INTERNATIONAL DESIGN CONFERENCE - DESIGN 2012
Dubrovnik - Croatia, May 21 - 24, 2012.

DESIGN 2012

A STANDARDIZED MEASUREMENT TOOL FOR EVALUATING AND COMPARING TEAM REFRAMING CAPABILITIES

G. Kress, M. Schar and M. Steinert

Keywords: reframing, redesign, teams, design methodology

1. Introduction

The process of innovative design is inherently nonlinear in nature, and teams must cope with a high degree of ambiguity and sometimes conflicting information. Though the design process is often represented as a series of steps that appear linear (see Figure 1), it is widely assumed that a successful design process is iterative in nature and often involves revisiting the various phases of the process time and time again. The visual links between phases and the backward looping process suggest both iteration and reframing.

Bruno Latour, a French sociologist and anthropologist who is influential in the study of science and technology has stated, "to design is always to redesign." [Latour 2008] Latour suggests that the design process never begins from scratch; rather, designers begin with a specific framing of a problem and solution and then rework both toward "something more lively, more commercial, more usable, more user friendly, more acceptable, more sustainable and so on, depending on the various constraints to which the project has to answer."

This suggests that it is important to understand and specifically focus onto the process of redesign as a central component of a successful team design process, particularly as evidenced by reframing. Presented with a problem within a specific framework, how does the designer deconstruct these and then reconstruct them toward an innovative solution? When presented with new information, how does the designer revise his or her previous frame? The Stanford Design Thinking Exercise (SDTE) is a tool that measures reframing behavior in less than one hour. The exercise, based on sequentially ranking a series of design options, is standardized and yields quantitative and reliable scores. We can use it to identify whether reframing behavior is affected by certain team characteristics. We will also have a basis for team comparison that is far less subjective and time-consuming than established research instruments. For example, researchers have had success in video coding of short-term team interactions as an indicator of long-term performance. However, these methods are very time-consuming, resource-intensive, and require significant coder training if subjectivity bias is to be minimized. Contrary to these methods, scoring of the SDTE is simple, robust and may be automated.

The task is to assess team performance in the short term and in a controlled environment. A reliable and quick team assessment tool has major implications for team-based education and design in industry; managers and instructors alike could identify potential problems in and between teams early enough to positively impact the team setup, mindset and ultimately the design outcome. Our long-term goal is to establish the SDTE as a reliable measure of intra- and inter-team reframing behavior in order to positively influence design effectiveness.

