WIP: Assessing Creativity of Alternative Uses Task Responses: A Detailed Procedure

Mr. Amin G. Alhashim, University of Oklahoma

Amin G. Alhashim is a Ph.D. candidate at the School of Industrial and Systems Engineering, The University of Oklahoma. Amin is studying creativity in the field of engineering education and looking forward to leveraging machine learning to deliver more personalized learning for engineers to foster their creativity.

Ms. Megan Marshall, The University of Oklahoma

Megan Marshall is an M.S. Aerospace Engineering candidate at the School of Aerospace and Mechanical Engineering, The University of Oklahoma. Her research interests include the neuroscience of creativity and design, and using these insights to develop a person's creative and design ability.

Tess Hartog, University of Oklahoma

Tess Hartog is a graduate student in Mechanical Engineering at the University of Oklahoma. Her interests include creativity, engineering education, and neuroimaging. Her research focuses on understanding creativity and divergent thinking in engineering students via the use of electroencephalography (EEG).

Dr. Rafal Jonczyk, (1) Adam Mickiewicz University; (2) Pennsylvania State University

Rafał Jończyk (PhD) is an Assistant Professor of Linguistics at the Faculty of English of Adam Mickiewicz University in Poland. His main research interests concern the behavioural and neurocognitive correlates of emotion anticipation, perception, and production in the first (L1) and second (L2) language(s). His recent research interests include the investigation of brain dynamics during creative ideation and the extent to which creative ideation may be modulated by prior knowledge and training.

Danielle Dickson

Dr. Dickson received a B.S. in Cognitive Science and a B.A. in Linguistics from UC San Diego in 2003. She earned a Ph.D. from the University of Illinois at Urbana-Champaign in 2016 with a dissertation examining the memory system's representation of numerical information, using behavioral and electrophysiological (EEG, brainwaves) measures. She extended this work into comparisons of children and adults' arithmetic processing as a postdoctoral scholar at The University of Texas San Antonio. Presently, she is incorporating more flexible forms of creative thinking as an area of postdoctoral research at The Pennsylvania State University to contrast with more fact-based arithmetic numerical comprehension.

Prof. Janet van Hell, Pennsylvania State University

Janet van Hell (PhD, University of Amsterdam) is Professor of Psychology and Linguistics and Co-Director of the Center for Language Science at the Pennsylvania State University. She is interested in the neural and cognitive mechanisms underlying language processing in monolingual and bilingual children and adults, including creative language use.

Dr. Gül E. Okudan-Kremer, Iowa State University of Science and Technology

Gül E. Kremer received her PhD from the Department of Engineering Management and Systems Engineering of Missouri University of Science & Technology. Her research interests include multi-criteria decision analysis methods applied to improvement of products and systems. She is a senior member of IIE, a fellow of ASME, a former Fulbright scholar and NRC Faculty Fellow. Her recent research focus includes sustainable product design and enhancing creativity in engineering design settings.

Prof. Zahed Siddique, University of Oklahoma

Zahed Siddique is a Professor of Mechanical Engineering at the School of Aerospace and Mechanical Engineering of University of Oklahoma. His research interest include product family design, advanced material and engineering education. He is interested in motivation of engineering students, peer-to-peer learning, flat learning environments, technology assisted engineering education and experiential learning. He is the coordinator of the industry sponsored capstone from at his school and is the advisor of OU's FSAE team.

WIP: Assessing Creativity of Alternative Uses Task Responses: A Detailed Procedure

Abstract
Creativity is the driver of innovation in engineering. Hence, assessing the effectiveness of a curriculum, a method, or a technique in enhancing the creativity of engineering students is no doubt important. In this paper, the process involved in quantifying creativity when measured through the alternative uses task (AUT) is explained in detail. The AUT is a commonly used test for divergent thinking ability, which is a main aspect of creativity. Although it is commonly used, the processes used to score this task are far from standardized and tend to differ across studies. In this paper, we introduce these problems and move towards a standardized process by providing a detailed account of our quantification process. This quantification process takes into consideration four commonly used dimensions of creativity: originality, flexibility, fluency, and elaboration. AUT data from a preliminary case study were used to illustrate how the AUT and the quantification process can be used. The study was performed to understand the effect of the stereotype threat on the creativity of 25 female engineering students. The results indicate that after the stereotype threat intervention, participants generated more diverse and original ideas.

1 Introduction

The observation of the need for creativity in engineering is far from new. In fact, the idea of creativity as a key competency in engineering was identified as far back as the 1960s [1-3]. Multiple recent reports have also recognized the need for engineers to be "creative" and "innovative," in addition to having sound technical skills [4-6]. Despite its importance, creativity has been in a steady decline since the beginning of this century [7-8]. Students graduating from engineering fields are lacking the creative ability [9-11] even though creativity and innovation are assumed to be hallmarks of engineering [12-13].

Research has shown that creative ability can be enhanced through various ways (e.g., use of various cognitive aids, or methods to enhance creativity). Though many articles have been written on different ways to enhance creativity, there is no universally accepted standard metric(s) to assess the effectiveness of these methods. In order to study the different methods claiming to enhance creativity, a standard creativity assessment approach or metrics are needed to systematically measure relative creativity level, or gains in creativity. While there is no direct way to measure creativity, a long-standing tradition in creativity research is to use divergent thinking task outcomes to assess an individual's creative thinking ability, or individual's creativity for short [14-20].

