DESIGN OUTCOMES: HOW DESIGNERS AND TOOLS INFLUENCE DESIGN QUALITY AND CREATIVITY

A STUDY OF INDIVIDUAL DESIGNERS

by
ANDERS HÄGGMAN

Funding for doctoral work has been provided in part by:
Tekniikan edistämissäätiö – Finnish Foundation for Technology Promotion
IK Investment Fund – The League of Finnish American Societies
Fulbright – Technology Industries of Finland Grant
Suomen kulttuurirahasto – Finnish Cultural Foundation
Emil Aaltosen säätiö

This thesis is typeset in:
Porter Bold, Porter Medium, Avenir Medium, Avenir Light, Baskerville, and Bookman Old Style

Product sketches by Nathan Cooke.
Illustrations by Anders Häggman, with modified base material from freepik.com and flaticon.com

DESIGN OUTCOMES: HOW DESIGNERS AND TOOLS INFLUENCE DESIGN QUALITY AND CREATIVITY – A STUDY OF INDIVIDUAL DESIGNERS

Anders Häggman

S.M. Mechanical Engineering, Teknillinen korkeakoulu, Finland, 2009

Submitted to the Department of Mechanical Engineering
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Mechanical Engineering
at the
Massachusetts Institute of Technology
February 2017

Signature redacted

Author
Anders Häggman
Department of Mechanical Engineering
4 October 2016

Signature redacted

Certified by
Maria C. Yang
Associate Professor of Mechanical Engineering
Thesis Supervisor

Signature redacted

Accepted by
Rohan Abeyaratne
Quentin Berg Professor of Mechanics
Chairman, Department Committee on Graduate Studies

Design outcomes:
how designers and tools influence design quality and creativity
– a study of individual designers

Anders Häggman

Submitted to the Department of Mechanical Engineering
on 4 October, 2016 in partial fulfilment of the requirements for
the degree of Doctor of Philosophy in Mechanical Engineering

ABSTRACT

The design process can be seen as a complex, ambiguous, ill-defined problem with no clearly correct answer. At the same time, the early stages of the design process carry importance with regard to design outcomes, sometimes with far reaching consequences. With the proliferation of computer modelling tools, designers are moving away from traditional design tools such as sketching, and begin designing in CAD earlier than before. This thesis focuses on the early stages of the design process, and on how selected design tools — sketching, foam prototyping, and computer modelling — influence the design outcomes of an individual designer in the early conceptual phases of the process.

Through the use of controlled design experiments with experienced design practitioners, this thesis seeks to examine how different design tools impact the design outcomes. Analysis of video and audio recordings, interviews, and talk-aloud protocols are used to gain insights, and investigate how different tools impact the design outcomes and decision making of individual designers in the early stages of the design process. As an example, does a designer who creates foam models — thereby receiving tactile feedback as they are creating the model — consider ergonomics more than a designer working in CAD?

Results suggest clear differences in quantity and quality of concepts depending on which design tool was used, as well as between designers themselves, highlighting the importance of using an appropriate design process and set of tools in the early conceptual stages of a design task.

Thesis supervisor: Maria C. Yang
Title: Associate Professor of Mechanical Engineering

DOCTORAL COMMITTEE

Thesis Supervisor
Professor **Maria C. Yang**
Massachusetts Institute of Technology

Professor **Warren Seering**
Massachusetts Institute of Technology

Professor **David Wallace**
Massachusetts Institute of Technology

Professor **Larry Leifer**
Stanford University

ACKNOWLEDGEMENTS

There are too many people to possibly thank them all here, but I will do my best to mention those who have been most influential.

I moved to Cambridge several years ago with my young daughter Emilia and wife Anna to pursue a PhD at MIT. I have learned a lot and experienced much during that time – none of which would have been possible without the support of my professors, colleagues, dear friends, and family. The friendships, laughs and support have kept me going.

My time here has been shaped by my PhD committee, and I have been incredibly fortunate to have four professors who are passionate about design and mentoring serve on it: professors Maria Yang, Warren Seering, and David Wallace from MIT, and Larry Leifer from Stanford University.

My thesis advisor and mentor, Maria, is the reason I came to MIT – and more importantly, endured. It is hard to imagine what my time here would have been like without Maria; the freedom to shape my own research has been rewarding, and her flexibility has made it possible to combine (by now) three small children with doctoral studies. Maria has been incredibly inspiring, understanding, and supportive in so many ways. She has been in one word – *super!*

I would like to thank Warren for his questions and comments, which were very insightful (and kept me honest) while at the same time supportive and kind. I enjoyed our conversations about design research and will try to remember his advice going forward.

David is an inspirational teacher, enthusiastic about design, with a strong creative vision and attention to detail. His teaching style is something I aspire towards and try to emulate, and it is always fun to talk with David about design or our shared links to The Soo.

If Maria is the reason I chose to come to MIT, then Larry is the reason I was able to come to MIT. Larry gave me an opportunity almost ten years ago, which has made all the difference in the world. The simple description would be to say that he changed my life. Larry is one of the warmest human beings I have ever met, with a sense of humour and wild creativity.

I owe a great deal to my friend Geoff Tsai. We have completed our doctoral journeys together – from qualifying exams to defending a few days apart. Geoff is an inspiration and has pushed me to do better design and research work. He has been a valued source of feedback on all matters of design, and I have appreciated his friendship.

I would also like to thank my other two research colleagues Tomonori Honda and Catherine Elsen, who were an integral part of my research and shaped it in many ways. Tomo's keen analytical mind kept the analysis on track, while Catherine was an important part of the design of the experiments. Catherine Fox was also a great help with data processing.

The Ideation Lab has been an unforgettable home, and the people in it exceptional. Everyone I have crossed paths with has shaped my time at MIT in some way, but I would especially like to thank Jesse Austin-Breneman, Jim Christian, Alison Olechowski and Janet Yun for their friendship. You have been an important part of my life at MIT. I would also like to thank my friends at Eastgate, for creating a community.

Before coming to MIT, Kalevi 'Eetu' Ekman, Matti Hämäläinen and Lauri Repokari at Teknillinen korkeakoulu were instrumental in introducing me to product design, and were there in the beginning.

T. Shawn McGrath introduced me to science, engineering and most of all, critical thinking. Thank you for believing in me when others did not, and for being a mentor and a friend. I will always think back fondly to our time in Windhoek. Thank you for everything you taught me – I would have had so much more to learn. I wish you could have seen me graduate, although I am sure in your mind there was no doubt I would.

Sten-Erik Häggman, my uncle, for all the conversations, help and encouragement, I thank you. Our farewell came all too soon, but I am happy for the time we had.

I would also like to thank my parents, Pirjo and Bjarne, for the endless support throughout the years. Thank you for giving me the freedom to chart my own path, even though at times the destination was unclear. Thank you for all the support and help during our time at MIT. Without you none of this would have been possible.

A big thank you to my little brother, Anton. Although five years younger, you have never let that bother you and with your example you have always challenged me to do better. I admire your courage to pursue your dreams.

Finally, I would like to thank my family; my three lovely daughters, Emilia, Matilda, and Lina – who mean everything to me, for being inquisitive, outrageous, hilarious, and empathetic – you have demanded much, but given infinitely more, and my wife, Anna, for making it all possible. Without all your sacrifices I would not have been able to pursue my dreams, and for that I am eternally grateful. Thank you. I love you all.

The Ideation Lab has been an unforgettable home, and the people in it exceptional. Everyone I have crossed paths with has shaped my time at MIT in some way, but I would especially like to thank Jesse Austin-Breneman, Jim Christian, Alison Olechowski and Janet Yun for their friendship. You have been an important part of my life at MIT. I would also like to thank my friends at Eastgate, for creating a community.

Before coming to MIT, Kalevi 'Eetu' Ekman, Matti Hämäläinen and Lauri Repokari at Teknillinen korkeakoulu were instrumental in introducing me to product design, and were there in the beginning.

T. Shawn McGrath introduced me to science, engineering and most of all, critical thinking. Thank you for believing in me when others did not, and for being a mentor and a friend. I will always think back fondly to our time in Windhoek. Thank you for everything you taught me – I would have had so much more to learn. I wish you could have seen me graduate, although I am sure in your mind there was no doubt I would.

Sten-Erik Häggman, my uncle, for all the conversations, help and encouragement, I thank you. Our farewell came all too soon, but I am happy for the time we had.

I would also like to thank my parents, Pirjo and Bjarne, for the endless support throughout the years. Thank you for giving me the freedom to chart my own path, even though at times the destination was unclear. Thank you for all the support and help during our time at MIT. Without you none of this would have been possible.

A big thank you to my little brother, Anton. Although five years younger, you have never let that bother you and with your example you have always challenged me to do better. I admire your courage to pursue your dreams.

Finally, I would like to thank my family; my three lovely daughters, Emilia, Matilda, and Lina – who mean everything to me, for being inquisitive, outrageous, hilarious, and empathetic – you have demanded much, but given infinitely more, and my wife, Anna, for making it all possible. Without all your sacrifices I would not have been able to pursue my dreams, and for that I am eternally grateful. Thank you. I love you all.

Signature redacted

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

1 INTRODUCTION

1.1 The design process

Unlike many other tasks that people encounter in their work or daily lives, the design process does not have a well-defined 'correct answer' towards which to strive, and some would argue that the problem itself is similarly not well defined (Simon 1973). One example of this understanding in popular culture is the joke below, of which there are several variations:

> ***'How many designers does it take to change a light bulb?'***
> ***'Does it have to be a light bulb?'***

This illustrates the basic tenet of the design process: there is no correct answer – the first step is finding ~~the correct~~ an ***appropriate*** question. The design process is not a search for the absolute truth, but rather an exploration of several different possible solutions to a complex problem. One could say that the solution and the problem develop together (Cross 1999).

Design is the process of finding better solutions to current problems. There exists a diverse set of different design processes, which differ based on their goals, executors and who formulated the process. They are all, however, a sequence of steps (that may loop back to earlier steps) focused on achieving a goal. Some have described the design process as a purposeful activity that includes both complex internal steps (thinking, evaluating and deciding) as well as external steps (writing, drawing and speaking). (Römer, Leinert, and Sachse 2000)

As mentioned, there are numerous different theoretical models that try to describe the design process, although variations in actual practice make it hard to do so precisely. These are, none-the-less, valuable tools in helping to think about the process of designing, and how one might go about improving it. Some commonly referenced design processes are presented by Ulrich and Eppinger (Ulrich and Eppinger 2003), and Pahl and Beitz (Pahl and Beitz 1996). The 'Compendium of models' by Dubberly (Dubberly

2004) also gives a general overview of several different models suggested by design researchers and practitioners.

My personal favourite is perhaps still the one created by Damien Newman (Newman n.d.), which visually describes the chaotic first exploratory steps of the process, with more clarity emerging as the process moves left to right, towards the final concept at the end.

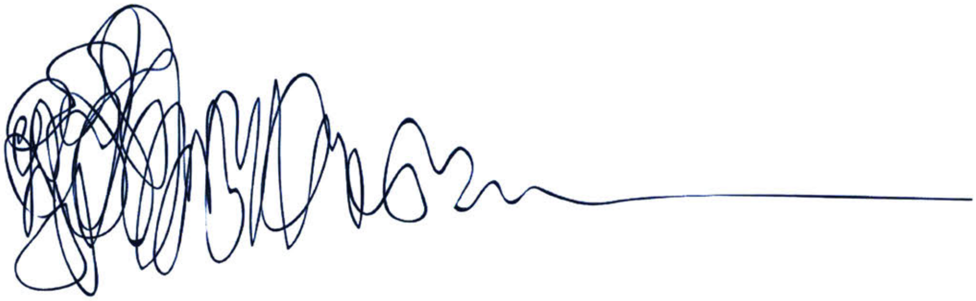


Figure 1 – The Design Process as Described by Damian Newman

In the context of this thesis 'design process' will refer to solution-focused human-centred design processes, which do not set out to solve a specific problem, but rather have a solution or goal in mind, namely a better future situation. This type of design process is often called 'design thinking'.

According to Meinel and Leifer (Meinel, Leifer, and Plattner 2011), there are four main principles to design thinking, namely:

> ***The human rule – all design activity is ultimately social in nature***
>
> ***The ambiguity rule – design thinkers must preserve ambiguity***
>
> ***The re-design rule – all design is re-design***
>
> ***The tangibility rule – making ideas tangible always facilitates communication***

Distinct from analytical thinking, in design thinking one of the early phases is called 'brainstorming', during which several early ideas are explored with little or no limits on breadth. Little investment in terms of time and effort has been put into them, reducing the cost of failure and increasing the willingness to make changes. The focus of this thesis is on these initial phases of the design process when an individual designer formulates their mental models into tangible or visual concepts, and therefore the principles proposed by Meinel and Leifer (which pertain more to actions following the formalisation of ideas, or to design teams) will not be discussed in detail. This thesis will focus on the ways that ideas that designers have are formalised, and why that matters.

1.2 The early stages matter

Once initial needs and requirements are formalised, one of the first actions a designer does is to brainstorm and create initial concepts. The early stage concepts are important as they lay the foundation for the design process. Several studies (Asiedu and Gu 1998; Ben-Arieh and Qian 2003; Dowlatshahi 1992; Duverlie and Castelain 1999; Huthwaite 1988; Schütze, Sachse, and Römer 2003; Wu 1998) mention that the early phases of the design process influence a majority of the total cost of a product (many mention figures around 70%-80%), although most of these studies only reference a handful of original research conducted – some of it anecdotal – in the automotive industry in the 1980's (Corbett and Crookall 1986; Whitney 1988). Some researchers also suggest manufacturing decisions are of equal importance in many contexts (Ulrich and Pearson 1998).

Regardless of the true number, it seems widely accepted that the early stages of the design process are important for the overall success of a given design, and that although creative problem solving is valuable at any stage of the design process, 'it is of critical importance in the conceptual design stage' (Robertson and Radcliffe 2009). Or as Huthwaite put it, the design of a product has a 'ripple effect' that 'reaches into every company function and moves forward into time', and that 'the penalties imposed by mistakes made during design reach every company employee from order clerk to floor supervisor' (Huthwaite 1988).

In real world situations, designers typically have a deadline to meet with a limited amount of resources at their disposal, and therefore, need to create drawings and models that elicit the maximum amount of useful information for the designer, with the minimum amount of spent time and effort

(Macomber and Yang 2011). Knowing which design processes and tools are most effective to use is important. Ulrich and Pearson also point out that 'design is hard' – different design teams will exhibit differences in design capability and produce quite different outcomes (Ulrich and Pearson 1998). In other words, there is skill involved, and the expertise of the designer matters.

Ulrich and Pearson also found that perceived industrial design quality and reliability quality were not significantly correlated with manufacturing costs. They called this 'design is free' (Ulrich and Pearson 1998), which is to say that with better processes and methods, designers can create products that have a higher aesthetic quality and are more reliable, without incurring additional manufacturing costs.

Although many factors are pre-requisites for a successful design, novel concepts or new creative ways of combining or utilising existing components or solutions are often required. Creativity is therefore an important part of any design process.

1.3 Research questions

As noted earlier, the design process is complex and ambiguous, with no clear answer and results that are difficult to judge. None-the-less, there seems to be wide agreement that in addition to the skills of the designers, the process itself is an integral part of determining the quality of the design outcome. Although products and services are in most cases designed by a design team (which can consist of product designers, engineers, industrial designers, anthropologists, and a number of other professions, that in this case will all be simply referred to as designers in a design team), significant amounts of initial ideation occur individually, after which the results are shared with the design team – although group brainstorming sessions are also common.

However, in order to focus on a manageable number of metrics, and since a lot of research on design teams already exists, the focus of this study is on individual designers working alone in the early conceptual phases of the design process.

With the proliferation of computer tools designers have begun moving to computer-aided modelling (CAD) earlier in the design process, instead of relying on hand sketching further into the design process. CAD poses a unique set of constraints that, it is assumed, negatively affect the quality and breadth of the ideas that are produced due to the constraining nature of current computer tools. Software companies are aware of this issue, and have begun designing tools aimed at the earlier stages of the design process, to facilitate more free-form unconstrained idea generation.

Similarly, making physical prototypes is often cited as an important part of the design process, but little research exists on the effect these different design tools have on individual designer (although several studies have examined the role different design representations play in facilitating communication and idea exchange between designers in a design team).

In order to study the influence that a select number of design tools (free-hand sketching, foam prototyping, and computer modelling software) have on the concept generation process in individual designers, the following research questions were formulated:

1. *Are there significant differences in terms of the creative value of concepts between the chosen design tools?*

Does the use of a specific design tool influence the mental processes of a designer to the extent that the concepts produced are of a different creative value? Simply put, are ideas produced by sketching, for example, more 'creative' than concepts produced while using computer modelling software?

2. *How does the rate of idea generation differ between the chosen design tools, in terms of quality of design outcomes?*

Although it is obvious that concepts created by computer modelling software will take longer to make than concepts created by foam models and hand-sketching, how is this associated with the quality of the design outcomes. In other words, when examining the concept output of designers in terms of quality and time, how do the tools differ?

3. ***Do the different chosen design tools influence designers to create a certain type of concept, distinct from the other design tools examined?***

Does the tactile feedback of making foam models make the designer more aware of ergonomic issues, thereby influencing them to produce – presumably – more organic shapes? Do designers using computer modelling software create more rectangular designs? Do designers who hand-sketch create more ambiguous concepts?

2 LITERATURE REVIEW

Extensive research exists on the design process and the design tools used. As this study focuses on comparing three common design tools – sketching by hand, foam prototyping and computer modelling – the literature review will focus on how these are used in the design process. Additionally, as crowd sourced data forms the basis for a large part of the study, crowd sourced data collection methods will also be discussed.

2.1 Design tools

Sketching and drawing are some of the oldest forms of visual communication known to man, and precede written language by thousands of years (Tversky 2011). From cave paintings to Picasso, humans draw naturally. Young children will draw if given a medium and canvas to express themselves with, and sketching is initially at least, a very natural form of expression. It is a quick and flexible way to explore design ideas (Cross 2000), although severe societal constraints may be imposed as children grow up ('I can't draw') through self-criticism and censoring. None-the-less, it is assumed that designers assigned to the sketching group will be less constrained in expressing their ideas, leading to larger breadth and variety.

Sketching by hand is an effective technique for early stage design (Eckert et al. 2012), and an integral part of the conceptual phases of the design process. There are several different types of sketching, suited best for different phases of the design process, and serving different purposes. Kudrowitz et al. categorise sketching into three main types: thinking, explanative, and persuasive (Kudrowitz, Te, and Wallace 2012), although other researchers have also proposed alternative categorisations. As these different types of sketching aim to achieve different goals, the type of sketching employed may also have an effect on the results, as will be seen later.

One of the perhaps most obvious functions of sketches are to convey ideas and meaning; 'this is what it will look like' (explanative sketches). Maps convey information about terrain, portraits about physical appearance, and blueprints give instructions on how to build. In the same way concept sketches externalise thought (Tversky 2011) and convey the idea of what the concept will look like, and possibly how it will function.

Freehand sketching is a common way for designers to express and develop their design concepts (thinking sketches), and some research has found that designers who sketched on paper were better at formulating and analysing the design problem (Bilda and Demirkan 2003; Cross 1999; Jonson 2005; Römer et al. 2000), more efficient with their time use, and produced more alternative solutions in the conceptual design phase than designers who sketched in digital media (Bilda and Demirkan 2003).

When designers sketch, they have a 'discussion' with the sketch while working through the design problem, and may realize things they would not otherwise have (Visser 2006). The 'roughness' of quick sketches can induce creativity, as the designer re-interprets their own sketches (Tseng and Ball 2011), and the ambiguity allows for the exploration of different design alternatives (Goel 1995), which is in line with Meinel's and Leifer's second design principle, 'design thinkers must preserve ambiguity' (Meinel et al. 2011).

Sketching can also act as a repository of information, by allowing the designer to unburden themselves of their ideas (by sketching them down on paper) thereby lightening the mental load, and freeing up resources for processing new ideas or thinking of ways to combine existing concepts into new combinations. Sketching down ideas also enables designers to return to them later in the design process (Pan, Kuo, and Strobel 2013).

Some argue that drawing in classrooms may aid the learning experience, enhance engagement and make students understand concepts on a deeper level (Ainsworth, Prain, and Tytler 2011). Unfortunately, students in one study felt they did not have sufficient opportunity to sketch, that their classes did not provide them with sufficient opportunities to improve their skills, and that their teachers were lacking as role models for sketching (Jonson 2005). Some research does suggest, however, that sketching may not necessarily improve the design outcome for expert designers in all situations (Bilda, Gero, and Purcell 2006).

Building physical models offers an alternative way to formalise ideas, offering tactile information not available in sketching. Foam modelling was chosen as one of the design tools to study, due to the fact that it is easy to work with, does not require special expertise or tools, and offers an avenue to create organic shapes easily – something not all physical prototyping methods are suited for (think for example about Lego). When one hears the word 'prototype', it may trigger different mental images of people in white lab coats and frizzy hair working on a complicated apparatus, pre-production cars shown at car shows, or a prototype aircraft taking off for its maiden flight. However, as this study focuses on the early stages of the design process, prototypes in this context are often significantly less detailed, and focus on a specific question rather than demonstrating system level performance. There is a widely circulated image of an early prototype for a medical device conceived of at IDEO, that demonstrates the level of fidelity that is referred to when talking about prototypes in this study (Kelley and Littman 2005) – on the left, and early prototype of the medical device, on the right, a computer rendering of the final product (Figure 2).

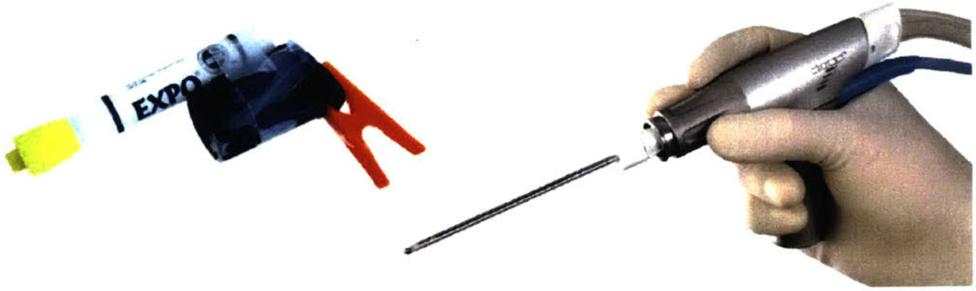


Figure 2 – IDEO prototype for medical device

Low-fidelity prototypes (Yang 2005) can be quick to fabricate, and reduce uncertainty in the design process (Gerber 2009). The question then arises, how low-fidelity can a prototype be? 'Is a brick a prototype?' The question posed by Houde and Hill elegantly presents the problem of defining what a prototype is. As they point out, the answer to the question depends on how the brick is used. Simulating the scale or weight of an object with the brick would make it a prototype, demonstrating how prototypes are not necessarily self-explanatory. What matters is how the artefact is used to help the designer or designers explore or demonstrate a certain aspect of the concept, not what it is made of, or how it was made (Houde and Hill 1997). Although outside the scope of this thesis, research also suggests the importance of prototypes in design team settings (Edelman et al. 2009; Schrage 1999).

In addition to sketching and prototyping, computer modelling was considered. Since their inception years ago, computer-aided design (CAD) tools have become more and more popular at all stages of the design process. Newly graduating students are more competent and comfortable with computer tools, and as they have developed to become more efficient and intuitive, they have also begun to replace the traditional notepad in the early conceptual phases. (Veisz et al. 2012)

CAD tools have become popular in part due to the fact that they are very useful for communication and visualization of concept ideas (Robertson and Radcliffe 2009). However, one of the challenges with computer models is that there is a cognitive bias to accepting detailed representations as being superior to abstract representations (Veisz et al. 2012) as they can convey an illusion of being complete, and thereby discourage creative thought in group situations (Robertson and Radcliffe 2009). Some design practitioners also reported using computer modelling as a result of the expectations of clients to see photo-realistic images early on in the design process (Jonson 2005).

Typical problems that may arise while designing with computers include circumscribed thinking, premature fixation and bounded ideation. Circumscribed thinking refers to situations where the design tool limits the designer through interfering with the designer's intent. (Veisz et al. 2012)

Premature fixation refers to situations where the designer becomes devoted to a certain idea prematurely, without exploring a full range of other possible design avenues, because of a high level of detail or complexity of the model they have created (Linsey et al. 2010; Veisz et al. 2012). In a case study, Robertson and Radcliffe found that designers were wary of incorporating modifications to existing CAD models if it involved making too many changes to the model or its underlying structure, even if these modifications would solve several problems at once, or reduce overall project risk (Robertson and Radcliffe 2009). Some research also suggests that using CAD tools too early in the design process encourages a focus on detailed design, rather than concept exploration, in other words a depth rather than breadth approach (Fixson and Marion 2012; Ullman, Wood, and Craig 1990).

Bounded ideation, on the other hand, refers to an overuse of a design tool, in this case a computer modelling program (CAD), which reduces the designer's motivation and creative abilities (Veisz et al. 2012). Some research suggests that teams that use advanced 3D CAD tools for a larger proportion of the workday create fewer ideas than teams who use less time working with computer modelling programs (Robertson and Radcliffe 2009).

Although a lot of work has gone into developing the functionality of computer tools, they still often limit the solutions available to the designer, and it is possible that CAD tools may never match the imaginative capabilities of designers (Robertson and Radcliffe 2009). Robertson and Radcliffe, however, note an even more serious problem with computer tools in that they may not only limit creativity by what is possible, but also tend to push design decisions towards what is easiest to create with the available tools (Robertson and Radcliffe 2009).

According to Robertson and Radcliffe, there is growing evidence that the over-use of computer modelling tools is influencing the ability of designers to solve engineering problems creatively. There also seems to be a tendency to over-use CAD, even in the conceptual design stage, where other design tools might be more appropriate. (Robertson and Radcliffe 2009)

In order to study the nuanced differences between concept generation using the aforementioned three design tools, it was determined that a large pool of design reviews would be needed to reliably determine any possible differences in outcomes. Online data collection was determined to be the most feasible method of collecting reliable data, which will be discussed next.

2.2 Mechanical Turk as a research tool

Amazon's Mechanical Turk (later simply Mechanical Turk, or M-Turk) is an online crowdsourcing tool that co-ordinates the supply and demand of, usually relatively short and uncomplicated, tasks that require human intelligence to complete. Mechanical Turk is named after an 18th century chess playing 'automaton' that was in fact operated by a concealed person. (Paolacci, Chandler, and Ipeirotis 2010)

Tasks, called HITs (Human Intelligence Tasks), are posted on the Mechanical Turk website, where workers can browse posted HITs and complete them for usually very modest financial rewards (Gosling et al.

2004; Mason and Watts 2010) in a short amount of time (Buhrmester, Kwang, and Gosling 2011; Paolacci et al. 2010). Some research suggests that the compensation amount does not impact data quality, just the rate at which responses are collected (Buhrmester et al. 2011; Mason and Watts 2010).

In their experiments Horton and Chilton found that Mechanical Turk workers had a median reservation wage (the wage at which they will no longer accept the job) of $1.38/hr. (Horton and Chilton 2010). There doesn't seem to be a minimum absolute wage though, as many respondents were willing to complete HIT's for even 1¢ – assuming the task was short enough (Buhrmester et al. 2011). Although varying from month to month, (Ross et al. 2010) estimated average wages to range from $1.50/hr. to $2.00/hr., with a bit over half reporting working less than 5 hours per week and around 60% of US workers declaring that the Mechanical Turk income had no impact on their financial situation, implying that they were completing the HITs due to some combination of other inducements, such as entertainment, information, the chance to be altruistic or gaining attention from others (Horton and Chilton 2010).

The Mechanical Turk workforce has previously been predominantly from the United States (Ross et al. 2010), but although greater proportions of Indian subjects have become available in recent years (Eriksson and Simpson 2010), it is assumed that the demographics of the US population of M-Turk workers has not changed, and is still representative of the US population as a whole. To avoid cultural and standard of living differences impacting the results, only reviewers residing in the US were chosen to complete the tasks, or HITs, by only accepting IP addresses from the United States.

Research suggests that online subject pools, although not perfect, are diverse, and ***more*** representative of the US population as a whole (and less prone to biases) than traditional subject pools recruited through universities (Buhrmester et al. 2011; Gosling et al. 2004; Paolacci et al. 2010).

In addition to a subject pool more representative of the general populace, a major benefit of online surveys is that they allow research to obtain a sample size that far exceeds those obtained with most traditional techniques. Some less obvious benefits also include reduced need for manual data entry (as the data is gathered electronically) (Gosling et al. 2004) and ease of payment and subject pool collection (Chandler, Mueller, and Paolacci 2014).

Paolacci et al. found that non-response error was the most challenging aspect of collecting data through Mechanical Turk, while at the same time noting that this error was likely higher in traditional web studies. In a study by Paolacci, subjects recruited through M-Turk were far more likely to complete the survey than subjects recruited from online discussions forums (91.6% and 66.7% respectively). (Paolacci et al. 2010)

In all other aspects, apart from multiple response errors, (Paolacci et al. 2010) found that M-Turk data was less susceptible to errors ranging from heterogeneity of samples to experimenter effects and dishonest answers, than was data collected in laboratory settings and was overall a reliable source of experimental data in judgement and decision making. M-Turk data quality also met or exceeded the psychometric standards associated with published research (Buhrmester et al. 2011).

Mechanical Turk was chosen due to the benefits associated with collecting data through Amazon Mechanical Turk, the most important of which was the fact that large populations representative of the general US populace could be reached quickly and inexpensively.

3 METHODOLOGY

Design is a highly complex, constantly changing process, affected by a multitude of inputs that are hard to describe, many of which are unknown. As such, it is clear that all facets of the process cannot be studied at once. The focus of the research was on early stage concept generation, and hence the methods were chosen to emulate – as far as possible – an actual early stage design experience, so that insights could be gained that could then be used to improve design practice.

3.1 Introduction

Some of the key features that were identified as important in re-creating a design scenario that was as realistic as possible, were ***motivation*** to complete the task in a serious manner, and a ***sufficiently long observation period***. To address the first point, careful consideration was given to specific ways of motivating the participants. Without any external pressures to perform, it was feared that participants might take the experiment as an opportunity to have careless fun, and create 'wilder' ideas than they typically would in a work environment.

As the participants were anonymous, peer recognition could also not be utilised as motivation, although several of the participants were curious and wanted to know 'how they had performed' compared to other designers who had taken part in the study. In the end, a monetary incentive and an (anonymous) competition were devised to provide an appropriate level of motivation and pressure on the participants. The design experiments were conducted in two phases, and data was collected through controlled experiments and surveys, in two sections labelled part I and part II, as seen in Figure 3.

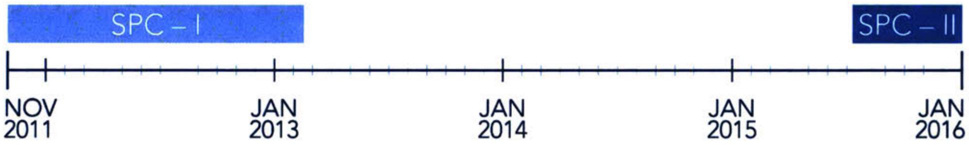


Figure 3 – Data collection periods

Participants in both parts were informed that there would be roughly thirty designers that they would be competing against, and that the designer who created the 'best' design – as judged by an independent panel of judges – would be awarded an additional $75 in addition to the $50 that was already given to everyone as a token compensation for their time.

There were twenty-one participants in part I, and six in part II. Additionally, some participants in part I were disqualified due to the fact that they did not produce any concepts for products (instead producing new service and business models), further improving the odds in favour of the participants. Although, it seemed that for many of the designers, winning the competition was more motivating than the additional cash prize.

One issue that design experiments face is the disconnect between studies performed in the lab, which typically range from a few minutes to an hour, and real working environments, where designers may be working on the early conceptual stages of the design process for days. Further complicating the issue is the fact that designers may be thinking of the challenges and problems even while not at work.

Reconciling the challenges of recording reliable data on participants not tightly monitored in a controlled experiment environment (but providing data over a longer period of time) with data collected during a shorter duration task in a tightly controlled environment (providing better quality data, although in a less realistic environment) it was decided that in order to control as many variables as possible the experiment data would be collected in a laboratory environment, but that the length of the experiment would be pushed as far as reasonably possible.

Based on previous experience and initial probes sent via e-mail to design firms and design students, it was assumed that the sample size of suitable designers for the study would be relatively limited. The number of dependent variables was kept small, in order to be able to make meaningful insights into the design process with the (assumed) relatively small sample size. Factors that were predicted to influence the design process, but which could not be studied with the current sample size were identified and their impact minimised as far as possible, some examples of which are given below.

For instance, length of work experience, design background (architect, industrial designer, product designer, mechanical engineer) and age were factors that were not examined – due to the scope of the study, it was not possible to make meaningful comparisons. Therefore, background factors that could not be studied were mitigated by distributing designers into one of the three groups in an even fashion, to minimise their impact on the data.

One of the distinguishing features of this research compared to many others is the use of, mainly, design professionals as research subjects, and the longer than usual observation period. Although teamwork is an essential part of many well functioning design teams, significant amounts of ideation and early concept development also happens in single-person settings (alone). Therefore, since research directly comparing different design tools is limited, and due to the fact that design teams have been more readily studied, only individual designers working alone were observed. This was done in order to decouple the substantial effects that team dynamics have on the design process; the design process in team settings was left outside the scope of this study.