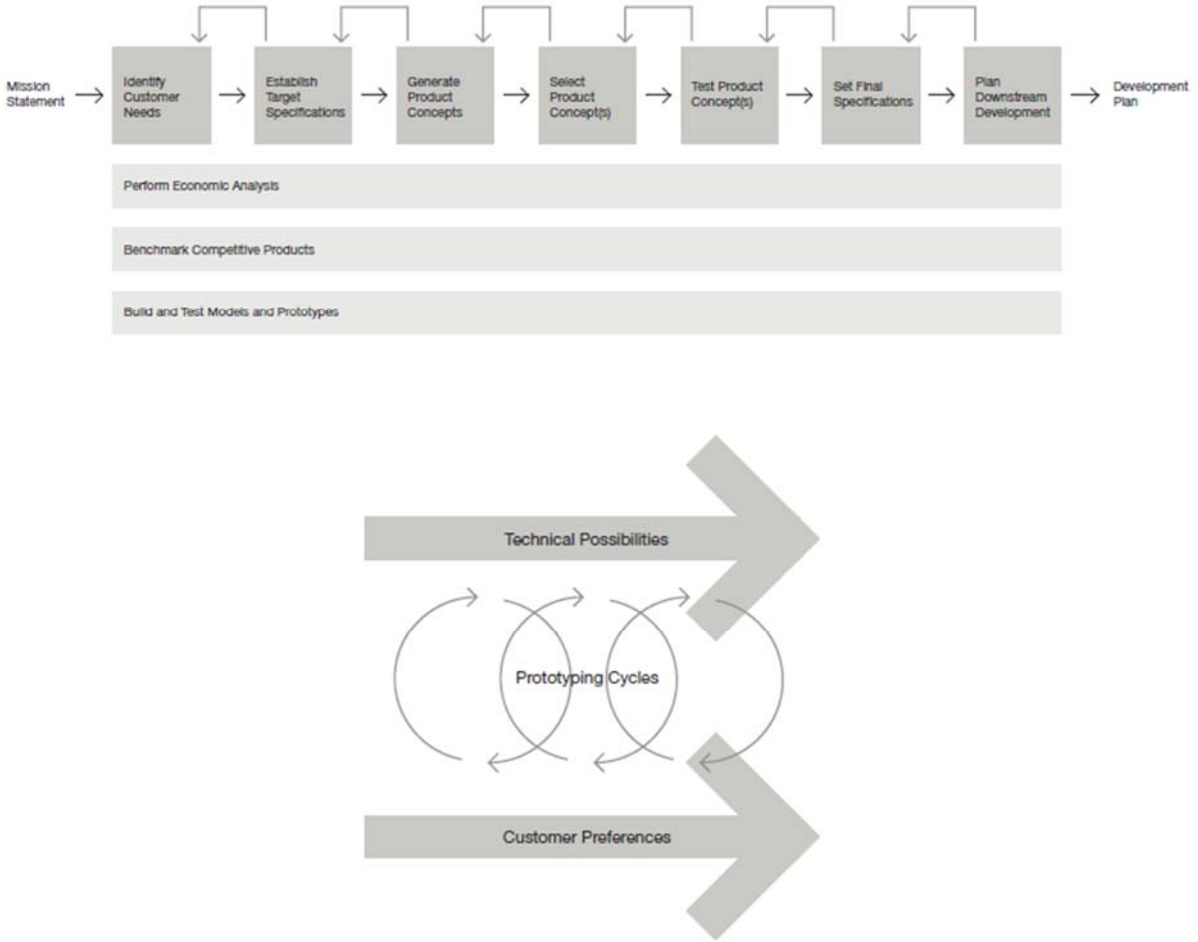

Figure 1. Steps in a design process as visualized by Eppinger & Ulrich [Dubberly 2005]

2. Background

2.1 Framing

In social theory, a frame is a schema of interpretation, the method by which an individual organizes experience and knowledge to solve problems. Erving Goffman described frame analysis as the "understanding available in our society for making sense out of events and to analyze the special vulnerabilities to which these frames of reference are subject." [Goffman 1974] Goffman describes framing as an examination of how an individual organizes experience, how they subjectively structure involvement in an event or circumstance. Simply stated, Goffman believes the frame analysis is the "way we take it that our world hangs together."

Designers bring their individual frames to the problems and solutions of design. These range from simply an understanding of "the way the world works" to more sophisticated perspectives around functionality and performance based on prior experience or expertise. If, indeed, "all design is redesign," then the ability to reframe, or break apart existing frames and reassemble new frames of understanding becomes a critically important activity within the context of design thinking.

2.2 Reframing

We define "reframing" as the disassembling of an existing problem solving frame set and reassembling a new frame set, possibly leading to a new problem solution. This definition has its roots in neuro-linguistic programming, a psychotherapeutic technique designed to change patient behaviors through reinterpretation of subjective experiences. [Bandler et al. 1982] This is also similar to work in

cognitive linguistics, particularly the work of George Lakoff, which examines the shifting public opinion on political issues through language that promotes reframing of issues. [Lakoff 2008]
We are also interested in the designer's propensity to reframe; their willingness to shift an existing problem-solving frame set based on the introduction of new information. In some cases, the new information can be important to shifting the problem-solving frame set, in which case reframing is an appropriate activity. In other cases, the new information may not be important and reframing the problem solving set may not lead to a better answer. The designer is responsible for deciding how this new information will (or will not) contribute to the reframing of the problem and solution. An individual's natural inclination to reframe (or not to reframe) may be tied to specific cognitive characteristics. We have extensively examined reframing behavior on this exercise from the cognitive perspective in an upcoming book chapter, which is based on the same data set. [Kress and Schar 2012]

