The Co-evolution of Theory and Practice in Design Thinking – or – "Mind the Oddness Trap!"

Julia von Thienen, Christine Noweski, Christoph Meinel*, and Ingo Rauth

Abstract In Design Thinking, theory and practice are closely interconnected. The theory serves as a blueprint, guiding companies in general and design teams in particular through the design process. Given such a close interrelation of theory and practice, we argue that Design Thinking research needs to be set up in a particular way too. This setup ties in with Design Thinking process models: To attain ever more befitting design solutions, prototypes are supposed to be tested and refined. Correspondingly, Design Thinking research should help to test and refine theory elements of Design Thinking. Researchers may serve as "dialogue facilitators," aiding the community of Design Thinkers to intensify their "dialogue" with empirical reality.

To provide reliable data on issues of central concern, we have tested experimentally two widely held convictions in the field of Design Thinking: (1) Multidisciplinary teams produce more innovate design solutions than monodisciplinary teams. (2) Teams trained in Design Thinking (by the D-School) produce more innovative solutions than untrained teams. In addition, degrees of communication problems were assessed. While both "multidisciplinarity" and "D-School training" have been associated with more unusual design solutions, with respect to utility a different picture emerged. Thus, hotspots have been identified that may stimulate some productive refinements of Design Thinking theory.

1 From Design Thinking to Design Thinking Research

How should teams approach design challenges? What do students need to learn to tackle design challenges successfully? With increasing frequency, *Design Thinking* is called upon to help answer these questions. Used by multiple big companies such as SAP, P&G, IDEO or GE Healthcare, accompanied by a lot of media attention and

J. von Thienen (✉), C. Noweski, C. Meinel, and I. Rauth
Hasso-Plattner-Institute, Campus Griebnitzsee, P.O. Box 900460, 14440 Potsdam, Germany
e-mail: e.valuate@hpi.uni-potsdam.de; meinel@hpi.uni-potsdam.de

* Principal Investigator

H. Plattner et al. (eds.), *Design Thinking: Understand – Improve – Apply*,
Understanding Innovation, DOI 10.1007/978-3-642-13757-0_5,

propelled by an increasing number of training institutions, Design Thinking seems on its way to become the state-of-the-art innovation method. And yet, we understand only little about what really matters for it to be successful.

In the armchair we may think about these issues, but many crucial questions will remain unanswered. Those who truly wish to know will have to confront the real world: Careful empirical analyses are in place! With this thought in mind, we decided to make a real job of it – and put fundamental assumptions of Design Thinking to an experimental test.

Naturally, in the booming, buzzing field of Design Thinking there are innumerable aspects that warrant careful scientific investigations. Of course, one might just cherry-pick some questions, selecting the issues according to personal interests. Yet, the research ought to take into account the interests of people working in the field as well, or shouldnt it? So, we made it our first empirical research task to scan in a somewhat broader fashion the interests, hopes and worries of experts in the field. But sure enough, there was some trouble ahead: While the term "Design Thinking" seems to allude to a common set of practices and a common theoretical matrix, the experts held ready an astonishing variety of understandings. What does that imply for the task of testing empirically central assumptions of Design Thinking theory? Our answer will be an outlook on research endeavours particularly designed to match the characteristic relation of theory and practice in Design Thinking. It will be the basis we start from and return to in our experimental work.

2 Experts Revealing What They Think About Design Thinking

In the winter of 2009, we had the opportunity to speak to a number of Design Thinking experts and conducted a series of guideline interviews of about $1\frac{1}{2}$ h each. In this context, we wish to thank once more members of IDEO, the Design Services Team of SAP, design consultants from Procter & Gamble and Palm as well as members of the staff and teachers of the Design Schools in Potsdam and Stanford. The interviews focussed on three major issues:

1. The definition and understanding of Design Thinking (the process and its methods) as well as prototypical conflicts in Design Thinking projects
2. needs regarding the work environment and tools
3. successful team orchestration and its specific needs

Key insights were synthesized using storytelling and clustering techniques within the project team. Papers have, or will be published on each of the topics. Here, we shall briefly review those issues that helped to shape our further approach within the *HPI* research program.

What stroke us as most momentous for the whole enterprise of Design Thinking research was the grand variety of understandings across experts in the field: The interviewees did not convey a common understanding of Design Thinking. They specified differing process models and named differing methods as crucial elements of the design process.

We found, for example, opposite beliefs regarding the question whether design work should be outsourced or not. According to some experts, design teams need to work outside of common business contexts to avoid being "captured" in their routines. These experts argue that creative freedom needs to be maximized. Ideally, the development of new design ideas should therefore be outsourced. Other leading experts prefer integrative approaches where managers set up teams by bringing together employees from different departments. This way, a single team may attend a project from the earliest up to the latest stages. While different departments are responsible for different steps in the design process (e.g., idea generation versus final implementation), representatives of all departments are joined in the responsible design team right from the start.

To mention another point of divergence, some experts highlight the pivotal importance of individual genius. Others believe, however, that individual genius is comparably unimportant when it comes to predicting the success of a design project. Instead, they say, teams need to be assembled according to sophisticated theories so as to combine particularly "matching" characters and competences.

Interestingly, the experts did not only differ in the concrete approaches they preferred. They explained their understanding of Design Thinking on different scales and reflected upon differing academic discourses. Obviously, there is no common set of beliefs (yet) associated with Design Thinking. Rather, there are differing lines of debate as well as differing practices. To what extent we should strive to bring them together is an interesting question by itself.