3.2 Controlled design experiments

The following section describes controlled design experiments that were conducted on professional designers that were individually engaged with a design task; using either sketching, foam prototyping or computer modelling. In both parts I and II, the designers were interviewed, the artefacts they created were collected, and audio- and video recordings were made. Additionally, further data was collected through online surveys for artefacts created in part I, as shown in Table 1.

Table 1 – Type of data collected during parts I and II

	video	audio	interviews	on-line surveys	design artefacts
part I	●	🔵	●	●	●
part II	🔵	●	●		🔵

● data used

🔵 data collected, but not used

◌ data not collected

The experimental set-up and design brief remained nearly identical in both parts to allow for direct comparisons between experiments, with the exception of a few minor differences, described next.

During the experiments in part I the designers were allowed to design freely, whereas in part II, the designers were asked to verbalise their thoughts continuously throughout the experiment. As both experiments were nearly identical, unless otherwise explicitly specified, the methods described apply for both parts. In cases where the experimental conditions differ, specific mention will be made whether the condition applies to part I or part II.

In addition to the two time periods during which the design experiment was run, additional data was also collected through an online survey based on artefacts from part I. The methodology for the online survey is described later in section 3.3.2 on page 31. The design experiment itself is described next.

3.2.1 Experimental set-up for part I

The design experiment was devised to study the impact of three different design tools on the design process and design outcomes. To that end, participants were divided into three groups – sketching, foam prototyping, and computer modelling. Due to the expectation that qualified participants for the study would be hard to find, the number of design tools being studied was limited to three, so as not to spread the participants too thin over several design tool categories.

Sketching, being an essential tool in any ideation process, was chosen as one of the tools to be studied. Due to the proliferation of computer tools in the last couple of decades computer-aided design tools have become ubiquitous in the design field – even for very small companies or individual consultants. The main challenge with studying CAD tools, however, is that there are a large number of different variations with different fields favouring different programs. Some common examples include programs made by Dassault

Systèmes, Autodesk, Siemens PLM Software, PTC and the Blender Foundation.[1]

Based on a few sample contacts within the pool of possible candidates for the study, it was found that SolidWorks (by Dassault Systèmes) was a widely used tool amongst the pool of designers who were candidates to take part in the experiment. To remove bias caused by the use of different CAD programs, all participants in the CAD group used SolidWorks. Furthermore, only subjects who were familiar and comfortable with SolidWorks were recruited to take part in the computer-aided design tool group.

It was hypothesized that the tactile aspect of building physical models may have an impact on the thought processes of designers, and therefore a third design tool that offered tactile feedback to the designer was sought. Several different alternatives from foam core and Lego to clay were considered, but ultimately polystyrene foam, referred to simply as 'blue foam' or 'foam' in this study, was chosen based on its relatively widespread use in the early stages of the design process and suitability for making quick mock-ups in both geometric and organic shapes, with fewer restrictions on the shape than some of the other quick, physical mock-up materials.

Clay was also considered as an alternative, but due to it being less commonly used it was assumed that it would be easier to find designers who were comfortable working with blue foam than ones who were used to working with clay (which was corroborated by early inquiries with people who were considered possible participants in the study).

Designers in the sketching group were provided with a variety of pencils, markers, ruler, eraser and pencil sharpener. Designers in the foam prototyping group were given a variety of rasps and grits of sandpaper, a metal ruler, toothpicks and a variety of glues to join foam pieces, a hot wire cutter, and a marker for drawing out cutting lines. Brainstorming or sketching out ideas with the marker was expressly prohibited. Participants in the computer modelling group were provided with a workstation and copy of SolidWorks. For a more detailed discussion on the specific equipment used in the study, see (Häggman et al. 2015).

[1] See www.3ds.com, www.autodesk.com, www.plm.automation.siemens.com/en_us, www.ptc.com, and www.blender.org/foundation. All accessed 2 September 2016.

Participants were provided with a working table, quiet location to work in, water and design tools according to the group they were assigned to. For participants in part I of the study, two cameras were positioned to record the design experiments, as seen in Figure 4. One camera was angled top-down to give a view of the working surface (labelled A), and the other (wider angle) camera, was positioned to capture the working table and hands of the participant from a lower angle (labelled B).

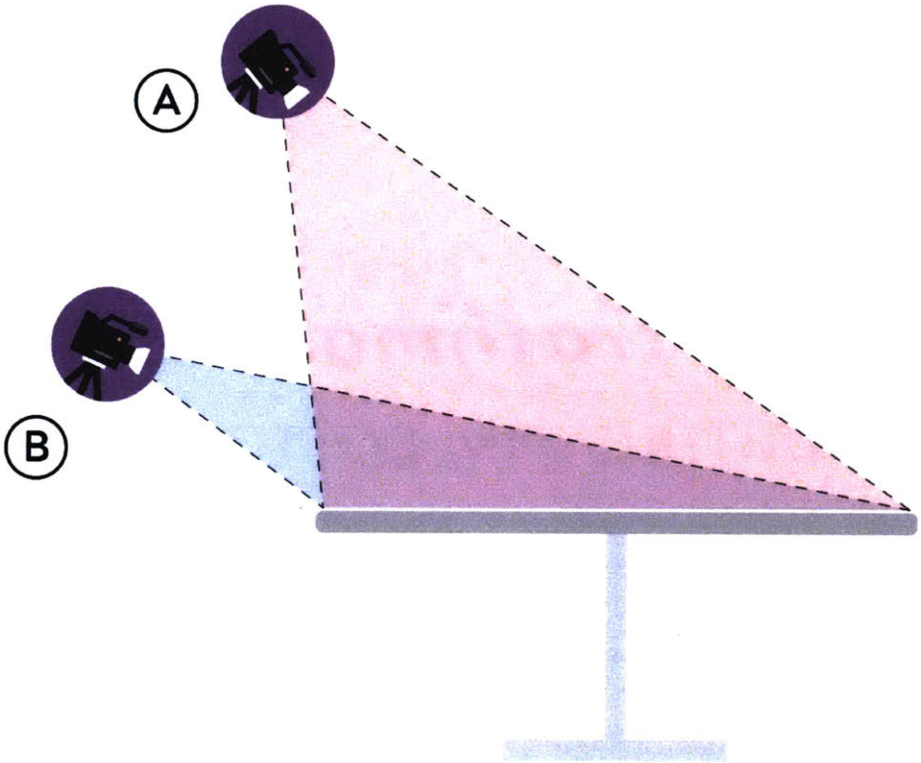


Figure 4 – Experimental setup for part I

The working area for the sketching and prototyping groups was roughly 80cm x 90cm or larger, but varied slightly based on experiment location. Participants in the computer modelling group were provided a conventional computer workstation – as they were prohibited from sketching they did not need table space in the same fashion that participants in the sketching or prototyping groups.

Participants worked alone in a quiet room with few distractions. They also did not know who the other participants of the study were (the experiments were not run consecutively), and more importantly, were not recruited from the same social circles (for the most part, the designers worked in different companies, lived in different parts of the city, and had few friends in common) – therefore one can safely assume, that the designers had not heard details regarding the experiment beforehand.

3.2.2 Experimental set-up for part II

The experimental set-up for part II was largely the same as in part I. However, due to the implementation of a talk-aloud protocol, there were some minor additions. The main differences were the addition of a high-quality microphone (to ensure the audio quality was good), and the placement of a permanent experiment facilitator who sat in the room with the participant for the duration of the experiment. The main task of the observer was to monitor the test subject, and make sure they kept verbalizing their thoughts, by gently reminding the participants to 'please, continue speaking' whenever they fell silent. If the test subject was silent for roughly 5-10 seconds, they would be reminded to speak, although depending on the frequency of previous reminders and what they were doing, the time after which a reminder to speak would be expressed could be increased slightly, to avoid having to repeatedly remind the test subject and thereby annoying them and possibly influencing the data by affecting their mood.

As it was foreseen that having an experiment moderator sitting in the experiment area observing the participants might make some participants self-conscious or uncomfortable, a few abbreviated trial studies were conducted to test the effects of moderator placement. These trial studies lasted roughly half an hour. The data collected during these sessions was not used in the actual experiment, as the session durations were incompatibly short, and the experiment moderator moved around trying different seating arrangements, undoubtedly affecting the concentration of the participants.

The participants knew that the experiment was a trial study and that they would be abbreviated in length, but were not made aware of what specifically was being tested. The trial experiments were abbreviated to roughly one design session, instead of the three sessions in the actual experiment. After the design session the participants were asked if they had noticed the different seating arrangements that the experiment moderator had sat in, and if so, how the different seating arrangements had affected them.

In addition to the general discomfort that most participants felt due to the awkwardness of verbalising their stream of consciousness and talking in incomplete sentences, there were three specific findings as well. The three main qualitative findings from these abbreviated trial studies dealt with: a. the positioning of the experiment moderator, b. the facial expressions of the moderator, and c. the importance of reminding the participants to speak.

Although some subjects did not seem to care about the positioning of the moderator, some participants had surprisingly strong emotional reactions and expressed anxiety and pressure caused by the placement of the moderator when in their field of view. Three different seating arrangements were tested with each trial participant. Based on verbal feedback, a position behind the test subject, out of the field of view of the participants (labelled 'C'), was chosen as the least obtrusive, as depicted in Figure 5.

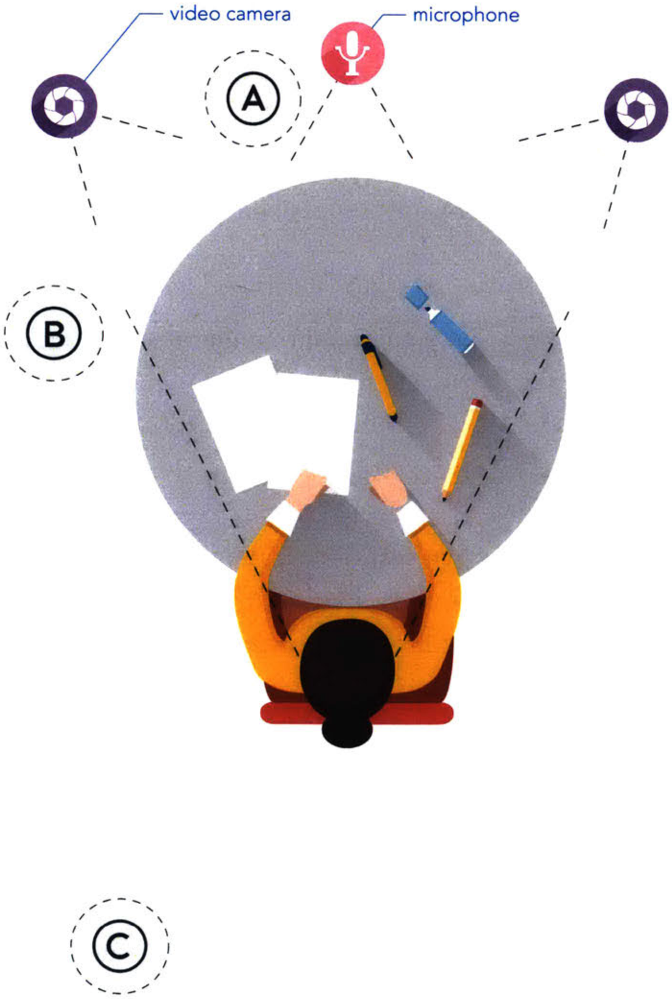


Figure 5 – Experimental setup for part II, with placement of observer

Even though position C caused some anxiety in some of the participants, it was none-the-less unanimously seen as the least obtrusive in all of the trial studies. Although not a concern in the chosen seating arrangement (as the test subjects could not see the experimenter), the trial experiments revealed that when seated in the field of view of the test subject, some subjects felt a need to glance at the experimenter from time to time, and some explicitly mentioned that they observed the facial expressions of the experimenter and used them as immediate feedback in their design process. This effect was not quantified in any rigorous manner as interviews unanimously suggested that seating position C was the least obtrusive, and therefore the possible effects of facial expressions while seated at positions A or B were irrelevant.

The final key learning from these trial studies was that it should be expected that some participants would need a considerable amount of reminding to keep talking. In the actual experiments this did not turn out to be an issue to the same extent as during the trial study.

Additionally, due to the large amount of talking – nearly three hours of non-stop speaking – the participants were instructed to re-schedule the experiment date in case they got a sore throat or any other condition developed that could negatively affect their ability to talk for prolonged periods of time. Bottled water was made available for all the participants in case they wanted to drink during the experiment.

3.2.3 Participant recruitment

Designers for the study were recruited in the Greater Boston area in Massachusetts, USA and in Liège, Belgium, through e-mail advertisements sent to design-related e-mail lists, and through direct contact with designers at different design consultancies. A few graduate design students from MIT were used to pilot the study, and some students with work experience were also used in the main study. Participants were compensated $50 for their time, with the possibility of winning an additional $75 if their design was chosen as the best one.

3.2.4 Assigning participants into groups

Prior to the experiment, the designers were asked for their educational background and work experience, and were asked to self-report their perceived skill level on a five-point Likert scale for free-hand sketching, foam prototyping and SolidWorks (CAD). They were assigned to one of the three groups based primarily on their skill level – participants were usually assigned to the group that they reported having the best skills in, although this did not always necessarily mean they had given their skills a score of 5 (on a scale from 1 to 5, where 5 is the best).

In some cases participants asked whether they should judge themselves compared to other professionals in their field of work, or to an 'average' person. In these cases the participants were informed that they should rate their skills in terms of how much they limit their ability to express their ideas. In other words, if they felt that the tool in question did not hinder their ability to express their ideas at all, they should choose 5, regardless of how they imagined they ranked against others in their field, or against 'average' people. Put another way, they were asked to judge their fluency with the tool and how much it hindered their concept creation process, not to try to evaluate their skills against their peers.

In case the participant felt very confident with several of the design tools, a second consideration was the distribution of prior participants into the three groups, in other words, if a certain group was lacking participants future participants were sought for that group to keep the number of participants in each group relatively even. Interestingly, even though computer modelling software is ubiquitous in the design workplace, it was challenging to find participants who felt very confident in their computer modelling skills. Even participants who could be considered 'digital natives' did not necessarily feel at ease with CAD software for concept development. Part of the reason for this was that the software we provided was SolidWorks (by Dassault Systèmes), and not everyone was familiar with it, even though it is widely used. There were also many participants who said that SolidWorks was their primary CAD tool, but even so, were arguably aware of the limitations of CAD, and therefore felt more comfortable with either blue foam or hand sketching.

For the experiment in part I, seven participants were assigned to the sketching group, six to the prototyping group, and five to the CAD group.

Three sketching participants (not included in the seven mentioned above) were excluded from the data as they did not follow the prompt and did not produce any concepts. Instead, they envisioned a new service or business model for the television viewing experience.

3.2.5 Additional warm-up task for part II

The main objective of the warm-up task was to assess the level of comfort the participants had with the talk-aloud protocol (also referred to as a think aloud protocol). Initial trial studies conducted with volunteers found significant improvement in verbalising fluency after the warm-up tasks, underlining their importance before the actual experiment. A three-step procedure was devised to prime the designers for the talk-aloud protocol before the beginning of the actual experiment.

The first step in the process was describing what was expected of the participants, in terms of vocalising their thoughts, not worrying about talking in complete sentences, and most importantly, not censoring their thoughts. They were instructed to say whatever came to their mind and to try not to censor themselves in any way; instead of 'describing' what they were doing (in which case an additional level of mental processing had happened) they were instructed to simply say whatever came to their mind (see Table 2 for clarification). They were also specifically instructed that incomplete sentences were acceptable, and that they should not worry about their verbalisation 'making sense'.

Table 2 – Examples of verbal output

acceptable answer	Hmm... let's see... this is pretty sharp... shiny... looks a little Apple-esque, expensive... pretentious... I want that modern look, and I think this edge here will give it to me... I just don't want to make this too round... there... that's looking better...	*Incomplete sentences and single words, which indicate less self-censoring. Timing of the words and pauses with the actions the participant is doing is also used as an indicator of whether or not the participant is mentally editing their verbal output.*
unacceptable answer	I decided to try and keep this edge quite sharp, to give the shape a modern look. It reminds me of an Apple mouse. I think in general Apple produts have a modern look that conveys quality, and that's what I am going for.	*The responses seem too thought out. Especially if they are delivered after the shape has been made. If the participant is silent while shaping the foam or sketching, then 'sits back' and explains what they did and why, it indicates that they are not narrating their thoughts as they go along, but rather giving a mentally edited version afterwards.*

The participants were assured that regardless of what they said, only the research team would be privy to the recordings and that they need not be worried about swearing or saying inappropriate things.

After the talk-aloud procedure had been explained to the participants and they had been given an opportunity to ask questions, they were asked a simple question to confirm they had understood the basics of the procedure. They were asked – and instructed to answer in line with the talk-aloud procedure – 'how many windows are there where you live?'. The role of the experiment moderator was to listen carefully to the participants, and make sure that they followed the procedure and answered the question in an acceptable fashion, as demonstrated in Table 3. If they answered simply by saying the number of windows, it would be considered incorrect. If the participant would have answered the first example question unsatisfactorily, the procedure would have been explained again, but all participants answered the first example question as intended.

Table 3 – Acceptable and unacceptable answers for talk-aloud task

acceptable answer	Well... let's see. I have one window in my kitchen... there are three in the living room, so that's four so far. Then there are two windows in my bedroom, and one in the bathroom, so what is that... seven in total.	*This answer indicates that the participant is thinking through the problem and vocalising it at the same time. Unless the number of windows is small enough that it is reasonable to assume one would remember it by heart, it is expected that the participant will go through the process of counting smaller sub-sections, such as per room, and then adding them.*
unacceptable answer	[silence]... I have nine windows in my apartment.	*Unless there is a plausible reason for why the participant would remember the numebr of windows by heart (if there are few enough, for example) it indicates that the particiant did not follow the talk-aloud protocol. In their mind, the participant probably went through room by room, added up the windows, and simply gave the answer.*

The desired answer would entail a verbalization of a mental process where the person goes through, in his or her own way, counting the windows in their apartment. In other words, an answer such as '...uh, well, there are three in the living room, another two in the bedroom... that makes five... and then there's one in the bathroom, so six. There are six windows in my apartment.' would be considered acceptable. All of the participants performed as expected, counting through the windows in their home, instead of replying with a simple answer. After the description of the talk-aloud protocol, and the cursory window counting test, a five-minute warm-up task was conducted.

The design brief for the warm-up task was to design their 'dream house' in five minutes using a set of provided Lego blocks. The task emulated the actual design task, in that it was conducted in the same physical work space as the actual experiment, it was video- and audio recorded, the designer was presented with a clear design brief, and it was timed. The fact that the data collected during the warm-up task would not be used later on was not communicated to the participants, in order to preserve the illusion of a design experiment, and make the warm-up task more similar to the actual design experiment that was to follow, and thereby more effective at uncovering possible issues that could be problematic during the actual design task.

In case the participant failed to verbalise their thoughts sufficiently, a third mock design task was also devised, at the end of which the participant would move on to the actual design task or be dismissed due to difficulties in verbalising their thoughts. However, although there were some differences in the level of encouragement needed for the participants to keep talking, at the end of the second warm-up task all participants were considered sufficiently fluent in verbalising their thoughts that the third reserve warm-up task was not needed.

3.2.6 Description of the main design task

The experiment followed a 'competition' structure – to motivate the designers to try harder, and to give some additional pressure to create designs that they actually thought would be viable (similar to ones they would present to a client who had asked them to create concepts for a new remote). The participants were told that an impartial jury would decide the winner, who would then be awarded an additional $75 in addition to the $50 that every participant received initially.

Without the additional motivation of a competition with a panel of judges and a monetary reward, it was hypothesized that participants may create more radical and exploratory designs than they would in real-world design situations, as they would have 'nothing to lose', so to say. It was hypothesized that the competition format added back an element of realism to the experiment, thereby, hopefully, making any findings and results more applicable to actual design scenarios.

The designers were not given any instructions on the type and fidelity of representations that they should produce, although the participants were told that they would have an opportunity to explain their design – how their concepts function did not need to be self-explanatory to someone simply looking at the concepts.

The design prompt was to create a remote control for a family of four – two adults, one teenager and one small child. The purpose of the remote was not strictly confined, and the participants were told that the remote could control 'a television, DVD player, digital video recorder, streaming console, game console, or any other device you feel appropriate'. A remote control was chosen to minimize the impact of any background knowledge or previous experience that the designers may have. In other words, a task was chosen where it was assumed that all participants had a comparable level of personal experience using the device, and no one had an unfair advantage due to having some specialist knowledge regarding the device in question.

Even if the participants did not have any remote controls at present, it is a fair assumption that they had all used a remote control several times during their lives, and were completely aware of the basic functionality of remote controls. Since the design task was situated into the very early concept development stages of the design process, knowledge of the technical details of how remote controls send commands to devices was not required, and it was assumed to not provide the participants with any significant advantage even if they had such knowledge. Furthermore, a remote was deemed to be of relatively low complexity, making it suitable for a short design task.

The designers were told that they had three 40-minute sessions to complete the task, that they could create as many concepts as they wanted, but that they could only submit a maximum of three concepts that would then be considered in the final competition. In reality all of the concepts they generated were considered. The requirement for the participants to choose between one and three concepts to enter into the competition was added to gain information on how they evaluated their own concepts, and to simulate a selection process.

The design task was initially intended to be 3 hours long, but based on the feedback from the pilot participants, the duration was decreased to 2 hours of design time, with an additional hour's worth of interviews combined. As can be seen in Figure 47 on page 81, even the reduced length experiment was too long for some participants, and not all of them used up the allotted time.

3.2.7 Running the experiment

Apart from the experiment moderator who would remain in the room at all times and prompt the designers to keep talking in case they fell silent, the experiment procedure was the same for both parts I and II. A step-by-step checklist was also created for the experiment moderators to follow, to decrease the likelihood of making mistakes in the procedure.

Participants were informed of their rights and were asked to sign a consent form. The participants were briefly interviewed before the beginning of the experiment, and again after each of the three 40 minute design sessions (for the pilot study, the sessions were 60 minutes long). A semi-structured interview format was used, where a list of questions was used as the basis for the interviews, but the interviewers were free to ask follow-up, clarifying, or probing questions when deemed appropriate. At the end of the last interview, participants were asked if they had any questions or feedback. Once any possible questions had been addressed, the participants were compensated for their time and the experiment was concluded on their part. The concepts they had created were logged (sketches were scanned, foam prototypes photographed and the originals were archived, and CAD files and screen capture files stored onto backed-up hard drives).

3.2.8 Processing data from design experiment

In order to prepare the collected data for later analysis and remove bias caused by the varying presentation methods, all of the concepts were re-drawn and annotated based on their explanations by a professional industrial designer – something the participants were unaware of. Annotations were added to give a comparable level of operational detail for all of the concepts, thereby making it possible to comprehend how they worked, so that a panel of judges could evaluate them without access to the interviews.

As far as the participants knew, their concepts would be presented to the jury 'as is', with their added explanations that they gave during the interviews. Participants were not made aware of the fact that their concepts would be re-drawn, as this could have potentially influenced how they created the concepts, and specifically the fidelity of the concepts. In other words, if a participant would have known that their concepts would be re-drawn, they may have in many cases spent significantly less time on the final appearance of their concepts.

The eighteen approved participants produced a total of 83 concepts. A standard remote control was also added to the dataset as a baseline comparison. The standard remote was the 'best-seller' at the time when searching for 'remote control' on Amazon.com.

Annotations for concepts produced in the prototype and CAD groups were added based on the interviews with the designers themselves, during which they explained the different functions and parts of the remote control concepts they had created. Participants in the sketching group tended to annotate their concepts, in which case their own annotations were generally used. However, there are some exceptions to this as well – some sketching participants used annotations very sparingly, or not at all, and relied on explaining the parts and functionality of their concepts during the interviews, in which case annotations were added in a similar fashion to those in the prototyping and CAD groups.

In cases where it was unclear what the specific annotations should be to make the information content comparable to the other concepts in the study, the decision was made by a panel of two university design professors and two PhD students. The re-drawn concepts were then used in an online survey to gather feedback on the concepts, discussed further in the following chapter.

3.3 User preference survey

The author recognises that design is a complex process, and that design outcomes are hard, if not impossible, to objectively evaluate. Design artefacts and their inherent value are sometimes vigorously debated. Imagine for instance a debate centring on the differences between an Android telephone and an Apple one. There are specific metrics that one can compare – battery life, weight, price – but which one is better? That all depends on who is asking, and what they value. A design can win an award at a design competition, which surely means something, but what?

3.3.1 What is good design?

Although there may not be a clear-cut answer, design evaluations are routinely conducted, and essential for making comparisons between design processes. In order to evaluate the effectiveness of a treatment, one must have metrics to measure the difference in outcomes. Although end-users may find it hard to specify exactly what they value – some factors may even be subconscious, some may argue that a good design is one that sells well. However, there are often several other considerations, which may be more important in determining whether an artefact is well designed.

User satisfaction, environmental impact, durability, reparability, user experience – to name a few – are all important factors in determining the design value of an artefact, but are not necessarily correlated with sales. Many non-design related factors such as existing market share of the company producing the product, advertising, government regulations and so on, can affect sales of a product.

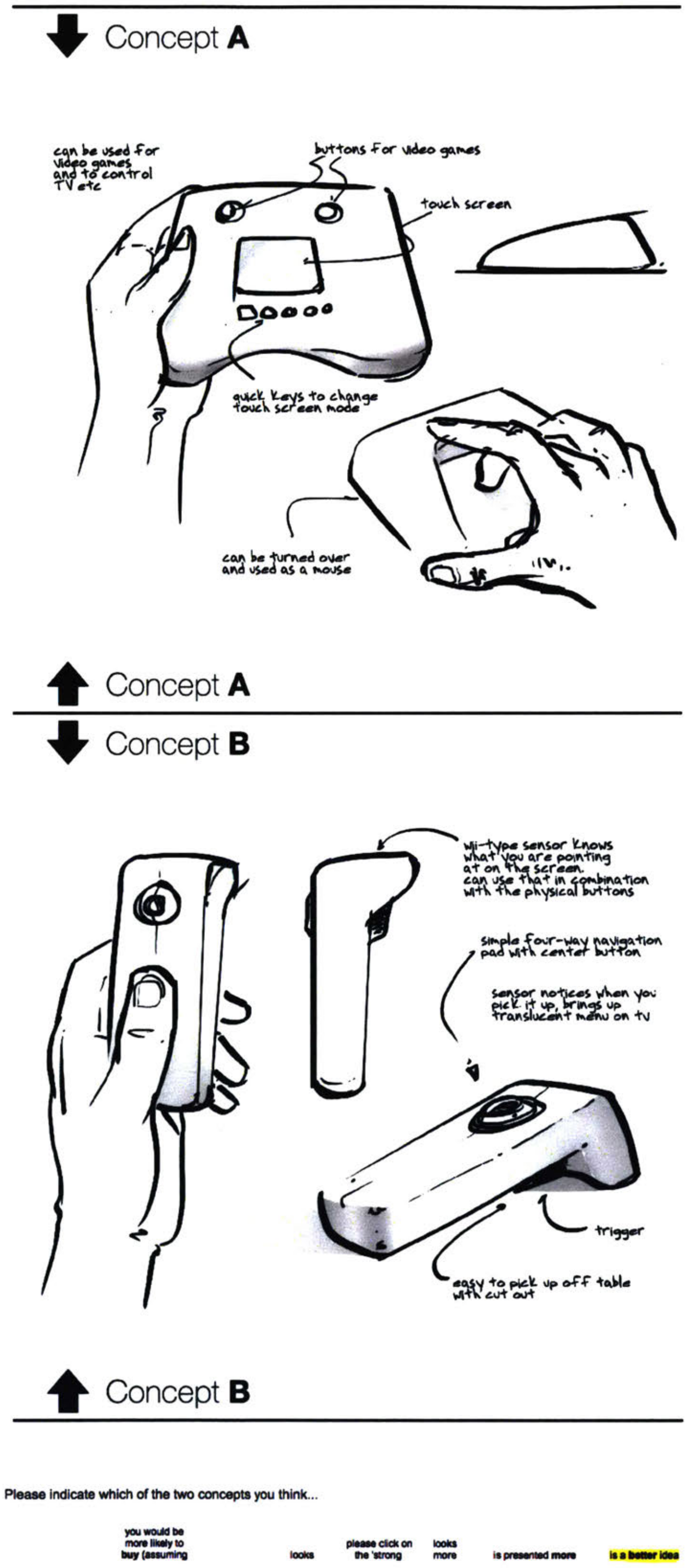
In many cases expert panels rate design outcomes, but often inter-rater reliability is low. The design artefacts created during the experiments were very early stage concepts with a high level of ambiguity, and none of the metrics were easy to evaluate in absolute terms. The approach adopted for this study centred around the notion that a large pool of reviewers would give a more reliable metric for success as the effect of a single reviewer decreased – in essence delivering an 'average' rating for each concept. This will be discussed in more detail later on.

3.3.2 Survey design of pairwise comparison

Ideally, the respondents reviewing the concepts would rank all 83 created concepts +1 baseline concept, but unfortunately that is infeasible due to significant amount of time it would take, as well as being a massive cognitive burden (imagine ranking 84 of your music albums into seven different rank ordered lists – it is highly likely that at several points in this exercise you would be severely challenged to decide which of two albums is the better one).

In lieu of ranking all 83+1 concepts, it was decided that the reviewers would attempt to rate a sub-set of concepts instead. In other words, give scores for concepts on the seven metrics, instead of trying to decide which concept is the best of all of the concepts, which is the second best, and so on. However, this presents a different type of challenge in that it is hard to judge a single concept on an absolute scale – once the reviewer sees more of the concepts, they get a better understanding for the general quality level of the concepts, and may want to adjust earlier scores, or may simply change their rating criteria as they go along, leaving concepts that are presented later on to receive higher or lower scores than they would have were the concept presented earlier on to the reviewer. When rating multiple concepts, it is often the case that if two concepts are later compared head-to-head, and the reviewer is asked to pick their favourite, they may choose a concept that, based on the scores they gave earlier, should not have won.

In order to gather good quality data whilst taking into account the challenges reviewers face rating multiple items on several different metrics, it was concluded that a pairwise comparison with a Likert scale would be the best option. Randomly selected pairs of concepts would be presented to the reviewers, who then had the task of choosing which concept was better, while also providing them a means to indicate on a subjective scale, ***how much better*** the chosen concept was. An illustration of the pairwise comparison can be seen in Figure 6.



Concept **A**

can be used for video games and to control TV etc
buttons for video games
touch screen
quick keys to change touch screen mode
can be turned over and used as a mouse

Concept **A**

Concept **B**

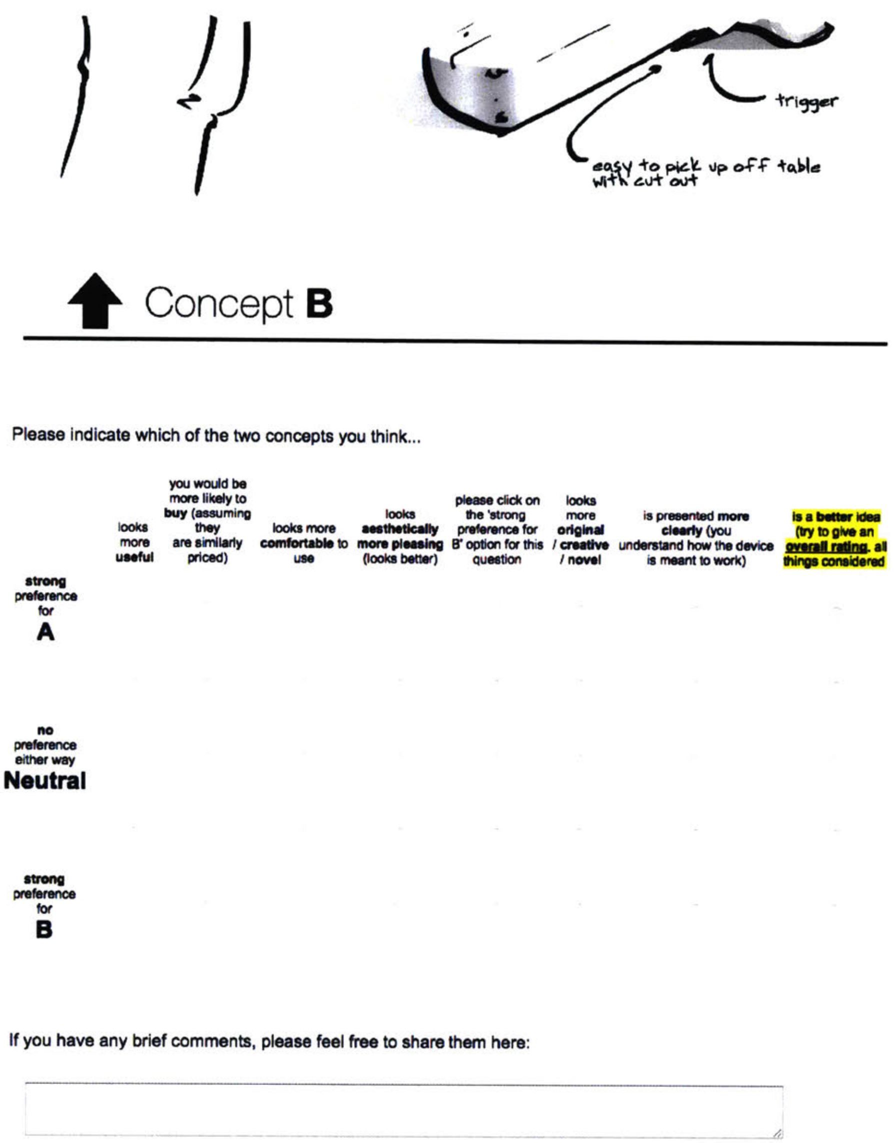
wii-type sensor knows what you are pointing at on the screen.
can use that in combination with the physical buttons
simple four-way navigation pad with center button
sensor notices when you pick it up, brings up translucent menu on tv
trigger
easy to pick up off table with cut out

Concept **B**

Please indicate which of the two concepts you think...

you would be more likely to buy (assuming	looks	please click on the 'strong	looks more	is presented more	is a better idea

Figure 6 – Pairwise comparison page of online survey, top section



Concept **B**

Please indicate which of the two concepts you think...

	looks more **useful**	you would be more likely to **buy** (assuming they are similarly priced)	looks more **comfortable** to use	looks **aesthetically more pleasing** (looks better)	please click on the 'strong preference for B' option for this question	looks more **original / creative / novel**	is presented **more clearly** (you understand how the device is meant to work)	**is a better idea** (try to give an **overall rating**, all things considered
strong preference for **A**								
no preference either way **Neutral**								
strong preference for **B**								

If you have any brief comments, please feel free to share them here:

Figure 7 – Pairwise comparison page of online survey, bottom section

The reviewers were presented with a randomly picked pair of concepts, one above the other. Although there was some concern that placing one image above the other may sub-consciously indicate a rank order (the better one being on top), there was no guarantee that a left-right ordering would not have done the same (as text is read left-to-right in western countries). Furthermore, as the images were in most cases – depending on the size and resolution of the screen of the person viewing the concepts – too large to be displayed at the same time, quick, easy scrolling between images was seen as a priority, hence favouring the top-bottom arrangement over the left-right arrangement.

