WE NEED A UNIVERSAL DESIGN PROJECT OUTCOME PERFORMANCE MEASUREMENT METRIC: A DISCUSSION BASED ON EMPIRICAL RESEARCH

Philipp Skogstad[1], Martin Steinert[2], Karl Gumerlock[1], and Larry Leifer[1]
(1) Stanford University, USA (2) University of Fribourg, CH

ABSTRACT

This paper aims to contribute and channel a discussion on the need for a universal outcome performance measurement for design projects. After introducing the problem verbally, an example situation of a missing universal success metric is given based on research in the context of a project-based engineering curriculum. The lack of such metric poses two problems: 1) when facing the choice between design projects of fundamentally different nature and industries, enterprises make suboptimal resource allocation decisions; 2) since a general common denominator in form of a success metric is missing, current design research is restricted to analyzing and comparing similar design projects. Currently there is no agreement on a general construct or variable that defines and measures the success of design projects. Hence a universally accepted dependent variable may be needed in order to create, test and verify/falsify hypotheses and theories through methods such as meta-analyses.

Keywords: design metrics, performance measurement and evaluation, inter project comparison, engineering design education

INTRODUCTION – THE INTER-PROJECT DESIGN MEASUREMENT PROBLEM

The aim of this paper is to contribute to the discussion amongst design research scholars and practitioners on the need of a universal design project outcome performance measurement metric that allows comparison of design projects with different natures. The authors claim that without such a common denominating measurement, resource allocation for diversified companies and venture capitals must remain rather suboptimal, and the creation of design research theory is severely hindered.

Measuring design performance is a key and ongoing issue for designers and companies employing designers inside or outside of their organization. The same applies to venture capitalists investing in new product/service concepts. Last but not least, measurement of design performance is crucial for design researchers, especially those whose approach is based on empirical research comparing different projects. Hewlett-Packard even proclaims a quest for “the holy grail of design measurement” [1], though only a framework and rules of thumb are given.

While intra-project performance measurement may be addressed quite specifically, using project time, budgets, design team member satisfaction user relevant performance criteria, number of parts, production and assembly cost, feasibility etc., measurement of inter-project outcome performance is a substantially trickier task. When comparing projects of identical nature or from within a particular industry, e.g. chip manufacturing, it is possible to identify more successful design teams and their projects based on established criteria. However, when comparing projects with *fundamentally different tasks*, the established criteria will not identify the most successful design team or project. Based on which criteria can you compare an automobile design project and a software design project? Which measurement should be used in order to determine which design project has been more successful? How do you define outcome success for various design tasks?

This paper is aimed at both academic and entrepreneurial audiences. These groups are seriously constrained by the fact that generally accepted measurement metrics to compare the outcome performance of arbitrary design projects are missing.
With this paper, the authors would like to contribute and channel this debate by

1) making the problem explicit,
2) providing an empirical example and
3) highlighting the negative effect of this situation for both, academia and business.

The paper builds on Skogstad 2009 who observed that "the difficulty of accurately measuring design process performance is a key challenge in design research, as this study confirms again" [2].
Methodologically, the authors argue based on data from a project-based engineering curriculum at the department of Mechanical Engineering at Stanford University. Correlations between existing performance measurement metrics and output indicators based on external expert opinion are absent, exemplifying that a common output performance measurement or success construct is not obvious to construct.
Data from the curriculum shows that inter-project comparison between substantially different design tasks is difficult to achieve. Venture capitalists and large companies are therefore unable to directly compare projects using fundamentally different early stage design prototypes that address different customer needs and markets. Consequently, it is difficult to optimize resource allocation e.g. in an investment portfolio.
A lack of universal criteria is equally problematic for design research, as it is currently not possible to statistically compare fundamentally different design projects. Hence, design paradigms such as the optimum amount of iterations, optimum team size, optimum team composition, optimum organizational structure cannot be statistically verified, because the common dependent output performance variable is missing. Researchers are thus unable to test the different paradigms and models under varying design project circumstances. As a result, design research may be stuck in a situation where statistically founded theory cannot emerge.