3. Development of the measurement tool

3.1 Objective

The Stanford Design Thinking Exercise (STDE) aims to measure reframing at both the individual and group levels. Subjects express their frame through a series of explicit decisions or choices, by which they translate their mental frames into observable behavior. In this manner, the frame can be implicitly described as a unique collection of the decisions or choices that comprise a solution. This concept is known as "choice structuring" and has long been associated with organizational learning and strategic business planning. [Artiz and Walker 2010] Argyris defines "double loop learning" as a process in which prior choices are reconsidered in the context of the next choice. In a double loop learning choice structuring process, any disagreement will be made explicit in the group context and immediately corrected through the mutual modification of the group's objectives. [Argyris 1976] The STDE was designed to be a double loop learning choice structuring instrument that could simultaneously measure both an individual and group reframing of a potential problem solution.

3.2 Exercise approach

The instrument is designed as a multi-round exercise that requires choice structuring in both an individual and a group context. The prototype exercise was based on the topic of an "urban public share bicycle" in which subjects are asked to design a bicycle suitable for an urban environment to be shared by the general public. This topic was chosen because we expected it would be relatively familiar to all of our subjects but not particularly polarizing to any of them. We introduced the "public shared" notion because this was probably a less familiar concept, and would thereby stimulate discussion and, hopefully, innovative thinking.
The exercise begins with a short text introduction describing the roles of the subjects (as designers in a bicycle-manufacturing firm) and presenting the subject with 10 initial design choices. The subject is then asked to independently rank those design choices from 1 (most important) to 10 (least important) to the design solution they envision. This process of choice structuring within a case study has been used in prior research, typically involving survival scenarios (e.g. "Lost at Sea" or "Lunar Survival"). Artiz and Walker employed a similar case study based choice structuring process to study decision making within intercultural teams. [Artiz and Walker 2010]
The exercise is comprised of a series of four rounds where new information is introduced, providing the opportunity for new choice structuring (or reframing) activity to be observed and measured. The STDE has four rounds: the first three present the subjects with 20 items for choice structuring in total; the fourth round allows for additional choice structuring by letting subjects introduce their own items. A complete discussion of the STDE instrument and application can be found in Section 4: The Exercise in Application.

3.3 Item definition and validation

The importance ranking of items is meaningful to us when it comes to "tuning" the instrument. We can present items in a known order of importance to the subjects over subsequent rounds; items presented in ascending order of importance should correspond with maximal reframing of the solution

during these rounds, whereas items presented in decending order of importance should stimulate a minimal amount of reframing. This allows us to precisely quantify the magnitude of the reframing stimulus that we are giving the team, and so have a baseline to which to compare their reframing behavior. This stimulus measure also allows us to normalize results across rounds with different levels of stimuli. For a complete discussion of the tuning process, see Section 4.3: Tuning the Exercise.

The first step in developing items for subject ranking was to brainstorm a list of potential design choices. In order to stimulate choice activity on the instrument, it is necessary to provide items that are in some way meaningfully different or distinct from one another. To this end, we decided to include both object-oriented and user-oriented items. These reflect two important but different perspectives on the problem and also represent two distinct, implicit "strategies" toward creating a solution.

Object-oriented items are associated with particular physical aspects of the bicycle (e.g. "The frame is made from thin-walled high grade aluminum."). User-oriented items are associated with the user experience with the bicycle (e.g. "The bicycle is easy to ride"). We developed a total of 31 items and tested with an independent subject group of 30 individuals to validate their reliably falling into either object- or user-orientated categories. Items with less than 85% agreement within that category were discarded, resulting in a final validated set of 10 object-oriented and 10 user-oriented items. In further iterations, we also introduced several "ambiguous" items and provided an assortment of all three types to the subjects.