Most computer mice have a way to scroll easily in the up-down direction, but fewer facilitate easy left-right scrolling, and therefore, the images were placed one above the other to make it as easy as possible to alternate between images while comparing them. However, the images were consciously labelled A and B, instead of 1 and 2, in order to try to avoid any reference to one being better than the other.

The respondents only needed to consider two concepts at a time, and decide which one was 'better', on seven different design metrics, presented in Table 4. A first set of reviewers were presented with six pairs of concepts, but based on feedback from the reviewers (a question was placed at the end of the survey to ask about the length of the survey), the length of the survey was extended to eight concept pairs.

The reviewers were initially tasked with comparing concept A to concept B using a 7-point Likert scale, but as analysis showed little difference to responses received from a 5-point Likert scale, a 5-point scale was utilised to reduce the mental burden on the respondents and make it quicker to fill in the form.

Table 4 – Evaluation metrics used in the pairwise comparisons of concepts

looks more **comfortable to use**	Although genuine comfort is, if not impossible, then at least exceedingly difficult to judge based simply on an image, the question was included as a proxy to gauge how organic the shapes of the concepts were.
looks more **useful**	In accordance with US patent law, inventions need to be novel, non-obvious and useful to be patentable. To that end, this question attempts to address the usefulness aspect.
you would be more likely to **buy** (assuming they are similarly priced)	This question offers another angle – a check of sorts – on which concept reviewers thought was the best (in addition to the last question). It was assumed that reviewers would buy the concept they thought was the best.
looks more **original / creative / novel**	During the idea generation process designers aim to create a variety of novel concepts, or new ways of combining existing ideas. This question was included to give insight into the novelty of a particular concept.
looks **aesthetically** more **pleasing** (looks better)	It was proposed that the design tool would impact the shape of the created concepts, and it was therefore hypothesised that different tools may create visually more appealing concepts than others.
is presented **more clearly** (you understand how the device is meant to work)	Research shows that the clarity of a concept, or the presentation of said concept, has an impact on how the concept is evaluated. It is therefore necessary to know whether the reviewers understood the concept.
is a **better** idea (try to give an **overall rating**, all things considered)	Reviewers were given the opportunity to – based on their own metrics and weightings – give the concepts an 'overall score'. In other words, all things considered, how 'good' a certain concept is.

The rating criteria were created loosely based on Garvin's (Garvin 1984) eight dimensions of product quality: performance, features, reliability, conformance to existing product standards, durability, serviceability, aesthetics, and perceived quality. In formulating attributes for this survey, an important consideration was whether a respondent could reasonably be expected to make meaningful judgments based on two-dimensional line drawings of the concepts viewed on a computer screen.

Reliability, conformance, durability and serviceability were deemed to either require expert knowledge that the reviewers would not have, or were attributes that could not be judged simply by looking at an image. Furthermore, these four dimensions were not the core focus of this research, and thereby were not included in the questionnaire. The study then focused on performance, features, and aesthetics. Performance was expressed as 'usefulness' and 'comfort during use'.

One of the challenges with collecting anonymous human subjects' data is whether or not you can trust the data that you collect. Based on previous experience, discussions with fellow researchers, and anecdotal and unpublished research found on websites, it was expected that without some precautions, a large percentage of the data collected would be unusable. Several quality control methods were developed based on similar quality checks found online in discussion forums, in un-published research, and informed by past experience.

In order to increase the confidence in the collected data, numerous quality control checks were implemented. Failing a single quality control question was not used to discredit a data entry, but rather for every quality control question that the respondent failed, a 'red flag' was marked. If a respondent only received one or two red flags, the respondents' data was put under closer scrutiny. If on closer inspection, however, that data seemed thoughtfully entered it was included in the dataset. If the respondent received multiple red flags, or received only a few red flags but on closer inspection it was obvious that the answers had not been thoughtfully provided (such as answering all questions with the same option) then the data was disregarded. In unclear situations, the predilection was to keep the data.

The quality control questions consisted of three different types of checks: a. filters, b. binary questions, and c. more subtle and nuanced questions for which it was not always possible to say if the answer was 'correct' or 'incorrect', but it could be, for example, 'mostly correct'. An example of a filter quality control was that only Mechanical Turk workers with an approval rating of 99% or over were allowed to take the survey. When rating two concepts, there was an additional 8th question (in addition to the seven previously discussed ones) that instructed the reviewer to 'select the strong preference for B option for this question'. It was possible to objectively determine if the reviewer had answered this correctly or not – this is an example of a binary control question. The third kind of controls utilized were more subjective, but gave supporting information regarding the quality of the data. An example of this type of question is that twice throughout the survey, the reviewer was asked to describe in writing what the two previous concepts they had seen looked like (they were unable to go back in the survey at this point).

Although these types of free form answers in theory could be hard to judge (was the reviewers description of the remote control concept 'close enough to being correct?') in practice it was almost always very obvious if the reviewer had thoughtfully considered the concepts. If they had actually spent the time to compare and evaluate the two different concepts, they would be able to describe them in sufficient detail. If a respondent said the remote control looked like a 'magic wand' or a 'teddy bear', it was quite easy to verify that they had indeed spent at least some time considering the concepts, if however, they simply did not answer, or wrote 'it looked like a remote control', it was an indication that they had not thoughtfully considered the concepts before rating them.

Another example of subtle information that was used to filter responses was the time that the reviewer spent on a specific page. In other words, if a reviewer spent 3 seconds on a page (in essence clicking through the answers as quickly as possible) it was clear that they had not reflected on their answers. In edge cases where the answers were given quickly, but they were still in the realm of possibility in terms of giving the concepts sufficient thought, then the other control checks were used to determine the quality of the data.

In other words, if the reviewer was towards the quicker end of the spectrum, but had not failed any of the other control checks, then the data was used. However, if in addition to being on the quicker side they also failed other control checks, then their data was not used. Although in most cases the choice was simple to make, in the rare event that reviewers were suspect but it was not evident whether or not they had filled in the questionnaire with a sufficient level of effort, a panel of two university professors and two PhD students made the decision after discussing the situation. For a full list of the control questions employed in this survey, see Table 5.

Table 5 – Quality control questions

FILTERS	
>99% worker approval rating on Mechanical Turk	As the survey was run through Mechanical Turk, it was specified in the survey that only workers with at least a 99% approval rating would be able to access the survey.
100 completed HIT's on Mechanical Turk	This requirement was instituted in order to ensure that the workers had completed a sufficient number of tasks so that their approval ratings were not based only on a handful or reviews.
BINARY AND MULTIPLE CHOICE QUESTIONS	
How wide does your screen need to be at a minimum? (Question about instructions reviewer had just been presented.)	The minimum screen width was specified to deter people from completing the survey on a smartphone (due to the large images that would be hard to view). They were asked about the resolution to check that they had read the instructions.
What was the user group? (Answer given in instructions on previous page.)	After being presented with the design prompt, the reviewers were asked to recall who the user group was (they could not return to the previous page to check). They were shown five multiple choice options, with one correct answer..
What devices was the remte control intended to control? (Answer given in instructions on previous page.)	Similar to the user group question, this question was used to check that reviewers had read the design prompt and instructions carefully. They were presented with eight multiple choice questions, with four correct answers.
Choose 'strong preference for B' (Reviewers were instructed to choose a specific answer in questionnaire.)	In addition to the seven pairwise comparison questions, an eighth question that instructed the reviewer to pick the 'strong preference for B' option was added to ensure reviewers were reading the metrics they were evaluating, and not just clicking randomly.
PROBING AND OPEN-ENDED QUESTIONS	
Describe the concepts on the previous page.	Participants were asked to describe the concepts they had just evaluated on the previous page succinctly, with less than 400 characters.
Please choose any concepts you saw during this questionnaire.	Participants were presented with fourteen randomly selected concepts and asked to choose the ones they had seen during the questionnaire. If they had carefully considered the concepts they were evaluating, it was assumed they would be able to recall them.

In addition to the control questions listed in Table 5, the time until first 'click', as well as the total time spent per page were recorded, and used as supporting data when determining if answers had been given in a thoughtful manner. In order to encourage participants to read the instructions carefully, the next button was hidden for 10-20 seconds (depending on the amount of text on a specific page), so participants could not move on to the next page before waiting a pre-determined time. The assumption was that if reviewers could not advance to the next page, they would be more likely to read the instructions – the waiting period was sufficiently short that it was unlikely the reviewers would engage in other activities while waiting for the time to run out.

3.3.3 Design practitioner survey

Once the survey layout and questions had been decided – based on short trial studies with design students – reviewers for the study were sought. Reviewers were recruited by sending e-mail advertisements to design firms and to design related e-mailing lists in the United States. Analysis on the responses (see Figure 19 on page 52) revealed that it would be very challenging to collect hundreds of responses from design practitioners, and therefore a copy of the survey was distributed to 'average Americans' through Mechanical Turk. Comparisons between the data collected from design practitioners and the first round of novice respondents revealed that there were no significant differences between these data, and it was therefore decided that a novice population – not design professionals – would be used as reviewers.

3.3.4 Novice population survey

As no significant differences were found between the reviews of practicing designers and Mechanical Turk workers based in the United States, and due to the challenges of collecting large amounts of survey data from design practitioners (with, as previously mentioned, no obvious benefit), it was decided that M-Turk workers would be used to review the concepts from the study. These data would later be used to evaluate the design artefacts and compare the effects of different design tools on the design process.

3.4 Design attributes

In parallel to the evaluation of the designs on M-Turk, a set of attributes was created to describe the generated designs. Two professors and two PhD students in the field of design created the initial list of attributes. Each created a list of possible attributes individually, after which the lists of attributes were consolidated into a single list during a meeting.

Table 6 – Initial list of design attributes

form factor

- standard remote
- smartphone / tablet
- game controller
- mouse
- novelty / other

input

- buttons
- joystick
- scroll wheel
- touchpad / touchscreen
- gestural (ie. waving hands, moving head)
- novelty / other

interaction

- hands
- body
- eyes
- novelty / other

aesthetics / look & feel

- visceral (concerned with sensory input, the way things look, feel, sound)
- behavioural (how things function and usability, the pleasure of use)
- reflective (eg. liking a knife that doesn't cut well, but has been passed down from your parent)

ease of use

- easier than standard remote
- comparable to standard remote
- harder than standard remote

After some discussion, however, the 'aesthetics / look & feel category, as well as the 'ease of use' category, were removed due to the excessive difficulty in judging these attributes based on the re-drawn sketches. The final three categories used were hence ***form factor***, ***input***, and ***interaction***.

Once the list of attributes was agreed upon, two master's students in the field of design also joined the four initial panel members. All six members of the expert panel reviewed all 83 concepts and indicated if a concept possessed a certain attribute. As an example, one of the attributes was whether or not a concept had buttons or not. Although this example seems quite clear and easy to assess, this was not necessarily always true. Furthermore, many of the other attributes were significantly harder to quantify, such as whether or not a remote control has the physical appearance of a 'game controller'. All six panel members reviewed all of the concepts with regards to the fifteen design attributes, on two separate occasions. After having reviewed the concepts once, the panel members took a roughly two-month long break before returning to the concepts to review them again.

Instead of searching for an absolute truth – 'does this concept have a button?' – the data from the attribute questionnaire was used more in terms of probability. In other words, as an example, one could say that concept X was seen to have buttons 50% of the time, and the appearance of a game controller 75% of the time, if in all twelve reviews (six reviewers, on two separate occasions) six out of twelve times the attribute 'buttons' was chosen (and six times it was not), and 'game controller' was chosen nine out of twelve times.

There were two specific reasons for this approach; firstly, in many cases an absolutely objective evaluation of an attribute was not possible, due to ambiguous concepts or the slightly subjective nature of an attribute. Secondly, even if an objective evaluation could be made (as in the case of being objective able to say a certain concept has a button), when reviewing 83 concepts on fifteen different attributes, there is a certain probability that the reviewer will miss details. Therefore, instead of deciding in a binary fashion whether a concept had a certain design attribute or not, fractions, or probabilities, for each attribute were assigned to each concept. These would later be used in the analysis to compare attributes, design tools and design outcomes. The initial set of design attributes can be seen in Table 6, of which three categories – form factor, input, and interaction – were used.

As can be expected, 'form factor' describes the physical shape and appearance of the remote control, and is divided into sub categories based on common items that the form resembles, such as 'mouse' or 'game controller'.

'Input' is concerned with how information in inputted into the remote, and refers to the physical hardware (buttons and sensors, for example) used to transmit information to the remote control – how a user inputs their commands. The traditional input method would be with buttons, but designs also included eye-tracking, brain waves (although specifics of how it would function were not described), touchscreens and joysticks, to name a few.

The third category, 'interaction', describes the way that a user of the remote would convey information about their intent to the remote control. In other words, 'buttons' is an input, whereas 'hands' would be an interaction. Some remote controls could conceivably have more than one interaction method, and therefore, only the 'primary' interaction method was considered. Additionally, most, if not all, remotes required the use of eyes as one would need to look at the remote to pick it up, for example, but such obvious interactions were ignored. In other words, for remotes where 'eyes' was a method of interaction, it would require that the movement and position of the eyes was in fact the interaction method (that is to say, eye-tracking).

The design attributes would later be used when examining the review scores for each concept, to differentiate between scores that may be due to the use of a specific design tool, versus the existence of a specific design attribute. Put another way, were concepts A, B and C judged to be creative because they were all created by sketching, or because they happened to have touchscreens? Some examples of design attributes are given for selected concepts in Figure 8 to Figure 10.

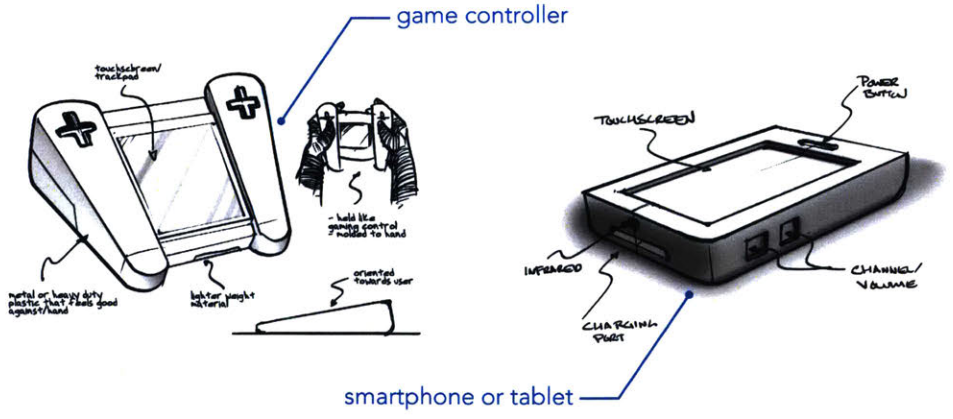


Figure 8 – Design attributes, form factor example

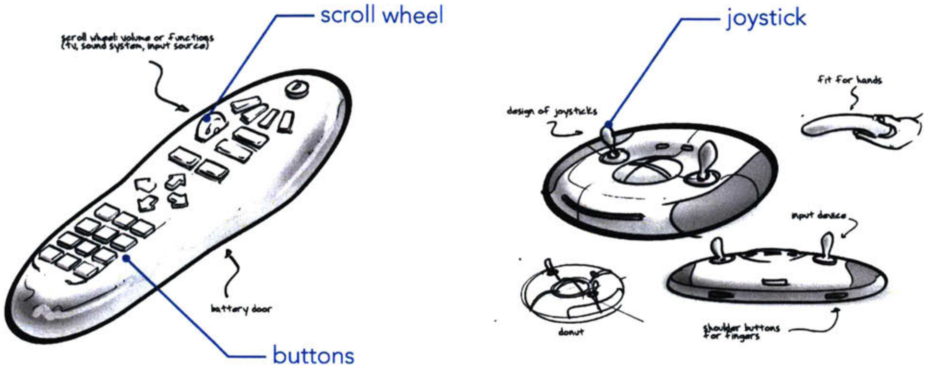

scroll wheel
scroll wheel: volume or functions (tv, sound system, input source)
buttons
battery door
joystick
design of joysticks
fit for hands
input device
donut
shoulder buttons for fingers

Figure 9 – Design attributes, input example

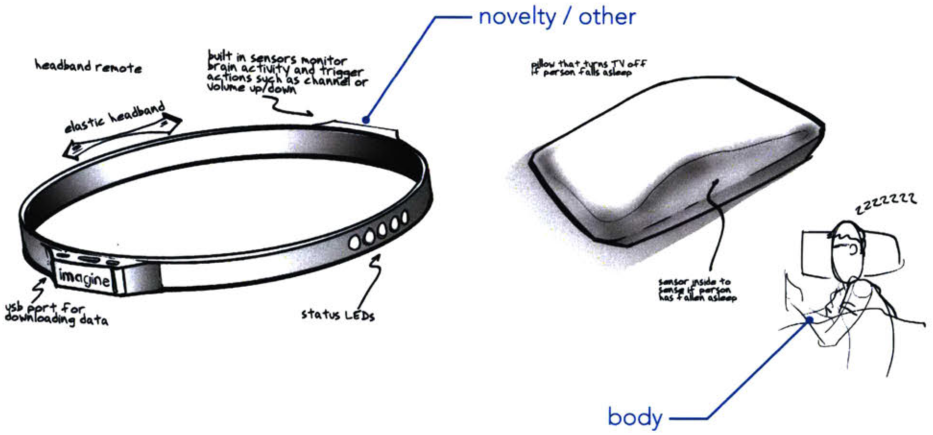

novelty / other
headband remote
built in sensors monitor brain activity and trigger actions such as channel or volume up/down
elastic headband
imagine
usb port for downloading data
status LEDs
pillow that turns TV off if person falls asleep
sensor inside to sense if person has fallen asleep
zzzzzzz
body

Figure 10 – Design attributes, interaction example

4 ANALYSIS & RESULTS

4.1 Introduction

In accordance with the objective of this research to examine the interplay between designers, design tools and the concepts they create in the early stages of the design process, the analysis focuses on what impact, if any, the choice of design tool had on the created concepts. Data for this research was collected in two rounds, the first between November 2011 andJanuary 2013 (part I), and the second round of data collected between August 2015 and January 2016 (part II). In later chapters the source data will be distinguished by referring to part I or part II.

All in all, twenty-one designers participated in part I of the study, with six designers taking part in part II of the study. Three of the designers in part I were omitted from the data analysis due to the fact that they did not follow the design prompt and did not create any concepts.

Results and analysis from experiments in these two parts of the study will be presented next. The results have been arranged thematically, and do not necessarily follow a chronological order nor do they necessarily adhere to the division into parts I and II.

4.2 Who decides what is good design?

Although designers such as Dieter Rams, have offered categories for good design – see (Rams n.d.) for example – making absolute evaluations on the quality of a design is, in some sense of the word, impossible. One need not try harder than to bring up the topic of a new Apple device, a rebranding of the logo for a company, or perhaps the latest Tesla model, to find opposing views on the quality of the designs in question. There will be those on both sides of the argument, but who is right? As design quality is to some degree subjective, it is hard to say. However, to evaluate the effectiveness of different treatments, design processes or methods, some metrics for success need to be defined.

One typical approach to determining design quality is to employ an expert panel of a half-a-dozen experts or less, who then evaluate the design outcomes, thereby determining how good a design is. Inter-rater reliability in these panels may be an issue, and a panel consisting of a different set or reviewers may come to a significantly different conclusion due to natural human variation in preferences and in how different design attributes are evaluated. More importantly however, an expert panel can only rate a limited number of concepts before becoming fatigued – an issue for this study.

One can also argue that design quality is dependent on the user population, in other words, how good a design is also depends on the users' needs. Depending on which group of people the design targets (young children, the elderly with poor eyesight, everyone), what is considered good design may also change. Although an absolute measure of design quality is perhaps unattainable, and despite variation between reviewers, given a sufficiently large panel of reviewers it is assumed an approximation of the design quality for an average user can be uncovered. To this end, the approach employed in this study was to use a large number of reviewers to find a reliable average rating for the design quality of a specific concept.

In order to rank the 83+1 concepts in the study, a pairwise comparison questionnaire was administered, where reviewers viewed two concepts at a time, and indicated on seven design metrics which concept they thought was better – as described in chapter 3.3.2.

4.2.1 Concept reviews – design practitioners

In addition to assuming that a large pool of reviewers would give a more repeatable and reliable evaluation of the concepts at hand, the notion of using novice reviewers in place of expert reviewers was also explored. This was in large part due to the difficulty of finding a sufficient number of expert reviewers with a design background to review the concepts. Initially a survey with six concept pairs was sent out to multiple design related e-mail lists and design companies. 105 participants opened the survey, with 74 beginning the survey, and 72 completing the whole survey. Data from the incomplete responses was used for the concept pairs that had been reviewed. The design practitioners' ages, gender, educational background, and geographical location can be seen in Figure 11 and Figure 12.

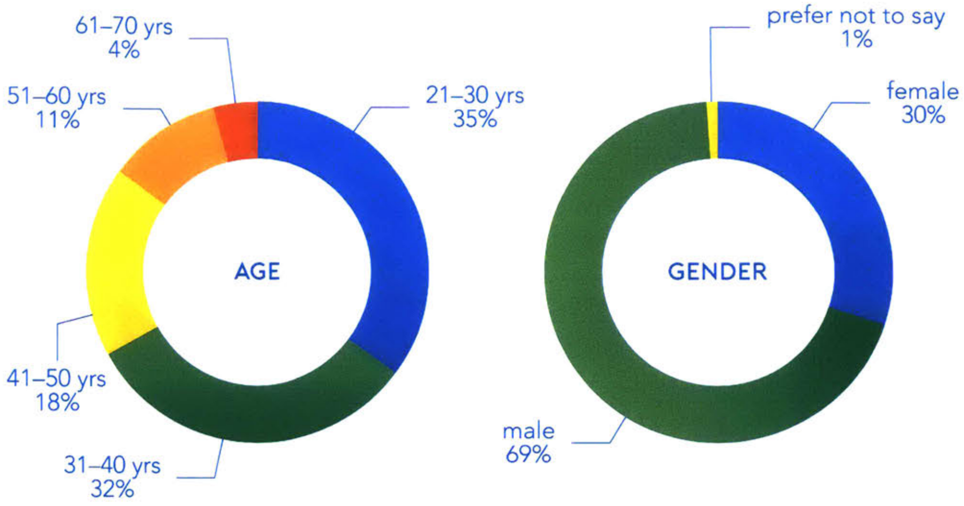


Figure 11 – Practitioners' ages and genders

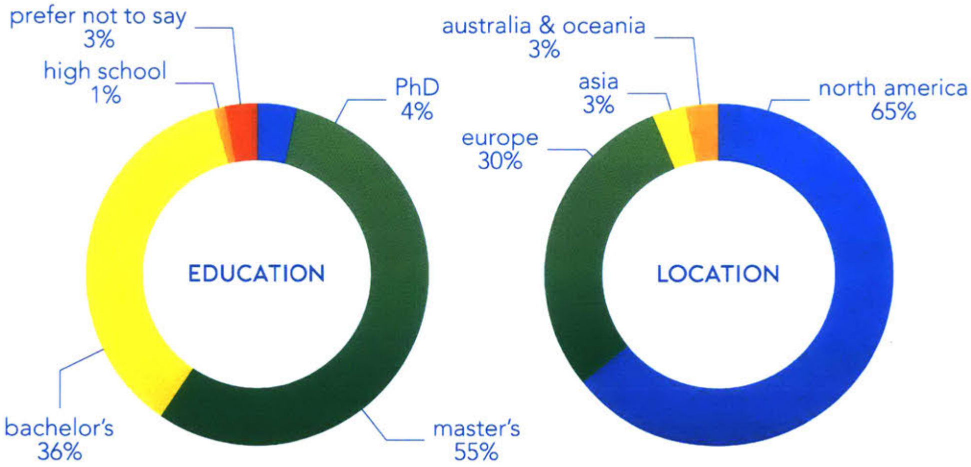


Figure 12 – Practitioners' educational and geographical backgrounds

Additionally, the reviewers were asked about their design background. Allowing for multiple categories to be selected, the most frequently chosen areas were 'product design', 'engineering', and 'industrial design'. All of the options and their respective percentages can be seen in Figure 13.

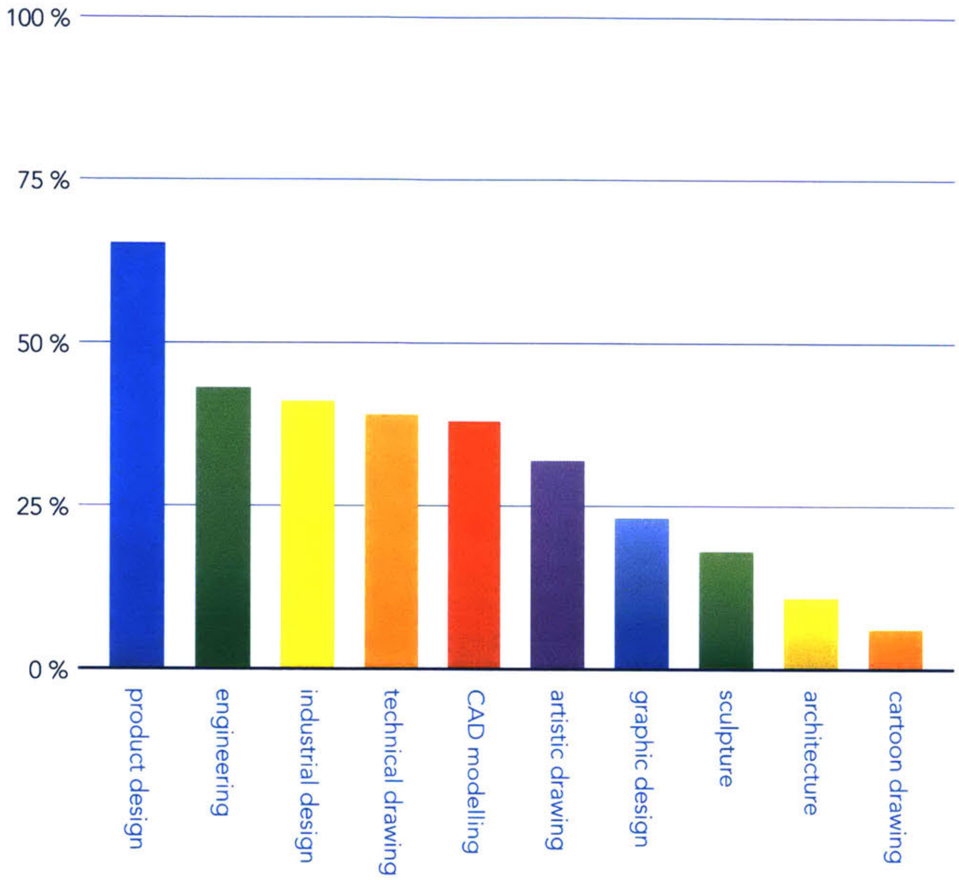


Figure 13 – Practitioners' design background

Finally, the practitioners were asked about their years of work experience (Figure 14). As this question was optional, some respondents opted not to answer. Assuming that the 45 reviewers who did answer the question constitute a representative sample of all respondents, one can see that the reviewers had a significant amount of work experience.

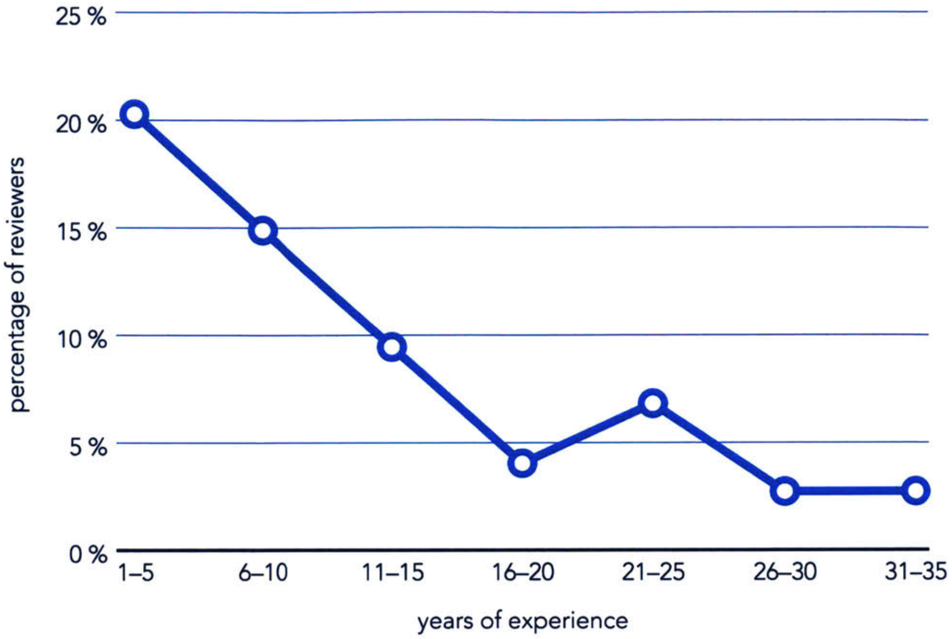


Figure 14 – Practitioners' work experience in a design related field

Several design companies and e-mailing lists were utilised in the initial solicitation for reviewers, and yet, after two-and-a-half months only 72 completed the whole survey – corresponding to a 68.6% response rate. More importantly, however, after the first three weeks the rate of responses decreased dramatically, as can be seen in Figure 15.

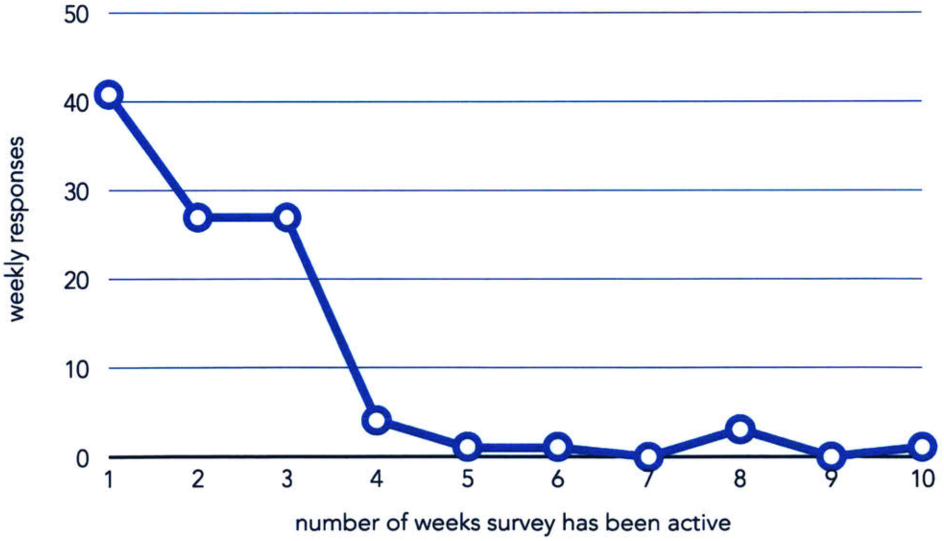


Figure 15 – Weekly responses by design practitioners

It is immediately clear that simply waiting longer would have produced few additional survey responses. Considering the total number of concept views, 857 (or averaging out evenly over all concepts, 10.3 views per concept), it becomes evident that each concept – on average – does not receive many views. Although it is not feasible that each concept is compared multiple times against every other concept, or even once against every other concept ($_{n}C_{r}$ = 3403 possible combinations), it was hoped that each concept would at least get multiple views to make comparisons against other concepts – and ranking the concepts – more accurate.

In order to increase the number of times a concept is viewed and compared, it became clear that utilising the mass reviewer approach for judging design quality would not be feasible solely using design practitioners. Therefore, the same questionnaire was distributed to a novice population through M-Turk, which will be described in the following chapter.

4.2.2 Concept reviews – novice population

The novice population of reviewers was recruited through Amazon Mechanical Turk and had a similar age distribution as the design practitioners used in the earlier survey, although a much more balanced gender distribution (compare Figure 11 and Figure 16).