DATA FROM PROJECT-BASED ENGINEERING DESIGN CURRICULUM

Stanford University's 'Mechanical Engineering 310 – Project-Based Engineering Design, Innovation & Development' course (henceforth 'ME310'), a graduate level project-based curriculum, served as a living laboratory for this research. In this engineering design class, Stanford students collaborate with students at other universities around the world to develop innovative solutions to open-ended problems posed by industry partners. The duration of the course is nine months. The nature of the projects and the structure of the teams closely resemble the working mode of start-ups and "tiger teams" [3] in industry. Like many Silicon Valley initiatives, design teams start with a vague idea of an area that allows for the creation of innovation. The problem briefs are purposely phrased broadly to challenge the students to determine, isolate, and pursue a particular opportunity for innovation.
Student teams tackle the industry-posed problems over the course of seven months, and develop fully functional product prototypes of their solutions. Several milestones, intermediate prototype reviews with a teaching team and presentations to industry help structure the project and ensure that results fulfill expectations.
Similar to start-ups or design teams in industry, the course has its own designated workspace, called the 'Loft'. It resembles a design studio space more closely than a classroom. The space is much like an incubator—each team has its own 'private' work area to outfit and personalize. The Loft also houses a variety of basic hand tools and building materials for prototype construction. Information technology infrastructure includes computer terminals for computer-aided-design (CAD) and for video editing, printers, phones, and video conferencing equipment.
The project nature, timeline, and the interactions such as intra-team, team/instructor, team/coach, team/company liaison, inter-team and within the community show that the design teams and projects in this curriculum closely resemble those found in industry. Unlike industry, however, ME310 is uniquely suitable for research studies as the milestones, deadlines, as well as project breadth and depth are common between teams. These commonalities make ME310 a controlled environment compared to industry settings, which makes the effect of team circumstances on design performance observable in ways not possible elsewhere.

The data for this study stems from the 2007/2008 offering of the curriculum and comprises eleven participating design project groups. The general topic and budget for each team were provided by an industry partner that also supplied a liaison to facilitate communication between the company and the students. Each group is comprised of students from Stanford University and students from an international partner university. This reflects the situation of having to work in a dispersed international setting with a multitude of communication tools and applications. In total 71 students participated and were willing to complete questionnaires. The following design challenges were given to the student design teams:

- Design an intelligent system to assist drivers (driver-assist)
- Design a tool that facilitates distant design collaboration (design-collaboration)
- Design a system to store small items in a car (car-storage)
- Design a tool to support maintenance personnel in the field (maintenance-assist)
- Design an new automotive center stack (center-stack)
- Design a device that extracts drinking water from ambient air (water-extraction)
- Design a new digital camera (digital-camera)
- Design a method to control wearable electronic devices (wearable-control)
- Design a new way to use RFID in a retail environment (RFID-retail)
- Design an industrial controller based on the Wii technology (Wii-industrial)
- Design a virtual convertible (virtual-convertible)

Internal Design Performance Measurements

The following performance indicators data were collected on the individual level. Using a standardized questionnaire, the students supplied information concerning their satisfaction levels, their individual and team energy level, the state of their collaboration, and team performance. Additionally, external experts were consulted:

Satisfaction (1)

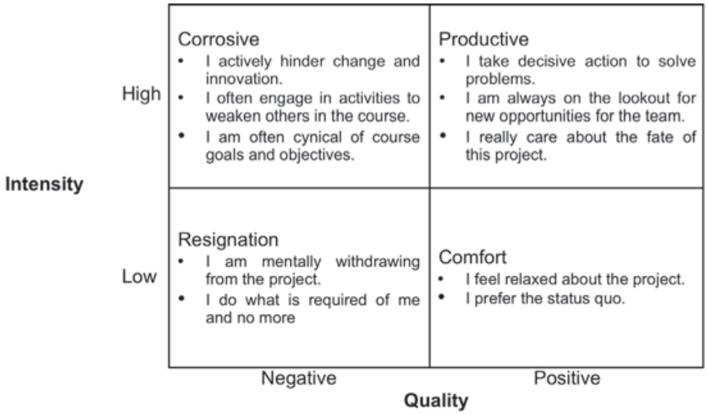
Three questions asked the designers how

a) satisfied they were with their own progress
b) satisfied they thought their liaison was
c) satisfied they thought their teaching team was.

The questions were given on a four-point scale from “very unsatisfied” to “very satisfied,” with the median omitted to require a decision. An additional question tested the students’ confidence in their ability to complete the project successfully.

Individual and team energy level (2)

Two questions asked the designers to estimate their a) team’s and b) their own attitude towards the project and towards each other. The questions were based on the framework of organizational energy [4]. Following suggestion by Bruch, the framework shown below was presented and respondents were asked to place a mark for themselves and a mark for their team in the table in figure 1.
This mark was coded in two dimensions, 0 to 20 for intensity and -10 to 10 for quality, thus allowing for a statistical analysis.



