Proceedings of the ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC2020
August 16 – 19, 2020 St. Louis, Missouri

DETC2020-19435

BEST FITS AND DARK HORSES: CAN DESIGN TEAMS TELL THE DIFFERENCE?

Daniel Henderson
Pennsylvania State University,
University Park, PA

Thomas Booth
Pennsylvania State University,
University Park, PA

Kathryn Jablokow
Pennsylvania State University,
Malvern, PA

Neeraj Sonalkar
Stanford University
Stanford, CA

ABSTRACT

Design teams are often asked to produce solutions of a certain type in response to design challenges. Depending on the circumstances, they may be tasked with generating a solution that clearly follows the given specifications and constraints of a problem (i.e., a Best Fit solution), or they may be encouraged to provide a higher risk solution that challenges those constraints, but offers other potential rewards (i.e., a Dark Horse solution). In the current research, we investigate: what happens when design teams are asked to generate solutions of *both* types at the same time? How does this request for dual and conflicting modes of thinking impact a team's design solutions? In addition, as concept generation proceeds, are design teams able to discern which solution fits best in each category? Rarely, in design research, do we prompt design teams for "normal" designs or ask them to think about both types of solutions (boundary preserving *and* boundary challenging) at the same time. This leaves us with the additional question: can design teams tell the difference between Best Fit solutions and Dark Horse solutions?

In this paper, we present the results of an exploratory study with 17 design teams from five different organizations. Each team was asked to generate both a Best Fit solution and a Dark Horse solution in response to the same design prompt. We analyzed these solutions using rubrics based on familiar design metrics (feasibility, usefulness, and novelty) to investigate their characteristics. Our assumption was that teams' Dark Horse solutions would be more novel, less feasible, but equally useful when compared with their Best Fit solutions. Our analysis revealed statistically significant results showing that teams generally produced Best Fit solutions that were more useful (met client needs) than Dark Horse solutions, and Dark Horse solutions that were more novel than Best Fit solutions. When looking at each team individually, however, we found that Dark Horse concepts were not *always* more novel than Best Fit concepts for every team, despite the general trend in that direction. Some teams created equally novel Best Fit and Dark Horse solutions, and a few teams generated Best Fit solutions that were more novel than their Dark Horse solutions. In terms of feasibility, Best Fit and Dark Horse solutions did not show significant differences. These findings have implications for both design educators and design practitioners as they frame design prompts and tasks for their teams of interest.

Keywords: design teams, education, design theory

INTRODUCTION

As design practitioners, we want to influence the solutions that designers generate through thoughtful prompting and direction. Similarly, as design educators, we want to teach our students how to interpret and respond to design specifications responsibly and mindfully. With these aims in mind, we note that the ability to discern one type of design concept from another in terms of a specific design characteristic (say, novelty or feasibility) will impact how designers generate solutions in response to a given set of constraints and specifications. If the instructions for a design task prompt a design team to strive for (say) the most novel solution, we expect the designers can discern differences in novelty among their design ideas so they can choose the one that best meets this criterion. But is this the case? By analyzing design outcomes, this paper will investigate how well design teams can distinguish between their design concepts in terms of two general solution categories with very different expectations—i.e., Best Fit and Dark Horse solutions.

Ideation and prototyping for so-called Dark Horse (high risk, high reward) solutions have been explored in students [1,2] as a means of promoting radical, innovative ideas in the design process. By encouraging the exploration of these risky

FIG. 1: DESIGN TEAM EFFECTIVENESS WORKSHOP FLOW

solutions, designers can expand their design space to new, unexplored areas. This contrasts with Best Fit (i.e., standard- and constraint-consistent) ideation and prototyping, which promote more traditional, incremental solutions. It is interesting to note that for the past few decades in the design research world, the dominant paradigm has been focused on the importance of engaging designers in Dark Horse-like ideation to "challenge boundaries" or to "be creative" [3,4] rather than Best Fit-like ideation that encourages careful attention to problem constraints. Rarely, if ever, do we explicitly admonish research study participants to "color *inside* the lines" and then closely examine the qualities of the design solutions that follow. While prompting designers to generate high risk/high reward solutions and studying the results is certainly important for many reasons, it raises the question of whether designers at any level of expertise are tacitly aware of the nature of their design solutions with respect to risk and reward as they carry out the design process. What's more, what happens when designers are asked to provide solutions of these two different types at the same time? Will the juxtaposition of the contrasting prompts help them generate solutions that are markedly different in the expected ways, or will their design concepts become muddied by the need to consider conflicting dual design prompts simultaneously?

To address these questions, we examine here how design outcomes differ when teams are asked to generate *both* Best Fit (BF) and Dark Horse (DH) conceptual prototypes in the solution of the same design task. Specifically, a BF concept was defined as "the best chance for fulfilling the design requirements," while a DH concept was defined as "crazy, but if successful, could revolutionize" the given design context. Evaluating these solutions through the lenses of feasibility, usefulness, and novelty, we investigate whether design teams, given little instructional scaffolding, display an implicit understanding of what we expect in a design concept that is Best Fit or Dark Horse when they consider the generation of both at the same time.

PROJECT CONTEXT

The work presented here is part of an NSF-funded effort in which we are mapping the individual characteristics of design team members and their interactions to the team's design performance to identify the behavioral building blocks of high-performance design teams (HPDTs). The identification of such behavioral building blocks will lead to scientific cognitive-behavioral models of design teams that are applicable in academic and industry environments, as well as new tools for improving the effectiveness of those teams. In that context, our aim is to identify and map the behavioral building blocks of HPDTs through two research objectives: (1) Identify the behavioral interaction sequences and individual characteristics that characterize high performance design teams; and (2) map those sequences and characteristics to team design outcomes.

The project utilizes a unique team interaction measurement system called the Interaction Dynamics Notation (IDN) [5] to characterize interaction behaviors between individuals on a team, as well as the Kirton Adaption-Innovation inventory (KAI®) [6,7] to measure the cognitive styles of those team members. Team outcomes are measured in two ways: (1) the conceptual prototypes delivered by each team are analyzed in terms of various design metrics (e.g., feasibility, novelty); and (2) team members' reflections on their design performance are recorded via a debrief survey. Other designer characteristics include demographics, such as gender and age. Thus far, data have been collected from 31 design teams across academia, industry, and the military. Initial results from our analyses of team interaction behaviors, individual characteristics, and team outcomes from a sample of the data can be found in [8,9]; the current paper focuses on 17 teams who participated in a series of design team effectiveness workshops (see Fig. 1) in Year 2 of our project.