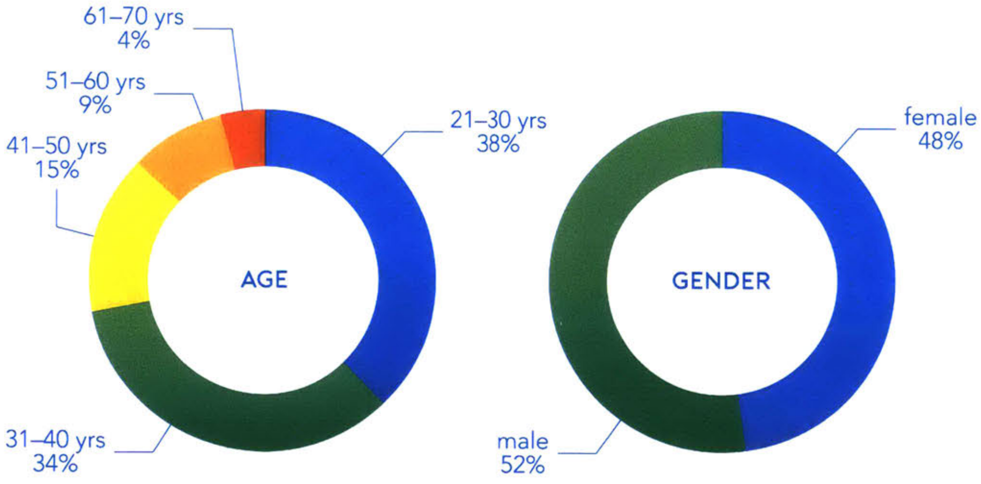


Figure 16 – Novice reviewers' ages and genders

Unsurprisingly, the sample of novice reviewers had a clearly lower level of academic education (compare Figure 12 and Figure 17), and more specifically, almost no design related experience (see Figure 18).

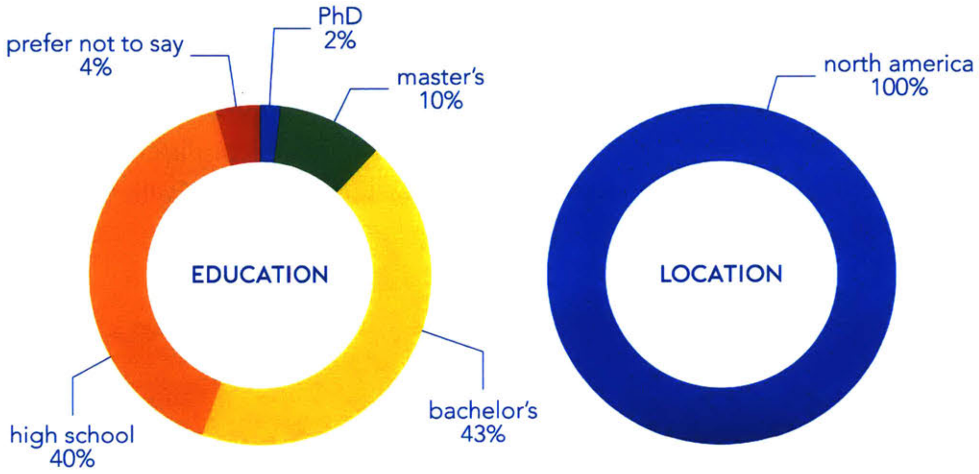


Figure 17 – Novice reviewers' educational and geographical backgrounds

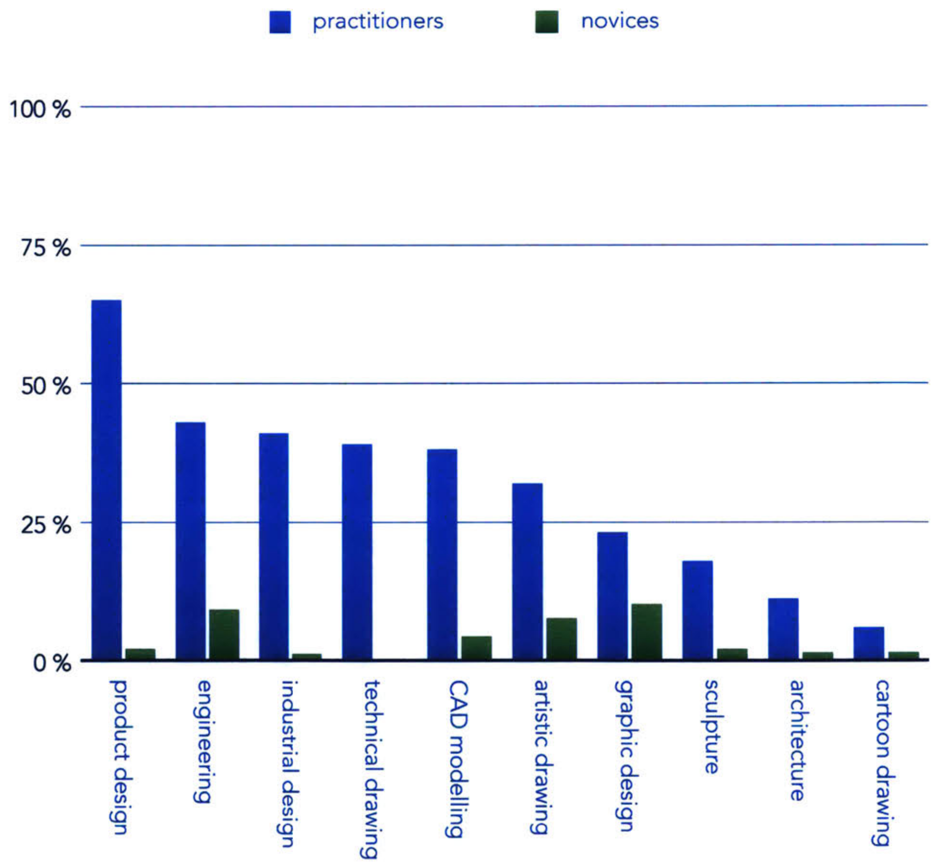


Figure 18 – Practitioner and novice reviewers' design background

Since analysis showed that there was not a significant difference in rater opinions between the population of design practitioners and novice reviewers – and due to the superior availability of novice reviewers – a novice population was used for subsequent analyses of the concepts.

Figure 19 shows the clear difference in response rates between the responses from design practitioners (shown in red), and the four different surveys distributed to novice reviewers through Amazon (shown in different shades of blue and purple), giving further grounds for the decision to collect reviews from a novice population through M-Turk.

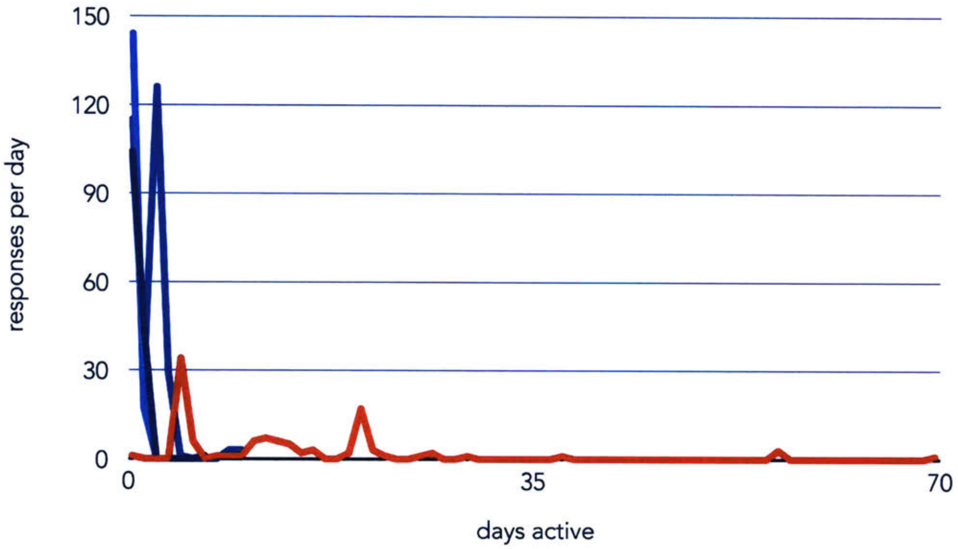


Figure 19 – Daily response rate for concept review questionnaires

Therefore, instead of using a small panel of expert reviewers, several hundred online reviews of the concepts created during part I of the controlled experiment were collected from novice reviewers. Data collection from the reviewers was done through four separate, but identical, surveys. The numbers of responses over time are shown in Figure 20.

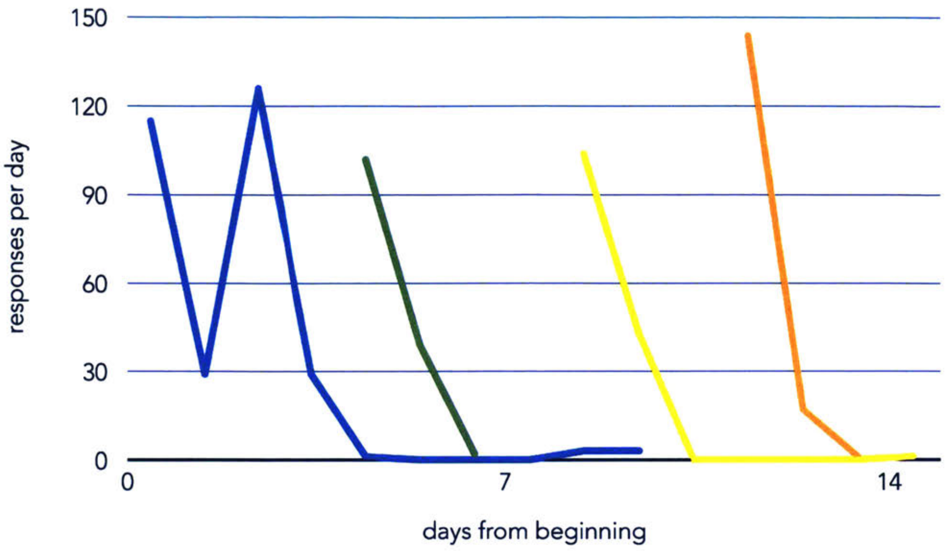


Figure 20 – Response rate for four Amazon surveys of novice reviewers

Apart from the first survey, which for an unknown reason received a second surge of reviews, the rate of responses for the concept reviews declined sharply after the second day – 98.2% of the responses were collected in the first two days of a survey being open, if one discounts the first peak of survey nr.1 (blue line) in Figure 20. This is most likely due to the way surveys are presented on M-Turk, where surveys or tasks posted recently have a higher chance of being seen by potential workers. Due to this realisation, instead of keeping a survey open for a long time, a new identical version of the survey was posted as soon as the response rate of the previous survey dropped – while also making sure that the same M-Turk worker did not re-take the survey.

Due to the rapidly declining response rate, four separate surveys were administered. The initial survey contained six concept pairs. When asked about the survey length, reviewers rated its length as 2.8, with a standard deviation of 0.9 (where 1 corresponded to 'it was very short and quick to fill in' and 5 being 'too long', the assumption being that 3 is the 'perfect length'). Since the average was slightly on the short side, and only 22.4% of respondents rated the survey length as a four or a five (varying degrees of being too long), it was decided that two additional concept pairs could be added to the survey to gather more data in a shorter timeframe. The three remaining surveys were administered with eight concept pairs in each survey. The average evaluation of the length rose only slightly, to 3.0, with a standard deviation of 0.9. Although the completion rate was not adversely affected, the percentage of reviewers who felt the survey was too long rose to 28.8%, and therefore further concept pairs were not added.

The survey was originally designed to take roughly 15 minutes to complete; the first round questionnaire with six concepts pairs took 17 minutes 28 seconds on average, and the eight concept pair questionnaire took 20 minutes 30 seconds on average, when discounting surveys completed by reviewers that were presumed to have taken breaks during the survey.

Whether a reviewer had taken a break during the survey or completed it in one sitting was estimated by first inspecting the duration times of all completed reviews, and then determining a cut-off time above which reviewers were assumed to have completed the survey in different sessions. When calculating the average time to complete a survey, these times were disregarded, and decisions were based solely on the survey duration times for surveys that were assumed to have been completed in one sitting. As can be seen in Figure 21 and Figure 22, a vast majority of surveys were completed in a similar time span, with a few reviewers taking significantly longer to respond, presumably due to leaving the survey for a while, and doing something else.

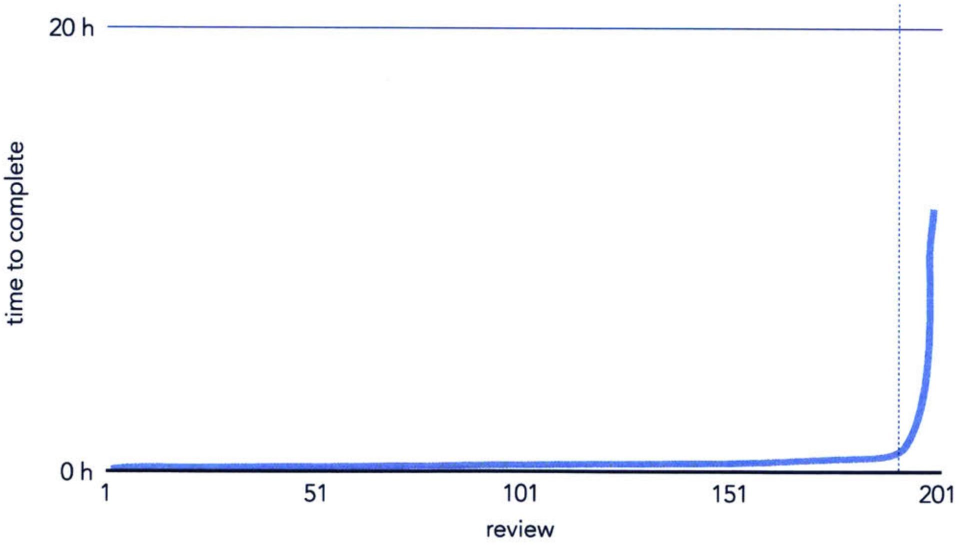


Figure 21 – Total duration to complete survey; six concept pairs

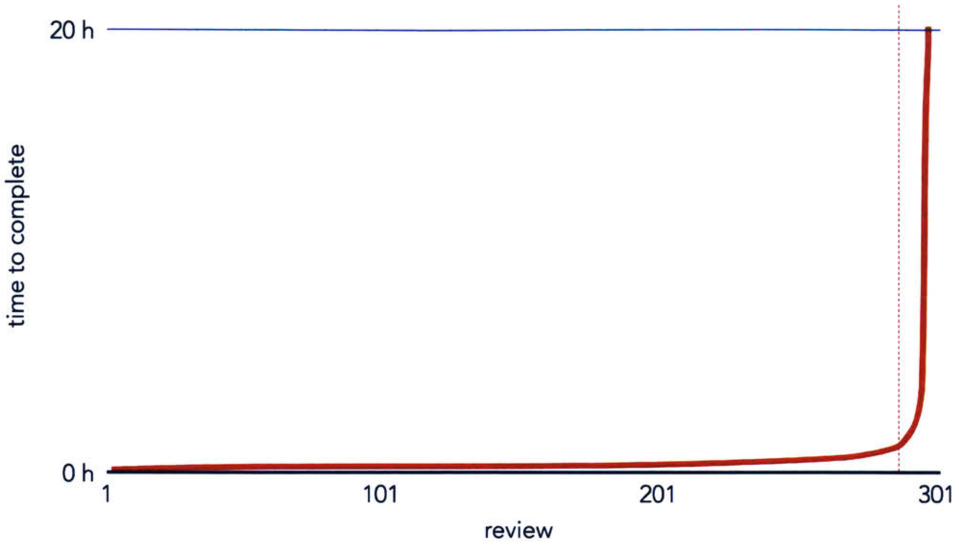


Figure 22 – Total duration to complete survey; eight concept pairs

The cut-off point for values included in the average duration calculations were based on a visual inspection of the graphs above, a calculation of the rate of change and the duration time itself. Although not necessarily the largest rate of change, the cut-off was placed at the first instance of a significant change in duration between reviewers. Figure 23 shows an enlarged area of the graphs above, imposed on each other, while normalising for the different numbers of reviewers.

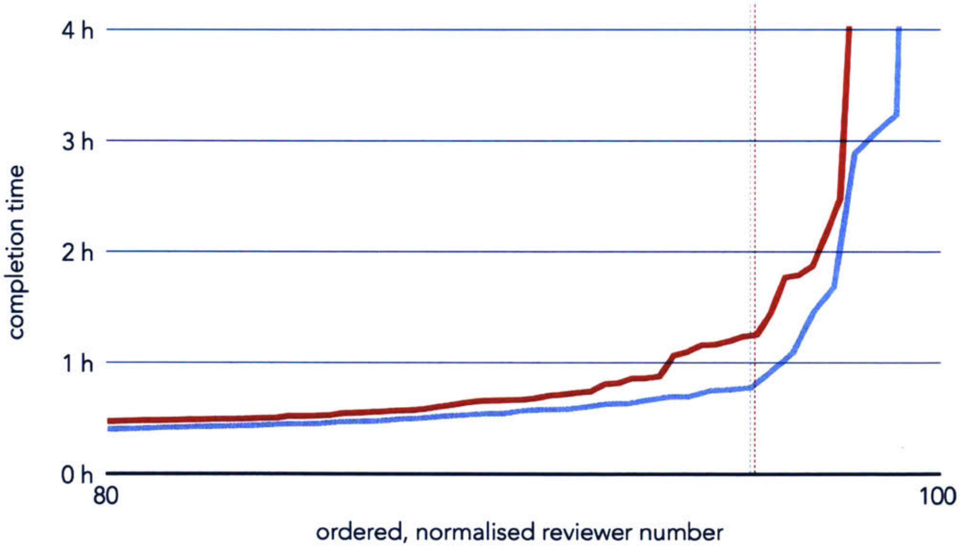


Figure 23 – Durations for six and eight concept pairs; selected area enlarged

As can be seen, the cut-off points were nearly identical for both sets, although an argument could be made for an earlier cut-off for the surveys with eight concept pairs (red line), especially since a total survey time of 1 hour 15 minutes seems somewhat excessive, but in the interest of making conservative assumptions the later cut-off location was kept. The current cut-off points correspond to a 95.8% single-sitting completion rate for both configurations. In other words, adding two concept pairs did not seem to impact the number of people who completed the whole survey in one sitting. However, as the number of reviewers who felt the survey was too long began to approach a third of all respondents, it was decided that the number of concept pairs would not be increased further.

In total, 759 reviewers began the survey, of which 506 completed the whole survey. Based on the quality control questions described in chapter 3.3.2 on page 31, a further 100 responses were rejected due to several failed quality control questions and suspicious data, as shown in Figure 24.

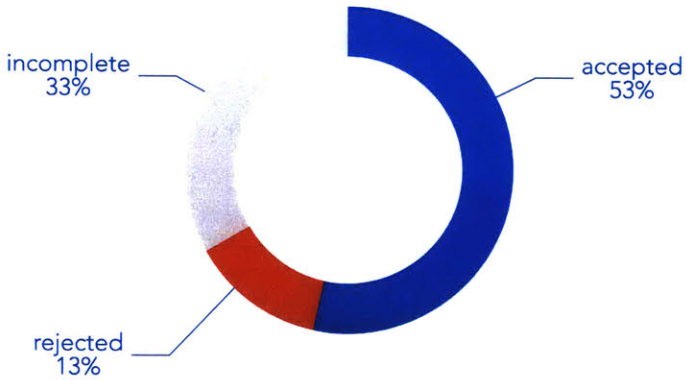


Figure 24 – Percentage of accepted, rejected, and incomplete reviews

Even opening the survey to take a look was counted as a started survey, and categorised as 'incomplete' if the reviewer did not finish it – even if they decided not to answer any questions. Of the reviewers who completed the survey, 19.8% were rejected due to inconsistencies in their answers – in other words, failing too many quality-control questions and having suspicious answers. This was somewhat surprising since an approval rating of 99% on Amazon Mechanical Turk was required of the reviewers in order for them to even be able to see the survey.

4.2.3 Concept reviews – survey statistics

This section will briefly discuss some of the main statistics pertaining to the administered questionnaire and describes how the concepts were displayed to reviewers. As the concepts were shown in random pairs generated by Qualtrics, not all concepts received an equal number of views; concepts were viewed between 58 and 78 times (see Figure 25).

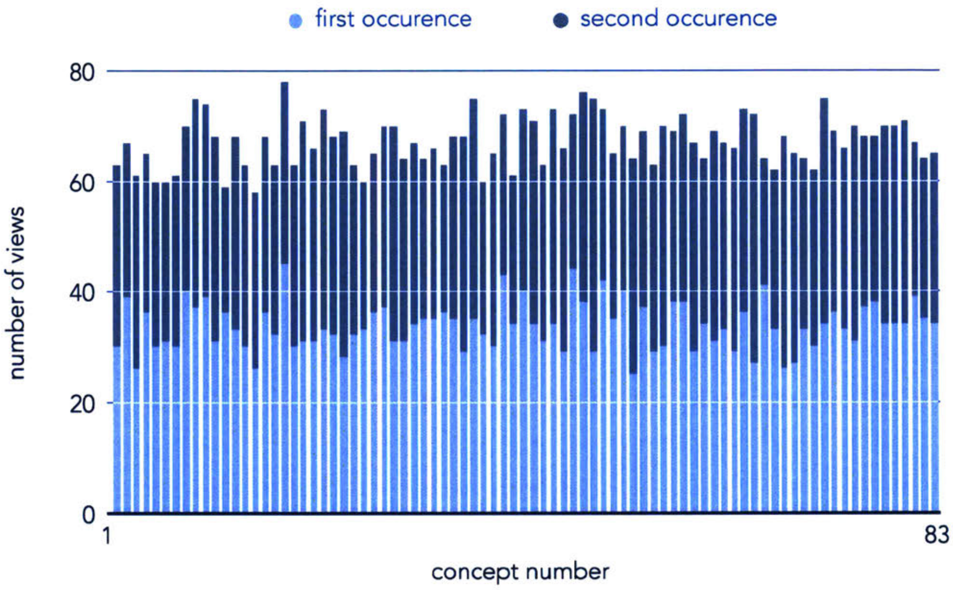


Figure 25 – Frequency of views per concept

Figure 25 above shows how many times each concept was placed as the top (first occurrence) or bottom (second occurrence) concept in the questionnaire. Since the placement of the concept has the potential to create bias in the data, it was important to check that concepts received similar views. On average, concepts received an equal number of first and second views, with a standard deviation of 7.71 for the difference in number of top and bottom views. The distribution of views between the top (first image) and bottom (second image) can be seen in Figure 26 – negative values indicate that the concept was placed as the lower image more often than at the top, and a positive value indicates that the image was placed as the top image more often.

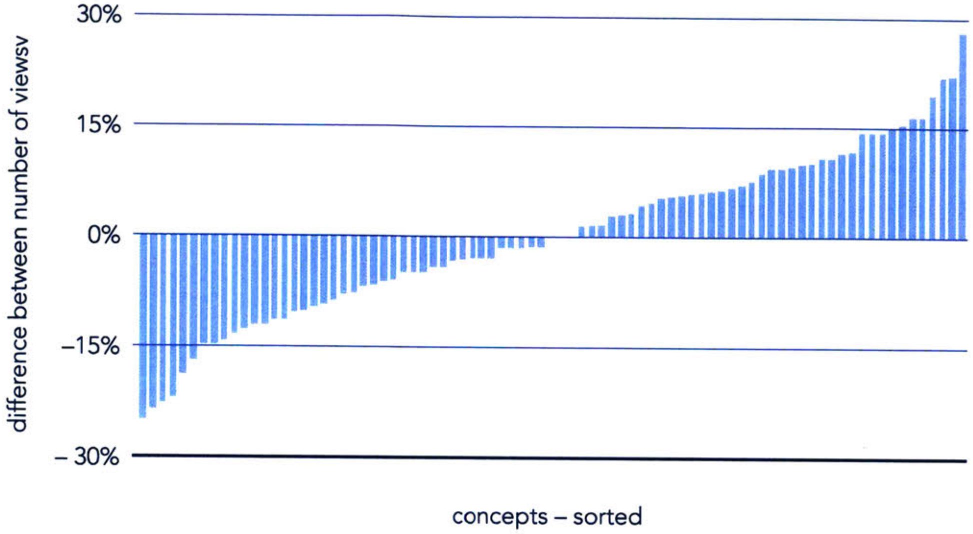


Figure 26 – Difference in number of views between top and bottom concepts

There were also some differences in total number of views – counting both first and second views – between the concepts. Figure 27 below, shows the minimum number of views a concept received (58), as well as the 1st, 2nd (median), and 3rd quartiles, as well as the maximum number of times a concept was viewed (78).

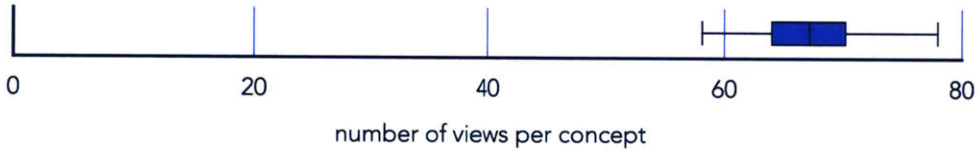


Figure 27 – Whisker diagram for the number of views a concept received

4.3 Which design is best?

As reviewers only saw a subset of twelve or sixteen concepts in a pairwise comparison (out of a total of eighty four concepts), this data needed to be converted into a ranking that could be used to compare how the use of a specific design tool influenced the outcome. The ranking method employed in this thesis will be discussed next.

4.3.1 Colley- and optimised concept rankings

Once the collected concept reviews had been examined and deficient responses had been removed from the dataset, the survey responses from the remaining 406 reviewers were used to rank the 83+1 concepts on all seven created design metrics discussed earlier in chapter 3.3.2. (see Table 4, page 36).

A Colley matrix (Colley 2002) based ranking was used to determine the 'quality' of each concept. The Colley ranking matrix was developed as a bias-free method to rank a large amount of college American football teams in the Bowl Championship Series system, which only play a handful of games (and play no games at all against most opponents). Further complicating the issue is the fact that the schedule is unbalanced; some teams play against 'easy' opponents, while other teams face 'tougher' teams. A team that plays an 'easy' schedule may have more wins, but if they were to switch to a 'tougher' schedule they might have lost more often – something that the Colley method takes into account.

In other words, it was specifically developed for situations where there is a limited amount of comparisons between the items being ranked, with unbalanced comparisons – a similar situation to this study, where a number of concepts are ranked against each other, without every concept pair being compared against every other one. (Boginski, Butenko, and Pardalos 2004; Govan, Langville, and Meyer 2009) The Colley ranking method is increasingly used in academic research, and offers an attractive way to rank the concepts in this study.

With the Colley ranking algorithm, a probability is computed that a certain concept will be ranked higher than another one, given the 'strength' of the concept being compared. As the sample size grows, the comparisons become more balanced. A rank order was calculated for the concepts in all seven design categories. The following formula was used to calculate the final ranking accuracy for the concepts:

Equation 1 – Ranking accuracy

$$\textit{ranking accuracy} = \frac{\textit{count of higher ranked concept winning}}{\textit{total number of comparisons}}$$

That is to say, in what percentage of cases did concept 'A' win concept 'B', assuming concept 'A' was ranked higher in the final rankings? This is further clarified in Figure 28 below. In the three example cases below, concept 'A' has been ranked higher than concept 'B' in the rankings of all the concepts – this ranking takes into account all the other concepts that concept 'A' and 'B' were compared to, and so even though 'A' is ranked higher overall, concept 'B' may have won concept 'A' when compared against each other. In the example case on the left, concept 'A' is always the victor, and hence the accuracy is 100%. In the middle example, one reviewer rated concept 'A' as inferior to concept 'B', and therefore the accuracy is 67% (two reviewers 'correctly' reviewed concept 'A' to be better than concept 'B', although one review placed 'B' above concept 'A' – an 'incorrect' answer when compared to the overall rankings). In the example on the right, none of the comparisons follow the general computed rankings; hence the accuracy for the selected three reviews would be 0%.

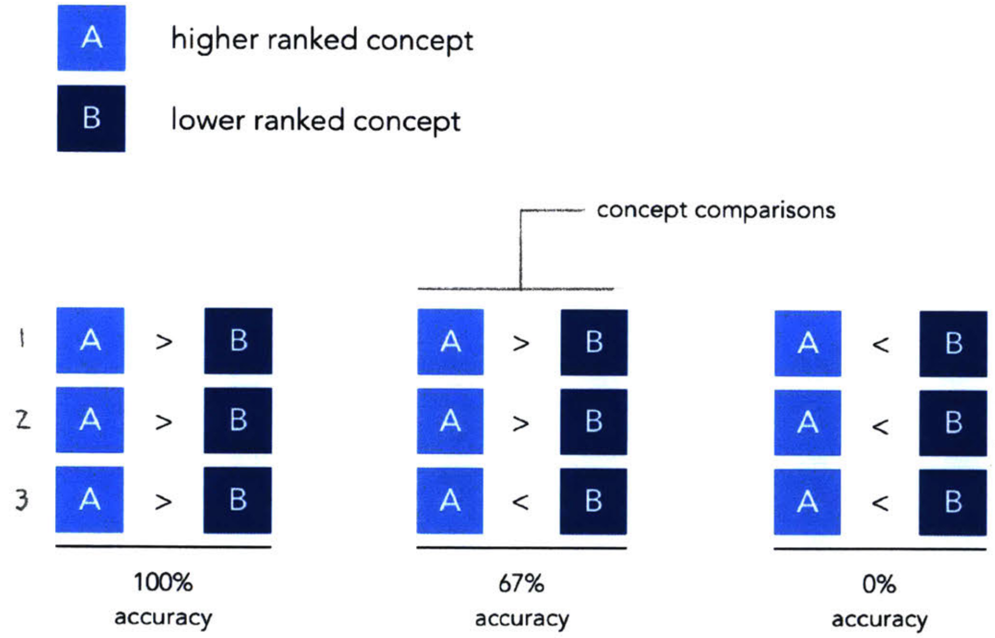


Figure 28 – Ranking accuracy example

If the ranking accuracy were 100%, then in every comparison between two concepts, the one ranked higher in the final rankings would have been chosen as the 'better' one. The rank accuracies are presented in Table 7.

There is also a possibility that no consensus within the population exists, and that certain subsets very consistently rate concepts differently, therefore, the homogeneity of the reviewer population needs to be inspected.

The analysis is based on the assumption that the novice M-Turk reviewers represent a single homogenous population, although it is conceivable that a segment of the population may consistently value other attributes of the design than the remaining population, in which case it does not make sense to try to explain why a certain concept is 'better' than another one (see Figure 29).

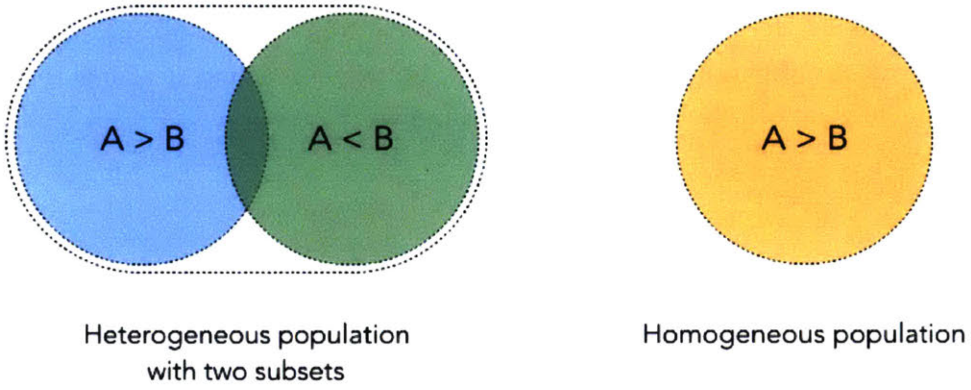


Figure 29 – Heterogeneous vs. homogeneous populations

To check whether the initial assumption of a homogeneous population was valid, a pairwise consistency check was performed. The consistency of a population was defined as a percentage of:

Equation 2 – Pairwise consistency

$$\frac{\sum_{\text{all pairwise comparisons}} max(count(a>b),\ count(b>a))}{count(\text{all pairwise comparisons with multiple reviewers})}$$

The consistency metrics were predominantly above 85%, indicating a single homogenous population with normal random variation, instead of two separate populations with different preferences. After the Colley rankings were computed, a brute force heuristic optimisation technique was used to increase accuracy – the final rank accuracy for the optimised rankings and initial Colley rankings for the seven design categories can be seen in Table 7.

Table 7 – Rank accuracy summary

	useful	creative	comfortable	purchase	looks	clarity	overall
Colley rank weighted accuracy	0.779	0.716	0.760	0.756	0.731	0.764	0.764
optimised rank weighted accuracy	0.817	0.751	0.791	0.793	0.776	0.803	0.795

The optimised rankings were used in all later analysis as the metric of quality for the concepts.

4.3.2 Correlations between design category ranks

Once all of the concepts had been ranked, the Spearman correlations between the different design categories (initially presented in Table 4 on page 36) were calculated, and presented in Figure 30.