Apart from considerable differences in the general understanding of Design Thinking, there were – fortunately! – a number of important commonalities too. Without any such visible connecting factors it would be hard to see how Design Thinking could be studied as a collective enterprise.

A strong focus on **user needs** is considered essential across the board and the aim of true **innovation** is a shared concern. Design teams should not just head for quantitative improvements (such as devising a memory stick with yet more storage capacity, applying well known technologies). They should also be able to bring about qualitative improvements (e.g., by devising new technologies that are more potent or by developing solutions that make memory sticks superfluous altogether). That is, design teams should reconsider initial design challenges ("reframing"): They should try to understand what the users' true needs are. Then, they should consider a whole variety of approaches, including (and quite essentially so) uncommon ones, the so called "wild ideas." In a continuous dialogue with the users, a solution shall finally be worked out that suits the users' needs particularly well.

Another aspect that many Design Thinkers view as central is the academic diversity of design teams. Commonly, **multidisciplinarity** is considered a good choice. Teams are supposed to be academically diverse so that they may integrate impulses from many different domains. It is assumed that multidisciplinarity is particularly well-suited to foster true innovation.

Next to multidisciplinarity, other factors are thought of as crucial for team performance too. In particular, many interviewees stressed the importance of a positive **communication culture**.

In sum, the experts named a number of common concerns. But, strikingly, they did not sketch out a common theoretical matrix associated with the term "Design Thinking." This is a finding that should occupy us! Given the cloudy theory structure of Design Thinking, what are we to expect of Design Thinking *research*?

3 Telling Differences, Illuminating Parallels

Traditionally, theories are considered to be systems of axioms: There are a couple of fundamental propositions from which everything about the field of interest may be deduced. When a scientist refers to "the theory," he refers to its set of axioms. Correspondingly, accepting a theory means to accept "the axioms." With this classical picture in mind, there seems to be something quite worrisome about Design Thinking. If it is a theory – or builds on a theory – where are its axioms? As became all too clear in the expert interviews, there is no common set of propositions that Design Thinkers accept in virtue of their expertise. There are *some* shared convictions that may be understood as guiding theoretical ideas. But, they certainly do not cover the whole domain of interest. Apart from that, rather than there being fundamental assumptions, there are shared *centres of concern*: Usability, multidisciplinarity, unusualness ("go for the wild"), reframing of original tasks – to name some in a random order. Experts occasionally disagree as to how important each issue is in differing project phases. But they routinely monitor and discuss them. Now, what does the lack of a classical theory-structure mean for Design Thinking? Is it non-professional after all? Is it in such an early stage of its development that it has not even managed to produce a meagre axiomatic system?

We, in contrast, believe the "axiomatic system" is a misguided ideal for Design Thinking. There are good reasons for the open theory-structures that characterize Design Thinking today. These open structures are sensible, but nonetheless they may – of course – be improved. To see how the structures make sense and what likely aims there may be for improvements, it seems a good idea to scan the academic field for domains with similar challenges.

Musicology, for instance, does have some interesting parallels to Design Thinking. First of all, its subject is something *productive* and *creative*: Musicologists study pieces of music and their composition just like academic Design Thinkers study design solutions and their coming about.

When looking at – say – pop songs, music theory serves a dual function. On the one hand, it *describes* songs. On the other hand, by working out and comparing song patterns the theory provides a *blueprint* how songs may be composed (successfully). For example, there typically is an intro, then strophes and the chorus alternate, there are bridges, breaks and, finally, an ending. Longer instrumental interludes are typically placed in the second half of a song, not the first.

Yet, such a scheme is not enough for a song. Individual musicians have to fill in the blanks. Novices in particular may profit from following strictly the blueprint they are given. But experts (or: visionaries) may produce masterpieces by breaking

the rules. Some of the time, they thus establish new patterns that other musicians will use fruitfully in the future.

In Design Thinking, things are not all that different. Design Thinking theory serves a dual function as well. It helps to describe and analyse design projects (e.g., does reframing happen at some point? What does the team do and when to ensure usability?). Design process models convey standards as to which phases there are and in which order they be put. They also encompass methods that may be invoked.

When a design team orchestrates its own project, it may well profit from given schemes. But sure enough there are blanks to fill in. (For instance, "Here we are in the research phase. We have methods *A* through *H* at our hands. Which shall we pick? How exactly shall we proceed?")

As Design Thinkers grow more and more experienced, they may identify circumstances in which unconventional procedures seem more promising than standard ones. Out they move of common schemes. They break the rules! If this happens, it is an interesting case for Design Thinking theory. Such a "breaking of rules" should not be generally damned. It is a precious test case. Maybe it fails. But if it doesn't, Design Thinking theory hits on an alternative whose potential is yet to be explored.

The parallels between musicology and Design Thinking illuminate two important issues that we need to keep in mind to avoid working towards an inadequate theoretical ideal.

The co-evolution of theory and practice. According to the classical understanding, a theory is true if it describes the empirical world correctly. An unbridgeable gap separates theory and world. Changing the theory will not change the world.

In the case of Design Thinking, as in the case of musicology, the gap is being crossed all the time. Since the theory provides blueprints to practitioners, a change in the theory is likely to change the empirical world itself. Theory and practice co-evolve. In consequence, the question of whether or not Design Thinking theory is true does not "function" in a conventional way. In many respects, Design Thinking theory may be true for trivial reasons: Because it serves as a scheme according to which practitioners proceed. Truth is cheep to have for Design Thinking theory in these regards. And truth does not suffice.

