The evaluation of creative ideas

– analysing the differences between expert and novice judges

Judit Pétervári

Thesis submitted in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

November 2017

Biological and Experimental Psychology

School of Biological and Chemical Sciences

Queen Mary University of London

STATEMENT OF ORIGINALITY

I, Judit Pétervári, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below, and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

ABSTRACT

The evaluation of creative ideas is a special case of judgment and decision making. It is difficult to objectively evaluate creative products because most people possess an internalised model of creativity which is usually neither verbalised nor explicitly defined. Also, one of the main assessment dimensions of creativity, originality varies as a function of the evaluator's previous experience. For these reasons, previous research has provided practical rather than theoretical grounds for studying the evaluation process. The present thesis examines the conceptual basis on which people evaluate creative ideas. The aim is to identify factors and conditions which enhance the detection of creative ideas.

A novel paradigm was created to test how creativity-related features influence the assessment of creativity. In six experiments, experts' and non-experts' judgment was examined regarding urban design. Two experiments established the expert ratings of the stimuli. Further two experiments explored the extent to which non-experts relied on four features (originality, utility, scalability, and riskiness) for judging the creativity of novel project ideas while the level of motivation was controlled. Overall, the findings show that non-experts' creativity judgment relied on all four characteristic features. Their ratings of the features predicted a substantial part of the variance in the creativity ratings. In another experiment, the effect of providing explicit task-related information was tested. Such information did not make a solid difference in the creativity ratings. A final experiment assessed the differences between making relative and absolute judgments about creativity. There was a large overlap between the selection of best and worst ideas regardless of which way the judgment was made.

In conclusion, non-experts were found to possess a robust internal model of creativity and not to make random choices. Experts and non-experts were found to judge creativity vastly differently, they only agreed that utility is the most important criterion.

ACKNOWLEDGMENTS

First, I wish to thank my supervisors - Magda Osman and Joy Bhattacharya, I am forever grateful for your great advice and openness, for always having my back and trusting me; altogether, for your precious support and stimulating suggestions during this entire research journey. I also would like to thank Aimee Bright, with whom I started this journey and who taught me many lessons. Further, I am indebted to Amory Danek, who helps my career from the first moment on.

I am really happy that I could discuss research and all the related and unrelated issues with the QMUL PhD students, including my fellow lab members from the Dynamic Learning & Decision Making Laboratory. I appreciate our lengthy chats and the many times you saved the cupcakes for me. I am also thanking the London CogSci crew and the QMUL Thinking Writing team for helping me through some professional impasses. Elisa Piccaro, you were the best Research Services Officer I could ever have. Thanks to all participants who contributed to the research presented here.

I am also very grateful to my loved ones, especially to my family members, Anya, Apa, Jancsi & Csilla; and to my lovely flatmates for tiptoeing around the house when I had stressful times. You gave me food and love when I needed you. My dear friends, Anja, Dalma, Dia, Dius, Kriszti, Nóri, Pepe & Philipp, thank you for keeping me motivated and giving me space when I asked you. I was also helped and inspired by many great people, Bence, Tomi, Krisztián, Josh, Joe, & Marci - thank you for sending me interesting papers, suggesting edits, and for the strategic stat meetings. I owe you all a lot.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

A REVIEW OF THE THEORETICAL APPROACHES REGARDING THE EVALUATION OF CREATIVE IDEAS

1.1 Preface

One of the leading creativity scholars, Mihály Csíkszentmihályi notes with a bit of frustration that "psychologists tend to see creativity exclusively as a mental process" (1999, p. 313). He does not think that this approach would do justice if one set out to solve the big questions (such as what is creativity and where does it come from?). To justify his statement, he explains one of the biggest Aha! moments of his decade-long career. It was a pivotal point for him, when, after years of ineffectual attempts of understanding creativity as process situated within the mind, he finally realised that researchers need to look beyond the level of the individual to understand how creativity happens and zoom out to see the bigger picture. That is how he was inspired to come up with a systems approach for the study of creativity (Csíkszentmihályi, 1988, 2014), which highlights the need to analyse the interaction between the creator and the audience of the creative work and also takes the many social factors which determine how creative endeavours turn out in real life into account.

While the dynamic, intersubjective nature of creativity should not be neglected, the position this project takes is that both high- and low-level approaches should be used to answer the big questions stated at the start. Keeping the discourse on a high, systematic level provides us only with an approximate understanding about the exact processes, therefore, in the present project, intrapersonal aspects of the creative process are investigated to gain a more fine-grained picture about how evaluations are made. The present thesis is focused on exploring factors which influence how judges of creativity react, and the experiments were aimed to find manipulations which could help to make more rigorous and consistent judgments of creativity.

As it turns out from this short prologue already, conceptualising creativity is key to understanding its underlying components. This chapter gives an overview of the theoretical approaches to the study of creativity, whereas the second chapter reviews the wide range of assessment methods regarding the evaluation of creative products. Below, the major research trends in creativity research are summarised, then the essential definitions from judgment research are clarified. The chapter also outlines how these two fields can be combined and what kind of framework is necessary to build their union.

In the next section, I am introducing the field by presenting the most relevant theories associated with the thesis research. An account of the creativity research trends is provided by summarising the cognitive, economic, and problem solving & expertise-based types of theories about creativity.

1.2 An Account of Creativity Research Trends

There are many theories available on creativity, however, the ones which provide overarching explanations about what creativity is are difficult to test empirically (e.g., personal and societal level of creativity, Csíkszentmihályi, 1998; Kaufman & Beghetto, 2009; Simonton, 2013; or domain-specificity, Baer, 2015; Plucker & Beghetto, 2004) and those which are suitable for experimental investigations are addressing the sub-questions, not the big questions of creativity. A few examples of the big questions are the following (based on Mayer, 1999): what is creativity (how should we define it)? Is creativity a personal or a social phenomenon? Is creativity common or rare? Is creativity domain-general or domain-specific? Is creativity quantitative or qualitative? These questions are all addressed by the research community but due to a lack of integration of theories, there is very little consensus regarding the answers. The "let a hundred flowers bloom" approach provides scholars with a wide array of theoretical models and pragmatic techniques for investigations, however, due to the mixed methods, the comparability of the empirical studies remains low and even if great meta-analyses are

produced, their impact on the selection of research methods is limited (cf. the guidelines provided by Dietrich & Kanso, 2010; Long, 2014).

As Kozbelt and colleagues (2010, p. 21) put it:

> "The variation is compounded by the fact that creativity involved a multitude of definitions, conceptualizations, domains, disciplines that bear on its study, empirical methods, and levels of analysis, as well as research orientations that are both basic and applied - and applied in varied contexts."

This quote was the conclusion of an extensive sorting job, in which the numerous available theories were grouped into ten categories: *developmental, psychometric, economic, stage and componential process, cognitive, problem solving and expertise-based, problem finding, evolutionary, typological and systems theories of creativity*. While *Chapter 2* appraises the psychometric approach to creativity, here I discuss the basics of the other three related categories. Namely, the *cognitive*, the *economic*, and the *problem solving & expertise-based theories of creativity*.

1.2.1 Cognitive Theories of Creativity. Cognitive theories of creativity address the big question of whether creative cognition is an extraordinary functioning of the human mind or creative products are results of a *business-as-usual* functioning. The consensus in psychology is that creative responses result from the same cognitive structures as non-creative responses, i.e., creativity does not require largely distinct processes as non-creative thinking does (e.g., Perkins, 1981). The central difference between the creative and non-creative functioning is not that creative people do have certain structures which non-creative people do not; it is rather that creative people use their mind to ignore some commonly used knowledge base while they access some not typically used knowledge base (Ward, 1994). The cognitive processes used for creativity, such as metaphorical thinking, the use of analogies, imagery, conceptual combination, or conceptual expansion, can be used in non-creative endeavours too (Ward, 1994; Ward, Smith, & Finke, 1999). This theory is called the ‘creative cognition approach’ (Finke, Ward, & Smith, 1992; Smith, Ward, & Finke, 1995), and changed the way how scholars

thought about creativity in ‘90s. Since the approach is chiefly concerned with how problem solvers generate ideas and then explore the implications of these responses, and since these two processes are intertwined, Finke and colleagues (1992) came up with the geneplore model of creativity (generate + explore).

Regarding the underlying cognitive mechanisms of creativity, the same big question can be raised: are these mechanisms of normal or extraordinary kind? If we think of the mind as a computer, the principle of parsimony dictates that creative products are the results of the same hardware and code language as non-creative products. What might be different is the script which gets executed and the programming of these scripts, which is the creative effort of the person. As the cognitive background of this, two antithetical types of thinking, convergent and divergent thinking (Guilford, 1956, 1967) are speculated to underlie the creative problem-solving process. It has been proposed that problem solvers use convergent thinking for selecting a single (best) solution in response to a well-defined problem by applying standard procedures to existing knowledge. By contrast, divergent thinking can be utilized in more ambiguous situations, where a range of alternative solutions are possible, therefore responses may vary individually (Cropley, 2006). The popularity of the concept of divergent thinking has meant that for some it has been translated into a measurement tool of creativity itself (Zeng, Proctor, & Salvendy, 2011; Kaufman and Baer, 2012); though this approach has been severely criticized (e.g., Dietrich, 2007; Piffer, 2012). Among others, Cropley (2006) reset the balance by noting that both convergent and divergent thinking are necessary for producing creative ideas and that it is not contingent on divergent thinking alone.

Divergent thinking, along with high ideational fluency, is often equated with creative cognition (Policastro & Gardner, 1999), and all humans are capable of these, even if to a varying degree. The model of generative cognitive style (Policastro & Gardner, 1999) includes three components, about which the same can be said. For creativity to happen, imagination, a sense

of domain relevance, and intrapersonal intelligence are required. In the words of the authors (p. 217), "imagination leads to originality, sense of domain relevance leads to high quality, and intrapersonal intelligence checks illusory and/or emotional inferences in the process of constructing a novel but appropriate representation".

Although these properties, such as intelligence, are rather dispositional than situational, creativity trainings can also rely on explicit strategies which advocate creative thinking. Such guidelines are usually related to metacognition and include tactics such as '*what if the opposite would happen what you expect', 'identify and question your assumptions*', *'turn the problem upside down', 'shift your attention to something else and incubate on the problem'* (Kozbelt, Beghetto, & Runco, 2010). Metacognition is an introspective process in which the individual can reflect on and control his or her cognitive performance. There are two main components: the knowledge about cognition and the regulation of cognition. The regulation strategies involve voluntary cognitive shifting or changing the zoom at which the problem is considered. Employing effortful planning and control of thinking and memory processes can be used to boost the generation of creative responses. E.g., divergent thinking is influenced by metacognitive strategies since a strenuous search is conducted to find original responses, where the strategy is to ignore all the trivial answers. Thinking about thinking fosters self-reflection and the conceptualization of abstract terms. This technique is often utilized for creating mind maps.

1.2.2 Economic Theories of Creativity. There are several principles of economics, which can be adapted to creativity research. The advantage of using economics for psychological research is that clear-cut hypotheses can be tested (Kozbelt, Beghetto, & Runco, 2010). Here, I will shortly discuss predictions based on investment strategies. One example of explaining creativity-related phenomena with economical concepts concerns experts and posits that the reason why experts are less susceptible to shifting to disruptive methods is that doing

so would increase their costs and would grant a smaller return of their previous investment (cost of education, working hours associated put towards the reigning paradigm) at the same time (Sternberg & Lubart, 1991). Another example of the economic view of creativity is that incentives associated with certain behaviours reinforce them, while costs associated with other behaviours reduces the chance of exhibiting them. If standing out from a group comes with the cost of becoming labelled or stigmatised, benefit of showcasing an original response will be outweighed by the potential cost if the size of the group will be large in contrast to a scenario in which the size of the group is small. Amabile's (1996) alternative explanation of this phenomenon is that by a highly original idea, uncertainty grows too and there can be reasonable doubts whether the idea will be feasible and implementable eventually. In fact, standing by to originality comes with a potential of risk (Rubenson & Runco, 1995).

A final example of adapting concepts from economy to creativity research is provided by Sternberg & Lubart (1992, 1995). They introduced the investment theory of creativity, according to which creative individuals are trying to buy low and sell high on ideas, and use intellectual processes, knowledge, intellectual styles, personality, motivation, and environmental context to do so. Similarly to investment making, the resources of a creative person are scarce and must be allocated wisely. Discernment abilities are crucial for picking the best ideas to pursue (Silvia, 2008), and thereby making the investment with the highest potential return.

1.2.3. Problem-Solving and Expertise-based Theories of Creativity. Both for the present conceptualisation and operalisation of creativity, it is useful to build on the problem-solving literature, as it can be interpreted as a wider framework for creativity which involves the solving process of both ill-defined and well-defined problems. Both by creativity and problem-solving, information is coordinated toward reaching a specific goal (Wiggins and Bhattacharya, 2014). Concepts from Newell & Simon (1972), such as the stages of the

problem-solving process (initial state, operations state, goal state), or problem representation, problem space, operator, as well as solving strategies, can be applied towards tackling the big questions regarding creativity (e.g., Langley, 1987). The use of these terms enables a precise discussion of how creativity-related phenomena, such as an insight is triggered in the mind (e.g., Ohlsson, 1992). Creative solutions are provoked by ill-defined problems (e.g., Pétervári, Osman, & Bhattacharya, 2016), and creativity is often assessed by providing participants with difficult problems (more details about this will be provided in *Chapter 2*), for the solution of which strategy use and the level of expertise can be important.

Expertise is defined as a collection of experiences acquired during a longer time span, during which the person engages in deliberate practice in their chosen domain. There is even a ten-year rule of thumb in the literature as it has been shown that a decade is the minimum time investment with which a person can become a domain-specific expert (Bloom, 1985; Chase & Simon, 1973; Ericsson, 1999; Ericson, Roring, Nandagopal, 2007; Hayes, 1989; Kozbelt, 2005; Newell, Shaw, & Simon, 1962; Simonton, 1991, 1997). More details about why the knowledge of a domain is crucial for creativity will be provided in *Chapter 2* and *3*. Here, I focus on the role of expertise in idea evaluation as these aspects are the most relevant for the purposes of the thesis.

Expertise was articulated as a requirement for successful evaluation of creativity by many scholars (Cropley & Kaufman, 2013; Kaufman & Baer, 2012; Kaufman, Baer, Cropley, Reiter-Palmon, & Sinnett, 2013). However, empirical studies do not reveal which part of expertise makes experts' judgment more accurate than of lay people's. In creativity studies, the case is usually that experts and non-experts need to evaluate creative products, which can be anything from mousetrap designs to poems (e.g., Kaufman & Baer, 2012; Kaufman, Baer, Cropley, Reiter-Palmon, & Sinnett, 2013) and the difference in their evaluations is deduced to be the result of the difference in expertise. Novices are found to rate in accordance with the

experts for products they are familiar with/frequent user of, e.g. mobile phone holders by Haller and colleagues (2011), and mildly or not at all in accordance with experts in case of products which are complex (Galati, 2015) and come from a highly structured, codified domain (Kaufman et al., 2013; Simonton, 2009).

When we look for differences between experts and non-experts in the literature, we find that expertise is more than a large amount of information collected from a specific domain: in the present research, novices educated about what constitutes a creative product could not substitute experts (e.g., Storme, Myszkowski, Çelik, & Lubart, 2014). Experts are theorised to outperform lay judges of creativity because their accumulated experiences grant them multiple benefits (e.g., Bonnardel, & Marmèche, 2004; Turnbull, Littlejohn, & Allan, 2010). They can operate with more information at the same time as their chunking span of domain-relevant information is higher than non-experts' and can also remember these chunks more accurately than non-experts do; experts also possess a more detailed mental model about creativity and therefore can construct a more accurate problem representation than what lay people are capable of (Ericson & Charness, 1994). Additionally, experts approach a problem differently from naïve people when it comes to the evaluation of creative ideas. They use different problem-decomposing strategies (Ho, 2001) and can detect more information sources than novices (Björklund, 2013). They also have alternative internalized standards of what is creative (Kaufman, Baer, Cole, & Sexton, 2008), including a fine-grained consideration of more sophisticated criteria compared to judges with no expertise (Kaufman et al., 2013). Finally, they can adapt to both the availability and the lack of decision criteria (Bettman & Sujan, 1987), and their internalized standards make them weigh relevant criteria differently than novices do (Silvia, 2013).

From the summary of the essential research trends for conducting this research, we learned that creative responses are the results of the same cognitive structures and mechanisms

as non-creative responses (e.g., convergent and divergent thinking). Due to the tendency of trying to keep the costs low and the benefits high, creative people must allocate their resources wisely; also, the judges of creative products must focus on the projects bringing the highest return on the resources invested into them. Another takeaway from previous research is that the problem-solving literature can be applied well to discuss creativity as a process. Finally, expertise was articulated as a requirement for successful evaluation of creativity, however, it is unclear what aspect of expertise is enabling professionals to make better evaluations than naïve observers. Is it only their knowledge of the domain or are there conditions under which lay people could perform just as well as they do? We will look for further information to answer this question.

As for the next step, a more nuanced explanation will be given about what constitutes creativity and what does not. The next section outlines the various definitions of creativity and what the conceptualisation of creativity could be based on them.

1.3 Definition(s) of Creativity

To achieve a scientific measurement of creativity, working with precisely defined terms is key. One of the biggest problems with the science of creativity is that it has been problematic to outline a clear definition of creativity for a long time. Fortunately, this problem seems to get resolved recently, the research community is getting settled on the few, consensually accepted definitions. For the purposes of this thesis, the standard definition of creativity (Runco & Jaeger, 2012) was interpreted for evaluation purposes: a product is creative if it is judged to be highly original and useful/effective by observers. These two criteria were complemented with two additional ones, low risk and high scalability, which were appropriate for the task domain but had a much smaller weight toward overall creativity.

Historically, there are many different definitions of creativity, but this chapter will only give a short overview of the definitions. Two kinds of definitions can be differentiated: one of

them provides a short, dense definition and focuses solely on the crucial criteria for creativity, while the other kind of definitions are attempting to give a full account of creativity and include its socially determined nature by describing the contextual factors too.

Regarding the first kind, Runco & Jaeger's definition (2012, p. 92) is the one accepted as "the standard definition of creativity": “Creativity requires both originality and effectiveness”. This definition is comprised of two components, similarly of Bruner's intertwined definition regarding creativity as an "effective surprise" (1962, p. 18). However, other componential definitions of creativity are built up from three different components: the additional one is linked to the interaction with the observer of the product, i.e., it is highlighted that the product should trigger a surprise or should be non-obvious (Simonton, 2016). Examples of the three-component definitions are Simonton's (2012), who posits that a creative product is new, useful, and nonobvious; Kaufman & Sternberg's (2010) who argue that a creative product is novel, good, and relevant; or Boden's (2004) who describes a creative product as novel, valuable, and surprising.

The more detailed definitions of creativity describe the context in which creativity is embedded as well as the relations between the contextual factors. For instance, Plucker et al. (2004) define creativity as “the interaction among aptitude, process, and environment by which an individual or group produces a perceptible product that is both novel and useful as defined within a social context” (p. 90). Amabile's (1982, p. 1001) seminal definition goes around the criteria and describes the process in which creativity is recognised: "A product or response is creative to the extent that appropriate observers independently agree it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated. Thus, creativity can be regarded as the quality of products or responses judged to be creative by appropriate observers, and it can also be regarded as the process by which something so judged is produced". Later (1996, p. 35), she provided a slightly different

definition: “A product or response will be judged as creative to the extent that (a) it is both a novel and appropriate, useful, correct or valuable response to the task at hand, and (b) the task is heuristic, rather than algorithmic. By definition, algorithmic tasks have a clearly identified goal, but heuristic tasks might or might not have a clearly identified goal; the important distinction is that, for heuristic tasks, the path to the solution is not completely straightforward.” On a similar account, Csíkszentmihályi's (1996, p. 28) definition of creativity also details the system in which creativity unfolds rather than the properties which a creative product must possess: "Creativity is any act, idea or product that changes an existing domain or that transforms an existing domain into a new one" and it cannot happen "without the explicit or implicit consent of a field responsible for it".

To sum up, a maximally creative product is as high on originality, utility, and surprise as possible and the creator is producing the work in a cultural and social system which influences the outcome. We must also note that creative products are created on a large spectrum, which results from the many possible combinations of the constituting dimensions and the lack of cut-off criterion. What is more, the process of creation is dynamic and a product which might not look creative at the first glance can end up as an outstanding product through the multiple iterations characteristic of many creative activities (e.g., writing poems, lyrics, painting pictures, etc.)

To address which categories exist between fully creative and not creative at all, and to cover the entire spectrum of creative products, Simonton (2016) has developed a comprehensive framework of creative outcomes, supported by his three-component conceptualisation of creativity (Table 1). His mathematical formula is built up using three different components. The first one is initial probability. Unoriginal ideas seem highly probable at the beginning, while original ideas seem highly improbable initially. The second one is final utility which denotes the effectiveness or appropriateness of the creative product. The third one

is called prior knowledge and regards the creator's awareness of the utility of the product. The higher this value is, the stronger the "hunch" or the feeling of knowing is about what the outcome will be. Conversely, a low level of prior knowledge means a low level of anticipation and can be conceptualised as surprise.

Table 1

Creative and non-creative outcomes according to the three-criterion definition.

Initial probability	**Final utility**	**Prior knowledge**	**Outcome**
1	1	1	Routine, reproductive, or habitual ideas or responses
0	1	0	Creative ideas or responses (*c* -> 1)
1	1	0	Fortuitous response bias (e.g., "lucky guesses")
0	0	1	Rational suppression (e.g., extinguished responses)
1	0	0	Problem finding (surprising expectation violation)
0	1	1	Irrational suppression
1	0	1	Irrational perseveration
0	0	0	Blissful ignorance

Note. This framework is based on the following equation: $c = (1 - p) * u * (1 - v)$, where p is the initial probability, u is the final utility, and v is the prior knowledge of that utility. This table was adapted from "Defining Creativity: Don't We Also Need to Define What is *Not* Creative?" by D. K. Simonton, 2016, *The Journal of Creative Behavior, n/a–n/a.* Copyright 2016 by the Creative Education Foundation.

In this framework, the three-criterion multiplicative definition of personal creativity is outlined as a mathematical formula: $c = (1 - p) * u * (1 - v)$, where c is creativity, p is the initial probability, u is the final utility, and v is the prior knowledge of that utility (p. 11). By articulating the necessary and sufficient conditions of creativity, he resolves the earlier dilemma of whether a product with no originality or no utility could be deemed as creative (Diedrich,

Benedek, Jauk, & Neubauer, 2015; Weisberg, 2015). His resolution of the issue is that if either originality or utility drops to zero, then the product or response cannot be called creative.

What we can see from the diversity of creativity definitions is that although people have an internalised model of creativity, this is usually not verbalised and therefore lacks an explicit definition. Even for professionals, it took many years to present an all-inclusive conceptualisation of creativity and there was a lengthy debate about the boundary conditions, since it is not clear where creativity ends, and *non-creativity* begins. In the present thesis, the focus was not on finding the boundaries of creativity but on identifying features associated with it. The main assumption underlying our creativity definition was that observer's perception of creativity is linked to their perception of the product's originality, utility/effectiveness, riskiness, and scalability. However, the latter two factors were complimentary and were assumed to carry less weight toward forming creativity judgments than the first two factors.

After specifying the definition of creativity, let us discuss an overarching framework for the psychological research on creativity. This theoretical framework is called the 4 Ps of Creativity (Rhodes, 1961; Runco, 2004), and is comprised of a systematic categorisation of the aspects of creativity. The 4 Ps stand for *process*, *product*, *person*, and *place*, which was subsequently expanded to the 6 Ps of Creativity (Simonton, 1990; Runco, 2003), by adding *persuasion* and *potential* to the mix. The thesis research is focused on the creative process and its outcome, the creative product. Therefore, these two categories will be detailed below.

1.4 Process and Product Theories of Creativity

Two of the Ps, process and the product theories will be summarised below as these are the subject of investigation in the present thesis.

1.4.1 Process theories. There are various ways of conceptualising the creative process, but most theorists assert that the creative process consists of several consecutive stages (e.g., blind variation and selective retention model, Campbell, 1960; associative hierarchy theory,

Mednick, 1962; three-process theory of creativity, Davidson and Sternberg, 1986; geneplore model, Finke et al., 1992). The number of stages differs by theory, and this is largely dependent on the ways in which theorists describe the critical components of the stages (e.g., preparation, incubation, illumination, and verification by Wallas, 1926; whereas problem formulation, preparation, idea generation, idea evaluation, and idea selection by Amabile, 1983). However, regardless of these variations, researchers agree on two main essential operations of the creative problem-solving process: (1) the generation of ideas and (2) the evaluation and selection of (an) appropriate outcome(s) (e.g., Finke et al., 1992; Lubart, 2001; Reiter-Palmon and Illies, 2004).

Given that these two stages are common to all theories of creativity, and are relatively uncontroversial, it is for these reasons that most empirical studies focus on these two stages as central to the creative problem-solving process.

However, it is worth noting that there are far more empirical investigations conducted on creative idea generation than on idea evaluation (Amabile and Müller, 2008; Rietzschel, Nijstad, & Stroebe, 2010; Silvia, 2008). Separating the two phases is necessary to disentangle the distinct cognitive processes applied in them, however, both idea generation and evaluation are critical for shaping the creative product of the creative process, and the two stages are tightly linked (neither makes sense without the other). Further, we must note that the creative process is a dynamic one which can involve several iterations of generation and evaluation of ideas that a problem solver goes through before reaching an end state (Lonergan, Scott, & Mumford, 2004; Kozbelt and Durmysheva, 2007).

1.4.2 Product theories. Shifting now to the product theories of creativity, one of the widely accepted frameworks is the so-called propulsion model of creativity; this theory identifies eight types of creative products as propulsions in a conceptual space (Sternberg, Kaufman, & Pretz, 2002). The least significant creative contributions can be labelled as (1) replications or (2) redefinitions of established products. They are necessary starting points but

do not move the field forward. On the other hand, products which serve as (3) forward incrementation and (4) advance forward incrementation can move the field forward in the direction which it is already going towards. What is more, with (5) redirection, (6) reconstruction/redirection and (7) re-initiation, it is possible to start moving the field to a new direction, to return from a new direction to an earlier one, or to shift to a new starting point and initiate movement from there, respectively. Finally, (8) synthesis is about combining and merging the essence of different paradigms into a new one. Using this theory, the contribution of each creative product can be accounted for. Focusing on creative products yields the strongest evidence-based approach on this field. Next, I will continue listing the merits of using a product-based approach for creativity research. Subsequently, the disadvantages will be discussed too.

On the plus side, creative products can be quantified and are tangible, even if the processes leading to such outcomes are yet ineffable. The biographical and historiometric approaches to creativity (i.e., studying the chronicle of human lives to see how significant creators produced their work) can make a good use of archival data and make it possible to reconstruct how geniuses' career and personality evolved during their lifetime (e.g., Simonton, 1998; Kozbelt, 2007). Importantly, the topic of this thesis is evaluation and how different groups of people make judgments about creativity, and dividing creators from their products provides a clean state for assessment. This approach was chosen because outside the lab, creative products are often competing for funds and recognition independently from their creators. Since the products are the means by which creativity receives recognition, we focus our attention on understanding by which mechanisms a product can achieve that. One caveat for researchers here is that products are not of interest by themselves, it is their interaction with the audience by which they are found to be creative (Csíkszentmihályi, 1996, 1999). Thus, when studying the products, it is really that the judges who evaluate the products must be drawn

under scrutiny. It is arguable whether the accuracy of creativity judgments can be interpreted (e.g., Silvia, 2008), thus, usually the agreement in raters is assessed (*Chapter 3* discusses this issue in more detail). Taking the products as the central aspect of creativity lends itself as a good angle for the study of evaluation.

The other side of the coin is that treating products as separate entities from their creators is rather problematic when creative idea generation is discussed. In that scenario, the division between the person and the product might entail the problem of making reverse inferences (cf. Hutzler, 2004; Poldrack, 2006), since the products are used to gain information about their creators. Although the assumption is that highly creative products are constructed by highly creative people, establishing this link does not yield any predictive power for the quality of the next creation of the same person.

After this short introduction about the field, now we turn to the particular topic of this thesis, which is how people evaluate creativity. The next section recaps what evaluation is and what cognitive mechanisms are linked to the assessment of ideas. The review of the literature will bring us closer to understanding how lay people and experts inform their judgments made about creativity.

1.5 Evaluating Creativity

Evaluation can be defined as the convergent phase after generating ideas (Basadur, 1995), when implementation, rejection, or revision can take place (Mumford, Lonergan, & Scott, 2002). This section provides a summary of the cognitive mechanisms linked to the assessment of ideas. Discussing these mechanisms in detail is necessary because despite the theories of creativity including the evaluation of ideas as a critical component (Finke, Ward, & Smith, 1992; Simonton, 1999; Sternberg, 2006), we know more about the errors in creative thought (Mumford, Blair, Dailey, Leritz, & Osburn, 2006) than about the process how these judgments are built.

Lay people are inexperienced judges of creativity and the goal of the researchers is usually to identify manipulations which can make them more similar to experts. First, I discuss the spontaneous behaviour of lay people in terms of the criteria they use for evaluation and the cognitive biases which might impair their performance. Subsequently, studies trying to bridge the gap between experts and non-experts are explored.

As it is not clear what lay people use as criterion for judging creativity, there are multiple studies investigating the effect of external manipulations on the internal criteria. Rietzschel, Nijstad, & Stroebe (2010) investigated whether providing explicit selection criteria can improve selection efficacy (this was also tested empirically in *Chapter 6*), whether participants choose either desirable and feasible or original ideas spontaneously, and whether processing the ideas on a deeper level correlates with improved selection performance. No evidence was found for a linear relationship between the level of idea processing and selection performance. Those who were instructed to select creative ideas rather used originality for the basis of their selection, as opposed to those who had to pick the best ideas: then participants preferred the ideas they found implementable. Overall, usefulness and the potential impact the idea had overrode the need of originality. In a different experiment, prompting participants to use a broad problem scope, as opposed to a narrow problem scope, facilitated to generate more creative ideas but not to select more creative ones (Rietzschel, Nijstad, & Stroebe, 2014). Moreover, it was found that people thinking with a more abstract mindset rated ideas as more creative than those with a more concrete mindset (Müller, Wakslak, & Krishnan, 2014). Inducing promotion focus as opposed to prevention focus during the assessment of creativity led participants to more accurately rate the originality of their own responses (Herman & Reiter-Palmon, 2011). The use of criteria was assessed also by creative products in science tasks (Long, 2014). The analysis of qualitative interviews showed that judges used appropriateness, novelty, thoughtfulness, interestingness, and cleverness as their assessment criteria. Notably,

these were not independent factors and their use varied by task type and by judge groups which employed them.

Finding different results in different tasks served as another evidence for the domain-specific nature of creativity and motivated researchers to investigate the influence of tasks on the judgments made about creativity. Runco, Illies, & Eisenman (2005) showed that realistic creativity tasks elicited more appropriate solution ideas than unrealistic ones, whereas unrealistic tasks provoked more original and flexible ideas.

As for the common thinking biases associated with the evaluation of creative ideas, lay people were found to be less accurate in evaluating the originality of exceptional as compared to less novel ideas. This difference was found to be a function of the complexity of the setting at the time of idea development (Licuanan, Dailey, & Mumford, 2007). However, it was also shown that actively focusing on originality and the appraisal of interactional processes led to reduced errors in undergraduates' evaluations (as compared to the baseline expert judgments). Lay judges of creativity also advocated easily understandable, widely usable ideas which were in alignment with the the social norms, as opposed to risky, time consuming, and original ideas (Blair & Mumford, 2007). It was not only that people performed sub-optimally in assessing the ideas themselves but they also underestimated the resources needed for implementing the creative ideas presented to them (Dailey & Mumford, 2006). In a comprehensive summary (Mumford et al., 2006), multiple possible sources of biases, such as pre-existing knowledge, limitations in processing capacity, patterns of information use, and the strategies applied in process execution, have been classified as contributors to the evaluation process.

Turning now to the role of expertise in the evaluation of creative ideas, expertise is often defined as a domain-specific, specialized knowledge acquired through more thousand hours of experience and focused practice (e.g., Chase & Simon, 1973; de Groot, 1978; Ericsson,

Krampe, & Tesch-Römer, 1993). It was found that experts' tacit knowledge affects their pattern recognition abilities (Wagner & Sternberg, 1985; Gurteen, 1998; Eraut, 2000; Cianciolo et al., 2006) and that this knowledge allows them to make rapid decisions, such as to recognize creativity when they see it (Amabile, 1982; Baer 1994) (more information above the properties of experts can be found above in the *Problem-Solving and Expertise-based Theories of Creativity* section). Although creativity is always evaluated in a specific socio-cultural context and may defy empirical objectivity on an individual level (Schaffer, 1994), the added benefit of using trained experts for judging creative ideas is that there seems to be a rather substantial agreement among them when it comes to evaluating creative products (e.g., Hennessey, 1994; Kaufman, Lee, Baer, & Lee, 2007; Kaufman et al., 2008). In the absence of a clear-cut criterion for what qualifies as creative and what does not, such a consensus in experts' opinion may be considered as the measurement tool for creativity (the use of expert ratings as substitutes for criteria will be discussed in *Chapter 3*).

Since acquiring expert evaluators may be very demanding for both researchers and companies, a great body of literature is targeting the exact differences between novice and expert evaluators, and aims to bridge the gap between them. Currently, there are a variety of trainings constructed to enhance judges' evaluation skills (e.g., Storme et al., 2014). Trainings were found as methods which can increase the similarity of lay participants' judgments with expert opinions, however, there are still problems with the inter-rater reliability of lay judges and the shift towards expert opinions is not sufficient to actually replace experts by the trained lay judges (Bruer, 1993; Kaufman, Gentile, & Baer, 2005; Kaufman et al., 2008; Kaufman, Baer, & Cole, 2009; Kaufman, & Baer, 2012; Storme et al., 2014). Despite the trainings, there are vast qualitative differences between the creativity ratings provided by expert and non-expert judges.

Namely, the standards and criteria by which ideas are judged appear to differ significantly between experts and non-experts (Kaufman, Baer, Cole, & Sexton, 2008; Kaufman et al., 2013; Silvia, 2013). There is a high degree of internal reliability of expert evaluations of creative products (Amabile, 1982; Hennessey & Amabile, 1999; Kaufman, Lee, Baer, & Lee, 2007), which does not tend to be the case in non-experts. Experts are equipped to handle contexts where strict criteria of assessment of creativity are available, as well as contexts when there are no decision criteria (Bettman & Sujan, 1987). In contrast, non-experts tend to have difficulty in developing reliable criteria for assessing creative ideas, which in turn limits their ability to identify ideas that experts would evaluate as genuinely creative (Galati, 2015).

In this section, the behaviour of both lay people and experts were described while completing creative evaluation tasks. Since it is difficult to recruit experts for conducting academic research, many studies attempted to bridge the gap between expert and non-expert evaluations. In multiple studies, the aim was to bring lay participants closer to experts via various manipulations (e.g., trainings, information briefs, mindset inductions). Despite finding results for an increased alignment between non-experts and experts after applying these manipulations, vast qualitative differences remained between the creativity ratings provided by expert and non-expert judges. Although prototypical situations could be trained, non-experts were not expected to be able to adapt when facing a novel situation, whereas experts were found to quickly transfer their knowledge even in stressful situations. While numerous boundary conditions have been investigated, it is difficult to organise and compare the findings due to the different paradigms and domains involved in the studies. Much information is available, however, it is scattered and not integrated into one theoretical framework. The next step is to outline which part of the aforementioned issues is this thesis addressing.

1.6 The Focus of The Thesis

The key insight drawn from the sections above is that while most research is conducted on idea generation (cf., Long, 2014), there are fewer studies focusing on how ideas are evaluated (e.g., Basadur, 1995; Silvia, 2008). In particular, little is known about the cognitive mechanisms of creative idea assessment, except maybe that lay people are not very good at it (Blair & Mumford, 2007; Dailey & Mumford, 2006; Rietzschel et al., 2010). This is the gap the present thesis set out to start filling. Now, I shortly outline the rationale behind the thesis research.