RESEARCH QUESTIONS AND RELATED WORK

In the context of our research objectives, the following research questions were explored:

RQ1: In terms of *feasibility,* how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?
RQ2: In terms of *usefulness,* how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?
RQ3: In terms of *novelty,* how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?

We assume that DH concepts would be less feasible and more novel than BF concepts, while achieving equal usefulness. Similar studies have explored relationships between the perceived risk of a concept and the likelihood of a team to select that concept [10-12], as well as the evolution of ideas during the design process through the lenses of common design metrics [13,14]. In particular, these studies showed that teams

tended to prioritize feasibility over creativity/novelty when selecting a final concept, seemingly defaulting to what we call a Best Fit solution. If asked specifically to generate a Dark Horse solution alongside a Best Fit solution, we assume they will be able to overcome those tendencies. Other studies have used problem framing to present a design challenge in multiple ways [15,16], but did not require participants to complete conflicting prompts for the same task. This study is unique in that we assign a single design task to a team, require both high risk (Dark Horse) and low risk (Best Fit) solutions, and examine the differences between those outcomes with familiar design metrics.

METHODOLOGY

To address our research questions, the Best Fit and Dark Horse design concepts collected from 17 design teams across five organizations were examined. These design concepts were generated as part of a half-day workshop focused on design team effectiveness (see Fig. 1 for the workshop flow). The remainder of this section provides details on the study design.

Participants

In total, 64 participants (51 males, 13 females) took part in this study; they were arranged in 17 teams of 3-5 members each to complete the workshop tasks. The participants represented a diverse array of students, faculty, and working professionals, all of whom had similar levels of mechanical design experience. We examined seven teams of defense instructors, three teams from a communications company, one team from an energy startup, one team from a mid-sized manufacturing company, and five NSF I-Corps® teams from a Midwestern university.

Data Collection

Participants in our workshop completed the KAI® cognitive style inventory [6,7] prior to the workshop, and also signed an informed consent form. The workshop began with a brief introduction to team effectiveness principles and the Interaction Dynamics Notation (IDN), after which the teams completed a one-hour design challenge called the Lifting Water Design Challenge (LWDC) (see Table 1).

The LWDC asks participants to design a solution for rural farmers in Myanmar to lift water from at least 50ft below ground to ground level, use only human power, cost less than US $50, and work with existing tube wells in the ground that are 2 inches in diameter. Each team received a packet of supplemental information to aid them in understanding the LWDC context. These documents included a set of slides describing what life is like for rural farmers in Myanmar, an ad for the manually-operated Red Rhino pump (cost: US $36), and an explanation of the physical limit (33.9 ft) that water can be pumped under a total vacuum. The teams could consult the internet for basic facts and ask questions of the workshop organizers (project PIs).

TABLE 1: DESIGN CHALLENGE DETAILS

Design Challenge
- Must lift water at least 50 ft - Must have sufficient discharge for farming use - Should be easy and intuitive to use and repair - Must be affordable (≤ US$50) - Must utilize 2-inch diameter tube wells already in the ground - Must use human power only
Cultural and Contextual Details
- Farmers each own small plots of land. - Farmers currently have no means to access water during the dry season and cannot be expected to do so otherwise. - Farmers currently do not have affordable pumps for aquifers beyond 30 ft deep. - Farmers must travel to cities in search of work while they cannot farm during the dry season. - Farmers want to increase the number of crop cycles (to >1). - Farmers currently earn less than one dollar per day. - Farmers desire to improve the quality of the crops (larger fruits, higher grade flowers). - Farmers desire to increase the acreage farmed (farmers can maximize every corner of their farm). - Farmers want to take advantage of the higher prices of produce in the dry season.

The LWDC activity was video recorded. Each team was asked to develop two solutions to the given challenge: (1) a Best Fit solution (i.e., one that "your team believes has the best chance of fulfilling the design requirements"); and (2) a Dark Horse solution (i.e., one that "your team believes is crazy, but if successful, could revolutionize irrigation in Myanmar"). These definitions were the only specific guidelines teams received regarding Best Fit and Dark Horse solutions, and very few questions arose about them. The teams were free to use any strategy to generate and select concepts. They were required to create deliverables of their final concepts, including sketches and low-fidelity prototypes from craft materials. At the conclusion of the one-hour challenge, each team gave a spoken presentation to their camera.

Afterwards, the participants completed a brief survey to collect their perceptions of the team experience and their design results. Each participant responded individually to the survey questions. Following this debrief, the teams met as a large group to receive feedback on their KAI (cognitive style) results. The workshop concluded with a discussion on integrating the day's learnings into their future team activities.

Assessing the Design Prototypes

We assessed the sketched, physical, and oral depictions of each design prototype using three design outcomes metrics (feasibility, usefulness, and novelty) with rubrics based on the LWDC. Feasibility, usefulness, and novelty have all been explored and used by many researchers [17]; we found that these metrics led us to holistic, complete impressions of the conceptual prototypes. Our rubrics also contain links to other common design criteria, such as manufacturability, acceptability, implementability, and effectiveness. Two coders independently evaluated each concept for all the teams in the

study and then established a consensus assessment that resolved any disagreements between their independent results [18]. The coders were both experienced design researchers with advanced engineering degrees. The guidelines and methodology used for each design outcome assessment are described in detail in the following sections.

1. Feasibility To assess feasibility, we established a list of specific requirements related to the broad question: *Does the design concept work technically and physically?* We derived these feasibility requirements (see Table 2) from the documents provided to the teams for the workshop (see Data Collection). Requirement F1 is general (i.e., whether or not the scientific principles used are sound), while the other three requirements are specific to the LWDC (e.g., the design concept lifts water 50 ft.). In our view, feasibility necessitates context-specific guidelines, as general requirements (e.g., F1) alone cannot accurately portray how feasible a concept may be for a task as specific as the LWDC.

TABLE 2: FEASIBILITY ASSESSMENT GUIDELINES

Feasibility Requirements
Yes or No:
F1. The principles utilized in the design concept are scientifically sound. • When design concepts rely on multiple scientific principles, all of the principles must be legitimate in order to satisfy F1.
F2. The design concept successfully lifts water 50 feet. • This requirement distinguishes between concepts that can bring water completely to the surface versus those that may only be able to lift water partially (or not at all). • While concepts that satisfy F2 can reasonably also be expected to satisfy F1, there are possible (and observed) cases where this is not the case. o E.g. if a concept utilizes a redundant scientific principle that is not sound, but has other sufficient scientific principles that can lift water 50 ft, it can satisfy F2 without satisfying F1.
F3. The design concept utilizes 2-inch diameter tube wells already in the ground. • This requirement primarily pertains to whether or not the existing infrastructure for obtaining water would still be in use. • Solutions unable to utilize the existing infrastructure are less feasible.
F4. The design concept can be operated solely with human power. • Design concepts that necessitate electricity, petroleum fuels, or animal power are considered less feasible.