Consider the two claims:

(a)	The theory is true.	True or false?
(b)	The theory is (most) useful.	True or false?

Conventionally, scientists ask whether claim (a) is maintainable. In the case of Design Thinking, claim (b) seems to be the more fundamental, the more demanding. It is the one whose correctness calls for rigorous empirical investigations.

Since theory and practice *are meant* to co-evolve, empirical evidence for a lack of utility will not (and should not) lead to the rejection of claim (b). Instead, careful analyses need to follow. Design Thinking theory – in particular: aspects of its process model – may have to be modified to become ever more useful.

The researcher as a dialogue facilitator. What is the second issue we may – and should – learn from the parallels between musicology and Design Thinking? In our understanding, one more point is particularly important for a proper setting of

goals. The example of musicology teaches us how fruitful it can be to have both at the same time: An overall-open theory structure that may seem cloudy – yet a rigorous precision in analytic conceptions.

On the one hand, it is clear that there are many ways to produce felicitous pieces of music; and there are different music styles that may be just as appealing. In this sense, it would be detrimental if musicology would specify one single theoretical matrix according to which music ought to be produced. Musicological theory needs to be open; it needs to be able to handle plurality and to incorporate new developments that the future will (hopefully) bring. This openness in theory structure does not, however, imply that it is necessary or helpful to work with cloudy concepts and claims. For example, think of notes and rhythms that do a marvellous job in documenting and structuring something as elusive as played music! (Do you think you could come up with just *two concepts* such that whole design projects could be reconstructed on their basis? If you have some spare time, maybe sitting in a bus or plain, why not give it a try?)

The aim of potent and precise analytic conceptions – despite of an overall open theory structure – is, we think, an excellent target for Design Thinking as well. While it is clear that Design Thinking theory needs to remain open to allow for new developments, we should still strive to refine our analytical conceptions so that they be ever more potent systematizing factors. We should also try to learn more about our individual versus collective claims – and how well they are substantiated.

With this background understanding, we feel that some rather peculiar role befits us, the researchers. We wish to serve as dialogue facilitators: We wish to help Design Thinkers enter in an intense dialogue with empirical reality. What concepts, what assumptions work well, which do not work all that well yet? The research ought to put Design Thinkers in a position to sharpen their vocabulary and their fundamental beliefs in a way that makes them ever-more adapt to reality, ever more fruitful.

4 Preparing a Look Behind the Curtain: Specifying Hypotheses

As there is no written out axiomatic system in Design Thinking that specifies crucial assumptions one after the other, it is the researchers' first job to pin down crucial beliefs in the field. Our take in the last year was this: In general, it is assumed that Design Thinking fosters innovation. After all, Design Thinking is supposed to be an innovation method (or even: *the* state-of-the art innovation method). So, people who have been trained in Design Thinking should produce more innovative solutions than people who have not been thus trained.

Of course, there are multiple institutes who offer Design Thinking education. As the *Design Thinking Research Program* in Potsdam and Stanford enjoys a close cooperation with the D-Schools in Potsdam and Stanford, the Design Thinking education we shall look at will be a D-School training. Our starting hypothesis may thus be formulated more specifically: It is assumed that D-School trained teams produce more innovative solutions than teams without this training. Additionally, to consider one rather confined factor, we shall test the widespread belief that

multidisciplinarity enhances innovation. If the belief is correct, multidisciplinary teams produce more innovative solutions than monodisciplinary ones on average.

While the two hypotheses concerning D-School training and multidisciplinarity are viable starting points, they need to be further refined. In particular, "innovation" is such an abstract notion that it is too remote from potential measurement operations. In such a case, it is usually a good idea to break the abstract concept down into disparate factors that may be assessed more easily. This is our take:

A design solution S_1 is considered more innovative than a solution S_2 if S_1 is more unusual as well as more useful than S_2.

Given this clarification of what "innovative" means, both of the starting hypotheses split into two more specific claims. These are the assumptions regarding D-School education:

1. D-School trained teams produce *more unusual* solutions than teams without this training.
2. D-School trained teams produce *more useful* solutions than teams without this training.

Accordingly, two hypotheses may be formulated concerning multidisciplinarity:

3. Multidisciplinary teams produce *more unusual* solutions than monodisciplinary teams.
4. Multidisciplinary teams produce *more useful* solutions than monodisciplinary teams.

While there are ample reasons to believe that multidisciplinary teams will indeed produce more innovative solutions than monodisciplinary ones on average, there is at least one notable reason to believe the opposite – and it may be fruitful to consider these reasons distinctly.

Experts who have been trained in the very same way of analyzing and approaching a subject matter are likely to invoke the strategies they are all used to when working on a new problem. Whatever work strategies are being used, by and large they pave the way for some particular type of result while detracting from other options. For example, imagine a team of chemists and a team of classical philologists who are to analyze a painting. While the chemists might take tiny samples of the paint and find out which material components have been used, the philologists might identify a scene from Greek mythology and reason backwards to the exact literary sources the painter had been exposed to. Given the specialized knowledge and training of the experts, there seems no way that the philologists could hit on the work results that chemists get and vice versa. Limiting oneself to a fixed set of (common) work strategies usually means limiting oneself to particular types of (common) results. In multidisciplinary teams, however, the approaches that team members are familiar with are likely to differ. Thus, there will be no immediate way of setting about the task. Rather, team members will have to (re-)consider the approaches they find convenient. In bargaining how to move on, they will have to detach themselves from common practices – melding, merging, blending the strategies they know in

a way that seems appropriate in the context of their current challenge. The broader the domain of strategies experts are willing to consider, the broader is the domain of results that their team may obtain. Insofar as new approaches are tried, the odds increase that something rather unusual results. Thus, it seems likely that multidisciplinary teams produce more unusual results than monodisciplinary teams.