Given the modern technological advancements, there is now an almost unlimited supply of ideas (Bayus, 2013), curating them became a key role to manage for success (Bakker, 2014). Often, companies ask for expert opinion when selecting ideas (e.g., Magnusson, Netz, & Wästlund, 2014). Experts of a domain (e.g., digital technology or visual arts) are chosen as curators because they are better suited than lay people to recognize the potential of ideas due to their experience with predicting the reaction of the market (Basuroy, Chatterjee, & Ravid, 2003) and to their potential to influence trends (Reinstein & Snyder, 2005). Therefore, there is motivation for researchers to tease apart in which ways expert evaluators deviate from the general population. Given the limited understanding of this topic, the present thesis aims to examine the basis on which creative ideas are evaluated by people with and without relevant expertise. Specifically, three main research questions are investigated in this thesis. First, the use of which inner criteria is fed into the judgment made about creativity? Second, what is the weighting of the criteria informing the judgment? The experiments were designed to gain data about which criterion is more important than the others to confirm or falsify current theoretical conceptualisations of creativity. Third, what contextual factors are influencing the evaluation of creative ideas? These questions will be outlined in detail in *Chapter 4*.

The motivation for this research is the hope that by understanding more about the cognitive underpinnings of making judgments about creativity, it will become possible to

reduce the amount of guesswork ratings given by non-professional evaluators as well as the mythical bubble of "divinity" (Cropley, 2010) surrounding creativity. The wider goal is to start an optimisation for creative evaluations in terms of the guidelines and preparation judges are receiving. It is in the interest of all of us that the judges of both potentially paradigm-shifting products and everyday inventions get to incorporate findings from the cognitive science literature. Hopefully, with time, a more data-driven and less incidental approach will become the mainstream by awarding grants and prizes.

Now that the rationale of the present research was sketched out, let us take a look on how previous studies addressed the aforementioned issues. In the next section, an organised list of earlier studies is provided, in which I discuss both the theoretical conceptualisation of creativity for the research purposes and the way in which the concepts were measured.

1.7 Different Approaches for Uniting Creativity and Judgment Research

The body of literature is limited on the evaluation of creative ideas, however, there are a handful of studies investigating how judgments are made about creativity. In this section, I discuss the theoretical conceptualisation of empirical studies investigating the evaluation process. Most of the available literature applies the short, two-component definitions of creativity (examples of this are listed in the *Definition(s) of Creativity* section), i.e., they use novelty, originality, or uniqueness on one side and usefulness, appropriateness, utility, relevance, or value on the other side as the criteria of creativity. Since both criteria are required for a creative response or product, it is worthwhile to study how these two components are weighted relatively to each other and to what degree they are integrated. The previous findings can provide valuable insights to what weighting between originality and utility can be expected *a priori* when looking for answers to the second main research question ("*What is the weighting of the criteria informing the judgment?*"). Thus, these are the aspects I analyse while describing

the theoretical framework of the available assessment methods. (Further assessment criteria for the judgment of creativity will be listed in both *Chapter 2* and *Chapter 4*.)

The earliest assessment tool similar to the paradigm used in this thesis is Susan P. Besemer's Creative Product Semantic Scale (1998; Besemer & O'Quin, 1986, 1987, 1993; Besemer & Treffinger, 1981). The theoretical model behind the scaled is called the Creative Product Analysis Matrix. This model can be used for the measurement of domain-general creativity by products and consists of three creativity-related criteria: Novelty, Resolution, and Elaboration & Synthesis. Novelty includes originality and surprise; Resolution is comprised of value, logic, usefulness and understandability. The third dimension is called Elaboration & Synthesis and it includes the facets of being organic, elegant, and well crafted. For the evaluation, these nine facets are used in form of pairs of adjectives and participants need to rate each criterion on Likert scales while being presented with pictures of objects (e.g., chair designs). Horn & Salvendy (2009) conducted two studies using similar stimuli to Besemer's. First, lay judges of creativity evaluated chairs and lamps on a web-based platform, then individually selected products were judged using paper and pencil in the second study. An exploratory factor analysis could explain 72% of the variance and identified three factors related to creativity: Affect, Importance, and Novelty. This data-driven conceptualisation aligns well with the three componential definitions of creativity. An interesting finding was that Affect was an equally good predictor of creativity as Novelty, which draws attention to the role of emotion in creativity assessments.

Another category of consumer products which is frequently used in creativity research as the stimuli for evaluation is advertisement. E.g., the Creative Product Semantic Scale was used for judging print advertisement campaigns (White, Shen, & Smith, 2002). It was found that while experts (advertising professionals) and non-experts (college students and members of the general public) agreed on the originality and logicality of the ads, they disagreed on how

well crafted and well executed the products were. Since the results were presented on an ordinal scale only, no data about the weighting of originality and appropriateness was available.

Another study (Caroff & Besançon, 2008) fulfilled exactly this gap: the role of originality and appropriateness was investigated while making judgments about automobile advertisements. In three experimental conditions, participants had to judge the creativity of fifteen ads either without any instructions, with using explicit criteria, or with receiving training for the task. The findings suggested that originality is a necessary but not sufficient condition of creativity. Appropriateness modulated the creativity ratings in interaction with originality and originality influenced creativity more when appropriateness was on a low level, while when appropriateness was on a moderate or high level, the impact of originality on creativity ratings was reduced. Regarding the experimental conditions, when participants had to rate creativity without receiving any special instructions, appropriateness was in a linear relationship with creativity. However, when they received explicit instructions on how to judge creativity, appropriateness did not predict creativity anymore. Finally, when participants received training for the task, the highest ratings of creativity were associated with a moderate level of appropriateness.

Another aspect of investigating the evaluation of creativity in advertisements is to look at how certain personality characteristics influence the weighting of creativity-related criteria (Storme & Lubart, 2012). Here, again the Creative Product Semantic Scale was applied as the measurement instrument (20-degrees Likert-scale). The authors found Novelty as the most influential predictor of creativity (not Resolution, or Elaboration & Synthesis). The expressed and implemented creativity concepts of lay participants (explicit and implicit in the authors' wording) were also studied and the findings showed that Novelty was the most heavily weighted dimension in both recalling participants' creativity definitions and in the experimental task.

A different method for evaluating creative responses was applied by Runco & Charles (1993), who asked participants to judge ideas by sorting cards containing a set of ideas (each card contained 8 responses to a divergent thinking task) to piles of high/low originality, appropriateness, and creativity. Originality and appropriateness seemed to work as trade-offs to each other: the lowest originality ratings were given to the idea sets with the highest amount of appropriate ideas and the lowest appropriateness ratings were given to the sets with the highest amount of original ideas. Similarly to the paradigm used in this thesis, both non-expert and expert participants had to rate originality and appropriateness and originality as rated by the lay participants themselves predicted creativity ratings. Testing the boundary condition of reducing originality to zero, creativity ratings decreased when the number of appropriate ideas in the set increased. A high level of originality and appropriateness as judged by experts was linked to high creativity ratings.

Another noteworthy paradigm of investigating creativity is instructing participants to draw alien creatures and then asking them to write a short description about them (Ward, 1994). In a study (Kozbelt & Durmysheva, 2007), lay judges had to evaluate the creativity of these products based on their internalised model of creativity; independently, a code system was developed to analyse the attributions of the products. Unfortunately, the originality and appropriateness of the products were not registered. However, both the drawings and the paragraphs were coded according to multiple comprehensive criteria (e.g., creature as a whole, features of the creature, creature's emotional expression). Paragraph coding categories were found to be the most predictive of creativity. The characteristics of alien creatures which was linked to a higher creativity score showed a systematic pattern: greater conceptual combination resulted in more creative products.

After discussing laboratory experiments, let us close with a real-life example of conducting research about creativity. Although Frederiksen & Knudsen's (2017) paper is placed

in the innovation literature, their methodology investigates the precursors of innovation, namely, creativity, thus their paradigm belongs to this section. In a study with a high ecological validity, industry and market experts were asked to evaluate some of the 106 student project proposals (one of the fifteen experts evaluated all of them). The dimensions for assessment were novelty, usefulness, and market potential of the projects and experts used a 0-100% scale for rating. Unfortunately, only descriptive statistics and the correlation between the dimension are communicated, thus there was no information revealed about the weighting of the dimensions. According to their conceptual model, all three dimensions were equally important, however, since the stimuli were real business ideas regarding new energy sources and offshore wind power, a longitudinal study could follow up which ratings were associated with the best outcomes in the implementation phase.

After flashing out how other scholars have united creativity and judgment making research to measure the creativity of products, now I introduce the conceptualisation of judgment used in the present research and propose a possible framework taken from judgment research for the investigations of this thesis.

1.8 A Framework from Judgment Research

Judgment can be defined as the result of forming beliefs about the likelihood of uncertain events (Hardman & Macchi, 2003). It can be also defined as a tautology: judgment is what judges do (Martin, 2006). In this sense, judgment can be defined as an activity exerted by cognitive agents: judging something is soliciting evidence, as well as weighing, interpreting, and assessing the evidence to make sense of the world (Martin, 2006).

In this thesis, judgments stand for estimations of values; they are made regarding the degree of creativity found in project proposals. Given the socially determined nature of creativity, the Theory of Social Judgment (Brehmer & Joyce, 1988; Hammond, Rohrbaugh, Mumpower, & Adelman, 1977; Hammond, Stewart, Brehmer, & Steinmann, 1975) was chosen

as the theoretical (and mathematical) framework for conducting the research. The core idea of this approach was laid down more than 60 years ago, almost at the same time when the scientific study of creativity started, by Egon Brunswik (1952, 1956), a functionalist psychologist interested in the adaptiveness of judgment making. In his writings about representative design, he portrayed judgment analogously to perception (Goldstein, 2004): humans perceive both objects and events by relying on incomplete information and are constructing the full picture based on sensory cues. A mathematical model of this notion was later formalised as a prism of cues between the actual values and the judgments made about these values. The Lens model is simple and elegant, it does not only capture what is happening in a person's head while forming a judgment but also includes the environment of the person, the surrounding ecology (see Figure 1).

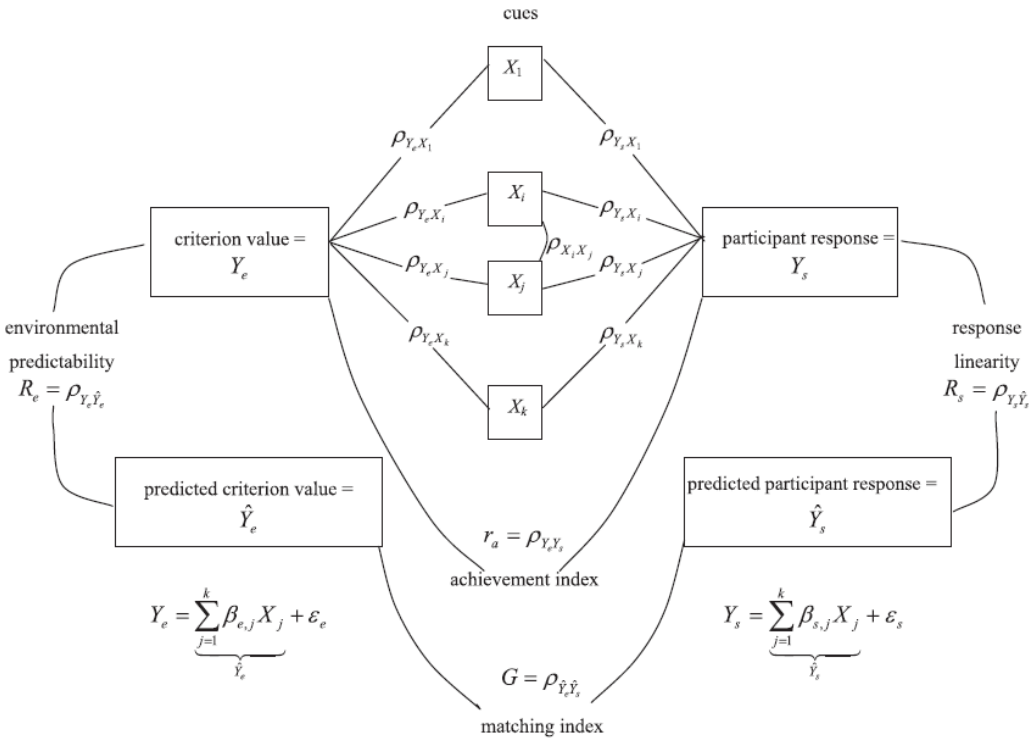

Figure 1. Outline of the Lens model. Adapted from "Heuristic and Linear Models of Judgment: Matching Rules and Environments" by R. M. Hogarth & N. Karelaia, 2007, *Psychological Review, 114*, p. 734. Copyright 2007 by the American Psychological Association.

The basic logic behind the model is that judgments are formed based on pieces of information (called cues), which are probabilistically related to the judgment, and get incorporated to varying degrees. The environment involves a criterion value which can be interpreted as a 'true score' of the judgment and it can be investigated how much the participant's actual response is corresponding to this; the metric is called achievement index and can be calculated via a correlation analysis. Brunswik (1952) noted that the cues are always in a probabilistic relationship with the criterion, as well as they are associated with various degrees of uncertainty with regards to their reliability and validity, thus they cannot be interpreted as deterministic factors. In fact, making inferences based on the cues involves two types of noise: one of them is associated with perceiving the cues, while the another one with applying the information extracted from the cue to the judgment (Kirlik, 2009).

According to Funder's (1995, 2001) Realistic Accuracy Model, four steps must be completed in sequence in order to reach a realistic judgment. The path to making an accurate judgment starts with relevance: the target of judgment must signal relevant information chunks to the observer. The second step regards availability as the conveyed information must be made available to the observer. The third step is about detection as the available information must be detected by the observer. Finally, the fourth step concerns utilisation and posits that the detected information must be interpreted and weighted in the correct way to reach an accurate judgment. Since the steps are built on each other, any mistake on a lower level will distort the judgment and amplify the inaccuracies. This model was described because the Lens model operates on the same principles and assesses the accuracy of the judgment. Brunswik's conceptualisation was picked up by Hammond and his colleagues (1977) and they extended the ground principles into the mathematical framework of social judgment research. In fact, the issues with the possible imperfections also became quantified and termed in the extended Lens model. The term cue utilization denotes the assessment of what cues are to which extent governing the

participant's judgment process, while cue validity is associated with the process of determining how predictable is the criterion using the given set of cues (Stewart, 2001). Part of the elegance of the model is that each of its components can be calculated with simple statistical analyses. As Stewart put it (2001, p. 35), the contribution of the Lens model is that it became „possible to analyze cognitive performance into the part that was due to the environment, the part that was due to the judge, and the relation between them".

The application of this model to creativity research, along with the linear model equation, can be found in *Chapter 3*. In sum, the Theory of Social Judgment lends itself as a suitable framework for investigating how people rate creativity. The Lens model offers a simple quantitative model for assessing which fallible cues are informing the judgments made about creativity. By adapting the model to creativity research, it will become possible to analyse whether expert and lay judges of creativity use the same criteria to inform their judgments and how both groups are weighting each criterion.

After outlining the selected framework from judgment research, let me reiterate what is the proposed theoretical framework for studying creative evaluation as an applied case of judgment making. Creativity is discussed from a cognitive point of view, i.e., in the focus of interest are the cognitive mechanisms involved in creative idea evaluation. Creative products are defined as both novel and useful, and these two components are interlinked. Further, creativity can only be interpreted within a social context; the judgment made about creativity is the result of the interaction between the product and its observer, the judge. Regarding the judges, there are two distinct groups examined, experts and lay people. This setup serves the purpose of investigating the role of expertise in creativity judgments. To date, there are several studies comparing experts' and non-experts' judgment about creativity, however, these studies do not offer a cohesive explanation to the question what experts are doing differently to become better judges of creativity than people without relevant expertise. The present research

addresses what information sources are used to inform judgments made about creativity (criteria) and how they are used (weighting). The next chapter will introduce the conceptual framework for investigating these questions and the biggest difficulties related to crafting measurement instruments.

CHAPTER 2:

A REVIEW OF THE METHODOLOGICAL APPROACHES REGARDING THE EVALUATION OF CREATIVE IDEAS

2.1 Introduction

Creativity is a multifaceted construct and notoriously difficult to capture by a single definition (Runco and Jaeger, 2012). Creativity is conceptualised as a process that is broadly similar to problem solving, in which, for both, information is coordinated toward reaching a specific goal (Wiggins and Bhattacharya, 2014), and the information is organized in a novel, unexpected way. Problems which require creative solutions are ill-defined, primarily because there are multiple hypothetical solutions that would satisfy the goals (Reitman, 1965). Because there are no objective rules on how to reach a solution to a creative problem, a combinatorial explosion of possible choices occurs (Simon, 1989; Simonton, 2010). Therefore, embarking on a solution to an ill-defined problem necessitates the problem solver to frame and interpret what might be relevant as a possible goal and then to establish a solution that meets that goal (Hayes, 1989; Mumford, Reiter-Palmon, & Redmond, 1994).

For a creative problem, an original solution is often unthinkable in advance, thus assessing creative solutions (i.e., creative ideas) occurs in the absence of objective criterion/criteria against which a creative product can be measured up to. As Amabile (1983, p. 359) put it, "current definitions of creativity are conceptual rather than operational; their conceptualizations have not been translated into actual assessment criteria" yet. Due to this 'criterion problem' (McPherson, 1963, Shapiro, 1970), it is difficult to objectively evaluate the extent in which a particular goal is met (Runco and Smith, 1992; Runco and Chand, 1995). One of the aims of this thesis is to establish criteria which can predict reliably what people will deem as creative. The lack of usable criteria can be detected on higher levels too. Due to the problems with measurement, there is only a limited pool of available paradigms about the evaluation of

creativity – this chapter gives a summary of the existing paradigms linked to the evaluation of creative ideas. In order to develop a new measurement tool, first the previous research is presented in an organised form. The paradigms are reviewed from two angles: first, according to the type of the paradigm, and second, according to the complexity of the paradigm. From the overview of the paradigms, several methodological issues come to light. For the purposes of the present thesis, two big issues are introduced, both related to the difficulties with the scientific measurement of creativity. After discussing the issues with the subjective and domain-specific nature of creativity, we move on to the focal point of establishing a novel paradigm: to the selection of creativity-related criteria. An overview of the existing taxonomies is provided to inform the reader which criteria have been previously used to measure the sub-components of creativity. Finally, the reasons for selecting the criteria for the present research will be fleshed out, keeping the applied research context in mind.

2.2 Research Approach

Since Guilford's famous APA Presidential Address held in 1950, the psychometric perspective became the mainstream approach in creativity research, launching the first golden age of creativity research (Plucker & Renzulli, 1999). Other scholarly approaches include the biographical approach (e.g., Gruber's and Gedo's work); the historiometric approach hallmarked by Dean Keith Simonton's work, the biometric approach favoured by Richard Haier, and the experimental approach used by Theresa Amabile and many others.

Here, we discuss the psychometric and the experimental approaches, as these two are relevant for the thesis work; but before delving into the topic, perhaps it would be worthwhile to spell out what are the exact differences between these two approaches. The psychometric and the experimental approach are indeed quite similar, however, they differ in terms of the research design employed in them; in psychometric investigations, the research questions are scrutinised with correlational and causal-comparative designs, while by experiments,

(quasi)experimental designs are used to generate data (Plucker & Renzulli, 1999). Another major difference, as identified by the authors, is that while experiments are focusing on the cognitive, problem-solving, and product aspects of creativity, psychometric investigations focus on the personality and environmental aspects of creativity. The thesis research can be sorted into the experimental approach.

There are four main areas of research in the psychometry of creativity: one might investigate the creative processes, the personality and behavioural correlates of creativity, the characteristics of creative products, and the attributes of creativity-fostering environments (Rhodes, 1961). From these, naturally, the characteristics of the creative products will be presented in this chapter.

There are not too many measurement instruments for creative idea evaluation. Below, I summarise the currently available research methods while sorting them according to multiple viewpoints. However, please note that the 'market leader' paradigm called the Consensual Assessment Technique (Amabile, 1982) will be only shortly discussed in this chapter as *Chapter 3* is dedicated to a thorough analysis of this paradigm.

2.3 Sorting the Paradigms According to Type

There are many ways to administer the evaluation of creative products. Previously, it was noted that common design competition assessment methods vary to the degree in which they are formalised - some judges rely on their gut feelings while others use external reference points and interpersonal discussion to refine their judgment (Yen & Sun, 2008; Lu & Luh, 2012). The previously identified types could be categorised according to three axes: (1) whether the assessment is structured or not; (2) whether the evaluation is absolute or relative; and (3) whether the assessment is provided with or without discussion amongst the judges. This section describes each axis in detail.

First, an example of a non-structured assessment can be a judgment made using gut feelings. (Pétervári, Osman, & Bhattacharya, 2016). Evaluating creativity requires the complex assessment of multiple dimensions, however, these dimensions are not articulated often. Also, lay people were found to have rather implicit models regarding creativity (Lim & Plucker, 2001; Runco & Johnson, 2002). Given that it is difficult to explicitly verbalise the content of the models, judges of creativity might have an easier time with rating creativity based on their ineffable feelings than with forcing themselves to come up with an explanation to their judgment. Another example of a non-structured assessment can be a funding agent who feeds back some thoughts on a pitch he watched, highlighting what was his personal impression and his non-official advice. This assessment is subjective but also quick and efficient, focusing on the essential information in a communicational context.

In scientific investigations, usually more formalised assessments are used. This is to offer a more in-depth coverage, to avoid ignoring any important aspect of the evaluation. Structured assessments require more paperwork but result in richer, less biased output data. A good example of a highly structured assessment method is the Creative Solution Diagnosis Scale (Cropley & Cropley, 2012), which inventories all noteworthy creativity-related criteria.

Second, creativity can only be assessed compared to a certain set of other products. The reason for this is that a requirement of creativity is for the product to be new and novelty emerges only in comparison. Thus, this axis differentiates between two reference systems: when a creative product is measured up ‘in general’ and when it is compared to other products which were designed to fulfil the same goal. In the first case, an absolute judgment is made, where the judges utilise their expertise and taste acquired through many years of being presented to more or less similar items. This type of judgment is less dependent on the exact context, therefore, can be deemed as a “standalone”, reliable rating. However, in real life, creative items are usually produced for open tenders where the entries are contrasted with each

other to find the winner (e.g., Science magazine's Dance Your Ph.D. competition, Creative Review's Annual Competition, or the Shorty Awards 'honouring the best of social media'). This other type of judgment is called relative judgment. Here, the question changes from '*how creative is this product?*' to '*which product is more creative than the others?*'. An example assessment method of making relative judgments is the frequently used sorting comparison method called Consensual Assessment Technique (CAT), which operates by asking several experts to rate the items and establishes rankings through quantifying the agreement among the raters. This method is often referred as the 'gold standard' of creativity evaluation (Carson, 2006) and is therefore thoroughly discussed in *Chapter 3*. The differences between making absolute versus relative judgments are further demonstrated in *Chapter 7* through empirical data.

Third, similarly to the idea generation phase, there are numerous techniques for getting to the most creative idea in the evaluation phase too. Findings vary in suggesting whether judges are the most equipped to spot great ideas if they do the job by themselves, or if they consult fellow judges and balance their opinion. According to the CAT, judges should be working independently from each other in order to ascertain they are not influenced in any way in their judgment. This method is very similar to the procedure of how prestigious prizes are awarded. However, in more practical settings, people rarely work in isolation. During brainstorming processes, the judges and the creators of the ideas are often the same people. For the self-selection of creative ideas, a moderate frequency of communication and a decentralised communication pattern were found as the most advantageous settings (Leenders, van Engelen, & Kratzer, 2003).

The insights drawn from here is that there are various practices on how to evaluate creativity. Although most researchers favour to use the CAT, sometimes in real-life settings, it would be more desirable to provide absolute ratings about creativity. A scientific measurement

tool cannot be based on hunches, however, often it would be troublesome to require judges to verbalise why they find something creative. To combat this issue, detailed measurement instruments have been constructed, however, some judges of creativity prefer to assess creative products rather holistically than analytically due to the complexity of the task. The next section analyses the available measurement techniques according to their complexity.

2.4 Sorting the Paradigms According to Complexity

Plucker and Renzulli (1999) suggested that creative evaluation techniques can be sorted according to their degree of complexity. With a simple grouping of similar methods, they identified conceptually less and more sophisticated evaluation techniques. Several straightforward rating scales are comprised of a few but clearly demarcated dimensions. Examples for this are the Creative Product Inventory (Taylor, 1975) measuring Generation, Reformulation, Originality, Relevancy, Hedonics, Complexity, and Condensation; the Creative Product Semantic Scale (Besemer, 1986), whose conceptualisation was discussed in *Chapter 1* already, or the Creative Product Analysis Matrix (Horn & Salvendy, 2009) which consists of six main dimensions related to product creativity: Novelty, Resolution, Emotion, Centrality, Importance, and Desire. Additionally, some researchers have developed criteria which can be employed to assess specific tasks they use in their studies, such as the criteria constructed by Kozbelt and Serafin (2009) for the dynamic assessment of drawings.

As for the conceptually more complex techniques, creativity researchers often not only rely on specific scales but on persons who have the necessary expertise to take all relevant dimensions into account. Raters may vary in terms of the tasks they need to judge: in educational settings, teachers often register the scales mentioned above and additionally, provide comprehensive comments based on their knowledge of the student's individual development. Similarly, experts of a domain, such as film, technology, or marketing, are often

employed to judge the creativity of a product corresponding to their specialised knowledge not only in a quantitative but also in a qualitative form. When experts are asked to share their professional opinion, some techniques offer instructions which are guiding the judges (Csíkszentmihályi & Getzels, 1971) and some do not provide additional information to avoid biasing the judge in any direction (e.g., Jeffries, 2017; Lee, Lee, & Youn, 2004; Lu & Luh, 2012). A lengthy discussion of using expert ratings with and without accompanying criteria can be found in *Chapter 3*.

The takeaway from this section is that creativity remains ineffable for many and while standardising its measurement is crucial from an academic point of view, in practice, many times the ratings are made while considering the constellation of various dimensions and cannot be made only by evaluating single dimensions and adding up the scores. Feelings and hunches are important for the judges of creativity, which might be linked to the lack of explicit conceptualisations about creativity in folk psychology. The next section touches upon two major methodological issues related to creative evaluation, which will shed further light on the challenges associated with the scientific measurement of creativity.

2.5 Issues

The two most common methodological problems every creativity researcher is facing when it comes to the evaluation of ideas are the selection of rating criteria and the subjectivity found in the ratings; this latter is called the rater bias (Hung, Chen, & Chen, 2012). I will first discuss these two problems, then continue with the issue of domain specificity since the domain-specific nature of creativity vastly influences the selection of rating criteria for its measurement.

2.5.1 Variability in the Creativity Judgments. There are considerable individual differences in the rating of creative ideas (e.g., Caroff & Besançon, 2008), which poses several

theoretical questions about the measurement of creativity. First, one assumption of the scientific measurement is that while measuring a construct, the measured aspect must reveal itself identically to all observers (Messick, 1989; Long & Pang, 2015). Thus, it must be ensured that the ratings are outcomes of a systematic process and not assigned randomly. Further, since the variance in the creativity ratings might reflect more the judges than the construct in some cases (e.g., Silvia et al., 2008), the sources from which this variance stems from and their share of the variance must be determined. Specifically, the relation between the product and its judge must be established and there are two major conceptualizations for this. One of them regards the judges of creative products as 'rating machines', while the other considers them as independent experts (Hung, Chen, & Chen, 2012). If the judges are thought of as rating machines, they are expected to cast their ratings objectively, according to well-defined criteria, and therefore are expected to reach an almost full agreement. If the judges are considered as independent experts, the assessment becomes more subjective as the judges are expected to apply their own understanding of the criteria and the goal is not to reach a high degree of consensus but to apply the criteria consistently. As a quality control of the ratings, the use of a large number of judges is recommended, thus the effect of individual differences can be cancelled out. However, acquiring ratings from many judges is rarely feasible in practice (Kaufman & Baer, 2012) (see *Chapter 3* for more details).

At the core of the two outlined approaches lies the way in which the idiosyncratic interpretations of creativity are considered. The intra-individual variability is influenced by both situational and dispositional factors, e.g., demographic variables and expertise were found to be related to the judgments made about creativity (White, Shen, & Smith, 2002). Even a general term, the so-called *rater effect* was coined to describe the individual rating tendency of a judge which influences the assessment of creativity (Wolfe, 2004; Hung, Chen, & Chen, 2012). In most creativity studies, the rater effect is dealt with as a type of noise related to the

data which should be estimated and corrected for, along with other sources of error (Long & Pang, 2015). Ultimately, the rater effect is understood as the human factor or the social constituent in the creativity ratings. While there is a general attempt to minimize it as it might pose a "major threat to construct validity" (Long & Pang, 2015; Messick, 1995), it is an open question whether the judges are indeed making systematic errors while assessing the ideas. And if so, what could be learned about creativity from these errors.

Taking a different approach, the variance due to the raters could also be observed as meaningful data and not as a source of measurement error. The rater effect might reflect the variability of the creativity conceptions internalized in each judge. Thus, while the first general model of creativity outlines that

creativity rating = creativity of the product + measurement error

in the second case, the model could be expanded to

creativity rating = creativity of the product + rater's understanding of the criteria + measurement error.

As it becomes apparent from this section, the most important question to address is whether it would be a reasonable expectation from the judges of creativity to cast the same rating about a product (within a confidence interval). For a scientific measurement, such a universal “creativity value” would be required. Or, if giving such absolute values are more influenced by the judges’ rating tendencies than by the creativity of the product, would it be reasonable to expect from the judges of creativity to make the same ranking of the products? The CAT assumes that this latter is feasible to expect from the judges.

Before constructing a novel paradigm, one must think through what scale should be used for measuring creativity. Even more so, the variability found in the creativity ratings must be interpreted on a theoretical level. In the present research, the idiosyncratic interpretations of

creativity are acknowledged as meaningful data and this data source is part of our measurement theory.

Next, a different issue with the measurement is discussed: the question whether creativity is rather domain-general or -specific (or perhaps neither or both) and how this influences the construction of a measurement tool for creative evaluation.

2.5.2 Domain-Specificity. The degree to which a given domain constrains a creator is one of the hot debates in creativity research. As Baer (2010, p. 321) noted it, "whether creativity is a general, domain-transcending set of skills, aptitudes, traits, propensities, and motivations that can be productively deployed in any domain – or, conversely, whether the skills, aptitudes, traits, propensities, and motivations that lead to creative performance vary from domain to domain – is a key question in creativity research and theory". There are many different points raised in the debate; e.g., since a general factor, which could be utilised to many different domains, was found in intelligence research (Neisser et al., 1996), it seemed reasonable to think that there would be a personality trait or cognitive skill associated with generating creative products, regardless of the domain (Baer, 2015). The broad adoptability of general intelligence also inspired the construction of creativity tests and trainings for many years. From all the related issues to domain-specificity, two key questions were selected corresponding to the purposes of this thesis: (1) are domain-general evaluators good enough judges of creativity or domain-specific experts are needed? (2) are the results obtained from one domain of creative evaluation transferable to reason about other domains and are they comparable with the results acquired from different domains?

To answer these questions, we must note that for creative achievements, broad abilities (such as problem-solving skills) must be used in a specialised way or, differently put, in a narrow direction. Thus, even if domain-general judges of creativity are experienced and proficient in evaluating ideas or proposals, they would need domain-relevant information to

assess the feasibility of a project from a highly codified domain (Kaufman et al., 2013). Based on this logic, if the creative product is related to a more accessible domain, then domain-general judges are acceptable to use. Regarding the second question, even the advocates of the domain-general approach acknowledged that creativity is content-specific, and in many cases, even task-specific (Plucker, 1998). These are not good news for research efficiency and a compromise might be to compare results obtained from the same branch of creativity research (e.g., functional judgments or aesthetic judgments).

Csíkszentmihályi (1999) argued that one reason for the controversies observed in this debate is that, similarly to chemistry before the periodical system or to physics before the quantum theory, without a paradigm proposing a symbolic system for a domain, it is very difficult to resolve conceptually intertwined issues. On a similar note, Plucker & Beghetto (2004) resolved the Gordian knot by declaring that the question of whether creativity is domain-general or domain-specific is most likely to be posed improperly since this kind of division is not clearly interpretable. They bring the example of humans having the potential to become an eminent creator in multiple different domains (domain-general abilities) but then acquiring the necessary expertise in only one of these possible domains (domain-specific performance).

Clearly, the chosen domain of a creative endeavour influences what measurement technique can be used for assessment, also what measurement instrument can be constructed to account for all relevant dimensions present in a task. In the present thesis, these challenges were overcome by selecting both domain-general and -specific criteria for assessment, which will be discussed in detail in *Chapter 4*.

Now we move on to the next section, in which I collect the criteria used for creativity measurement in experiments, observing whether they are domain-specific or domain-general.

2.6 Existing Taxonomies of Features

A review of 90 creativity studies (Dean, Hender, Rogers, & Santanen, 2006) was used to identify the features by which a creative product is evaluated. The authors adopted MacCrimmon & Wagner's (1994) multi-attribute taxonomy and classified the features related to creative products as typically falling into categories that include novelty, workability, relevance, and specificity. This four-dimension framework mapped well with 51 out of the 90 studies. Novelty was defined based on whether an idea has been expressed before. Workability could be interpreted as feasibility: whether the idea could be easily implemented (did it violate known constraints?). Relevance was fulfilled if the idea satisfied the goals set by the problem solver. Finally, thoroughness was closely related to resolution – an idea was found to be thorough if it was worked out in detail. It is conspicuous that all four features were domain-general criteria of creativity, accordingly, the studies were collected from various domains.

A full list of all features stated in the studies goes as the following: Novelty, Creativity, Originality, Unusualness, Paradigm Relatedness, Non-obviousness, Imaginativeness, Innovativeness, Excitement, Rarity related to Novelty. Workability, Feasibility, Implementability, Logical, Adoptability, Non-violation of known constraints, Practicality, Social Acceptability, Probability related to Workability. Relevance, How well it dealt with problems, Effectiveness, Appropriateness, Ability to solve the problem, Potential plausibility, Impact, Value addition, Applicability, Business potential, Utility/Usefulness, Relation to topic, Importance, Realistic, Magnitude of impact of policy on stakeholders, Goodness or usefulness for purpose, Validity, Significance, Quality related to Relevance. Finally, Specificity, Thoroughness, How well described, Solution based on facts and possibilities, Generality, Detail, Depth, Clarity related to Specificity. This list is provided to show that a wide variety of creativity-related features exist and there is no consensus on which ones to use. In fact, there are several problems stemming from the irregular use of features associated with the assessment of creativity: (1) they make it difficult to train the judges, (2) they result in raters being

inconsistent in their individual ratings, (3) they can lead to inconsistencies between the raters, and (4) they make the studies' comparability and generalisability low (Dean et al., 2006).

The conclusion of the review was that different studies measure different constructs: measuring idea quality and idea creativity are not equal. Also, Dean et al. (2006) found that for a systematic sampling, the best is to score each creativity-related dimension separately than to provide a holistic, overall creativity score too.

In the recent years, another ambitious initiative set out to construct a comprehensive scale for the evaluation of functional creativity (Cropley & Kaufman, 2012, 2013; Haller, Courvoisier, & Cropley, 2011). The Creative Solution Diagnosis Scale consists of four main features: Relevance and Effectiveness, Novelty, Elegance, and Genesis. The 30-item measurement instrument was validated through a study of judging mousetrap designs and was shortened to a 24-item scale (Cropley & Kaufman, 2012). There are several indicators associated with each feature. Relevance and Effectiveness consists of correctness, performance, appropriateness, operability, safety, and durability. Novelty includes diagnosis, prescription, prognosis, replication, combination, incrementation, redirection, reconstruction, re-initiation, redefinition, and generation. Elegance involves recognition, convincingness, pleasingness, completeness, gracefulness, harmoniousness, and sustainability. Finally, Genesis, which is a dimension not mentioned in the review above, includes foundationability, transferability, germinability, seminality, vision, and pathfinding.

These partly domain-general, partly domain-specific indicators were collected and validated to offer a fine-tuned resolution about the creativity judgments and to cover all possible contributing factors. Although Cropley & Kaufman's scale covers more dimensions than any other evaluation inventories, it is questionable whether it would become a widely used measurement instrument due to the time and energy constraints imposed by the administration of it. The takeaway message for this thesis is that a good balance should be kept between

covering *all* dimensions and covering too few dimensions: several but not too many criteria should be selected for the present research, which can be handled by both expert and lay judges of creativity. Next, I discuss which criteria were selected for the applied purposes of the present research.