In order to determine the relative importance of each of these items, a second independent sample of 95 individuals was presented with the exercise and asked to rate each of the items in order of importance to the design solution on a five-point Likert scale. This was created as an online task via Amazon Mechanical Turk for 100 unique respondents, 5 of which were subsequently rejected for errors or anomalies in data entry. This resulted in a general consensus on the order of importance of each item within the context of the case study task. Statistical t-tests showed a significant difference for items separated by about 3 positions on the rating scale (e.g. the item ranked #7 is significantly different than the items ranked #10 and higher) at the 95% confidence level.

4. The exercise in application

4.1 Explanation of the exercise

The SDTE is a paper-based team design challenge. It is comprised of four rounds, each with an individual and a group component. Team members are prompted to role-play as a design team within a bicycle manufacturing company. They are tasked with designing an "urban public share bicycle" by selecting and ranking features, components and design elements from a list of items that is provided to them. In each of the following two rounds, new options are revealed to the group on additional lists and they are given the option to revise their previous ranking. For each round, subjects first prepare a ranking on their own, and then enter into discussion to reach a group consensus. In the fourth round, subjects are given the open option to create up to five new options (in contrast to the previous three rounds, where all information is provided in a closed format). In theory the exercise is not limited specifically to four rounds; so, in total, the combined individual and group rounds of the exercise provide $m(n+1)$ rankings that can be scored independently, where n represents the size of the subject group and m the number of rounds.

The four rounds of the experiment each contain new choices, creating the opportunity for the team to revise their previous solution (reframing). However, subjects are always limited to selecting a top ten, even as the item set grows beyond 20. This was intended to force some tough decision-making and also encourage more reframing. A complete breakdown of how the items are presented is given in Table 1.

Table 1. Items sets by round

Round	Set Name	New Options	Cumulative Options	Choice Set
1	Initial Set	10	10	10
2	Stimulus Set	5	15	10

3	Final Set	5	20	10
4	Open Set	5*	25*	10

*The group may create up to five new options, but are not required to create any.

4.2 Scoring the exercise

At the end of each round, each individual writes a rank ordering of ten items into a score sheet (see Figure 2). The entries on the score sheet reflect both the individual and group consensus rankings. Reframing is scored by tallying the total amount of change from one round to another, as measured by the number of rank differences in the score chart that a given item moved. Which rounds are compared is determined by what type of comparison is being made (e.g., looking at individual changes only, or group changes only, or a combination thereof). Taking for example the scorecard in Figure 2, we can score the amount of reframing that the participant did to match the group solution in Round 1. To do so, we compare the entries in the two columns of Round 1. Between the individual and group rounds, seven of the items (A–E, H & J) each moved one slot up or down (the direction is not recorded). Item G moved two slots, item I three slots, and item F four slots. These motions are summed to yield a total reframing score for this round given by 7(1) + 1(2) + 1(3) + 1(4) = 16, which can then serve as a basis for comparison to other rounds, other individuals and other teams. Where new items are introduced to the list (as in item M, Round 2), they are counted as if being "brought up from the bottom" of the ten-item list (i.e. slot 11); in this case item M would be scored 9 in Round 2, as it moved from "slot 11" to 2.

For our validation trials, we automated the scoring process by means of a MATLAB routine such that analyzing a given trial takes a few minutes at most (primarily for data entry). The simplicity of this scoring rubric allows for the complete reframing activity of all trials, at both the individual and group level, to be recorded in a straightforward spreadsheet format.

Round 1			Round 2			Round 3			Round 4		
Rank	Individual	Group	Rank	Individual	Group	Rank	Individual	Group	Rank	Individual	Group
1	A	F	1	F	M	1	M	M	1		
2	E	A	2	M	F	2	F	F	2		
3	C	E	3	A	A	3	A	Q	3		
4	I	C	4	I	N	4	Q	A	4		
5	F	D	5	C	E	5	N	N	5		
6	D	H	6	O	C	6	R	R	6		
7	H	I	7	K	L	7	E	E	7		
8	B	G	8	L	D	8	C	C	8		
9	J	B	9	N	H	9	T	L	9		
10	G	J	10	D	K	10	L	D	10		

Figure 2. One participant's scorecard for three subsequent rounds

Theoretical minimum and maximum scores are determined by the structure of the instrument. Minimum reframing scores for all rounds are zero (indicating no change). Maximum scores for reframing are 50 for group rounds (where no new information is introduced) and 65 for individual rounds (where 5 new items are introduced per round). Knowing this range allows us to place individuals and teams fall on a continuum of reframing behavior, and also allows us to normalize our results. The exercise can be scored in a number of ways:

Individual-to-Individual. Individuals exhibit different amounts of reframing from round to round, even within a team. Comparing reframing across individual rounds allows us to observe whether certain individuals have a propensity to reframe more than others.