One divergent thinking task used to assess an individual's creativity is the Alternate Uses Task (AUT) [21]. Though the AUT is commonly used to assess creativity, the literature lacks an appropriately detailed explanation of the processes involved in quantifying the responses obtained through this task. Filling this void, this paper reports a method to score the AUT responses to assess an individual's creativity. It is vital that this process is standardized so that

scores generated through the scoring process are consistent across studies. This allows the AUT to be a useful tool in assessing the outcomes of a creativity-enhancing intervention.

In the remainder of the paper, first the AUT will be described along with a discussion of the scoring processes currently used to analyze the data generated. Noted drawbacks of these scoring processes will also be introduced. In Section 3, we discuss the specifics of the process we propose to score the AUT responses. Section 4 presented the results of applying this scoring process on a dataset obtained from a case study investigating the effect of a stereotype threat on female participants' AUT performance. The limitations of the scoring process and the experimental design used to collect the data along with suggested future directions are discussed in Section 5. Section 6 summarizes the conclusions.

2 Background

2.1 Alternative Uses Task and Scoring

In the Alternative Uses Task (AUT), developed by Torrance [21], the participant is asked to generate as many alternative uses as possible for a common object such as a pen, a brick, or a paperclip. For instance, alternative uses for a brick could include a bed riser, a place mat, or a weapon. This task may be repeated for several objects, one object at a time, with each object recorded as a separate trial. Within the literature, there are various constraints applied to this task, such as time limits and number of trials. These constraints are usually dependent on the objective of the study [22]. Once participants' answers are collected, they are analyzed and scored to assess each participant's creativity in divergent thinking.

Over the years, researchers have proposed several ways to score responses obtained through the AUT. The most commonly used dimensions to score responses are originality, or how rare the responses are; flexibility, or how different the responses are; fluency, or how many responses are generated; and elaboration, or how informative the responses are [7, 18; 20, 23, 24]. Other dimensions used to score the AUT include appropriateness/usefulness, how well an idea would work; and creativity, the first impression of how creative a response is on a numeric scale [25-26]. Recently, an automated creativity assessment (SemDis) has been developed to score AUT responses in terms of semantic distance (i.e., the extent to which an idea is conceptually distant from common ideas) by computing the similarity between concepts in large corpora of natural language [27].

Not only different dimensions are being used in different studies, but the dimensions themselves can be calculated differently. For example, the dimension originality usually represents the uniqueness of a response within its data set. But how does one quantify originality? In *Testing for Creativity: Three Tests for Assessing Creativity,* Bayliss [28] presents the following method for scoring originality: comparing within the entire data set, responses that occur more than one percent of the time but less than five percent receive one point; responses occurring less than one percent receive two points; all others receive zero points. One might realize, though, that this kind of scoring would depend greatly on how those analyzing the data categorize similar answers, and details of this scoring process are missing in the literature. Other strategies that have been used to score originality include using trained experts to individually judge the

originality of an answer and averaging their ratings, proportionally weighting each use by its frequency of occurrence, and selecting the top two responses and rating them further [22].

Besides different dimensions and various ways to calculate these dimensions, there is no standard method for compiling the scores for each dimension into an overall metric of creativity in divergent thinking. For instance, studies may use weights when compiling dimensions to account for their interdependence, or place emphasis on a certain dimension. Many have also noted the problem of fluency affecting originality. The more responses a participant generates, the more likely they include an original response. Also, the number of responses in a data set will affect any proportional weightings used to score originality. Some studies use corrective factors to account for this effect while others do not. In many cases, if a corrective factor is used, it is not always apparent, as no detailed explanations are provided. In other cases, the complex interplay between these two factors is completely ignored [22].

Lastly, the subjectivity of those scoring the responses plays a large role in the outcome of the results. As mentioned above, before scoring can take place a categorization process is sometimes necessary. This is done in order to calculate the flexibility dimension. Scorers judge the meaning or intent of a response based on their understanding, and hence, subjectivity is inherent to the process. In this paper, we use the dimensions of originality, flexibility, fluency, and elaboration, as presented by Bayliss [28], as the foundation for our scoring method. We provide details of the proposed scoring method in Section 3, along with a reflection on the problems encountered and how we overcame them. We justify why certain processes were followed and present our systematic approach to scoring the metrics chosen to measure participants' creativity. Although the chosen metrics were not compiled into an overall creativity metric in this paper, it is offered as an important direction to pursue.

3 AUT Assessment Process

In order to quantify the responses obtained through an AUT implementation or episode, the responses were first "coded." This "coding" process, described in Section 3.1, is how our team categorized the responses. Based on the assigned codes, originality, flexibility, fluency, and elaboration were computed; the process followed is explained in Section 3.2. A flowchart of the overall process is depicted in Fig. 1.

3.1 Coding Responses

AUT responses were categorized or "coded" based on the nature of the function given in the response; responses with similar functions received the same code. For example, "paper weight," "sink treasure down ocean," and "lighten your load on hot air balloon," were all coded as "weight." We coded responses based on the function of the response, since participants in the AUT are asked to list alternate uses, i.e., functions, for various objects.