Intensity	Negative	Positive
High	Corrosive • I actively hinder change and innovation. • I often engage in activities to weaken others in the course. • I am often cynical of course goals and objectives.	Productive • I take decisive action to solve problems. • I am always on the lookout for new opportunities for the team. • I really care about the fate of this project.
Low	Resignation • I am mentally withdrawing from the project. • I do what is required of me and no more	Comfort • I feel relaxed about the project. • I prefer the status quo.

Figure 1: Framework used to ask about Attitude in Questionnaire (adapted from [5])

Global Collaboration (3)

Three questions asked team members about their identification with the group as a whole since this is positively associated with the willingness to collaborate: [6] [7] [8]

a) Sense of belonging to the entire global vs. local sub-team.
b) Communication between sub-teams is very fluent vs. not occurring.
c) Cooperation within global team is smooth and enjoyable vs. dysfunctional.

The closeness of the two parts of the team, the quality of their communication, and the quality of their cooperation were asked and tested on a five point Likert scale.

Design process performance (4)

A special survey was administered after completion of the project to measure process performance from the perspective of the designer. The five point Likert scale questions have been taken from an tested team diagnostic survey (TDS) diagnosis instrument [9]. Three performance-relevant aspects of teamwork, which are controlled by the team and its members, were collected. Each item is based on three to four, sometimes reversed questions: [9]

a) Measure for process criteria of team effectiveness

Effort-related process criteria:

- Members demonstrate their commitment to our team by putting in extra time and effort to help it succeed.
- Everyone on this team is motivated to have the team succeed.
- Some members of our team do not carry their fair share of the overall workload. (R)

Strategy-related process criteria:

- Our team often comes up with innovative ways of proceeding with the work that turn out to be just what is needed.
- Our team often falls into mindless routines, without noticing any changes that may have occurred in our situation. (R)
- Our team has a great deal of difficulty actually carrying out the plans we make for how we will proceed with the task. (R)

Knowledge-and-skill-related process criteria:

- How seriously a member's ideas are taken by others on our team often depends more on who the

person is than on how much he or she actually knows. (R)
- Members of our team actively share their special knowledge and expertise with one another.
- Our team is quite skilled at capturing the lessons that can be learned from our work experiences.

b) Measure for team interpersonal processes

Quality of team interaction:
- There is a lot of unpleasantness among members of this team. (R)
- The longer we work together as a team, the less well we do. (R)
- Working together energizes and uplifts members of our team.
- Every time someone attempts to correct a team member whose behavior is not acceptable, things seem to get worse rather than better. (R)

Satisfaction with team relationships:
- My relations with other team members are strained. (R)
- I very much enjoy talking and working with my teammates.
- The chance to get to know my teammates is one of the best parts of working on this team.

c) Measure of the individual's learning and well-being

Internal work motivation:
- I feel a real sense of personal satisfaction when our team does well.
- I feel bad and unhappy when our team has performed poorly.
- My own feelings are not affected one way or the other by how well our team performs. (R)
- When our team has done well, I have done well.

Satisfaction with growth opportunities:
- I learn a great deal from my work on this team.
- My own creativity and initiative are suppressed by this team. (R)
- Working on this team stretches my personal knowledge and skills.

General satisfaction:
- I enjoy the kind of work we do in this team.
- Working on this team is an exercise in frustration. (R)
- Generally speaking, I am very satisfied with this team.

This survey provides a measurement of the design process quality from the designer's point of view. It accounts for the fact that design performance must include more than just the project result, because no organization will survive if the designers are consistently unsatisfied.

Résumé Indicators

All four indicators measure a certain concept and aspect of design project performance. 2) and 4) are academically well established instruments, 3) is a combination based on reviewed works and 1) is an attempt to get the performance information at its source, the participating design team members. These metrics are exemplary for conducting longitudinal analysis of projects and/or for comparing projects of similar nature. That way, the best team structure or project setup may be discovered.
For the line of argument in this paper, it is hypothetically assumed that these indictors address factors influencing overall performance. Hence it is assumed that these indictors may be used as independent variables for a dependent outcome success metric. Such an external success measurement might e.g. be based on independent expert opinion.

External Expert Opinions

Additionally to the designer centric data, external experts were asked to judge the outcome of the design projects. The goal was to measure the output of the design process in an unbiased manner.