Regarding the second facet of innovation – usefulness – multidisciplinarity may be all the more advantageous. After all, the development of useful solutions depends upon knowledge, e.g., knowledge concerning the situation of users or knowledge about technical options for realizing some particular idea. Imagine experts who are equally well trained. Clearly, if they are all trained in the very same domain, the knowledge their team disposes of is rather limited compared to the knowledge of a team whose members differ in their fields of expertise. Thus, multidisciplinary teams seem better equipped for developing useful solutions.

Yet, at the same time, there is a reason to believe that, on average, multidisciplinary teams will produce less innovative solutions than monodisciplinary ones. Why that? Even if multidisciplinary teams have a greater potential for innovation, communication problems might hinder them. It seems reasonable to expect that communication will be more challenging in multidisciplinary than in monodisciplinary teams. Just as people with differing academic backgrounds have been trained to use different strategies when approaching a problem, they have also been trained to use different concepts. The words they use may differ, the categories by which they sort things in the world may differ and the implications associated with one or the other categorization may differ as well. If design teams are unable to work out a common conceptual ground, they may not be able to make good use of the wide-ranging expertise of their team members. Thus, we decided to consider a fifth hypothesis that may shed some light on important team processes in the design process:

5. Multidisciplinary teams experience more communication problems than monodisciplinary teams.

At the same time, D-School training might well make a difference with respect to communication success. D-School trained team members might – or rather: they should – be able to handle potential communication problems, whether or not working multidisciplinarily. After all, it is assumed that they are particularly apt for design work. Thus, they must not be thwarted or halted by potential communication obstacles. A sixth and final hypothesis is therefore:

6. D-School trained teams experience less communication problems than teams without this training.

5 Why Experiments Matter

As preliminary considerations have been formulated, a choice needs to be made as to how the subject matter shall be tackled empirically. In principle, two alternatives are available. Investigations can be experimental or non-experimental. Both

approaches have their advantages as well as their disadvantages. The experimental method has been devised to fade out or "oppress" all the factors potentially relevant to an outcome except for those factors whose influences are to be investigated (as specified by the hypotheses). Thereby, the relationship between the factors that one takes interest in becomes maximally clear. But, naturally, one doesn't find out anything about the other factors (not addressed by the hypotheses) that one is at such pains to fade out in the experimental setting. In non-experimental studies, on the other hand, one may explore all the facets of real-life situations in their full booming buzzing mix-up. Thus, you may come to consider aspects you would never have thought about in your office armchair, extrapolating from the data hypotheses as to how they *might* be interrelated. Yet, whether these putative causal relations truly exist, one cannot really tell.

In our case, factors have been selected that are of primary interest. The crucial question is whether or not they are causally related. If D-School training and multidisciplinary actually do enhance innovation (as is hypothesized), a hook-up question may be how strong their effect is. These are questions to which experiments alone provide thoroughly compelling answers.

6 The Challenge

In every experiment, the setup requires thorough considerations as it sets the upper limit of what can be found out. In our case, a challenge needs to be formulated concerning a topic that all participants are about equally familiar or unfamiliar with. Otherwise, some teams might dispose over a lot of knowledge regarding the subject matter right from the start as some members would be experts, while other teams would have laypersons only. Regardless of whether one believes that teams profit from an expert (due to their knowledge) or whether one considers experts as a threat to innovation (because they might act as rigorous sensors), the teams with versus without experts would not be working under comparable conditions. Let's assume that, in the end, the presented solutions actually differ in their quality. These differences could not be clearly attributed to the factors of multidisciplinarity versus monodisciplinarity or D-School training versus no such training if the teams had differed in other respects as well, such as expert knowledge versus no such knowledge.

In addition, the scope of the challenge should be somewhat grand, or at least not minute. It should be "open" enough so that it would be possible to come up with a technical or a social solution or an artistic or political or yet other type of solution. A related demand is that there should be the possibility of using knowledge from diverse fields. If, on the other hand, only people with one particular academic training could complete the task (e.g., implement a certain computer algorithm), this would probably forestall successful Design Thinking right from the start.

The challenge that was chosen to meet these needs was this: Come up with something that helps traumatized people to manage their everyday lives!

Indeed, the participants of our experiment (40 students) indicated that their pre-experience with the subject matter, trauma, was basically negligible. For example, no one had ever been a practitioner in the field or had had a considerable training in the domain. Only one student had ever encountered the subject matter in her university studies.

7 Operationalization or: Let's Be Concrete!

Now that a challenge has been specified the question of how to asses, how to "measure" the attributes of interest needs to be considered. Each team will present its suggestion for how to help traumatized people. What is to be done so that reliable measures result, i.e. estimates of the unusualness of each solution?

When invoking numbers in every day life, we often ask questions about concrete things. For example, how many eggs are left in the fridge? In cases like these, we may start counting right away. In our study, on the other hand, the factors of interest are quite abstract. This does make a difference for the procedure of assessing or "measuring" those factors. How is one to count the unusualness of a design solution, for instance? Obviously, some further steps need to be taken.