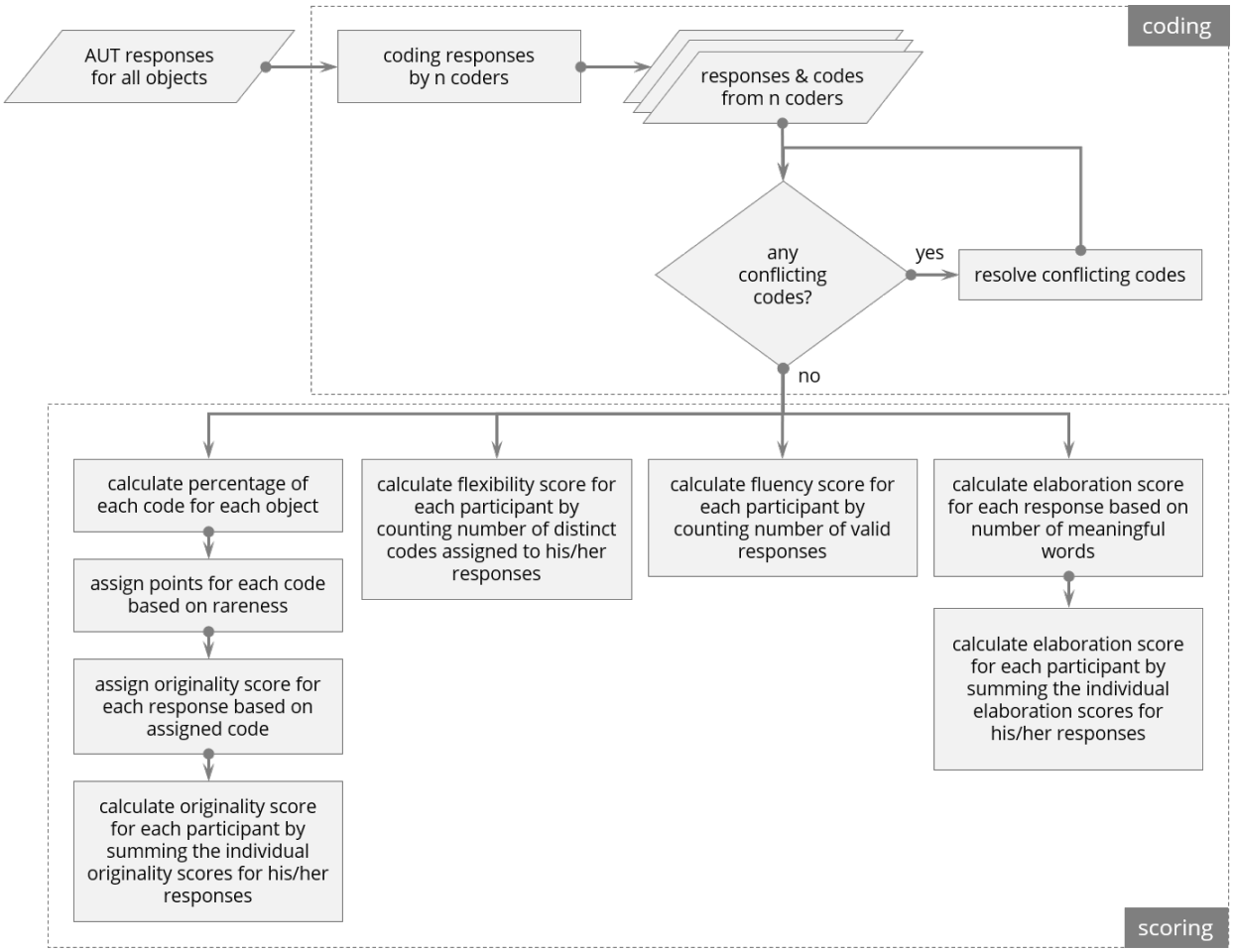

Figure 1: Flowchart of AUT assessment process.

First, each member of our team individually coded responses for at least three of the eight objects, or trials. During the initial meeting to compare our codes, we realized that we each had a different number of codes. One team member had around 80, another had around 30, and the third had around 10. At this point, the team agreed to establish an appropriate granularity for the codes. With too many codes, the flexibility scores would be high and the originality scores would be less distinctive across answers. Given too few codes, the opposite would occur. This led us to an objective of about 20-40 codes per trial (object) for our data set.

While coding, we also noticed that some answers did not make sense. For example, one participant listed “can’t pay attention”, “unappealing day” and “need food” as responses for their given object. These responses were considered invalid in the end and were not scored. Additionally, some responses could not be associated to a function or a use. Given the object hanger, a participant listed “talking stick” as a response. Evaluators could not come to a consensus on the meaning/intention of this response, and therefore it was not scored.

When the coding team members met again, it was realized how semantics affected the coding process. Some uses were coded completely differently. Many responses were coded to similar functions, but individual inferences from the words led to disagreements. To aid these

discussions, the team created definitions for each code, which could also be used for future trials and scorers. It was also realized that because the coding for each trial was done separately, across trials codes changed. To resolve this, the team adopted the "one-code book" approach, and applied it across trials. Overall, this was an iterative process that inherently included subjectivity. However, experiences from this process helped design the following strategy for coding.

The responses obtained from all the test takers are aggregated and coded individually by multiple coders. Regardless of the trials (different objects), the coding should be completed aggregately to ensure consistency across trials. The number of codes should also be set to an appropriate value based on the data set to ensure appropriate calculations of the chosen dimensions. The coders should then come together and collectively resolve any conflicts in coding in order to ensure consistency in scoring. The coders should exclude the responses that do not make sense from further calculations. Once codes have been established, the scoring process for each dimension can commence.

3.2 Scoring Creativity Dimensions

Computation of the values for originality, flexibility, fluency, and elaboration was based only on valid responses. Originality and flexibility were computed based on the assigned codes, while fluency and elaboration were computed based on the raw responses. A detailed account is provided below for how each dimension was scored. An example score for one participant is provided in Table 1 for illustration.