Based on two-page project summaries, nine experts with no direct contact to the projects were asked to grade the projects from three different perspectives, as:

a) an investor,
b) a user, and
c) a gadget lover.

Judgment was indicated using the standard US grading system (A=4, B=3, C=2 D=1, F=0). The judges were selected to be familiar with the structure and goals of the course to ensure that they could evaluate the designs based on the organizational circumstances under which they were created. At the same time, it was ensured that the judges did not have prior knowledge of the details of the projects or of the design teams, so that they were not biased by the evolution of the design.

Table 1 Project grades from an investor's perspective

INVESTORS	***expert***	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*		
project											***avg***.	***rank***
driver-assist		**4**	*2*	3	*1*	*3*	*0*	3	3	**3**	**2.44**	**6**
design-collaboration		3	*2*	3	2	**4**	1	**4**	3	*2*	**2.67**	**4**
car-storage		**4**	*2*	1	2	*3*	*0*	2	*1*	**3**	**2.00**	**10**
maintenance-assist		*2*	**4**	*0*	**3**	**4**	2	3	**4**	**3**	**2.78**	**3**
center-stack		3	*2*	2	*1*	**4**	*0*	3	3	*2*	**2.22**	**9**
water-extraction		3	**4**	3	**3**	**4**	2	**4**	3.7	*2*	**3.19**	**1**
digital-camera		3	3	1	2	*3*	*0*	3	3.7	**3**	**2.41**	**7**
wearable-control		3	3	3	2	*3*	*0*	2	2	**3**	**2.33**	**8**
RFID-retail		3	3	**4**	2	*3*	*0*	2	3.3	**3**	**2.59**	**5**
Wii-industrial		**4**	3	2	**3**	**4**	**3**	*1*	3.7	**3**	**2.97**	**2**
*virtual-convertible***		*2*	*2*	2	2	*3*	1	*1*	3	*2*	**2.00**	**10**
grader average		**3.09**	**2.73**	**2.18**	**2.09**	**3.45**	**0.82**	**2.55**	**3.04**	**2.64**	**2.51**	

Table 2 Project grades from a user's perspective

USERS	***expert***	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*		
project											***avg***.	***rank***
driver-assist		**4**	**4**	**4**	*2*	*2*	2	**4**	2.7	3	**3.08**	**3**
*design-collaboration**		3	**4**	3	3	**4**	1	**4**	**3.3**	3	**3.14**	**1**
car-storage		3	3	**4**	3	*2*	**4**	3	*2*	**4**	**3.11**	**2**
maintenance-assist		**4**	3	*0*	2	3	2	3	**3.3**	*2*	**2.48**	**7**
*center-stack***		3	*2*	3	3	3	*0*	2	2.7	*2*	**2.30**	**9**
water-extraction		**4**	*2*	3	3	**4**	2	**4**	3	*2*	**3.00**	**5**
digital-camera		3	3	2	3	3	*0*	2	**3.3**	3	**2.48**	**7**
wearable-control		*2*	**4**	3	*2*	*2*	*0*	*1*	*2*	**4**	**2.22**	**10**
RFID-retail		3	3	**4**	*2*	3	1	3	**3.3**	*2*	**2.70**	**6**
*Wii-industrial**		**4**	**4**	1	**4**	3	3	2	**3.3**	3	**3.03**	**4**
*virtual-convertible***		3	*2*	1	*2*	3	3	*1*	3	*2*	**2.22**	**10**
grader average		**3.27**	**3.09**	**2.55**	**2.64**	**2.91**	**1.64**	**2.64**	**2.90**	**2.73**	**2.71**	

Table 3 Project grades from a gadget lover's perspective

GADGET LOVERS	***expert***	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*		
project											***avg.***	***rank***
driver-assist		3	3	**4**	*2*	3	1	*3*	2.7	3	**2.74**	**8**
design-collaboration		**4**	**4**	3	3	3	3	**4**	3	*2*	**3.22**	**4**
*car-storage***		*1*	*0*	2	*2*	*2*	*0*	*3*	*2*	3	**1.67**	**11**
maintenance-assist		2	1	*0*	3	**4**	1	**4**	3.3	*2*	**2.26**	**10**
center-stack		2	3	2	*2*	3	1	**4**	2.3	3	**2.48**	**9**
water-extraction		**4**	2	3	**4**	**4**	3	*3*	3	*2*	**3.11**	**5**
digital-camera		**4**	2	2	3	**4**	1	*3*	3.3	**4**	**2.92**	**6**
wearable-control		**4**	**4**	**4**	3	**4**	2	*3*	*2*	**4**	**3.33**	**1**
RFID-retail		3	2	3	*2*	3	3	**4**	3.3	3	**2.92**	**6**
*Wii-industrial**		**4**	3	2	3	**4**	2	**4**	**3.7**	**4**	**3.30**	**3**
*virtual-convertible***		3	3	**4**	3	**4**	**4**	**4**	3	*2*	**3.33**	**1**
grader average		**3.09**	**2.45**	**2.64**	**2.73**	**3.45**	**1.91**	**3.55**	**2.87**	**2.91**	**2.84**	