2.7 The Implementation of Features in an Applied Context

The previous section shed light on the wide range of features used as creativity-related criteria and on the lack of consensus among the researchers on which ones to use. Creativity-related criteria are not standardised, and the selection of features is most likely to be determined by the specifics of the task and by the information available to the researcher, e.g., the consciously chosen research tradition which the scholar set out to follow. In psychological research, efforts are made to cover all relevant dimensions related to creativity (e.g., the Creative Solution Diagnosis Scale, Cropley & Kaufman, 2012). However, in applied research fields, such as consumer research, or in practical settings, such as multinational corporations' product development units, fewer criteria are selected to focus on the relevant features and to make the process as efficient as possible. For example, in the innovation literature, the requirements of creativity are novelty, usefulness, and value produced for the firm, while in the creativity literature, these requirements are novelty, usefulness, and value produced for the user of the product (Frederiksen & Knudsen, 2017). In the words of the authors, "the usefulness-to-the-recipients requirement is rarely found in the innovation literature" (p. 3). These examples show that the goal of the evaluation must be kept in sight while constructing assessment instruments. Additionally, the review analysed above (Dean et al., 2006) outlined guidelines for feature selection. The main decision points are the selection of the construct which one aims to measure and the selection of the features which are relevant for the experimental task. The purpose of the research determines how many and which criteria are chosen to be included in the assessment tool.

Applying these notions to the present premises, the rationale behind choosing the focus of the research will be explained. First, creativity was chosen as the construct to be measured (and not, for example, the quality of the ideas). This is due to the theoretical framework outlined in *Chapter 1*. The present research addresses the cognitive mechanisms involved in creative idea evaluation, not the ideas presented as the stimuli. The research is designed to learn more about what information sources are used to inform judgments made about creativity and how they are used.

Second, the domain of urbanism was chosen for the investigations. Here, the aim was to produce realistic stimuli. The target group of the first few experiments were university students in London, who are coming from vastly different backgrounds. A common experience shared by all of them is living in one of the biggest cities in the world. Thus, the domain related to improving urban lives was selected, as presumably the students have some basic knowledge and interest regarding this topic.

Finally, the assessment criteria were selected. Four criteria are used in this research and they were selected with a goal of both covering the most relevant dimensions linked to creativity and constructing a pragmatic, off-the-shelf tool for assessing urban ideas. Along these lines, two criteria, originality and utility, are domain-general and core to the definition of creativity. The first criterion measures how novel, unique, and surprising is the idea and the second criterion assesses the functionality of the idea. The other two criteria are specific to the domain, which is improving urban lives. These are concerned with the feasibility of the creative idea: scalability and riskiness (i.e. whether a project would be implementable and sustainable). Scalability stands for the likelihood of the creative product to penetrate multiple regions, which is a requirement for spreading rapidly. Low risk means a high probability for the product to get implemented; this criterion can be measured by how positive the outlook on the project is.

These criteria were drawn from field studies observing the criteria used by venture capitalists (Kaplan & Strömberg, 2000).

In sum, this section explained the principles based on which criteria for the assessment of creativity are selected at different premises, then used the insights to describe the purposes of the present research and outlined the thought process behind the main decisions made about the research design.

Zooming out, the present chapter inventoried the most important issues which must be considered for selecting the methodological approach of the present research. The takeaways from this chapter are that both domain-general and domain-specific features should be included in the judgment; that researchers must come up with a research design which accounts for the variability found in the ratings; and that the aim of the research could be to explain as much variance in the ratings as possible and to inform cognitive theory about the mechanisms behind creative idea evaluation.

The reasoning about the selected assessment method will be continued in *Chapter 4* with the description of the creation and validation of the paradigm. Before that, *Chapter 3* continues the outline of methodological considerations, however, with a narrow focus on one question. The question the next chapter is considering is that given the criterion problem of creativity research (McPherson, 1963, Shapiro, 1970), what alternatives could be used as substitutes for the criterion? It will enumerate the existing solutions and attempt to propose new ones.

CHAPTER 3

THE CRITERION PROBLEM OF CREATIVITY RESEARCH – USING ALTERNATIVE METHODS AS SUBSTITUTES

3.1 Introduction

To measure creativity, one needs to know what creativity is. However, finding the common denominator in all aspects of creativity is a challenging task. Creativity researchers spent decades of efforts on boiling down what the essence of being creative is (as demonstrated by the abundance of theories and definitions outlined in *Chapter 1*); and in the meantime, the construct needed to be measured at various fields, from school to workplace, even without a consensual definition. As *Chapter 2* documented it already, there is a lack of objective criterion/criteria against which a creative product can be measured up to. This chapter goes into further details about the problems with the measurement and explores what could be done to get around these problems.

Historically, to resolve the lack of conceptualisation which could have been translated into an operational definition of creativity (Amabile, 1983), proxies of creativity were used in experimental settings. The protocols included either divergent thinking (DT) tests, which were designed to capture only one aspect of the construct but could be scored objectively, or other creative production tasks which were more ecologically valid but were difficult to evaluate. To resolve the latter issue, expert judgments were adopted to academic research. The reason for asking experts to evaluate creative products is that according to the mainstream approach, 'one can recognize creativity if one sees it' (Amabile, 1982; Csíkszentmihályi, 1990; Cropley & Cropley, 2008). Instead of relying on criteria, experts were confided with the evaluation of creativity.

The reason why tests used to not involve direct criteria for creativity was the lack of such criteria. The infamous 'criterion problem' (McPherson, 1963, Shapiro, 1970) denotes the methodological struggle in which researchers and practitioners should come up with pre-defined objective criteria against which the creative product could be measured. The issue here is that the most creative solutions are often unthinkable in advance, thus constructing criteria to measure them before they exist is troublesome. A further problem pointed out by Amabile (1982) is that even if there are criteria for assessment at hand, they are still not going to be used in an objective manner but rather according to the internalised conception of creativity possessed by the judge. She concluded that an *objective* scoring of creativity is not possible since the judgment is largely based on the specific social context. Similarly, Csíkszentmihályi noted that "we cannot study creativity by isolating individuals and their works from the social and historical milieu in which their actions are carried out" (2014, p. 47). Hence, a measurement technique avoiding the use of criteria was crafted and, in a few decades, it became the mainstream method of evaluating creativity in academic research.

3.2 Avoiding the Use of a Criterion: the Consensual Assessment Technique as the Mainstream Assessment Tool of Creativity

Amabile advocated that the judges of creativity should embrace the subjective aspect of the process and rely on their ability to detect the signal of creativity when they perceive it. She has defined creativity from a practical, measurable point of view: "A product or response is creative to the extent that appropriate observers independently agree it is creative. Appropriate observers are those familiar with the domain in which the product was created or the response articulated" (1982, p. 1001). With this approach, the Consensual Assessment Technique (CAT, Amabile, 1982, 1983, 1996) was born.

The procedure of administering this methodology comprises of two basic steps (Baer & McKool, 2009). First, a set of creative products is generated, which might be anything from musical pieces through sculptures to mousetrap designs. All participants, who are usually non-experts or quasi-experts, receive instructions about what they need to produce and are provided with the required materials if required. Subsequently, experts of the corresponding field must judge how creative each product is. Usually, creativity is judged using a Likert-scale whose range is set by the creativity researcher. Judges are encouraged to use the full scale while rating multiple products. Seemingly, it is a straight-forward process.

Additionally, some ground rules were laid down to standardize how the expert judgment should be made. First, I describe these rules, then I interpret them from a critical point of view. To begin, the experts must work independently and cannot interfere with each other's judgment in any way. More importantly, they cannot be asked to explain how or why they cast their votes. Any choice made is considered as final and cannot be questioned under any circumstances. The only guideline provided is that judges should use their "mysterious expertise" to judge creativity, that is, only their understanding of the field should guide their judgments, nothing else. That is, providing with any additional instruction, explanation, or assessment criteria is prohibited.

One should also note that ratings are cast in a relative, not in an absolute manner (for further details about how this might influence judgments, see *Chapter 7*). In other words, judges are only ranking the products from the given pool, the judgments cannot be and should not be interpreted as standalone creativity ratings. This methodology is only suitable for making *internal*, group-level ratings, i.e., even the highest received score means only that the item is the most creative in the pool but can be nevertheless quite weak if contrasted to any *external* standard (Baer & McKool, 2009).

To capture the innovative value of this methodology, it will be outlined now what exactly was gained by starting to use it and how did it become the mainstream paradigm of evaluating creativity. Then, the next section will detail what is on the other side of the coin, i.e., how did the success of the CAT lead to a methodological impasse in creativity research.

There are several factors accounting for the popularity of CAT. Above all, it takes the bull by the horns: it is one of the few evaluation methods which directly assesses the full construct of creativity. In other words, the CAT is not restricted to measuring a subcomponent of creativity or only a related aspect. Further, it makes the lab assessment of creativity largely similar to the real-world assessment of creativity. Adding these two aspects together suggests that it is an ecologically valid way of measuring creativity. Due to its simple design, it is applicable to many different domains, more recently its use has even been expanded beyond artistic creativity (Lee, Lee, & Youn, 2005; Tan et al., 2015). If appropriate experts are in reach, the conduct of CAT becomes uncomplicated. One of its beneficial aspects is its feasibility.

On the other hand, several concerns can also be raised regarding the CAT. This section plots the problematic aspects of using this methodology, while the next section outlines the wider implications of how using CAT has affected creativity research, with a focus on studying the evaluation of creativity. The first group of concerns regards the sampling of the experts. From the description of the method, it is not transparent what exactly the threshold of expertise should be – the decision seems to be left to the creativity researcher. Since convenience is a major factor by sampling the participants of any research endeavour, the selection of not sufficient judges might not only compromise the results but also blur the findings included in the wider literature. It is also not clear how many experts are required for each investigation. Although naming them a group implies there must be at least three experts involved, there is no explicitly stated minimum which each creativity study should consider. Since the number of experts is almost never higher than fifteen (e.g., Kaufman, Gentile, & Baer, 2005; Kaufman et

al., 2008; Runco, McCarthy, & Svenson, 1994), perhaps a meta-analysis of the studies employing the CAT could be conducted to check whether such a low number of judges is acceptable to draw conclusions from.

The next concern is related to the notion that the experts should work independently from each other. It is trivial that if the consensus among independent judges is the dependent variable, judges should not communicate in any way during the test. However, the independence of the judges cannot be secured with this single instruction. Realistically, the expert pool of the CAT is often recruited from the same lab or company. Thus, there is a good chance that the independent experts know each other quite well, have professional discussions regularly, have formed each other's tastes and knowledge, as well as share one to many group memberships. These experiences are likely to lead to shared mental models which might stand behind their consensus. Given these factors, many possible confounds might be involved by the expert judgments.

The second group of concerns is philosophical and regards the way in which creativity must be judged by the experts when administering the CAT. Here, the underlying assumption is that creativity is an objective quality, which is revealing itself in different products and appropriate judges are equipped to recognise it (Csíkszentmihályi, 1999). This assumption induces several questions: is creativity really a similar quality to a colour, which might be detected visually? Is creativity something which can be perceived or is creativity rather something which is constructed on a higher cognitive level? Further, if creativity is linked to its recognition, if a product does not get public recognition, then does it cease to be creative?

It is difficult to locate *where* creativity happens because neither the individual, nor the social context or the domain can be singled out (Csíkszentmihályi, 2014). According to the systems view of creativity, all these components are entangled since "without a culturally defined domain of action in which innovation is possible, the person cannot even get started.

And without a group of peers to evaluate and confirm the adaptiveness of the innovation, it is impossible to differentiate what is creative from what is simply statistically improbable or bizarre" (p. 48). Investigations using the CAT are focusing on the product instead of the person. Although the domain is taken into account, the same cannot be said about the social context.

Finally, the third group of concerns regards what the judgments of the projects imply. On a practical level, the guidelines of the CAT explicitly state that the creativity judgments should not be taken at face value but that these judgments are only orienting inside the pool of products and cannot be translated into standalone assessments. However, one might question that if obtaining absolute creativity ratings would be the ultimate goal of any stakeholder/researcher and given that the CAT is declaratively not apt to provide such, then should not be another method which can provide this information too?

In the studies featuring CAT, the products under scrutiny are created for the purpose of studying creativity. It is explicitly stated that the judgments provided cannot be generalised outside of the pool of products. Conducting measurements by making 'mock-ups' of creativity assessment raises a theoretical issue. Namely, the logic of psychological research is that paradigms are sought to organize empirical data in such way that through analysis the assumption of a theory can be verified or falsified. The CAT does not belong to the group of such paradigms, as after administering it, researchers do not get to know anything about the process, only about the outcome. It is almost as if creativity would happen below an opaque bell jar - once the process has finished, the jar gets lifted and creativity can reveal itself to the public. Some reasonable doubts can be drawn about this. First, is it always possible to judge creativity divided from the story behind the creative product? The risk is that only simple products (e.g., drawings created by pupils) can be assessed using this method. If expert judges cannot be provided with any accompanying information, then more abstract products might remain unseen. If it is not evident what a creative product represents (e.g., a dance, installation),

then a crucial step of understanding might be missing without providing any additional explanation of the context.

To sum up, although the CAT is a simple and compelling method to evaluate creativity, its applicability is limited, and its use does not inform cognitive theory.

3.3 Reasons for Using Criteria for Creativity Measurements

The main argument of this section is that historically, having had no agreement on what constitutes creativity, the CAT was an adequate methodology to measure creativity in the best possible way. However, making it the ultimate tool of evaluating creativity also avoided the core questions academic research should address and reduced the level of investigations to a purely practical, atheoretical level. The CAT is applied since more than 30 years, and during the long decades the understanding of creativity in the scientific community has increased with a large extent. From this point of view, it seems counter-intuitive why an old paradigm would be still in the main stream of research instead of being replaced by methods which are informed by the theoretical advancements. It seems that the messiness associated with finding objective, ultimate criteria for the evaluation of creativity has kept most researchers away – apart from a handful of attempts (e.g., Cropley & Kaufman, 2012; Haller, Courvoisier, & Cropley, 2011). The CAT can be used as a shortcut to evaluate creativity but from a theoretical point of view, it is rather a temporary solution which might give the false comfort as if finding an ultimate solution would not be required. Even if it is accepted that the CAT is the optimal methodology which should be adopted on a large scale, one should note that doing so cannot be feasible due to the limited number of experts available and due to the limited amount of funds available to compensate these experts for their efforts. On top of that, even if there would be more financial sources to support a widespread use of the CAT as a general research methodology in all

creative evaluation studies, the time-consuming nature of the paradigm would still prompt researchers to look for a more scalable solution.

Thus, while it is not difficult to see that even on practical grounds, the CAT cannot offer an ultimate solution, numerous studies are still occupied with providing a more and more fine-grained account of how it can be used (e.g., Baer, Kaufman, & Gentile, 2004; Jeffries, 2017). In the meantime, only limited resources are left to work on the still unresolved theoretical and methodological issues. Focusing on the underlying principle of the CAT, which is that one can recognise creativity if one sees it, the question arises: how does this recognition happen? It would be desirable to decompose the recognition process to its components and this process should start with the identification of the psychological processes contributing to it. Another reason for why a scientific understanding of how creativity emerges is required in the near future is that, sooner or later, this job will be outsourced to machines. If researchers give up on finding an at least partly objective way to quantify the amount of creativity in a product, the computerisation of the evaluation process cannot be started. Amabile's observation regarding the rating of creativity being socially determined was spot on and it means that a good model of creativity should include a quantified measure of the interaction between the social context, the judge’s knowledge and the creativity signal.

Indeed, she has acknowledged that it might be possible "to identify particular objective features of products that correlate with subjective judgments of creativity or to analyse the nature of subjective correlates of those judgments" (Amabile, 1982, p. 1001). This can be only corroborated in 2017, as nowadays there is a consensus about the defining traits of creativity in the literature, at least one of the novelty/originality and the usefulness/appropriateness components being involved in all of them (e.g., Corazza, 2016; Runco, & Jaeger, 2012; Simonton, 2012). Still, the consensus in definition does not translate to a consensus in approaches in the selection of creativity-related evaluation criteria (cf., Kudrowitz & Wallace,

2013, Table 1). As demonstrated in the Existing Taxonomies of Features section of *Chapter 2*, the variability of approaches in selecting creativity-related criteria shows that there is no overall agreement on what the standard criteria for measuring creativity should be. Often, it is up to the researchers what criteria to use as there are only a few suitable measurement tools and most of them are general tests which need to be adopted to the specific domain and task. To combat these challenges, in the present thesis two domain-general and two domain-specific features have been selected, apart from measuring creativity directly (see *Chapter 2* and *4*). In the remainder of this chapter, alternative methodological frameworks will be considered. First, the Lens model is discussed, which is rooting back to the '50s and has been successfully used in multiple domains of judgment and decision-making research ever since.

3.4 Probing Whether Expert Judgment Is a Suitable Substitute of the Criterion via the Lens Model

This section deals with Brunswik's Lens model (1952; Hursch, Hammond, & Hursch, 1964) which can be used to assess the accuracy of human judgment. The model is presented here as it is a good tool to analyse what information is extracted from multiple cues to form the basis of a judgment. Also, because it quantifies how much each criterion is contributing to the judgment. The Theory of Social Judgment, on which the model is based, was already introduced in *Chapter 1*. Below, the model's application to creativity research is discussed.

The Lens model was selected as a possible framework for the thesis research since it has been applied widely and because some intriguing similarities were found in the structure of the model with our spontaneously constructed data analysis strategy. The Lens model lent itself as a suitable mathematical framework to take a closer look at the raters' internalised criteria, i.e., to explore where does the difference between experts' highly informed judgments and non-experts' naïve judgments stem from. In fact, this model allowed for gaining a deeper

understanding about how both the group of experts and non-experts interpreted the provided criteria and compared what patterns emerged, which helped to determine the extent of objectivity in the evaluations (the model is illustrated in Figure 2).

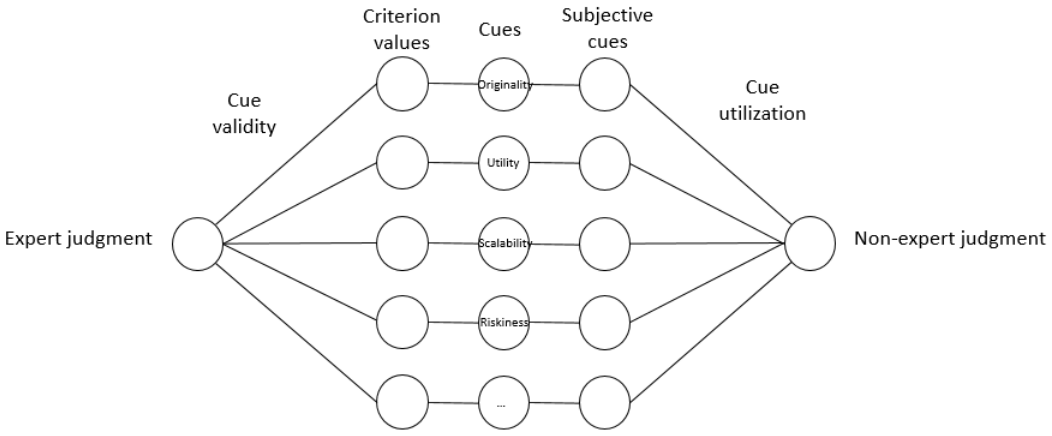

Figure 2. The Lens model applied to the current research paradigm. Cues are the creativity-related features.

To our knowledge, this thesis is the first attempt to apply the Lens model to creative evaluation, despite the model being consistently used to study social judgments on fields stretching from medical decisions to moral judgments.

As for the application of the model, there was a clear match between the creativity-related criteria used in the empirical chapters with the cues used in the model. On the other hand, choosing the criterion value proved to be somewhat problematic. The word criterion was already mentioned in the sections above, specifically because creativity research struggles with not having a clear-cut criterion to mark the threshold between what is creative and what is not. Thus, the best available proxy was selected as the criterion: the judgment of experts. One could raise the issue that this approach seems largely similar to the CAT, however, while the CAT explicitly states that judges should not be provided with any specific criteria or training, here,

exactly the opposite happens. Expert judges need not only to rate the creativity of products but all cues contributing to the judgment too. With the present application of the model, the use of criteria between two groups (experts and non-experts) becomes comparable.

Although it was argued in this chapter that using expert ratings can substitute the missing criterion for creativity ratings, a few caveats must be noted. I do not claim that the use of expert ratings can replace a criterion drawn from the environment – I am aware that this solution is not ideal. Nevertheless, to make a step forward, the investigation was started with the best available proxy. When making this choice, it was acknowledged that there are some serious limitations associated with the use of expert ratings. The first one of these limitations is using simulated/hypothetical values for the cues and the criterion (Karelaia & Hogarth, 2008).. The ideal workflow would be to generate predictions, then verify or falsify them by obtaining direct measurement values from the environment. The problem with using expert ratings as the criterion of creativity is that when the ratings of unexperienced judges are tested, the expert ratings cannot inform about the true state of the environment; they only provide with another set of predictions. Assumedly, these predictions are more closely representing the true state of the environment than the non-experts' ratings but the expert ratings are subjective (e.g., Amabile, 1982) and hypothetical too.

Based on this, one could say that since creativity is only a social construct and can only be captured as a subjective preference, then there is no evidence that experts are more *accurate* judges of creativity than lay people are. It has not been shown empirically what aspect of expertise is what enables experts to judge creativity better than lay people do. We might even come short of direct evidence about how experts are able to judge creativity more correctly, not only differently (e.g., Cropley & Kaufman, 2013; Kaufman & Baer, 2012). In practice, experts are thought to be more accurate estimators of creativity than lay people because experts have more experience, which enables them to judge both whether a project proposal is original and

whether it would be feasible to implement it. Due to the time spent in a certain domain, they are better equipped to estimate the associated risk with a project. A further argument for using expert judgments is related to a key concept of Brunswik's work. This is representative design, which means that any experiment should be created by representing the outside world in order to be able to generalise its results to a larger population (Brunswik, 1952; Hammond, 2001). In other words, a lab experiment should be designed to mirror how values and structures are observed in the ecology. Applying this principle to creativity research, studies investigating how creativity is evaluated should be streamlined with the factors of practices of "real-world" evaluation of creativity. That is why expert ratings were taken as the assessment method of creativity.

A possible criticism for using expert ratings as substitutes for the criterion of creativity is that experts are not unerring either. There is plenty of evidence showing that expertise as a mental set can impede the solving of a problem (Wiley, 1998; Bilalić, McLeod, & Gobet, 2008). Anecdotal evidence is also available about the close-mindedness of experts; e.g., art history provides us with numerous stories regarding the immediate rejection of new trends in painting (such as impressionism, cubism, or fauvism) coming from the experts. As Kaufman & Baer (2012) notes it, “How could we expect experts to judge those creations that might be changing the very rules that helped establish their own standing in their field?” (p. 84). Again, the unavailability of objective metrics makes it impossible to check the true state of the environment, therefore, the consensus in the subjective ratings is the only method which can make these ratings ‘relatively objective’. Whether there is a ’true score’ of creativity could only be found out by taking a closer look at the environment and taking direct indicators of creativity from there (possible ideas for this are outlined later in this chapter).

Apart from the criterion value, there are certain issues related to the use of cues in the research context of creativity. One of them is related to the inability to capture the criteria

directly and not as an estimate cast by judges. The result of this difficulty is a lack of hard numbers. In the words of Karelaia and Hogarth (2008, p. 407),

> "an important dimension of many tasks involves identifying and assessing levels of relevant information (Einhorn, 1972). Therefore, one can distinguish between studies where cues are "given" as opposed to "achieved." For the former, decision makers are provided with the explicit values of the cues by the investigator. For the latter, the values of the cues need to be inferred—and often even identified—by decision makers."

Certainly, the latter is the case by creativity judgments. E.g., originality and appropriateness are quite difficult to quantify, and the judges need to get through a lengthy process to achieve the values. They need to identify and select the relevant information linked to each cue/criterion from a noisy environment.

Further, an appropriate criterion is linearly linked to the cues (the cues can be its predictor variables). To be able to use the experts' ratings as the criterion for non-experts' ratings, the two groups must use similar cues to inform their judgments. The validity of these concerns could only be warranted by conducting empirical research and checking which hypotheses are confirmed by the results.

Another concern linked to this methodological framework is that a criterion drawn from the environment is a single score (even if it includes the measurement error). This looks a bit different in case of the present procedure. Although direct creativity ratings were obtained from a set of domain-specific experts (n=8), their ratings are estimates and can only be used if they are highly consistent (i.e., if the standard deviation in a project's ratings is low). In other words, the judgments obtained from individuals with at least 10 years of experience in a domain related to cities such as urban planning or civil engineering can only be used as the criterion value if a high degree of consensus is found in their ratings. Otherwise, it cannot be determined what value should be selected. Although a lot of researchers have reported a rather strong consensus amongst the experts rating different creative products (Amabile, 1996; Baer, 1998; Baer,

Kaufman, & Gentile, 2004; Hennessey & Amabile, 1999; Kaufman, Lee, Baer, & Lee, 2007, Kaufman, Baer, & Cole, 2009), in a few studies, very little or no agreement was found (Hickey, 2001; Jeffries, 2017; Lee, Lee, & Youn, 2005).

Even if it can be assumed that experts agree on the evaluation of a given product, employing the Lens model to investigate the evaluation of creativity still does not become problem-free. The logic of the Lens model is that it assesses the match between multiple judgments made about the environment and the indicator of the environment (criterion) (Karelaia & Hogarth, 2008) or the so-called judgmental achievement. Amabile argues that the judgment of creativity is determined by the "test constructor's or scorer's intuitive assessment of what is creative and not according to the objective criteria of novelty, appropriateness, satisfyingness, and so on" (Amabile, 1982, p. 999). Based on this, if experts' and non-experts' judgments are made according to their internalised conceptualisation of creativity, then the Lens model analysis becomes only a comparison between two groups of subjective judgments, not a comparison between subjective judgments and the criterion drawn from the environment. Since the structure of the model is to compare estimates to one fixed value, comparing two sets of estimates might be troublesome. Translating this problem to mathematical terms, the Lens model equation's (presented as Equation 1) outcome measure is the correlation between the judgment and the criterion (Tucker, 1964, p. 528 as cited by Karelaia & Hogarth, 2008):

$$r_a = G\ R_s R_e + C\sqrt{1-R_s^2}\sqrt{1-R_e^2} \tag{1}$$

The components are depicted in Figure 1 presented in *Chapter 1* and can be defined as the following (based on Karelaia & Hogarth, 2008; Kaufmann, Reips, & Wittmann, 2013):

r_a=achievement index (i.e., the correlation between a person's judgments and the criterion),

G=knowledge index or the correlation between the predictions of both models: the predicted levels of the criterion and the predicted judgments,

R_e=environmental predictability (i.e., multiple correlation of the cues with the criterion or the degree to which a judgment can be made based on the cues),

R_s=consistency, response linearity, or the reliability of judgments (i.e., the multiple correlation of the cues with a judge's estimate or in other words, the extent to which a judge reliably reaches the same judgment based on the same pieces of information),

C=an unmodeled knowledge component that signifies the correlation between the residuals from the environmental predictability component and the consistency component.

To calculate the achievement index, one criterion value is compared with one judgment and this step is iterated many times. A requirement for using expert judgments as the criterion value of creativity is that the range of the ratings cannot be too wide or otherwise there would be no basis on which to decide what should be the single value used as the indicator of the environment. The descriptive data about the ratings collected from the domain-specific expert judges showed a large variance in the creativity ratings (cf. Table 4 in *Chapter 4*). A moderate degree of absolute agreement was found after conducting a two-way mixed intra-class correlation. The coefficient was .497 with a 95% confidence interval from -.008 to .800 ($F(14,98)= 2$, $p=.026$). Krippendorff's alpha was obtained as another measure for inter-rater agreement. Taking all 15 projects, $\alpha=.058$, which signals a low degree of agreement. Item analyses were also conducted to identify projects whose evaluation indicated a higher level of agreement. Inter-judge and judge-total correlations were computed to assess the internal consistency of the domain-specific judges (see Table 6 & 7 in Ch. 4). Inter-judge correlations ranged between .08 and .33, indicating vast individual differences in the experts' judgments. Judge-total correlations ranged from .22 to .78, signalling mixed interpretations. All these

results confirmed that the variance in the ratings were too high to use averages as the criterion values.

The other requirement of conducting the Lens model analysis is that experts and non-experts must use the same cues to inform their judgments. This can be tested by using only the 'cue utilization' side of the model (as depicted in Figure 2) and treating both experts and non-experts as participants providing two sets of judgments. The data presented in Experiment 1 of *Chapter 5* shows that the non-expert and the domain-specific expert participants of this research were not found to inform their judgments based on the same criteria. While non-expert participants were found to incorporate all four cues (features) to their judgments made about creativity, no evidence supported that domain-specific experts would be using any of the cues. (The definition of the four cues were provided to all participants before rating them.)

In sum, the Lens model is a potential tool for assessing the similarities and the differences between experts' and non-experts' judgments of creativity, and to scrutinize what might stand behind the inconsistencies found between them. It served as a suitable to model to test the requirements whether experts' judgments could be used as the objective criterion of creativity; this was not found to be the case. (However, studies measuring creativity with a different methodology, e.g., Likert scales, might be suitable for further investigations using the model.)

The results indicate that another method should replace expert ratings in the quest for searching objective criteria to evaluating creative ideas. As the empirical investigations should not be stopped until the criterion problem (Shapiro, 1970) gets resolved, for now, a good alternative might be to acquire criterion values directly from the environment, and possibly, from multiple sources (such as amounts of investment received, rankings in different competitions, generated donations, etc).

3.5 Further Alternatives for the Criterion of Creativity

An issue with aiming for direct creativity assessments is that they might not exist in real life. People only rarely score products according to their creativity, that is why it is problematic to gain inspiration to the construction of research methodologies by real-life practices. However, funding decisions related to creative products indicate in a straight-forward manner what their perceived values are, therefore, they could be used (and are used such as in Mollick & Nanda, 2015) as a good proxy of the 'true score' of creativity. These investments or budget allocations indicate the result of a very complex judgment process. The task of psychologists is to disentangle the different factors contributing to the final judgment, including not only the weighting of the factors but also the relationships amongst them.

In conclusion, the way creativity is regarded is central to its measurement. This thesis was designed for the study of the recognition/perception aspect of creativity. However, others have questioned whether there is anything to perceive at all, that is, it was suggested that the creativity of the product might not be there first but is the result of the judgment process itself. This assumption poses an interesting constraint because it states that a product hidden from an observer is not creative; namely, "product creativity only exists if there is a judgement of a product (based on a set of criteria) and products cannot be inherently creative (without judgement)" (Horn & Salvendy, 2006, pp. 396-397). This statement is in sharp contrast with the ’rater effect’ approach of creativity assessment (Hung, Chen, & Chen, 2012; Long & Pang, 2015), in which the judge of creativity is a source of potential error, i.e., the subjective component of the process is treated as one which blurs the picture, not as one which creates it. Csíkszentmihályi warns that research isolating the product from its evaluator can never capture what is at the heart of being creative. He notes (1999) that many evaluation research efforts are conducted based on a false assumption. The assumption is that there would be an objective quality called „creativity” to be found in the product. Then, the job of the judges of creativity

would be to perceive and recognise this quality. Instead, he draws attention to the notion that „expert judges do not possess an external, objective standard by which to evaluate creative responses. Their judgments rely on past experience, training, cultural biases, current trends, personal values, idiosyncratic preferences. Thus, whether an idea or product is creative or not does not depend on its own qualities, but on the effect it is able to produce in others who are exposed to it" (p. 314). Therefore, he recommends that researchers should not measure the creator, the product, or the evaluator in isolation, but all these agents of creativity should be included in one system. This suggestion is acknowledged by conducting the present research, however, the approach taken here is to focus on the product while also registering the idiosynchronicities of the evaluators.

The current methodological problems might be connected to the conceptual controversies outlined in this chapter. E.g., the variability found in the creativity ratings, which gives a lot of headache to the researchers, can be interpreted, according to the systems view, as stemming from the interactional, dynamic nature of creativity. If creativity lies in interactions, then creativity researchers should come up with tools to investigate the dynamic process of creative systems. One possibility for doing so could be the application of network research to creativity. As Csíkszentmihályi (2014, p. 49) notes it:

> "Where does the information that gives us the ability to make sophisticated judgments come from? The information does not seem to be in the object itself. If we think about it, the reason we believe that Leonardo or Einstein was creative is that we have read that that is the case, we have been told it is true; our opinions about who is creative and why ultimately are based on faith. We have faith in the domains of art and science, and we trust the judgment of the field, that is, of the artistic and scientific establishments. There is nothing wrong with this, because it is an inevitable situation. But by recognizing it, we must also accept some of its consequences, namely, that any attribution of creativity must be relative, grounded only in social agreement. And from this it also follows that social agreement is one of the constitutive aspects of creativity, without which the phenomenon would not exist."

To wrap up, this chapter has outlined how expert ratings are and could be used for evaluating creativity, as well as why people's internalised conceptualisation of creativity should be brought to the surface by rating creativity-related features and thereby externalising what cues are implemented to form judgments about creativity. Finally, theoretical assumptions underlying creativity research were discussed to find new potential methods for investigating creativity.

What needs to be carried forward from this section is that creativity cannot be rated with one value only, however, asking raters to do so might reveal meaningful information about how the evaluation process takes place. In the empirical chapters, different contextual manipulations are introduced which are aimed to enhance the agreement between different groups of participants. This approach was motivated by the current research trends in which conditions are searched for aligning non-expert participants' judgments with experts' judgments. Using one side of the Lens model, it became possible to test the requirements of using experts' judgments as the criterion of creativity. The requirements were not fulfilled in case of the current research, however, other methodologies applied on a different domain of creativity might be able to fulfil them.

The next chapter is getting to the nuts and bolts of the paradigm developed for conducting the empirical studies. The entire creation and validation process of the paradigm, as well as the rationale of each study will be outlined. A list of all hypotheses which were tested empirically will also be enclosed.

CHAPTER 4:

RESEARCH PARADIGM AND HYPOTHESES

4.1 Introduction

This chapter outlines the methodological considerations behind the construction of the research design. It also presents the details of the pilot work that was used to determine the materials, design, and procedure for the main 4 experiments that will be presented in the three empirical chapters (*Chapter 5-7*). Previously, *Chapter 2* summarised how creativity could be conceptualised for the measurement purposes in. In *Chapter 3*, further methodological concerns were introduced, alternatives of the criterion of creativity were discussed. Now, this chapter describes how the measurement of creativity was operationalised for the purposes of the present premises.

The objective of this research project is to holistically measure the evaluation of creative ideas based on selected criteria. A crucial aspect of the measurement is to avoid aesthetic judgments and focus on tangible projects. Thus, our attention was turned to functional creativity (Cropley & Cropley, 2005, 2008). Creative evaluation studies were found to range widely regarding the type of stimulus they use: advertisements (Caroff & Besançon, 2008), apparel design illustrations (Freeman, Son & McRoberts, 2015), websites (Zeng, Salvendy, & Zhang, 2009), restaurant interior designs (Horng, Chou, Liu, & Tsai, 2013) or even sketches of toasters (Kudrowitz, Te, & Wallace, 2012) were all used to study what counts as creative. We also had to choose a specific domain to specify the paradigm and looked for a relatively highly codified domain (Kaufman et al., 2013), which at the same time also seemed to be approachable for naïve judges. Although a few paradigms were considered for conducting the research, a novel one was created. This decision was motivated by the realisation that there are only a few

research paradigms available, also that it is a common practice among creativity researchers to adopt the assessment criteria which are the most relevant for their research projects.

As the research was conducted in London, one of the biggest cities in the world, the idea to measure how creativity is demonstrated in cities, i.e. researching creativity in the domain of urbanism lent itself. By the time this thesis was written, it was noted with delight that a largely similar paradigm to ours emerged from the innovation literature (Frederiksen & Knudsen, 2017). As discussed in *Chapter 1*, although the structure of the creativity assessment and the selected criteria in the Danish study are comparable to ours, the chosen domain (technical innovation of renewable energy) and the analysis of the data paved ways for making different inferences than in our case. This example illustrates the need of adapting the measurement tool for the exact research purposes: main research questions of this thesis regard the criteria informing judgments made about creativity and their respective weighting; and the already established paradigms were not used to measure the weighting of the criteria the way as it was aimed in the present project (details will be provided in *Chapter 5*).