Individual-to-Group. Individuals must alter their own solutions to different extents to match the group consensus. Often, a strong persona can dominate the discussion, resulting in a group consensus

that closely matches their own solution. This is apparent when individuals have very imbalanced amounts of reframing relative to the group consensus, which can be measured by comparing a group consensus round to the preceding individual round.

Group-to-Group. Different groups exhibit different amounts of reframing, even with a stimulus of constant magnitude. Groups can be compared to one another either by the mean of all team members' individual reframing, as the average overall reframing between group consensus rounds alone, or both.

Group-to-Consensus. Given importance rankings for all the items, we have some sense of what the general consensus solution would be. There is the expected outcome that at the end of each round, the ranking will reflect the general consensus. This would simply be the top ten ranked items of the given item set in descending order. Particular group consensus scores can be compared on their overall difference from this expected outcome.

Fourth Round. The fourth round of the exercise has an open-response component that allows for additional scoring opportunities such as counting the total number of ideas generated, the balance in contributions toward the consensus, etc.

4.3 Tuning the exercise

The goal of the exercise is to reveal a group's natural tendency to reframe. As such, it is necessary that the exercise allow for and encourage enough reframing so that teams predisposed to reframe will do so, but not so strong that to force reframing behavior on teams that would otherwise not do so. We would like to tune the exercise to set some medium level of expected reframing that will maximize our experimental variation. We can do this by using the item importance rankings to control in what order salient pieces of information are delivered to the team.

As an example, a team that receives the most important items in the last round is likely to exhibit significantly more reframing behavior in that round. The opposite is true if all the important items are given in the first round. Of course, there is very large number of permutations for how the 20 distinct items can be subdivided into three rounds (of 10, 5 & 5, respectively). The importance rankings allow us to measure how much reframing is expected for each of those permutations, which we call the magnitude of the stimulus. We have automated this process by means of a MATLAB routine that generates an appropriate permutation for the desired stimulus size.

5. Results & analysis

We conducted thirty trials of the exercise in two phases: pilot testing and validation. The first twelve trials were for pilot testing and refining the instrument. The remaining eighteen trials were conducted with the final version of the exercise tuned to two different stimuli magnitudes and were scored and analyzed according to all the methods listed in Section 4.2.

5.1 Pilot test results

The initial trials of the exercise were conducted with random groups of undergraduate and graduate students. These trials were tuned to minimum, maximum and several intermediate levels of reframing stimulus. Initially, we found that individuals and groups were both responsive to the size of the reframing stimulus in a fairly consistent pattern (see Figures 3 & 4). We also found that there was substantial variability in reframing behavior apparent between individuals and groups. This was an encouraging result, as our tuning of the stimulus appeared to affect the outcome as expected, and also that, regardless of stimulus size, there is sufficient variability from which to draw meaningful conclusions. Additionally, each group's end result was distinct, allowing a final comparison of the team's design solution against other teams', and against the general consensus.