In order to highlight the level of subjectivity and the difference of the grading for each grader, the best and worst teams have been highlighted. A quick glance at the data in tables 1, 2 and 3 makes it obvious that there is no clear ranking or homogeneous opinion concerning the output performance of the design teams amongst the experts.
A first observation concerns the average of the grader - the experts seem to have quite different opinion on how to use the available grading scale. Their averages grading range from:

0.82 to 3.09 with an average of 2.51 (INVESTORS)
1.64 to 3.27 with an average of 2.71 (USERS)
1.91 to 3.55 with an average of 2.84 (GADGET LOVERS)

Secondly, from the 33 observed cases, 26 have been graded with both, the best and the worst grade at least once. Only 3 cases (*)[1] show a tendency towards a higher grade and only 4 cases (**)[2] show a tendency towards a lower grade.

In addition to giving grades, the experts were asked to produce a forced ranking of the design projects, as depicted in table 4.

Table 4 Project ranking based on a forced ranking

FORCED RANKING	***expert***	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*		
project											***avg.***	***rank***
driver-assist		2	6	3	*11*	8	3	4	8	5	**5.56**	**6**
design-collaboration		4	3	8	7	2	4	**1**	6	10	**5.00**	**4**
*car-storage***		9	*11*	9	6	*11*	9	7	*11*	3	**8.44**	**10**
maintenance-assist		*11*	4	*11*	5	4	10	5	2	7	**6.56**	**7**
center-stack		7	9	6	9	5	*11*	3	9	8	**7.44**	**9**
water-extraction		6	2	4	**1**	**1**	**1**	2	5	9	**3.44**	**1**
digital-camera		3	7	10	4	6	5	8	3	2	**5.33**	**5**
wearable-control		**1**	**1**	**1**	3	10	6	10	10	**1**	**4.78**	**3**
RFID-retail		8	8	5	8	9	7	6	4	6	**6.78**	**8**
*Wii-industrial**		5	5	7	2	3	2	9	**1**	4	**4.22**	**2**
*virtual-convertible***		10	10	2	10	7	8	*11*	7	*11*	**8.44**	**10**

[1] * have been graded by at least one expert with the best grade

[2] ** have been graded by at least one expert with the lowest grade

Again, strong deviations between the individual rankings come to the foreground. When comparing the results of the forced ranking with the ranks derived from the grades, the picture remains contradictory (see left part of table 5).

Table 5 Comparison forced ranking with investors, users and gadget lovers ranking

RANKING COMPARISON	***expert***	*INVESTORTS*	*USERS*	*GADEGET LOVERS*	*avg. rank*	*FORCED RANKING*	*delta*
project							
driver-assist		6	3	8	4	6	2
design-collaboration		4	1	4	3	4	1
*car-storage***		10	2	11	11	10	1
maintenance-assist		3	7	10	9	7	2
center-stack		9	9	9	10	9	*1*
water-extraction		1	5	5	2	1	1
digital-camera		7	7	6	7	5	2
wearable-control		8	10	1	6	3	3
RFID-retail		5	6	6	5	8	3
*Wii-industrial**		2	4	3	1	2	1
*virtual-convertible***		10	10	1	8	10	2

However, an average created from the three perspective ranks compares reasonably well with the overall forced ranking (max. rank deviation of 3). Although this result might hint to the notion of crowd based wisdom, [10] the final verdict emerges rather strongly—external expert-based outcome performance measurement does not necessarily result in a valid and reliable metric that would allow for comparison of design projects of different nature.

CORRELATING PERFORMANCE INDICATORS?

In order to test this assumption, the average forced ranking as depicted in table 4 has been defined as outcome performance measurement metric. This measurement forms the dependent variable that is influenced by the four performance concepts measured before.