Table 1: Example of AUT assessment process for the item 'helmet'

Alternative usages ideas for Helmet	**Code**	**Originality**	**Flexibility**	**Fluency**	**Elaboration**
Weapon	Weapon	1	Code 1	Idea 1	1
Dish to **eat soup**	Container	0	Code 2	Idea 2	3
Watering plant	Container	0	Code 2	Idea 3	2
Sell it **for money**	Money	2	Code 3	Idea 4	3
Football	Entertainment	0	Code 4	Idea 5	1
Safety goggle	Protect	0	Code 5	Idea 6	2
Place on sculpture to **vandalize**	Entertainment	0	Code 4	Idea 7	4
Net score for participant		**3**	**5**	**7**	**16**

Originality. The originality score was calculated for each response based on the rareness of the assigned code. The rareness of a code was calculated based on the number of responses assigned to a code compared to the total number of responses for a given trial. A response that belonged to a code that appeared less than or equal to 1% of the total number of responses, i.e., a very rare code, received two points. If a response belonged to a code that appeared more than 1% but less than or equal to 5% of the total number of responses, i.e., a relatively rare code, it received one point. A response that belonged to a code that appeared more than 5% of the total number of responses, i.e., not rare, received no points.

For example, in Table 1, the participant listed seven alternative uses for a helmet. These included the responses "weapon," which was coded as "weapon," and "to sell it for money," which was coded as "money." Within the dataset of all participants for the trial "helmet," the code

"weapon" appeared between 1-5% of the time, so the participant received one point for originality for that response. Within this same dataset, the code "money" appeared less than 1% of the time; accordingly, the participant received two points for that response. The overall originality score for one trial is the sum of the originality scores assigned to each response. In the case of multiple trials, the total originality scores from each trial the participant completed are averaged together to obtain the average originality score.

Flexibility. The flexibility score represents the number of distinct codes that appear within the set of responses generated during a trial. This meant that each unique code within an individual participant's dataset received one point, and the sum of those points was the flexibility score. For example, in Table 1, though the participant provided seven answers, there were only five different codes, so the participant received a flexibility score of five. In the case of multiple trials, the total flexibility scores from each trial the participant completed are averaged together to obtain the average flexibility score.

Fluency. The fluency score was a count of the number of responses given for an object. For example, in Table 1, the participant gave seven responses, the fluency score would be seven. In the case of multiple trials, the total fluency scores from each trial the participant completed are averaged together to obtain the average fluency score.

Elaboration. This score depended on the number of meaningful words given in the response. Generic words such as "something" or "people" as well as the object name were not considered meaningful and were not included when calculating the elaboration score. A response received one point for each meaningful word it contained. For instance, given the object helmet, "cover things" would receive one point for "cover" but would not get a point for the general word "things." Another example from Table 1 was "Place on sculpture to vandalize." This response would get four points: one point each for "Place", "sculpture", "vandalize" and "on." "Place" refers to a specific action, "on" and "sculpture" refer to a specific place, and "vandalize" refers to a goal or objective. To obtain a total elaboration score for a trial, the elaboration scores were first evaluated for each response, then summed. In the case of multiple trials, the total elaboration scores from each trial the participant completed are averaged together to obtain the average elaboration score.

4 Case Study: Effect of the Stereotype Threat on Creativity

This section presents a study that investigated the effect of a stereotype threat on creativity of female students. The design of the experiment is presented in Section 4.1 followed by an application of the AUT evaluation process in Section 4.2. Section 4.3 presents the results of the evaluation process for the creativity dimensions.

4.1 Method Design and Procedure

Twenty-five female participants (M_{age}=19.1 years, SD_{age}=.89) participated in the study. The participants were recruited from the Penn State University campus. Only one participant was excluded from the study due to giving invalid responses to the given tasks.

A pre-post design was employed to test the effect of a stereotype threat on participants' creativity. In a predetermined time, each participant was asked to generate up to ten alternative uses for eight objects: foil, a hanger, a key, a pipe, a brick, a helmet, a magnet, and a pencil. The objects were shown to the participants one at a time in random order while their brain activity was recorded using electroencephalography (EEG; the EEG results are not reported in this paper). When generating alternative ideas for an object, the participants first mentally thought of an alternative usage and when done pressed a button and verbalized their response. The process continued until ten usages were generated or the time was up for that object.

Four of the objects were shown before the intervention while the remaining four objects were shown afterward. Objects shown were counterbalanced across participants using a Latin-square design. The intervention, i.e., the stereotype threat, was administered by a male student after the first part of the experiment. The male student said the following to each of the participants:

"We're looking how you're doing. What we've seen so far is that women tend to struggle with this task, so please try to do the task to the best of your ability after the break"

All participants received a stereotype threat regardless of their performance on the first part of the experiment that they just finished, i.e., the task of generating alternative uses for the four objects. The time taken by any participant to generate alternative ideas for each object was not considered as a factor and hence was not recorded.

4.2 Results and Discussion

In this section, exploratory statistics of the collected responses as well as the calculated metrics of creativity dimensions are provided. A paired-samples t-test was used to compare the creativity of the responses before and after the stereotype threat intervention for originality, flexibility, fluency, and elaboration. An alpha level of 5% was used to determine the level of significance in differences in the four creativity dimensions. Statistical analysis was performed using R [29].

4.2.1 Responses

A total of 1,390 responses were collected from the participants for the eight objects. Eleven responses, .79%, were excluded from analysis because they did not make sense, i.e., they were coded invalid and excluded from calculations. This left a total of 1,379 valid responses. The total number of valid and invalid responses received for each object is shown in Table 2. It is interesting to note that the number of invalid responses decreased after the stereotype threat. This could point to participants better understanding of how to complete the AUT after some practice and/or participants rising to the challenge to provide appropriate responses after receiving the challenge to provide appropriate responses after receiving the stereotype threat. The ranges of the number of responses given by the participants for each object are shown in Fig. 2. Most participants provided between 5-10 responses, before and after the stereotype threat.