2. Usefulness The usefulness requirements shown in Table 3 were developed in a fashion similar to those for feasibility to answer the question: *Does the design concept meet the needs of the client?* Like feasibility, one requirement (U1) is more general than the others. Requirements U2 through U4 pertain to specific details of the LWDC, while requirement U1 indicates that for a design concept to be useful, it should be appropriate generally within the context of the client's culture.

TABLE 3: USEFULNESS ASSESSMENT GUIDELINES

Usefulness Requirements
Yes or No:
U1. The design concept is contextually appropriate for the existing farms in rural Myanmar. • Concepts that majorly change the nature of rural Myanmar farming via restructure of society/land/resources are considered less useful for the clients.
U2. The design concept is affordable (<US $50). • Based on available cost estimates, some scenarios of excessive cost (>US $50) would include: o Uses two Red Rhino pumps (US $36 each) o Uses equipment that requires specialized manufacturing o Requires significant labor
U3. The design is easy and intuitive to use and repair. • This requirement considers two related aspects of the design concept, both of which must be satisfied: o The concept does not require specialized knowledge in order to operate it or diagnose problems. o When a component fails, it is easy to replace.
U4. There is sufficient discharge for farming use. • This requirement presupposes that the concept is feasible enough to work physically, so it is distinct from the feasibility assessment • Some concepts as described would deliver a relatively small volume of water. A significant volume of water is needed to be useful for farmers.

3. Novelty We chose to evaluate novelty using a rarity-based approach. An assortment of novelty assessments are described in the literature [19-24]; some note that measuring rarity can be more straightforward than novelty [19,24]. In this work, we developed an integrated method for assessing rarity by adapting existing methods to meet our needs in evaluating the given data. We began with Shah's framework for variety [23] and utilized a sorting approach similar to Linsey's [20] to group like ideas together at each of Shah's functional principle levels. First, two coders independently assessed the physical principles (PP), working principles (WP), and embodiment principles (EP) [23] of the concepts in our data set, grouping together like concepts at each level. We did not include the "detail" level of Shah's framework [23] in our assessment, since at that minute level of inspection, all concepts tend to appear distinct from one another. We are primarily interested in the more meaningful distinctions that arise at the PP, WP, and EP levels. Our evaluation of rarity is based only on the final concepts selected and identified as BF or DH by each team. If a team selected a concept that was generated and discarded by other teams during the workshop, it was still considered rare for that concept *to be selected*.

The rarity of a concept at each level was calculated as the ratio of the number of concepts grouped at that level to the total number of concepts in the data set. For instance, if four concepts were determined to exhibit the same physical principle within a data set of twenty concepts, the rarity score for those four concepts would be 4÷20 or 0.2. Using this method, lower numbers indicate rarer concepts. Some methods of measuring novelty prescribe weightings and combinations of scores [19-21,23], but we opted to keep the scores unweighted

and separate for each level. Thus, each concept receives three rarity scores—one for each level. A detailed description of the rarity evaluation at each level is provided below.

At the physical principle level, the coders created categories based on the physical processes presented in the concepts. The physical principles that appeared more than once in our data set are shown in Table 4. Many concepts relied upon a single physical principle, but some appeared to combine multiple principles. We chose to define the combination of two physical principles as its own category of physical principle, rather than splitting a concept across two component categories. Concepts differing at the physical principle level are quite distinct from one another; the scientific underpinnings of each category can vary widely.

TABLE 4: PHYSICAL PRINCIPLES APPEARING MORE THAN ONCE IN THE DATA SET

Physical Principles
Negative Pressure (suction, often using pumps)
Positive Pressure (creating pressure to bring water up)
Manual Drawing (a physical mechanism lifts the water)
Negative Pressure + manual drawing

At the working principle level, the coders determined how the physical principles were being implemented differently between concepts. Table 5 shows physical principles and the associated working principles identified in our data set. Concepts differing at the working principle level often still display noticeable distinctions from one another. Even when the scientific underpinnings of a physical principle are shared among concepts, the concepts themselves can use the principle in different ways.

TABLE 5: WORKING PRINCIPLES (WP) AND ASSOCIATED PHYSICAL PRINCIPLES

Physical Principle: Negative Pressure	
WP1: Two stages of pumping	WP2: Solar-powered single-stage pumps
Physical Principle: Positive Pressure	
WP1: Pressurize the water table	WP2: Use pressure and check valves to sequentially draw up water
Physical Principle: Manual Drawing	
WP1: Cups	WP2: Archimedes Screw
Physical Principle: Negative Pressure + manual drawing	
WP1: Two stages: pumping + manual mechanism	

Finally, at the embodiment principle level, the coders examined the finer differences among concepts that share a working principle. These concepts are still similar to one other, but they may differ in key features. For instance, for the "Two Stages of Pumping" working principle, some concepts required two pumps (one at ground level, and one at a platform below ground level), whereas others indicated that one pump with various valves, switches, and hoses was sufficient.

DATA ANALYSIS

Tables 6 through 8 display sample results of applying the feasibility, usefulness, and rarity assessments, respectively, to the design concepts, organized with respect to Best Fit (BF) and Dark Horse (DH). In Tables 6 and 7, the requirements (F1-F4 and U1-U4, respectively) are each given a value of 1 (requirement met) or 0 (requirement not met). In Table 6, the rightmost column (Fs) shows the total Feasibility Score for each design concept as the sum of the values (1,0) for each of the requirements F1, F2, F3 and F4. Similarly, in Table 7, the rightmost column (Us) shows the total Usefulness Score for each design concept as the sum of the values (1,0) for each of the requirements U1, U2, U3 and U4.