A priori, the Kolmogorov–Smirnov test (K–S test) on normal distribution has been conducted. The independent performance measurements “global collaboration” and “design process performance” as well as “team energy level” are normally distributed.[3] The dependent variable “forced ranking” is also normally distributed.[4] For those variables, the Pearson product-moment correlation coefficient, also known as "Pearson's r" may be applied for correlation testing.

The variables “satisfaction” and “student energy level” are not normally distributed.[5] Hence Spearman's rank correlation coefficient or Spearman's rho (ρ), and Kendall tau rank correlation coefficient or Kendall's τ or tau test must be used when testing for correlations (rank only).

Since the linear regression analysis of the forced ranking described by the four performance indictors did not lead to meaningful results, the correlation between “forced ranking” (outcome performance metric) and the four sub performance concepts were tested. *No correlation could be found!*[6]

[3] K-S test asymp. sig. 2-sided: 4) (.280), 3) (.157), 2a) intensity (.300) and quality (.319)

[4] K-S test asymp. sig. 2-sided: Forced Ranking (.108), is normally distributed.

[5] K-S test asymp. sig. 2-sided: 1) (.000), and 2b) intensity (.041) and quality (.001)

[6] A similar result is obtained when testing the correlation of the four indictors and the projects grades as given by the teaching team. No correlation can be found.

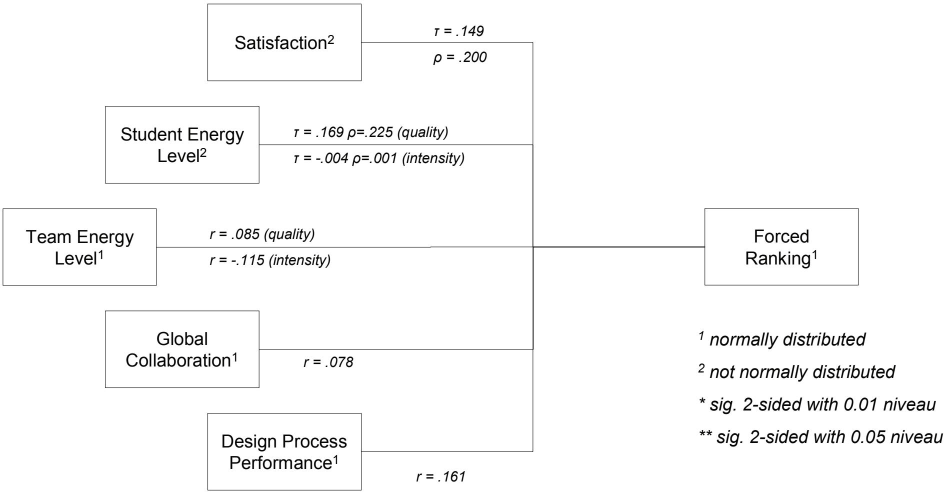


Figure 2: No significant correlations between outcome variable and standard performance metrics are apparent

The satisfaction from the perspective of the team members, their individual and team energy level, the state of their collaboration and their design process performance do not correlate with the evaluation of their design project outcome as seen by independent external experts. Although we might be able to measure which team collaborated best and which team was most satisfied, we cannot indicate which project was the most successful at the end.

Table 6. Correlation matrix of the variables

CORRELATION MATRIX	forced ranking[1]	satisfaction[2]	student energy (quality)[2]	student energy (intensity)[2]	team energy (quality)[1]	team energy (intensity)[1]	global collaboration[1]	design process performance
forced ranking[1]	***1***	***τ=.149 ρ=.200***	***τ=.169 ρ=.225***	***τ=-.004 ρ=.001***	***r=.085***	***r=-.115***	***r=.078***	***r=.161***
Satisfaction[2]		1	τ=.255* ρ=.311*	τ=.061 ρ=.076	τ=.206* ρ=.252*	τ=.120 ρ=.149	τ=.294** ρ=.356**	τ=.223* ρ=.276*
student energy (quality)[2]			1	τ=.352** ρ=.425**	τ=.533** ρ=.641**	τ=.345** ρ=.416**	τ=.340** ρ=.439**	τ=.227* ρ=.315*
student energy (intensity)[2]				1	τ=.446** ρ=.551**	τ=.548** ρ=.671**	τ=.105 ρ=.145	τ=.320** ρ=.429**
team energy (quality)[1]					1	r=.578**	r=.513**	r=.407**
team energy (intensity)[1]						1	r=.309*	r=.234
global collaboration[1]							1	r=.246
design process performance[1]								1

[1] *normally distributed*
[2] *not normally distributed*
* *sig. 2-sided with 0.01 niveau*
** *sig. 2-sided with 0.05 niveau*

Each of the four independent indictors does excellent work in its own right and together they give a good picture of the performance of the teams, as the correlation matrix shows. However, when confronted with the task of comparing the outcome success of different projects, they all seem to fail.