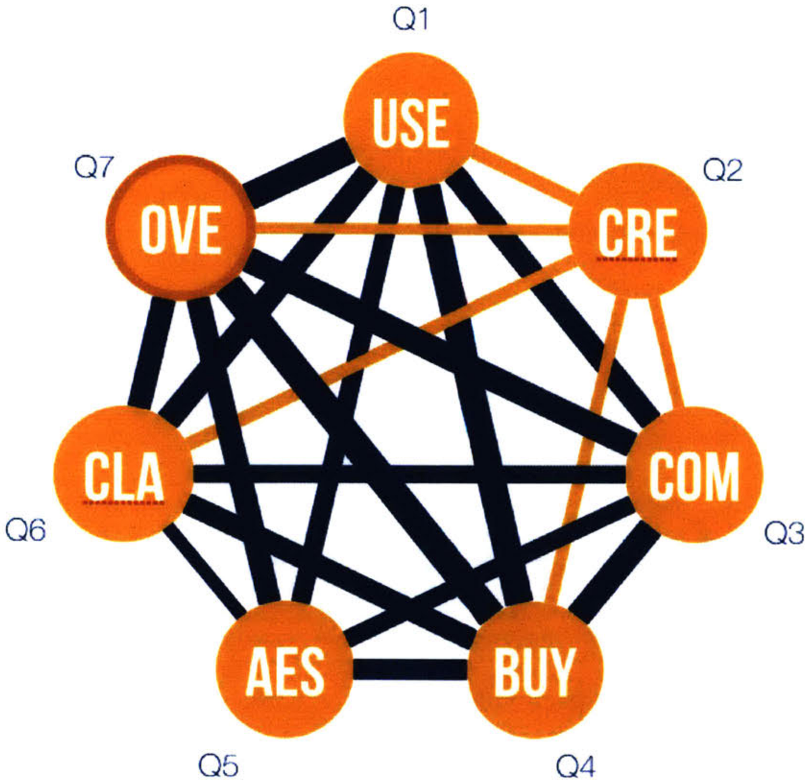


Figure 30 – Correlations between metrics

The thickness of the line indicates the strength of the correlation, with thicker lines indicating a stronger correlation between categories. Dark blue indicates a positive correlation, with orange indicating a negative correlation. A visual inspection of the data reveals that many of the categories are highly correlated, in essence indicating that reviewers did not differentiate between certain categories. In other words, if a concept received a high ranking for 'overall score', it is likely it also faired well in the 'useful', 'comfortable', 'likely to purchase', 'aesthetics' and 'clarity' categories. Creativity was the only category negatively correlated with 'overall score'.

Further clustering analysis revealed that the categories could be condensed into three distinct categories – aesthetics, creativity and overall score. The concept rankings for creativity and overall score were chosen as the main metrics for design quality, as the study focused on creativity and design quality. Creativity is an essential part of the early stages of a design process, while the 'overall score' gave respondents the ability to determine which they thought was the 'better' concept using their own set of metrics. It is important to note that 'overall score' is not calculated based on any other metrics; it is a category in and of itself, similar to the other categories – with the difference that the reviewer decides how to weight different aspects of the design. Another way to think about the 'overall score' category is asking someone 'Which is the best design?'.

Additionally, although the 'overall score' is not calculated based on other categories, it did align well with many of them. A few selected correlations can be seen in the figures on the subsequent pages – the graphs show the correlation between the rankings for overall score and: ...usefulness (Figure 31), ...creativity (Figure 32), ...likelihood of purchase (Figure 33), ...aesthetics (Figure 34), and clarity (Figure 35).

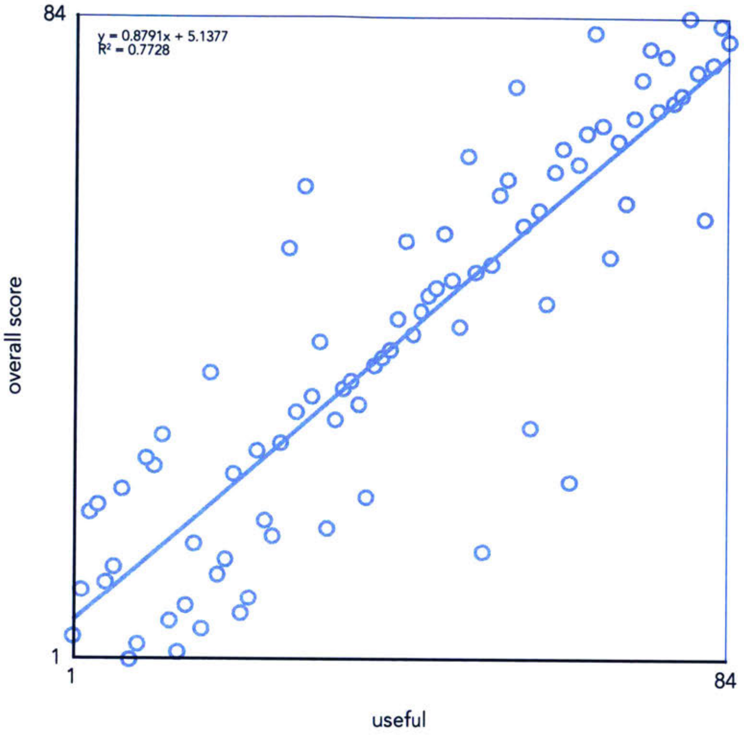


Figure 31 – Ranking for overall score v. usefulness

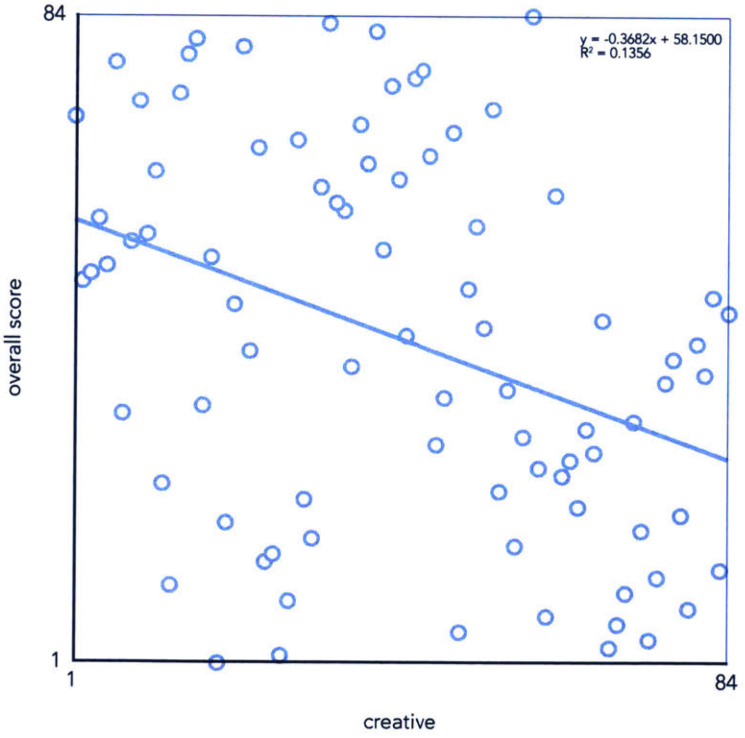


Figure 32 – Ranking for overall score v. creativity

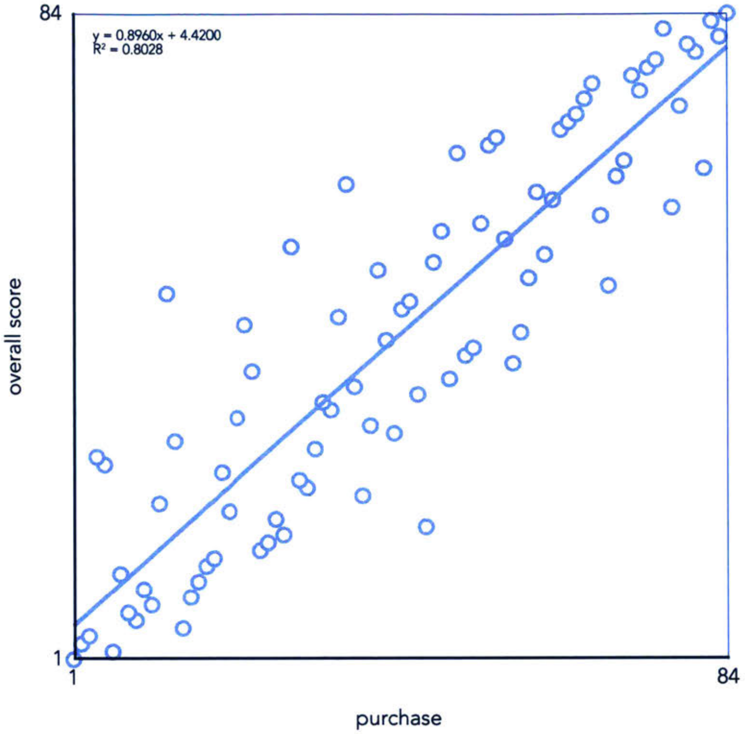


Figure 33 – Ranking for overall score v. likelihood of purchase

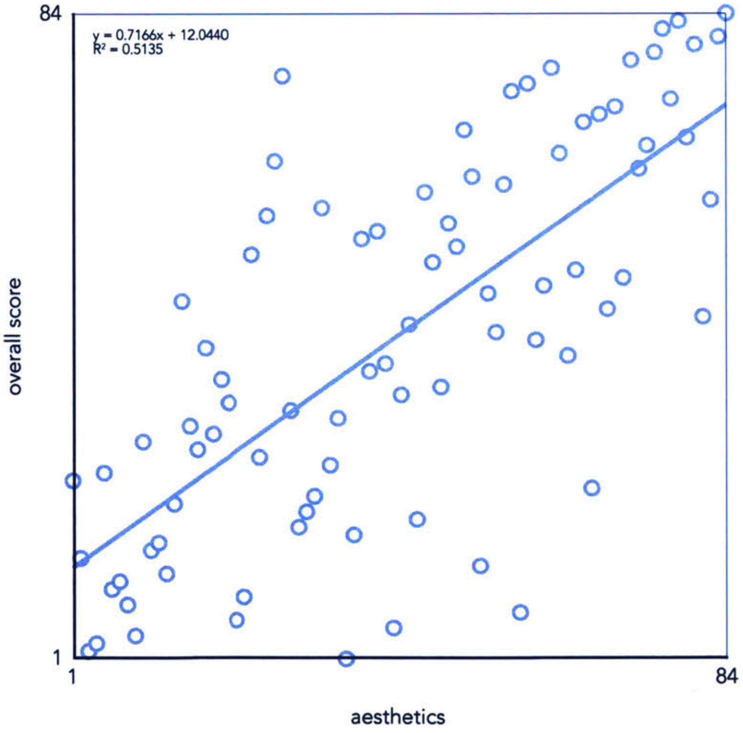


Figure 34 – Ranking for overall score v. aesthetics

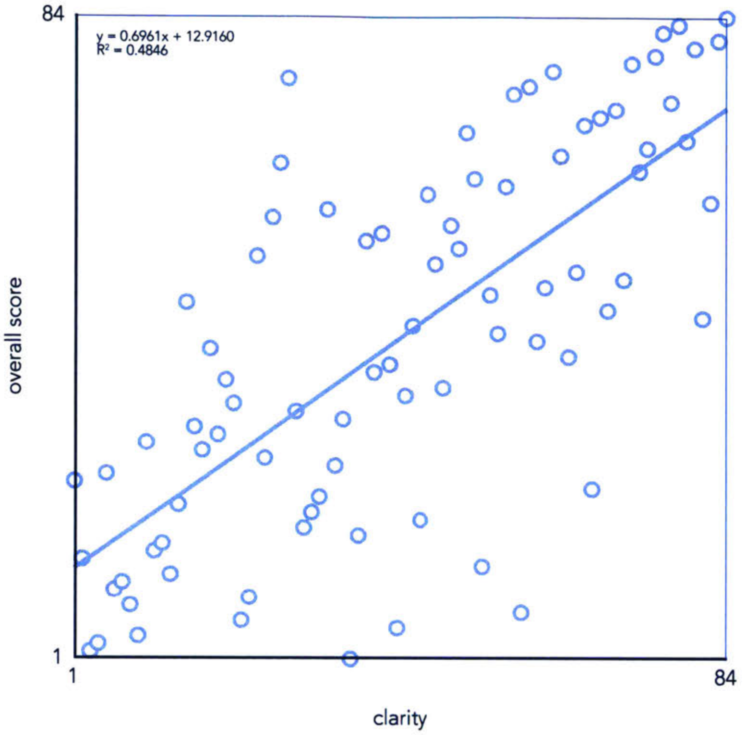


Figure 35 – Ranking for overall score v. clarity

Of the selected graphs that were just presented, the correlation between the overall ranking and likelihood of purchasing was the strongest. This seems like a reasonable result, as one would expect reviewers to want to buy the remote, which they thought was overall the best one. Although less conclusive, remotes that were ranked higher in aesthetics or usefulness were also more likely to be ranked highly overall. Underlining the importance of the role of representations in reviewing designs, in general, concepts that were more easily understood were also ranked higher overall.

4.4 How do designers design?

To gain further insights into how the designers in the study created concepts, the video recordings from part I, and the audio recordings from the talk-aloud protocol in part II of the study were analysed by three researchers.

4.4.1 Video analysis of design process

Exclusively looking at the end products (the concepts) that were created and interviewing the designers gave scarce information on what the designers were doing during the actual design task – between the interviews – and what their design processes were. In order to get more detailed data on what the designers were doing, the video recordings from the experiments in part I were analysed in detail.

In addition to being able to observe what designers were using their time on, analysing the video also provided information about ***when*** certain concepts were created. This could be of interest in several cases, for instance, were more creative concepts created later with less creative concepts created first? Did designers spend more time on concepts that were rated 'better'? or were there perhaps differences in how designers alternated between working on different concepts; did they simply finish one concept, and then move on to the next one?

This poses the question of how to measure ***when*** a concept is created. Although superficially similar, the creation processes behind two different concepts may vary greatly, as illustrated in Figure 36.

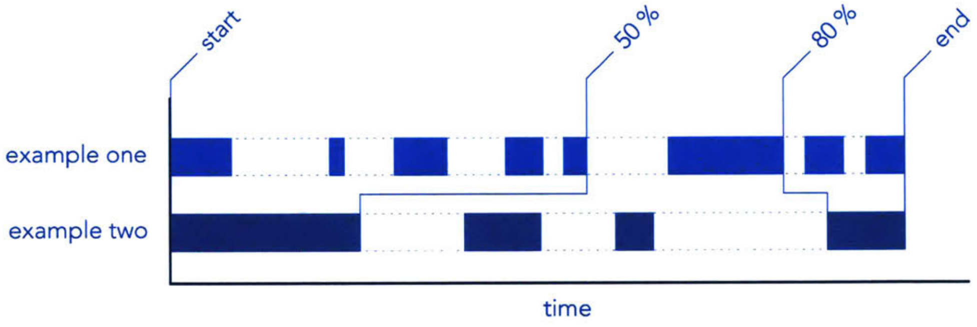


Figure 36 – Two fictitious examples of time-spent working on concepts

As can be seen in this fictitious example, although the beginning and ending times are similar, the bulk of the design is allocated at very different times in the design process. If the bulk of the concept design happened early on in the experiment, for example in the first ten minutes of the experiment, but a minute before the end the designer decides to change a small feature of the concept, was that design completed at the ten minute mark or a minute before the end?

In order to be able to address the issue of when a design is regarded as completed, and to see how it affects later analysis, four different times were evaluated for each concept: the start time, the point at which 50% of the concept was done, the point at which 80% of the concept was done, and the end time. In this context, the phrase '50% of the concept was done' refers to the point at which 50% of the total time spent working on a specific concept had been spent. In other words, it only refers to a percentage of the total time that was used to work on the concept, and is not a judgement of 'how complete' the concept was. The last five minutes of working on a concept may well be the most important, but as there was no reliable way of judging progress other than in terms of time spent working on a concept, percentage of total time was used.

In addition to analysing the time-use in terms of 'was the concept being worked on', a more detailed analysis of the video footage was conducted, to determine not only ***if*** a concept was being worked on, but also ***what*** was being done. To this end, the actions of each designer were categorised into four separate categories: thinking, exploring, doing and evaluating, as shown in Table 8.

Table 8 – Four designer action categories used in video analysis

thinking	any time where the designer sat still in thought, or looked around, but did not use the design tools at hand was categorised as 'thinking'
exploring	time when designer is actively making something – sketching, shaping foam, or making a software model – but whatever they made does not become part of any submitted concept
doing	making the concept, with whatever design tool they were assigned to – the difference between the 'doing' and 'exploring' categories is simply whether or not the artefact that was made became part of a concept or if it was discarded
evaluating	holding and looking at a design artefact the designer made earlier (either a fully formed concept, or one that is mostly complete)

The video footage for each designer was analysed with one-second precision and categorized into one of the four categories above. An example of a segment from one of the experiments can be seen below in Figure 37.

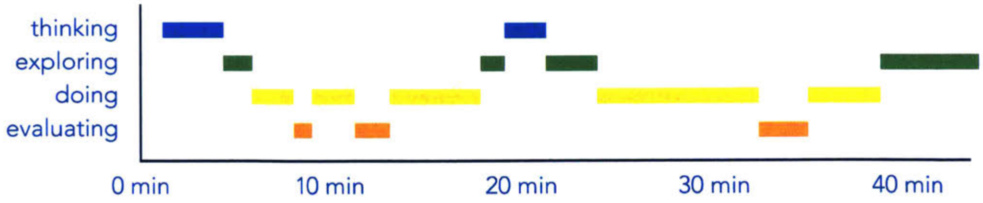


Figure 37 – A segment of the time-use profile for one of the designers

In the interest of readability and space, the time profiles (as shown in Figure 37 were compressed into a one-line representation, as shown in Figure 38 below, with colour-coding to differentiate between work categories.

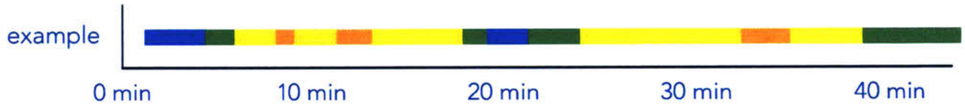


Figure 38 – One-line representation of the concept time-use profile

The time spent working on each concept was categorised into one of the four groups. Initial qualitative and quantitative analysis on the time-use profiles of the designers in part I of the study did not reveal any interesting results. No correlations were found between the time structure and design outcome. The variation in designers' backgrounds and relatively limited sample size could make possible correlations difficult to uncover. However, how designers spend their time during a design process may warrant further exploration with a more targeted study specifically designed to examine designer behaviour – this was however outside the scope of this thesis.

4.4.2 Creating categories for design activity types

Although the video analysis in the previous section revealed what the designers were ***doing***, it gave little insight into what the designers were ***thinking*** during the experiment. In order to gain insight into what designers were thinking during the experiment, a talk-aloud protocol was implemented for the second set of experiments. Of particular interest was whether there were differences between what designers focused on based on the tool at their disposal. For example, did designers working with foam think more about ergonomics due to the fact that they were constantly handling physical models and receiving tactile feedback? Did designers working on the computer consider dimensions and measurements more often, due to the fact that the modelling software required designers to make decisions regarding dimensions in a very different way than designers working with foam or sketching?

As described in the methodology section (3.2 Controlled design experiments), audio recordings from six designers in part II of the experiment were sent out to be transcribed by a professional transcription service.

In order to analyse the results from the talk-aloud design sessions, a coding scheme was needed to make sense of the data. Previous categorizations of design processes were initially studied to see if an existing classification rubric could be found that would be directly applicable for this case. Unfortunately, none of the studied categorisation schemes seemed appropriate, mostly due to their focus on later stages of the design process – on categorising more developed concepts. Therefore, a custom categorisation scheme was devised for this research.

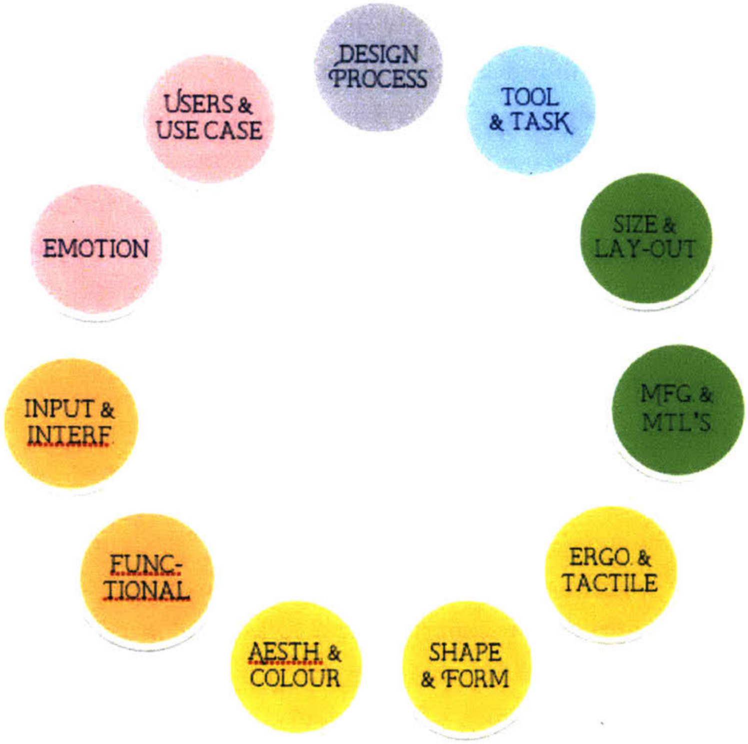
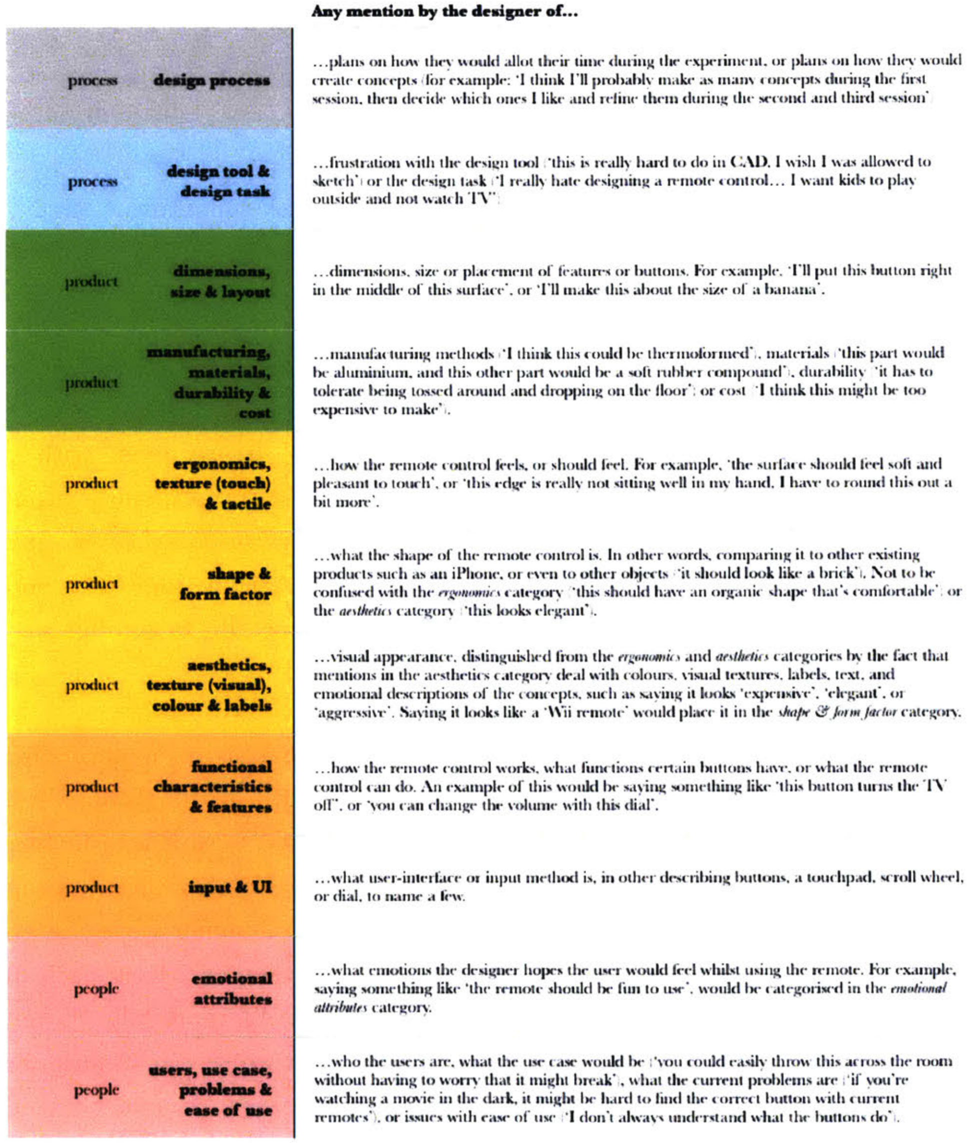


Figure 39 – Talk-aloud protocol verbalisation categories

The categorisation was created in an iterative fashion, by first creating a 'best guess' based on other previously observed categorisations, then testing it on the transcript data from this study, and then revising and modifying it until the categories were well defined, and all the sample segments that were being tested fit clearly into a specific category. The final categories are shown in Figure 39, with more detailed descriptions of the categories in Table 9.

Table 9 – Talk-aloud protocol verbalisation categories

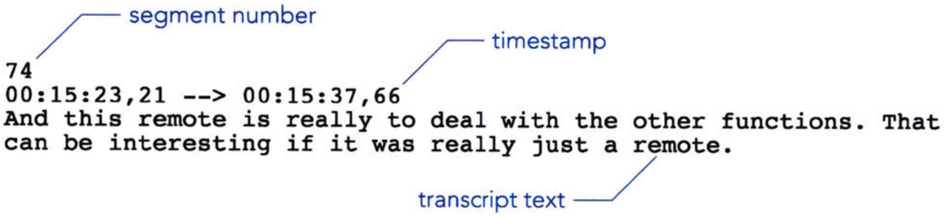


		Any mention by the designer of...
process	**design process**	...plans on how they would allot their time during the experiment, or plans on how they would create concepts (for example: 'I think I'll probably make as many concepts during the first session, then decide which ones I like and refine them during the second and third session')
process	**design tool & design task**	...frustration with the design tool ('this is really hard to do in CAD. I wish I was allowed to sketch') or the design task ('I really hate designing a remote control... I want kids to play outside and not watch TV')
product	**dimensions, size & layout**	...dimensions, size or placement of features or buttons. For example, 'I'll put this button right in the middle of this surface', or 'I'll make this about the size of a banana'.
product	**manufacturing, materials, durability & cost**	...manufacturing methods ('I think this could be thermoformed'), materials ('this part would be aluminium, and this other part would be a soft rubber compound'), durability ('it has to tolerate being tossed around and dropping on the floor'; or cost ('I think this might be too expensive to make').
product	**ergonomics, texture (touch) & tactile**	...how the remote control feels, or should feel. For example, 'the surface should feel soft and pleasant to touch', or 'this edge is really not sitting well in my hand. I have to round this out a bit more'.
product	**shape & form factor**	...what the shape of the remote control is. In other words, comparing it to other existing products such as an iPhone, or even to other objects ('it should look like a brick'). Not to be confused with the *ergonomics* category ('this should have an organic shape that's comfortable') or the *aesthetics* category ('this looks elegant').
product	**aesthetics, texture (visual), colour & labels**	...visual appearance, distinguished from the *ergonomics* and *aesthetics* categories by the fact that mentions in the aesthetics category deal with colours, visual textures, labels, text, and emotional descriptions of the concepts, such as saying it looks 'expensive', 'elegant', or 'aggressive'. Saying it looks like a 'Wii remote' would place it in the *shape & form factor* category.
product	**functional characteristics & features**	...how the remote control works, what functions certain buttons have, or what the remote control can do. An example of this would be saying something like 'this button turns the TV off', or 'you can change the volume with this dial'.
product	**input & UI**	...what user-interface or input method is, in other describing buttons, a touchpad, scroll wheel, or dial, to name a few.
people	**emotional attributes**	...what emotions the designer hopes the user would feel whilst using the remote. For example, saying something like 'the remote should be fun to use', would be categorised in the *emotional attributes* category.
people	**users, use case, problems & ease of use**	...who the users are, what the use case would be ('you could easily throw this across the room without having to worry that it might break'), what the current problems are ('if you're watching a movie in the dark, it might be hard to find the correct button with current remotes'), or issues with ease of use ('I don't always understand what the buttons do').

The verbalisation categorisation scheme described above was used for subsequent analysis of the audio data from the talk-aloud design experiments conducted in part II of the study.

4.4.3 Quality of the transcription data

As described earlier, the audio recordings were sent out to be transcribed, and the resulting data was used to gain insights into what designers were thinking during the design task. The processed transcription data included sentence chunks with associated time codes, as shown in Figure 40 below.



```
74
00:15:23,21 --> 00:15:37,66
And this remote is really to deal with the other functions. That
can be interesting if it was really just a remote.
```

segment number
timestamp
transcript text

Figure 40 – Sample segment of transcribed data

The audio was transcribed into sentence 'chunks' or sections, with a beginning and ending time code. The length of the transcribed sections seemed to follow the structure of complete sentences in some cases, and natural pauses in speech in others, although any specific logic was not verified and it is unclear how the transcribers chose the beginning and ending times for segments.

This transcribed audio data formed the basis for analysis on experiments conducted in part II of the study, and since the length of the sections has significance for the later analysis, a random sample of section lengths was collected from the transcripts, to check how much variability there was in their length and word count. Six sentences from each designer were chosen with a random number generator (Haahr n.d.) – two sentences from each of the three design sessions per person, for a total of 42 randomly chosen sentence samples. Figure 41 shows the speech rate in words per minute, as derived from the time codes in the transcribed passages. As can be seen from the graph, most samples were similar in speech rate, with a few samples that were clearly outliers.

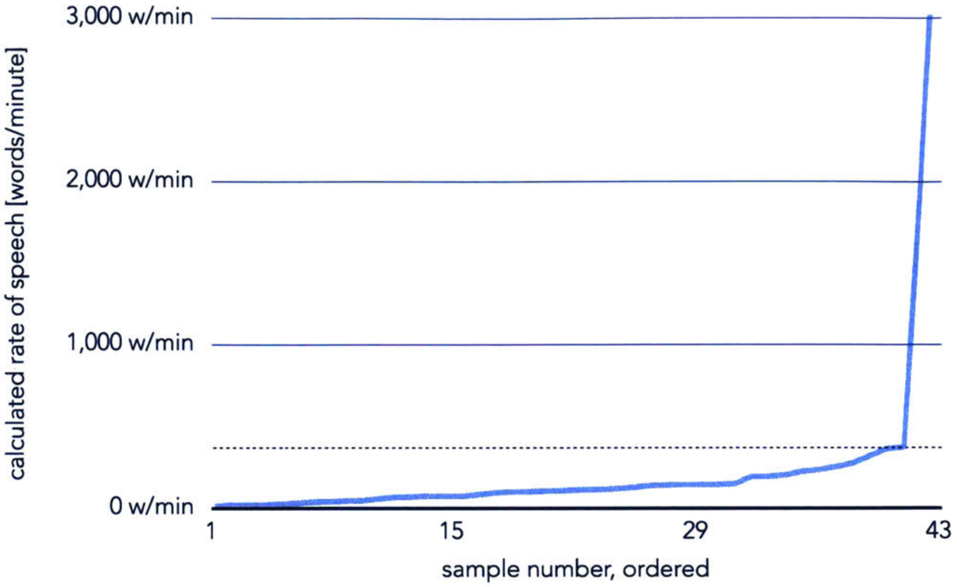


Figure 41 – Speech rate of selected transcription samples – ordered

An enlarged view of the lower range for speech rate can be seen in Figure 42. The orange line denotes values that were clearly outliers (two out of the forty-two samples) with speech rates over six times quicker than what is considered quick conversational speech. The light-blue area shows the range of speech from 'slow clear' speech, to 'quick conversational' speech – a normal conversational rate being around 150–200 words per minute (Krause and Braida 2002; Wong 2015) – with the dotted line denoting the cut-off for data points that were considered outliers.

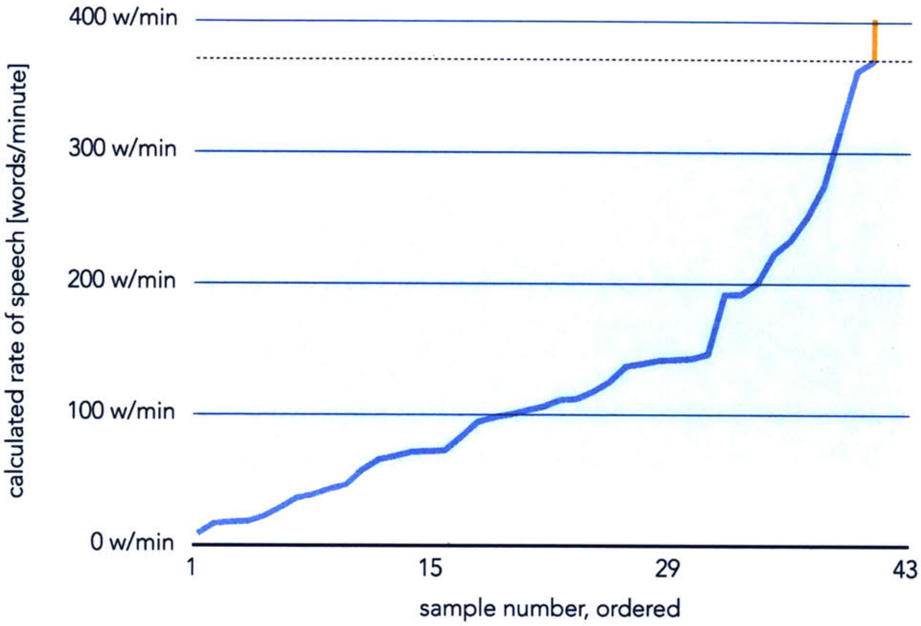


Figure 42 – Speech rate of selected transcription samples for selected range

Many of the samples fall in the range of 'normal' speech rate, although several samples also have a slower speech rate, which is easily explained by the fact that designers took small breaks between words, as they were thinking through their design. In other words, the speech during the design experiment did not follow 'traditional' speech patterns, and a significantly slower speech rate is understandable. However, the two outliers – shown in orange in Figure 43 – clearly had speech rates that were unrealistic. This would translate into 4.5% of the time codes being incorrect, assuming these two instances were the only ones with incorrect time stamps.