TABLE 6: FEASIBILITY RUBRIC RESULTS

Team	*Feasibility Requirement Met*									
	Best Fit					*Dark Horse*				
	F1	**F2**	**F3**	**F4**	**Fs**	**F1**	**F2**	**F3**	**F4**	**Fs**
1	0	0	0	1	1	0	0	1	0	1
2	1	1	0	1	3	1	1	0	1	3

TABLE 7: USEFULNESS RUBRIC RESULTS

Team	*Usefulness Requirement Met*									
	Best Fit					*Dark Horse*				
	U1	**U2**	**U3**	**U4**	**Us**	**U1**	**U2**	**U3**	**U4**	**Us**
1	1	1	1	1	4	1	0	1	0	2
2	1	0	1	1	3	1	0	1	1	3

It is interesting to note that for *all* 17 teams, the total usefulness of their BF concepts was higher (in most cases) or equal to the total usefulness for their DH concepts. (one team did not complete a DH concept, so they are excluded from this observation.) A similar observation can be made for feasibility, with the exception of one team who produced a DH concept that was more feasible than their BF concept.

Finally, Table 8 shows sample rarity scores for the three components of rarity: Physical Principle (PP), Working Principle (WP), and Embodiment Principle (EP). Within our data set, all concepts exhibited equal or rarer values at each subsequent level; we did not observe any cases where a concept was rarer at its PP level than at its WP or EP level. Again, for our rubric, a lower numeric value indicates a rarer concept. A rarity score of 0.03 (1 concept ÷ 33 total concepts) indicates that the concept was considered to be different from *every* other concept at that level. We refer to this concept as "totally rare" for that level.

TABLE 8: RESULTS FOR THE THREE COMPONENTS OF RARITY (PP, WP, EP)

Team	*PP Value*		*WP Value*		*EP Value*	
	BF	*DH*	*BF*	*DH*	*BF*	*DH*
1	0.33	0.03	0.21	0.03	0.12	0.03
2	0.15	0.33	0.15	0.21	0.06	0.09

The concepts determined to be totally rare at the PP level were also found to be totally rare at the WP and EP levels. For example, Team 1's DH concept utilized "evaporation and condensation" as its physical principle; no other concepts utilized evaporation and condensation. Its working principle and embodiment principle were similarly not found in any other concept. This situation can be described as a concept being totally rare at all three levels.

RESULTS

We analyzed the collected data using a combination of descriptive techniques and t-tests; the significance threshold was set at .05. Results of our analyses for each Research Question (RQ), beginning with feasibility, are presented in the remainder of this section.

RQ1: In terms of feasibility, how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?

Fig. 2 shows the distribution of DH and BF concept feasibility scores for the 17 teams. The mean DH score for the dataset was 1.875 (for 16 teams, since one team did not submit a DH concept), and the mean BF score was 2.29. The difference in the means of the two distributions was not statistically significant, which suggests that the DH concepts were equally feasible when compared to the BF concepts.

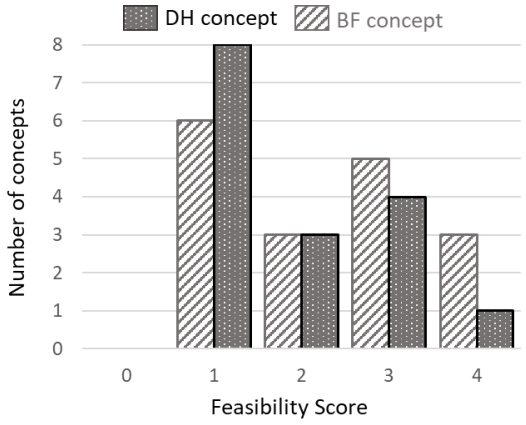

FIG. 2: DISTRIBUTION OF DH AND BF FEASIBILITY SCORES IN THE DATASET

If we consider how the BF and DH feasibility scores differed *within* each team and not across the entire dataset, we find that 68.75% of the teams (11 of 16; one team was again omitted for this analysis) had zero difference in the feasibility scores of their BF and DH concepts (see Fig. 3). This finding is interesting, since the LWDC prompt encouraged participants to consider DH solutions that might not be feasible now, whereas they were encouraged to create immediately feasible BF solutions. Based on our results, it appears that this prompting did not shift the participants' thinking toward higher risk solutions in terms of design feasibility when generating DH concepts, nor did it shift their thinking to lower risk solutions when generating BF concepts.

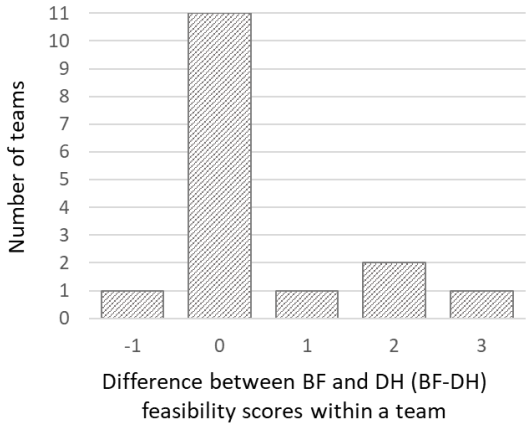

FIG. 3: DISTRIBUTION OF DIFFERENCES BETWEEN BF AND DH (BF - DH) FEASIBILITY SCORES WITHIN A TEAM

RQ2: In terms of usefulness, how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?

Fig. 4 shows the distribution of DH and BF concept usefulness scores for the 17 teams. The mean DH score for the dataset was 2.25 (for 16 teams, since one team did not submit a DH concept), and the mean BF score was 3.05. The difference in the means of the two distributions is statistically significant ($p=0.01$) for a standard two-tailed t-test assuming unequal variances. In other words: when considering whether the client's needs were met (usefulness), the BF and DH solutions were distinct, with the DH concepts exhibiting a lower mean usefulness score than the BF concepts across the data set. Thus, it appears that teams implicitly focused more on moving away from the client's needs (usefulness) when it came to DH solutions than they did on challenging technical feasibility.

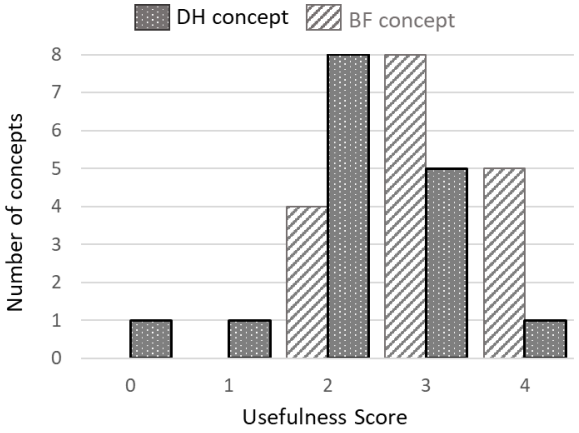

FIG. 4: DISTRIBUTION OF DH AND BF USEFULNESS SCORES IN THE DATASET

If we consider how the usefulness scores differed *within* each team and not across the entire dataset, we find that 50% of the teams (8 of 16—one team was again omitted for this analysis) had zero difference in the usefulness scores of their BF and DH concepts (see Fig. 5).