In order to assess abstract factors they need to be *operationalized*. The question to be pondered is this: Given the context of your particular study, what could you observe straightforwardly to find out about the factor(s) of interest? Your task is to find concrete entities that one can look at to arrive at reasonable statements about the abstract notions of interest.

In the setup of an experiment, the operationalization is a crucial step. If one's operationalization is unconvincing, one's data will fail to bear on the issue that one sets out to investigate! Thus, in the case of our experiment as well as in general, we want to invite you to take a very careful look at the operationalizations: What do people (we) actually observe when they (we) make claims about highly abstract matters? Is the step they (we) take from observed entities to theoretical entities actually warranted? In our case, on the level of theory there are five factors of interest: (1) D-School training, (2) academic diversity, (3) the unusualness of design solutions, (4) the usefulness of design solutions and (5) communication problems.

While the factors (3)–(5) truly call for discussion, for reasons of completeness we shall mention the first two as well. There was a very convenient way of assessing the academic background of participants: We basically asked them. In the case of Design Thinking experience we consulted official lists of D-School trainees and alumni.

What is "unusual"? While the "unusualness of a design solution" is too abstract to be looked at and counted directly, we may ask people questions and attain concrete answers, counting how many times particular replies are given. To arrive at a pertinent question, the following consideration seems reasonable: In the context of our experiment, a group presents an unusual solution if the other teams (who have worked on the same challenge, after all) failed to consider that particular possibility when discussing options for helping.

In the course of the experiment, every team has to present its solution. All the participants need to fill out a questionnaire including the following question – regarding each single presentation (of the other groups):

Item 1 Has the presented solution been discussed in your group as well?

- ☐ Yes, exactly in this form (1)
- ☐ Yes, in about that way (2)
- ☐ More or less (3)
- ☐ No, but that may have been a coincidence (4)
- ☐ No, we would never have hit on it (5)

The brackets show our coding. Thus, the statistical values obtained range from 1 to 5. Greater values indicate a greater degree of unusualness.

Of course, the participants of our study are not the only people to ever think about how one could help in the case of traumatisation. There are experts in the field, trauma therapists in particular, whose job it is to help traumatized people. In addition, there are people who have suffered a traumatisation, of course. They too may have thought about options for improving their situation. Accordingly, these experts shall be contacted, introduced to one design solution after the other and asked a question quite similar to *item 1*:

Item 2 Have you ever considered this option for helping before?

- ☐ Yes, exactly in this form (1)
- ☐ Yes, in about that way (2)
- ☐ More or less (3)
- ☐ No, but that may have been a coincidence (4)
- ☐ No, I would never have hit on it (5)

Again, values range from 1 to 5. Greater values indicate a greater degree of unusualness.

What is "useful"? While the design teams may contribute information regarding the unusualness of a design solution, they are hardly in a position to specify utility. Of course, members of design teams can say something about *what they think* how useful their solution is (and we did ask them this question). Yet, whether or not a tool is actually helpful is not decided by the developers but by the users. In our context, the users are traumatized people or therapists who work with traumatized people. (Many teams actually developed tools that would aid the therapists in helping their clients.)

To attain judgements of how useful each solution is experts have been asked the following question:

Item 3 What do you think, how helpful is this approach for the target group?

- ☐ Very helpful (5)
- ☐ Quite helpful (4)
- ☐ Somewhat helpful (3)
- ☐ Barely helpful (2)
- ☐ Not helpful (1)

Again, values range from 1 to 5. Greater values indicate a greater degree of usefulness.

When working with operationalizations, disposing over a second estimate for each factor of interest is commonly quite advantageous. It helps you check whether the numbers you attain actually represent what they are supposed to. If two different indicators of the very same factor point in the same direction this gives you some (further) evidence for their working properly. If, on the other hand, indicators for the same subject matter point in different directions, this is ample evidence for there being something wrong with your assessment procedure(s). Thus, a second item was formulated that ought to cap onto the factor "usefulness."

Item 4 Which approaches should be realized by all means?

Please mark up to five approaches!

Marked (1)
Not marked (0)

Again, the brackets show our coding. Values range from 0 to 1. Greater values indicate a greater degree of usefulness.

How to assess "communication problems"? Communication problems, of course, would have to be estimated by the team members and not by the experts (who were contacted after the experiment). At the end of the workshop, the participants were asked to fill out a questionnaire containing three items to assess potential communication problems.

Item 5 Was it easy or difficult for your group to reach an agreement?

☐ Very easy (1)
☐ Easy (2)
☐ Neither nor (3)
☐ Difficult (4)
☐ Very difficult (5)

Item 6 Have there been group decisions that you felt uncomfortable with?

☐ Not at all (1)
☐ Very few (2)
☐ Some (3)
☐ Several (4)
☐ Plenty (5)

Item 7 Have there been communication problems in your team?

☐ Not ever (1)
☐ Rarely (2)
☐ Sometimes (3)
☐ Often (4)
☐ Very often (5)

Table 1 The constructs of interest and their operationalization

	Variations (Independent variables)		Outcome (Dependent variables)		
Level of theory					
Of interest	Team setup		Innovation		Communication
	D-School training	Academic diversity	Unusualness of solution	Usefulness of solution	Problems
Level of observation (operationalization)					
Who rated			Experts and teams	Experts	Teams
Observable	Statements, list	Statements	Item 1 (team) Item 2 (experts)	Item 3 (aid) Item 4 (choice)	Item 5 (agreement) Item 6 (decisions) Item 7 (problems)

In all three cases, values range from 1 to 5. Greater values are taken to indicate more communication problems.

