ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE
Paper ID #37269

Utilization of Automatized Creativity Ratings in Linguistically Diverse Populations: Automated Scores Align with Human Ratings

Danielle S. Dickson

Danielle Dickson received her PhD 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 electro-physiological (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. Her most recent research examines creative thinking processes as an area of postdoctoral research at The Pennsylvania State University.

Gul E. Okudan Kremer (Wilkinson Professor and Senior Director)

Gül E. Kremer is Dean-elect of Engineering at University of Dayton. Kremer served as chair of the Department of Industrial and Manufacturing Systems Engineering (2016-2021) and Senior Director Presidential Projects (2021-2022), in addition to past leadership roles at Penn State. Dr. Kremer has degrees in industrial engineering from Yildiz Technical University, a masters in business from Istanbul University, and a PhD in Engineering Management from Missouri University of Science and Technology. She was a National Research Council-US AFRL Summer Faculty Fellow in the Human Effectiveness Directorate (2002-2004), a Fulbright Scholar (2010-2011), and Program Director in NSF's Division of Undergraduate Education (2013-2016). Dr. Kremer's research interests include applied decision analysis to improve complex products and systems, and engineering education. Her research has appeared in 3 books and over 360 refereed publications. She is a Fellow of the American Society for Mechanical Engineers and senior member of the Institute of Industrial Systems Engineers. In addition, she has significant contributions to research efforts that are directed toward improving engineering education.

Zahed Siddique (Professor)

Zahed Siddique is a Professor in the School of Aerospace and Mechanical Engineering at the University of Oklahoma.

Elif Elcin Gunay

Elif Elçin Günay is an assistant professor in the Department of Industrial Engineering at Sakarya University, Turkey. She received her B.S.E., M.S.E., and Ph.D. degrees in Industrial Engineering from Sakarya University, Turkey, in 2007, 2009, and 2016 respectively. She worked in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University as a post-doctoral research associate between 2017-2018. Her research interest includes optimization in manufacturing and service systems, stochastic processes, and engineering education. Her recent research interests focus on enhancing creativity in engineering classrooms.

Janet Van Hell (Professor of Psychology and Linguistics)

Utilization of Automatized Creativity Ratings in Linguistically Diverse Populations: Automated Scores Align with Human Ratings

Abstract

Measuring an individual's creativity typically relies on labor-intensive subjective ratings of the quality of ideas and solutions to problems. In the Alternate Uses Task (AUT), frequently used in engineering design education for concept generation and to gauge creative function-object relationships, participants generate as many novel uses of everyday objects as possible within a given time frame. Unfortunately, objective and rapid evaluation of AUT responses for levels of originality and usefulness is difficult. Recently, an automatized method for generating scores has been developed, the freely accessible Semantic Distance (SemDis) tool [1]. Given the linguistic and cultural diversity of engineering students in the U.S., it seems fair to question how well this type of automatic rating system, based on prototypical language models, captures the creativity of engineering students who may be nonnative speakers of English. We extensively trained human raters to score the AUT responses of multilingual engineering students living in either a non-English environment or in the US, and the AUT responses of monolingual English engineering students. We found that the human ratings of all three groups of engineering students correlated strongly, and positively, with the automatic SemDis ratings. This forms proof of concept for using automatic rating systems such as SemDis in engineering classroom settings. In addition to saving evaluators' time, this method may also be preferred because it is unbiased to cultural and linguistic features of responders' answers that might reveal their gender, race, ethnic or linguistic background information.