Table 2: The total number of valid and invalid responses collected for each object.

Object	Before the Intervention?	After the Intervention?	Total Number of Responses	Total Number Valid	Total Number Invalid
Foil	x		185	185	0
Hanger	x		171	164	7
Key	x		174	173	1
Pipe	x		182	180	2
Brick		x	183	183	0
Helmet		x	171	171	0
Magnet		x	153	153	0
Pencil		x	171	170	1

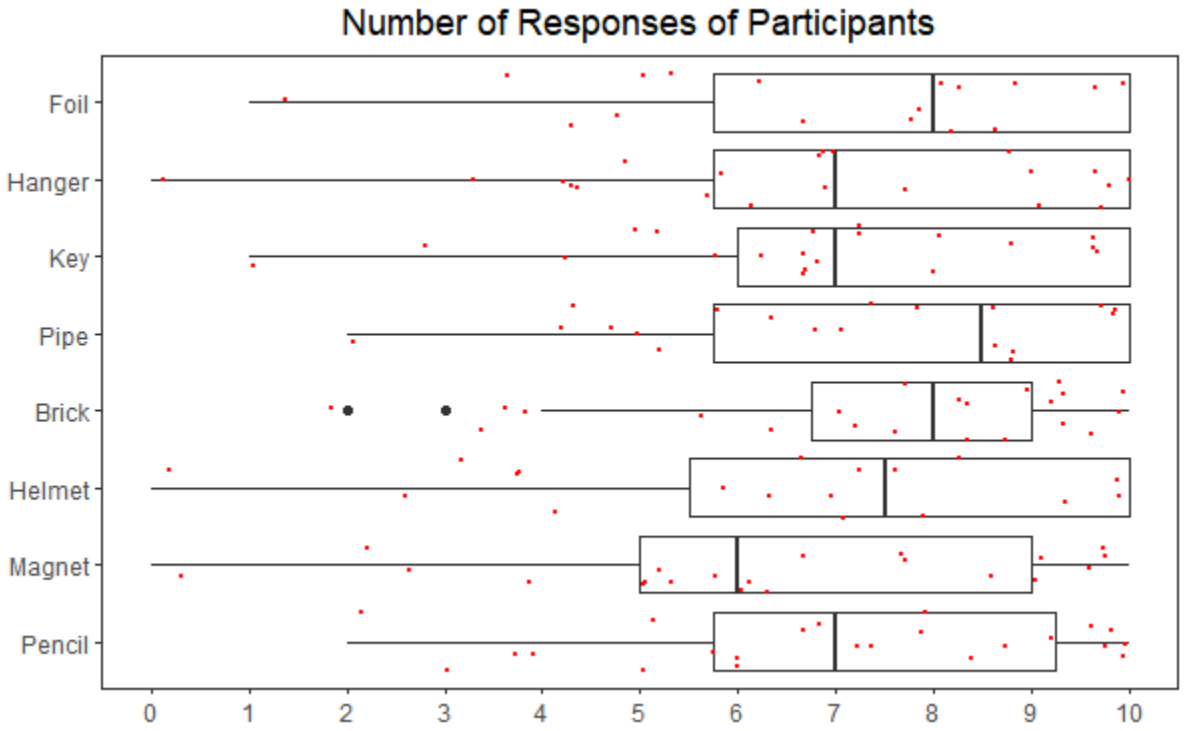

Figure 2: Range of number of responses given by participants for each object (boxplot) with individual points (small red dots). The first four objects (foil, hanger, key, and pipe) were administered before the intervention while the last four objects (brick, helmet, magnet, and pencil) were administered after the intervention.

4.2.2 Codes

The number of codes used to label the responses for each object is shown in Table 3. As mentioned in Section 3.1, the number of codes for each object was kept to about 20-40, with consistent codes for each trial. The lowest number of codes was associated with the helmet trial, as a vast majority of the responses were coded to "container," "entertainment," and "protect." The commonality of these codes suggests that it was hard to deviate from these uses to find more creative, unexpected uses. The highest number of codes was associated with the hanger trial. This suggests that the versatility or transformability of an object may play a role in a

participants' ability to think of creative uses. For example, answers included using a hanger to pick a lock, perform surgery, reach something, as a pirate hook, as jewelry, etc. The range of the number of responses assigned to each code for each object is shown in Fig. 3. In most trials, 25% of the codes were coded to responses that were given less than 1% of the time, and over 75% of the codes were associated to responses that were given between 1-5% of the time. The majority of responses fell within less than 25% of the codes.

Table 3: The number of codes used to label responses for each object.

Object	Before the Intervention?	After the Intervention?	Number of Codes
Foil	x		32
Hanger	x		44
Key	x		29
Pipe	x		35
Brick		x	28
Helmet		x	19
Magnet		x	34
Pencil		x	33