The next section outlines the construction of the experimental paradigm. The aim was to collect a representative set of tangible ideas on how cities can be more creative; to achieve this, project outlines were pooled together.

4.2 Creating the Paradigm

The outcome of the construction phase is a final set of fifteen project ideas, which were created based on proposals collected from an open-source platform, OpenIDEO (2011). All project ideas were entries for a competition on “How might we restore vibrancy in cities and regions facing economic decline?” originally. They were collected from the website and were edited into two paragraphs of texts. Subsequently, the validation process took place, and the entries collected from the website were transformed into proposals describing a single initiative, the set including samples ranging between 'hardly creative at all' to 'very creative'.

In all empirical chapters presented in this thesis (*5-7*), participants were instructed to become invested in the outcome of the realization of the projects evaluated by them. Realistic stimuli were selected for the purposes of the research despite of being aware of the potential problems stemming from the higher amount of noise associated. The stimuli were picked to ascertain a high level of ecological validity. In line with this, the stimuli included project proposals about opening local museums, creating guided tours, or establishing rooftop gardening projects. After outlining the objective which the paradigm was set up to achieve, I report the process of collecting and validating the project proposals below.

4.3 Selection of the Features

4.3.1 Construction of the Paradigm. To control for the influence of visual appearance, and to make the implementation of the experimental manipulation easier, the project outlines included a one-page brief without any graphical illustration. Each project was rated by a pool of domain-general experts (n = 16) on four features: originality, utility, scalability and riskiness (see Table S1). Admittedly, the criteria were self-selected to learn more about the judgment making process. The rationale behind selecting the features, as outlined in *Chapter 2* already, was to cover the most relevant dimensions but also to construct a pragmatic, off-the-shelf tool for assessing urban ideas. Along these lines, two of the four selected criteria, originality and utility, are domain-general and core to the definition of creativity, while the other two of them are domain specific because they are concerned with the functionality of the creative idea: scalability (i.e., the opportunity of growth) and riskiness (i.e. whether a project would be implementable and sustainable). Table 2 contains the explanations of the four features provided to the expert raters.

Experts validating the presence of the criteria in the paradigm were domain general experts of evaluation and the discernment of ideas: they were recruited with a snowball

methodology from junior and senior science staff at Queen Mary, University of London. Each feature was rated by them with one of the three responses: high level, low level and unsure. Experts were also prompted to indicate if any of the projects were diverging from the others in the pool so much that comparing them would become problematic.

Table 2

Feature definitions as presented to the experts and non-experts.

Feature	**DEFINITION**
Originality	The quality of being novel or unusual, unique. In the present case, whether the project is unprecedented in its environment.
Utility	The (perceived) ability of something to satisfy needs or wants. In the present case, one has to determine whether a project would satisfy the needs of a city.
Scalability	The ability of a proposed project to be enlarged to accommodate growth. In the present case, whether a project would be implementable on a larger scale (e.g. worldwide).
Riskiness	The probability that the project will be successfully implemented. In the present case, whether the project outcome is unsafe, uncertain or precarious.

Note. The riskiness feature denotes "low risk", thus the higher the riskiness rating, the lower the perceived risk is.

4.3.2 Validation Process. To establish the final set of stimuli, *N*=20 projects were rated in three consecutive rounds. After the first round of evaluation, preliminary results showed the categorical judgment of each feature by each project. Subsequently, the content of these projects was changed in order to cover the entire matrix of the possible combinations (e.g., low level of originality, high level of utility, low level of riskiness, high level of scalability), aiming for a final set of *N*=16 project proposals.

After the second round of evaluation, the optimal differentiation, defined as an $\alpha > .80$ consensus regarding the categorical ratings made by the experts, of these 16 projects has not been reached. Wording of the texts were changed for clarity on the level of features. *N*=1 project was excluded due to being substantially distinct from the rest in the set, as reported by the experts.

After the third round of evaluation, *n*=19 projects were rated and *n*=15 projects have been selected as the final set due to the consistency in their ratings. Fifteen projects were selected because a) one combination of the features, namely a low level of originality, utility, scalability, and a high level of low risk was found to be not applicable even after several attempts and because b) these projects received consensual ratings, $\alpha > .75$.

At the time, the expert ratings were obtained to validate the stimuli. However, later we wanted to compare whether experts and non-experts are rating creativity in a similar way and for doing so, obtaining creativity ratings from experts on the same scale as from non-experts became crucial. The first approach was to use the data we already had as a proxy to creativity ratings, thus it was checked in what extent can their feature ratings predict non-experts' creativity ratings.

4.3.3. Domain-specific Experts. However, in order to make a Lens model analysis possible, as outlined in *Chapter 3*, a criterion variable was needed, thus direct creativity ratings from the experts had to be obtained. Since the additional data collection was unavoidable, it was taken as an opportunity to kill two birds with one stone. Not only the missing ratings were collected but the expert pool was extended to domain-specific experts. Domain-specific experts were selected based on the following criteria: (1) they must be UK residents and proficient English speakers, (2) they must have at least 10 years of expertise in their domain, (3) their domain must be related to cities, e.g., urban planning, architecture, civil engineering, policy

making to cities, etc., and (4) they also must have some understanding of creativity, e.g., writing poems or publishing novels, composing music, coding programs, etc.

Domain-specific experts were identified via multiple rounds of Google X-ray searches for 'creativ* & city', filtering to UK results, then finding further links and people with a snowball methodology. The professional platform LinkedIn was also exhaustively researched to locate suitable experts. After collecting the contact information of all candidates using various customer relationship management techniques (e.g., lead generation databases), they have been contacted in a personalised e-mail shortly describing the purpose of the research. They were kindly asked to contribute to the research and were offered with a £10 Amazon voucher as compensation. The ratio of rejection to completion was 94:10. This resulted in a sample size of N=10 domain-specific experts, from which two participants had to be discarded due to providing incomplete data, resulting in a final data set of n=8 experts with M=20.38±10.37 years of relevant expertise. They are aged M=47.38±12.22 years, 3 of them are males.

Finally, please note that 2.5 years have passed between the data collection of the domain-general judges and of the domain-specific judges and they have not been matched according to any demographical variable. These might mean a relevant limitation toward the comparability of their ratings.

After outlining the methods of establishing the paradigms, the focus is moved now to the research questions which the paradigm was designed to address. First, the research questions will be identified, subsequently, the hypotheses related to each empirical study will be listed.

4.4 Research Questions

This thesis is exploring the basis on which people evaluate creative ideas. Three broader questions were explored. First, (Q1) in a noisy environment, what information do judges use to

evaluate creative ideas? What criteria is applied internally to form an overall creativity judgment? The present research investigated the link between creativity-related criteria and holistic creativity ratings. Criteria were selected to test to what degree can the evaluation of creative ideas be predicted by using them. This leads us to the second broader question of this thesis, which delves into (Q2) the weighting each one of the criteria is carrying towards forming the overall creativity rating. Both domain-general and domain-specific criteria were tested, and it was examined whether being original or useful is more important in an urban domain. The third large question addressed in this thesis (Q3) was seeking to understand how certain contextual factors influence creative idea evaluation.

As for the contextual factors, three research themes emerged. It was investigated how the level of motivation, the amount of available information about the study, and the manner in which the rating is cast influences the evaluation of creative ideas. In the next section, the hypotheses of each empirical study are outlined.

4.5 Rationale of each Study & Hypotheses

This section offers a full overview of the hypotheses which were investigated in the empirical studies (*Chapter 5-7*). They are numbered consistently, and a summary of the rationale behind conducting each study is provided here to make the connection across the studies explicit.

4.5.1 Chapter 5. In *Chapter 5*, the effect of motivation was tested. Extrinsic and intrinsic motivation were both enhanced by using incentives. Motivation was manipulated for two purposes. (1) In order to mimic conditions outside of the laboratory, in which experts are motivated to evaluate creative ideas on the basis of intrinsic and extrinsic motivation, we introduced manipulations comprising of the two types of motivation on the creative evaluation process. In addition, (2) in order to assess the extent to which experts and non-experts can be

brought into alignment, goal oriented motivational manipulations were introduced. The intrinsic incentive made it the goal of the participants that they invest the funds of their investment company in the most effective way, while the extrinsic incentive made it to their goal to do their best on the task to bring home the highest possible sum as a reward for their participation.

The following hypotheses were expected to be confirmed as the result of this study:

(H_1) Overall creativity ratings can be predicted from the four features we have identified as critical for the assessment of creativity of ideas (originality, utility, scalability, and riskiness), i.e. all four of them are contributing to the judgments made about creativity.

(H_2) Due to the subjective nature of creativity assessment, lay participants' creativity ratings can be predicted better from their feature ratings than from the expert judges' creativity ratings. I.e., the internal model comprising of how the four features are weighted differs as a function of expertise.

(H_3) Enhancing the intrinsic and extrinsic motivation of lay participants through the use of incentives will result in non-experts' task performance becoming more similar to experts' ratings as compared to non-incentivized baseline performance.

4.5.2. Chapter 6. In *Chapter 6,* the effect which the amount of available information has on rating creativity was tested. Particularly, two manipulations were used: first, the role of providing task-relevant information on task performance was investigated. Second, the role of meta-information with respect to the task's characteristics was investigated.

The description about the task-relevant features was manipulated by dividing participants to two groups: the first started the experiment with the Feature rating task, while the second group started with the Investment task. It was a within-subject design, so all participants completed both tasks eventually. Critically, in the feature rating task, participants

were provided with the name and definition of the relevant criteria for making their judgment about creativity in the second part of the experiment. The participants who started with the Investment task were not informed about the criteria before casting their creativity ratings. Two hypotheses were outlined about the potential effects:

(H_4) Those participants who were made aware of the name and content of criteria linked to creativity will cast feature and creativity ratings more closely linked to each other than the participants who were not made aware of which criteria are connected to creativity ratings.

(H_5) The different order of completing the Feature rating task and the Investment task provides a different amount of available information to the participants at the time of casting their creativity ratings. Therefore, the mean of the creativity ratings will be affected by the order in which the experimental tasks are completed.

Introducing different meta-information about the experiment conducted was assumed to influence the outcome in the following ways:

(H_6) It was assumed that if people need to evaluate the creativity of ideas, they make less certain judgments than what they would make if they were informed they have to evaluate business ideas per several dimensions, not mentioning creativity explicitly (‘nothing special’ condition).

(H_7) In addition, we expected people to give less coherent creativity ratings than viability ratings as they would have more expertise in estimating the usefulness/feasability than the creativity of an idea. For this, we expected to find a difference in the average of creativity ratings between the 'creativity' and the 'nothing special' groups, whilst no difference in the average of the viability ratings between the groups.

4.5.3. Chapter 7. In *Chapter 7*, the effect of making absolute vs. relative judgments about creativity was tested. The aim was to extend the scope of the evaluation process from absolute creativity judgments to comparative judgments, i.e., participants needed to rank which projects are more/less creative than others instead of simply judging them one-by-one. This procedure was motivated by ecological validity as the comparative approach models how creative ideas are selected most of the time, e.g., how grants are awarded or how the creative industry works. One group of participants had to shortlist the best 6 ideas, while the other group needed to shortlist the worst 6 ideas.

Three types of metrics were collected regarding the evaluation of creativity: (1) absolute creativity ratings were obtained (the same way as outlined in the earlier chapters), (2) the projects were shortlisted, which resulted in a relative ranking, and finally, (3) a budget of 100 coins was allocated amongst the selected 6 projects, which resulted in a relative weighting.

First, internal consistency in the ratings were checked to estimate their similarity, that is, to see whether the different type of creativity ratings would result in the same ranking of the projects (please note that the absolute creativity ratings were obtained in randomized order).

(H_8) The ordering of the absolute creativity ratings corresponds to making relative judgments about how creative the stimuli are compared to each other. That is, more than half of the 6 shortlisted projects was expected to be listed in the top/bottom six ideas too when converting the absolute judgments to an ordinal scale.

(H_9) The explicit rank ordering established in the shortlisting task aligns with the weighted rank ordering established in the budget allocation task with regard to the ranking across the projects. That is, the projects should be aligned in their ranking position to show the judgments are reliable.

Subsequently, the data was analysed according to task condition to determine which type of creativity ratings are aligned with each other.

(H_{10}) We expected the absolute creativity ratings to not differ between the two task conditions (shortlisting the best vs. the worst ideas), as the task instructions did not concern these ratings.

(H_{11}) However, we expected the shortlists to be severely influenced by the task condition. It was hypothesized there would be no overlap between the 'best ideas' and the 'worst ideas' shortlists, i.e., they would share less than one idea on average.

Finally, we wanted to explore whether experts' rankings of the projects would be similar to non-experts' rankings. That is, even if their absolute judgments were found to be largely dissimilar, it is possible that the ranking of the projects is better aligned between the two groups.

(H_{12}) It was expected that the rankings provided by domain-specific experts and non-experts would largely overlap. That is, an overlap of at least half of the projects in the best six ideas and the worst six ideas was expected between the two groups. It was not required the projects to be listed in the same ranking position.

To sum up, the three empirical chapters of the thesis are investigating three contextual factors which might influence how creative ideas are evaluated. First, it is probed whether motivation is a decisive factor contributing to the different way in which experts and lay people judge creative ideas. Second, it is investigated how task-related information modulates creativity ratings. Third, it is explored whether making judgments about creativity in an absolute as compared to a relative manner would result in distinct outcomes.

The next section of this chapter provides an account of how each project proposal presented as a stimulus was rated on average. This is to inform the reader what are the baseline tendencies of non-expert participants when rating the stimuli.

4.6 Project-wise Evaluation of the Stimuli

The creativity ratings were calculated by averaging all ratings in case of each project by all non-expert participants (*n*=447) contributing to any of the experiments. When ideas were ranked, the rankings stretched from the highest to the lowest mean in case of the best ideas and from lowest to highest mean in case of the worst ideas.

First, the raw ratings show quite a bit of fluctuation in the values cast by non-expert participants. The standard deviations range between 21 and 28, which is, counted towards both directions, covers approximately half of the scale. Based solely on the aggregated data, one might conclude that, on a group level, non-experts do not have a consistent concept on how to judge creative ideas. However, the mean ratings are roughly following the domain-general expert ratings, thus although most of the projects were classified to the mid-part of the rating scale and therefore the comprehensive creativity ratings cannot be set apart from each other distinctly, a direct comparison across the projects seems like a promising alternative method to find the threshold between more and less creative project proposals. Table 3 shows the descriptive data collected from all non-expert participants.

Table 3

Descriptive data for the ratings provided by non-expert participants, pooled together from all experiments

Project's name	Project's number	Minimum rating	Maximum rating	Mean rating	Std. error of the mean	Std. deviation
SIP Veggie Farm	1	0	100	70.14	1.07	22.71
The Neighbourhood Museum Collective	2	0	100	56.26	1.21	25.50
Intermodal	3	0	100	66.24	1.02	21.65
Street Signs With Stickers	4	0	100	36.60	1.27	26.82
miLES	5	0	100	65.26	1.06	22.42
School of ideas	6	0	100	50.00	1.20	25.43
Sidewalk Chalk Arts	7	0	100	46.38	1.25	26.50
Time capsules	8	0	100	29.32	1.24	26.19
Free Neighbourhood Blocks	9	0	100	44.29	1.35	28.48
Feel the city	11	0	100	44.47	1.37	28.86
A.R.T.S.	12	0	100	57.16	1.11	23.56
Pop-up Cultural Hub	13	0	100	50.83	1.12	23.74
Fablab	14	0	100	57.39	1.18	24.85
Movies to the park	15	0	100	33.10	1.21	25.60
Blogger Café	16	0	97	36.46	1.19	25.24

To check the reliability of the expert ratings, domain-specific experts were also asked to judge the creativity of the stimuli. Table 4 and 5 show the descriptive data collected from the domain-specific expert judges (*N*=8). What can be deduced from this data is that these experts were more severe judges than the non-expert participants; also, that there was a similar amount of variability, or in other words, agreement in their ratings as in the non-experts' data.

Table 4

Descriptive data for the creativity ratings provided by the domain-specific expert judges

Project's name	Project's number	Minimum rating	Maximum rating	Mean rating	Std. error of the mean	Std. deviation
SIP Veggie Farm	1	33	75	55.13	5.34	15.11
The Neighbourhood Museum Collective	2	24	75	48.88	7.24	20.48
Intermodal	3	16	80	44.25	7.66	21.67
Street Signs With Stickers	4	5	85	37.25	9.85	27.86
miLES	5	28	90	63.00	7.63	21.57
School of ideas	6	19	81	48.13	7.96	22.52
Sidewalk Chalk Arts	7	14	75	45.38	7.43	21.00
Time capsules	8	2	74	24.38	7.95	22.48
Free Neighbourhood Blocks	9	20	85	50.63	7.81	22.08
Feel the city	11	0	89	50.88	11.37	32.17
A.R.T.S.	12	34	75	53.75	5.90	16.69

Pop-up Cultural Hub	13	18	80	45.25	8.73	24.70
Fablab	14	9	95	56.00	9.28	26.24
Movies to the park	15	18	75	39.50	7.29	20.62
Blogger Café	16	1	60	30.63	7.63	21.57

Table 5

Descriptive data for the feature ratings provided by the domain-specific expert judges

Project's name	**Project ID**	**Originality (*M*±*SD*)**	**Utility (*M*±*SD*)**	**Scalability (*M*±*SD*)**	**Riskiness (*M*±*SD*)**
SIP Veggie Farm	1	36.88±19.56	42.63±30.05	69±23.76	34.13±30.3
The Neighbourhood Museum Collective	2	35±24.08	52.25±26.55	56.63±23.3	33.38±21.17
Intermodal	3	37.5±23.37	60.5±21.73	65.5±16.48	46.5±28.21
Street Signs With Stickers	4	50.88±27.33	61.63±30.16	55.88±19.32	44.25±22.45
miLES	5	44.75±24.75	65.25±27.74	63.38±25.24	51.75±21.68
School of ideas	6	48.88±31.71	64±24.72	76.25±15.95	36.13±20.72
Sidewalk Chalk Arts	7	43.63±35.76	47.13±28.71	69.88±24.42	34.75±25.84
Time capsules	8	29.25±25.09	64.38±28.13	54.13±24.53	48.25±29.26

Free Neighbourhood Blocks	9	42.25±27.89	55.63±36.78	48.75±28.89	31.88±20.29
Feel the city	11	53.13±29.79	59.75±22.49	58.75±30.62	42±22.08
A.R.T.S.	12	32.5±22.25	55±31.55	53±31.67	41.88±30.87
Pop-up Cultural Hub	13	42.25±31.06	49.88±26.23	73±15.03	41.25±31.14
Fablab	14	21.88±14.46	52±29.6	50±30.53	48±29.66
Movies to the park	15	44.63±27.26	57.88±34.54	60.75±24.35	38.75±29.51
Blogger Café	16	54.5±27.34	61.5±33.14	53.88±25.14	45.88±28.74

A moderate degree of absolute agreement was found by the domain-specific experts regarding the creativity of the stimuli. The two-way mixed intra-class correlation coefficient was .497 with a 95% confidence interval from -.008 to .800 ($F(14,98)= 2$, p=.026). Due to the not sufficiently high inter-rater agreement, item analyses were considered. Inter-judge and judge-total correlations were computed to assess the internal consistency of the judges. An overall reliability coefficient was also calculated. The results are summarized in Table 6 and 7.

Table 6

Internal consistency of the domain-specific experts

Judge	***M* of ratings**	***SD* of ratings**	**Inter-judge correlation**	**Judge-total correlation**
D-S Expert 1	51.93	22.73	-.21	-.20
D-S Expert 2	52.47	35.41	.16	.64
D-S Expert 3	54.73	22.02	.21	.60

D-S Expert 4	37.93	17.35	.31	.76
D-S Expert 5	63.33	14.80	.33	.81
D-S Expert 6	30.27	21.23	.16	.61
D-S Expert 7	47.47	14.61	.13	.44
D-S Expert 8	31.47	14.89	.08	.29

Table 7

Reliability analysis of the domain-specific experts

Judge	**Standardized alpha**	**Average inter-item correlation**	**Judge-total correlation corrected for item overlap and judge reliability**	**Judge-total correlation if judge deleted**
D-S Expert 1	.72	.27	.52	.44
D-S Expert 2	.66	.21	.78	.67
D-S Expert 3	.73	.28	.36	.21
D-S Expert 4	.66	.22	.71	.59
D-S Expert 5	.65	.21	.78	.65
D-S Expert 6	.69	.24	.56	.55
D-S Expert 7	.72	.27	.45	.35
D-S Expert 8	.77	.32	.22	.03

Note. D-S Expert 1 was negatively correlated with the other judges. His scores were automatically reversed.

Second, in Chapter 7, participants were prompted to directly compare the project proposals instead of providing absolute judgments. It was hypothesised that although non-expert participants might be more uncertain with ‘stamping’ a project with an exact number, they might be more apt to rank the projects similarly to the experts, and thereby locate them on an ordinal scale instead of placing them on a ratio scale. These results are discussed in detail in *Chapter 7*.

This chapter continues with a measurement of how non-expert participants spontaneously conceptualised their internal model of creativity. The data was grouped according to the criteria used in the present research to see whether there is an overlap in the ratios, i.e., whether participants weigh those dimensions as the most crucial for the judgment making which they spontaneously report as the most relevant dimensions of their conceptualisation of creativity.

4.7 Non-experts’ Spontaneous Creativity Definitions

Prior to any experimental tasks, all participants were asked to name a few factors which according to their opinion contribute to perceive something as creative. This free recall task was intended to register the ‘expressed conception of creativity’ - the part of a participant's internalized, naive conception which can be easily recalled and put into words. This was then contrasted with the so-called ‘implemented conception of creativity’; it was examined in multiple studies how underlying factors of creativity influence the ratings. A prior study investigating the structure in self-reported creativity conceptions was also uncovered. Storme & Lubart (2012) demonstrated that the expressed creativity conceptions matched the implemented creativity conceptions vastly, as in both cases, originality had the strongest predictive power for creativity. They found that subjects were mentioning originality more frequently than any other related factor - whether it was the most important or the most easily accessible component (or perhaps both) was not clear though.

Our sample consisted of n=416 participants. Each of them was prompted by the instructions to fill in a middle-sized text box with creativity-associated factors. The string variable, which was 1 to 127 words long, was then re-coded by N=2 independent raters according to the four features used in the experimental tasks: raters could sort the words or syntaxes into the following categories: Originality, Utility, Scalability, Riskiness. If the response did not fit any of these four categories, it was sorted into the Other category. As typically multiple expressions were provided, the input from one participant could be coded into multiple categories. However, one expression could only belong to one category. When frequencies were quantified, it was counted whether the participant used a given category in a binary manner. E.g., they could either provide some input to utility or not – fluency was not measured thus it did not matter whether the participant mentioned 1 or 5 words linked to the given category. The agreement between the two non-expert raters (neither of them was involved with any prior phase of the research and had any expertise related to creativity research; they only received detailed instructions on the categorization task) was κ = .460. Since the value indicates a moderate agreement, both categorizations were used for the further calculations.

Notably, originality was found to be the most popular response, however, this finding might be heavily confounded by the instruction providing the word originality as an example for the task. Utility was mentioned substantially less frequently than originality. Only the content of the 'other' category is comparable to the dominance of originality-related words. Both scalability- and riskiness-related items were recalled more marginally, although the two raters differed a lot in judging the gap between these two frequencies. The raw as well as the relative frequencies are displayed in Table 8.

Table 8

The categorisation of free associations according to the measured features by two independent raters

		Originality	***Utility***	***Scalability***	***Riskiness***	***Other***
RATER #1	***Raw frequency***	317	117	15	84	326
	Relative frequency	76.2%	28.1%	3.6%	20.2%	78.4%
RATER #2	***Raw frequency***	403	86	27	35	125
	Relative frequency	96.9%	20.7%	6.5%	8.4%	30%

Note. *N = 416 by each column. 'Raw frequency' denotes the number of non-empty cells in the categorization, while 'relative frequency' is the percentage of non-empty cells divided by the total number of cells.*

The next section outlines a few suggestions for further criteria which could be implemented in future research efforts.

4.8 Suggestions for Further Features

The suggestions listed here are based on the evaluation of free association data obtained from non-expert participants (*n*=416). Table 9 contains the concepts mentioned more than ten times to showcase what other factors contribute to creativity according to lay participants. Words used in the instructions were eliminated from the list (creativity, ideas, something).

Table 9

Participants' free recalls regarding creativity-related features, ranked by frequency

Word	Frequency of appearance
Different	44
Thought, thinking, thought-provoking	43 (=21+16+6, resp.)
Colourful	37

Imagination	36
Inspiring	27
Interesting	26
Artistic, art	25 (=20+5, resp.)
Personality, people	20 (=13+7, resp.)
Details	16
Design	13
Emotional	13
Fun	13
Aesthetics	12
Surprising	10

Note. Please note that the displayed frequencies were gathered combined from the two independent raters, thus are drawn from a duplicate of the data.

From the elements of this list, a few are already covered by our features (e.g., different and surprising are covered by the originality feature). However, thoughtfulness and imagination could be combined together into a ‘cleverness’ feature. Further, colour, inspiration, art, emotion, and aesthetics could be applied as an ‘artistic’ feature and be used when appropriate for the domain. What can be seen from this list is that spontaneous definitions of creativity are largely associated with artistic features, emotions, and smartness.

In this chapter, the methodological considerations leading to the creation of the paradigm, the entire process of constructing the paradigm, the baseline values obtained by using the paradigm, as well as its mapping with non-experts' spontaneous definitions about creativity were presented. The construction of the paradigm was detailed to familiarise the reader with the methodology used in the empirical chapters. The upcoming three chapters (*Chapter 5-7*) will seek answers to the main research questions by using the basic paradigm introduced here.

CHAPTER 5: EMPIRICAL STUDY I

5.1 Introduction

5.1.1 Overview. The present research seeks to answer three questions: (Q1) what information do judges use to evaluate creative ideas? What criteria is applied internally to form an overall creativity judgment? (Q2) What weighting each one of the criteria is carrying towards forming the overall creativity rating? And finally, (Q3) how do certain contextual factors influence creative idea evaluation? In response to Q1, this chapter outlines the results of using linear models to predict creativity scores based on creativity-related feature ratings. In response to Q2, data was collected to find out which criterion is more important than the others to confirm or falsify current theoretical conceptualisations of creativity. To address Q3, the experimental protocol includes the investigation of motivation via the use of incentives. The role of motivation influencing creative idea evaluation as a contextual factor is addressed. Motivation is a candidate factor for explaining part of the differences found between experts' and non-experts' judgments. Here, we set out to align experts' and non-experts' level of motivation, to bring them to a more similar starting point, in two experiments.

Experiment 1 and 2 reported in this chapter are almost identical as the second one is a replication of the first one. What is following now is the first application of the novel paradigm introduced in *Chapter 4*. The data analysis reported here is similar to the Lens model framework outlined in *Chapter 3*, however, no direct comparison is made between the experts and non-experts in this chapter. The rationale for this is that the judgments provided by experts and non-experts are treated as two distinct sets. To make a comparison possible, first it needs to be confirmed that they are informed by the same information. This is the reason why, similarly to

the Lens model, identical criteria are linked to the creativity ratings in case of both experts and non-experts to see to what extent they are informing the judgments made by both groups.

5.1.2 General Introduction. A creative product can transform the life of millions - smartphones and their apps, drones, tablets and game consoles have changed the way we work and spend leisure time. These products start out as one of many ideas in a pile, and through a long process of refinement, which requires identifying their potential, the final product eventually reaches the market, if successful, even goes viral (Thompson, 2017). It is widely accepted that the best judges of creativity are experts with domain-specific knowledge (Amabile, 1983), and indeed, businesses employ expert opinion of this kind during idea evaluation (e.g., Magnusson, Netz, & Wästlund, 2014). But how does one decide between ideas that are creative and investment-worthy from ideas that are not? Despite significant advancements in the domain of creativity and ideation (Runco & Pritzker, 2011), the creative idea evaluation process is treated almost like a black box; experts are characterized as relying on their gut feelings and non-experts are considered to have a noisy and unreliable judgment process. In this study, we systematically attempt to peer into the black box of creative evaluation in both experts and non-experts. By revealing the inner workings of the box, we aim to explain how a truly creative idea is recognized.

5.1.3 Experts. When it comes to the assessment of creativity, much of the work tends to favour the view that experts and novices are qualitatively different (Kaufman & Baer, 2012). The standards and criteria by which ideas are judged appear to differ significantly between experts and non-experts (Kaufman et al., 2008; Kaufman et al., 2013; Silvia, 2013). Unlike non-experts, experts show a high degree of internal reliability in their evaluations of creative products (Amabile, 1982; Hennessey & Amabile, 1999; Kaufman, Lee, Baer, & Lee, 2007). Experts are equipped to handle contexts where strict criteria of assessment of creativity are

available, as well as contexts when there are no decision criteria (Bettman & Sujan, 1987). In contrast, non-experts tend to have difficulty in developing reliable criteria for assessing creative ideas, which in turn limits their ability to identify ideas that experts would evaluate as genuinely creative (Galati, 2015). Further, compared to non-experts, experts use different problem-decomposing strategies (Ho, 2001), and utilize and apply their knowledge differently because they draw from diverse sources of information (Björklund, 2013). Despite these qualitative differences, it is not clear how precisely experts and non-experts differ with respect to the degree to which they utilize specific features of an idea/product in order to evaluate its creativity. We addressed precisely this issue by comparing experts and non-experts on identical judgment criteria.

5.1.4. Criteria. Though there is no threshold or gold standard, there is a good consensus around two core features of assessment of creativity, novelty and usefulness (Runco & Jaeger, 2012). Several creativity assessment protocols have been developed using these two, as well as other features. For instance, the Creative Product Analysis Matrix includes features such as elaboration and synthesis, as well as novelty, and resolution (O'Quin and Besemer, 2006). The Creative Solution Diagnosis Scale consists of features such as relevance and effectiveness, novelty, elegance, and genesis and are employed for assessing functional creativity (Cropley & Kaufman, 2012, 2013). A meta-analysis has identified novelty, workability, relevance, and specificity as the most common features that inform creative evaluations of ideas (Dean, Hender, Rodgers, & Santanen, 2006). Consistent with this, we have argued earlier that there are two broad categories of features that underpin evaluative judgments (Pétervári, Osman, & Bhattacharya, 2016). The first refers to how unique an idea is (novelty, originality, surprise) and the second refers to how functional the idea is (utility, effectiveness, appropriateness) (e.g., Bruner, 1962; Runco, & Jaeger, 2012).

Therefore, from the work reviewed, in our current study we consider four features: originality, utility, scalability, and riskiness. The first two are domain-general and are core to the definition of creativity, while the latter two are domain specific because they are concerned with the functionality of the creative idea: scalability (i.e., the opportunity of growth) and riskiness (i.e. whether a project would be implementable and sustainable) (Kaplan & Strömberg, 2000). The materials we presented participants were project ideas on improving urban lives with varying degrees of these four characteristic features. Experts and non-experts evaluated the projects based on the four features, as well as provided an overall creativity assessment based on their internalized construct of creativity.

5.1.5. Motivation. We examined both extrinsic and intrinsic motivation, given that both types have been strongly linked to the creative process (Baer & Kaufman, 2005). Apart from the level of expertise (Kaufman et al., 2013), we identified motivation as an essential factor for optimal evaluation performance (cf. de Jesus, Rus, Lens, & Imaginário, 2013). In real life, the evaluation of creative products is always goal-directed, and judges are engaged in the process, aiming for the best possible decision maximizing the impact of their limited resources. During the pilot phase of this study, non-experts were found to be less interested in, or sometimes even confused about the outcome of their evaluations. Thus, we streamlined the process by introducing intrinsic and extrinsic incentives to simulate a creativity judge's environment. Incentives are rewards which are meant to increase performance by motivating the individual to exert more effort towards a task (Bonner, Hastie, Sprinkle, & Young, 2000). The incentives were aimed to guide the participants' thinking while aligning their motivation. In the present study, intrinsic motivation was induced by setting the task as a challenge, offering participants a sense of autonomy through making important decisions, as well as food for their curiosity, and finally, by encouraging the use of their fantasy through the introduction of a role-play (based on Lepper, & Hoddell, 1989). Extrinsic motivation involved a financial payoff scheme

in which there was a flat fee and further opportunity for a bonus dependent on task performance. The incentives were applied together and not separately because a, we were interested in their joint effects (reaching the highest task-engagement possible) b, the two conditions cannot be fully divided since the task was interesting enough to trigger intrinsic motivation spontaneously. We were not interested in whether participants would do the task without payment but rather wanted to make sure the participants were as much invested and motivated as possible. This manipulation was introduced despite being aware of many instances in which introducing (extrinsic) incentives did not equate to better results (Kamenica, 2012). There is a branch of research showing a negative effect of incentives on creativity, regarding the idea generation phase (e.g., Conti & Amabile, 2011). Yet inspiration to create is a vastly different state from making judgments about someone's creative idea, thus we did not generalize this finding in the present study. In fact, there is no prior data on how different types of incentives are affecting the evaluation of creative ideas.

5.1.6. Hypotheses. We expected that (H_1) the overall creativity rating of the project ideas would be predicted by the four features we have identified as critical for the assessment of creativity of ideas (originality, utility, scalability, and riskiness), i.e. all four of them are contributing to the judgments made about creativity. However, (H_2) the internal model, consisting of how the four features were weighted, was expected to differ by expertise.

Finally, it was assumed that task complexity is a differential component between non-experts' and experts' evaluation ability. If tasks demand high cognitive complexity and if participants do not possess the required skills for performing well on them, then incentives are less likely to improve task performance (Bonner et al., 2000). Based on this, we set out to conclude whether non-experts are performing poorly at evaluating creative ideas because the task is difficult, and they are not capable of making accurate judgments, or because they are usually not sufficiently motivated to embark on them. (H_3) enhancing the intrinsic and extrinsic

motivation of lay participants through the use of incentives was expected to result in non-experts' task performance becoming more similar to experts' ratings.

5.2 Experiment 1

5.2.1 Method

5.2.1.1 Participants. As for the non-experts, 80 healthy participants (60 females, age: *M*±*SD* = 20.18±2 years) were recruited from the Queen Mary University of London, UK. All participants gave written informed consent and were compensated between 6.5 – 16.8 USD (the exact amount depended on performance). As there are strongly diverging opinions on how sample size and power computations of linear mixed models should be done (or if they should be done at all), we followed the guidelines of Simmons and colleagues (2011) and set the sample size to 40 participants in each condition to explore the effects resulting from the use of a novel paradigm. To ensure the robustness of the effects, a replication study with 30 participants in each condition was conducted too (see Exp. 2).Domain-general experts (*N*=16) were recruited with a snowball methodology from junior and senior science staff at Queen Mary, University of London. As for the domain-specific experts, *N*=10 domain-specific experts were recruited, from which two judges had to be discarded due to providing incomplete data, resulting in a final data set of *n*=8 experts with *M*=20.38±10.37 years of relevant expertise. They are aged *M*=47.38±12.22 years, 3 of them are males. Domain-specific experts were selected based on the following criteria: (1) they must be UK residents and proficient English speakers, (2) they must have at least 10 years of expertise in their domain, (3) their domain must be related to cities, e.g., urban planning, architecture, civil engineering, policy making to cities, etc., and (4) they also must have some understanding of creativity, e.g., writing poems or publishing novels, composing music, coding programs, etc.

The study protocol was approved by the Local Ethics Committee at the Queen Mary, University of London (reference number: QMREC1566a).

5.2.1.2 Design and Materials. The experiment was comprised of two conditions, following a within-subject design. All participants completed a Feature rating task and an Investment task, but the order of presentation of the tasks was randomized across participants.