1. Introduction

Engineering education frequently involves warm-up activities for concept generation to emphasize creativity and facilitate students' experience of their creativity. For example, in the Alternate Uses Task (AUT), frequently used in research (e.g., [2], [3]) and in engineering/design education (e.g., [4], [5]) to gauge creative function-object relationships, students generate alternate uses of an existing object, e.g., a brick, a pencil. For research purposes, the AUT can be used as a measure of an individual's potential for divergent thinking. For instructional purposes, the AUT may instead be used as an exercise to show benefits from individual versus team level creative problem solving, as a warm up step before students tackle a more complex design task, or to exemplify the varying performance levels in creative problem situations as faculty teach about the importance of creativity. Typically, the novelty and originality of AUT responses are later evaluated by human raters, often professors and graduate students, according to a predetermined scale [6]. AUT ratings by humans are labor intensive, which makes rating creative thinking and ideation outcomes a time-consuming process. Moreover, human ratings are subjective and vulnerable to rater bias. For example, in evaluating creative thinking outcomes of linguistically diverse students and second language learners, human raters may notice the language being used and base their judgement on surface-level grammar use or phrasing choices instead of conceptual content.

Recently, Beaty and Johnson developed an automated method of generating originality scores for the AUT, SemDis [1], and demonstrated that SemDis scores correlated strongly with those of human raters. SemDis utilizes a database of word associations to determine whether an answer is more or less distant in semantic space from the prompt word, with more distanced response words receiving higher creativity scores. The highest performing model combines output from multiple semantic models of English word associations. However, the utility of SemDis has not been systematically tested on AUT outcomes produced across linguistically diverse students and second language learners.

This is relevant for two main reasons. First, the models underlying the automated rating system assume a uniformity of language exposure and ability that might not be reflected in more linguistically diverse populations. Depending on language experience, some multilinguals may have smaller, and less accessible vocabularies than monolinguals (e.g., weaker link hypothesis, [7]) and the structure of their semantic representations do not necessarily conform to expectations present in standard language models [8], [9]. Therefore, it is not clear how well SemDis would perform across English speakers with more varied language backgrounds.

Second, when we recruited undergraduate engineering students at Penn State University to participate in research, their language background screening revealed that a high number of engineering students have a bilingual/multilingual background and are second language speakers of English [10]. Indeed, according to the American Society for Engineering Education, 9.4% of undergraduate engineering students are foreign nationals, a number which rises to 64.9% in graduate school [11]. Additionally, 22% of households in the United States report that a language other than English is spoken at home [12]. Therefore, establishing the utility of SemDis across language groups is of specific interest in the engineering education context, and beyond, where a high proportion of students come from linguistically diverse backgrounds.

To that end, the present study aims to validate the reliability of the SemDis automatic rating method across groups with different language backgrounds. We took the AUT responses of engineering students from two studies, one conducted at Penn State University and another conducted at Braunschweig Technical University, to test how well human ratings of AUT responses align with the automatic ratings of these responses. If SemDis ratings correlate strongly with human ratings of students from a more diverse language background, then this outcome will support the incorporation of automated SemDis scores in practical classroom usage. This will enable instructors to generate faster and more unbiased measures of creativity in the classroom, which in turn contributes to the overarching goal of enhancing the training and assessment of students' divergent thinking and ideation skills.

2. Method

2.1 Participants

2.1.1 Braunschweig Technical University Study Sample

One group of participants were English-speaking bilingual engineering students at Braunschweig Technical University in Germany, who are living in a non-English environment (N=37). In order

to be included in the study, participants needed to be native speakers of German and be highly proficient in English, as the full experiment required them to perform the task in both languages. Participants had been learning English in school from the age of about 9, in line with the education system in Germany. Their high German and English language proficiency levels had been verified with proficiency tests (LexTALE task, [13]), and their verbal answers in the Alternate Uses Task were fluent and understandable in both languages. For this study, only their English responses were considered, as the SemDis database is limited to models of English words.

2.1.2 Penn State University Study Sample

A second set of engineering students were drawn from a larger-scale creativity study conducted at Penn State University in the US, a predominantly English-speaking environment. These students were recruited if they identified as proficient in English, and for the purposes of this study, were separated into two language groups. The first would be considered a traditional monolingual group by the standards kept by language researchers, wherein someone self-identifies as being fluent only in English and reports no early exposure (e.g., prior to age 5) to a second language at home or regularly in their environment (N=28). The second group are bilinguals (N=21), either by self-identification or classified as such by experimenters due to reported use of another language in childhood. These language profiles typically lead participants to being excluded from English language-focused studies involving neurocognitive measures, as brain and language development is thought to be especially shaped by early language learning environments. Unlike the German-English bilinguals recruited at Braunschweig Technical University, their language profiles were more varied and diverse, with varying levels of skills across multiple languages and with a wider range of native languages.