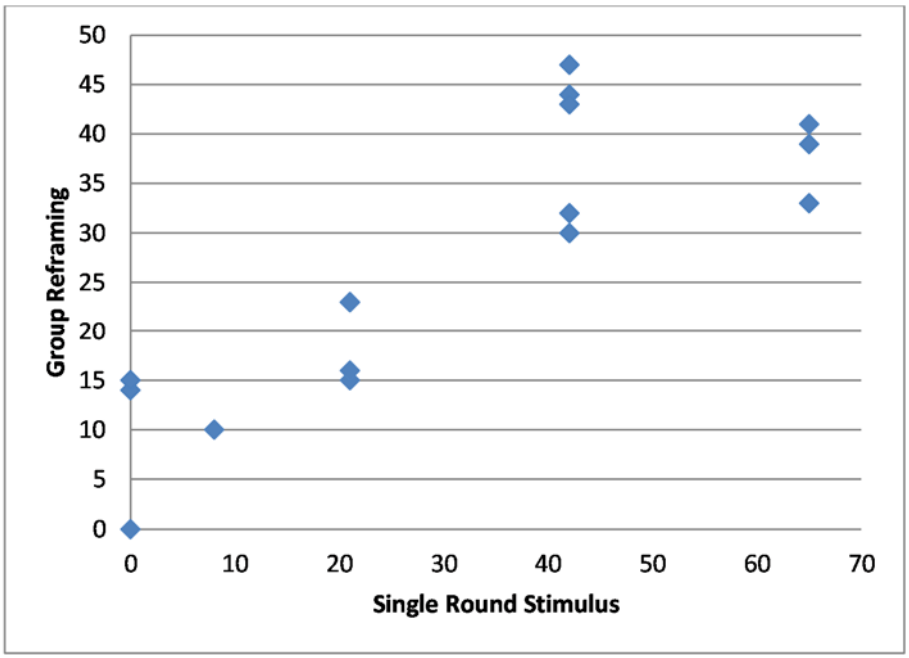

Figure 3. Group reframing as compared to stimulus size (computed as an average of individual reframing)

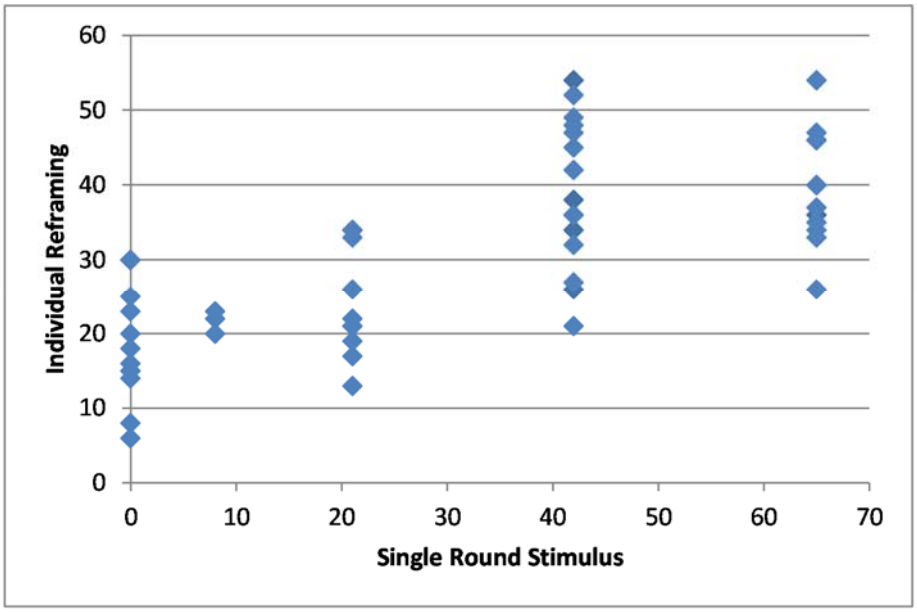

Figure 4. Individual reframing as compared to stimulus size

5.2 Validation trial results

We conducted eighteen validation trials of the exercise using a standard version tuned to a total reframing stimulus of either 63 (Case 1) or 84 (Case 2), which represents a difference in magnitude of about 20%. The trials were conducted with seventeen different teams of three individuals each and one team of four, or 55 subjects in total. There was a high degree of variance in reframing behavior at both the individual and group level, and this variance does not appear to be sufficiently explained by the difference in stimulus. The differences between individual- and group-level reframing behavior between the two cases were not found to be significant, though we did observe marginally higher reframing in Case 2 at both the individual and group levels (see Table 2). Though we expect to observe a normal distribution in reframing behavior, we were unable to confirm a normal distribution

in the individual or group reframing scores; this is likely due to insufficient sample size. This indicates that a 20% difference in stimulus is too small to be detected amidst the natural variation in reframing activity, or perhaps that the stimulus level is too high overall (encouraging "unnatural" reframing and creating a source of error).