The stimuli consisted of fifteen project ideas, which were created based on proposals collected from an open-source platform, OpenIDEO (2011); all project ideas were entries for a competition on "How might we restore vibrancy in cities and regions facing economic decline?" and were edited into two paragraphs of text. Each project was subsequently rated by a pool of experts (N = 16) on four features: originality, utility, scalability and riskiness. Experts had domain general expertise, consisting of junior and senior science staff at Queen Mary, University of London. Table 2 presented in *Chapter 4* outlines the four critical features on which the experts judged the 15 projects. A detailed account of the creation of the paradigm can be also found in *Chapter 4.*

5.2.1.3 Procedure. After giving their consent, all participants were asked to provide demographic information, as well as potentially relevant experience related to the task domain, their current motivational levels, and their subjective interpretation of creativity. Once completed, all participants were presented with two tasks: Feature rating task and Investment task. In the Feature rating task, participants were first familiarized with the four features of creativity (Table 2). After this, each project proposal (N=15) was presented for 60s, and for each, participants rated the project according to each of the four features; rating responses were provided on a visual analogue scale (VAS) from 0 (not at all) to 100 (most). In the Investment task, participants were presented with the same 15 projects, but this time they were instructed to indicate their willingness to invest in each project; they indicated their response on scale between 0 (no investment) to 100 (maximum investment). Participants were explicitly instructed to make their investment judgment solely on the basis of their subjective interpretation of creativity. In principle, the investment responses can be considered as a

reasonable proxy for overall creativity; thus, hereafter we refer to these responses as creativity ratings.

The reward schedule for each participant was the same. In order to extrinsically motivate participant, they were informed that the amount of they could earn (between £5 and £13) was dependent on their performance; this was actually assessed on the basis of a normative rating as provided by a separate pool of experts. Total earnings were presented to participants after completing the feature rating task and the investment task. In order to assess the extent to which participants were sufficiently motivated by the investment task, at the end of the experiment participants were asked the following: 1) Did you actually imagine yourself as an investor? (10-point Likert scale from "I stayed completely outside of the game" to "I was fully immersed, felt as an investor", measuring intrinsic motivation) (2) "How much did the potential earning in the experiment motivate you?" (5-point Likert scale from "I did not care" to "I tried to earn as much money as possible", measuring extrinsic motivation).

5.2.1.4 Data Analysis. The full data set (*N*=1200 trials) was analysed. To predict overall creativity ratings from the ratings of four features, a linear mixed-effects regression (LMER) analysis was conducted using the *lme4* package of R (Bates, Maechler, Bolker, & Walker, 2015). This analysis also helps delineating the effect from the variance stemming from possible biases in response tendencies (the so-called 'rater effect', Hung, Chen, & Chen, 2012). The significance of the four predictors was determined while controlling for the intercept of participants (random effect). Initially a null model was used including solely the random effects. Incrementally, each feature was added to the model as a fixed effect (Model 1-4), and its performance was assessed by the likelihood ratio tests. Next, we calculated the coefficient of determination to assess the explained variance (Nakagawa & Schielzeth, 2013). The formula accounting for the fixed effects was $R^2_{GLMM(m)} = \frac{\sigma_f^2}{\sigma_f^2+\sigma_\alpha^2+\sigma_\epsilon^2}$, in which $R^2_{GLMM(m)}$ is the percent of variation in the creativity ratings explained by the fixed effects, σ_f^2 is the amount of variance

due to all fixed effects, σ_{α}^2 is the amount of variation due to the random effect, and σ_{ϵ}^2 is the residual variance. Combining fixed and random effects was obtained by, $R^2_{GLMM(c)} = \frac{(\sigma_f^2+\sigma_{\alpha}^2)}{\sigma_f^2+\sigma_{\alpha}^2+\sigma_{\epsilon}^2}$.

5.2.2 Results

First, we analysed participants' self-reported responses on intrinsic and extrinsic motivations. We observed a moderate to high level of both motivations (intrinsic: 7.55±1.65 on a 10-point scale; extrinsic: 3.38±1.12 on a 5-point scale). Before proceeding with the main analyses, we also examined the possibility of order effects on creativity ratings, and found that an effect of order was present by only two out of the fifteen project proposals, project 3: t(78) = 2.14, p = .035; project 8: t(70.43) = 2.70, p = .009. This finding is not detailed here, as the first half of *Chapter 6* will discuss both the hypotheses about and the wider implications of this result.

To understand the effects of the four features on overall creativity, a LMER analysis was performed on the novices' data (see *Data Analysis*). We found that Model 4, including all four features, provided a better fit than Models 1-3 that included one, two, or three features, respectively. The increase in the goodness of fit was revealed by running likelihood ratio tests, Model 1 vs. Model 0: $\chi^2_{(1)}$ = 165.34, p < .001, Model 2 vs. Model 1: $\chi^2_{(1)}$ = 438.62, p < .001, Model 3 vs. Model 2: $\chi^2_{(1)}$ = 40.515, p < .001, Model 4 vs. Model 3: $\chi^2_{(1)}$ = 15.917, p < .001. However, the fixed intercept of Model 4 was not significantly different from zero (t = .937). Thus, a fifth model was tested; Model 5 differed from Model 4 only that it did not contain the fixed intercept. The LMER analysis showed that all four features remained significantly different from zero (see Table 10) and the amount of unexplained residual variance remained unchanged (Model 4: 436.7 and Model 5: 436.4), suggesting that the intercept was redundant. Thus Model 5 was kept as the most parsimonious fit.

Table 10

Fixed and random effects for Model 5 predicting the creativity ratings based on feature ratings of the participants

Parameter	Estimate (β)	Standard error	t-value
Fixed effects			
Originality	.275	.024	11.550*
Utility	.514	.024	21.691*
Scalability	.125	.021	5.912*
Riskiness	-.103	.022	-4.745*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	110.7	10.52	
Creativity ratings	436.4	20.89	

Note. n = 1200, * = $p < .05$

Next, we computed the coefficient of determination and found that 39.7% of the total variance in creativity ratings was due to the fixed effects. Further, combining the fixed and random effects, the explained variance increased to 51.9%. This suggests that all four features contributed significantly towards the judgment of overall creativity of the projects as made by the non-experts. Utility was considered as the most influential, and riskiness the least.

To determine the extent to which categorical judgments of the features (high/low) made by domain-general experts predicted the creativity ratings of non-experts, a LMER analysis was carried out. In Model 1, the originality feature did not contribute to the model, as indicated by the likelihood ratio test, Model 1 vs. Model 0: $\chi^2_{(1)} = 0.33$, $p > .250$, thus it was omitted from further analysis. Model 2 including the utility feature and was a better fit than the null model, Model 2 vs. Model 0: $\chi^2_{(1)} = 375.61$, $p < .001$ and Model 3 including utility and scalability, and was a better fit than Model 2, $\chi^2_{(1)} = 4.31$, $p = .038$. However, adding the riskiness feature to Model 4 showed that the scalability feature no longer contributed significantly to the model ($t = 1.49$), in contrast to the intercept ($t = 12.75$), utility ($t = 21.17$) and riskiness ($t = 8.31$). When the scalability feature was excluded from Model 5, the residual variance did not change (Model 4: 525.2, Model 5: 525.7) and the goodness of fit did not decrease, $\chi^2_{(1)} = 2.24$, $p = .135$. To

conclude, Model 5 including the intercept, the utility and riskiness features as fixed effects were selected as the best fit for the data (see Table 11).

As for the coefficient of determination, the feature ratings provided by the experts were found to explain 23.4% of the variance in the creativity ratings provided by the non-experts. By combining the fixed and the random effects, 37.2% of the variance in the creativity ratings was explained.

Table 11

Fixed and random effects for Model 4 predicting the creativity ratings based on feature ratings of the experts

Parameter	**Estimate (β)**	**Standard error**	**T-value**
Fixed effects			
Intercept	23.963	1.646	14.56*
Utility	28.079	1.330	21.11*
Riskiness	11.227	1.330	8.44*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	115.6	10.75	
Creativity ratings	525.7	22.93	

Note. Standard errors are identical due to dummy variable coding. n = 1200, * = $p < .001$

Domain-specific experts' feature ratings were not suitable for predicting non-experts' creativity ratings. The lack of agreement in their data (see Ch. 4) did not allow for matching the two data sets. Similarly, the noisiness in the domain-specific experts' feature judgments (N=4 x 120) did not allow for using them as predictors for novice judges' feature judgments (N=4 x 1200).

However, overall creativity ratings provided by the domain-specific experts could be predicted from their ratings of the four features. An LMER analysis revealed that none of the fixed effects improved the goodness of fit of the null model. Adding each selected criterion to the model incrementally (in the order of originality, utility, scalability, and riskiness) did not make a difference as indicated by the non-significant likelihood ratio tests, Model 1 vs. Model

0: $\chi^2_{(1)} = .196$, n.s.; Model 2 vs. Model 0: $\chi^2_{(1)} = 1.797$, n.s.; Model 3 vs. Model 0: $\chi^2_{(1)} = 0.819$, n.s.; Model 4 vs. Model 0: $\chi^2_{(1)} = .067$, n.s.. The null model (summarised in Table 12) could explain only 1.8% of the variance in the creativity ratings, and the difference due to participants accounted for further 24% of the variance in the ratings.

It was also probed whether domain-specific experts' creativity ratings can be predicted based on domain-general experts' feature ratings. In Model 1, the originality feature did not improve the model's fit, which was shown by the likelihood ratio test: Model 1 vs. Model 0: $\chi^2_{(1)} = 0.56$, $p = .455$, thus further models did not include this feature. Model 2 including the utility feature and was a better fit than the null model, Model 2 vs. Model 0: $\chi^2_{(1)} = 9.80$, $p = .002$, while Model 3 including utility and scalability did not turn out as a better fit, Model 3 vs. Model 2, $\chi^2_{(1)} = .02$, $p = .877$. Adding the riskiness feature to utility in Model 4 also proved to be redundant, Model 4 vs. Model 2, $\chi^2_{(1)} = .08$, $p = .960$. Thus Model 2 including the intercept and the utility feature as fixed effects proved to be the best fit for the data (Table 13), accounting for 6.3% of the variance in the creativity ratings, and a further 19.5% of the variance stemming from the domain-specific experts. Taken these results together, the domain-specific experts demonstrated no linear relationshop between how they rated creativity and the features. On the other hand, there was a link between domain general experts' utility ratings and domain-specific experts' creativity ratings.

Table 12

Fixed and random effects for Model 0 predicting domain-specific experts' creativity ratings based on their feature ratings

Parameter	**Estimate (β)**	**Standard error**	**t-value**
Fixed effects			
Intercept	46.2	3.91	11.81*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	91.9	9.59	
Creativity ratings	457.68	21.39	

Note. n = 120, * = $p < .001$

Table 13

Fixed and random effects for Model 2 predicting domain-specific experts' creativity ratings based on domain-general experts' feature ratings

Parameter	**Estimate (β)**	**Standard error**	**t-value**
Fixed effects			
Intercept	39.8	4.64	8.58*
Utility	12.0	3.76	3.19*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	111.7	10.57	
Creativity ratings	423.1	20.57	

Note. n = 120, * = $p < .001$

5.2.3 Discussion

In sum, Exp. 1 revealed that all four features significantly contributed towards non-expert creativity ratings and accounted for almost 40% of the variance. While we found that all four creativity-related features contributed to the overall creative evaluation, previous work has shown that originality is typically considered as the core feature of creativity (Runco & Jaeger, 2012). In the present study, for non-experts, the parameter estimate of utility was more than twice as large as the parameter estimate of originality (Table 10). This disproportionate weighting of utility stands apart from previous studies in which utility was less relevant than other features for creative evaluation (Storme & Lubart, 2012), or meaningful only when ideas were also judged to be original (Diedrich, Benedek, Jauk, & Neubauer, 2015).

What is more, it was found that domain-general judges' ratings of the utility and riskiness features could predict non-expert participants' creativity ratings, which makes the domain-general expert ratings candidate for the criterion values. In contrast, when modelling the domain-specific judges' predictors of creativity ratings, it was found that domain-specific judges do not use any of the cues the non-expert participants do. This suggests that domain-

specific experts and non-experts are not suitable for pairwise comparison, as their judgments are formed on different bases.

To examine how robust these findings are, whether they are stable also if other participant population is asked to complete the experiment, the protocol was replicated on a different sample.

5.3 Experiment 2

5.3.1 Method

5.3.1.1 Participants. 60 healthy participants (24 females, age: $M \pm SD$ = 29.87±11 years) were recruited using Prolific Academic. Only non-student, native English speaker, residents of the United Kingdom, participants were eligible to participate; none of them took part in the Experiment 1. All participants gave written informed consent and were compensated between £5-13 (the exact amount depended on performance). One participant was excluded for not following the instructions properly, resulting in a final sample size n=59. The domain-general export pool (n=16) and the domain-specific expert pool (n=8) were identical to the ones described in Experiment 1.

5.3.1.2. Materials, Design, & Procedure. The design and procedure, along with the stimuli used was identical to Exp. 1 except for a minor change to the presentation style of the total reward at the end of the experiment.

5.3.2. Results

First, participants' self-reported responses on intrinsic and extrinsic motivations were assessed. Similarly to Exp. 1, a moderate to high level of both motivations were reported (intrinsic: 7.78±1.56 on a 10-point scale; extrinsic: 3.4±1.25 on a 5-point scale).

Next, an LMER analysis was implemented to predict non-experts' creativity ratings from their feature ratings. Here, we found that Model 4 including all four features and the

intercept as fixed effects provided the best goodness of fit, as indicated by the likelihood ratio tests: Model 1 vs. Model 0: $\chi^2(1) = 87.18$, $p < .001$, Model 2 vs. Model 1: $\chi^2(1) = 120.71$, $p < .001$, Model 3 vs. Model 2: $\chi^2(1) = 31.68$, $p < .001$, Model 4 vs. Model 3: $\chi^2(1) = 16.83$, $p < .001$, Model 5 vs. Model 4: $\chi^2(1) = 8.90$, $p = .003$. The amount of unexplained residual variance was lower in Model 4 (506.4) than Model 5 (510.1). Further, the intercept of Model 4 contributed significantly to the model ($t = 2.983$) and subsequently, this model was selected. Compared to the results of Exp. 1, the parameter estimate of the originality feature was reduced, while the estimate of the utility feature marginally increased (Table 14 summarizes the values).

Table 14

Summary of effects for Model 4 predicting non-experts' creativity ratings based on features.

Parameter	**Estimate (β)**	**Standard error**	**t-value**
Fixed effects			
Intercept	10.218	3.426	2.983*
Originality	.194	.034	6.954*
Utility	.351	.031	20.262*
Scalability	.138	.034	5.040*
Riskiness	-.130	.032	-4.264*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	108.5	10.42	
Creativity ratings	506.4	22.50	

Note. $n = 864$, * = $p < .05$

Calculating the coefficient of determination revealed that 28.6% of the total variance in creativity ratings was due to the fixed effects; by adding random effects, 41.5% of the variance was explained.

We turn now to an examination of whether experts' feature ratings predict domain-general non-expert's creativity ratings. As before, categorical judgments (high vs. low) of the relevancy of the four features made by the experts were used as predictors. An LMER analysis was conducted. Again, Model 1 with the originality feature was not a better fit as the null model,

Model 1 vs. Model 0: $\chi^2_{(1)} = 0.46$, $p = .498$, thus originality was excluded from subsequent models. Model 2 included the utility feature and was found to increase the goodness of fit: Model 2 vs. Model 0: $\chi^2_{(1)} = 238.54$, $p < .001$. In Model 3, the scalability feature was not significant ($t = .81$), thus Model 4 was introduced including the intercept, as well as utility and riskiness features. All fixed effects were significant in Model 4 (see Table 15), and the goodness of fit was higher than Model 3: $\chi^2_{(1)} = 27.47$, $p < .001$, and Model 2: $\chi^2_{(1)} = 28.13$, $p < .001$. As in Exp. 1, the model including the intercept, as well as utility and riskiness features served as the best fit of the data.

The feature ratings were found to account for 20% of the variance in the creativity ratings, shown by the coefficient of determination. Together with the random effects, 35.28% of the variance in the creativity ratings was explained by this analysis.

Table 15

Summary of effects for Model 4 predicting the creativity ratings based on feature ratings of the domain-general experts

Parameter	**Estimate (β)**	**Standard error**	**T-value**
Fixed effects			
Intercept	23.804	1.921	14.56*
Utility	25.061	1.522	21.11*
Riskiness	8.129	1.522	8.44*
Random effects	**Variance**	**Standard deviation**	
Within-person variability	120.1	10.96	
Creativity ratings	507.4	22.53	

Note. Standard errors are identical due to dummy variable coding. n = 885, * = $p < .001$

Exp 2. replicated the pattern of results observed in Exp. 1. All four features when rated by the non-experts contributed significantly towards their own creativity judgment. Furthermore, utility and riskiness rated by the experts were stronger predictors of non-experts' creativity judgment than originality and scalability. Thus, the results of the two experiments can be combined. The next section discusses the findings of both experiments.

5.3.3. Discussion

As it stands, given the progress of research in creativity, the process of evaluation of creative ideas/products resembles a black box. The present study is the first of its kind to open the 'black box' in order to examine the evaluation process of creative ideas. It has done this by exploring how experts and non-experts make creative judgment based on features typically associated with creativity, and model the process. In sum, in two experiments it was found that non-experts' relied on all four features (Originality, Utility, Scalability, Riskiness). In Experiment 1, this explained 39.7% of the variance in creativity judgments, and 28.6% in Experiment 2. In addition, when domain-general experts rated the creativity of the projects, their weighting of utility and riskiness over originality and scalability in turn predicted non-experts' creativity judgments. The latter of which provides an important insight into the alignment between expert and non-expert evaluations of creativity.

All four creativity-related features contributed to the overall creative evaluation, however, in both studies, the parameter estimate of utility was more than twice as large as the parameter estimate of originality. This disproportionate weighting of utility is unlike previous studies in which utility was less relevant than other features for creative evaluation (Storme & Lubart, 2012). However, other evidence also suggests that ideas are positively associated with practicality above creativity judgments (Müller, Melwani, & Goncalo, 2012), and that compared to experts, non-experts weight the usefulness of an idea over originality (Rietzschel et al., 2010). We speculate that the mixed findings with regards to the relevancy of utility/practicality in creative evaluations judgments might be explained by differences in the domain of the creative judgment task itself. Consistent with this, some have speculated that when the ideas being evaluated concern applied domains such as urbanism or design, a salient factor is the impact the ideas will have, referred to as *functional creativity* (Cropley & Cropley,

2005, 2008). Moreover, when people report an increased personal involvement with a problem domain the creative project is addressing, they tend to assess the more feasible and practical projects as creative (Illies & Reiter-Palmon, 2004; Rietzschel et al., 2010). Again, this suggests that the usefulness of a creative idea can be more relevant than originality in the process of creativity evaluation given a specific domain.

A possible criticism of our explanation is to question whether investment responses can be considered as proxies for overall creativity. Even though participants were explicitly instructed that they should rate the stimuli based on their creativity and nothing else, one might raise concerns that an 'investor mindset' is vastly different from one which would propagate creativity. Investors must make returns on their investments; thus, they might become rather conservative and loss-aversive when selecting investments, which might be an alternative explanation of why usefulness is outweighed at the expense of originality.

Moving onward, in both experiments, for non-experts, the higher the perceived riskiness of a project the more creativity it was judged to be. It is worth highlighting that for the same projects rated as low risk by experts, non-experts attached higher creativity ratings to them. This suggests that while the riskiness of a project is a relevant factor in creative evaluation, perceptions of risk are clearly differentiated on the basis of expertise. Of all four features, scalability made the least contribution to creativity judgments, though the reason for this may be a result of a lack of domain-specific knowledge non-experts possessed regarding evaluating the potential of growth of the project ideas.

In case of the domain-specific experts, it was found that none of the features could be used as predictors of their judgments made about creativity. Based on the findings that all four features contributed to the non-expert participants' creativity ratings and no features contributed to the domain-specific experts' creativity ratings, it is suggested that domain-specific experts and non-experts are not suitable for pairwise comparison since their judgments are formed on

different bases. However, domain-general experts showed similarities (mainly in the utility ratings) with both groups, which shows a gradual decline in agreement as domain knowledge increases.

As well as revealing which factors contribute to the evaluation of creative ideas, the present study also showed how taken together, the four features accounted for the variation in creativity judgments of non-experts. In Experiment 1, ratings of the four features explained 39.7% (28.6% Experiment 2) of the variance in creativity ratings, and a further 12.2% was accounted for by variability amongst the participants (including the rater effect), resulting in a total of 51.9% (41.5% Experiment 2) of the explained variance. While there is no direct comparison to the present study, to put our findings in context, a recent study comparing quasi-expert and novice creativity ratings for ideas designed to solve science problems reported that 32-40% of the variance in creativity ratings could be accounted for by the non-expert judges (Long & Pang, 2015). Moreover, other studies have also reported that non-experts tend to generate noisy and unreliable creativity judgments (e.g., Kaufman et al., 2008; Lee, Lee, & Youn, 2005). The general conclusion drawn from this work (as it was outlined in *Chapter 1* already) is that non-experts use features in an arbitrary way, which is reflected in the fact that they diverge significantly from the pattern of judgments of experts, as well as relative to other non-experts (see Kaufman & Baer, 2012, for a review). While it is difficult to make strong comparisons given the differences in between previous studies and the present study, in contrast to prior work, we show that non-experts do possess an internalised model of creativity that is robust and consistent. Previous studies that look at non-expert groups tend to show inconsistent judgments as a result of averaging across non-expert groups. We speculate that one key difference between previous studies and the present study is our attempt to individually map the basis on which creative judgments are made, by using a linear mixed model analysis. This analysis is grounded in the Theory of Social Judgment (Hammond, Stewart, Brehmer, &

Steinmann, 1975) and was an attempt to fit the Lens model on creativity judgments to garner more information about the cues informing the judgment. The results demonstrate that non-experts do not make random subjective judgments regarding the assessment of creative ideas and the variance found in creativity ratings can be explained if the "input values" of the model are adequately defined at the start.

When the internalised creativity model of non-experts was replaced with an external expert model, utility emerged as the dominant feature, suggesting that the perception of utility and its influence on creativity evaluation is largely aligned between experts and non-experts. Using the same approach to explore other features, when it came to originality, this feature was not a significant predictor of creativity ratings in non-experts, suggesting its potential relevance in evaluating creativity is likely to be shaped by expertise. It is worth nothing here that a possible limitation of the study is that expertise was treated as a distinct, binary construct (expert vs. non-expert). In reality, expertise is likely to be graded, thus future studies would could use the model we developed to examine how the magnitude of expertise affects creativity ratings. Another important factor that also requires further exploration is the generalizability of our findings to other creative idea evaluation in other domains. It may be the case that utility consistently predicting creativity ratings for non-experts and experts, because of the business context of the project ideas, and so further work is needed to explore the extent to which the consistency of the non-experts' internal model of creativity is demonstrated in other creative domains.

In sum, we showed that, if motivated sufficiently, lay participants' overall creativity ratings can be predicted from the four features selected in *Chapter 4*. Their creativity ratings could be predicted to a larger extent from their feature ratings than from domain-general expert judges' creativity ratings. The experiments revealed an internal model of the non-experts

comprising of how the four features are weighted. In contrast, domain-specific experts' creativity ratings were not found to be informed by their ratings of the features.

One of the takeaways of this research was that the available information at the time of casting the rating might be crucial for forming the judgment about creativity. A frequent feedback of the participants was that they would do a better job if they would be provided with more information about the task. Thus, this hypothesis was also investigated as part of Experiment 1 & 2, however, the analysis of this issue will be presented in the next chapter, as task-related information is a different contextual factor than motivation. The next chapter discusses in detail how manipulating the available information might influence the evaluation process. The manipulations are alternating task instructions and providing additional data to the participants.

CHAPTER 6: EMPIRICAL STUDY II

6.1 Overview

The previous chapter discussed key results drawn from Experiment 1 and 2 by focusing on the use of linear models. The present chapter builds on these results by providing an account of possible reasons which might impose challenges to naïve participants while judging creative ideas. This chapter addresses the third main research question (Q3), namely, it addresses the role contextual factors are playing in the evaluation of creative ideas. One of the main results from *Chapter 5* is that creativity ratings can be predicted based on creativity-related features. The two experimental manipulations reported below are appended to the previous chapter: they help to explore what factors are possibly informing the judgment making process.

The data presented in this chapter stems from three experiments; the first part of the chapter probes whether providing task-relevant information influences creativity ratings based on data from Experiment 1 & 2. Subsequently, Experiment 3 is introduced in the second part of the chapter. This one is testing whether meta-information regarding the nature of the task affects creativity ratings and the corresponding certainty ratings. The aim in case of both manipulations is to figure out how does providing and not providing participants with explicit information about the task affect how they cast their ratings about creativity. First, an information brief is provided to half of the participants about the criteria which can be used to assess creativity. It is assumed that those who receive additional information before rating creativity would cast different ratings than those who do not receive such information. In the second part, meta-information about the nature of the task is given to half of the participants: it is expected that they would be less certain in their judgments if it is known that creativity is the crucial dimension which needs to be rated in the experimental task.

6.2 The Effect of Providing Task-relevant Information

6.2.1 Introduction

Participants without expertise are often asked to complete difficult tasks; in the present research, they are instructed to evaluate the creativity of several project proposals, each of them outlining an idea about improving urban lives. This job is not only daunting because judging realistic ideas is a complex task and one needs to take numerous dimensions into account but also because a lay person has very little idea about what the relevant dimensions are. This might be also related to studies finding that people have implicit models of creativity (Lim & Plucker, 2001; Runco & Johnson, 2002). Given the lack of accessible knowledge, it is not rare for participants to feel unguided or that their responses can only be provided somewhat randomly, without a deeper understanding of the underpinnings, relying mostly on a gut feeling formed based on a first impression.

Thus, we attempted to supply non-expert participants with some essential information about the task to reduce their uncertainty and increase their success rate. The aim was to find out whether providing participants with explicit information about the creativity-related features would affect their creativity ratings. This question was motivated by the logic of the Lens model (Brunswik, 1952; Hursch, Hammond, & Hursch, 1964) outlining the judgment process as one in which individuals search for cues they can build in to inform their judgments.Previous research has also found different ratings of creativity when participants had to make judgments with no instruction, guided by explicit criteria, or task-related training (Caroff & Besançon, 2008). In this study, it was expected that providing the criteria for judgments would impact the evaluations.

Available information about the task-relevant features was manipulated by dividing participants into two groups: the first started the experiment with the Feature rating task, while the second group started with the Investment task. It was a within-subject design, so all

participants completed both tasks eventually. Critically, in the feature rating task, participants were provided with a brief pitch describing the features which have been theorised (see Table 2 in *Chapter 4*) to underlie the making of the judgment. Participants completing the Feature rating task were presented with the name and the definition of the four features implemented in the study (see *Procedure*). They were instructed to carefully read the presented information and keep it in mind while rating the features. It was expected that (H_4) those participants who were made aware of the name and content of criteria linked to creativity will cast feature and creativity ratings more closely linked to each other than the participants who were not made aware of which criteria are connected to creativity ratings. Further, participants who started with the Investment task were not informed about the criteria before casting their creativity ratings. Thus, (H_5) the different order of completing the Feature rating task and the Investment task was expected to provide a different amount of available information to the participants at the time of casting their creativity ratings. Therefore, the mean of the creativity ratings was expected to differ between the two conditions.

A potential caveat of the research design is that participants were not explicitly informed that the presented criteria should be used for judging creativity; it was only assumed that they would realise they could use their newly acquired knowledge during the second part of the experiment too. Also, it was not specified that all criteria should be used, participants could choose to incorporate only one or multiple ones to form their judgment about creativity. Based on previous findings, originality was expected to contribute the most to the selection of creative ideas (Rietzschel et al., 2010).

6.2.2 Method

6.2.2.1 Participants. For Experiment 1, 80 healthy participants (60 females, age: $M \pm SD$ = 20.18±2 years) were recruited from the Queen Mary University of London, UK. For

Experiment 2, 60 healthy participants were recruited via Prolific Academic. They were aged between 18 and 61 years (M = 29.87, SD = 11) and 24 of them were female. *N*=1 participant was excluded due to guessing the same number to all ratings. This resulted in an overall sample of *n*=139 participants. All of them gave their informed consent and were compensated with £5-13 (the amount depended on performance). The study protocol was approved by the Local Ethics Committee at the Queen Mary, University of London (reference number: QMREC1566a).

6.2.2.2 Design and Materials. The experiment was comprised of two conditions, following a within-subject design. All participants completed a Feature rating task and an Investment task, but the order of presentation of the tasks was randomized across participants. The data presented here is a sub-set of the data collected in Experiment 1 & 2, which were introduced in *Chapter 5* already.

The stimuli consisted of fifteen project ideas, which were created based on proposals collected from an open-source platform, OpenIDEO (2011); all project ideas were entries for a competition on "How might we restore vibrancy in cities and regions facing economic decline?" and were edited into two paragraphs of text. Each project was subsequently rated by a pool of experts ($n = 16$) on four features: originality, utility, scalability and riskiness. A detailed account of the creation of the paradigm can be found in *Chapter 4.*

6.2.2.3. Procedure. After giving their consent, all participants were asked to provide demographic information, as well as potentially relevant experience related to the task domain, their current motivational levels, and their subjective interpretation of creativity. Once completed, all participants were presented with two tasks: Feature rating task and Investment task. In the Feature rating task, participants were first familiarised with the four features of creativity. After this, each project proposal (*N*=15) was presented for 60s, and for each, participants rated the project according to each of the four features; rating responses were

provided on a VAS from 0 (not at all) to 100 (most). In the Investment task, participants were presented with the same 15 projects, but this time they were instructed to indicate their willingness to invest in each project; they indicated their response on scale between 0 (no investment) to 100 (maximum investment).

Critically, those who started the experiment with the Feature rating task received the additional information about the criteria before rating creativity as part of the Investment task. Participants who started with the Investment task were presented with the information only after they have rated creativity already.

6.2.2.4. Data Analysis. The data was analysed by using the Statistical Package for Social Sciences (SPSS) version 20 as well as Microsoft Office Excel 2010.

6.2.3 Results

In Exp. 1, a one-way multivariate analysis of variance (MANOVA) was conducted with the two conditions as the independent variable and the project proposals' creativity ratings as 15 dependent variables. This test can be considered an extension of the independent-samples t-test and was run to detect the effect of providing additional information on the ratings (n=40). The differences between the two experimental conditions on the combined dependent variables was statistically significant, $F(15, 64) = 2.239$, $p = .013$; Wilks' $\Lambda = .656$; partial $\eta^2 = .344$. A Bonferroni adjusted α level of .0001 with a simultaneous 99.9% confidence level was used. There were no significant differences in the pairwise comparisons as outlined in Table 16. Making the creativity investment with or without reading the feature descriptions did not result in different performance.

Table 16

Pairwise comparisons by each of the projects in Experiment 1.

Project ID	Mean difference between the investment first vs. the feature first condition	Std. error	*p*	99,9% CI for Mean Difference, Lower Bound	99,9% CI for Mean Difference, Upper Bound
1	70.750	3.521	58.710	82.790	70.750
2	52.500	4.165	38.257	66.743	52.500
3	61.925	3.444	50.148	73.702	61.925
4	29.275	4.290	14.603	43.947	29.275
5	70.150	3.260	59.003	81.297	70.150
6	43.375	4.277	28.750	58.000	43.375
7	36.275	4.049	22.428	50.122	36.275
8	33.525	3.586	21.260	45.790	33.525
9	24.625	3.840	11.494	37.756	24.625
11	36.100	4.750	19.855	52.345	36.100
12	60.900	3.811	47.868	73.932	60.900
13	56.850	3.724	44.116	69.584	56.850
14	45.425	4.332	30.611	60.239	45.425
15	24.275	3.109	13.643	34.907	24.275
16	33.575	3.779	20.651	46.499	33.575

Note. Differences between the average creativity ratings, starting with the Investment task vs. with the Feature rating task. *N*=40 in both conditions.

In Exp. 2, the same analysis was conducted and the independent-samples t-tests revealed no effect due to the manipulation, $F(15, 43) = .645$, $p = .821$; Wilks' $\Lambda = .816$; partial $\eta^2 = .184$. Table 17 displays the pairwise comparisons and that no significant difference was spotted by any of the projects.

Table 17

Pairwise comparisons by each of the projects in Experiment 2.

Project ID	**Mean difference between the investment first vs. the feature first condition**	**Std. error**	***p***	**99,9% CI for Mean Difference, Lower Bound**	**99,9% CI for Mean Difference, Upper Bound**
1	-1.244	6.279	.844	-23.028	20.540
2	-6.230	6.994	.377	-30.496	18.036
3	-.094	6.113	.988	-21.304	21.116
4	-3.848	5.667	.500	-23.509	15.812
5	-.475	6.226	.939	-22.077	21.127
6	-2.311	6.295	.715	-24.154	19.531
7	3.452	6.431	.594	-18.861	25.765
8	-10.061	7.504	.185	-36.095	15.973
9	-10.285	5.945	.089	-30.912	10.342
11	-3.148	7.532	.678	-29.281	22.984
12	-8.508	5.909	.155	-29.010	11.994
13	-4.259	5.593	.450	-23.663	15.145
14	-7.140	6.736	.294	-30.510	16.229

15	-2.221	6.794	.745	-25.792	21.351
16	-6.407	4.383	.149	-21.613	8.800

Note. Differences between the average creativity ratings, starting with the Investment task vs. with the Feature rating task. *N*=30 in both conditions.

6.2.4 Discussion

No effect resulted from providing the definition of criteria to the participants. This null result might be due to several reasons. One limitation is that participants were informed that the information brief was supposed to help them with the Feature ratings task, not with the Investment task, in which the creativity ratings were provided. It was assumed participants would connect the two parts of the experiment but there is no guarantee they indeed did so. Furthermore, a possible reason for the inefficacy of the manipulation is that the additional information was not presented in a form which could have been useful to the lay participants. The helpful information was laid out as a set of abstract definitions. On one hand, participants reported spontaneously at the end of the testing sessions that this information enriched their representations and reduced the vagueness of the task. On the other hand, the information did not make any impact on the mean creativity ratings of the projects. The paradigm was not suitable for detecting which level of processing formed the bottleneck the information could not get through.

Another issue with the research design might be that the guidelines outlined could not be effectively incorporated as there was only one shot to make each choice and no feedback was provided regarding the accuracy of the judgment made. A good test of whether the manipulation works might be a within-subject rather than a between-subject design as the participants were not matched according to any variables. A possible alternative of this manipulation could be an explicit instruction handed out to the participants, in which they are

told to rate creativity high if they think the level of a feature, e.g., originality, is high in the idea. A recent study found that such explicit instructions enable participants to overcome their original reluctance of selecting original ideas (Rietzschel et al., 2014). It is also speculated that revealing the relations between the cues of the features and the expected creativity ratings could inform participants better than presenting the definition of the features.

However, it is not only the connection of the criteria which might inform participants about how to judge stimuli. Information regarding the nature of the experimental task which participants must complete might also guide them in their efforts. The second half of this chapter lays out how this was studied in Experiment 3.

6.3 The Effect of Meta-information on Creativity Ratings

6.3.1 Introduction

During the data collection of Experiment 1 & 2, non-expert participants often reported that assessing creativity seems difficult for them because they do not have a clear idea on what is expected from them and the fact they need to complete a task they have never done before is making them nervous. This chapter is strongly built on this feedback collected from the participants. In the first part, it was probed whether providing more information to participants would affect how they judge creativity.

In contrast, Experiment 3 was designed to investigate whether possessing different meta-information influences the appraisal of creativity or not. Meta-information is understood in this thesis as contextual information which guides the participant in evaluating the task beforehand and creating a strategy to solve it. The aim was to check whether meta-information is one of the contextual factors which drives the judgments made about creativity. The assumption that it would be was motivated by the experience gathered during the data collection of the pilot work with the paradigm (not reported in this thesis). It was observed that delivering

the task instructions had discriminable effects on the participants: if it was highlighted that assessing creativity would be the point of the experiment, participants became visibly anxious and uncertain, while if only the 'judgment making about ideas' aspect was stressed, then participants seemed to have an easier time while completing the tasks. Thus, the framing of creativity assessments became one of the candidate contextual factors possibly influencing the judgment making process.

While meta-information implicates certain interpretations and associations and therefore evokes a context for the task, meta-cognition (discussed in Ch. 1) stands for understanding people's theories about their own cognition and reflecting on the strategies which govern their ways of thinking (Schraw, & Moshman, 1995). Meta-information helps to step into a process and to frame it appropriately, while meta-cognition facilitates the stepping outside of a process to consider it from an outer perspective.