2.2 Stimulus Materials

2.2.1 Braunschweig Technical University Materials

There were 16 common household objects used as prompt words: brick, button, car tire, chair, fork, hat, key, knife, microphone, napkin, newspaper, pencil, rope, shoe, spoon, toothbrush.

Two lists were created such that half of the words were presented with a German translation instead of in English. A given participant would therefore only be exposed to each object once. The data reported here is drawn from their English responses to whichever 8 English words they saw (two possible sets of words).

Practice item(s): plastic cup, paperclip, cardboard box, thumbtack.

Participants practiced only two of these items in English and the other two were used for their German practice block.

2.2.2 Penn State University Materials

There were 8 objects used as prompt words: key, hanger, foil, pipe, pencil, brick, magnet, helmet.

Half of the prompt words in this study were related to engineering and half were common household objects. All participants saw and responded to these same 8 objects.

Practice item(s): chair.

All participants saw and responded to this practice item.

2.3 Procedure

2.3.1 Creativity Measure

Participants completed the Alternate Uses Task. They were asked to speak as many creative and novel alternate uses of an object as possible in an assigned window of time. The task begins with instructions about generating creative responses to presented items, after which participants see a practice item (or multiple, in the Braunschweig Technical University study). The practice item provides an opportunity for the participants to clarify instructions and for the experimenter to intervene if responses are not appropriate due to misunderstanding of the task. Following orientation to the task, during the experimental phase each item prompt (a word) appeared visually and was then replaced with a "?" indicating that participants were to think about potential responses and provide an oral reply when a creative idea comes to mind. In the Braunschweig Technical University study, participants had 2.5 minutes to answer; in the Penn State University study, the time allotted was 2 minutes.

All participants saw the following instructions regarding the alternate uses task:

You will be presented with the names of objects on the computer screen, one at a time. Each object name will stay on the screen for a few seconds and will be followed by a question mark "?" At that stage, your task will be to silently generate an alternate use of the presented object and press the button whenever a creative idea comes to mind. You will then vocalize the idea and confirm this with another button press. You will be generating ideas to that object until you see a green cross (+), which will be followed by a presentation of another object.

Auditory recordings of spoken responses were assigned for transcription to trained laboratory assistants who did not later participate in the ratings procedure for those datasets. Raw full transcriptions included disfluencies (e.g., "um", "uh"), annotated timestamps where speech was difficult to comprehend (to be reviewed by the experimenter), and were structured to demarcate which item the numbered responses corresponded to. Prior to sending the items to raters for evaluation, these full transcriptions were lightly edited by the experimenter to remove grammatical or phrasing errors, disfluencies, repetitions, false starts that the participant corrected, and fragmented/incomplete thoughts. This editing was performed to avoid bias based on perceived language ability and to enable to raters to focus on judging complete and final thoughts. This editing also provided SemDis with a cleaner sample of relevant and critical words to score.

2.3.2 AUT Ratings

Raters were provided the same training procedure across studies. Five highly proficient German-English bilinguals rated the English responses of the participants from the Braunschweig Technical University study. A different group of four raters evaluated the responses for the Penn State University study (three raters were highly proficient bilinguals, each in different languages). The German-English raters were graduate students and the raters for the Penn State University study were undergraduate students. Critically, both groups of raters were similarly unfamiliar with creativity research and unfamiliar with the alternate uses task prior to receiving training in ratings. As such, they were provided extensive training as unpacked below.