Table 1 summarizes the variables of interest in the experiment and how the constructs have been operationalized.

Once the blueprint has been worked out and all the necessary provisions have been made, the experiment may begin. This is what happened:

8 Looking Behind the Curtain: The Experiment

The experiment spanned over five full days. It took place at the D-School on the Potsdam campus. The participants had to be present for the whole time, beginning from 9.30 each morning; on some days there were teams still working as late as midnight.

The project had been announced both as a "workshop on trauma" as well as an "experiment." It was made clear on all placards that the project was part of an experimental research program. Thus, the activities of participants would be observed and documented. At the same time, the program to be followed throughout the five days resembled that of a workshop. Participants would be supplied with information regarding trauma and had the task of developing some helpful approach.

40 students participated in the study, 15 men and 25 women. About half of the students had a technical background (software systems engineering). The background of the other students varied widely. Majors included business studies, languages, sports and others. On average, the participants were 22.71 years old and studied in the 4.82 semester. Half of the participants had been trained by the D-School, half of them not. We randomly assigned them to the mono- versus multidisciplinary team condition, making sure that there would be the same number of teams in each condition. Ideally, there should be three teams (of four members each) in all the four conditions:

1. D-School trained, multidisciplinary
2. D-School trained, monodisciplinary

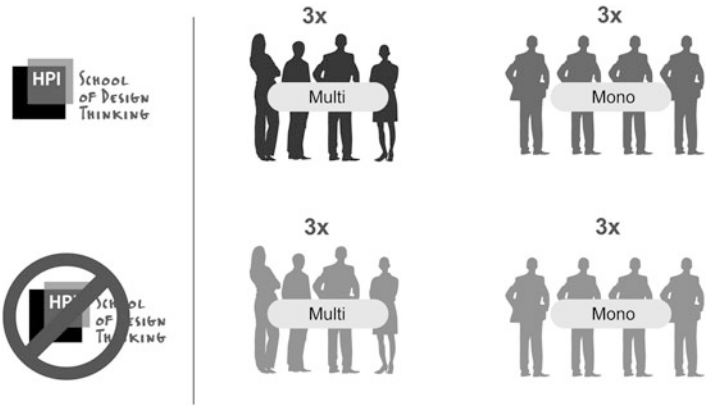

Fig. 1 The experimental setup allots three D-School trained multidisciplinary teams, three D-School trained monodisciplinary teams, three multidisciplinary teams without D-School training and three monodisciplinary teams without D-School training

3. Not-D-School trained, multidisciplinary
4. Not-D-School trained, monodisciplinary

Due to illnesses, there were some minor variations in the number of participants.

On each day of the experiment, multiple observations were made over and above those already specified. The participants filled out questionnaires regarding diverse issues such as their plan for proceeding, their satisfaction with their current standing, how they spent their time etc. A random sample of teams was filmed throughout the entire week, insofar as they were present at the D-School. Pictures were taken of all workspaces. The final presentations of all groups (approximately 10 min) were video-recorded. These video presentations as well as written summaries of the design solutions (1–2 pages) were made available online.

In the context of a lecture, the material was presented to trauma therapists and clients who had agreed to evaluate the solutions. The participants of the workshop/experiment were not present at that lecture so that personal sympathies or animosities would not bias the expert judgements (Fig. 1).

9 Design Thinkers Versus "Ordinary Students": Results

Of the two aspects of **innovation** that have been distinguished, lets consider **unusualness** first. D-School teams receive higher ratings than Non-D-School teams, as was hypothesized. The finding is consistent across experts and team members. Experts rate the unusualness of solutions by D-School teams with 2.80 on average; solutions by untrained teams 2.54. (Higher ratings indicate a greater degree of unusualness.) The participants themselves rate solutions by D-School teams 4.06 on average, solutions by other teams 3.65. The average unusualness ratings of experts

Table 2 Results regarding "usefulness" as estimated by the experts, comparing D-School trained teams with untrained teams

Question on usefulness	D-School	N	Mean	Mean diff.	p
What do you think, how helpful is this approach for the target group? (Experts, 1–5)	Trained	20	3.60	0.65	<0.05
	Untrained	24	4.25		
Which approaches should be realized absolutely? Please mark up to five approaches! (Experts, 0 or 1)	Trained	20	0.25	0.258	n.s.
	Untrained	24	0.42		

versus participants differ quite considerably in their height: Experts generally give lower ratings than participants. Thus, experts seem to have tapped the domain of potentially helpful interventions more completely than the project teams. Yet, the data consistently favors D-School teams in terms of unusualness.

Regarding the second facet of innovation, **usefulness**, all teams perform quite well. In none of the experimental conditions the average rating falls below "3," indicative of a "somewhat helpful" solution.

Just like the two measures of unusualness yield a consistent picture, the two measures of usefulness are consistent with one another too. However, the picture they suggest deviates from what had been expected. Not only does the data fail to show a significant superiority of D-School solutions. Indeed, Non-D-School teams outplay teams with D-School experience.

In Table 2, the column "N" specifies the number of ratings upon which the group averages are calculated. The column "p" specifies whether or not the difference between trained versus untrained teams is statistically significant. "N.s." means not significant, "<0.5" means significant and "<0.01" means highly significant.