The aim of this project was to find out whether possessing different meta-information about the task drives people to judge creativity differently, i.e., whether creativity ratings change if participants are aware that creativity is the important dimension in their evaluation. Deception is a technique often used in psychological research for discovering the effect of informed and uninformed scenarios on the dependent variable, however, here, no false information was provided to the participants, we only distorted the narrative of the instructions to create two alternatives. Namely, highlighting that creativity is the central dimension which needs to be assessed was compared to completing the same procedure without informing participants that the purpose of the task is to evaluate creativity. It was hypothesised that (H_6) if people need to evaluate the creativity of ideas, they make less certain judgments than what they would make if they were informed they must evaluate business ideas per several dimensions, not mentioning creativity explicitly ('nothing special' condition). In addition, (H_7) participants were expected to give less coherent creativity ratings than viability ratings as they

would have more expertise in estimating the usefulness/feasibility than the creativity of an idea. For this, difference was expected to be found in the average of creativity ratings between the 'creativity' and the 'nothing special' groups, whilst no difference in the average of the viability ratings between the two groups.

6.3.2 Method

6.3.2.1 Participants. 85 healthy participants (54 males, age: $M \pm SD$ = 31.85±7.19 years) were recruited using the online platform Prolific Academic. Half of them (n=43) were instructed to evaluate how creative the ideas are, whereas the other half of the participants (n=42) were instructed to evaluate the ideas taking a few criteria (as such, creativity happened to be one of them) into consideration. All participants gave written informed consent and were compensated with £3. The study protocol was approved by the Local Ethics Committee at the Queen Mary, University of London.

6.3.2.2 Materials. The stimuli consisted of fifteen project ideas, which were created based on proposals collected from an open-source platform, OpenIDEO (2011); all project ideas were entries for a competition on "How might we restore vibrancy in cities and regions facing economic decline?" and were edited into two paragraphs of text. Each project was subsequently rated by a pool of experts (N = 16) on four features: originality, utility, scalability and riskiness. A detailed account of the creation of the paradigm can be found in *Chapter 4*.

6.3.2.3 Design. The experiment followed a between-subject design: participants in two conditions had to complete an almost identical task – the only difference was the experimental manipulation, that is, the meta-information the task instruction suggested about the nature of the task to the participants.

6.3.2.4 Procedure. As outlined in the *Design* section, participants in both conditions completed an almost identical task. After giving their consent, all participants were asked to

provide demographic information, as well as potentially relevant experience related to the task domain, their current motivational levels, and their subjective interpretation of creativity. Subsequently, each project proposal (N=15) was presented for 60s. All participants were asked to consider each project on the basis of how creative, attractive and viable it is, and corresponding certainty judgments were also provided in case of each judgment. Rating responses were provided on a visual analogue scale (VAS) from 0 (not at all) to 100 (most). After completing the evaluation part, N=14 self-report statements were provided, each of them expressing possible strategies which one could use to evaluate creativity (e.g., „Although I was asked to judge the ideas objectively, at the end I relied on which one I liked and which one I didn't.”). Participants self-reported how much they agree with these statements/how much the statements described what they did in the main task on a 1-5 Likert scale (1 = not at all, 5 = fully).

Critically, the only difference between the two task conditions were the information participants received regarding the nature of the task. In the ‘creativity’ condition, participants were informed that the point of the experiment is to evaluate creativity. In contrast, in the ‘nothing special’ condition, participants were told that the aim of the experiment is to assess business ideas and there are several dimensions in which judgments must be made. As already stated above, the three criteria for evaluation were identical in both conditions.

6.3.2.5 Data Analysis. The data was analysed by using the Statistical Package for Social Sciences (SPSS) version 20 as well as Microsoft Office Excel 2010.

6.3.3 Results

A one-way MANOVA was conducted with the two conditions as the independent variable and the project proposals’ creativity ratings as 15 dependent variables. The aim was to reveal whether there were any differences in the creativity ratings between the 'creativity

instruction' and the 'nothing special instruction' groups. No significant difference was found between the two groups on the creativity ratings, F(15, 69) = .462, p = .952; Wilks' Λ = .909; partial η^2 = .091. A Bonferroni adjusted α level of .0001 with a simultaneous 99.9% confidence level was used for pairwise comparisons but there were no significant differences in any ratings between the 'creativity instruction' and the 'nothing special instruction' groups. The same can be stated about the certainty ratings: conducting another one-way MANOVA showed that the fixed effect had no main effect, F(15, 69) = 1.559, p = .109; Pillai's Trace= .253; partial η^2 = .253. Here, Pillai's Trace was used because the assumption of homogeneity of variance-covariance matrices was violated, as assessed by Box's M test (p < .001). Table 18 displays the creativity ratings, while Table 19 shows the certainty ratings.

Table 18

Pairwise comparisons between the creativity ratings provided by the two experimental groups.

Project ID	**Mean diff. between the nothing special vs. creativity instruction**	**Std. error**	***p***	**99,9% CI for Mean Difference, Lower Bound**	**99,9% CI for Mean Difference, Upper Bound**
1	70.750	3.521	58.710	82.790	70.750
2	52.500	4.165	38.257	66.743	52.500
3	61.925	3.444	50.148	73.702	61.925
4	29.275	4.290	14.603	43.947	29.275
5	70.150	3.260	59.003	81.297	70.150
6	43.375	4.277	28.750	58.000	43.375

7	36.275	4.049	22.428	50.122	36.275
8	33.525	3.586	21.260	45.790	33.525
9	24.625	3.840	11.494	37.756	24.625
11	36.100	4.750	19.855	52.345	36.100
12	60.900	3.811	47,868	73.932	60.900
13	56.850	3.724	44,116	69.584	56.850
14	45.425	4.332	30,611	60.239	45.425
15	24.275	3.109	13,643	34.907	24.275
16	33.575	3.779	20,651	46.499	33.575

Table 19

Pairwise comparisons between the certainty ratings provided by the two experimental groups

Project ID	**Mean diff. between the nothing special vs. creativity instruction**	**Std. error**	***p***	**99,9% CI for Mean Difference, Lower Bound**	**99,9% CI for Mean Difference, Upper Bound**
1	3.781	4.530	.406	-11.675	19.236
2	-.394	4.521	.931	-15.817	15.029
3	2.256	4.068	.581	-11.623	16.136
4	5.502	3.435	.113	-6.217	17.220
5	4.133	4.102	.317	-9.863	18.128
6	6.511	4.178	.123	-7.741	20.764
7	3.468	4.030	.392	-10.279	17.215
8	-2.497	3.642	.495	-14.923	9.930

9	3.669	3.596	.311	-8.598	15.936
11	-3.341	4.438	.454	-18.482	11.801
12	2.499	3.705	.502	-10.140	15.139
13	-.237	4.078	.954	-14.150	13.676
14	.408	4.215	.923	-13.974	14.790
15	.904	4.432	.839	-14.218	16.026
16	-.838	4.068	.837	-14.717	13.041

Upon test completion, 14 self-report statements were filled in by the participants to check for their strategy uses. A one-way MANOVA was conducted with the two conditions as the independent variable and the self-report statements as 14 dependent variables. The aim was to reveal whether there were any differences in the creativity ratings between the 'creativity instruction' and the 'nothing special instruction' groups. No significant difference was found between the two groups in what they thought about the task and what kind of strategies they used, $F(14, 70) = 1.606$, $p = .099$; Wilks' $\Lambda = .757$; partial $\eta^2 = .243$. A Bonferroni adjusted α level of .0001 was used for pairwise comparisons but no significant difference was revealed by the analysis (Table 20).

Table 20

Differences between the self-report statements provided by the two experimental groups

Statement	***Mean difference between the nothing special vs.***	***Std. error***	**p**	***99,9% CI for Mean Difference, Lower Bound***	***99,9% CI for Mean Difference, Upper Bound***

	the creativity instruction				
I found this task difficult to do.	.352	.210	.098	-.366	1.069
I made my rating somewhat randomly, didn't think about it too much.	-.007	.151	.962	-.521	.507
"Creative" is the opposite of "practical".	-.447	.229	.055	-1.229	.336
I don't think there is a good way to judge projects, the outcomes depend on luck anyway.	-.161	.207	.441	-.868	.547
I am certain that there are objective criteria with which great ideas can be detected.	-.239	.225	.291	-1.005	.528
Creative is something I have never heard of.	.020	.111	.853	-.357	.398
I think it's simple and straightforward to judge these ideas.	-.266	.221	.233	-1.020	.488
Creative is something I couldn't have come up with.	.262	.213	.223	-.465	.989
Only people who have more business-related expertise than me could do this task well.	-.022	.237	.926	-.829	.785

Only people who have creative talents could do this task well.	.214	.210	.311	-.501	.929
I'm not aware how this task should be done and felt kind of lost during the process.	-.058	.167	.730	-.626	.511
I know exactly how to judge project proposals such as the ones I read earlier.	.246	.202	.226	-.442	.934
Although I was asked to judge the ideas objectively, at the end I relied on which one I liked and which one I didn't.	-.164	.221	.460	-.920	.591
I know what I was doing and applied consistent criteria over all project evaluations.	.038	.255	.883	-.831	.906

A principal components analysis (PCA) was run on the self-report questionnaire including 14 statements. The PCA reveals to what extent are the original distances between the data points preserved while reducing the dimensions of this 14 dimensional data set. There is no agreement on whether a PCA is suitable for ordinary variables thus alternative methods, such as the factor analysis, may also be used.

PCA revealed five components that had eigenvalues greater than one; they explained 25.7%, 12.5%, 11.6%, 9.6% and 7.9% of the total variance, respectively. Visual inspection of the scree plot (see Figure 3) indicated that two components should be retained.

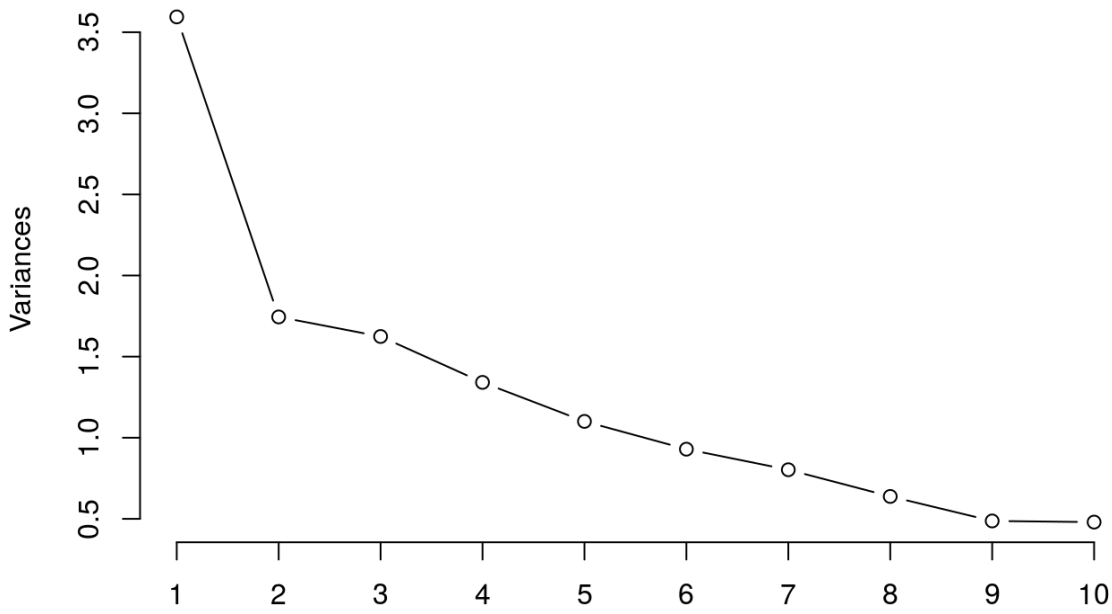

Figure 3. Scree plot of the principal components visualizing the retained variances.

The two-component solution explained 38.2% of the total variance and is depicted on Figure 4.

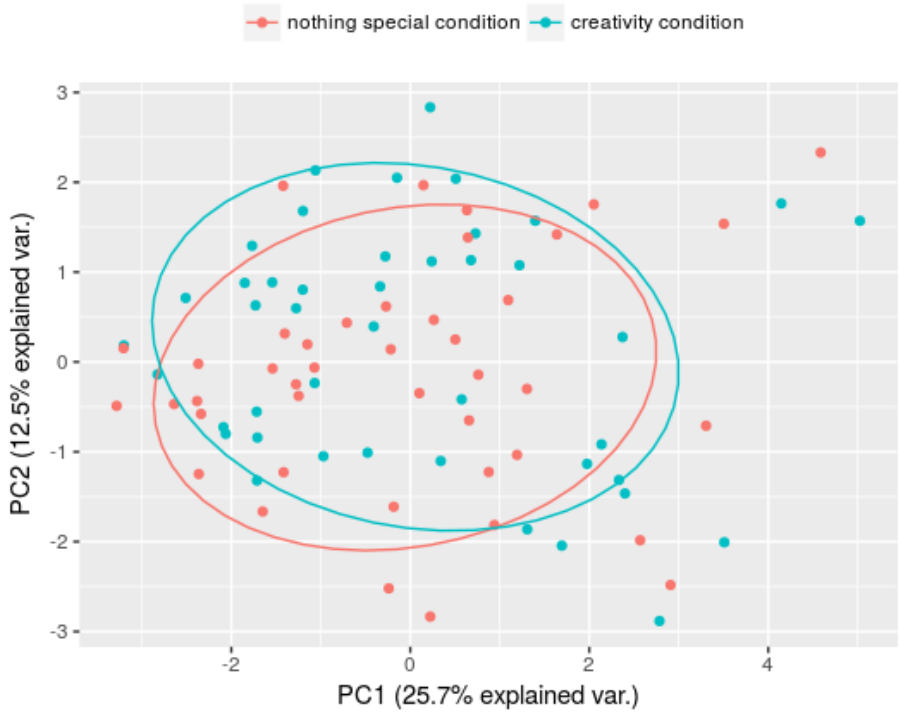

Figure 4. The two principal components in the two experimental conditions.

6.3.4 Discussion

In the present chapter, factors which might inform the judgment of creativity, and the lack of which might impede the successful assessment of creativity, were investigated. The first part of the chapter examined the role of task-relevant information in casting creativity ratings based on data from Experiment 1 & 2. It was expected that (H_4) those participants who were made aware of the name and content of criteria linked to creativity would cast feature ratings with more predictive power creativity ratings than the participants who were not made aware of which criteria are connected to creativity ratings. Also, it was hypothesised that (H_5) the different order of completing the Feature rating task and the Investment task would provide a different amount of available information to the participants at the time of casting their creativity ratings. Therefore, the mean of the creativity ratings would be affected by the order in which the experimental tasks are completed. No such effect was found, based on which the verity of the hypotheses cannot be supported. It might be that task-relevant information does not affect creativity ratings but also that applying a different methodology could achieve such effect.

Subsequently, Experiment 3 was introduced, and it tested whether meta-information regarding the nature of the task would affect creativity ratings and the corresponding certainty ratings, as well the subjective experience of the participants. In this experiment, (H_6) it was assumed that if people need to evaluate the creativity of ideas, they make less certain judgments than what they would make if they were informed they must evaluate business ideas per several dimensions, not mentioning creativity explicitly. In addition, (H_7) participants were expected to give less coherent creativity ratings than viability ratings as they are expected to have more expertise in estimating the usefulness/feasability than the creativity of an idea. For this, we expected to find a difference in the average of creativity ratings between the 'creativity' and the 'nothing special' groups, whilst no difference in the average of the viability ratings between the

groups. Given the pattern of the results, it can be stated that meta-information about the task did not affect the ratings. However, the effect might be only marginal, in which case a within-subject design could reveal it better. Based on the data, it seems also that participants were no more confident judges of feasibility than of creativity.

Apart from working with noisy stimuli, meeting the task requirements was also challenging for the lay participants. Based on their feedback, it seemed opaque to them how the degree of creativity should be computed for casting their rating. The lack of prior experience seemed like a gap being too wide to bridge over. Providing relevant information about the features *before* starting the task did not change the performance at all. Undoubtedly, providing information about the task performance *after* the completion of the task could not influence it, either. Finally, attempting the implementation of a trial-to-trial learning method would not feasible in case of the current version of the paradigm due to the limited amount of trials. The question of whether providing task-relevant information in a descriptive or experiential manner could train laypeople to perform more similar to experts remains open. The null result found after providing task-relevant information can be attributed to both not presenting the information in a sufficient way for processing it (e.g., it is too abstract thus not straight-forward enough to implement it immediately, there is not enough time for the information to sink in, etc.) or the presented information indeed not affected the judgment process considerably. Also, it was found that the notion itself that creativity must be evaluated is not making participants perplexed. They have not been found to become less certain in their evaluations as compared to evaluating other factors, as well as did not report daunting feelings associated with evaluating creativity.

To sum it up, the aim in case of both manipulations was to figure out how does providing and not providing participants with explicit information about the task affect their judgment about creativity. The null results can be interpreted in two ways: it is possible that alternative

task instructions are not sufficient to influence the metacognition regarding creativity, as well as the creative evaluation performance itself (however, this does not cross out the possibility that using different technique could achieve the effect). Another explanation might be that despite the anecdotal evidence, metacognition about creativity judgments is not influencing the task performance, or at least not in such manner which we could have captured by either the ratings or a simple questionnaire. It is difficult to make straight-forward conclusions, since there is no theory outlining the connection between meta-cognition and creativity. The construction of a theoretical framework can be informed by practical, data-driven research.

Similar things can be said about the research coming up in the next chapter. One of the difficulties with comparing the research paradigms is that some of them assess creativity in an absolute manner (by awarding scores to different criteria), while other paradigms establish rankings in a pool of creative products, where the scores must be interpreted relatively to each other. No theory has outlined yet how absolute and relative judgments made about creativity should be compared, or whether they are comparable at all. *Chapter 7* is considering this issue by discussing the way of casting judgments as a potential contextual factor influencing the evaluation of creative ideas.

CHAPTER 7: EMPIRICAL STUDY III

7.1 Overview

The present chapter investigates the similarities and the differences between making absolute and relative judgments about creativity. A short introduction about this topic is included in *Chapter 2*. Here, the focus is on the third research question of this thesis, which raises the possibility of certain contextual factors influencing the evaluation of creative ideas. A few key insights from the previous experiments are that the basic findings of the novel research paradigm regarding the weighting of criteria fed toward the overall judgment about creativity can be replicated (*Chapter 5*) and that the amount of available information about the features did not influence creativity ratings considerably (*Chapter 6*). To address whether creativity ratings are indeed consistent, this chapter probes whether different ways of evaluating creativity affects the ratings in a within-subject design. The aim of the empirical study was to extend the scope of the evaluation process from absolute creativity judgments to comparative judgments. Participants were prompted to 'shortlist', i.e., to rank which projects are more/less creative than others, instead of judging them one-by-one. An answer is sought to the question whether these two types of judgments could be treated as equivalents when applied to the present paradigm. The overarching hypothesis is that participants rating of the ideas should not be subject of the way in which they are judged. Further, the ranking of the ideas is supposed to be independent from asking them to select the best or the worst ideas. The procedure was motivated by ecological validity as relative judgment making models how creative ideas are selected in most occasions, e.g., how grants are awarded or how the creative industry works.

7.2 Making Absolute vs Relative Judgments

7.2.1 Introduction

Although measuring the same concept with various tasks is expected to yield essentially the same results regardless of the task, this assumption is rarely checked, at least in the creativity literature. Here we test in a within-subject design whether judging identical stimuli in three different ways modulate the findings. Participants were instructed to judge project proposals one-by-one such as in the chapters above to obtain absolute ratings of creativity. Additionally, they were prompted for the discernment of the ideas, that is, they needed to create a shortlist of selected ideas. This procedure is similar but not identical to screening processes, in which ideas are eliminated based on their compliance with industry standards (Bink & Marsh, 2000). While during the screening criteria is applied to make the judgment, here relative comparisons had to be made to produce the selection of ideas. Finally, to increase the ecological validity of the study and to gain useful information about the weighting of participants' rankings, they were instructed to allocate a limited budget to the shortlisted ideas based on the judged creativity of the ideas. To our knowledge, this study is the first to compare the making of absolute vs relative judgments in the creativity literature. There are a few previous examples testing how sorting ideas into selected sub-groups impacts creativity assessment, nevertheless, different methodologies of judgment making were not contrasted in them. An early example for the study of making relative judgments is when participants had to judge a set of eight ideas presented on cards to them (Runco & Charles, 1993). In this case, the stimuli were very simple, responses collected from divergent thinking tasks, but sorting them was still not straight-forward due to the multiple attributes one needed to consider while appraising creativity (and choosing the suitable pile for the card). Another empirical study shed light on the role of discernment by creative ideas – Silvia (2008) instructed his participants to choose the top two responses from the available alternatives. He preferred the protocol of asking people to choose a subset of

responses as “forced-choice measures of creative judgment could offer insight into how people make such real-world creative judgments” (p. 144).

Blair & Mumford’s (2007) paradigm lays the closest to the present research; lay judges of creativity had to evaluate project proposals which have been previously created to receive funding from a non-profit foundation. The participants were briefed about the scenario for which the stimuli were created, were even informed about the history of the funding body. This served as the instruction to align the judging criteria: “the Jackson foundation’s primary goal was to expand their program along new and useful avenues that might serve to lessen society’s significant problems” (p. 204). In the experimental task, participants had to make relative judgments and select the one from each of the 72 pairs of available project proposals which they considered for further development. Finally, the participants had to make a shortlist and select those ideas which they recommended for receiving funding. The manipulations employed in this study were the allowed time, the number of ideas which could be selected to receive funding (limiting them to 5 or imposing no limit), and the added pressure of social evaluation. The findings suggested that lay judges avoid choosing risky and original ideas, these options were not selected for further development. The sample of undergraduates preferred the ideas which were safe choices and consistent with societal norms and expectations. Participants who were limited in the number of ideas they could recommend for funding selected fewer original ideas than those who had no limit in the number of proposals they could recommend.

Finally, a very recent study investigated different contextual factors which have a potential to facilitate the selection of creative ideas (de Buisonjé, Ritter, de Bruin, ter Horst, & Meeldijk, 2017). From a pool of 18 ideas, participants had to select the 5 most creative ones, which is an identical procedure to the relative ranking task reported here. Before the shortlisting took place, participants in the experimental condition were primed with promotion focus and positive affect and a self-affirmation task was also introduced. More creative ideas were

selected in the experimental group than in the control group receiving no contextual manipulation, which suggests that contextual factors can aid the discernment of creative ideas.

7.2.2 Hypotheses

(H_8) The ordering of the absolute creativity ratings corresponds to making relative judgments about how creative the stimuli are compared to each other. That is, more than half of the 6 shortlisted projects was expected to be listed in the top/bottom six ideas too when converting the absolute judgments to an ordinal scale.

(H_9) The explicit rank ordering established in the shortlisting task aligns with the weighted rank ordering established in the budget allocation task with regard to the ranking across the projects. That is, the projects should be aligned in their ranking position to show the judgments are reliable.

Subsequently, the data was analysed according to task condition to determine which type of creativity ratings are aligned with each other.

(H_{10}) The absolute creativity ratings were not expected to differ between the two task conditions (shortlisting the best vs. the worst ideas), as the task instructions did not concern these ratings.

(H_{11}) However, we expected the shortlists to be severely influenced by the task condition. It was hypothesized there would be no overlap between the 'best ideas' and the 'worst ideas' shortlists, i.e., they would share less than one idea on average.

Finally, we wanted to explore whether experts' rankings of the projects would be similar to non-experts' rankings. That is, even if their absolute judgments were found to be largely dissimilar, it is possible that the ranking of the projects is better aligned between the two groups.

(H_{12}) It was expected that the rankings provided by domain-specific experts and non-experts would largely overlap. That is, an overlap of at least half of the projects in the best six

ideas and the worst six ideas was expected between the two groups. It was not required the projects to be listed in the same ranking position.

7.2.3 Method

7.2.3.1 Participants. 118 healthy participants (50 males, age: $M \pm SD$ = 31.92±8.01 years) were recruited using the online platform Prolific Academic. All participants gave written informed consent and were compensated with £5. *N*=5 participants were excluded due to their relevant expertise in cities. The study protocol was approved by the Local Ethics Committee at the Queen Mary, University of London.

7.2.3.2 Materials. The stimuli consisted of *N*=15 project ideas, which were created based on proposals collected from an open-source platform, OpenIDEO (2011); all project ideas were entries for a competition on "How might we restore vibrancy in cities and regions facing economic decline?" and were edited into two paragraphs of text. Each project was subsequently rated by a pool of experts (N = 16) on four features: originality, utility, scalability and riskiness. A detailed account of the creation of the paradigm can be found in *Chapter 4*.

7.2.3.3 Design. Experiment 4 employed a between-subject design. There were two conditions: in the 'best ideas' condition, participants had to select 6 ideas which they found the most creative in the shortlisting task, whilst in the 'worst ideas' condition, participants had to select 6 ideas which they found the least creative in the shortlisting task. *N*=55 participants were instructed to shortlist the best 6, whilst *n*=63 participants were instructed to shortlist the worst 6 project proposals. *N*=10 participants have shortlisted the best 6 ideas despite being instructed to do the opposite. Since these participants were not at all aware of the mix-up, their data was simply added to the 'best ideas' condition. This resulted in a final number of *n*=65 participants in the 'best ideas' condition and *n*=53 participants in the 'worst ideas' condition.

7.2.3.4 Procedure. After providing informed consent, participants completed a brief questionnaire on demographic information, relevant expertise, degree of current motivation, and their internalized interpretation of creativity (e.g., originality). The participants of the ’best ideas’ condition were instructed to shortlist the best 6 ideas, while the participants of the ’worst ideas’ condition were told to shortlist the worst 6 ideas. Subsequently, each project (*N*=15) was presented for 60 s (or could be skipped any time before). Afterward, participants were instructed to rate each feature on a VAS from 0 (not at all) to 100 (most), as well as to type in what was the best and the worst aspect of the presented project. Finally, they were asked whether they consider the presented project to be included in their shortlist. The presentation order of the projects and the features were randomized across participants.

Once each project was judged in an absolute manner, participants were presented with a list of all projects and had to drag-and-drop the 6 projects which they selected for the shortlist. On task completion, they were also prompted to explain with their own words how the selection was made (this was a built-in check to see whether they have completed what was asked from them).

In the next section, they needed to weight the ranked ideas. This was done through allocating a budget. The titles of the shortlisted 6 ideas were presented and participants had to divide 100 coins across them. In the 'best ideas' condition, the more creative an idea was, the more coins it should have received. Conversely, in the 'worst ideas' condition, the less creative an idea was, the more coins should be allocated to it. In both conditions, explicit instructions outlined based on what principle the projects should be weighted. In the ’worst ideas condition’, participants were instructed by the following statement: *"This might sound a bit strange but if a project is less creative than the others, it should receive more coins than the others. (So a 30 coin idea is less creative than a 15 coin idea.) Please focus only on the creativity of the project while making your choices. You will need to spend all 100 coins."* Finally, participants needed

to explain shortly the strategy behind the distribution of their budget (this was another built-in check too see whether the task instructions were comprehended).

7.2.3.5 Data Analysis The data was analysed by using Statistical Package for Social Sciences (SPSS) version 20, the *irr* package of R (Gamer, Lemon, Fellows, & Singh, 2015) as well as Microsoft Office Excel 2010. The full data set (*N*=1770 trials) was analysed. Three types of metrics were collected regarding the evaluation of creativity: (1) absolute creativity ratings were registered (the same way as outlined in the earlier chapters), (2) the projects were shortlisted, which resulted in a rank ordering, and finally, (3) a budget of 100 coins was allocated amongst the selected 6 projects, which resulted in a weighted rank ordering. To compare the three different scales, both the absolute ratings and the weighted ranks were converted to ordinal scales. In the *Results* section, the similarities between the acquired 2 (condition) x 3 (rating method) types of ratings are investigated.

7.2.4 Results

First, we show the results of the three creativity rating tasks. Subsequently, the results will be discussed according to the hypotheses outlined above. Starting with the absolute creativity rating task, the findings are displayed in Table 21 and 22. As expected in (H_{11}), there were no significant differences in the mean of the absolute creativity ratings between the ’best ideas’ and the ’worst ideas’ conditions (as participants completed the same task in both conditions).

Table 21

Absolute creativity ratings in the 'best ideas' condition

Project's name	Number	Mean	Standard deviation
SIP Veggie Farm	1	69.65	20.77
Intermodal	3	66.71	21.24
miLES	5	62.83	21.83
Fablab	14	62.06	21.82
The Neighbourhood Museum Collective	2	58.6	19.75
A.R.T.S.	12	56.83	24.33
Sidewalk Chalk Arts	7	52.28	25.08
School of ideas	6	50.6	23.5
Free Neighbourhood Blocks	9	50.49	26.31
Feel the city	11	49.38	25.81
Pop-up Cultural Hub	13	48.66	23.19
Street Signs With Stickers	4	40.69	25.18
Blogger Café	16	40.6	24.83
Movies to the park	15	35.91	23.88
Time capsules	8	28.09	26.55

Note. N=65. Projects are sorted from the highest mean to the lowest mean.

Table 22

Absolute creativity ratings in the 'worst ideas' condition

Project's name	Number	Mean	Standard deviation
SIP Veggie Farm	1	70.89	18.51
Intermodal	3	70	17.61
miLES	5	67.96	19.19
Fablab	14	63.02	16.41
The Neighbourhood Museum Collective	2	58.75	24.13
School of ideas	6	56.68	22.76
A.R.T.S.	12	55.68	21.46
Pop-up Cultural Hub	13	47.4	22.6
Sidewalk Chalk Arts	7	46.83	23.07
Feel the city	11	46.68	26.1
Free Neighbourhood Blocks	9	45.13	22.66
Street Signs With Stickers	4	38.19	24.78
Blogger Café	16	36.79	22.27
Movies to the park	15	34.89	22.18
Time capsules	8	25.38	22.61

Note. N=53. Projects are sorted from the highest mean to the lowest mean.

To get a cleaner measure of the absolute ratings, the descriptive data was also computed for the first responses given by the participants due in the randomized presentation order (Table 23 and 24). This was done to make sure that no relative judgment was made while rating the stimuli. Although the magnitude of the results matched the ones in Table 21 and 22, the rank order has changed. This must be due to the tiny sample sizes from which the descriptive statistics were calculated. The lack of big differences between Table 21 & 23, as well as between Table 22 & 24 verifies the validity of both measures.

Table 23

Absolute creativity ratings in the 'best ideas' condition by all of the firstly presented projects

Project ID of the project first presented	**Project title**	***n***	***M***	***SD***	**Variance**
1	**SIP Veggie Farm**	5	70.20	19.98	399.20
3	**Intermodal**	6	65.83	18.43	339.77
12	**A.R.T.S.**	3	61.00	18.52	343.00
14	**Fablab**	1	60.00	0.00	0.00
13	**Pop-up Cultural Hub**	4	57.75	18.43	339.58
7	**Sidewalk Chalk Arts**	5	56.80	16.75	280.70
5	**miLES**	5	56.40	27.83	774.30
9	**Free Neighbourhood Blocks**	3	54.33	8.96	80.33
2	**The Neighbourhood Museum Collective**	3	44.33	29.94	896.33
11	**Feel the city**	10	43.50	28.04	786.06
6	**School of ideas**	4	37.00	19.71	388.67
8	**Time capsules**	2	35.00	24.04	578.00
15	**Movies to the park**	6	34.17	21.48	461.37

4	**Street Signs With Stickers**	6	26.17	27.15	736.97
16	**Blogger Café**	2	18.00	4.24	18.00

Table 24

Absolute creativity ratings in the 'worst ideas' condition by all of the firstly presented projects

Project ID of the project first presented	**Project title**	***n***	***M***	***SD***	**Variance**
5	**miLES**	4	71.75	9.71	94.25
3	**Intermodal**	3	70.67	12.10	146.33
1	**SIP Veggie Farm**	3	66.33	5.51	30.33
7	**Sidewalk Chalk Arts**	3	65.33	15.01	225.33
2	**The Neighbourhood Museum Collective**	8	56.00	19.03	362.29
14	**Fablab**	1	55.00	0.00	0.00
11	**Feel the city**	4	47.75	20.69	428.25
13	**Pop-up Cultural Hub**	3	47.67	28.04	786.33
16	**Blogger Café**	4	35.75	22.72	516.25
12	**A.R.T.S.**	1	34.00	0.00	0.00
15	**Movies to the park**	6	33.33	20.68	427.47
9	**Free Neighbourhood Blocks**	3	31.67	16.07	258.33
6	**School of ideas**	3	28.33	2.08	4.33
4	**Street Signs With Stickers**	4	26.50	17.67	312.33
8	**Time capsules**	3	24.00	7.94	63.00

Next, the relative ratings were sorted into descending order in both conditions. This was done using the mean ranking of each project in the shortlisting task (Table 25 and 26).

Table 25

Relative creativity ratings in the 'best ideas' condition (ranks)

Project's name	Number	Mean ranking
SIP Veggie Farm	1	10.4
A.R.T.S. (Adaptive Reuse of Temporary Space)	12	9.92
miLES	5	9.38
Intermodal	3	8.83
Fablab	14	8.71
The Neighbourhood Museum Collective	2	8.32
Pop-Up Cultural Hub	13	8.12
Sidewalk Chalk Arts	7	7.94
School of ideas	6	7.83
Feel the city	11	7.48
Blogger Cafe	16	7
Free Neighbourhood Blocks	9	6.95
Street signs with stickers	4	6.54
Movies to the park	15	6.34
Time capsules	8	6.23

Note. N=65. Projects are sorted from the highest mean ranking to the lowest mean ranking.

Table 26

Relative creativity ratings in the 'worst ideas' condition (ranks)

Project's name	Number	Mean ranking
Time capsules	8	10.81
Street signs with stickers	4	10.58
Movies to the park	15	10.45
Blogger Cafe	16	10.08
Feel the city	11	8.87
Free Neighbourhood Blocks	9	8.13
Sidewalk Chalk Arts	7	7.98
The Neighbourhood Museum Collective	2	7.17
School of ideas	6	7.04
Fablab	14	6.94
Pop-Up Cultural Hub	13	6.77
Intermodal	3	6.53
A.R.T.S. (Adaptive Reuse of Temporary Space)	12	6.51
miLES	5	6.34
SIP Veggie Farm	1	5.79

Note. N=53. Projects are sorted from the highest mean ranking to the lowest mean ranking.

The mean rankings were obtained from Kendall's Coefficient of Concordance analysis. In the shortlisting task, each participant (*n*=65 in the 'best ideas' condition and *n*=53 in the 'worst ideas' condition) was to select 6 out of 15 possible projects and rank them. They selected the 6 idea they were considering as the most/least creative. Consequently, to the remaining 9 projects a zero rating was given, indicating that the participant was not interested in including them to the shortlist. Since the participants did not use a true ranking order for the projects, the coefficient was corrected for ties within raters.

Using the Kendall's *W* analysis, an overall low agreement was found amongst the lay participants on which projects to select as part of the 'best ideas' shortlist, as indicated by Kendall's $W = 0.11$, $\chi^2_{(14)} = 95.61$, $p < .001$. In the 'worst ideas' condition, a bit higher but still low agreement was found, Kendall's $W = 0.19$, $\chi^2_{(14)} = 141$, $p < .001$.

Finally, in the coin allocation task, participants used a budget to weight which projects, from the 6 ones they included in their shortlist, should be funded more than the others. To analyse the agreement in weighting the projects, the interval scale including the exact number of coins (which are displayed in Table 27 and 28) allocated was converted into an ordinal scale. The new scale ranked the projects in ascending order, i.e., the project with the first rank received the highest amount of funding from the participant. Subsequently, Kendall's *W* analyses were conducted to detect the degree of agreement regarding the weighting of each project amongst the participants. In the 'best ideas' condition, the agreement was tiny, Kendall's $W = 0.06$, $\chi^2_{(14)} = 54.7$, $p < .001$. In the 'worst ideas' condition, the agreement was twice as big, however, still a negligible amount, Kendall's $W = 0.12$, $\chi^2_{(14)} = 91.9$, $p < .001$.

Table 27

Relative creativity ratings in the 'best ideas' condition (weights)

Project's name	Number	Mean of allocated coins	Standard deviation
SIP Veggie Farm	1	23.24	12.96
Fablab	14	19.18	10.40
Free Neighbourhood Blocks	9	17.72	8.57
Feel the city	11	17.05	9.61
The Neighbourhood Museum Collective	2	16.25	7.98
miLES	5	15.83	7.51
School of ideas	6	15.83	9.52
A.R.T.S. (Adaptive Reuse of Temporary Space)	12	15.59	7.70
Intermodal	3	15.55	6.34
Sidewalk Chalk Arts	7	14.80	10.68
Movies to the park	15	14.58	8.65
Blogger Cafe	16	14.56	10.30
Time capsules	8	12.50	6.35
Pop-Up Cultural Hub	13	12.38	5.00
Street signs with stickers	4	11.69	7.54

Note. The budget was 100 coins for 6 projects.