The core instructions provided to raters were to score ideas for their creativity based on two primary criteria: 1) novel or unusual, and 2) fitting, clever, interesting, humorous and/or surprising. The reason for the second criterion is that an idea could be novel or unusual while also being senseless and unrelated to the prompt – in other words, just because an idea is new is not sufficient to make it a high quality response.

Table 1. Rating scale from 0-3 used in both studies. Raters were encouraged to use the full range of the scale. This was translated from German and lightly edited from an unpublished scale used by creativity research groups in Europe (provided in correspondence with Mathias Benedek).

Not creative	**Little creative**	**Quite creative**	**Very creative**
0	**1**	**2**	**3**
Completely uncreative idea	Obvious idea	Good idea	Very good idea
Very common or senseless	Somewhat unusual	Rather original and sensible	Original and sensible
Everyone can think of this	Many people may think of this sooner or later	Not something everyone comes up with	Very few people can think of this

We describe our step-by-step training procedure for raters below.

1) After receiving these core instructions, the experimenter held a 30 minute group meeting about expectations for ratings. This included guidance to use the full ratings scale because sometimes raters are overly cautious to assign the lowest/highest scores to responses. Instructions also encouraged raters to preview the general responses to a given item prior to rating to get a sense of what answers are common and what are more rare. A final instruction was to return to earlier ratings to evaluate their consistency with later ratings and ensure that ratings for similar responses do not drift to become more lenient/strict over time.

All raters were provided with a set of example responses to practice on that were independent of the set they would be asked to rate later. For example, the raters for the Penn State University study were given the English-language responses of four participants from the Braunschweig Technical University study.

2) After rating the practice sets independently, raters and the experimenter met to evaluate and compare ratings. The experimenter combined raters' responses into a single document to visually compare and identify items that had caused divergent responses. Raters discussed items that caused confusion during the ratings process with the experimenter and with each other. In addition, raters were instructed to adjust their ratings if they had consistently been too strict or too lenient relative to other raters. These group discussions provided an opportunity for raters to better understand expectations.

3) Raters were provided additional practice sets of responses to rate, again from an independent dataset. Note that after the first practice rating group evaluation, most raters were too inconsistent and all expressed a desire to have a second round of practice rating. This was the case for the graduate student as well as the undergraduate student raters.

4) The experimenter and raters met to review the second practice set and repeated step 2 for another 30 minutes. After these discussions, the experimenter consolidated the returned ratings and confirmed that responses for items were more consistent with each other than the prior round, in addition to confirming that overly strict/lenient judges had adjusted their personal criterions to be less skewed and more in line with other raters. By the end of this meeting, raters expressed high confidence in their ability to rate independently.

5) Throughout the training procedure, the trainer kept notes. These notes, the questions asked by the raters, and the discussion points were integrated into a Frequently Asked Questions document that now is used in our group as a supplemental guide for raters along with the standard rating scale (Appendix 1). Every research group conducting this type of research might find it useful to create such a document to clarify desired standards for rating judgments.

Raters were considered trained after completing steps 1-4 and were instructed to follow the guidelines they had used when they were independently rating the actual AUT responses. Each rater provided ratings for all of the responses by all of the participants in either the Braunschweig Technical University study (1958 total responses to be rated across 37 bilingual participants) or the Penn State University study (2361 total responses to be rated across 49 monolingual and bilingual/multilingual participants). They were additionally provided with the FAQ described in 5 (see Appendix 1).

2.4 Statistical Approach

2.4.1 Inter-Rater Reliability

Prior to directly comparing the collective human ratings with the output of the SemDis model, we validated the inter-rater reliability for each study. This is a process undertaken to statistically

validate that our human raters were producing ratings for each response that were consistent with one another. This inter-rater reliability measure was calculated with the IRR package in R [14], and is mathematically an intra-class coefficient (ICC, [15]). Higher ICC values (>0.5, where 1 is the largest value possible) indicate the degree to which each item was scored similarly by the judges.