Teams without D-School training receive higher ratings (4.25) on average than D-School trained teams (3.6). Higher values indicate a greater degree of usefulness; values may range between 1 and 5. The second measure of utility – whether or not a solution is chosen by the experts to be implemented "by all means" – points in the same direction. Solutions presented by teams without D-School training are selected more often (0.42) than solutions by D-School trained teams (0.25). Again, higher values indicate a greater utility; values may range between 0 and 1.

Now that we have considered trained versus untrained teams, lets take a look at the **mono-** versus **multidisciplinary** team condition.

Of all the groups, multidisciplinary D-School teams perform worst. Their average rating is close to 3 (somewhat helpful), whereas teams of all the other conditions receive an average rating above 4 (quite helpful) by the experts. Monodisciplinary teams outperform multidisciplinary teams, both in the D-School and in the Non-D-School condition.

Please note that statistical calculations for levels of significance depend not only on the size of the effect (here: the actual group difference) but also on the number of ratings. Thus, it is always a good idea to look at effect sizes over and above levels of significance. In Table 3, the average difference between mono- and multidisciplinary groups is greatest for D-School trained teams alone (first row in Table 3). It amounts to 1.083 as opposed to 0.167 for untrained teams (second row) or 0.633 for all teams

Table 3 Results regarding "usefulness" as estimated by the experts, comparing mono- versus multidisciplinary teams

	Teams	N	Mean	Mean diff.	p
D-School trained	Mono	8	4.25	1.083	0.05
	Multi	12	3.17		
Not D-School trained	Mono	12	4.33	0.167	n.s.
	Multi	12	4.17		
All teams	Mono	20	4.3	0.633	<0.05
	Multi	24	3.67		

together (third row). Yet, since the number of cases is halved when D-School teams are considered alone, the level of statistical significance is actually lower in the first row (for D-School teams only) than in the third row (where all the teams are considered).

Now, an interesting hook-up question may be whether there is some interrelation between unusualness and usefulness: Knowing that a solution is rather unusual (or usual), can you predict to some extent how useful the solution is? Or, vice versa, knowing that a solution is rather useful (or barely helpful), can you predict to some extent whether it is a rather unusual (or usual) solution?

Indeed, this is possible! The correlation between "unusualness" and "usefulness" is highly significant. It is negative: -0.547 ($p<.001$). This means, that the more unusual solutions are, the less they are helpful on average. (Correlations vary between -1 and 1. A value of zero indicates that there is no interrelation. A value of 1 indicates a perfect positive relation. A value of -1 indicates a perfect negative relation, that is: the higher the value of the first variable, the lower the value of the second and vice versa.) When only D-School teams are considered, the negative correlation between unusualness and usefulness becomes even more drastic: -0.700 ($p < 0.001$). This is an issue we will return to in the discussion.

Regarding **communication problems**, there is no statistically significant difference between mono- versus multidisciplinary teams; the effect sizes are negligible.

There is, however, a consistent difference between D-School trained teams versus untrained teams. According to all three indicators (items 5, 6 and 7), untrained teams experience more communication problems than teams with D-School training. This holds true both in the monodisciplinary as well as in the multidisciplinary team condition.

Teams without D-School training find it significantly more difficult to reach agreements (2.89 as opposed to 2.13). Members of not-trained teams report more group decisions they felt uncomfortable with (2.42 versus 1.88). Members of not-trained teams report more communication problems than members of D-School teams (2.53 as opposed to 1.88) (Table 4).

While some of the group differences fail to be statistically significant due to small N, it is noteworthy how consistent the picture is even when the mono- and multidisciplinary team condition are considered separately: All six comparisons indicate less communication problems in D-School teams (Table 5).

Table 4 Results regarding "communication problems", comparing D-School teams versus Non-D-School teams

Questions on communication problems	D-School	N	Mean	Mean diff.	p
Was it easy or difficult for your group to reach an agreement? (Item 5, teams, 1–5)	Trained	16	2.13	−0.77	<0.05
	Untrained	19	2.89		
Have there been group decisions that you felt uncomfortable with? (Item 6, teams, 1–5)	Trained	16	1.88	−0.546	n.s.
	Untrained	19	2.42		
Have there been communication problems in your team? (Item 7, teams, 1–5)	Trained	19	1.88	−0.651	<0.01
	Untrained	19	2.53		

Table 5 Results regarding "communication problems," comparing D-School teams with Non-D-School teams, multi- and monodisciplinary teams separately

		D-School	N	Mean	Mean diff.	p
Multi	Item 5	Trained	10	2.50	−0.600	n.s.
		Untrained	10	3.10		
	Item 6	Trained	10	2.00	−0.400	n.s.
		Untrained	10	2.40		
	Item 7	Trained	10	1.70	−1.00	<0.01
		Untrained	10	2.70		
Mono	Item 5	Trained	6	1.50	−1.167	<0.05
		Untrained	9	2.67		
	Item 6	Trained	6	1.67	−0.778	<0.05
		Untrained	9	2.44		
	Item 6	Trained	6	2.17	−0.167	n.s.
		Untrained	9	2.33		

10 Discussion

Regarding our two major experimental issues – **innovation** and **communication** – the second may be commented with greater ease as the findings approximate prior expectations. In terms of communication problems, no difference between mono- versus multidisciplinary teams has been found. Yet, D-School teams consistently report less difficulties than untrained teams. Does D-School training enhance communication skills so that communication obstacles may be handled more easily? Potentially. In pondering this causal claim, it needs to be considered that D-School trained team members generally knew each other in advance as they had studied together at the D-School. This familiarity yields an alternative explanation for reduced communication difficulties. Yet, quite a few of the untrained participants had known each other in advance as well. For example, most monodisciplinary teams comprised students of software systems engineering who knew each other from regular courses. Thus, there is some reason to assume that D-School training helps people to develop effective communication strategies. Whether the training does indeed have a causal effect in that regard, and what elements of the D-School experience most powerfully enhance communication skills, are issues that would have to be addressed by further studies.