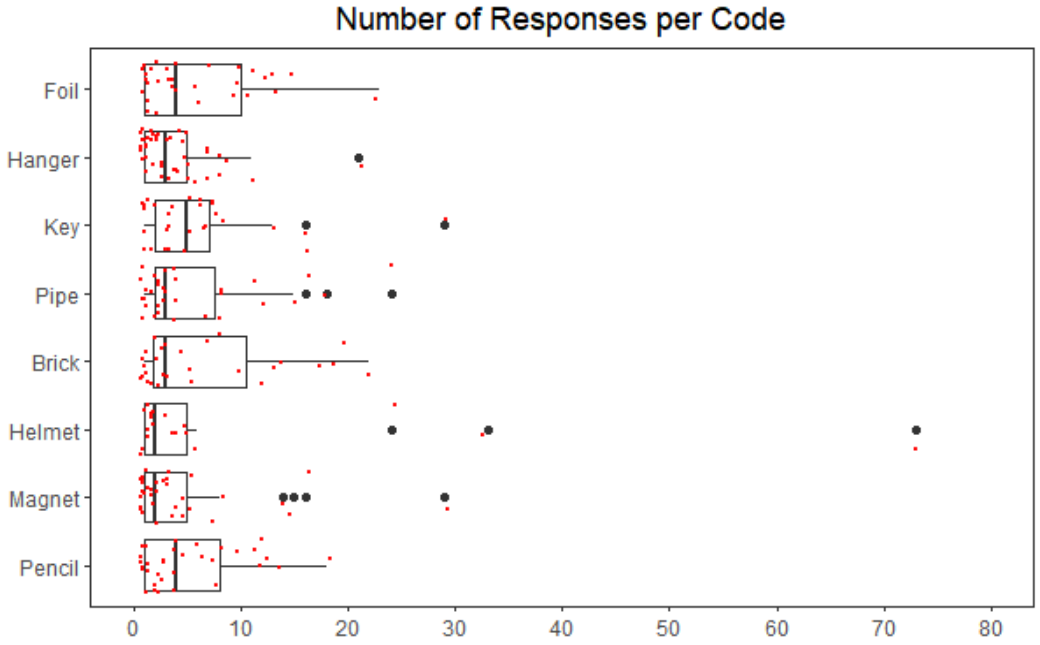

Figure 3: Range of number of responses per code for each object (boxplot) with individual points (small red dots). The first four objects (foil, hanger, key, and pipe) were administered before the intervention while the last four objects (brick, helmet, magnet, and pencil) were administered after the intervention.

4.2.3 Creativity Metrics Before and After the Stereotype Threat

To understand the effects of the stereotype threat on participants' creativity, a paired-samples t-test with an alpha level of 5% was used to check for any significant differences in the means of

participants' scores for each creativity dimension. The mean pre- and post-intervention scores for originality, flexibility, fluency, and elaboration, are shown in Fig 4.

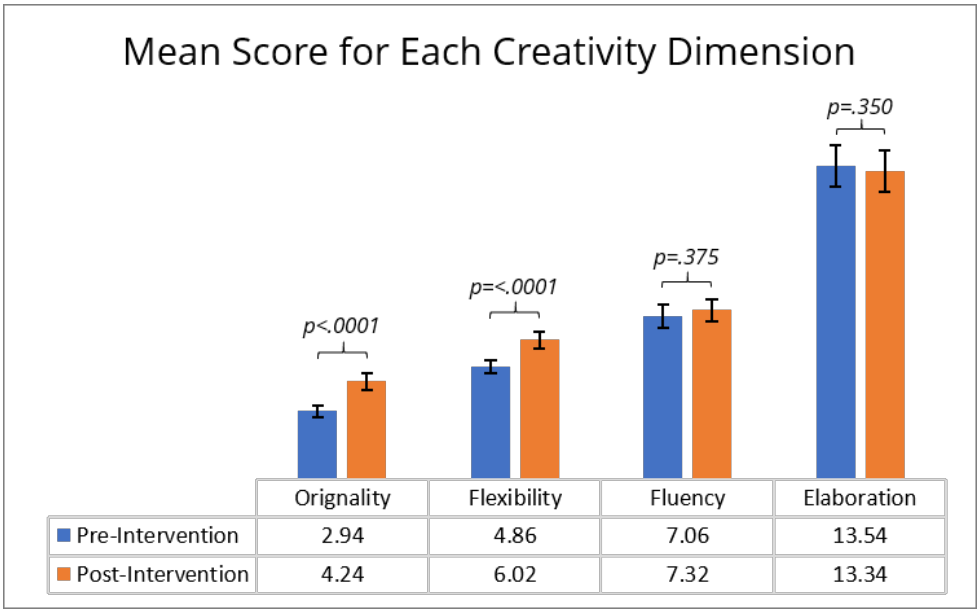

Figure 4: Mean pre- and post-intervention scores for each of the four creativity dimensions

Originality. There was a significant difference in the originality scores before the stereotype threat (M=2.94, SE=.26) and after the stereotype threat (M=4.24, SE=.37); $t(23)$=4.83, p<.0001. The higher originality score after the stereotype threat suggests that the stereotype threat may have actually spurred the participants to perform better, and this pressure resulted in more original ideas.

Flexibility. There was a significant difference in the flexibility scores before the stereotype threat (M=4.86, SE=.30) and after the stereotype threat (M=6.02, SE=.38); $t(23)$=4.21, p<.0001. Again, this result suggests that the stereotype threat may have pushed participants to stretch their minds, which in turn enabled responses that differed greatly from each other within a trial.

Fluency. There was no significant difference in the fluency scores before the stereotype threat (M=7.06, SE=.50) and after the stereotype threat (M=7.32, SE=.49); $t(23)$=0.91, p=.375. These results suggest that the stereotype threat does not have much effect on a participant's ability to produce more or fewer responses.

Elaboration. There was no significant difference in the elaboration scores before the stereotype threat (M=13.54, SE=.90) and after the stereotype threat (M=13.34, SE=.91); $t(23)$=0.35, p=.729. These results suggest that the stereotype threat does not have an effect on a participants' choice to elaborate a response.

5 Limitations and Future Work

The initial step of the quantification process, i.e., the coding step, described in this paper is laborious. All the coders need to go over all the responses one-by-one and assign a code for each.