In the interest of completeness, since the recommended SemDis output is also an average of multiple semantic distance models, we also derive the ICC of the SemDis models. Unlike human raters, this score is not needed or expected to be high because the models themselves were created based on different indices of semantic distance, and, in a sense, they have different "instructions" (algorithms) for judging semantic distances between words. In the original paper that established SemDis [1], certain models were revealed to work better at predicting scores for a given individual, and the combined output of the SemDis models was the best at predicting the human rating scores. Here, it is of interest to see if language backgrounds might influence the consistency of the individual SemDis models.

ICCs are independently produced for each language group, (1) the Braunschweig Technical University study of German-English bilinguals, (2) the Penn State University study subgroup of monolinguals with no knowledge of a second language, (3) the Penn State University study subgroup of highly proficient English bilingual speakers of linguistically diverse backgrounds.

2.4.2 Pearson Correlations

The average scores across human raters were correlated with the scores produced by the average of the SemDis models; this correlation is calculated at the individual item level across all participants in a given language group. These correlations were generated for each of the three groups listed above. Our approach here aligns with the correlations used in the original study establishing the SemDis method [1].

3. Results

3.1 Inter-Rater Reliability

For the Braunschweig Technical University study, the human raters showed high inter-rater reliability ($ICC_{raters}(C,5) = 0.82$)); SemDis models turned out to be equally reliable ($ICC_{SemDis}(C,5) = 0.77$). For the Penn State University monolingual group, the human raters and the SemDis model generated satisfactory reliability ($ICC_{raters}(C,4) = .64$; $ICC_{SemDis}(C,5) = .66$). Finally, for the Penn State University bilingual group, human raters showed satisfactory inter-rater reliability ($ICC_{raters}(C,4) = .66$) while SemDis models yielded very low reliability ($ICC_{SemDis}(C,5) = .27$).

ICCs of human raters for each language group were strong and suggest that our methodology generated trained raters that whose scores were consistent with one another. The ICCs for the SemDis models were strong for the bilinguals in the Braunschweig Technical University study and for Penn State University monolinguals, but quite weak for Penn State University bilinguals. This is notable and worthy of discussion, but does not present an issue for utilizing the average

score given individual models were found to vary and differ in their ability to predict scores of human raters in the original study [1].

3.2 Pearson Correlations

In the Braunschweig Technical University study, the Pearson's product-moment correlation between the average human and average automatic SemDis ratings was significant (r = .13, 95% CI [.08, .17], t(1953) = 5.68, p < .001; see Figure 1). Similarly, in the Penn State University study, the correlation between these ratings was significant in both bilingual (r = .15, 95% CI [.08, .21], t(798) = 4.33, p < .001) and monolingual (r = .19, 95% CI [.14, .24], t(1503) = 7.49, p < .001) groups (see Figure 2).

In sum, the correlations between human raters and SemDis were positive and strongly significant for each language group.

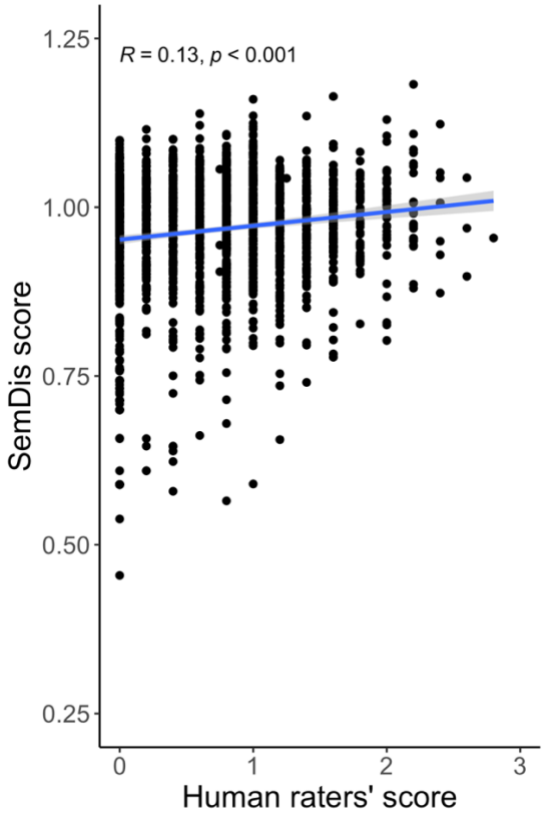

Figure 1. The correlation of automatically generated SemDis scores and the scores generated by highly trained human raters is plotted with data drawn from engineering students in Braunschweig Technical University.