However, apart from re-doing the transcriptions by watching the video recordings, or listening to the audio recordings there is no way to verify the actual accuracy of the data. Due to the significant time requirement for either of the two options, the estimated accuracy (based on the sample segments) was deemed sufficient, taking into account the scope of this thesis.

The general speech rate check does give some confidence that most of the time codes translate into reasonable speech rates, and in most cases, the data is at least not obviously incorrect. The light-blue area in Figure 43 denotes the area of normal uninterrupted speech, with the black cross showing the average word count and length of segment (not accounting for the two outliers).

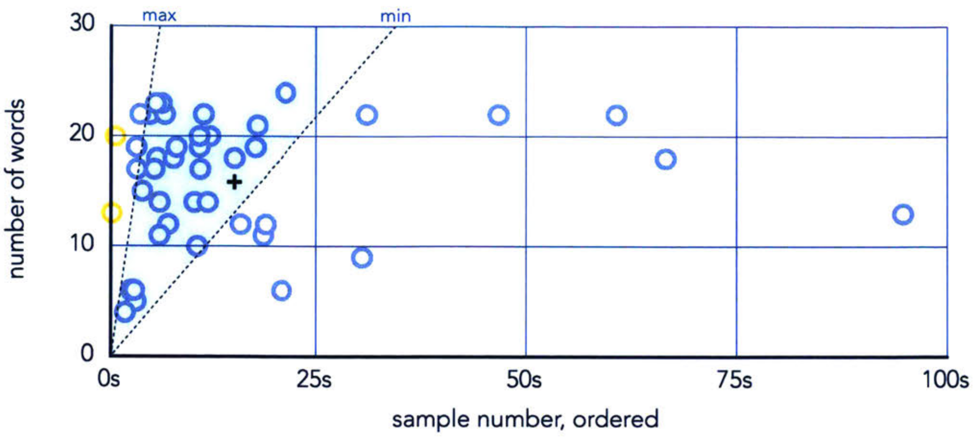


Figure 43 – Most samples had normal speech rates (light-blue wedge).

Additionally, the transcriptions were used with ± 5-second fidelity in later analyses, so small deviations were not presumed to be of any real significance.

4.4.4 Designer verbalisation during design task

Once the verbalisation categories (Figure 39) had been created and the audio recordings from the talk-aloud protocol transcribed, each section or 'chunk' of transcription was coded into the verbalisation categories. Depending on what the designer said, each section could be coded into multiple categories, or none of the categories if they talked about matters that did not fit into any of the categories, although this was very rare. An imaginary example of a sentence that would not fit into any of the categories would be, for example, if the designer said 'I'm supposed to have dinner with my mum tonight, I wonder what she's making?'. Sentences were categorised into 10-second bins, based on the beginning and ending time codes for the segments, as demonstrated by the example in Figure 44.

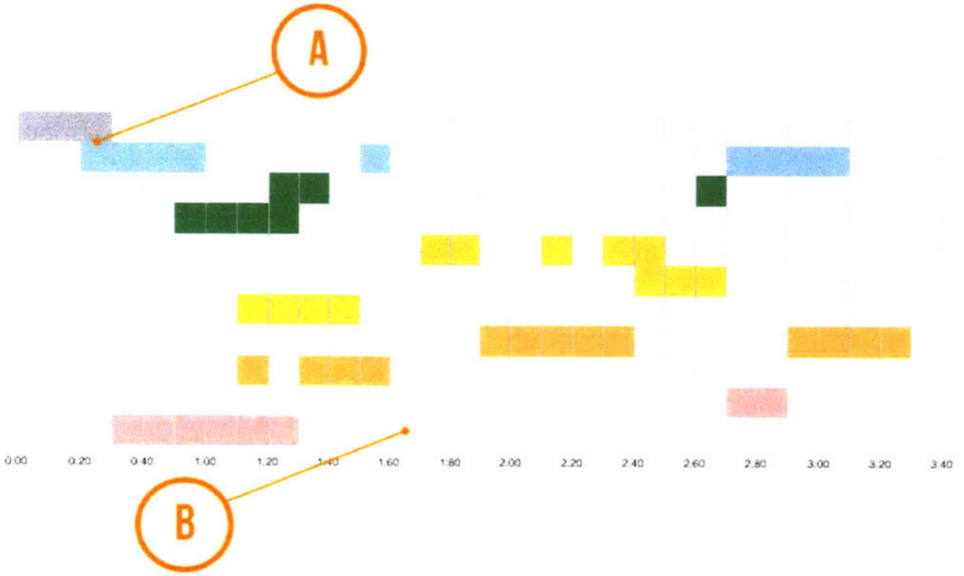


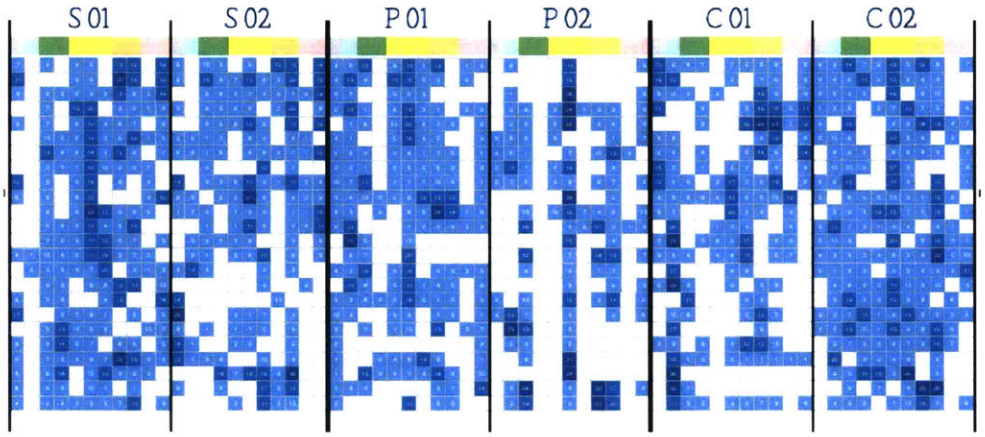
Figure 44 – Example categorisation of design activities.

As can seen in the image above, the same segment of an experiment could be coded into several different categories (A), and there could also be time periods with no coding (B), in case the designer spoke about something unrelated to the design process or experiment. The imaginary example above demonstrates what the first few minutes of the coded data would look like, this would consequently be done for all two hours of experiment time, for all six designers who were involved in part II of the study.

Once the data had been categorised, the 10-second data bins were combined to form larger 5-minute bins. A heat map was created, where a value between 0 and 30 was used, depending on how many of the 10-second long bins had initially been marked. In other words, if a designer talked about the design tool for the first thirty seconds, but then never mentioned it again in the first five minutes, the first 5-minute bin would receive a value of 3 (since there were three 10-second bins that would have been coded for the group 'design tool'). The data from the experiment, binned into 5-minute bins, can be seen in Table 10 below, with letters denoting the tool used (S – sketching, P – foam prototyping, and C – CAD), and the number identifying the designer.

Time runs vertically, with the start of the experiment at the top, and the end of the experiment at the bottom of the table. Different columns denote different design categories. Darker colours indicate a larger fraction of the earlier 10-second segments that were combined into 5-minute segments as having been coded for a particular category. A value of zero (indicating that none of the thirty 10-second segments had been coded for a certain category) would be coloured white, with progressively darker shades of blue for increasing values, up until thirty (indicating that all thirty of the 10-second segments had been coded for that category).

Table 10 – Heat map of designer verbalisation, in five-minute bins



Due to the sample size – although quantitative data was collected – the results were viewed as qualitative, to gain insights of the general trends and as a precursor when determining whether to conduct additional research on the topic in the future. One of the early hypotheses was that there would be clear differences between design tools. One example was the expectation that designers in the foam prototyping group would have perhaps talked more about ergonomics, and the shape of the concept, than designers in other groups. It was also expected that designers in the CAD group would talk about dimensions to a larger extent that designers in other groups. Although what the designers talked about – and thereby focused more on – are clearly different from one designer to another, there are no clear trends in terms of tools used, when comparing the data binned into 5-minute segments.

However, when examining the categories one by one (Figure 45), as percentages of the total experiment, one can see some general trends between tools, although further research is required to confirm the tentative findings. A value of 100% in the graphs below would correspond to each of the 720, 10-second segments containing speech that would be coded for that category. As an example, if one out of every three 10-second segments contains speech by the designer that would be categorised into 'aesthetics', then the bar for that category would extend to 33%.

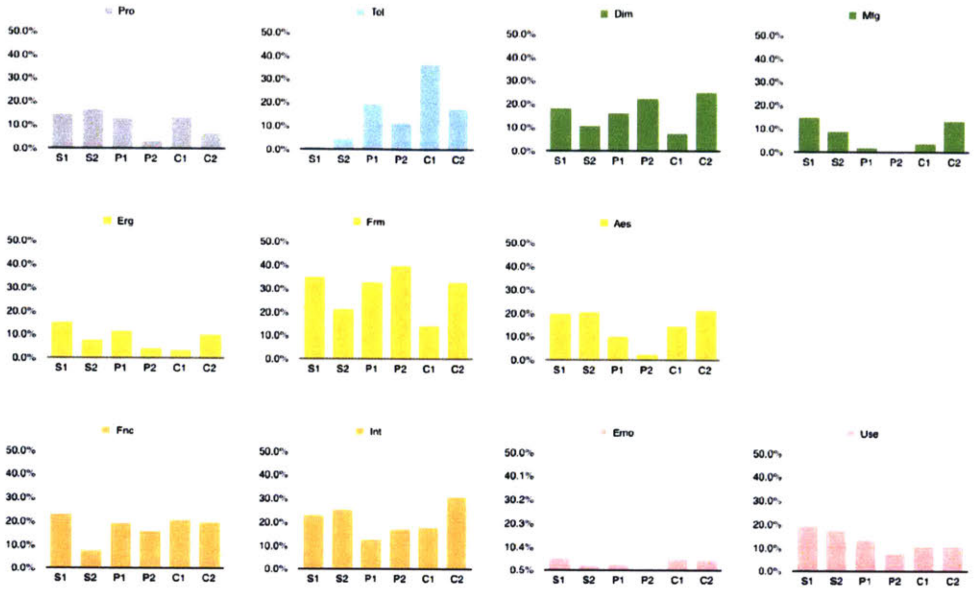


Figure 45 – Speech divided into categories; percentage of time v. designer

Although the sample size is very small, there seems to be a trend in some of the categories. Most notably, there were clearly more mentions of the design tool in the CAD group, with an average of 27% of the time segments containing at least one mention of design tool (often a negative one), compared to an average of 3% and 15% for the sketching and prototyping categories, respectively. Furthermore, perhaps somewhat surprisingly designers in the sketching group mentioned manufacturing in 12% of the time segments, with designers in the prototyping group barely mentioning it at all, and designers in the CAD group mentioning it in 8% of the time segments. The data would also suggest that designers in the foam prototyping group mentioned aesthetics and visual appearance clearly less (6% of time segments) than designers in the sketching or CAD groups, at 20% and 18%, respectively.

However, it is worth noting that too much significance should not be attached to the percentages, rather, they should be used as qualitative metrics to compare different tools with each other, but the percentage values in and of themselves carry little significance for two reasons. Firstly, due to the very limited sample size, the results are intriguing glimpses into possible trends, but far from conclusive, but also due to the fact that changing the time interval influences the results.

To demonstrate, in Figure 46 one can see how the interval size affects the error amount. Furthermore, since the differences are always rounded up (to avoid rounding down to zero and completely missing out on a section where a certain category was talked about) one can imagine taking it to its absurd limit, and only having one category, in which case there would be no difference between any of the categories – if a designer had talked about a category even once it would appear in the data the same as someone who had mentioned it on multiple occasions.

Clearly then, the higher the fidelity – in other words, the smaller the intervals – the better the data. However, as categorisation was extremely time consuming and tedious, it was not feasible to make categories that were exceedingly small. To that end, a 10-second interval was seen as sufficiently detailed; a 10-second interval was 1/720th of the total experiment time, in other words intervals with a magnitude of 0.14% of the total time.

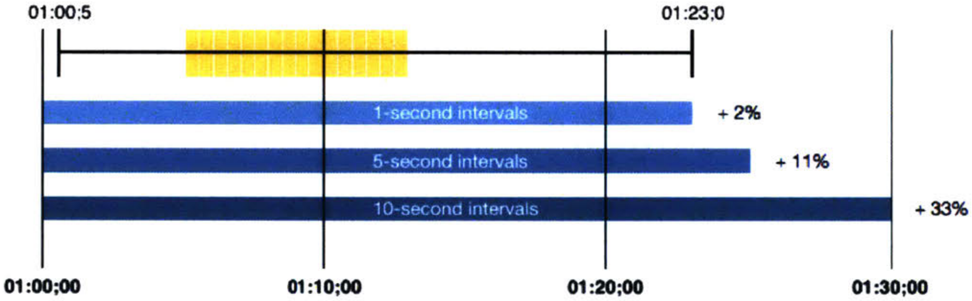


Figure 46 – Example time intervals and associated errors

4.5 How were good designs created?

As the goal of the study was to gain insights into the design process, once a rank order had been determined for the concepts, the next step was to determine which factors influenced the ranking of the concepts – in other words, what led to better design outcomes? How were good designs created?

4.5.1 Duration of design time

Compliant with internal review board requirements – and basic decency – participants were instructed that they were allowed to discontinue the experiment and leave at any point. Due to the long duration of the task, several test subjects did in fact leave before the allotted time was up, even though the experiment time was decreased after the pilot participants feedback, as can be seen in Figure 47. In fact, three participants declined the third design session altogether, while another participant did complete three sessions, but ended two of them prematurely and in the end only designed for the equivalent of two nominal length sessions (80 minutes in total). Two sketching participants also notified the experiment facilitator beforehand that they did not want to stay for the whole duration of the design task, and wanted an abbreviated schedule. Their session lengths were shortened to 25 minutes per session, although both decided to leave even before the condensed experiment time was up.

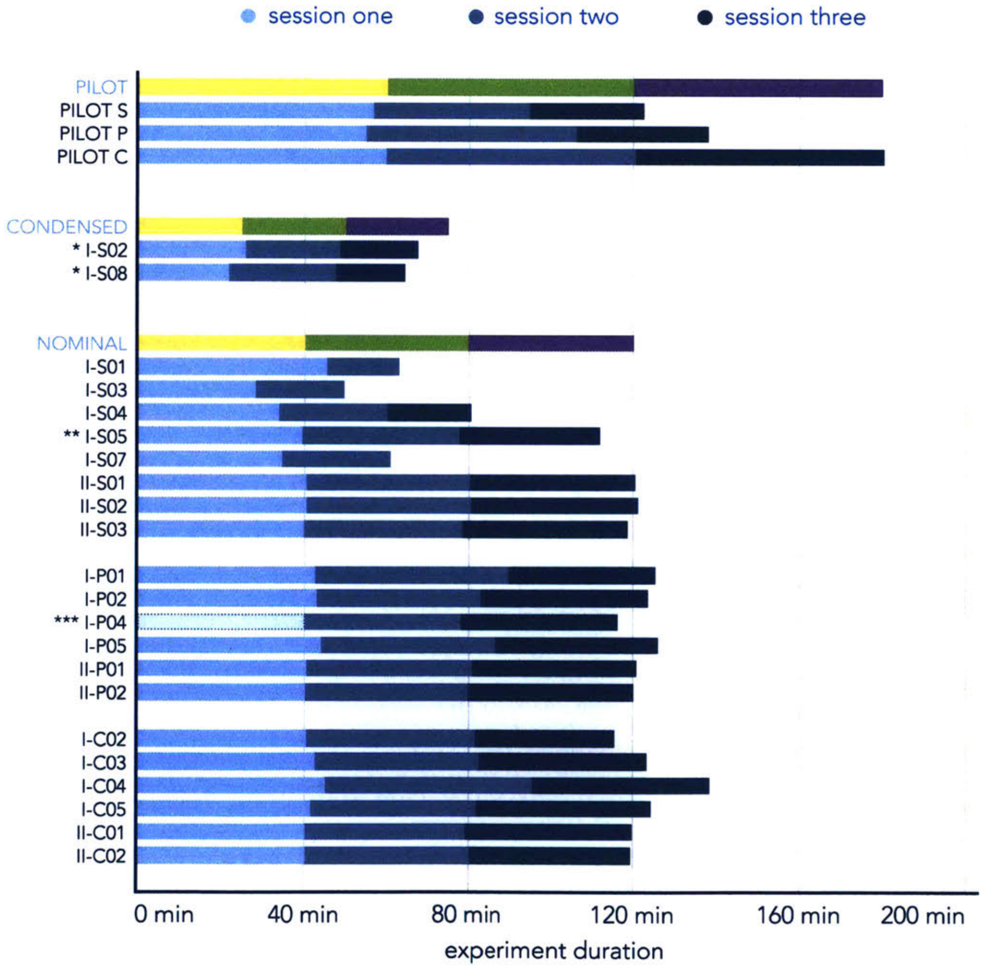


Figure 47 – Experiment durations – per designer

In Figure 47, four experiment durations have noteworthy deviations from the nominal experiment schedule, and are marked with asterisks as follows: * due to limited time availability, two sketching participants requested an abbreviated schedule of three 25 minute sessions, ** video file missing a few minutes of data, making exact timing impossible – based on progress of designer before and after missing footage, it is a question of a few minutes at the most, and *** due to multiple equipment failure, parts of the first section are missing (a camera charging cable became unplugged resulting in a dead battery mid-way through the first session of the experiment, and an over-full memory card for the backup camera prevented recording during the time the primary camera was turned off).

As timing for the experiment was kept independently with an additional stopwatch, it can be assumed that the length of the first section is close to the nominal 40 minutes, but the exact length is unknown. The participants with uncertain experiment durations were omitted in later calculations involving timing. The mean experiment times for part I and II, as well as a combined mean, are presented in Figure 48 with ± 1 standard deviation.

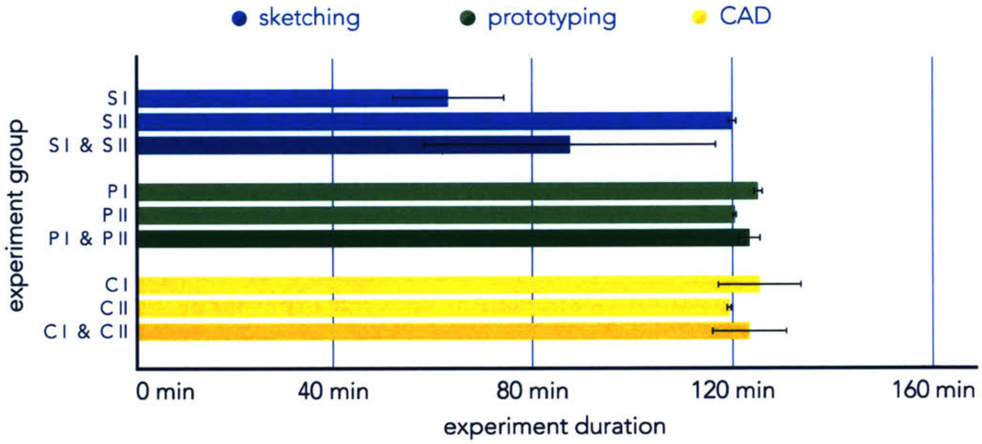


Figure 48 – Experiment durations – per tool

As can be seen from the graphs, there were significant differences in experiment lengths for the sketching participants between parts I and II. This can be, at least partially, explained by the differing experiment facilitators. Throughout the study, there were a total of six different experiment facilitators. Most of the sketching experiments were run by one of two facilitators, denoted as 'facilitator 1' and 'facilitator 2'. Figure 49 below shows the average experiment times for sketching sessions that were run by either facilitator 1 or facilitator 2.

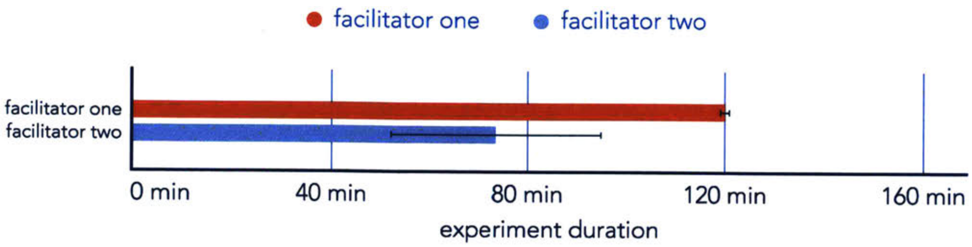


Figure 49 – Difference between experiment lengths between facilitators.

As can be seen, experiments run by facilitator 2 were significantly shorter. Some explanations for the observed difference could be cultural, since most of the participants that were overseen by facilitator 2 were located in Belgium, whereas all of the participants overseen by facilitator 1 were located in Massachusetts, USA. Analysing the video and audio recordings from the experiments, it also seems reasonable to conclude that although facilitator 1 definitely made the participants aware of their right to leave the experiment at any time, facilitator 2 brought it up more often. There were also subtle differences in the form of speech, whereby participants perhaps felt more comfortable leaving the experiment early when overseen by facilitator 2 (compare 'you are allowed to end the experiment whenever you want' versus 'do you still want to continue, or are you done?').

It was not possible to quantify the effect this may have had on the results, although analysis found no clear relation between experiment duration and concept quality, so it seems the possible bias introduced by the facilitators was not significant with regards to the general results of the study.

4.5.2 Impact of tool on design quality of concepts

Although there is little difference between the average quality of concepts created in the sketching, foam prototyping, and computer modelling groups, when looking at the top ranked ideas, some very clear differences emerge. Inspecting the top 10 most creative ideas out of the total 84 concepts, one can see that – when normalised to account for the uneven number of designers in the three different design tool groups – over half of the most creative ideas were created using foam prototypes, as seen in Figure 50.

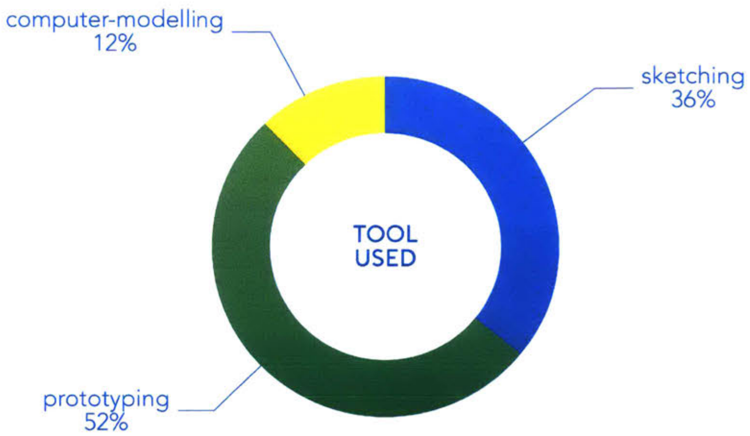


Figure 50 – Ten most creative ideas, normalised by the number of designers

At first glance one may conclude that by creating concepts using foam modelling, one has a higher probability of creating more creative concepts. However, if one looks at the data slightly differently, normalising by the number of concepts, not just the number of designers, one can see that on a per concept basis, there is little difference between the tools, even when observing just the top-ten most creative ideas, as seen in Figure 51.

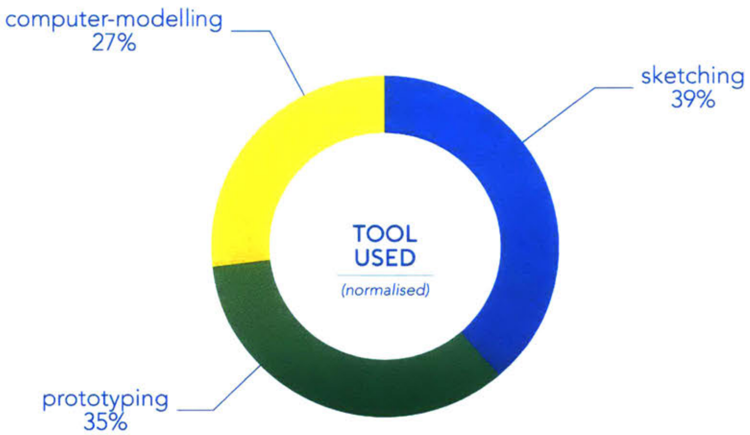


Figure 51 – Ten most creative ideas, normalised by the number of concepts

4.5.3 Impact of tool on number of concepts created

One of the more obvious effects of the design tool was on the number of concepts created. Somewhat surprisingly, designers in the 'prototyping' group created the largest number of concepts, with designers in the 'CAD' group creating the fewest, as expected. Figure 52 shows the average number of concepts created per group, and associated ± 1 standard error.

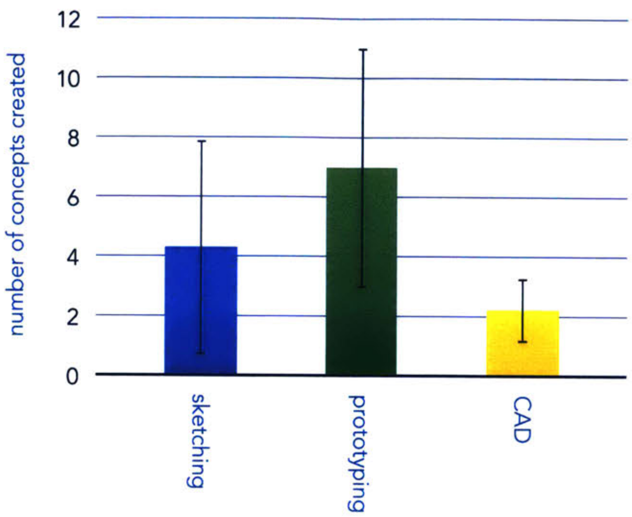


Figure 52 – Average number of concepts created per design tool group

Figure 53 shows the number of concepts created by each designer in part I of the study, as can be see, there are significant variations in the number of concepts created, although some general trends can also be observed. Designers in the computer modelling group created fewer concepts than designers in the other two groups. However, the difference between designers in the sketching and foam prototyping groups is subtler.

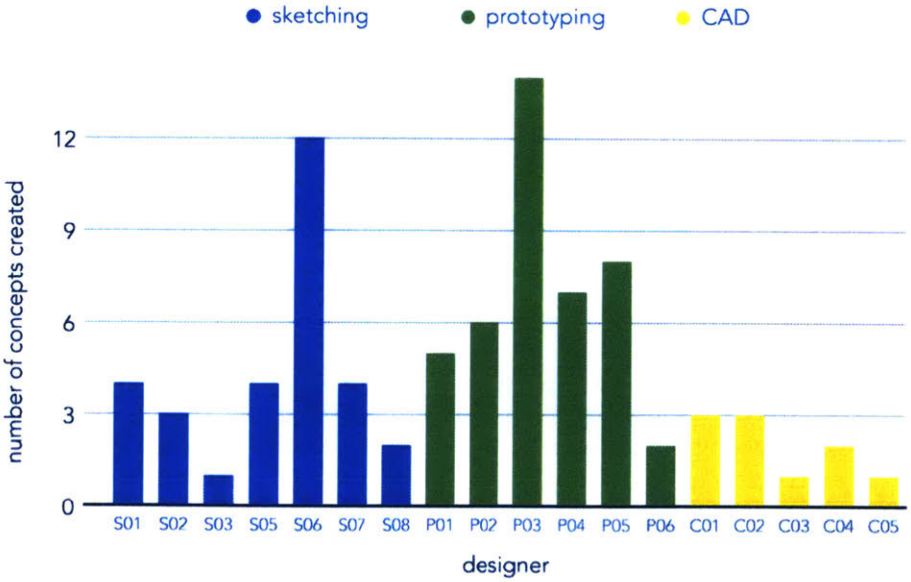


Figure 53 – Number of concepts per designer in part I of the study

4.5.4 Number of concepts and design quality

Using either of the two chosen metrics – creativity and overall score – one could calculate average creativity and quality ratings for design tools, although this may not provide any useful insights, as there is little difference between tools. It is important to realise that in a design process one does not necessarily care about the average quality of all the ideas, but rather, that you have a few – or at least one – very good one. Although having many concepts to choose from is beneficial, ultimately the goal of a designer or design team is to pick a single concept that becomes the final product. (Although care must be taken to explore the design space sufficiently broadly – a common mistake is to have a very limited exploration of the design space and prematurely settle on a specific concept. In other words, possible solutions – concepts – should not be excluded before there is a reason to do so.)

Creating a large number of concepts in the early stages of the design process is desirable since it provides different options for future development, although the main benefit is that it also increases the likelihood of creating better ideas in general, or as Linus Pauling put it:

> ***'The best way to have a good idea is to have a lot of ideas.'***

Figure 54 plots the total number of concepts that designers created against the highest ranked creative idea that they created – in other words, the points on the graph represent the best concept for each of the eighteen designers – whilst Figure 55 shows the total number of concepts against the highest overall score for a concept.

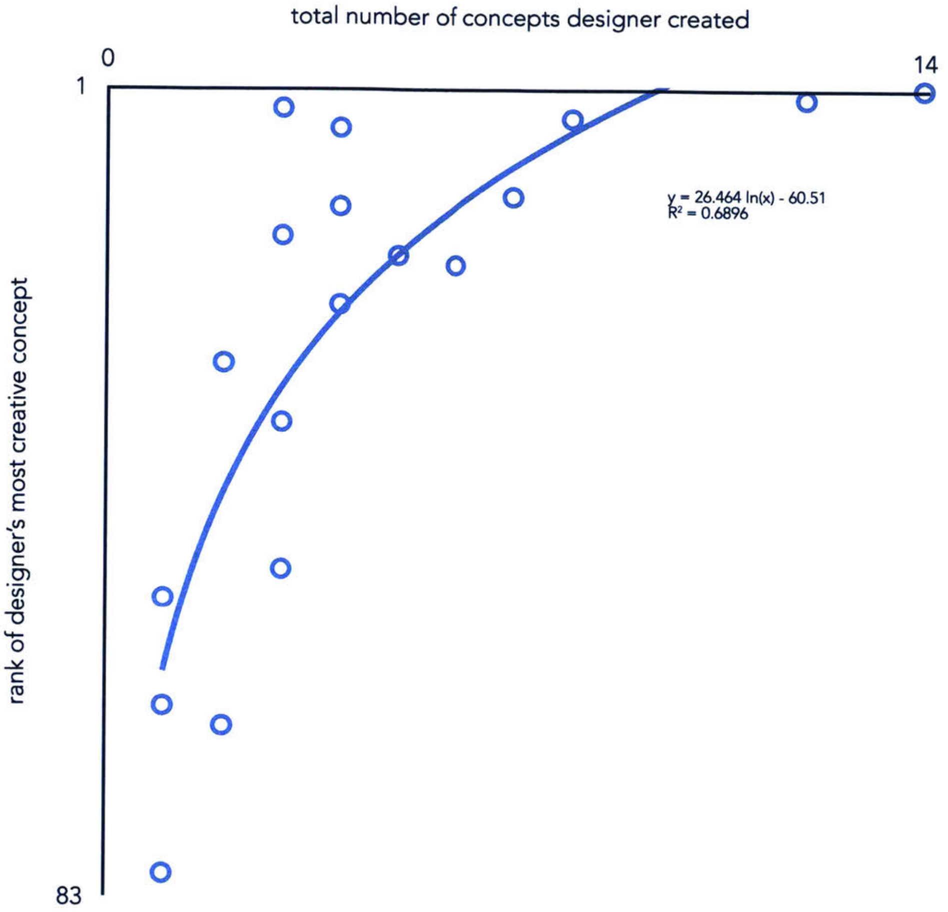


Figure 54 – Number of ideas versus the highest ranked creative idea

Even though the fit could be better, there is a clear trend, where the more concepts a designer created, the higher ranked their best idea was in the 'creativity' category. However, when looking at overall concept ranks, the relationship between the total number of ideas a designer had and the ranking of their best concept the relationship is much more vague, although a higher number of total ideas tended to mean a more highly ranked best idea here as well. It is also important to remember that the concepts in the two previous graphs are not the same. That is to say, the concepts shown in the two graphs represent the best concept in each respective category of the designers, but the concepts shown are not necessarily the same.

It is also worth noting that there is a strong negative correlation between the overall score rank and creativity rank – in other words, concepts that were ranked as creative received a low overall score rank, and vice versa (see Figure 32).