Table 2. Validation trial results

Case	Level	Sample Size	Stimulus Size	Reframing Actions (Avg.)	Standard Deviation
1	Individual	30	63	63.2	11.0
2	Individual	25	84	64.4	13.3
1	Group	9	63	56.4	12.8
2	Group	8	84	62.2	9.9

5.3 Accuracy results

In addition to comparing reframing behavior between individuals and groups, we sought a measure of "accuracy" of the solution that would be representative of the quality of the team's design outcome. Although there is no "right answer" in this exercise, we define accuracy as the extent to which an individual's or group's solution conforms to the importance rankings of the items as had been determined previously. This is not an entirely robust measure, as there is no *a priori* reason to believe that multiple groups would arrive at the same design result, and the general consensus importance rankings do not represent one coherent design solution (but rather a high-level average). In theory, however, we would expect that any solution iterated on long enough, and with input from enough individuals, would conform to the general consensus; thus, we expect that solutions that have been subjected to more reframing activity may reflect the consensus more accurately. We did find a slight correlation between the individual reframing score (adjusted for stimulus size) and individual accuracy on the exercise ($r = 0.31$, $p = 0.02$, $n = 55$). However, we also found that the "consensus" of item importance when calculated from these eighteen trials did not match the consensus from prior item validation; this is either the result of too few trials, or perhaps a substantial difference in interpretation between the the initial online consensus group and experimental trial participants.

6. Conclusions & discussion

We were able to confirm that different individuals and teams exhibit different amounts of reframing behavior, and that this behavior is captured accurately by our scoring rubric. Though we had early success in tuning the stimulus size, we find that overall the natural variation in reframing behavior appears to be far greater than the 20% differential in stimulus size between the two cases of validation trials. It is possible that with the stimuli were too similar in size and that our sample size was too small to observe the broader statistical trend. However, the wide variability in reframing among individuals and teams is encouraging; this offers some support for the SDTE as a measure of difference. However, it does not support the hypothesis that reframing behavior is reliably controlled by tuning the instrument, and calls into question the tuning used for validation trials. The stimulus may have been high enough to trigger reframing behavior in those who may not be naturally inclined to do so, creating noise in the data.

Altneratively, though we did not measure this explicitly, the temporary nature of the team/task in the trial case may have resulted in low personal investment in the solution, and thus made individuals more open to revisions than they may have otherwise been. Future trials could use a lower level of stimulus and perhaps be revised somehow to increase personal investment in the solution. Furthermore, our validation trials included a mix of participants, some of whom were familiar with each beforehand and some of whom were not. This data was not completely recorded so we are not able to distinguish the cases in which preexisting interpersonal dynamics may have influenced behavior on the exercise. Future trials should take care to separate the case of preexisting teams, for whom long-term interpersonal dynamics have had time to evolve, from the case of participants who are unfamiliar with one another. At the very least, the level of familiarity between participants should be recorded as a part of the trial data; the scoring rubric could then be used to search for differences in reframing between familiar and unfamiliar participants (even within the same team). This might be

interesting to see if, for example, participants who are familiar with one another are more willing to reframe toward one another's solutions.

Since our validation trial only examined one topic and one set of design choices, it was not possible to run the same participants and teams through the exercise multiple times. This would have been very helpful information to determine if reframing behavior on the exercise is stable over time (or, for example, if teams assembled for the second or third time behave significantly differently). If the willingness to reframe is in some sense a trait of the individual, then we would expect to see relatively little change over time (though development is possible, and this may also be worth observing, particularly in an educational context). Without this data, we cannot positively conclude that behavior on the SDTE is a meaningful "snapshot" that can be extrapolated to activity outside the lab and over time. There is the potential to develop alternate versions of the exercise that present new topics and design choices such that the same individuals and teams can be run through the exercise multiple times. This would also create the possibility of "rotating" individuals through trial teams to track their reframing behavior in different group contexts. Given the structure of the exercise, developing alternate versions would not be that difficult and would just require a brief brainstorming and item validation phase. Alternatively, we could develop an observational framework to record team reframing in the field, and establish a connection between these observed dynamics and performance on the exercise.