Table 28

Relative creativity ratings in the 'worst ideas' condition (weights)

Project's name	Number	Mean of allocated coins	Standard deviation
Time capsules	8	24.20	12.16
A.R.T.S. (Adaptive Reuse of Temporary Space)	12	19.80	29.84
Street signs with stickers	4	17.70	11.13
Fablab	14	17.31	17.45
School of ideas	6	16.92	9.26
Feel the city	11	16.17	7.15
Movies to the park	15	16.05	9.95
Pop-Up Cultural Hub	13	15.00	9.05
Blogger Cafe	16	14.57	8.83
Sidewalk Chalk Arts	7	14.10	8.07
The Neighbourhood Museum Collective	2	14.07	10.10
miLES	5	13.75	6.41
Free Neighbourhood Blocks	9	12.48	10.04
Intermodal	3	11.50	6.26
SIP Veggie Farm	1	10.00	5.00

Next, it was checked whether the hypotheses could be upheld considering the results.

(H_9) and (H_{10}) were outlined regarding the internal consistency in the ratings. That is, we looked whether the different ways of casting creativity ratings would result in the same ranking of the projects or not. In the 'best ideas' condition, there was an overarching agreement about the best project (#1, SIP Veggie Farm) regardless of the judgment task. Comparing the absolute judgment ratings to the relative ranks, all 6 top projects overlapped, and 2 of them were ranked in the same positions (#1 and #3). Comparing the relative ranks to the relative weights, 4 projects overlapped but there was no agreement on the ranking positions (except of #1). Comparing the absolute judgment ratings to the relative weights, 4 projects overlapped, and 2 projects were ranked in the same positions (#1 and #2).

In the 'worst ideas' condition, again, there was a consensus regarding which one was the worst project (#8, Time capsules). All projects overlapped between the absolute judgments and the relative ranks but only #8 was in the same rank position. Contrasting the relative ranks with the relative weights, 3 projects overlapped (#8, #4, and #11) but apart from the Time capsules, there was no agreement on the ranking positions. Finally, between the absolute judgments and the relative weights, 3 projects overlapped (#8, #4, and #11) and the position of two projects were identical (#8 and #11). Based on the data, (H_9) was confirmed as we found the different creativity judgment methods to be sufficiently reliable. (H_{10}) was disproved: although 4 and 3 projects overlapped in the best/worst ideas condition, respectively, apart from one idea standing out, no agreement was found in the ranking positions.

Next, the data was analysed according to task condition. While an overlap in the project was desired across the different judgment methods, no overlap was expected between the 'best ideas' and the 'worst ideas' shortlists (H_{11}). The hypothesis was confirmed for the shortlisting task (relative ranks). Additionally, we checked the budget allocation task (relative weights) too

but here, there were two projects (#11 Feel the city and #14 Fablab) which were part of both the *'should be funded the most'* and the *'should be funded the least'* lists.

Finally, it was expected that the rankings provided by domain-specific experts and non-experts would largely overlap (H_{12}). That is, an overlap of at least half of the projects in the best six ideas and the worst six ideas was expected between the two groups. As it can be seen from Table 4 (presented in Chapter 4), the six most creative projects rated by the domain-specific experts are in descending order: miLES, Fablab, SIP Veggie Farm, A.R.T.S., Feel the city, and Free Neighbourhood Blocks. The six least creative projects according to them (from worst to not that bad): Time capsules, Blogger Café, Street signs with stickers, Movies to the park, Intermodal, Pop-Up Cultural Hub. In both conditions, an overlap of four projects was found. In the light of the findings, H_{12} was upheld.

7.2.5 Discussion

To our knowledge, this study is the first of its kind compare absolute and relative judgment methods of creativity. The comparison was necessitated to raise the ecological validity of lab experiments - usually, only one or the other approach is applied (e.g., an absolute rating scale used by Plucker, Kaufman, Temple, & Qian, 2009; or a relative evaluation technique as seen by Kaufman et al., 2008). Here we provided one pool of stimuli to the participants, which they then needed to rate in three different ways. What is more, there were two task conditions: in one of them, participants had to select the best ideas, while in the other, they were instructed to select the worst ideas.

We found a sufficient degree of reliability across the three rating methods. Participants rated the stimuli largely similarly, which is an indirect evidence of them possessing a stable internalised concept of creativity. The ranking positions were swapped around between the relative ranks and the relative weights tasks, but these might be also due to the nomothetic

approach of the research. An analysis of individual differences could show that each participant is consistent in rating the ideas.

Another finding corroborating with this notion is that the best ideas and the worst ideas were clearly separated in the shortlists. There was no overlap between the projects in the 6 highest positions in the shortlisting task.

Finally, a good agreement was found between the domain-specific experts and the lay participants about which ones are the best and the worst ideas in the pool. This evidence is supporting the claim that when people are asked to evaluate creativity in a more polarised way (i.e., 'which ones are the best ideas based on their creativity?' instead of 'how creative is this idea?'), their opinion fall more closely with the experts'.

In conclusion, non-expert participants were found to possess a stable internalised model of creativity, their ratings were reliable even if they needed to make judgments according to differing task instructions. By making relative judgments, a good agreement was found between the lay and the expert participants.

CHAPTER 8: GENERAL DISCUSSION

8.1 Overview

The aim of this thesis is to examine the basis on which people with and without relevant expertise evaluate creative ideas. In six experiments, experts' and non-experts' judgment was examined regarding urban design. Two experiments established the expert ratings of the stimuli. Further two experiments explored the extent to which non-experts relied on four features (originality, utility, scalability, and riskiness) for judging the creativity of novel project ideas while the level of motivation was controlled. In another experiment, the effect of providing explicit task-related information was tested. A final experiment assessed the differences between making relative and absolute judgments about creativity.

The discussion is structured as follows: first, the key hypotheses, methodologies, and the main findings will be summarised. Then, the relevance of these results will be considered: based on (a) what they have contributed theoretically and (b) what they are implying practically. The theoretical aspect is divided to the appraisal of the full methodology, to the comprehension of the results in case of each type of group contributing to this research (non-experts, domain-general experts, domain-specific experts), and then a global comparison of the results with earlier studies. The practical aspect is divided to guidelines on implementing criteria for judgment making, a description of optimal environment for evaluating creativity, as well as opportunities for optimisation. Once the results and their implications have been assessed, the limitations of this research will be catalogued too. The paradigm, the experimental manipulations, as well as relevant difficulties characterising the conduct of creativity research will be considered. Informed by both the prospects and the constraints, appropriate directions for future research will be laid out. Lastly, a conclusion reviewing the thesis research will be offered.

In the present research, three broader questions were explored. First, (Q1) in a noisy environment, what information do judges use to evaluate creative ideas? What criteria is applied internally to form an overall creativity judgment? The thesis research investigated the link between creativity-related criteria and overall creativity ratings. Criteria were selected to test to what degree can the evaluation of creative ideas be predicted by using them. This is related to the second broader question of this thesis, which delves into (Q2) the weighting each one of the criteria is carrying towards forming the overall creativity rating. Both domain-general and domain-specific criteria were tested, and it was examined whether being original or useful is more important in an urban domain. The third large question addressed in this thesis (Q3) was seeking to understand how certain contextual factors influence creative idea evaluation. Three research themes emerged, each of them was studied in a separate empirical chapter. Namely, it was investigated how the level of motivation, the amount of available information about the study, and the manner in which the rating is cast influences the evaluation of creative ideas.

These questions were imposed because there is little known about the basis on which creative ideas are judged. Creativity research suffers from the 'criterion problem' (McPherson, 1963, Shapiro, 1970), which makes the measurement of the concept troublesome (the details of this are discussed in *Chapter 2* and *3*). As other research paradigms of creativity (cf., Long, 2014), the present paradigm also applied creativity-related criteria to approximate creativity. However, the four selected criteria, originality, utility, scalability, and riskiness, were not used to substitute a creativity score. Instead, they were used to see to what extent can these values predict an overall creativity rating provided by an evaluator. *Chapter 1* and *Chapter 3* provided an overview about the main issues regarding the judgment making about creativity. Briefly put, using experts' estimates for assessing creativity is not an optimal solution but the best we have

at hand. The aim of this research was to further unfold the yet ineffable evaluation process by identifying and quantifying relevant criteria using a novel paradigm.

Overall, our procedure included a diverse toolkit of creativity measurements. Four creativity-related criteria were combined: two of them were extracted from the consensual definition of creativity (Runco & Jaeger, 2012) and they were expected to be the primary driving forces behind the judgments. The second two criteria were selected based on practical grounds - they were reckoned by investors while evaluating creative ideas (Kaplan & Strömberg, 2000). It was assumed that the rating of a criterion is in a linear relationship with judging an idea as creative and it was probed using the Lens model's framework whether expert judgments can be used as proxy measures for the criteria.

To break more ground for the modelling process, pilot work was started with predictions drawn from the literature and from anecdotal evidence, and an equal weighting was given to the features as the null hypothesis. Apart from the criteria, expert ratings were used as a proxy to creativity. Although this method is the gold standard of measuring creativity (Carson, 2006), using expert ratings was found to be a problematic proxy for creativity research. In the present research, expert judgment was applied with a twist: instead of using qualitative analyses or Likert-scales for the assessment, the judges' opinion was captured on a visual analogue scale comprising of a 0-100 interval scale. Finally, a data-driven methodology was also employed, as the expressed and the implemented conceptualizations of creativity were contrasted with each other using frequency analysis. Participants had to list what counts as creative to them (expressed values), which was then compared with how they evaluated the criteria (implemented values).

As for the contextual factors, the ones related to the typical problems associated with performing a complex, daunting task were focused on. First, participants were offered with various incentives to align their cognitive efforts. The impact of motivation on creative

evaluation was investigated. From the feedback received, it turned out that lay judges cannot enhance their performance due to a lack of confidence and understanding regarding what creativity is exactly and how it should be evaluated. Thus, explicit information was supplied to the participants before embarking on the task to see whether that would help the judgment making process and ease the frustration. Despite of implementing such a manipulation, judgments did not become more certain, neither the task performance more similar to those of the experts. The next 'candidate' source of problem was the available meta-information on the specific task. It was checked whether declaring that the task is to 'evaluate creativity' and not only to 'make judgments on various factors' would confuse participants already, or even make them anxious regarding the task, resulting in less systematic creativity ratings as compared to other, more frequently practised ratings. Finally, after investigating the absolute judgments of creativity extensively, the scope of the investigation was extended to gain more ecological validity. It was probed whether it is easier, more difficult or the same to evaluate creativity in a 'standalone' version or by rating a batch of projects together. The latter seemed to be more frequently applied in everyday judgments of creativity, thus the consistency of creativity judgments was examined: are they differently provided in contrast to evaluating projects independently from each other?

The findings of the set of studies carried out were partly confirming our hypotheses outlined in *Chapter 4* and partly revealed some unexpected insights. The mini-discussions dedicated to the interpretations of the results at the end of each empirical chapter already touched upon the key takeaways of the presented research, but this chapter will further synthesise the insights by inserting the extracted information into a larger picture.

8.2 Main Findings

The main findings of this thesis can be summarised with respect to five notions.

The first one is titled as ’opening the black box’. While people without relevant expertise in a certain domain were deemed previously as such who would make noisy, unreliable judgments about creativity, our findings suggest that they are in fact possessing a robust internalised model of creativity. In *Chapter 5*, based on the ratings of four creativity-related criteria which participants gave ‘subjectively’, 51.9% and 41.5% (in Experiment 1 and 2, respectively) of the variance in the creativity ratings could be explained. In *Chapter 7*, a sufficient degree of reliability was found across the three rating methods. This means that participants rated creativity consistently, regardless of the method in which they were asked to cast their ratings. Additionally, separate groups of lay participants had a good agreement on which projects are the best and the worst in a given pool of ideas. The best and the worst ideas were clearly separated in the shortlists, i.e., there was no overlap between the projects in the 6 highest positions in the shortlisting task.

The second one is titled as ‘function is key’. In *Chapter 5*, the parameter estimate of utility was more than twice as large for the prediction of creative ideas as the parameter estimate of originality. This finding was interpreted as the lay participants’ preference for usefulness in a domain of functional creativity (Cropley & Cropley, 2005, 2008). In other words, judging proposals about urban design evoked an enhanced weighting for functionality and impact. This explanation was also supported by the analysis of lay participants’ expressed creativity conceptions (*Chapter 4*). When they were asked “*what is creativity to you?*” in general, without the influence of the stimuli, originality was mentioned the most frequently, many more times than utility. Although this measure was confounded by the task instruction implying originality as one of the possible answers, the difference between the frequency of the recalls was so large

that originality can be assumed to be a more integral part of lay participants' spontaneous creativity definition than usefulness, feasibility, or any other concept related to utility.

The third one is titled as 'usual judgment making manipulations do not work here'. *Chapter 5 and 6* listed several attempts of finding methods which would make lay people more similar to experts in judging creativity. Neither the manipulation of motivation, the metacognition about creativity, or providing explicit information about the criteria changed the way how lay people judged creative ideas. A possible explanation for the lack of change in the ratings might be that below a certain level of competence, using incentives and other manipulations cannot prompt a substantial change. It is speculated that participants want to do their best and even if they receive extra information on *what* they should do, they will still not know *how* to do it. Further, it might not be a fair expectation from lay participants to ask them to rate creative products as experts do. The next notion outlines the reasons for this.

The fourth point is titled as 'both experts and non-experts make subjective judgments, but experts are more informed'. As discussed above, attempts were futile in aligning non-expert participants with our expert pool (*Chapter 5* and *6*). *Chapter 4* and *7* showed that the ratings of domain-general and domain-specific experts differed too. Apart from utility being key for creativity, there was hardly any agreement between the domain-general experts and the non-experts. Specifically, it was found that when experts rated the creativity of the projects, their weighting of utility and riskiness but not their ratings of originality and scalability predicted non-experts' creativity judgments (*Chapter 5*). Domain-specific experts were not found to inform their judgments using the four features. Although there is no clear explanation for this, it can be speculated based on the domain-specific experts' qualitative responses that prior holistic experiences shaped their judgment more than individual feature ratings. It might be the case that the features are not simply added to each other but are interfering with each other while forming an overall judgment. E.g., legislative regulations were mentioned as determining

factors, which affect both the scalability and the riskiness features. It must be also noted that the evidence from the thesis research does not allow for making sweeping statements: the domain-specific expert sample was small and heterogenous, therefore the influence of individual preferences and biases might have reached a greater magnitude. Additionally, domain-general experts' utility ratings could predict domain-specific experts' creativity ratings, which suggests that utility is the main driving force behind making judgments about functional creativity. Combining these two results, the conclusion is that domain-specific experts have a different internal conceptualization of functional creativity than novices and domain-general experts. They are affected by their specific experiences and do not rely on adding up individual features for forming holistic judgments. Also, there are many kinds of domain-specific expertise and each should be studied in homogenous groups to delineate the underlying effects.

Chapter 3 considered the measurement issues and went into extensive details about the feasibility of using expert judgments as the criterion of creativity. There, the conclusion was that according to the logic of how concepts are measured in psychological science, expert judgment is one of the best proxies available, even if not an optimal one, for the 'true state' of creativity. This is because all judgments of creativity are subjective but domain-specific experts are possessing the most information for making accurate predictions about creativity. However, both descriptive data and agreement measures of the domain-specific experts' ratings showed a large variance in the creativity ratings, making the ratings no good substitutes for the criterion values.

The other requirement of conducting the Lens model analysis is that experts and non-experts must use the same cues to inform their judgments. This was tested while treating both experts and non-experts as participants providing two sets of judgments. Non-expert and the domain-specific expert participants of this research were not found to inform their judgments based on the same criteria. While non-expert participants were found to incorporate all four

cues (features) to their judgments made about creativity, no evidence supported that domain-specific experts would be using any of the cues. A silver lining might be the finding from *Chapter 7*, in which more agreement was found between experts and non-experts if the task was to make relative, not absolute, judgments about creativity. Selecting the best and the worst ideas in the pool, i.e., evaluating creativity in a more polarised manner, streamlined the lay judgments more with expert judgments than previous efforts. (It might be that lay participants are more familiar with this type of task.)

Finally, the fifth notion is titled as 'a more creative approach is necessary to move creativity research forward'. This section shortly recaps what are the novel methods and findings in this thesis. First, in *Chapter 5*, using the four creativity-related features and linear models to predict creativity, a bigger proportion of the variance could be accounted for than in previous studies. Second, a new insight was gained on lay people's creativity model. As outlined in the same chapter, previously, non-experts were thought as those who would receive meaningful data as input but would generate noisy and unreliable creativity judgments as output. We looked into the 'black box' and found that although there are vast individual differences regarding the judgment of creativity, non-experts have a robust internalised model of it. Third, a new model was suggested as a potential method to compare experts' and non-experts' information utilisation regarding creative ideas. In *Chapter 3*, Brunswik's (1952) Lens model was considered as a novel method for creativity assessment. Fourth, the first comparison was provided about making absolute vs. relative judgments regarding creativity in a within-subjects design (*Chapter 7*) and comforting results were found: a large overlap was observed between the results obtained from the two methods. And finally, using spontaneous recalls of lay participants' creativity concepts, two additional features were identified which could be used as creativity-related criteria in future studies about judging creative ideas: 'cleverness' and 'artisticness' (*Chapter 4*).

In the next two sections, I discuss both the theoretical and the practical implications of the results.

8.3 Theoretical Contribution

First, the thesis investigated the mechanisms behind the subjective nature of creativity evaluations. Creativity is measured up to previous experiences, thus, its judgment is relative because the extent in which something is seen as creative does not solely depend on the product but on the interaction between the product and its evaluator (Csíkszentmihályi, 1999). The evaluation process is situational and differs from person to person. What can be deduced here is that the judgment of creativity is inherently subjective, and that only extremities will be judged unanimously (as seen in *Chapter 7*).

This thesis set out to look behind the subjective nature of rating creativity and provide a model based on which creativity can be predicted better than previously. The most important novel finding is that lay people too were found to possess a robust internal model of creativity: their creativity ratings could be predicted by looking inside their 'black box', i.e., if their perception about the input variables were accounted for (*Chapter 5*). Also, they were found to rate creativity with high consistency across different rating methods (*Chapter 7*), which provided another proof of the reliability of the internal model. The difficulties regarding the measurement stem from a theoretical issue: although nowadays there is a growing consensus regarding what constitutes creativity (e.g., Runco & Jaeger, 2012), in practice, the judged degree of creativity is the result of comparing the product in question with the similar products encountered with before. Standard paradigms used in other fields of judgment research, such as the Lens model, were not suitable to measure how judgments are informed by cues extracted from the environment. Therefore, there is no direct evidence for judges being better, not only different judges of creativity than novices.

Reasons for why experts are considered as better judges of creativity than those without expertise are that a) they have access to a bigger inventory of previously encountered products and thus possess a more structured internal model of creativity than laypeople do and b) when they encounter a product which cannot be compared to anything they have seen before, then they can be more certain that the product is truly extraordinary and creative, while laypeople in the same situation would be in doubt whether the creative product is only novel to them or would be also novel to an expert. The point here is that while the functionality of an object or a proposed plan is something which can be judged based on the knowledge possessed by a lot of humans, the judgment of novelty is extremely dependent on the socio-cultural background of the judge and this kind of expertise cannot be trained easily. Novelty is experience- and timing-specific, thus it is not enough for a judge to recall the relevant information for the judgment, but that information will also be turned into a different output value in 2007 and 2017.

The perceived amount of risk depends also on the richness of the judge's experience (e.g., previous malfunctions can enrich the representation of risk). A person who has been troubleshooting a lot can judge future situations more accurately by calculating the risk in those dimensions too which might be a blind spot for others. Another aspect here is that there are common myths surrounding what creativity means and the folk psychological view of creativity involves taking big risks. However, the risk in the present research was not presented in the problem selection (where it is preferred to take risks) but in the execution of the solution (where it is not). Therefore, the amount of risk desired for an idea to be creative was different by the experts and non-experts. Finally, the concept of scalability (interpreted as the potential for growth) is a dimension which was included due to experts using it and one which was opaque to most of our lay participants. When non-experts were prompted to take this dimension into

account, they did a good job with it, but this feature was hardly ever mentioned when asking about their internal model of creativity.

There are quite a few papers assessing the typical fallacies lay people make when evaluating creativity (Blair & Mumford, 2007; Dailey & Mumford, 2006; Licuanan et al., 2007; Lonergan et al., 2004) and the reasons for these errors are most likely to stem from the information deficient state in which non-experts have to cast their ratings as contrasted to the experts. Especially since creative products are often complex and therefore difficult to understand at the first view, there is a lot of degrees of freedom when it comes to their appraisal. Lay people code the available information differently (and are also likely to code the more available layers when more layers of interpretation are offered), can recall fewer, and sometimes less relevant, items from their memory storage than their expert counterparts, and are also speculated to have a less sophisticated internal model comprising of fewer dimensions than what experts have. It is no wonder then that the information sampled from the stimulus and compared with the personal experiences gets weighed in varied ways, and this weighting of the dimensions is the one based on which the judgment is manufactured. Tracing the multiple ill-defined steps of evaluating creativity, the likelihood of two persons to rate a creative product in the same manner seems low.

In sum, the theoretical contribution of this thesis is that it investigates the mechanisms behind evaluating creativity. Although the aim of cognitive science is to provide explanations, there is no good explanation yet of why experts are evaluating creativity differently from non-experts. Here, a methodology was developed to disentangle how criteria is applied towards rating creativity in project proposals. By acquiring this information, the results of the investigations can be now placed into the wider matrix of creativity literature.

This thesis yielded a few findings conflicting the earlier results in the literature. Previously, testing the boundary conditions of finding a product as creative (Runco & Charles,

1993), originality was reduced to zero in an idea set to control for that dimension. Creativity ratings were found to decrease when the number of appropriate ideas in the set increased, which is the opposite of the finding in our studies.

In another study (Caroff & Besançon, 2008), when participants had to rate creativity without receiving any special instructions, appropriateness was in a linear relationship with creativity. Our results align with this finding, however, when participants of the French study received explicit instructions on how to judge creativity, appropriateness did not predict creativity anymore. Although whether our manipulation in *Chapter 6* can be interpreted as providing explicit instructions is up to debate, it must be concluded that the same effect was not found in our case.

Finally, when investigating the internalised creativity model of lay participants, Storme & Lubart (2012) found too that novelty is the most frequently mentioned criterion (called as expressed value above). The alternative result here is that they replicated the dominance of novelty in the experimental task too; in fact, novelty had the highest weight ($\beta = .65$) towards finding an advertisement creative. In our case, utility was a twice as good predictor of creativity than novelty. Although the employed paradigms were largely dissimilar in term of their structure and the stimuli presented, obtaining highly conflicting results in different studies highlights the issue of generalisability of the research results.

8.4 Practical Implications

First, the applicability of the present paradigm is discussed, then, the applicability of using expert judgments for the evaluation of creativity. Finally, guidelines are outlined on how to implement criteria for judgment making about creativity.

To begin, only one paradigm was used in all studies, the one which was newly developed as part of the research project. Therefore, its validation is not complete yet. Although face validity is high, criterion validity should be tested further. Next, both the convergent and the

discriminant validity of the paradigm should be studied by contrasting the present paradigm with other related paradigms. However, to our current knowledge, there is no other paradigm which would be tightly linked with the one presented here. This paradigm achieves a direct measurement of creativity using visual analogue scales, while the stimuli simulate the environment in containing both relevant cues and distractor words. Therefore, the potential pool of related studies is narrow and on top of this, the findings obtained from different creativity-related domains are hardly comparable (Long, 2014). Given our paradigm is linked to the domain of urbanism, critical decisions could be difficult to draw if only supported by the data acquired from a different domain due to the domain-specific nature of creative performance. Nevertheless, there are several things which can be deduced based on the findings obtained using this paradigm.

If we look at the mean rating associated with each project (Table 3), it is striking that all projects with a low level of utility received mean ratings lower than 50. The same cannot be said about novelty: if an idea was feasible, even if not new, it could still earn a relatively high mean rating (up to 70). Please note that the project titled 'Time capsules' received the lowest creativity rating ($M \pm SD = 29.32 \pm 26.19$). This project was neither novel nor useful but was low on risk and easily scalable in turn. Still, the two projects, which had the same low-novelty, low-utility parameters, and additionally higher risk and lower scalability, received slightly higher mean ratings (see 'Movies to the park' and 'Blogger Café' in Table 3).

The conclusion here is that the paradigm is still in its early phase; the pool of ideas as well as the array of features could be expanded. However, its core principle, which is about using ecologically valid stimuli and linear models to predict creativity ratings based on creativity-related criteria is simple and effective to use. With obtaining individual data on the input variables, the computation process of the output value becomes clearer. The outcome is a reduced portion of noise and a higher extent of explained variance in the ratings, which means

we are one step closer to a realistic modelling of how people evaluate creative ideas. The structure of the task could be also implemented in different domains of creativity, with this method, even aesthetic judgments of visual stimuli could be understood better. Applying the Theory of Social Judgment to the measurement of creative evaluation would also mean that there is a way to directly compare experts' and non-experts' utilization of available cues. Using Brunswik's Lens model (1952), it is becoming tangible what are the distinct information sources behind the ratings given by the two groups of judges. Testing the same cues to assess experts' and non-experts' ratings, the weight of each cue becomes comparable. Another plus side is that the model can be extended very easily, thus, an unlimited number of features could be fed in as fallible cues related to the judgment.

A bigger question is whether taking expert judgments is the best way to assess how creative a product is. If so, then what would be the approximate number of judges required to make valid assessments? The present research did not collect data for resolving this issue, however, it showed that there is a far-from-perfect agreement amongst domain-specific experts and also that lay people can never replace experts, at least in the relatively highly codified domain of urbanism. What follows from this is that if researchers wish to use experts as „assessment tools" of creativity, then they need to convince at least one, and ideally multiple, group(s) of experts to volunteer for the study. A group is recommended because it seems that only a pool of similarly sampled participants is providing similar ratings. Further, contrasting the ratings obtained from multiple groups of experts could shed light on both the reliability and the validity of the judgments. Nevertheless, it is acknowledged that achieving such a desirable state is difficult. Then, empirical research efforts should be made towards testing what would be the rule-of-thumb on using experts, and this side of the assessment should be standardised; the sooner, the better. An alternative of using high quality evaluators for the assessment could be potentially (so only in some cases) to use a high quantity of evaluators. Here the idea is that

a) there are domains of creativity, such as the performing arts or the innovation of common products, where lay people are the main consumers of the products and therefore their preference should be served at least partly and b) with modern technology, accessing a big pool of people is no longer an issue. This is how crowdsourcing evaluations and recruiting judges of idea markets is turning into a trend (Mollick & Nanda, 2014; Soukhoroukova, Spann, & Skiera, 2012). Recently, massive data sets became available to researchers, so evidently it is easier to create algorithms which can analyse the hard metrics of an ecologically valid data set than spending a lot of time and effort on trying to recruit scarce and expensive human labour to produce such metrics.

Finally, after exploring what the future will bring, let us turn back to the present. Based on the research presented in this thesis, a handful of guidelines can be plotted for real-life evaluations. Creators should note that in the domain of urbanism, function is above all and developing a novel design can never happen at the expense of usefulness. On a similar note, evaluators should take into account that if time is scarce or they do not have the resources to give an overarching feedback to the creator, then they should start with appraising the usefulness of the product as that one is the crucial dimension which needs to be prioritised. This means that the outer appearance of the product is, at least in principle, inferior to functionality of the product. Another thing evaluators should aim for is making a comprehensive registry of their judgment criteria. Both a holistic impression of creativity, likely to be given on a ratio or an ordinal scale, and creativity-related criteria should be registered. For the latter, Cropley & Kaufman (2012, 2013) developed a domain-general tool which includes a large-scale inventory of potential rubrics. Once a substantial amount of data becomes available for analysis, with the linear models presented in *Chapter 5*, it will get traceable what sub-components of creativity are the most influential to perceive a product as creative, and this information can be fed back to forthcoming tenders as production criteria. Funding criteria can also be informed by the

properties which are most typical of products later deemed as creative. Turning now to the interest of companies and other businesses, one recommendation would be to not to spend a lot of resources on trying to train people with absolutely no expertise to quickly become comparable to experts in creativity assessment. Chances are good they will not, especially if the training includes written instructions instead of hands-on experience. On the other hand, the evaluation of creativity can be optimised in an environment where a) the goal for which the product is created is clear to all, b) the criteria for the assessment is already provided to the creators at the time of the production, and c) the shaping of an idea or product can be followed through the entire production process and feedback from the evaluator can be provided multiple times, from iteration to iteration (e.g., Kozbelt & Serafin, 2009; Serafin, Kozbelt, Seidel, & Dolese, 2011).

8.5 Limitations

The research presented here is an original contribution to the field, which yields both strengths and weaknesses. This section is about the latter. As the issues with the validity of the new paradigm developed for the present purposes is already discussed in the *Practical implications* section of this chapter, these concerns will be not repeated here.

Another concern regarding our newly developed methodology is that while it gets beyond the prior research efforts in scrutinizing what is underlying the differences between the perception of creativity across experts and non-experts, it does not register all the steps of the spontaneous process through which a judgment is formed. The paradigm proposed in Ch. 3 was not suitable for studying how cues support the forming of a judgment about creativity. It can only be assumed that the relevant information gets extracted from the presented text by detecting it during the provided reading time. Further, even if the relevant information is perceived by the participant, that is still no guarantee that the information bit gets used for

making the judgment, or even if so, that it gets used in the 'correct' way, i.e., with an appropriate weighting.

Another issue concerns the judges who use the information extracted from the stimulus. Namely, one could criticise that the test-retest reliability of judges has been not tested in any of the experiments. It has been not tested whether the judges would make the same ratings if their task would be to rate the same trials multiple times. This is not only due to the constraint of time and resources; a major limitation of our paradigm is that it only includes a few trials (*N*=15), moreover, the project proposals are quite memorable, so it would not make sense to present them repeatedly. Given that there is considerable variability in the judgments given about creativity (for details, see *Chapter 2*), the reliability of the judges must be checked. This issue was overcome by conducting a replication study was conducted to probe whether the task performance would be stable on a group level. The results confirmed that the ratings are vastly similar even if different samples are tested. Thus, although the reliability of individual judges was not followed up, the reliability of the ratings was examined on a group level.

A further issue with this research is an assumption which was made about it. The motivation behind the studies was to identify criteria and/or factors which enable people to evaluate creative ideas efficiently, ultimately, to equip lay judges with tools to become more similar to expert judges. The search for such tools is based on the tacit notion that such tools *do exist*, i.e., practising for multiple thousand hours is not the only way to become reasonable good at evaluating creative ideas but a proxy can be found for expertise. Thus, we set out to find a method which would align the ratings given by judges without expertise largely with those who have expertise. In judgment making research, the manipulations used in the presented studies usually work (e.g., Chang, Chen, Mellers, & Tetlock, 2016; Donovan, Güss, & Naslund, 2015; Kudryavtsev, & Pavlodsky, 2012). However, when applying classical manipulations to the present framework, the results included from zero to moderate effects. At this point, it is open

to debate whether the observed effects are due to the actual manipulations not being optimally designed (assuming that other manipulations could achieve the desired outcome) or to the evaluation process not being pliable to the same manipulations which are usually used in judgment and decision-making research. Conducting more research, including both other manipulations using this paradigm and using the same manipulations on other creative evaluation paradigms, could bring some enlightenment for this issue.

8.6 Directions for Future Research

The findings of the research in this thesis warrant possibility for conducting multiple types of follow-up research. Regarding the open questions, qualitative information could be collected to shape future quantitative studies. Theoretical and practical issues, especially blank spots and found inconsistencies, should also be tackled.

First, our experimental manipulations were drawn mostly by relying on prior protocols of judgment making research. Although trying them has led to many insights, they were not sufficient to cover all key differences between how non-experts and experts judge creativity. Remarkably, domain-specific experts were found to not use the same features which domain-general experts used to inform their judgments (see *Chapter 4* and *5*). Thus, before conducting further quantitative studies to explore what's in the shadow of their decisions, qualitative studies should be carried out to gain a deeper understanding of domain-specific experts' perception of creativity. It is acknowledged that accessing such experts is not easy, however, with good timing and a bit of luck, one could collect invaluable information. Our research granted the insight that instead of trying to approach experts one-by-one, it is a more valid and perhaps less time-consuming strategy to arrange data collection with a group of experts. This group may be formed by participating in a common cause - specific conferences might be good places to approach experts, as well as ongoing competitions might grant a good opportunity to observe

how decisions are made outside the lab environment (examples of this could be administering multiple short interviews during the evaluation process, at each stage of filtering the participants, or administering structured interviews at the beginning and after the final decision with curators of creative products). The interviews could guide the researcher's attention to systematic social and cognitive factors lying beneath the seemingly subjective decisions.

Second, there are several theoretical concerns needing resolution. Although basic research is dedicating a lot of resources to find out how the relevant information is filtered out from a noisy environment (e.g., Peelen, & Kastner, 2014), no previous study has addressed this issue specifically in creativity research. Even if the complex integration of information cannot be made visible to researchers, there are excellent methodologies already to register up to a certain point how the information is being scanned into the system. That is, an eye tracking study should be conducted to find the exact words participants spend longer time processing (also those which are overlooked by them albeit containing crucial information), then these words could be clustered to features. This would be a data-driven, bottom-up way to create features and could be complemented with the data obtained from interviewing the experts about what aspects are central to their judgments in general. Another study could determine how many and which additional features should be introduced, based on the frequency with which the features are used. The optimal number of features could be explored by factor analysis, then pilot studies using the selected features could establish the ranking of the features, as they contribute to the judgment making with different weights.

It should be also tested what is the basis of most studies finding a moderate-to-strong inter-rater reliability (e.g., Amabile, 1996; Cheng, Wang, Liu, & Chen, 2010) among experts about how to rate creative ideas, while the domain-specific experts with 10-35 years of experience participating in the thesis research showed almost no agreement at all. A not-yet-investigated explanation might be that there are factors other than expertise which influence the

consensus across experts. These moderating variables should be mapped, as they might bring us closer to understanding the specific role which expertise plays in evaluating creativity. Given the lack of empirical evidence, one can only speculate and rely on hunches to establish what the confounding variables might be. One suspect behind the agreement amongst a group of people could be their group membership, or more generally speaking, the sampling of participants. An empirical study should compare the consensus amongst experts rating the same stimuli if they are member of the same group as compared to half of them or all of them not sharing any group membership. While keeping the degree of expertise constant, it is assumed that experts sharing group membership would agree more with each other. Another study could investigate whether experts and non-experts evaluate creative ideas more similarly if they share group memberships (such as religious or political views but not expertise related to the domain) as compared to the data presented in this thesis, where they do not.

Third, while the current research protocol is to ask experts to rate the 'actual amount of creativity' in a product, even experts often err (e.g., Harry Potter was first rejected several times by publishing houses as mentioned by Licuanan et al., 2007). One proposition to correct for this is to run studies to determine a rule of thumb of how many judges are needed (approximately) in a certain domain to judge creative products with a sufficient validity. Another proposition would be to run more explorative studies outside the lab and measure which creative ideas are the most viable when implementing the criteria in a realistic situation. „Crowd evaluation" of ideas is becoming more and more popular (e.g., Soukhoroukova, Spann, & Skiera, 2012) and apart from the number of judges, the number of measurement points can be also increased. Regarding the current paradigm, an example would be to rate the project proposals' creativity from their first pitch through the final realization of the project. A good opportunity for doing something like this could be a start-up competition, where different project ideas are competing (or a hackathon, where different solutions compete with each other to solve the same problem).

After an initial evaluation, the projects should be monitored at regular intervals to gain metrics about how creative they were perceived by the investors and by the public. This would yield simultaneously acquired data points from both experts and non-experts.

8.7 Conclusion

In conclusion, this thesis set out to start filling the gap in knowledge regarding the cognitive mechanisms behind the judgments made about creativity. In a series of experiments, factors and conditions were searched for which could enhance the detection of creative ideas. People with and without expertise in urban design were compared with each other based on how they evaluate creative ideas. A novel methodology was created, with which different aspects of information processing were manipulated to see how these variables influence lay people's and experts' judgment making process.

