiting** from this action and...*
- ***Why** is this being done.*

You gave the following answers:

(b) Ideation screen for thematic role explicit prompt condition

Figure C-5: Ideation screens for no prompt and explicit conditions in Gorilla



What can the individual do to increase their general level of health?

Example ideas:

- *Invest more into designing energy efficient and human-friendly office buildings, which can affect mental health.*
- *Volunteer at a charity to do something good which will help you feel better about yourself, also this might get you out of the house and being more active.*
- *Use apps or fitness trackers which can help keep you on track and motivate you, e.g. you can compare number of steps walked on each day.*

You gave the following answers:

(a) Ideation screen for implicit example prompt (Health)

Your phone can tell you exactly where you are, come up with functions or apps that can use this.

Example ideas:

- *An application could match you with people who also visit the same area at the same time, maybe you could car-pool.*
- *Your phone can bring up extra content about your environment; for example, when you walk past a monument, your phone can tell you about its history.*
- *Connect with social services to not have to tell the location when calling for an ambulance, a fire engine, etc.*

You gave the following answers:

(b) Ideation screen for implicit example prompt (Phone)

Figure C-6: Ideation screens for implicit example prompt (health and phone questions) in Gorilla

Appendix D

General appendices

D.1 Ethics

UNIVERSITY OF BATH

Department of Computer Science

13-POINT ETHICS CHECK LIST

This document describes the 13 issues that need to be considered carefully before students or staff involve other people ("participants") for the collection of information as part of their project or research.

1. *Have you prepared a briefing script for volunteers?* You must explain to people what they will be required to do, the kind of data you will be collecting from them and how it will be used.

2. *Will the participants be using any non-standard hardware?* Participants should not be exposed to any risks associated with the use of non- standard equipment: anything other than pen and paper or typical interaction with PCs on desks is considered non-standard.

3. *Is there any intentional deception of the participants?* Withholding information or misleading participants is unacceptable if participants are likely to object or show unease when debriefed.

4. *How will participants voluntarily give consent?* If the results of the evaluation are likely to be used beyond the term of the project (for example, the software is to be

deployed, or the data is to be published), then signed consent is necessary. A separate consent form should be signed by each participant.

5. *Will the participants be exposed to any risks greater than those encountered in their normal work life?* Investigators have a responsibility to protect participants from physical and mental harm during the investigation. The risk of harm must be no greater than in ordinary life.
6. *Are you offering any incentive to the participants?* The payment of participants must not be used to induce them to risk harm beyond that which they risk without payment in their normal lifestyle.
7. *Are any of your participants under the age of 16?* Parental consent is required for participants under the age of 16.
8. *Do any of your participants have an impairment that will limit their understanding or communication?* Additional consent is required for participants with impairments.
9. *Are you in a position of authority or influence over any of your participants?* A position of authority or influence over any participant must not be allowed to pressurise participants to take part in, or remain in, any experiment.
10. *Will the participants be informed that they could withdraw at any time?* All participants have the right to withdraw at any time during the investigation. They should be told this in the introductory script.
11. *Will the participants be informed of your contact details?* All participants must be able to contact the investigator after the investigation. They should be given the details of the Unit Lecturer or Supervisor as part of the debriefing.
12. *Will participants be de-briefed?* The student must provide the participants with sufficient information in the debriefing to enable them to understand the nature of the investigation.
13. *Will the data collected from the participants be stored in an anonymous form?* All participant data (hard copy and soft copy) should be stored securely, and in anonymous form.