Multiple trained evaluators need to be used to ensure consistency in coding. A significant amount of time is also required to resolve any conflicts between the evaluators in the assigned codes. One way to mitigate this is to move from an open coding scheme where the coders can label the responses with any label found appropriate to a closed coding scheme where the coders pick from a curated list of codes. The curated list of codes should be comprehensive enough to capture common and uncommon uses. A way to further expand while retaining consistency is to have a special code for which the responses that do not belong are associated. However, the coders should not fall into the trap of quickly assigning responses to this special code. The responses assigned to the special code are advised to be reviewed by another coder to make sure that the assignment is legitimate. Another approach that could prove effective in the coding process is to use unsupervised machine learning techniques, i.e., unsupervised clustering and unsupervised classification. The unsupervised clustering resembles the open coding scheme while the unsupervised classification resembles the close coding scheme.

Besides improving the coding process, new ways of scoring the originality dimension for individuals or a small group of individuals are necessary. Another problem with the current scoring mechanism for originality is that when the sample size increases, the probability for ideas to be considered creative reduces. One possible solution is to have a standard list of alternative uses for each object against which the originality of the responses obtained via the task is evaluated. This solution may not be appropriate if the standard list is not comprehensive enough or does not suit the cohort of participants performing the task. To overcome this problem, a list of alternative uses can be collected from a sample of the target population who will later not be part of the planned study in the future. The list can then be used to evaluate the originality of the alternative uses obtained from the participants who are doing the task for the first time.

In addition to considering improved scoring mechanisms for the originality dimension, one creativity dimension that can be considered in the future is the "usefulness/appropriateness." The purpose of the usefulness/appropriateness dimension is to distinguish between random responses and those that had been generated with a purpose in mind. The evaluation of this dimension will require passing through all the responses one by one and evaluating and indicating whether a response can be considered useful or not in one setting or the other. Although the coders in this study are coming from different backgrounds (engineering and non-engineering), some of the responses being evaluated could be outside the familiarity scope of the coders and hence end up unevaluated[1]. One way to mitigate this problem is to have people with more diverse backgrounds do the coding. Having such a group of coders, however, is not easy. One may turn into crowdsourcing to achieve such diversity but may end up dealing with the validity of the provided codes. Although a plausible solution, crowdsourcing should be accompanied with a rigid validation strategy.

Since coding will exclude invalid responses and those that the coders do not come to a consensus about, the participants must be made clear on the task that they are about to do. This can be

[1] One of the reviewers of the manuscript gave a plausible explanation for the alternative use "talking stick" for the hanger object given by one of the participants that the coders could not decipher. The reviewer explanation was "a 'talking stick' is an object that gets passed around as people in a group want [to] speak. Only the person holding the object can speak and the others must be silent until th[e]y have a turn with the object."

achieved by making sure that the participant does not have any question about the study or its prompts and allow him/her to go through a training trial before doing the actual task. For example, the participant must be made clear that they need to be as informative as possible whenever they provide an alternative use for objects. This will remove any ambiguity when assigning codes to the response as well as ensure fair scores for the elaboration dimension. The case study did not administer any personality tests and hence concluding that the stereotype threat is an effective method in encouraging female students to be more creative is not definite.

6 Conclusion

In engineering design education, the divergent thinking tasks can also be utilized to evaluate the effectiveness of certain pedagogic styles, methods, or techniques on the creative ability of engineering students. Having such insights can help administrators to take informed decisions to foster creativity in engineering curricula and help students be more aware of their creative skills and work on enhancing them [30-33].

In this paper, the Alternative Uses Task (AUT), a commonly used divergent thinking task, to measure individuals' creative abilities was used to experiment with and design a robust way of quantifying dimensions of creativity. The paper introduced the AUT and how others have scored the responses and noted the subjectivity required in scoring. The paper also pointed to the lack of a detailed standardized scoring process. In an effort to standardize the process and clarify its execution, a new scoring process is proposed and applied on a case study that investigates the effect of the stereotype threat on creative ability of female participants.

Acknowledgments
We are grateful to the reviewers who commented on the draft (as well as those who commented on the abstract) and helped in improving the manuscript. This material is based upon work supported by the National Science Foundation under Grants No. 1561660 and 1726358, 1726811, and 1726884. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Jones, F. E. (1964). Predictor variables for creativity in industrial science. *Journal of Applied Psychology*, 48(2), 134–136. https://doi.org/10.1037/h0047167
[2] McDermid, C. D. (1965). Some correlates of creativity in engineering personnel. *Journal of Applied Psychology*, 49(1), 14–19. https://doi.org/10.1037/h0021658
[3] Sprecher, T. B. (1959). A study of engineers' criteria for creativity. *Journal of Applied Psychology*, 43(2), 141–148. https://doi.org/10.1037/h0047763
[4] Robinson, M. A., Sparrow, P. R., Clegg, C., & Birdi, K. (2005). Design Engineering Competencies: Future Requirements and Predicted Changes in the Forthcoming Decade. Design Studies, 26(2), 123-153.
[5] National Academy of Engineering. (2004). *The Engineer of 2020: Visions of Engineering in the New Century*: National Academies Press Washington, DC.