Résumé

Of course, the data from the curriculum does not allow for generalization, because external validity is not given. Nevertheless, it serves as an example that although we do have well established diagnosis instruments for measuring certain aspects of design projects, the overall outcome performance measurement is rather elusive. External experts and their forced ranking do e.g. not necessarily serve as a general success indicator.

If such a general indictor is missing, and if we are not able to identify the successful design teams when comparing projects of fundamentally different nature, we are confronted with two major problems.

ENTREPRENEURIAL IMPLICATIONS

The first problem is of entrepreneurial nature. Companies have to make a resource allocation decision during the fuzzy front-end phase of innovation, often based on the results and experience gained in design projects. Venture capitalists (VCs), for example, must decide which team to fund. A similar stage-gate problem is inherent in larger companies that are tasked with deciding which innovation project may receive further investment in order to proceed to the next stage. [11] Typically the majority of projects do not make the cut. If the VCs and the companies are acting in multiple industries, they have to compare projects of entirely different natures. The standard measurement metrics do not help in identifying the innovation with the greatest potential of success.

Therefore, companies and VCs tend to first fall back on investment theory and cash flow analysis such as net present value (NPV) and discounted cash flow analysis (DCF). Unfortunately, these monetary and risk based analyses are, though precise at first sight, inherently vulnerable to future uncertainty and are based on subjective assumptions (future cash flow in year 20xy based on a risk premium supplement interest *r*). Second, at the stage gates some metrics on the projects are usually collected. The subsequent decisions though are often based on a biased scoring model, experience, and/or gut feeling.

VCs and companies would thus greatly benefit from a general output performance measurement, as it might allow them to allocate their resources on a more objective basis.

ACADEMIC IMPLICATIONS AND RESEARCH AGENDA

The lack of general design project success measurement metrics creates another problem: the emergence of a general empirically based design paradigm may be hindered. There are two ways for establishing a paradigm.

Deductive reasoning produces valid assumptions and constructs, as long as their conclusion is a logical consequence of the true premises. In social sciences, but also in natural sciences, when confronted with a complex problem such as planetary climate, often the prior model assumptions are constrained in relation to reality. The resulting theories are therefore not necessarily externally valid and may only be generalized with caution.

The second approach, inductive reasoning, tries to counter this by building theory based on observations. These may be taken *in vitro* from experimentation or *in sito*, e.g. from case research. The results are then generalized into hypotheses or used as analogies. Over time a causal relation paradigm may emerge, though based on singular observations only. In order to increase the strength of the reasoning, the different hypotheses are tested logically or statistically. Should they withstand falsification, [12] i.e. when no contradicting evidence is produced and the results are statistically significant, the paradigms may result in general theory, waiting to be tested further by another generation of researches that posses other means and techniques.

In design research we are faced with a multitude of different design tasks and projects. Based on certain observations and experiments, hypotheses have emerged, such as the superiority of a heterogeneous design team composition, and the importance of iterations in the design process. In order to test these hypotheses for general design projects, it is necessary to have a universal performance criterion. This criterion, which should allow differentiating between more or less successful design projects, may act as dependent variable. Individual studies and observations from multiple data sources may be meta-analyzed using it. Researchers would then be able to attempt to falsify current design research assumptions, allowing the emergence of general, project and industry independent theory and paradigms.

CONCLUDING REMARKS

This paper attempted to convince the reader that a general applicable success construct for design projects is needed. Although plenty of tried and tested performance indicators and instruments exist, when comparing design projects of fundamentally different nature, the identification of the best projects poses serious problems because no commonly accepted success construct exists. Due to this, resource allocation in companies may be suboptimal. Furthermore, since no common success denominator exists, the creation of design theory might be hindered, since researchers are unable to statistically compare different design projects. Hence, the Popper cycle of hypotheses creation, testing and falsification/verification usually remains within a specific design domain or industry. By introducing a common design project success measurement construct, design research would benefit from universally applicable operational and testable paradigms.

The authors are quite aware that such a “silver bullet” metric does probably not exist. However, the primary aim of this paper is to motivate the reader to comment on this train of thought and to make suggestions toward a possible design project outcome performance measurement metric. Contributions are more then welcome and will be collected i.a. during the ICED’09 conference at Stanford.