First, non-experts' evaluations of creativity were modelled based on four related features (originality, utility, scalability, and riskiness). Their judgment of creativity relied significantly on all four characteristic features when rated by themselves and the amount of explained variance was substantial. We found that in project proposals related to cities, functionality was preferred above all - utility was twice as likely to predict creativity ratings than the originality of the product. These findings were replicated, and good reliability was found on a group level. The results suggest that experts and non-experts disagree in almost all aspects of assessment except of the perceived utility of a product.

When accounting for the individual differences in the input variables, non-experts were found to possess a robust internal model of creativity – despite previous research classifying them as making noisy, random choices. This notion was further confirmed by contrasting absolute with relative judgment making: lay people provided consistent ratings. In general, there was a substantial overlap between the selection of best and worst ideas even if they were

sorted in various ways, however, a full agreement in the ranking positions was only found by the extremities. Providing different task instructions did not make a solid difference in the creativity ratings.

Overall, the ‘black box of creativity assessment’ was opened and it was shown that although the judgments made about creativity are subjective due to drawing the value from an interaction between the creator and the evaluator, objective measurement tools accounting for the individual differences can be designed.

REFERENCES

Akturk, A. O., & Sahin, I. (2011). Literature review on metacognition and its measurement. *Procedia-Social and Behavioral Sciences*, 15, 3731-3736.

Amabile, T. M. (1982). Children's artistic creativity: Detrimental effects of competition in a field setting. *Personality and Social Psychology Bulletin*, *8*(3), 573-578.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of personality and social psychology*, *45*(2), 357.

Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Westview Press.

Amabile, T. & Müller, J. (2008) Assessing Creativity and its Antecedents: An Exploration of the Componential Theory of Creativity. In Zhou, J. & Shalley, C.E. (Eds.), *Handbook of Organizational Creativity* (pp. 31–62). BocaRaton, FL: Taylor & Francis.

Anders Ericsson, K., Roring, R. W., & Nandagopal, K. (2007). Giftedness and evidence for reproducibly superior performance: An account based on the expert performance framework. *High Ability Studies*, *18*(1), 3-56.

Baer, J. (1998). The case for domain specificity of creativity. *Creativity research journal*, *11*(2), 173-177.

Baer, J. (2015). *Domain specificity of creativity*. Academic Press.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, *16*(1), 113-117.

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

Baer, J., & McKool, S. S. (2009). Assessing creativity using the consensual assessment technique. In C. S. Schreiner (Ed.*), Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 65-77). Hershey, PA: IGI Global.

Bakker, P. (2014). Mr. Gates returns: Curation, community management and other new roles for journalists. *Journalism Studies, 15*(5), 596-606.

Basadur, M. (1995). Optimal ideation-evaluation ratios. *Creativity Research Journal, 8*(1), 63-75.

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing, 67*(4), 103-117.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., & Singmann, H. (2015). lme4: Linear mixed-effects models using Eigen and S4, 2014. *R package version, 1*(4).

Bayus, B. L. (2013). Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management Science, 59*(1), 226-244.

Besemer, S. P. (1998). Creative product analysis matrix: Testing the model structure and a comparison among products – Three novel chairs. *Creativity Research Journal, 11*(4), 333-346.

Besemer, S. P., & O'Quin, K. (1986). Analyzing creative products: Refinement and test of a judging instrument. *Journal of Creative Behavior,* 20, 115-126.

Besemer, S. P., & O'Quin, K. (1987). Creative product analysis: Testing a model by developing a judging instrument. In S. G. Isaksen (Ed.), *Frontiers of creativity research: Beyond the basics* (pp. 341-357.) Buffalo, NY: Bearly Limited.

Besemer, S. P., & O'Quin, K. (1993). Assessing creative products: Progress and potentials. In S. G. Isaksen, M. C. Murdock, R. L. Firestien, & D. J. Treffinger (Eds.), *Nurturing and developing creativity: The emergence of a discipline.* (pp. 331-349). Norwood, NJ: Ablex.

Besemer, S. P., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal, 12*(4), 287-296.

Besemer, S. P., & Treffinger, D. J. (1981). Analysis of creative products: Review and synthesis. *Journal of Creative Behavior, 15*, 158-178.

Bettman, J. R., & Sujan, M. (1987). Effects of framing on evaluation of comparable and noncomparable alternatives by expert and novice consumers. *Journal of Consumer Research, 14*(2), 141-154.

Bilalić, M., McLeod, P., & Gobet, F. (2008). Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology, 56*(2), 73-102.

Bink, M. L., & Marsh, R. L. (2000). Cognitive regularities in creative activity. *Review of General Psychology, 4*(1), 59.

Björklund, T. A. (2013). Initial mental representations of design problems: Differences between experts and novices. *Design Studies, 34*(2), 135-160.

Blair, C. S., & Mumford, M. D. (2007). Errors in idea evaluation: Preference for the unoriginal?. *The Journal of Creative Behavior, 41*(3), 197-222.

Bloom, B. S. (Ed.). (1985). *Developing talent in young people*. New York, NY: Ballantine.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Psychology Press.

Bonnardel, N., & Marmèche, E. (2004). Evocation processes by novice and expert designers: Towards stimulating analogical thinking. *Creativity and Innovation Management, 13*(3), 176-186.

Bonner, S. E., Hastie, R., Sprinkle, G. B., & Young, S. M. (2000). A review of the effects of financial incentives on performance in laboratory tasks: Implications for management accounting. *Journal of Management Accounting Research, 12*(1), 19-64.

Brehmer, B., & Joyce, C. R. B. (Eds.). (1988). *Human judgment: The SJT view* (Vol. 54). Elsevier.

Bruer, J. T. (1993). The mind's journey from novice to expert. *American Educator, 17*(2), 6-15.

Bruner, J. S. (1962). The conditions of creativity. In *Contemporary Approaches to Creative Thinking, 1958, University of Colorado, CO, US; This paper was presented at the aforementioned symposium.*. Atherton Press.

Brunswik, E. (1952). The conceptual framework of psychology. *Psychological Bulletin, 49*(6), 654-656.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments.* Berkeley: University of California Press.

de Buisonjé, D. R., Ritter, S. M., de Bruin, S., ter Horst, J. M. L., & Meeldijk, A. (2017). Facilitating creative idea selection: The combined effects of self-affirmation, promotion focus and positive affect. *Creativity Research Journal, 29*(2), 174-181.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review, 67*(6), 380.

Caroff, X., & Besançon, M. (2008). Variability of creativity judgments. *Learning and Individual Differences*, *18*(4), 367-371.

Carson, S. (2006, April). Creativity and mental illness. In *Invitational panel discussion hosted by Yale's Mind Matters Consortium*. New Haven, CT.

Cianciolo, A. T., Grigorenko, E. L., Jarvin, L., Gil, G., Drebot, M. E., & Sternberg, R. J. (2006). Practical intelligence and tacit knowledge: Advancements in the measurement of developing expertise. *Learning and Individual Differences*, *16*(3), 235-253.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making*, *11*(5), 509.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55-81.

Cheng, Y. Y., Wang, W. C., Liu, K. S., & Chen, Y. L. (2010). Effects of association instruction on fourth graders' poetic creativity in Taiwan. *Creativity Research Journal*, *22*(2), 228-235.

Conti, R. and Amabile, T. (2011). Motivation. In M.A. Runco and S.R. Pritzker (Eds.), *Encyclopedia of Creativity, 2nd Edition* (pp. 147-152). Oxford: Elsevier.

Corazza, G. E. (2016). Potential originality and effectiveness: the dynamic definition of creativity. *Creativity Research Journal*, *28*(3), 258-267.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal*, *18*(3), 391-404.

Cropley, D. H., & Cropley, A. J. (2005). Engineering creativity: A systems concept of functional creativity. *Creativity across domains: Faces of the muse*, 169-185.

Cropley, D., & Cropley, A. (2008). Elements of a universal aesthetic of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 155.

Cropley, D., & Cropley, A. (2012). A psychological taxonomy of organizational innovation: Resolving the paradoxes. *Creativity Research Journal*, *24*(1), 29-40.

Cropley, A. J. (2010). The Dark Side of Creativity: What Is It?. In Cropley, D. H., Cropley, A. J., Kaufman, J. C., & Runco, M. A. (Eds.). *The dark side of creativity* (pp. 1-14). Cambridge University Press.

Cropley, D. H., & Kaufman, J. C. (2012). Measuring functional creativity: Non-expert raters and the Creative Solution Diagnosis Scale. *The Journal of Creative Behavior*, *46*(2), 119-137.

Cropley, D. H., & Kaufman, J. C. (2013). 14. Rating the creativity of products. *Handbook of research on creativity*, 196.

Csíkszentmihályi, M. (1988). Society, culture, and person: A systems view of creativity. In R. J. Sternberg (Ed.), *The nature of creativity: Contemporary psychological perspectives* (pp. 325–339). New York: Cambridge University Press.

Csíkszentmihályi, M. (1990). The domain of creativity. In M. Runco & R. Albert (Eds.), *Theories of creativity* (pp. 190–212). London: Sage.

Csíkszentmihályi, M. (1996). *Creativity: Flow and the psychology of discovery and invention.* New York: HarperCollins Publishers.

Csíkszentmihályi, M. (1998). Reflections on the field. *Roeper Review, 21*, 80–81.

Csíkszentmihályi, M. (1999). Implications of a systems perspective for the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 313-335). New York: Cambridge University Press.

Csíkszentmihályi, M., (2014). Society, culture, and person: A systems view of creativity. In M. Csíkszentmihályi (Ed.), The Systems Model of Creativity (pp. 47-61). Dordrecht: Springer.

Csíkszentmihályi, M., & Getzels, J. W. (1971). Discovery-oriented behavior and the originality of creative products: A study with artists. *Journal of Personality and Social Psychology, 19*(1), 47.

Dailey, L., & Mumford, M. D. (2006). Evaluative aspects of creative thought: Errors in appraising the implications of new ideas. *Creativity Research Journal, 18*(3), 385-390.

Davidson, J. E., & Sternberg, R. J. (1986). What is insight?. *Educational Horizons, 64*(4), 177-179.

Dean, D. L., Hender, J. M., Rodgers, T. L., & Santanen, E. (2006). Identifying good ideas: constructs and scales for idea evaluation. *Journal of Association for Information Systems, 7*(10), 646-699.

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

Dietrich, A. (2007). Who's afraid of a cognitive neuroscience of creativity?. *Methods, 42*(1), 22-27.

Dietrich, A. (2015). *How creativity happens in the brain*. Springer.

Dietrich, A., & Kanso, R. (2010). A review of EEG, ERP, and neuroimaging studies of creativity and insight. *Psychological Bulletin, 136*(5), 822.

Donovan, S. J., Güss, C. D., & Naslund, D. (2015). Improving dynamic decision making through training and self-reflection. *Judgment and Decision Making, 10*(4), 284.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, *7*(1), 86-106.

Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology*, *70*(1), 113-136.

Ericsson, K. A. (1999). Creative expertise as superior reproducible performance: Innovative and flexible aspects of expert performance. *Psychological Inquiry*, *10*(4), 329-333.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, *49*(8), 725.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Finke, R. A., Ward, T. B., & Smith, S. M. (1992). *Creative cognition: Theory, research, and applications*. Cambridge, MA: MIT Press.

Freeman, C., Son, J., & McRoberts, L. B. (2015). Comparison of novice and expert evaluations of apparel design illustrations using the consensual assessment technique. *International Journal of Fashion Design, Technology and Education*, *8*(2), 122-130.

Funder, D. C. (1995). On the accuracy of personality judgment: a realistic approach. *Psychological Review*, *102*(4), 652.

Funder, D. C. (2001). The realistic accuracy model and Brunswik's approach to social judgment. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 365-369). Oxford, UK: University Press.

Galati, F. (2015). Complexity of judgment: What makes possible the convergence of expert and nonexpert ratings in assessing creativity. *Creativity Research Journal*, *27*(1), 24-30.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2015). *The irr package for R: various coefficients of interrater reliability and agreement.* R Foundation for Statistical Computing, Vienna. Available from http://cran.rproject.org/web/packages/irr/ (accessed September, 2017)

Goldstein, W. M. (2004). Social judgment theory: Applying and extending Brunswik's probabilistic functionalism. In: D. J. Koehler, & N. Harvey, (Eds). *Handbook of judgment and decision making. (pp. 37-61).* MA, US: Blackwell Publishing.

de Groot, A. D. (1978). *Thought and choice in chess* (Vol. 4). Walter de Gruyter GmbH & Co.

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

Guilford, J. P. (1956). The structure of intellect. *Psychological Bulletin, 53*(4), 267.

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

Gurteen, D. (1998). Knowledge, creativity and innovation. *Journal of Knowledge Management, 2*(1), 5-13.

Haller, C. S., Courvoisier, D. S., & Cropley, D. H. (2011). Perhaps there is accounting for taste: Evaluating the creativity of products. *Creativity Research Journal, 23*(2), 99-109.

Hammond (2001). Representative Design in Action in the Middle of the Twentieth Century In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 67-105). Oxford, UK: University Press.

Hammond, K. R., Rohrbaugh, J., Mumpower, J., & Adelman, L. A. (1977). Social Judgement Theory: applications in policy formulation. In M.F. Kaplan & S. Schwartz (Eds.), *Human judgement and decision processes in applied settings* (pp. 1 - 30). New York: Academic Press.

Hammond, K. R., Stewart, T. R., Brehmer, B., & Steinmann, D. O. (1975). Social judgment theory. M. F. Kaplan and S. Schwartz (Eds.), *Human Judgment and Decision Processes* (pp. 271-312). New York: Academic Press, 1975.

Hardman, D., & Macchi, L. (2003). Thinking: Psychological Perspectives on Reasoning. Judgment and Decision Making. New York: Wiley.

Hayes, J. R. (1989). Cognitive processes in creativity. In *Handbook of creativity* (pp. 135-145). Springer US.

Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, *7*(2), 193-208.

Hennessey, B. A., & Amabile, T. M. (1999). Consensual assessment. *Encyclopedia of creativity*, *1*, 347-359.

Herman, A., & Reiter-Palmon, R. (2011). The effect of regulatory focus on idea generation and idea evaluation. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(1), 13.

Hickey, M. (2001). An application of Amabile's consensual assessment technique for rating the creativity of children's musical compositions. *Journal of Research in Music Education*, *49*(3), 234-244.

Ho, C. H. (2001). Some phenomena of problem decomposition strategy for design thinking: differences between novices and experts. *Design Studies*, *22*(1), 27-45.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review*, *114*(3), 733.

Horn, D., & Salvendy, G. (2009). Measuring consumer perception of product creativity: Impact on satisfaction and purchasability. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *19*(3), 223-240.

Horng, J. S., Chou, S. F., Liu, C. H., & Tsai, C. Y. (2013). Creativity, aesthetics and eco-friendliness: A physical dining environment design synthetic assessment model of innovative restaurants. *Tourism Management*, *36*, 15-25.

Hung, S. P., Chen, P. H., & Chen, H. C. (2012). Improving creativity performance assessment: A rater effect examination with many facet Rasch model. *Creativity Research Journal*, *24*(4), 345-357.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review*, *71*(1), 42.

Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *Neuroimage*, *84*, 1061-1069.

Illies, J. J., & Reiter-Palmon, R. (2004). The effects of type and level of personal involvement on information search and problem solving. *Journal of Applied Social Psychology*, *34*(8), 1709-1729.

Jeffries, K. K. (2017). A CAT with caveats: is the Consensual Assessment Technique a reliable measure of graphic design creativity?. *International Journal of Design Creativity and Innovation*, *5*(1-2), 16-28.

de Jesus, S. N., Rus, C. L., Lens, W., & Imaginário, S. (2013). Intrinsic motivation and creativity related to product: A meta-analysis of the studies published between 1990–2010. *Creativity Research Journal*, *25*(1), 80-84.

Kamenica, E. (2012). Behavioral economics and psychology of incentives. *Annual Review of Economics*, *4*(1), 427-452.

Kaplan, S. N., & Strömberg, P. (2000). How do venture capitalists choose investments. *Working Paper, University of Chicago*, *121*, 55-93.

Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: a meta-analysis of lens model studies. *Psychological Bulletin*, *134*(3), 404-426.

Kaufman, J. C., & Baer, J. (2012). Beyond new and appropriate: Who decides what is creative?. *Creativity Research Journal*, *24*(1), 83-91.

Kaufman, J. C., Baer, J., & Cole, J. C. (2009). Expertise, domains, and the consensual assessment technique. *The Journal of Creative Behavior*, *43*(4), 223-233.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, *20*(2), 171-178.

Kaufman, J. C., Baer, J., Cropley, D. H., Reiter-Palmon, R., & Sinnett, S. (2013). Furious activity vs. understanding: How much expertise is needed to evaluate creative work?. *Psychology of Aesthetics, Creativity, and the Arts*, *7*(4), 332.

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity. *Review of General Psychology*, *13*(1), 1.

Kaufman, J. C., Gentile, C. A., & Baer, J. (2005). Do gifted student writers and creative writing experts rate creativity the same way?. *Gifted Child Quarterly*, *49*(3), 260-265.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the consensual assessment technique: New evidence of reliability. *Thinking Skills and Creativity*, *2*(2), 96-106.

Kaufmann, E., Reips, U. D., & Wittmann, W. W. (2013). A critical meta-analysis of lens model studies in human judgment and decision-making. *PloS One*, *8*(12), e83528.

Kaufman, J.C., & Sternberg, R.J. (2010). Preface. In J.C. Kaufman, & R.J. Sternberg (Eds.) *Cambridge handbook of creativity* (pp. 13-15). New York: Cambridge University Press.

Kirlik, A. (2009). Brunswikian resources for event-perception research. *Perception*, *38*(3), 376-398.

Kozbelt, A. (2005). Factors affecting aesthetic success and improvement in creativity: A case study of the musical genres of Mozart. *Psychology of Music*, *33*(3), 235-255.

Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music*, *35*(1), 144-168.

Kozbelt, A., Beghetto, R. A., & Runco, M. A. (2010). Theories of creativity. In J.C. Kaufman, & R.J. Sternberg (Eds.) *Cambridge handbook of creativity* (pp. 20-47). New York: Cambridge University Press.

Kozbelt, A., & Durmysheva, Y. (2007). Understanding creativity judgments of invented alien creatures: The roles of invariants and other predictors. *The Journal of Creative Behavior, 41*(4), 223-248.

Kozbelt, A., & Serafin, J. (2009). Dynamic evaluation of high-and low-creativity drawings by artist and nonartist raters. *Creativity Research Journal*, *21*(4), 349-360.

Kudrowitz, B. M., & Wallace, D. (2013). Assessing the quality of ideas from prolific, early-stage product ideation. *Journal of Engineering Design, 24*(2), 120-139.

Kudrowitz, B., Te, P., & Wallace, D. (2012). The influence of sketch quality on perception of product-idea creativity. *AI EDAM, 26*(3), 267-279.

Kudryavtsev, A., & Pavlodsky, J. (2012). Description-based and experience-based decisions: individual analysis. *Judgment and Decision Making, 7*(3), 316.

Langley, P. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT press.

Lee, S., Lee, J., & Youn, C. Y. (2005). A variation of CAT for measuring creativity in business products. *The International Journal of Creativity & Problem Solving*, *15*(2), 143-153.

Leenders, R. T. A., Van Engelen, J. M., & Kratzer, J. (2003). Virtuality, communication, and new product team creativity: a social network perspective. *Journal of Engineering and Technology Management*, *20*(1), 69-92.

Lepper, M. R., & Hodell, M. (1989). Intrinsic motivation in the classroom. *Research on Motivation in Education*, *3*, 73-105.

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *The Journal of Creative Behavior*, *41*(1), 1-27.

Lim, W., & Plucker, J. A. (2001). Creativity through a lens of social responsibility: Implicit theories of creativity with Korean samples. *The Journal of Creative Behavior*, *35*(2), 115-130.

Lonergan, D. C., Scott, G. M., & Mumford, M. D. (2004). Evaluative aspects of creative thought: Effects of appraisal and revision standards. *Creativity Research Journal*, *16*(2-3), 231-246.

Long, H. (2014). An empirical review of research methodologies and methods in creativity studies (2003–2012). *Creativity Research Journal*, *26*(4), 427-438.

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity*, *15*, 13-25.

Lu, C. C., & Luh, D. B. (2012). A comparison of assessment methods and raters in product creativity. *Creativity Research Journal*, *24*(4), 331-337.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity Research Journal*, *13*(3-4), 295-308.

MacCrimmon, K. R. & Wagner, C. (1994). Stimulating Ideas through Creativity Software. *Management Science (40)*11, 1514-1532.

Magnusson, P. R., Netz, J., & Wästlund, E. (2014). Exploring holistic intuitive idea screening in the light of formal criteria. *Technovation*, *34*(5), 315-326.

Martin, W. (2006). *Theories of judgment: Psychology, logic, phenomenology.* Cambridge University Press.

Mayer, R. E. (1999). Fifty Years of Creativity Research. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 449-460). New York: Cambridge University Press.

McPherson, J. H. (1963). A proposal for establishing ultimate criteria for measuring creative output. In C.W. Taylor & F. Barron (Eds.), *Scientific creativity: Its recognition and development* (pp. 24-29). New York, NY: Wiley.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American psychologist, 50(9), 741.

Mollick, E., & Nanda, R. (2015). Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science*, *62*(6), 1533-1553.

Müller, J. S., Melwani, S., & Goncalo, J. A. (2012). The bias against creativity: Why people desire but reject creative ideas. *Psychological Science, 23*(1), 13-17.

Müller, J. S., Wakslak, C. J., & Krishnan, V. (2014). Construing creativity: The how and why of recognizing creative ideas. *Journal of Experimental Social Psychology, 51*, 81-87.

Mumford, M. D., Blair, C., Dailey, L., Leritz, L. E., & Osburn, H. K. (2006). Errors in creative thought? Cognitive biases in a complex processing activity. *The Journal of Creative Behavior, 40*(2), 75-109.

Mumford, M. D., Lonergan, D. C., & Scott, G. (2002). Evaluating creative ideas: Processes, standards, and context. *Inquiry: Critical Thinking across the Disciplines, 22*(1), 21-30.

Mumford, M.D., Reiter-Palmon, R., Redmond, M.R. (1994). Problem construction and cognition: applying problem representations in ill-defined problems. In M.A. Runco (Ed.), *Problem Finding, Problem Solving, and Creativity* (pp. 3-39). Norwood, NJ: Ablex Publishing Company.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Neisser, U., Boodoo, G., Bouchard Jr, T. J., Boykin, A. W., Brody, N., Ceci, S. J., ... & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77.

Newell, A., Shaw, C. J., & Simon, H. A. (1962). The processes of creative thinking. In H. E. Gruber, G. Terrell, & M. Wertheimer (Ed.), *Contemporary approaches to creative thinking* (pp. 63–119). New York: Lieber-Atherton Inc.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. Keane & K. J. Gilhooly (Eds.), Advances in the psychology of thinking (pp. 1–44). London: Harvester-Wheatsheaf.

O'Quin, K., & Besemer, S. P. (1989). The development, reliability, and validity of the revised creative product semantic scale. *Creativity Research Journal*, *2*(4), 267-278.

O'Quin, K., & Besemer, S. P. (1999). Creative products. In M. A. Runco & S. R. Pritzker (Eds.), *Encyclopedia of creativity* (Vol. 1, pp. 413–425). San Diego, CA: Academic Press

O'Quin, K., & Besemer, S. P. (2006). Using the Creative Product Semantic Scale as a Metric for Results-Oriented Business. *Creativity and Innovation Management*, *15*(1), 34-44.

Peelen, M. V., & Kastner, S. (2014). Attention in the real world: toward understanding its neural basis. *Trends in Cognitive Sciences*, *18*(5), 242-250.

Perkins, D. N. (1981). *The mind's best work: A new psychology of creative thinking.* Cambridge, MA: Harvard University Press.

Pétervári, J., Osman, M., & Bhattacharya, J. (2016). The role of intuition in the generation and evaluation stages of creativity. *Frontiers in psychology*, *7*.

Piffer, D. (2012). Can creativity be measured? An attempt to clarify the notion of creativity and general directions for future research. *Thinking Skills and Creativity*, *7*(3), 258-264.

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

Plucker, J. A., & Beghetto, R. A. (2004). Why creativity is domain general, why it looks domain specific, and why the distinction doesn't matter. In R. J. Sternberg, E. L.

Grigorenko, & J. L. Singer (Eds.*), Creativity: From potential to realization.* (pp. 153–168). Washington, DC: American Psychological Association.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational psychologist*, *39*(2), 83-96.

Plucker, J. A., Kaufman, J. C., Temple, J. S., & Qian, M. (2009). Do experts and novices evaluate movies the same way?. *Psychology & Marketing*, *26*(5), 470-478.

Plucker, J. A., & Renzulli, J. S. (1999). Psychometric approaches to the study of human creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 35-61). Cambridge, England: Cambridge University Press.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in Cognitive Sciences*, *10*(2), 59-63.

Policastro, E., & Gardner, H. (1999). From case studies to robust generalizations: An approach to the study of creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 213–225). Cambridge, England: Cambridge University Press.

Reinstein, D. A., & Snyder, C. M. (2005). The influence of expert reviews on consumer demand for experience goods: A case study of movie critics. *The Journal of Industrial Economics*, *53*(1), 27-51.

Reiter-Palmon, R., & Illies, J. J. (2004). Leadership and creativity: Understanding leadership from a creative problem-solving perspective. *The Leadership Quarterly*, *15*(1), 55-77.

Reitman, W. R. (1965). *Cognition and thought: an information processing approach.* Oxford, England: Wiley.

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2010). The selection of creative ideas after individual idea generation: Choosing between creativity and impact. *British Journal of Psychology, 101*(1), 47-68.

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2014). Effects of problem scope and creativity instructions on idea generation and selection. *Creativity Research Journal, 26*(2), 185-191.

Rhodes, M. (1961). An analysis of creativity. *The Phi Delta Kappan, 42*(7), 305-310.

Rubenson, D. L., & Runco, M. A. (1995). The psychoeconomic view of creative work in groups and organizations. *Creativity and Innovation Management, 4*(4), 232-241.

Runco, M. A. (2003). Commentary on personal and potentially ambiguous creativity: You can't understand the butterfly unless you (also) watch the caterpillar. *Creativity Research Journal, 15*(2-3), 137-141.

Runco, M. A. (2004). Everyone has creative potential. In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 21–30). Washington, DC: American Psychological Association.

Runco, M. A., & Chand, I. (1995). Cognition and creativity. *Educational Psychology Review, 7*(3), 243-267.

Runco, M. A., & Charles, R. E. (1993). Judgments of originality and appropriateness as predictors of creativity. *Personality and Individual Differences, 15*(5), 537-546.

Runco, M. A., Illies, J. J., & Eisenman, R. (2005). Creativity, originality, and appropriateness: What do explicit instructions tell us about their relationships?. *The Journal of Creative Behavior, 39*(2), 137-148.

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

Runco, M. A., & Johnson, D. J. (2002). Parents' and teachers' implicit theories of children's creativity: A cross-cultural perspective. *Creativity Research Journal*, *14*(3-4), 427-438.

Runco, M. A., McCarthy, K. A., & Svenson, E. (1994). Judgments of the creativity of artwork from students and professional artists. *The Journal of Psychology*, *128*(1), 23-31.

Runco, M. A., & Pezdek, K. (1984). The effect of television and radio on children's creativity. *Human Communication Research*, *11*(1), 109-120.

Runco, M. A., & Smith, W. R. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences*, *13*(3), 295-302.

Schaffer, S. (1994). Making up discovery. In Boden, M. A. (Ed.) *Dimensions of creativity*, (pp. 13-51). MIT Press, Cambridge, MA.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, *7*(4), 351-371.

Serafin, J., Kozbelt, A., Seidel, A., & Dolese, M. (2011). Dynamic evaluation of high-and low-creativity drawings by artist and nonartist raters: Replication and methodological extension. *Psychology of Aesthetics, Creativity, and the Arts*, *5*(4), 350.

Shapiro, R. J. (1970). The criterion problem. In P. E. Vernon (Ed.), *Creativity* (pp. 257–269). New York: Penguin.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas?. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(3), 139.

Silvia, P. J. (2013). Interested experts, confused novices: art expertise and the knowledge emotions. *Empirical Studies of the Arts*, *31*(1), 107-115.

Simon, H. U. (1989). Continuous reductions among combinatorial optimization problems. *Acta Informatica*, *26*(8), 771-785.

Simonton, D. K. (1991). Emergence and realization of genius: The lives and works of 120 classical composers. *Journal of Personality and Social Psychology, 61*(5), 829.

Simonton, D. K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. *Psychological Review*, *104*(1), 66.

Simonton, D. K. (1998). Gifted child, genius adult: Three life-span developmental perspectives. In R. C. Friedman & K. B. Rogers (Eds.), *Talent in context: Historical and social perspectives on giftedness* (pp. 151-175). Washington DC: American Psychological Association.

Simonton, D. K. (1999). Creativity as blind variation and selective retention: Is the creative process Darwinian?. *Psychological Inquiry*, *10*(4), 309-328.

Simonton, D. K. (2009). Varieties of (scientific) creativity: A hierarchical model of domain-specific disposition, development, and achievement. *Perspectives on Psychological Science, 4*(5), 441-452.

Simonton, D. K. (2010). Creative thought as blind-variation and selective-retention: Combinatorial models of exceptional creativity. *Physics of life reviews*, *7*(2), 156-179.

Simonton, D. K. (2012). Taking the US Patent Office criteria seriously: A quantitative three-criterion creativity definition and its implications. *Creativity Research Journal, 24*(2-3), 97-106.

Simonton, D.K. (2013). What is a creative idea? Little-c versus Big-C creativity. In J. Chan & K. Thomas (Eds.), *Handbook of research on creativity* (pp. 69–83). Cheltenham Glos, UK: Edward Elgar.

Simonton, D. K. (2016). Defining Creativity: Don't We Also Need to Define What Is Not Creative?. *The Journal of Creative Behavior*.

Smith, S. M., Ward, T. B.,& Finke, R. A. (1995). Cognitive processes in creative contexts. In S.M.Smith, T. B.Ward & R. A. Finke, (Eds.), *The creative cognition approach* (pp. 1-8). Cambridge, MA: MIT Press.

Soukhoroukova, A., Spann, M., & Skiera, B. (2012). Sourcing, filtering, and evaluating new product ideas: An empirical exploration of the performance of idea markets. *Journal of Product Innovation Management*, *29*(1), 100-112.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), 87-98.

Sternberg, R. J., Kaufman, J. C., & Pretz, J. E. (2002). *The creativity conundrum: A propulsion model of kinds of creative contributions.* Psychology Press.

Sternberg, R. J., & Lubart, T. I. (1991). An investment theory of creativity and its Development. Human development, 34(1), 1-31.

Sternberg, R. J., & Lubart, T. I. (1992). Buy low and sell high: An investment approach to creativity. *Current Directions in Psychological Science*, *1*(1), 1-5.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd: Cultivating creativity in a culture of conformity*. Free Press.

Stewart, T. R. (2001). The lens model equation. In K. R. Hammond & T. R. Stewart (Eds.), *The essential Brunswik: Beginnings, explications, applications* (pp. 357-362). Oxford, UK: University Press.

Storme, M., & Lubart, T. (2012). Conceptions of creativity and relations with judges' intelligence and personality. *The Journal of Creative Behavior*, *46*(2), 138-149.

Storme, M., Myszkowski, N., Çelik, P., & Lubart, T. (2014). Learning to judge creativity: The underlying mechanisms in creativity training for non-expert judges. *Learning and Individual Differences, 32,* 19-25.

Taylor, A. (1975). An emerging view of creative actions. In I. A. Tylor, & J. W. Getzels (Eds.), *Perspectives in creativity* (pp. 297-325). Chicago: Aldine.

Thompson, D. (2017). *Hit makers: The science of popularity in an age of distraction.* New York: Penguin Press.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch and by Hammond, Hursch, and Todd. *Psychological Review, 71,* 528–530.

Turnbull, M., Littlejohn, A. & Allan, M. (2010). Creativity in the design disciplines; learning from the practice of experts. In J. Herrington & C. Montgomerie (Eds*.), Proceedings of ED-MEDIA 2010--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 1574-1578). Toronto, Canada: Association for the Advancement of Computing in Education (AACE). Retrieved November 1, 2017 from https://www.learntechlib.org/p/34848/.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology*, *49*(2), 436.

Wallas, G. (1926). *The Art of Thought.* New York, NY: Franklin Watts.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar generation. *Cognitive Psychology*, *27*(1), 1-40.

Ward, T. B., Smith, S. M., & Finke, R. A. (1999). Creative Cognition. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 189–212). Cambridge, England: Cambridge University Press.

Weisberg, R. W. (2015). On the usefulness of "value" in the definition of creativity. *Creativity Research Journal*, *27*(2), 111-124.

White, A., Shen, F., & Smith, B. L. (2002). Judging advertising creativity using the creative product semantic scale. *The Journal of Creative Behavior*, *36*(4), 241-253.

Wiggins, G. A., & Bhattacharya, J. (2014). Mind the gap: an attempt to bridge computational and neuroscientific approaches to study creativity. *Frontiers in Human Neuroscience*, *8*.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory and Cognition, 26*, 716–730.

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychology Science*, *46*, 35-51.

Yen, J., & Sun, C. (2008). A study on the design teachers' and students' recognition of performance evaluation and evaluation criteria. *Journal of Science and Technology*, *17*(1), 41–57.

Zeng, L., Proctor, R. W., & Salvendy, G. (2011). Can traditional divergent thinking tests be trusted in measuring and predicting real-world creativity?. *Creativity Research Journal*, *23*(1), 24-37.

Zeng, L., Salvendy, G., & Zhang, M. (2009). Factor structure of web site creativity. *Computers in Human Behavior*, *25*(2), 568-577.

APPENDICES

Appendix 1

Table S1

Levels of features in the projects based on the categorical judgments of experts

Project's name	Number	Originality	Utility	Scalability	Low risk
SIP Veggie Farm	1	High	High	High	High
The Neighbourhood Museum Collective	2	High	High	High	Low
Intermodal	3	High	High	Low	High
Street Signs With Stickers	4	High	Low	High	High
miLES	5	Low	High	High	High
School of ideas	6	High	High	Low	Low
Sidewalk Chalk Arts	7	High	Low	Low	High
Time capsules	8	Low	Low	High	High
Free Neighbourhood Blocks	9	High	Low	Low	Low
Not applicable!	10	Low	Low	Low	High
Feel the city	11	High	Low	High	Low
A.R.T.S.	12	Low	High	Low	High
Pop-up Cultural Hub	13	Low	High	High	Low
Fablab	14	Low	High	Low	Low
Movies to the park	15	Low	Low	High	Low
Blogger Café	16	Low	Low	Low	Low

Appendix 2

A sample project proposal

Project #5:

miLES

Founded by Eric Ho, miLES (made in the Lower East Side) seeks a better way to utilize underused storefronts and "turn them into vibrant community hubs for working, learning, connecting, and starting up new projects". The platform is notorious, works just like Airbnb and Zipcar: people can rent out a service from an online database for temporary use.

miLES, likewise to its competition, opens up storefronts to new possibilities by changing short-term multi-use spaces into community hubs (such as a studio, classroom, cinema, library, etc.). There are more than 200 vacant storefronts in the Lower East Side and this ratio is typical for other cities, too. The project is trusted and funded by both local committees and successful crowdfunding campaigns. The service is getting more and more popular and is looking at a bright future.

All stimuli can be downloaded from the Open Science Framework:

https://osf.io/r3xch/?view_only=5a78847982ad434ea66167cbb8c9b861

LIST OF PUBLICATIONS

Peer-reviewed articles

Pétervári, J., Osman, M., & Bhattacharya, J. (2016). The role of intuition in the generation and evaluation stages of creativity. *Frontiers in Psychology*, 7, 1420.

Pétervári, J., Osman, M., & Bhattacharya, J. (2017). *Useful above all – Predicting how experts and non-experts evaluate creativity.* Manuscript submitted for publication.

Outreach activity

Pétervári, J. [Judit Pétervári]. (2017, September 29). Dance your PhD 2017 - Building up creativity [Video file]. Retrieved from https://www.youtube.com/watch?v=wXzPH3iQV8E