[6] National Academy of Engineering. (2013). *Educating Engineers: Preparing 21st Century Leaders in the Context of New Modes of Learning: Summary of a Forum*. Washington, DC: The National Academies Press.
[7] Kim, K. H. (2011). The Creativity Crisis: The Decrease in Creative Thinking Scores on the Torrance Tests of Creative Thinking. Creativity Research Journal, 23(4), 285–295. https://doi.org/10.1080/10400419.2011.627805
[8] Kim, K. H., & Pierce, R. A. (2013). Torrance's innovator meter and the decline of creativity in America. In The Routledge International Handbook of Innovation Education (Vol. 5, pp. 153–167). Routledge. https://doi.org/10.4324/9780203387146.ch11
[9] Bateman, K. (2013). IT students miss out on roles due to lack of creativity. Retrieved July 21, 2019, from https://www.computerweekly.com/blog/ITWorks/IT-students-miss-out-on-roles-due-to-lack-of-creativity
[10] Cropley, D. H. (2016). Creativity in Engineering. In G. E. Corazza & S. Agnoli (Eds.), *Multidisciplinary Contributions to the Science of Creative Thinking* (pp. 155–173). Springer Science+Business Media Singapore. https://doi.org/10.1007/978-981-287-618-8_10
[11] Kazerounian, K., & Foley, S. (2007). Barriers to Creativity in Engineering Education: A Study of Instructors and Students Perceptions. *Journal of Mechanical Design, 129*(7), 761-768. doi:10.1115/1.2739569
[12] Charyton, C. (2014). Creative Engineering Design Assessment. London: Springer London. https://doi.org/10.1007/978-1-4471-5379-5
[13] Richards, L. G. (1998). Stimulating Creativity: Teaching Engineers to be Innovators. In *FIE '98. 28th Annual Frontiers in Education Conference. Moving from "Teacher-Centered" to "Learner-Centered" Education. Conference Proceedings (Cat. No.98CH36214)* (Vol. 3, pp. 1034–1039). Tempe, AZ, USA: IEEE. https://doi.org/10.1109/FIE.1998.738551
[14] Cropley, D. H. (2019). Measuring Creativity. In Homo Problematis Solvendis–Problem-solving Man: A History of Human Creativity (pp. 9–12). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-3101-5_2
[15] Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. British Journal of Psychology, 98(4), 611–625. https://doi.org/10.1111/j.2044-8295.2007.tb00467.x
[16] Gilhooly, K. J., Fioratou, E., Anthony, S. H., & Wynn, V. (2007). Divergent thinking: Strategies and executive involvement in generating novel uses for familiar objects. *British Journal of Psychology*, 98(4), 611–625. https://doi.org/10.1111/j.2044-8295.2007.tb00467.x
[17] Hocevar, D. (1981). Measurement of Creativity: Review and Critique. Journal of Personality Assessment, 45(5), 450–464. https://doi.org/10.1207/s15327752jpa4505_1
[18] Park, N. K., Chun, M. Y., & Lee, J. (2016). Revisiting Individual Creativity Assessment: Triangulation in Subjective and Objective Assessment Methods. Creativity Research Journal, 28(1), 1–10. https://doi.org/10.1080/10400419.2016.1125259
[19] Runco, M. A. (2010). Divergent thinking, creativity, and ideation. In J. C. Kaufman & R. J. Sternberg (Eds.), The Cambridge handbook of creativity (1st ed., pp. 413–446). New York, NY, USA: Cambridge University Press.
[20] Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the

reliability and validity of new subjective scoring methods. Psychology of Aesthetics, Creativity, and the Arts, 2(2), 68–85. https://doi.org/10.1037/1931-3896.2.2.68
[21] Torrance, E. P. (1966). Torrance tests of creative thinking. Norms-technical manual. Research edition. Verbal tests, forms A and B. Figural tests, forms A and B. Princeton: Personnel Press.
[22] Abraham, A. (2018). The Neuroscience of Creativity. Cambridge: Cambridge University Press.
[23] Amabile, T. M. (1982). Social Psychology of Creativity: A Consensual Assessment Technique. Journal of Personality and Social Psychology, 43(5), 997–1013.
[24] Peterson, R. E., & Harrison, III, H. L. (2005). The created environment: an assessment tool for technology education teachers: creativity doesn't just happen by chance; the prepared environment nourishes it. The Technology Teacher: A Journal of the American Industrial Arts Association, 64(6), 7–10.
[25] Madore, K. P., Jing, H. G., & Schacter, D. L. (2016). Divergent creative thinking in young and older adults: Extending the effects of an episodic specificity induction. Memory & cognition, 44(6), 974-988.
[26] Wang, P., Wijnants, M. L., & Ritter, S. M. (2018). What enables novel thoughts? The temporal structure of associations and its relationship to divergent thinking. Frontiers in psychology, 9, 1771.
[27] R. E. Beaty, D. R. Johnson, Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Psyarxiv* (2020)
[28] Bayliss, Arlon. "Three Tests for Assessing Creativity." *Testing Creativity - Three Tests for Assessing Creativity*, Apr. 2016, arlonbayliss.com/creativity-tests/.
[29] R Core Team. (2019). R: A Language and Environment for Statistical Computing. Vienna, Austria.
[30] Charyton, C. (2014). Creative Engineering Design Assessment. London: Springer London. https://doi.org/10.1007/978-1-4471-5379-5
[31] Charyton, C., Jagacinski, R. J., & Merrill, J. A. (2008). CEDA: A Research Instrument for Creative Engineering Design Assessment. Psychology of Aesthetics, Creativity, and the Arts, 2(3), 147–154. https://doi.org/10.1037/1931-3896.2.3.147
[32] Daly, S. R., Mosyjowski, E. A., & Seifert, C. M. (2014). Teaching creativity in engineering courses. Journal of Engineering Education, 103(3), 417–449. https://doi.org/10.1002/jee.20048
[33] Howard, T., Culley, S. J., & Dekoninck, E. (2007). Creativity in the Engineering Design Process. In International Conference on Engineering Design, ICED'07 (pp. 329–330). Paris, France.