What remains to be shown, even if the SDTE proves to be a reliable measure of difference between teams, is that this difference is meaningful for long-term team performance. Our first attempt to reconcile reframing and performance was by comparing reframing scores to performance on the "Reading the Mind in the Eyes" test, which has been shown previously to be a reliable indicator of social sensitivity (a positive correlate with team performance). [Woolley, 2010] We expected to find a slight positive correlation between reframing behavior and social sensitivity, which would have suggested that these two are both measures of an underlying individual characteristic that is important for team performance. However, we in fact found a slight negative correlation between these scores ($r = -0.32$, $p = 0.085$, $n = 30$); upon visual inspection, it appears that reframing behavior is independent of social sensitivity. So, while this does not offer support for reframing as a predictor of performance, it at least suggests that it is a distinct measure from social sensitivity. As such, there is no support for the hypothesis that reframing behavior and social sensitivity are expressions of the same underlying characteristic, or that they would tend to influence performance in the same way.

Applying this exercise in the context of a longitudinal study would allow us to explore the link between short-term reframing behavior on the exercise and long-term design effectiveness. Until this link is shown, there will not be much of a compelling reason to adopt the SDTE in place of other methods. The exception is for applications in which reframing capability is explicitly part of the curriculum (for example in design). However, we believe that the theoretical link between design activity and reframing behavior is strong enough to lend credibility to the broader hypothesis, and that research into the means to measure and observe design team reframing is worth pursuing further.

Acknowledgments

We would like to acknowledge our research assistants J. D. Benner & A. H. Hoster for their contribution to performing experimental trials, the guidance of our P.I. Mark Cutkosky, as well as the generous support of the Hasso Plattner Design Thinking Research Program (HPDTRP) in this endeavor.

References

Argyris C., 1976, "Single-loop and double-loop models in research on decision making," Administrative Science Quarterly, 21(3), pp. 363–375.

Artiz J., and Walker R. C., 2010, "Cognitive Organization and Identity Maintenance in Multicultural Teams: A Discourse Analysis of Decision-Making Meetings," Journal of Business Communication, 47(1), pp. 20-41.

Bandler R., Grinder J., and Andreas S., 1982, Neuro-Linguistic Programming and the Transformation of Meaning, Real People Press, Moab, Utah.

Dubberly H., 2005, "How Do You Design? A Compendium of Models" Dubberly Design Office, 2005.

Goffman E., 1974, Frame Analysis: An Essay on the Organization of Experience, Harvard University Press.

Kress G., Schar M. F., (Pending publication 2012), "Applied Teamology: The Impact of Cognitive Style Diversity on Problem Reframing and Product Redesign Within Design Teams" Design Thinking Research: Studying Co-Creation in Practice (Series). Eds. Hasso Plattner, Christophe Meinel and Larry Leifer. Springer, 2012.

Lakoff G. P., 2008, The Political Mind: Why You Can't Understand 21st-Century American Politics with an 18th-Century Brain, Viking Adult.

Latour B., 2008, "A Cautious Prometheus? A Few Steps Toward a Philosophy of Design (with Special Attention to Peter Sloterdijk," Networks of Design: Proceedings of the 2008 Annual International Conference of the Design History Society (UK) University College, Falmouth UK, p. 13.

Woolley A.W. et al., 2010, "Evidence for a Collective Intelligence Factor in the Performance of Human Groups." Science. 2010 Oct 1; 6.

Gregory Kress
Ph.D. Candidate
Mechanical Engineering, Center for Design Research
424 Panama Mall (Bldg. 560)
Stanford, CA, 94305 USA
Email: glkress@stanford.edu