More demanding, and potentially more interesting is the issue of **innovation**. Why were D-School teams, and multidisciplinary D-School teams in particular, outperformed by teams with no D-School experience?

A first reply might highlight the shortness of time available for the task. In a Design Thinking process, teams are encouraged to explore the problem space copiously before actually deciding on one particular solution. Indeed, this is what D-School teams did in the experiment. Untrained teams, on the other hand, were much quicker to decide. Quite a few of them selected their approach on the first day of the workshop. This left them with a lot more time for developing and refining a prototype. Following this line of thought, one might argue that D-School teams would have performed much better had they had a few more days to work on the project. Yet, this line of reasoning does not seem to endure careful consideration. After all, the experts did not rate the prototypes presented by the teams. These prototypes were, as a matter of fact, all rather foreshadowing than usable. What the experts did rate were the *ideas* teams had come up with. (If the suggestions were to be carried out, how helpful would they be?) D-School teams spent a lot of time selecting their idea, so the process of evaluation applied in the experiment should not work to their disadvantage. Thus, the supremacy of Non-D-School teams in our experiment calls for another explanation.

One important hint may be the strong negative correlation between **usefulness** and **unusualness**. Wild ideas are explicitly encouraged in the D-School training. While there is no need to question this outlook in general, there certainly is a danger of what may be called an **oddness trap**. When much effort is put into devising a solution that others will find surprising, solutions may be surpassed that are rather self-evident and yet highly effective. Indeed, these likely solutions may be the most effective ones in some circumstances. A "go-for-the-wild" approach might be more productive in circumstances when basically all likely solutions have already been explored and something else is wanted. In our experiment, this was obviously not the case. In all conditions, the average expert rating of "unusualness" falls between 2 and 3. That is, the experts state they have already considered the solutions presented by the teams, just not in all details precisely as the groups would have them.

In general, awareness of the oddness trap – knowing that there may be a trade-off between **unusualness** and **usefulness** – is only a first step. What we ought to strive for are means, strategies and potentially even techniques for avoiding the trap. Falling in love with funny ideas must not deflect designers from the user's true needs.

11 What We Wish to Pass Back

Having been endowed with a number of considerations by the Design Thinking community, we focused on a few recurrent believes. Now that the experimental results are in, our theory prototypes may be refined. In the dialogue between Design Thinkers and empirical reality, some hotspots have been identified that certainly

span room for improvements. So, how can we sharpen our vocabulary? How can we refine our central believes so that they be ever more adapt to reality, ever more fruitful?

Regarding Design Thinking education, we might consider more explicitly what it is we wish to promote in differing circumstances. Certainly, there may be many situations in which fanciness or oddness is valuable in itself. In other cases, the users will want nothing but a working solution – whether fanciful or not. Maybe we can do a better job in systematising circumstances under which fanciness versus usefulness needs to be the ultimate standard. Maybe usefulness should always be the ultimate standard because fanciness trumps only when there is a major need for fanciness. In parallel to these theoretical issues, methodological considerations are likely as well: Should we equip students with (more) powerful methods to ensure a close(r) tie to the users' central needs? If so, ought we to provide a fixed procedure or would it suffice to make utility tests more explicit a factor in Design Thinking process models? Or, to name another possibility, should "carful utility tests" rather be taught as an overarching value/goal that students need to internalize?

Regarding the second experimental issue, we wish to turn to the advocates of multidisciplinarity in particular. Taking seriously the experimental results, some refinement in Design Thinking theory would seem helpful. This does not necessarily mean a major reorientation; some further specifications might due.

Perhaps multidisciplinarity does have a positive effect on innovation – but the effect is so small that it was easily overridden (and even "conversed") by chance variation in our experimental setting. If this is true, Design Thinking theory would surely profit from a realistic estimate of the effect size: If the effect size is small, we need to expect very limited gains with respect to innovation simply by assembling multidisciplinary instead of monodisciplinary teams. Or, to address another likely reasoning: Multidisciplinarity may have a considerable positive effect, but not in all contexts. For example, it comes to unfold its positive impact only after longer periods of time (months, not days). Another viable thought may be that multidisciplinary design teams provide more helpful prototypes than monodisciplinary ones when it comes to communicating design ideas to development divisions who work out final products. Such a handover was no subject of our experiment. Thus, there are many ways in which Design Thinking theory may be carried forwards by helpful specifications.

In sum, there is "experimental feedback" we may seek and use to refine Design Thinking theory – just as there is "user feedback" which design teams may seek and use to refine their prototypes. To be sure, this seeking and refining is a lot of hard work! And it may be a painful experience to see ones precious conceptions wobble under the pressure of an experimental test. But: We wouldnt be Design Thinkers if we were to duck out of the test, would we?