When looking at the mean scores for both feasibility *and* usefulness, we observe that the mean usefulness scores (DH = 2.25, BF = 3.05) were higher than the mean feasibility scores (DH = 1.875, BF = 2.29) across the data set. It is possible the teams generally prioritized useful concepts over ones that were feasible for the LWDC, but it is also possible that the usefulness requirements were generally more achievable than the feasibility requirements.

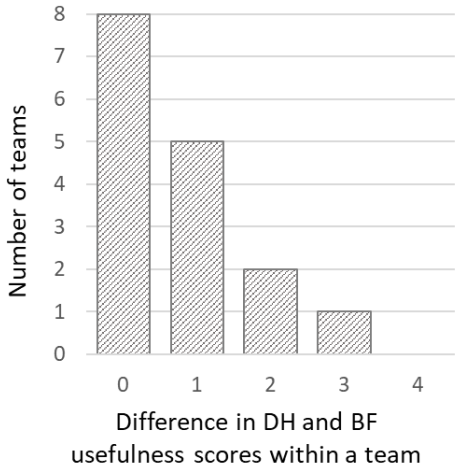

FIG. 5: DISTRIBUTION OF DIFFERENCES BETWEEN BF AND DH (BF - DH) USEFULNESS SCORES WITHIN A TEAM

RQ3: In terms of novelty, how do designers' Dark Horse (DH) concepts differ from their Best Fit (BF) concepts?

The analysis of novelty was operationalized in terms of measuring the rarity or the frequency of occurrence of concepts in the full dataset of 33 concepts. As explained previously, we deconstructed each concept into Physical Principle, Working Principle, and Embodiment Principle elements to compare concepts at each level. As a result, Research Question 3 (RQ3) was elaborated as follows:

- **RQ3a**: How do DH concepts differ from BF concepts in terms of their *rarity* at the Physical Principle level?
- **RQ3b**: How do DH concepts differ from BF concepts in terms of their *rarity* at the Working Principle level?
- **RQ3c**: How do DH concepts differ from BF concepts in terms of their *rarity* at the Embodiment Principle level?
- **RQ3d**: How do DH concepts differ from BF concepts in terms of their *rarity* differences across all three levels?

RQ3a: Physical Principle Level Fig. 6 shows the distribution of DH and BF concept Physical Principle (PP) rarity scores for the 17 teams. The mean PP rarity score for the DH concepts in the dataset was 0.14 (for 16 teams) and the mean PP rarity score for the BF concepts was 0.21. The difference in the means of the two distributions is statistically significant (p=0.048) for a standard two-tailed t-test assuming unequal variances. This shows that the DH concepts had an overall lower PP rarity score than the BF concepts, implying that the DH concepts were rarer overall (i.e., more novel) at the Physical Principle level than the BF concepts.

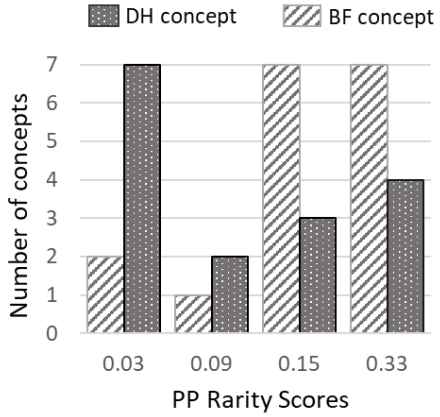

FIG. 6: DISTRIBUTION OF DH AND BF PP RARITY SCORES IN THE DATASET

If we consider how the PP rarity scores differed *within* each team and not across the entire dataset, we find that 31.25% of the teams (5 of 16—Team 9 was omitted for this analysis) had zero difference in the PP rarity scores of their BF and DH concepts (see Fig. 7); 50% of the teams (8 of 16) had a rarer DH concept than a BF concept, and the final 18.75% of teams (3 of 16) had a DH concept that was less rare than their BF concept. It appears that only half of the design teams in our sample successfully emphasized rarity (i.e., novelty) in their DH solutions at the Physical Principle level.

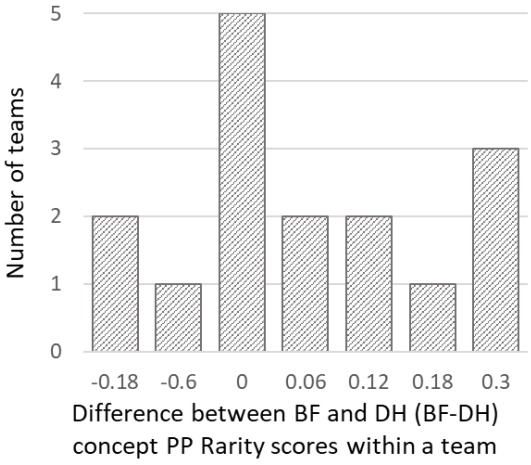

FIG. 7: DISTRIBUTION OF DIFFERENCES BETWEEN BF AND DH (BF - DH) PP RARITY SCORES WITHIN A TEAM

RQ3b: Working Principle Level Fig. 8 shows the distribution of DH and BF concept Working Principle (WP) rarity scores for the 17 teams. The mean WP rarity score for the DH concepts in the dataset was 0.07 (for 16 teams), and the mean WP rarity score for the BF concepts was 0.13. The difference in the means of the two distributions is statistically significant (p=0.0063) for a standard two-tailed t-test assuming unequal variances. This shows that the DH concepts had an overall lower WP rarity score than the BF concepts, implying that the DH concepts were rarer overall (i.e., more novel) at the Working Principle level than the BF concepts.

If we consider how the WP rarity scores differed *within* each team and not across the entire dataset, we find that the teams were clustered at the two extremes. Specifically, 25% of the teams (4 of 16) had zero difference in the WP rarity scores of their BF and DH concepts, while 62.5% of teams (10 of 16) had a rarer DH concept, and 12.5% of teams (2 of 16) had a rarer BF concept (see Fig. 9). Based on these results, the differences in rarity between the DH and BF concepts is even more pronounced at the WP level than the PP level.