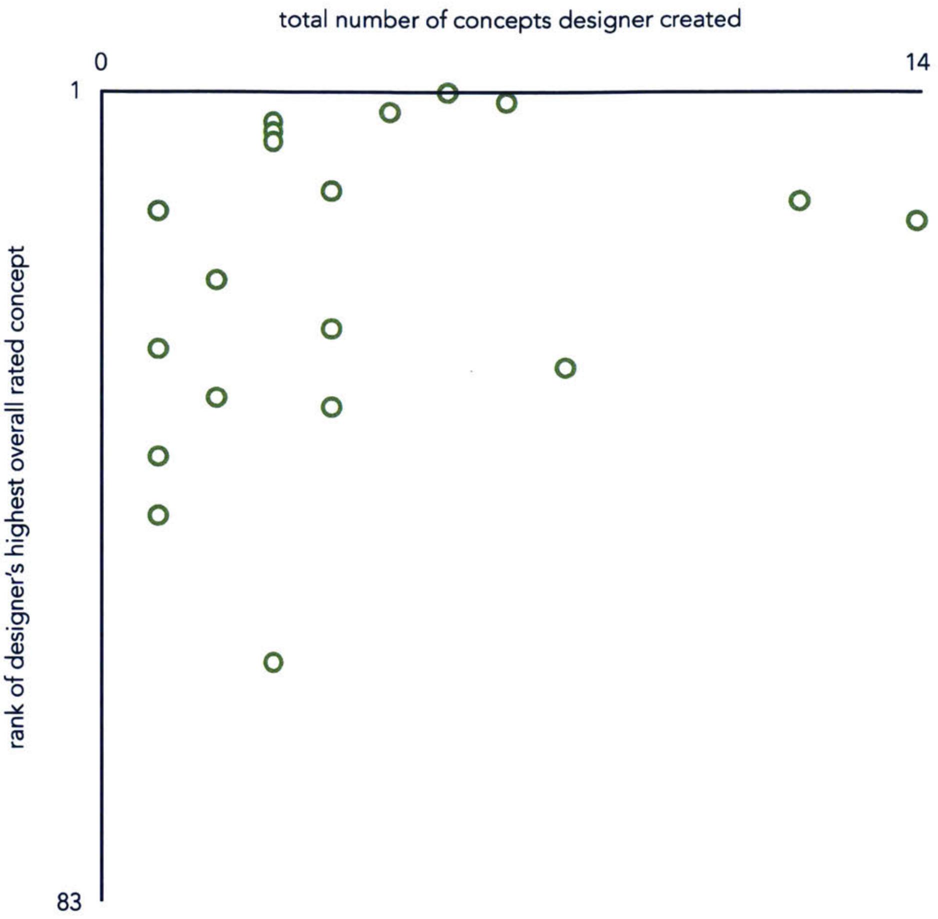


Figure 55 – Number of ideas versus the highest ranked overall score

4.5.5 Concept sequence and idea quality

Although prior research has shown that in divergent thinking tasks, ideas become more novel and original as time passes (Beaty and Silvia 2012), that was not the case in this study. As can be seen in Figure 56, creativity increases only slightly in later ideas. If every successive idea were more creative than the previous one, for every designer, all of the circles would lie on the dotted diagonal line. (Size of the circles is used to denote overlapping data points – larger circles indicate multiple data points; the largest circle denotes six overlapping data points.)

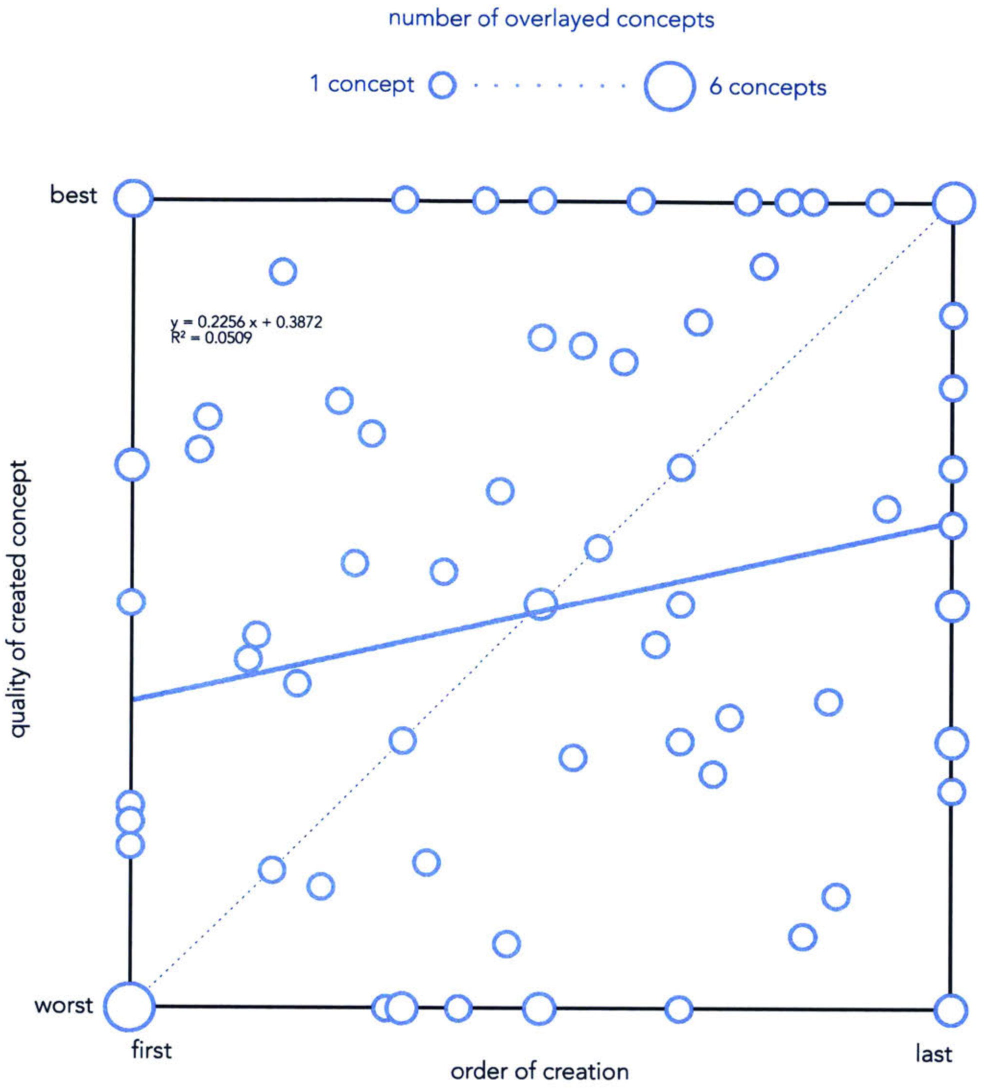


Figure 56 – Creativity rank against order of creation

One possible explanation for this deviation from prior research is that the design task at hand is actually not a divergent thinking situation, as ideas are not created in rapid succession. Furthermore, unlike a divergent thinking task (such as creating as many alternative uses for a brick as possible in a short amount of time) the goal of the designers was not to be as creative as possible – the goal for the designers was to create the three 'best' concepts that they could, to win the competition with. In other words, the designers tried to create designs they thought the panel of judges would appreciate, which may or may not be the most creative designs. Furthermore, apart from the sketching group, they would also have to be able to make it. In other words, it was not sufficient to be able to think of a creative design, they would also have to be able to make it with the tool that they were assigned to.

Having said that, when observing the timing data for concept creation, there were three designers (see participants S06, P03, and P05 in Figure 57) who created concepts more quickly – on average at a rate of about 9 minutes per concept, compared to the average of over 27 minutes per concept for all designers. Although still clearly not a brainstorming or divergent thinking rate of idea generation, compared to other designers, these were created more quickly. Two out of the three designers had a tendency to create more creative ideas as time passed, whereas one of them saw little change in creativity of ideas over time.

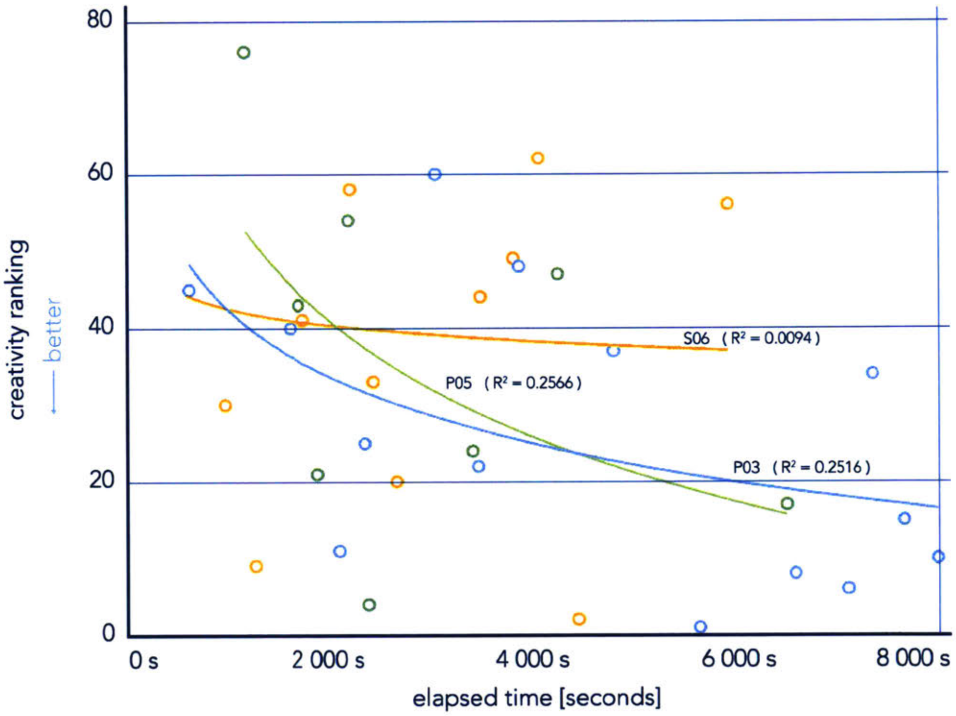


Figure 57 – Creativity rank as a function of time for selected designers

As mentioned earlier, even though the three designers in question created ideas more quickly, the idea generation rates are still far lower than what one would expect in divergent thinking tests, where participants create tens of ideas in a span of a few minutes. Another way to view the three designers in question is to think of them as having a different design process, where they went for quantity instead of (presumed) quality – creating on average 11.3 concepts, compared to just 3.3 by the remaining designers – with the intention to down-select to three concepts at the end. Even though the designers were under the impression that only three of their designs would be evaluated in the 'competition', they decided to create several concepts none-the-less, and then down select to their three favourite designs. As it turns out, these three designers created the first, second and fourth most highly ranked concepts in the creativity category, as well as concepts that were ranked in the top-thirty for overall score.

Another interesting discovery is that – based on somewhat scarce data from interviews – it seems that designers in this study were not very good at appraising which concepts would do well with the panel. The chosen concepts did not perform statistically significantly (t-test) better than the concepts the designers had chosen not to enter into the competition. Designers who made three or fewer concepts simply submitted them all, and therefore did not make any decisions on which concepts to keep – their data was not considered in the calculations.

4.5.6 Design attributes

Based on the set of attributes created in '3.4 Design attributes', a survey was run for six expert design reviewers, who went through the complete set of concepts, and assigned attributes to every concept. After several months, when it could be assumed that the reviewers had no recollection of what they had answered earlier, the survey was repeated for the same six experts. The experts had several years of experience with design practice, design research, or both.

The approach used for evaluating the results from the attribute survey was as follows: instead of assuming that the expert reviewers had to somehow uncover some absolute truth about the concepts (i.e. is there a button?) it was assumed that the Mechanical Turk reviewers would have some distribution of opinions for each of the attributes. In other words, it was not always obvious by looking at a concept whether a certain attribute existed or not. Sometimes the attribute was non-obvious. For the sake of argument, let us assume there was a small button in the corner of a remote control concept, and that the main surface area of the remote control was covered with a touch screen. Some reviewers may pick up on the button, but many wouldn't. Based on the expert reviewers an estimate would then be made on how obvious a certain attribute was, and how large a percentage of the

Mechanical Turk reviewers could be assumed to have been aware of that particular attribute.

It is this sort of distribution that was sought by having the six design experts evaluate the concepts for attributes. The output for this activity was a distribution of values for all of the attributes, for all of the concepts. What this specifically means is that every concept had a value for each of the attributes in the range of 0% to 100%. That is to say, how many per cent of the expert reviewers agreed that the concept in question possessed the attribute in question. This was based on a set of six design experts, doing the evaluation twice. Due to the time (several months) between evaluation rounds, it was judged impossible for the reviewers the remember any of their previous answers. Thereby, the range of values produced by the six experts can be equated to having been produced by a set of twelve reviewers. The reason why six experts were used twice, instead of simply having twelve reviewers evaluate the designs was that assigning attributes to 84 concepts was extremely arduous and took several hours to complete, and suitable design experts were hard to come by.

An inter-rater reliability score was calculated to see how consistently attributes were assigned to the concept sketches. Fleiss' Kappa was used since it can be used in cases with more than two reviewers. Using Landis and Koch's criteria it was found that there was substantial inconsistency between raters for certain design attributes. In other words, some of the categories that had been created were very unclear. (See original categories in Table 6 on page 41.) Some categories were removed altogether, while others were combined into larger categories. The final categorisation is presented in Table 11.

Table 11 – Final list of design attributes

form factor	input	interaction
• hands / eyes • body / other	• standard / game ctrl • phone • mouse • other	• buttons / touchpad • joystick • scroll wheel • gestural • other

Spearman correlations were calculated between each of the design attributes, as well as with the design tools used. The results are presented in Table 12, with shaded areas highlighting statistically significant correlations. It is also good to note that the table is symmetric. An example to demonstrate how to interpret Table 12: on the y-axis if one chooses the category 'input / buttons and touch' and on the x-axis the column labelled 'form / phone or tablet' one can see that these two categories are positively correlated (rho: +0.311, p-value: 0.001). In other words, concepts that were deemed to have a shape resembling a phone or a tablet were also more likely to have been categorised as having buttons and/or a touchpad.

Table 12 – Spearman correlations between attributes and design tools

		Interaction		Form				Input					Design tool		
		Hands and eyes	Body and other	Standard and game	Phone/ tablet	Mouse	Other	Buttons and touch	Joystick	Scroll wheel	Gestural	Other	CAD	Sketch	Proto
Interaction	Hands and eyes		−0.553 (0.000)	+0.346 (0.001)	+0.182 (0.099)	+0.168 (0.128)	−0.411 (0.000)	+0.538 (0.000)	+0.102 (0.357)	+0.173 (0.116)	−0.056 (0.615)	−0.497 (0.000)	+0.112 (0.310)	+0.033 (0.769)	−0.097 (0.379)
	Body and other	−0.553 (0.000)		−0.370 (0.001)	−0.278 (0.011)	−0.082 (0.458)	+0.401 (0.000)	−0.441 (0.000)	−0.165 (0.134)	−0.038 (0.730)	+0.400 (0.000)	+0.570 (0.000)	+0.015 (0.891)	−0.197 (0.073)	+0.152 (0.168)
Form	Standard and game	+0.346 (0.001)	−0.370 (0.001)		0.043 (0.700)	+0.076 (0.494)	−0.712 (0.000)	+0.607 (0.000)	+0.365 (0.001)	+0.010 (0.926)	−0.191 (0.083)	−0.582 (0.000)	+0.187 (0.088)	+0.099 (0.371)	−0.187 (0.089)
	Phone	+0.182 (0.099)	−0.278 (0.011)	−0.043 (0.700)		+0.043 (0.699)	−0.363 (0.001)	+0.311 (0.004)	−0.091 (0.413)	−0.131 (0.236)	−0.208 (0.058)	−0.385 (0.000)	+0.044 (0.690)	−0.222 (0.042)	+0.170 (0.123)
	Mouse	+0.168 (0.128)	−0.082 (0.458)	+0.076 (0.494)	+0.043 (0.699)		−0.333 (0.002)	+0.230 (0.036)	−0.049 (0.660)	−0.008 (0.946)	+0.275 (0.011)	−0.025 (0.824)	−0.132 (0.231)	−0.167 (0.128)	+0.246 (0.024)
	Other	−0.411 (0.000)	+0.401 (0.000)	−0.712 (0.000)	−0.363 (0.001)	−0.333 (0.002)		−0.701 (0.000)	−0.189 (0.085)	+0.072 (0.514)	+0.087 (0.430)	+0.610 (0.000)	−0.348 (0.001)	+0.108 (0.329)	+0.103 (0.352)
Input	Buttons and touch	+0.538 (0.000)	−0.441 (0.000)	+0.607 (0.000)	+0.311 (0.004)	+0.230 (0.036)	−0.701 (0.000)		+0.005 (0.967)	+0.053 (0.634)	−0.006 (0.956)	−0.804 (0.000)	+0.255 (0.019)	+0.001 (0.991)	−0.157 (0.155)
	Joystick	+0.102 (0.357)	−0.165 (0.134)	+0.365 (0.001)	−0.091 (0.413)	−0.049 (0.660)	−0.189 (0.085)	+0.005 (0.967)		+0.099 (0.373)	−0.056 (0.612)	−0.212 (0.053)	−0.082 (0.458)	+0.197 (0.073)	−0.143 (0.194)
	Scroll wheel	+0.173 (0.116)	−0.038 (0.730)	+0.010 (0.926)	−0.131 (0.236)	−0.008 (0.946)	+0.072 (0.514)	+0.053 (0.634)	+0.099 (0.373)		−0.045 (0.685)	−0.018 (0.869)	+0.212 (0.053)	+0.015 (0.890)	−0.175 (0.111)
	Gestural	−0.056 (0.615)	+0.400 (0.000)	−0.191 (0.083)	−0.208 (0.058)	+0.275 (0.011)	+0.087 (0.430)	−0.006 (0.956)	−0.056 (0.612)	−0.045 (0.685)		+0.126 (0.253)	−0.056 (0.611)	−0.014 (0.902)	+0.034 (0.762)
	Other	−0.497 (0.000)	+0.570 (0.000)	−0.582 (0.000)	−0.385 (0.000)	−0.025 (0.824)	+0.610 (0.000)	−0.804 (0.000)	−0.212 (0.053)	−0.018 (0.869)	+0.126 (0.253)		−0.135 (0.223)	−0.097 (0.378)	+0.159 (0.148)
Design tool	CAD	+0.112 (0.310)	+0.015 (0.891)	+0.187 (0.088)	+0.044 (0.690)	−0.132 (0.231)	−0.348 (0.001)	+0.255 (0.019)	−0.082 (0.458)	+0.212 (0.053)	−0.056 (0.611)	−0.135 (0.223)		−0.272 (0.012)	−0.386 (0.000)
	Sketch	+0.033 (0.769)	−0.197 (0.073)	+0.099 (0.371)	−0.222 (0.042)	−0.167 (0.128)	+0.108 (0.329)	+0.001 (0.991)	+0.197 (0.073)	+0.015 (0.890)	−0.014 (0.902)	−0.097 (0.378)	−0.272 (0.012)		−0.760 (0.000)
	Proto	−0.097 (0.379)	+0.152 (0.168)	−0.187 (0.089)	+0.170 (0.123)	+0.246 (0.024)	+0.103 (0.352)	−0.157 (0.155)	−0.143 (0.194)	−0.175 (0.111)	+0.034 (0.762)	+0.159 (0.148)	−0.386 (0.000)	−0.760 (0.000)	

4.5.7 Designer types

Based on an inspection of the number and quality of ideas in chapters 4.5.2 and 4.5.4, it would seem that – as Linus Pauling asserted – the best way to have a good idea is to have many ideas. Nevertheless, in an attempt to examine whether differences could be identified between different designers beyond merely the number of concepts they created, each individual designer was examined in more detail on the creativity metric. In other words for all subsequent analyses and results the designers have only been considered from the perspective of their creativity rank.

Figure 58 shows the creativity ranks for all of the created concepts, grouped by designer (each vertical line represents a different designer, with creativity increasing when going down along the y-axis).

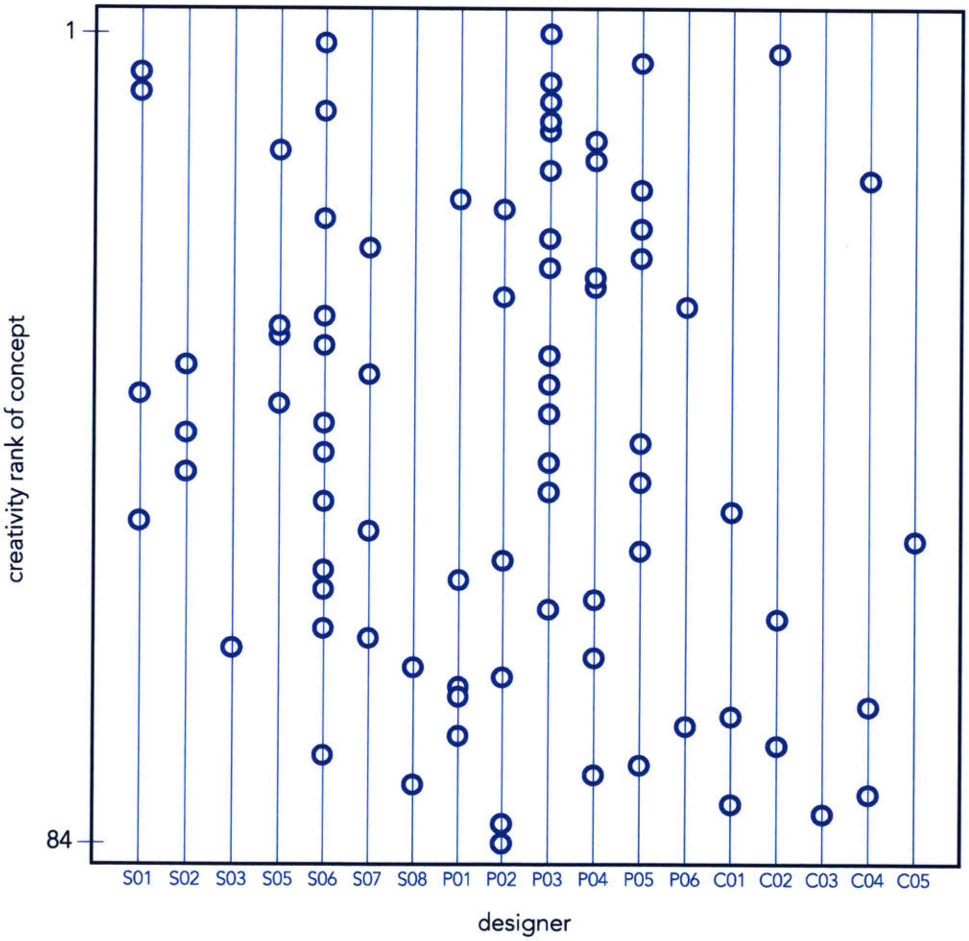


Figure 58 – Creativity ranks for concepts, grouped by designer

Although, on average, the more concepts a designer created, the higher the likelihood that their 'best' idea (ranked on the creativity metric) was ranked highly. One can also see that some designers, such as designer C02 in the previous graph, created a very creative concept, despite only creating a few concepts in total. In order to compare designers between each other, step functions were created to describe the creativity levels of the their concepts, such as in Figure 59.

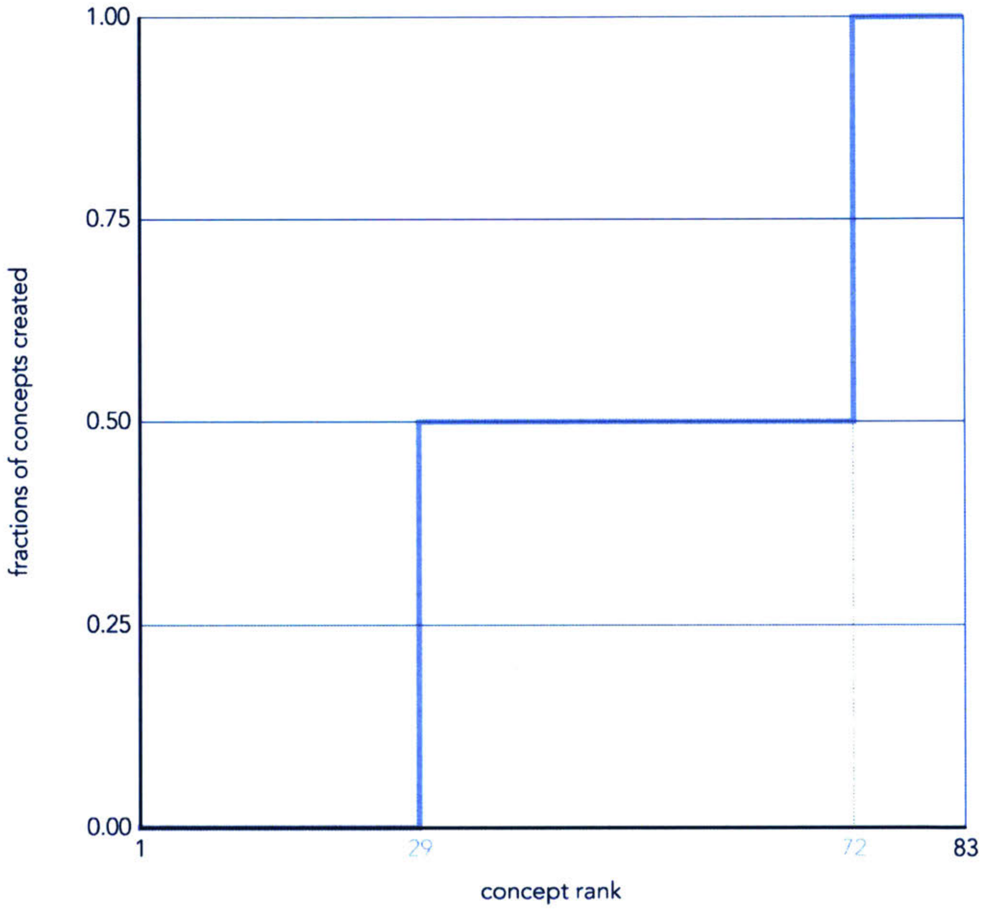


Figure 59 – Normalised cumulative function for a designer – example one

The graph above describes the creativity level of the concepts that the chosen designer created. They created two concepts, which were ranked 29th and 72nd in creativity.

Another example of a different designer can be see in Figure 60, where the second designer (green line), has created three concepts, ranked 35th, 42nd, and 46th most creative.

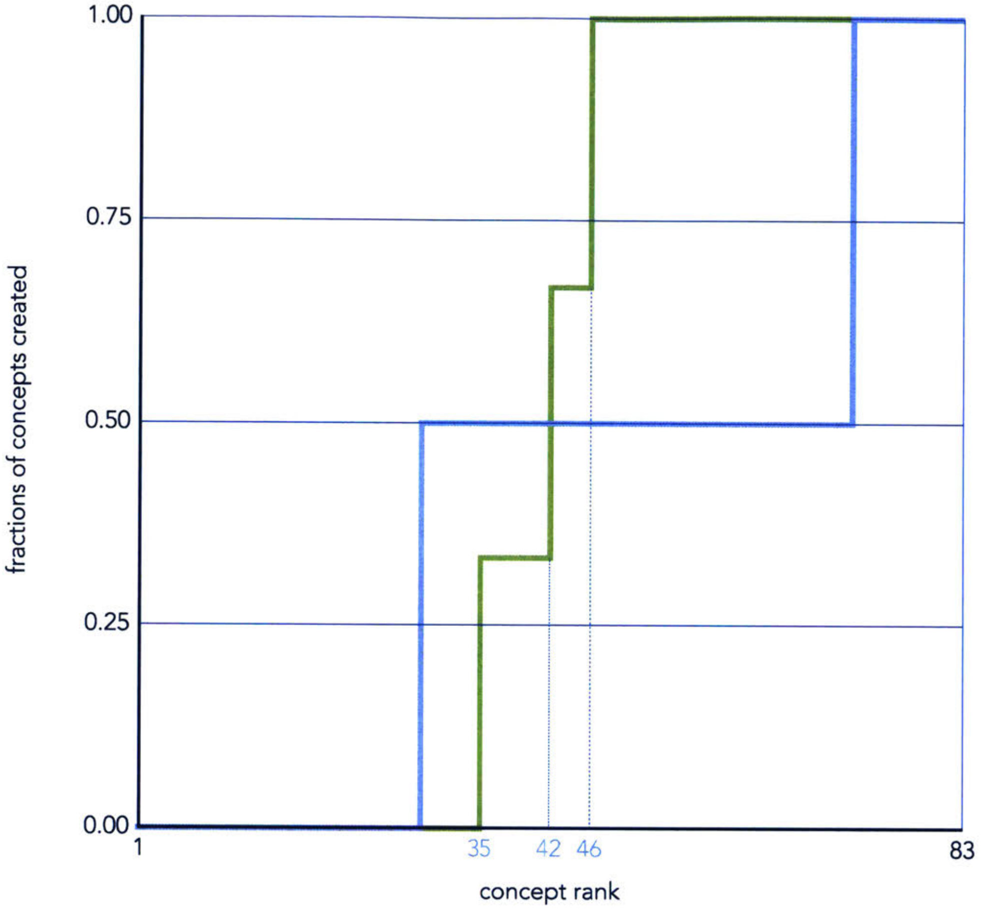


Figure 60 – Normalised cumulative function for a designer – example two

The y-axis denotes what fraction of the concepts has been taken into account at a certain point, or what percentage of the concepts that the designer created was above a certain threshold. For example, in Figure 61, if one looks at the rank value 37 (on the x-axis), one can see that the value of the function at that point is 0.70. The way to interpret this is, that 70% of the concepts that this designer created were ranked higher than 37 out of 83 (the baseline remote was not included in these analyses).

In other words, the steeper the graph at the beginning, and the quicker it reaches a value of 1 on the y-axis, the more highly ranked the majority of a designers' concepts are, as shown in Figure 62.

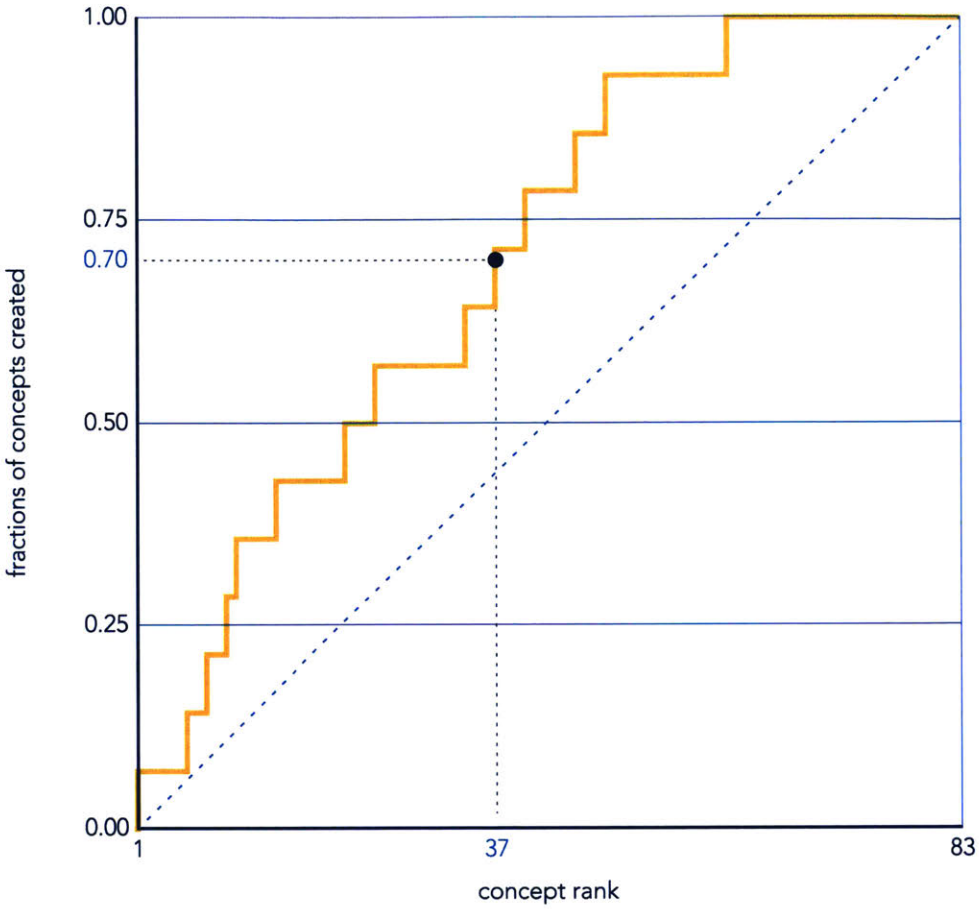


Figure 61 – Normalised cumulative function for a designer – example three

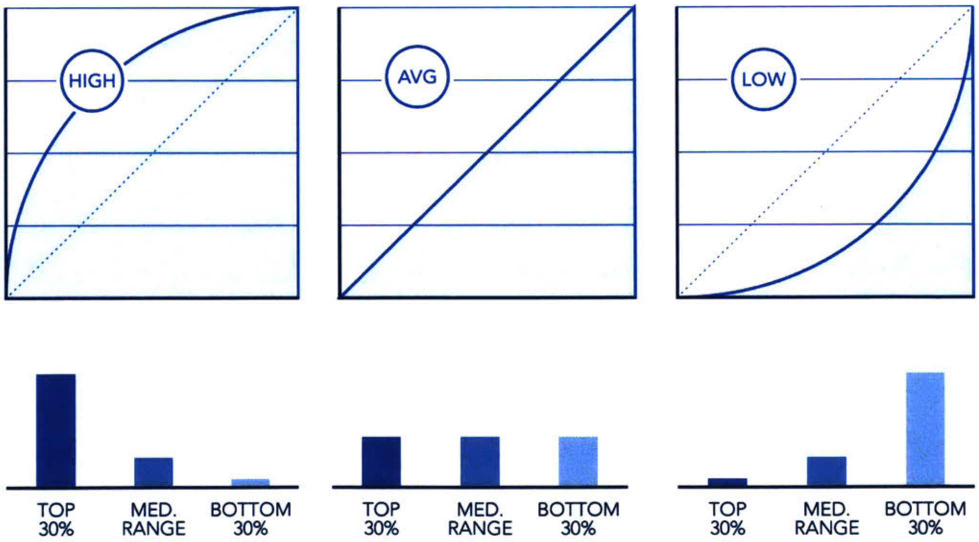


Figure 62 – Distribution of ranks base on graph shape

In order to see if there were distinct differences between the designers in terms of how creative their ideas were cumulative distribution functions were generated for each participant (see Figure 63 below), which were then clustered using DBSCAN (density-based spatial clustering of applications with noise) using a Manhattan distance metric.