REFERENCES

[1] Mrazek, D., Lucente, S., Sato, S., Menter, A., Wai, C., Wakid, K., and Hartley, P., *The Holy Grail of Design Measurement - to measure design, don't build just another measurement system*. Vancouver, OR: Hewlett-Packard, 2006.

[2] Skogstad, P., "A Unified Innovation Process Model for Engineering Designers and Managers," in *Department of Mechanical Engineering*. Doctoral Dissertation Stanford University, Stanford, CA, 2009.

[3] Wheelwright, S. C. and Clark, K. B., *Revolutionizing product development: quantum leaps in speed, efficiency, and quality*. New York, NY: Free Press, 1992.

[4] Bruch, H. and Ghoshal, S., "Unleashing Organizational Energy," *MIT Sloan Management Review,* vol. 45, pp. 45-52, (2003).

[5] Bruch, H., "Definition of Organisational Energy," St. Gallen, 2009, accessed 02.01.2009 2009, from http://www.ifpm.unisg.ch/org/ifpm/web.nsf/wwwPubInhalteGer/Results+and+Concepts.

[6] Hakonen, M. and Lipponen, J., "Antecedents and consequences of identification with virtual teams: Structural characteristics and justice concerns," *The Journal of E-working,* vol. 1, pp. 137-153, (2007).

[7] Riketta, M., "Organizational identification: A meta-analysis," *Journal of Vocational Behavior,* vol. 66, pp. 358-384, (2005).

[8] Tyler, T. R. and Blader, S. L., "The Group Engagement Model: Procedural Justice, Social Identity, and Cooperative Behavior," *Personality and Social Psychology Review,* vol. 7, pp. 349-361, (2003).

[9] Wageman, R., Hackman, J. R., and Lehman, E., "Team Diagnostic Survey - Development of an Instrument," *The Journal of Applied Behavioral Science,* vol. 41, pp. 373-398, (2005).

[10] Surowiecki, J. M., *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*: Little Brown, 2004.

[11] Cooper, R. G., "Doing it right: Winning with new products," *Ivey Business Journal,* vol. 64, pp. 54-61, (2000).

[12] Popper, K. R., *The logic of scientific discovery*, 2nd ed.: Routledge, 2002.

Contact:
Philipp Skogstad
Center for Design Research
Stanford University
424 Panama Mall
Bldg. 560
Stanford, CA 94305
tel. +1 650-799-0298
fax. +1 650-725-8475
skogstad@cdr.stanford.edu
skype: philippskogstad

Martin Steinert
University of Fribourg (iimt)
Boulevard de Pérolles 90
CH - 1700 Fribourg
Switzerland
tel. +41 26 300 8433 / 8430
fax +41 26 300 9794
martin.steinert@unifr.ch
skype: martin.steinert
http://www.iimt.ch/index.php?id=138

Karl Gumerlock
Center for Design Research
Stanford University
424 Panama Mall
Bldg. 560
Stanford, CA 94305
tel. +1 650 353-7684
karl@cdr.stanford.edu

Philipp is the Deputy Director of the Center for Design Research at Stanford University. He holds a Ph.D. in Mechanical Engineering from Stanford University and an M.B.A. from the University of St. Gallen, Switzerland. His research interest is in understanding how to form and manage high-performance engineering design teams. His experience includes work for major tier-one and two automotive manufacturers, in the defence industry, and in teaching design thinking in the US and abroad.

Martin is a Senior Research Associate at the international institute of management in technology (iimt) at the University of Fribourg, Switzerland where he received his doctorate in 2006. He researchers and teaches at the Master and EMBA levels on the topics of Technology and Innovation Management (TIM), Research methods (RM) and Strategy. His special interest is to combine empirical research methods of quantitative and qualitative nature to create a comprehensive research design. In 2008 Martin was invited as visiting scholar to the Center of Design Research at Stanford University.

Karl is a Ph.D. student at the Center for Design Research at Stanford University, having previously completed his undergraduate studies and his M.S. in mechanical engineering there. His research interests lie in the emerging field of open engineering and open intellectual property practices as they apply to engineering design.

Larry is a Professor in Mechanical Engineering at Stanford University and the Director of the Center for Design Research (School of Engineering). His design thinking and informatics research is concerned with understanding, supporting and improving design practice, including issues in research methodology, team dynamics (co-located and distributed), innovation leadership, interactive design spaces, collaboration technology, and design-for-wellbeing.