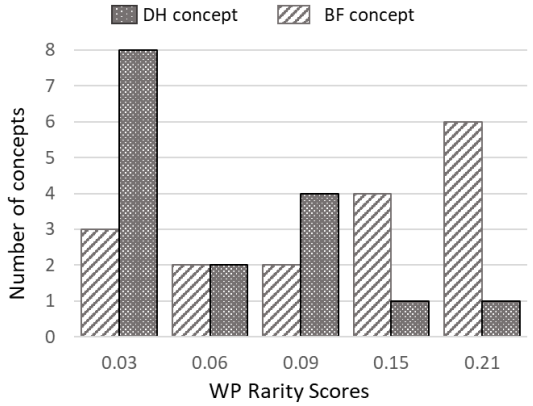

FIG. 8: DISTRIBUTION OF DH AND BF WP RARITY SCORES IN THE DATASET

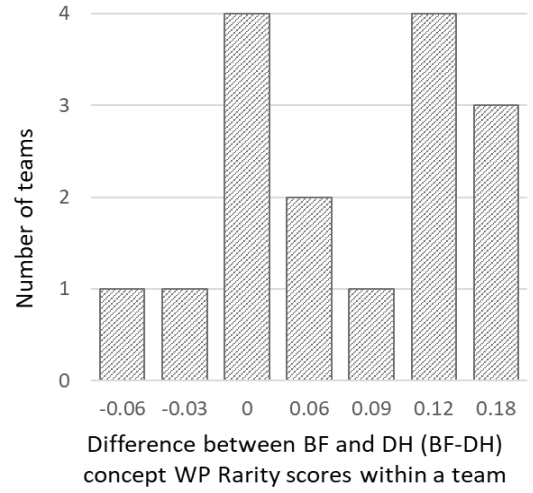

FIG. 9: DISTRIBUTION OF DIFFERENCES BETWEEN BF AND DH (BF - DH) WP RARITY SCORES WITHIN A TEAM

RQ3c: Embodiment Principle Level Fig. 10 shows the distribution of DH and BF concept Embodiment Principle (EP) rarity scores for the 17 teams. The mean EP rarity score for the DH concepts in the dataset was 0.035 (for 16 teams), and the mean EP rarity score for the BF concepts was 0.064. The difference in the means of the two distributions is statistically significant (p=0.0056) for a standard two-tailed t-test assuming unequal variances. This shows that the DH concepts had an overall lower EP rarity score than the BF concepts, implying that the DH concepts were rarer overall (more novel) at the Embodiment Principle level than the BF concepts.

If we consider how the EP rarity scores differed *within* each team and not across the entire dataset, we find that 37.5% of the teams (6 of 16) had zero difference in the EP rarity scores of their BF and DH concepts, 50% of the teams (8 of 16) had a rarer DH concept, and 12.5% of the teams (2 of 16) had a rarer BF concept (see Fig. 11). Based on these results, the differences in rarity between the DH and BF concepts at the EP level is similar to differences at the PP level.

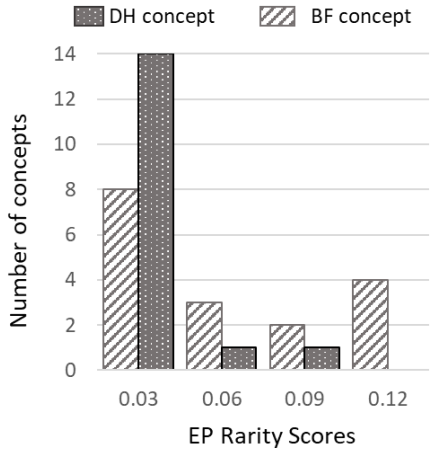

FIG. 10: DISTRIBUTION OF DH AND BF EP RARITY SCORES IN THE DATASET

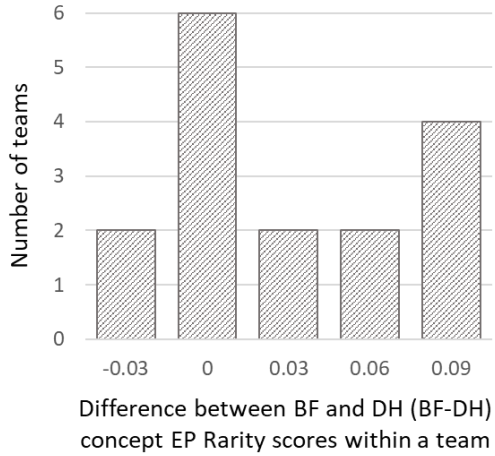

FIG. 11: DISTRIBUTION OF DIFFERENCES BETWEEN BF AND DH (BF-DH) EP RARITY SCORES WITHIN A TEAM

RQ3d: How do DH concepts differ from BF concepts in terms of their rarity differences across the three levels? Table 6 summarizes the mean rarity scores for the Physical Principle (PP), Working Principle (WP) and Embodiment Principle (EP) levels for both Best Fit (BF) and Dark Horse (DH) concepts across the 17 teams. For the BF concepts, both the difference in PP and WP mean rarity scores ($p=0.0012$) and the difference between WP and EP mean rarity scores ($p<0.001$) are statistically significant. Similarly, for the DH concepts, both the difference between the PP and WP mean rarity scores ($p=0.018$) and the difference between the WP and EP rarity scores ($p=0.017$) are statistically significant. The mean rarity scores decrease as we move from PP to WP to EP levels, indicating that both the BF and the DH concepts become rarer (more novel) as we shift from PP to WP to EP levels.

TABLE 9: MEAN RARITY SCORES ACROSS THREE LEVELS —PP, WP AND EP FOR BF AND DH CONCEPTS

	Mean PP Rarity	Mean WP Rarity	Mean EP Rarity
BF	0.21	0.13	0.06
DH	0.14	0.07	0.04

If we observe the number of concepts that remained equally rare across the three levels, we find that 43.75% of the DH concepts (7 of 16) were totally rare (rarity = 0.03) across all three levels. Concepts that are totally rare across all three levels can be considered to be the rarest concepts in the data set. In contrast, only 11.76% of the BF concepts (2 of 17) were totally rare across all three levels. Taken together, the RQ1a, RQ1b, RQ1c and RQ1d results indicate that while not all teams generated or identified concepts appropriate to the BF and DH definitions of rarity at all three functional principle levels, there was an overall trend toward doing so across the full data set.