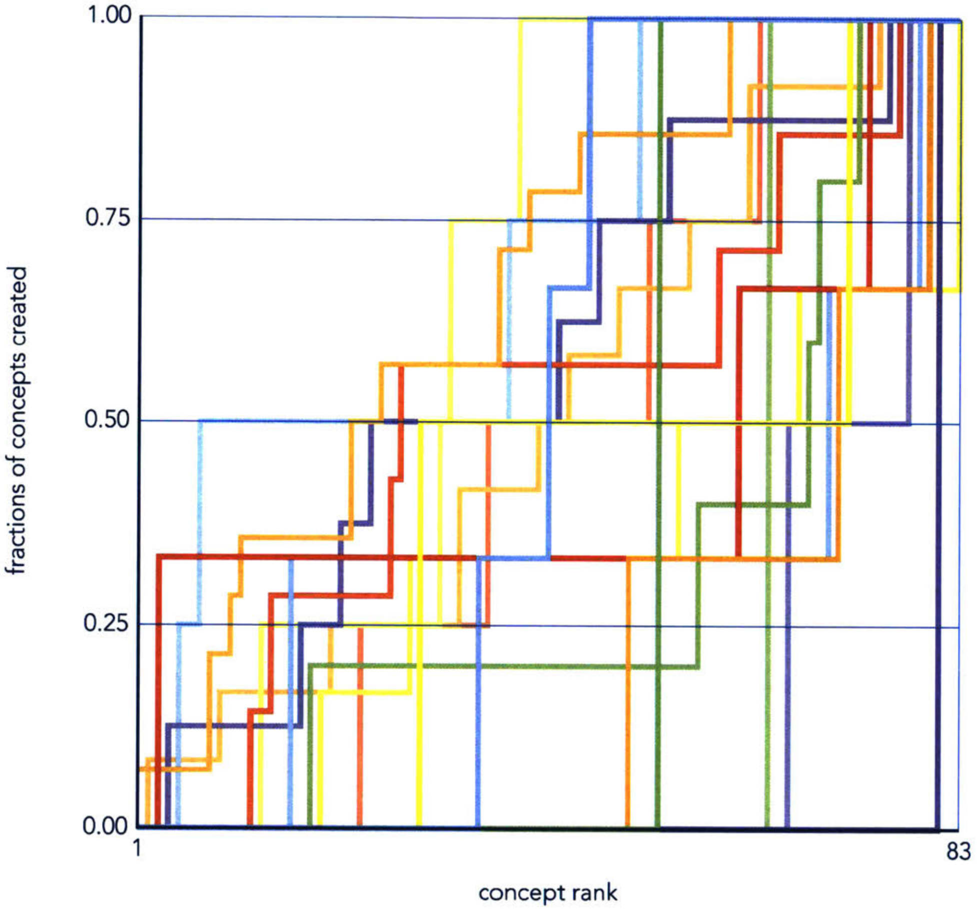


Figure 63 – Cumulative distribution functions for all designers

The clustering groups change depending on what values are chosen for the parameters 'minPts' (minimum number of designers per group) and ε (epsilon, which defines how close points need to be in order to be considered part of the same cluster). Choosing the parameters based on domain knowledge is a generally accepted practice when using density-based clustering.

Based on the sample size of designers used in the study, a minimum group size of three designers per cluster was deemed appropriate. Based on visual inspection of the data and clusters, and by trying different values, an epsilon value of 10 was used. Different epsilon and minPts values were used to see how the data behaved – with epsilon values that were too high, all of the designers were part of the same cluster, and with values that were too low, designers quickly became marked as outliers or their own individual clusters. When conducting the clustering analysis, the goal was to create meaningful clusters with a minimal number of designers that were marked as outliers. 'Meaningful' in this instance refers to group sizes that were at neither end of the spectrum (only one group that all designers are a part of, or separate groups for each individual designer). A visual inspection was also conducted to judge how well the designer profiles within each group matched.

With the parameters (ε:10, minPts:3) the designers group nicely into four main groups of designers, plus a group of three outliers (not shown) that do not fit into any of the four groups, as shown in Figure 64. The groups are characterized by clear differences in the creative quality of the ideas they produce. Although the total number of concepts a designer created correlated with quality of their most creative idea, in certain design situations there may be a need for a second or third creative concept. As the total number of concepts created did not correlate with which group a designer was clustered into, it would seem to indicate differences between designers beyond simply the number of concepts they create.

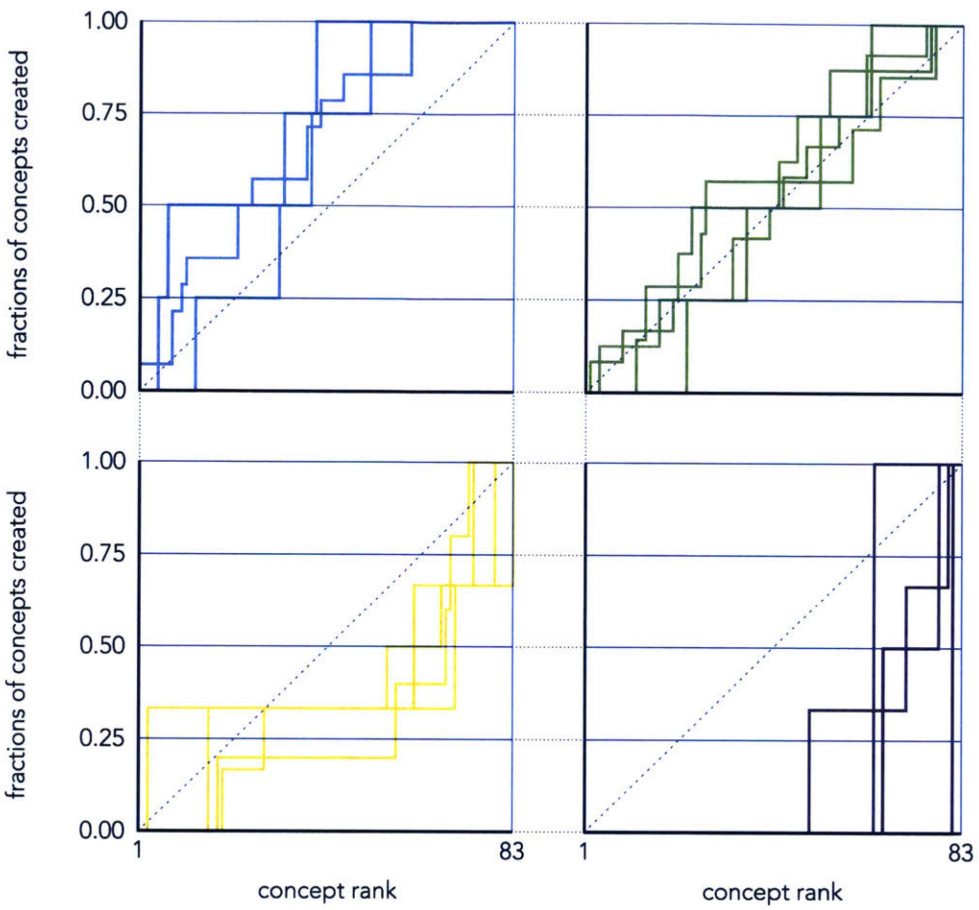


Figure 64 – Designers clustered into four groups based on creativity ranks

In other words, although the number of ideas created may be a good indicator for the creative quality of a designer's top idea, it is less of an indicator for the general profile of the quality of all of their ideas, which is evident when comparing the four different groups above. The same information – with the average step functions of each clustering group – can be seen in Figure 65.

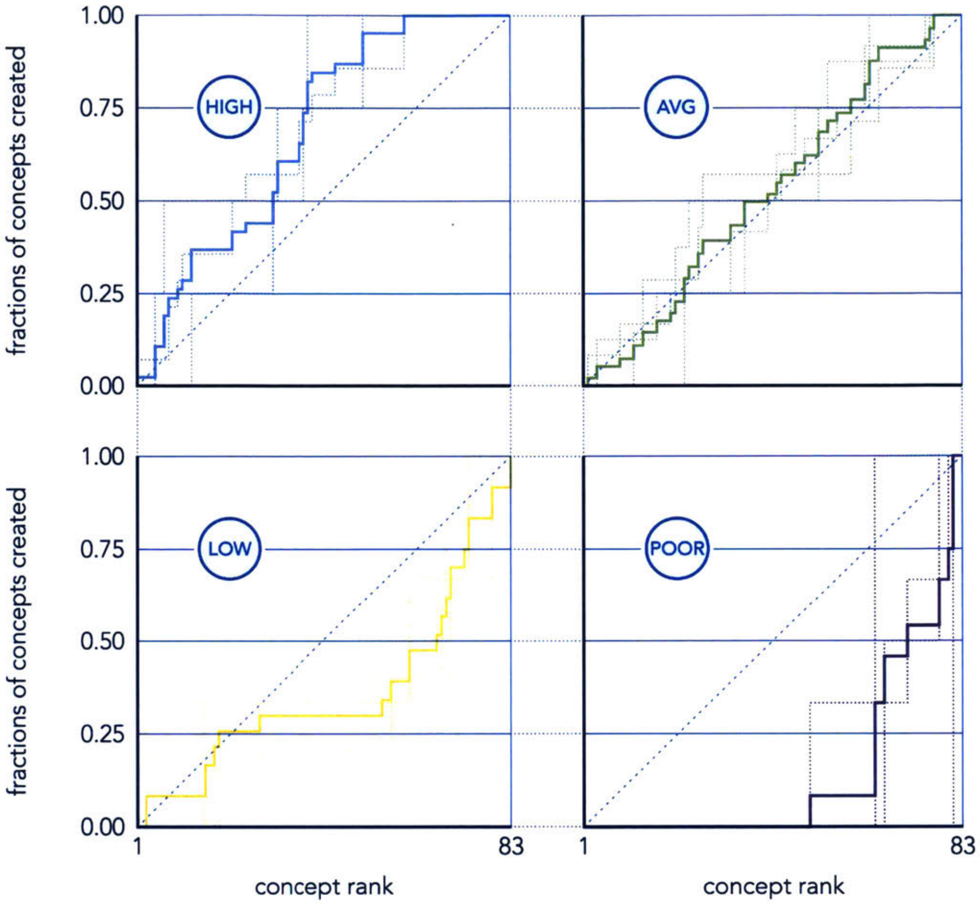


Figure 65 – Per group averages, based on performance in creativity

For easier comparison, the averages from each clustering group are also presented together in the same graph – see Figure 66. One can think of the groups in terms of the creative design output of the designers, with group one (light blue) being 'high-performing' designers, group two (green) being 'average' performing designers, group three (orange) being 'below average' performers, and group four (purple) being 'poorly performing' designers.

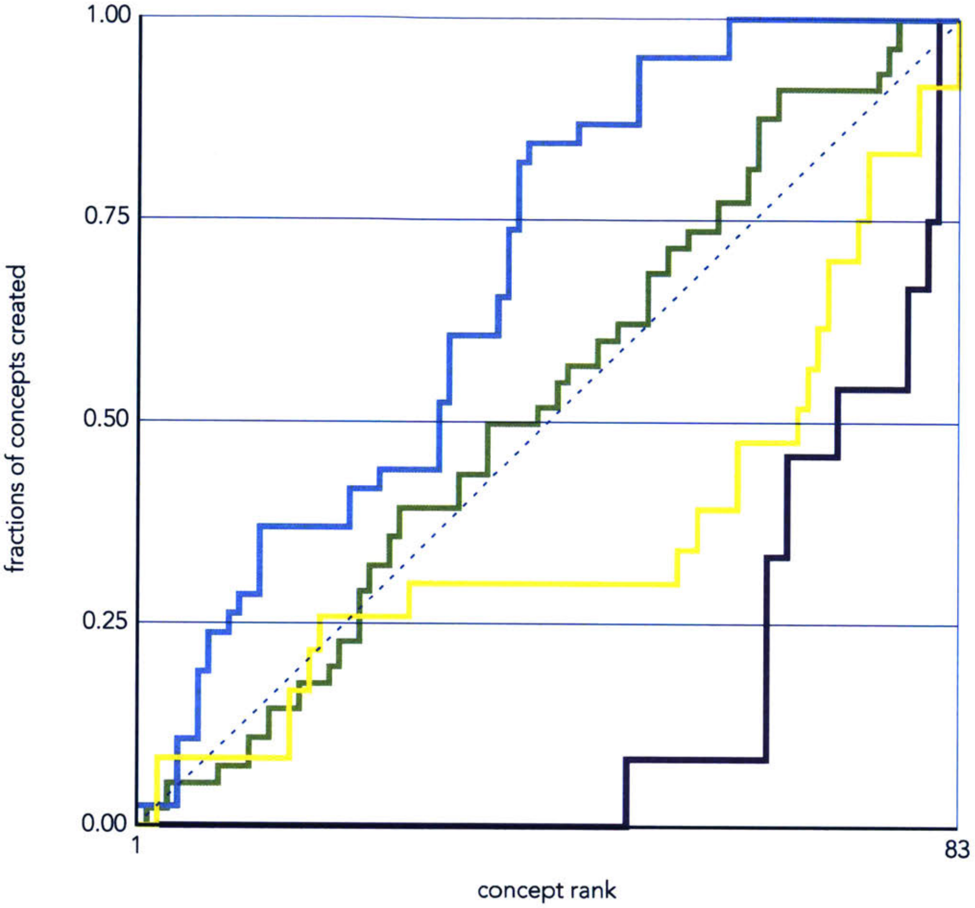


Figure 66 – Average cumulative distribution functions for the four groups

Even though the end goal of a design process is usually to create a single product or process, and thereby one could imagine only a single good idea (executed well) being enough, there are often unforeseen complications that require the design team to alter their design path, sometimes requiring another idea to be taken under consideration. At times like these, having more breadth in high quality creative ideas is desirable to provide the design team with multiple options and design avenues to pursue.

Therefore, although there is little difference in terms of the number of ideas created and the rank of the most creative ideas created between the top two groups (see Figure 67 and Figure 68), and to a lesser degree the third group, there is still a distinct difference in the quality of the top ideas if one considers more ideas than simply the single highest ranked concept. In other words, although both groups one ('high') and two ('avg') have nearly identical 'best' ideas, if looking at the top quartile of ideas, the quality of those produced by group one exceeds those produced by group two. Put more colloquially, if you want five good ideas, a designer from group one is more likely to produce them than a designer from group two.

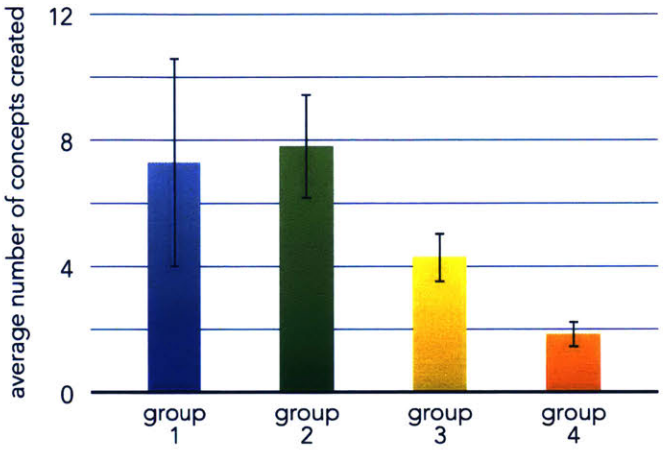


Figure 67 – Average number of concepts created in each of the four groups

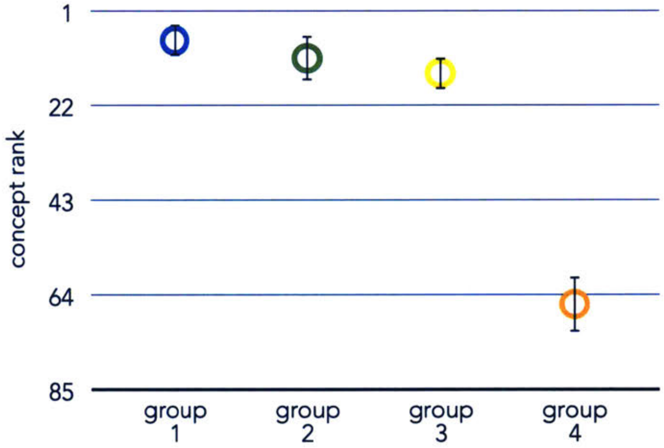


Figure 68 – Average of the most creative idea created in each of the groups

To re-iterate, Figure 69 shows that although there is little difference in the ***most*** creative ideas that designers in the top two groups created, there is a distinct difference in the average quality of their output, with designers in group one being more likely to create several top ranked creative ideas.

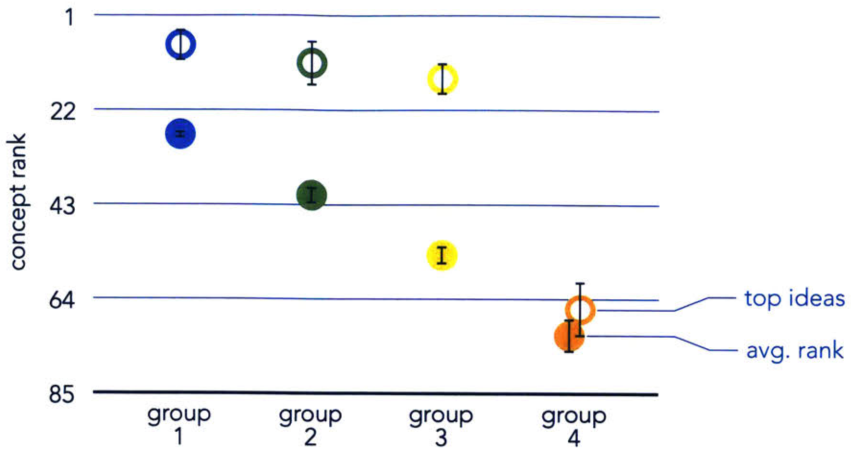


Figure 69 – Differences in the most creative, and average creativity ranks

The question then arises, what determines which group a designer ends up in? Unfortunately, the data is insufficient to answer this question, but this raises the opportunity for several directions for future research.

5 DISCUSSION

5.1 General remarks

As noted before, design is an open-ended, goal-oriented process where the aim is to create a future situation that is superior to the current one, but where the answer and the question itself, are both unclear. That is to say, there is no clear 'correct' answer and design outcomes are always debatable and subjective to some degree.

Although some studies rely on expert panels to evaluate design outcome, in this study the evaluations from several hundred 'average' users was collected as it was seen as a more reliable metric for design quality. This was also compared to a smaller sample of expert designers working in product, graphic and industrial design. Analysis showed that there was little difference in evaluation on the metrics measured.

5.2 Summary of results

The aim of this thesis was to study the influence that three selected design tools – sketching, foam prototyping, and computer modelling – have on design outcomes. The main results from the study will be addressed next, in terms of the original research questions posed.

> ***Research question 1: Are there significant differences in terms of the creative value of concepts between the chosen design tools?***

The average quality of the concepts did not differ significantly between the three studied groups – sketching, foam prototyping, and computer modelling. However, in design often one is not concerned with the average quality of concepts, but rather with the quality of the ***best*** concepts that one produces – when examining the top-ten designs on various metrics, it became evident that designers in the foam prototyping group created significantly more top-rated concepts in the creativity category. This categorical difference disappears, however, when the number of concepts the designers created is taken into account. In other words, on a per concept basis the ideas created in the three different groups did not differ significantly – concepts created in the foam prototyping group were in the top-ten of creativity, comfort and aesthetics due to the larger number of concepts that were created.

> *Research question 2: How does the rate of idea generation differ between the chosen design tools, in terms of quality of design outcomes?*

Unsurprisingly, designers in the computer modelling group created significantly fewer concepts than designers in the two other groups, with designers in the foam-modelling group creating the largest number of concepts. Closer inspection revealed that this was, at least partially, due to the high-fidelity of the sketches produced by the designers in the sketching group.

Clustering of the individual designers also revealed, that in addition to the observed trend of more concepts increasing the likelihood of creating a top-rated concept, irrespective of the design tool, the designers themselves could be categorised into four different groups based on their performance in this study.

> *Research question 3: Do the different chosen design tools influence designers to create a certain type of concept, distinct from the other design tools examined?*

Although not many, there were some statistically significant differences between the design tools. Concepts created in the computer modelling category were more likely to contain buttons, and ***not*** to have novel form factors. Although the form factors of the created concepts were not directly evaluated for the level of flowing, or organic, shapes, concepts created in the foam-modelling group tended to contain more form categorised as (computer) mouse-like, which qualitatively tended to be significantly more organic in their shapes.

The main contributions of this thesis can be seen in terms of a greater understanding of the influence of design tools on concept generation in an individual setting. The results caution against the overuse of computer modelling, especially in the early stages of the design process, due to the lower idea generation output and constrained idea space.

The results from the study also point towards inherent differences between designers – some designers were higher performing than others, irrespective of the design tool to which they were assigned. Due to the limited sample size, reasons for this could not be uncovered, but this presents an interesting opportunity for future research.

6 FUTURE WORK

This doctoral thesis presented several interesting findings, many of which pose opportunities for future research, some of which are briefly described below.

6.1 Expansion of original study

In this study, data from part I was used to draw conclusions regarding the type and quality of concepts produced, and data from part II was studied in an attempt to gain insight into the thought processes of designers using different design tools. However, the concepts created in part II were not studied in terms of their creativity (or other) ranks. Therefore this dataset presents a good opportunity for further data analysis and comparison against the results from part I. In addition to these two aforementioned sections, a third round of data from nine designers (two sketching, four prototype, three CAD) has already been collected, although this set has not been processed. Together, these three sets of data (part I, II & III) present a good opportunity to expand the original analysis, and confirm the findings from this study; repeating the same data collection and analysis for parts II and III provides a valuable opportunity to strengthen the findings from the data analysis on part I. Due to the fact that the experiment set-up and procedures are nearly identical in all three sections comparisons can be made with little bias.

6.2 Repeating study with a different task

One of the more interesting discoveries of this study (based on the dataset used, and the limitations that come with it) is the suggestion that although previous research, and analysis conducted for this study, suggest that quantity of ideas is strongly correlated with the likelihood of creating a creative idea, this study also discovered that in addition to the sheer number of concepts produced, there were also difference between designers not explained by the number of concepts they created.

In other words, some designers were better performing than others, and it would therefore be interesting to have the same designers from the initial study take part in another experiment, with a slightly different design task to confirm whether or not the high-performing designers from the initial study would still perform at a higher level than the other designers.

6.3 EEG

Part II of the experiment attempted to study what designers were thinking about and focusing on during the experiment through the use of a talk-aloud protocol. However, as the talk-aloud protocol undoubtedly interferes with the thought processes of the designers during the experiment, the use of EEG equipment to measure brain wave activity may provide a more suitable alternative to uncover differences in thought patterns due to the use of different design tools.

7 BIBLIOGRAPHY

Ainsworth, Shaaron, Vaughan Prain, and Russell Tytler. 2011. "Drawing to Learn in Science." *Science* 333(6046):1096–97.

Asiedu, Y. and P. Gu. 1998. "Product Life Cycle Cost Analysis: State of the Art Review." *International Journal of Production Research* 36(4):883–908.

Beaty, Roger E. and Paul J. Silvia. 2012. "Why Do Ideas Get More Creative Across Time? An Executive Interpretation of the Serial Order Effect in Divergent Thinking Tasks." *Psychology of Aesthetics, Creativity, and the Arts* 6(4):309–19.

Ben-Arieh, David and Li Qian. 2003. "Activity-Based Cost Management for Design and Development." *International Journal of Production Economics* 83:169–83.

Bilda, Zafer and Halime Demirkan. 2003. "An Insight on Designers' Sketching Activities in Traditional versus Digital Media." *Design Studies* 24(1):27–50.

Bilda, Zafer, John S. Gero, and Terry Purcell. 2006. "To Sketch or Not to Sketch? That Is the Question." *Design Studies* 27(5):587–613.

Boginski, Vladimir, Sergiy Butenko, and PM Pardalos. 2004. "Matrix-Based Methods for College Football Rankings." *Economics, Management and Optimization in Sports* 1–13.

Buhrmester, M., T. Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6(1):3–5.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46(1):112–30.

Colley, Wesley N. 2002. "Colley's Bias Free College Football Ranking Method: The Colley Matrix Explained." *Princeton University.* Retrieved March 22, 2014 (https://kassiesa.home.xs4all.nl/bert/uefa/files/ranking-colley.pdf).

Corbett, J. and J. R. Crookall. 1986. "Design for Economic Manufacture." *CIRP Annals - Manufacturing Technology* 35(1):93–97.

Cross, Nigel. 1999. "Natural Intelligence in Design." *Design Studies* 20(1):25–39.

Cross, Nigel. 2000. *Engineering Design Methods: Strategies for Product Desgin*. 3rd ed. Wiley.

Dowlatshahi, Shad. 1992. "Product Design in a Concurrent Engineering Environment: An Optimization Approach." *International Journal of Production Research* 30(8):1803–18.

Dubberly, Hugh. 2004. *A Compendium of Models*. Retrieved July 26, 2013 (http://www.dubberly.com/articles/how-do-you-design.html).

Duverlie, P. and J. M. Castelain. 1999. "Cost Estimation during Design Step: Parametric Method versus Case Based Reasoning Method." *International Journal of Advanced Manufacturing Technology* 15(12):895–906.

Eckert, Claudia, Alan Blackwell, Martin Stacey, Christopher Earl, and Luke Church. 2012. "Sketching across Design Domains: Roles and Formalities." *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 26(3):245–66.

Edelman, Jonathan Antonio et al. 2009. "Hidden In Plain Sight: Affordances of Shared Models in Team-Based Design." Pp. 395–406 in *International Conference on Engineering Design, ICED'09, 24 – 27 August*. Stanford University.

Eriksson, Kimmo and Brent Simpson. 2010. "Emotional Reactions to Losing Explain Gender Differences in Entering a Risky Lottery." *Judgment and Decision Making* 5(3):159–63.

Fixson, Sebastian K. and Tucker J. Marion. 2012. "Back-Loading: A Potential Side Effect of Employing Digital Design Tools in New Product Development." *Journal of product innovation management* 29(S1):140–56.

Garvin, David a. 1984. "What Does 'Product Quality' Really Mean?" *Sloan Management Review* 26(1):25–43.

Gerber, Elizabeth. 2009. "Prototyping: Facing Uncertainty Through Small Wins." Pp. 333–42 in *International Conference on Engineering Design, ICED '09, 24-27 August*. Stanford University.

Goel, Vinod. 1995. *Sketches of Thought*. Cambridge, MA, USA: MIT Press.

Gosling, Samuel D., Simine Vazire, Sanjay Srivastava, and Oliver P. John. 2004. "Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires." *American Psychologist* 59(2):93–104.

Govan, A. Y., A. N. Langville, and C. D. Meyer. 2009. "Offense-Defense Approach to Ranking Team Sports." *Journal of Quantitative Analysis in Sports* 5(1):1151.

Haahr, Mads. n.d. "Random.org." Retrieved March 10, 2015 (https://www.random.org/).

Häggman, Anders, Geoff Tsai, Catherine Elsen, Tomonori Honda, and Maria C. Yang. 2015. "Connections Between the Design Tool, Design Attributes, and User Preferences in Early Stage Design." *Journal of Mechanical Design* 137(7):71408.

Horton, John and Lydia Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." in *Proceedings of the 11th ACM conference on Electronic commerce*. Retrieved September 18, 2015 (http://arxiv.org/abs/1001.0627).

Houde, Stephanie and Charles Hill. 1997. "What Do Prototypes Prototype?" Pp. 367–81 in *Handbook of Human-Computer Interaction*, edited by M. Helander, T. K. Landauer, and P. Prabhu. Elsevier Science B.V.

Huthwaite, Bart. 1988. "Designing in Quality." *Quality* 27(11):34–35.

Jonson, Ben. 2005. "Design Ideation: The Conceptual Sketch in the Digital Age." *Design Studies* 26(6):613–24.

Kelley, Tom and Jonathan Littman. 2005. *The Ten Faces of Innovation: IDEO's Strategies for Defeating the Devil's Advocate and Driving Creativity throughout Your Organization*. 1st ed. Currency / Doubleday.

Krause, Jean C. and Louis D. Braida. 2002. "Investigating Alternative Forms of Clear Speech: The Effects of Speaking Rate and Speaking Mode on Intelligibility." *The Journal of the Acoustical Society of America* 112(5 Pt 1):2165–72.

Kudrowitz, Barry Matthew, Paula Te, and David Wallace. 2012. "The Influence of Sketch Quality on Perception of Product-Idea Creativity." *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 26(3):267–79.

Linsey, J. S. et al. 2010. "A Study of Design Fixation, Its Mitigation and Perception in Engineering Design Faculty." *Journal of Mechanical Design* 132(4):41003.

Macomber, Bryan and Maria C. Yang. 2011. "The Role of Sketch Finish and Style in User Responses To Early Stage Design Concepts." Pp. 1–10 in *Proceedings of the 2011 ASME International Design Engineering Technical Conferences & Information in Engineering Conference, August 28-31*. Washington D.C.

Mason, Winter and Duncan J. Watts. 2010. "Financial Incentives and The 'performance of Crowds.'" *ACM SIGKDD Explorations Newsletter* 11(2):100.

Meinel, Christoph, Larry Leifer, and Hasso Plattner, eds. 2011. *Design Thinking: Understand - Improve - Apply*. Berlin, Heidelberg: Springer.

Newman, Damian. n.d. "The Design Squiggle - Central." Retrieved August 2, 2013 (http://cargocollective.com/central/The-Design-Squiggle).

Pahl, Gerhard and Wolfgang Beitz. 1996. *Engineering Design: A Systematic Approach*. 2nd ed. edited by K. Wallace. London, UK: Springer-Verlag.

Pan, Rui, Shih Ping Kuo, and Johannes Strobel. 2013. "Interplay of Computer and Paper-Based Sketching in Graphic Design." *International Journal of Technology and Design Education* 23(3):785–802.

Paolacci, Gabriele, Jesse Chandler, and Pg Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision making* 5(5):411–19.

Rams, Dieter. n.d. "Good Design." Retrieved June 28, 2016 (https://www.vitsoe.com/gb/about/good-design).

Robertson, B. F. and D. F. Radcliffe. 2009. "Impact of CAD Tools on Creative Problem Solving in Engineering Design." *Computer-Aided Design* 41(3):136–46.

Römer, Anne, Sven Leinert, and Pierre Sachse. 2000. "External Support of Problem Analysis in Design Problem Solving." *Research in Engineering Design* 12(3):144–51.

Ross, Joel, Lilly Irani, M.Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. "Who Are the Crowdworkers?" P. 2863 in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10.* New York, New York, USA: ACM Press.

Schrage, Michael. 1999. *Serious Play: How the World's Best Companies Simulate to Innovate.* Boston: Harvard Business School Press.

Schütze, Martina, Pierre Sachse, and Anne Römer. 2003. "Support Value of Sketching in the Design Process." *Research in Engineering Design* 14(2):89–97.

Simon, Herbert A. 1973. "The Structure of Ill Structured Problems." *Artificial Intelligence* 4(1973):181–201.

Tseng, Winger S. W. and Linden J. Ball. 2011. "How Uncertainty Helps Sketch Interpretation in a Design Task." Pp. 257–64 in *Design Creativity 2010.* Springer.

Tversky, Barbara. 2011. "Visualizing Thought." *Topics in Cognitive Science* 3(3):499–535.

Ullman, David G., Stephen Wood, and David Craig. 1990. "The Importance of Drawing in the Mechanical Design Process." *Computers & Graphics* 14(2):263–74.

Ulrich, K. T. and S. D. Eppinger. 2003. *Product Design and Development.* Tata McGraw-Hill Education.

Ulrich, Karl T. and S. Pearson. 1998. "Assessing the Importance of Design Through Product Archaeology." *Management Science* 44(3):352–69.

Veisz, David, Essam Z. Namouz, Shraddha Joshi, and Joshua D. Summers. 2012. "Computer-Aided Design versus Sketching: An Exploratory Case Study." *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 26(3):317–35.

Visser, Willemien. 2006. *The Cognitive Artifacts of Designing.* Lawrence Erlbaum Associates. Retrieved (https://hal.inria.fr/inria-00526069).

Whitney, Daniel E. 1988. "Manufacturing by Design." *Harvard business review* 83–92.

Wong, Linda. 2015. *Essential Study Skills.* 8th ed. Stamford, CT, USA: Cengage Learning.

Wu, M. C. 1998. "Justification of Concurrent Engineering Environments Based on Fuzzy Mathematics." *International Journal of Production Research* 36(7):2025–41.

Yang, Maria C. 2005. "A Study of Prototypes, Design Activity, and Design Outcome." *Design Studies* 26(6):649–69.