Section Summary

To summarize, for the feasibility rubric, it was found that the DH concepts were not significantly different from the BF concepts. This was also the case when the data was analyzed *within* each team. However, when the usefulness rubric was used to analyze the data, it was found that the BF and DH concepts were distinct from each other, with DH concepts exhibiting a lower mean usefulness score than the BF concepts. When comparing the mean feasibility and usefulness scores to each other, it was observed that the mean usefulness scores were higher in both BF and DH concepts. The rarity analysis for the PP level found that the DH concepts had an overall lower rarity score than the BF concepts, indicating that they were more novel. This finding was also the case when the WP and EP principles of rarity were analyzed across the entire dataset. Despite that positive result, when the concepts were analyzed *within* the teams, it was found that only half of the teams successfully emphasized rarity (DH concepts being rarer than BF) at that PP and EP level. At the WP level, over half (62.5%) of the teams successfully emphasized novelty in their DH concepts.

Furthermore, though the differences in the mean usefulness scores and the mean rarity scores across the three levels (PP, WP, EP) between the DH concepts and BF concepts were statistically significant, the numerical differences in the means were relatively small; this may seem to indicate a limited impact. Nevertheless, if we observe the score extremes in Figures 4 (usefulness), 6 (PP rarity), 8 (WP rarity), and 10 (EP rarity), we see that BF concepts occur more often at the lower extreme than DH concepts, whereas DH concepts occur more often at the upper extreme. Thus, the relatively small difference in DH and BF concepts with regards to their usefulness and novelty (rarity) mean values is not just statistically significant, but it is also noticeable and meaningful in the context of the rubrics and research questions we are exploring.

IMPLICATIONS

In asking teams to generate both Best Fit and Dark Horse solutions to the Lifting Water Design Challenge in the same design session, our expectation was that teams would take higher risks with their DH solutions and also be able to discern which solution belonged in each category (BF or DH). Based on those expectations, our assumption was that their DH solutions would be more novel, less feasible, but equally useful when compared with their BF solutions. The analyses presented in this paper show that the DH solutions developed by the teams were indeed generally more novel than their BF solutions; our prompt was successful in this regard. However, their DH solutions were less useful, but equally feasible, when compared to their BF solutions; the push for greater novelty appears to have an adverse effect on concept usefulness but not on feasibility, at least in our dataset. This raises the question of whether teams can be primed towards greater novelty without sacrificing usefulness or feasibility. Moreover, if we examine the difference between BF and DH concepts in each team, we find that only half of the teams delivered a DH concept that was rarer at physical principle level or embodiment principle than the BF concept. Thus, for half the teams, the Dark Horse was not more novel than the Best Fit, when the design prompt has asked specifically for a concept that was 'crazy' and 'revolutionary'. This raises further questions about what factors might have influenced half the teams to distinguish between BF and DH successfully and the other half to not make a strong distinction between the two. Was it the cognitive climate of the teams, their experience or level of design skill, their domain knowledge or the nature of their interpersonal interaction? Further analysis of the data along the cognitive style and interpersonal interaction dimensions, which is currently on-going, might reveal some answers.

The overarching question to consider when completing this research was, *"Can design teams actually tell the difference between a Best Fit and Dark Horse solution?"* When looking at the results, it appears that the answer to this question is "yes, in *some* ways, but not in *every* way." Half of the design teams (or more than half, at the WP level) were able to successfully develop DH concepts that were more novel than BF concepts,

but this result came at the price of sacrificing usefulness rather than feasibility.

More broadly and looking ahead, these results point to a challenge that all design educators and practitioners face at one time or another. That is, our expectations of what a team (or individual) should do is not always what they ultimately do. Even more concerning is the idea that while we may ask a team for a "feasible (or useful or novel) concept," we may not agree with them on what a "feasible concept" looks like. While we as researchers have an archive of literature and data on which to rely, design teams rely on their own knowledge, experiences, and resources. What is "feasible" to them may not be "feasible" to us, or vice versa, which could result in unsatisfactory design outcomes. The lingering question is: how should we frame problems [15,16,25-28] and direct teams to best mitigate these issues, while still allowing teams to have freedom in their ideation?

LIMITATIONS AND FUTURE WORK

The primary limitation of this study is the small number of teams in our data set. While 17 teams were sufficient to reveal interesting preliminary results, larger samples of data could strengthen our findings and their significance. This fact also speaks to the difficulty of accessing large numbers of teams for design research. We might have chosen to examine individuals generating concepts rather than teams, but team dynamics and performance are paramount to our overarching research objectives. We plan to continue conducting these experiments with additional teams and to train additional coders to increase data processing capacity and reliability.

Evaluating design data is not without its challenges as well. When assessing the physical principles, working principles, and embodiments of concepts, differing amounts of elaboration between concepts can cause difficulties. Two concepts could be functionally the same in nature, but if one is elaborate and the other vague, the coders' interpretations could be affected. Additionally, rarity measures are inherently limited by the fact that they only consider the concepts already appearing in a given data set. It could be beneficial to conduct a more comprehensive analysis of rarity/novelty by including all generated concepts (rather than selected concepts) as well as concepts that may exist outside of the given data set. Our rubrics for feasibility and usefulness were also simple and holistic. More complex incarnations of these metrics may be worth exploring.

In the future, in addition to analyzing more teams and their BF and DH design concepts, we plan to compare participants' self-evaluations of their BF and DH concepts to the evaluations from our rubrics, expanding on the work from this paper and our prior work in [29]. We also wish to further explore framing and scaffolding of design problems in the context of the High Performance Design Team project. Not only is it important to describe the characteristics and interactions of high performance design teams, we also believe it is important to understand how the structure and guidance provided to teams can affect their ability to more or less effectively complete a design task.

CONCLUSIONS

For various reasons, in the design research world, we tend to focus more on developing design challenges that encourage norms to be "challenged", rather than encouraging careful attention to detail of existing norms. Here, we discuss the use of three rubrics—feasibility, usefulness, and rarity—to evaluate and differentiate design concepts that are referred to as Best Fit (BF) or Dark Horse (DH). The results found through this study, in combination with further research, will allow design researchers to understand whether design teams can confidently and correctly generate design solutions that exhibit different degrees of feasibility, usefulness, and novelty, as well as distinguish between these diverse concepts.

ACKNOWLEDGMENTS

This research was funded by the National Science Foundation through CMMI Grants #1635437 and #1635386.

REFERENCES

[1] Bushnell, T., Steber, S., Matta, A., Cutkosky, M., & Leifer, L., 2013, "Using A 'Dark Horse' Prototype to Manage Innovative Teams," 3rd International Conference on Integration of Design, Engineering and Management for Innovation, Porto, Portugal, September 4-6

[2] Durão, L. F. C., Kelly, K., Nakano, D. N., Zancul, E., & McGinn, C. L., 2018, "Divergent Prototyping Effect on the Final Design Solution: the Role of "Dark Horse" Prototype in Innovation Projects," 28th CIRP Design Conference, Nantes, France, May 23-25, pp. 23-25

[3] Milne, A., & Leifer, L., 1999, "The Ecology of Innovation in Engineering Design," International Conference on Engineering Design, Munich, August 24-26.

[4] Cropley, D. H., 2015, "Promoting Creativity and Innovation in Engineering Education," *Psychology of Aesthetics, Creativity, and the Arts*, **9**(2), 161.

[5] Sonalkar, N., Mabogunje, A., and Leifer, L., 2013, "Developing a Visual Representation to Characterize Moment-To-Moment Concept Generation in Design Teams," *International Journal of Design Creativity and Innovation* **1**(2), pp. 93-108.

[6] Kirton, M. J., 1976, "Adaptors and Innovators: A Description and Measure," *Journal of Applied Psychology*, **61**, pp. 622-629.

[7] Kirton, M. J., 2011, *Adaption-Innovation in the Context of Diversity and Change*, Routledge, London.

[8] Sonalkar, N., Jablokow, K., Edelman, J., Mabogunje, A., and Leifer, L., 2017, "Design Whodunit: The Relationship between Individual Characteristics and Interaction Behaviors in Design Concept Generation," ASME Paper No. DETC2017-68239

[9] Jablokow, K. W., Sonalkar, N., Avdeev, I., Thompson, B., Megahed, M., & Pachpute, P., 2018, "Exploring the Dynamic Interactions and Cognitive Characteristics of NSF Innovation Corps (I-Corps) Teams," ASEE Annual Conference, Salt Lake City, UT., June 24-27, Paper No. 21674.

[10] Starkey, E. M., Menold, J., & Miller, S. R., 2019, "When are Designers Willing to Take Risks? How Concept Creativity and Prototype Fidelity Influence Perceived Risk," *ASME J Mech. Des.,* **141**(3): 031104

[11] Toh, C. A., & Miller, S. R., 2016, "Choosing Creativity: the Role of Individual Risk and Ambiguity Aversion on Creative Concept Selection in Engineering Design," *Research in Engineering Design*, **27**(3), pp. 195-219.

[12] Toh, C. A., & Miller, S. R., 2016, "Creativity in Design Teams: the Influence of Personality Traits and Risk Attitudes on Creative Concept Selection," *Research in Engineering Design*, **27**(1), pp. 73-89.

[13] Starkey, E., Toh, C. A., & Miller, S. R., 2016, "Abandoning Creativity: The Evolution of Creative Ideas in Engineering Design Course Projects," *Design Studies*, **47**, pp. 47-72.

[14] Toh, C. A., & Miller, S. R., 2015, "How Engineering Teams Select Design Concepts: A View through the Lens of Creativity," *Design Studies*, **38**, pp. 111-138.

[15] Rechkemmer, A., Makhlouf, M., Wenger, J. M., Silk, E. M., Daly, S. R., McKilligan, S., and Jablokow, K. W., 2017, "Examining the Effect of a Paradigm-Relatedness Problem-Framing Tool on Idea Generation," ASEE Annual Conference, Columbus, OH, June 25–28, Paper No. 18507.

[16] Wright, S., Silk, E. M., Daly, S. R., Jablokow, K. W., Yilmaz, S., and Teerlink, W., 2015, "Exploring the Effects of Problem Framing on Solution Shifts: A Case Analysis," ASEE Annual Conference, Seattle, WA, June 14–17, Paper No. 11638.

[17] Dean, D. L., Hender, J. M., Rodgers, T. L., and Santanen, E. L., 2006, "Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation," *J. Assoc. Inf. Syst.,* **7**(10), pp. 649–699.

[18] Bakeman, R., & Gottman, J. M., 1997, *Observing Interaction: An Introduction to Sequential Analysis*, Cambridge University Press, Cambridge

[19] Brown, D.C., 2014, "Problems with the Calculation of Novelty Metrics," 6th Int. Conf. on Design Computing and Cognition, London, England, June 23-25.

[20] Linsey, J. S., 2007, "Design-by-Analogy and Representation in Innovative Engineering Concept Generation," PhD Thesis, The University of Texas, Austin, Texas.

[21] J. Peeters, P.-A. Verhaegen, D. Vandevenne & J.R. Duflou, 2010, "Refined Metrics for Measuring Novelty in Ideation," IDMME Virtual Concept 2010, Bourdeaux, France, Oct. 20-22.

[22] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N., 2000, "Evaluation of Idea Generation Methods for Conceptual Design: Effectiveness Metrics and Design of Experiments," *ASME J Mech. Des*., **122**(4), pp. 377–384.

[23] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003, "Metrics for Measuring Ideation Effectiveness," *Des. Stud*., **24**(2), pp. 111–134.

[24] Verhaegen, P.A., Vandevenne, D. and Duflou, J.R., 2012, "Originality and Novelty: A Different Universe," *DS 70:* DESIGN 2012, the 12th International Design Conference, Dubrovnik, Croatia, May 21 – 24, pp. 1961-1966

[25] Henderson, D., Jablokow, K., Daly, S., McKilligan, S., Silk, E., and Bracken, J., 2019, "Comparing the Effects of Design Interventions on the Quality of Design Concepts as a Reflection of Ideation Flexibility," *ASME. J. Mech. Des*, **141**(3): 031103.

[26] Silk, E. M., Daly, S. R., Jablokow, K. W., Yilmaz, S., and Rosenberg, M., 2014, "The Design Problem Framework: Using Adaption-Innovation Theory to Construct Design Problem Statements," ASEE Annual Conference, Indianapolis, IN, June 15–18, Paper No. 8781.

[27] Shergadwala, M., Bilionis, I., Kannan, K., and Panchal, J. H., 2018, "Quantifying the Impact of Domain Knowledge and Problem Framing on Sequential Decisions in Engineering Design," *ASME J. Mech. Des*., **140**(10), p. 101402.

[28] Yilmaz, S., Rosenberg, M. N., Daly, S. R., Jablokow, K. W., Silk, E. M., and Teerlink, W., 2015, "Impact of Problem Contexts on the Diversity of Design Solutions: An Exploratory Case Study," ASEE Annual Conference, Seattle, WA, June 14–17, Paper No. 11206.

[29] Jablokow, K. W., & Vora, A., & Henderson, D. A., & Bracken, J., & Sonalkar, N., & Harris, S., 2019, "Beyond Likert Scales: Exploring Designers' Perceptions through Visual Reflection Activities," ASEE Annual Conference, Tampa, FL, June 16-19, Paper No. 32150.