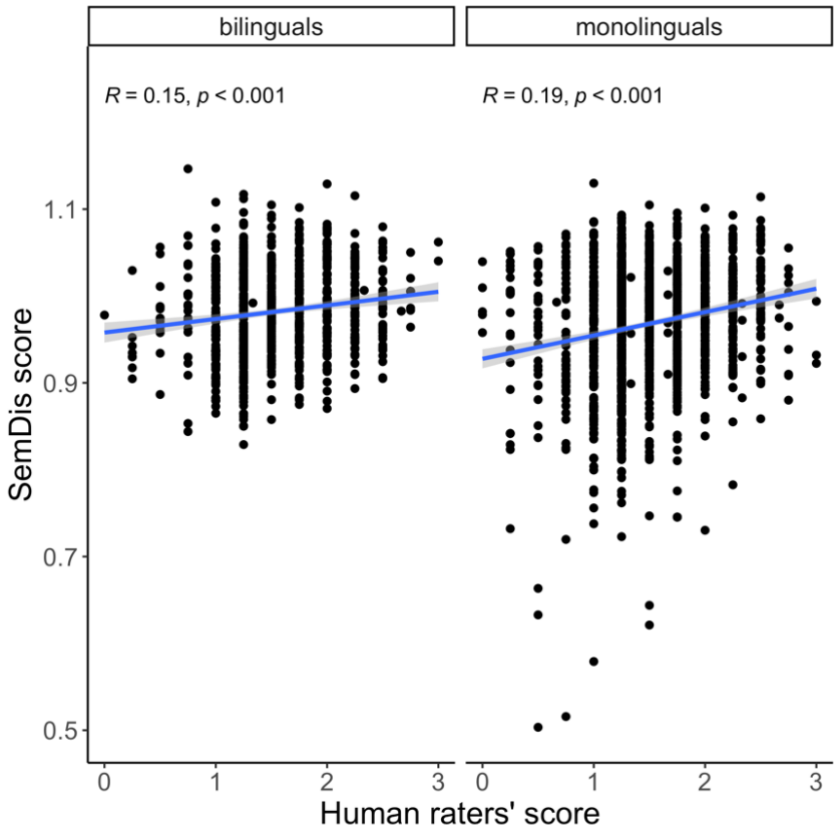

Figure 2. The correlation of automatically generated SemDis scores and the scores generated by highly trained human raters is plotted with data drawn from bilingual (left) and monolingual (right) engineering students from Penn State University.

4. Discussion

The Alternate Uses Task (AUT) is frequently used in engineering/design education to promote creative thinking and ideation in students. Instructors use AUT as an in-class exercise to teach about the importance of creativity, individual differences that support creativity, and the advantages of creative problem solving as a team. Its utility in classrooms is challenged by the demanding nature of its time-consuming evaluation process, where raters must be trained and knowledgeable about evaluating creativity and then also must individually rate each item one by one. Accordingly, a robust evaluation for each participant's responses is not possible, potentially lessening the full potential of an important lecture. Recently, an automated approach was provided for generating ratings through use of semantic distance models, the SemDis approach [1]. However, the practical use of this automatic system in linguistically and culturally diverse engineering classroom settings has not yet been systematically established. Here, we demonstrate that the creative output from highly proficient speakers of English on the AUT can be approximated by application of SemDis across three different linguistic populations: monolingual speakers of English, bilingual/multilingual speakers of English living in a predominantly English environment, and bilingual/multilingual speakers of English living in a predominantly non-English environment.

Notably, the SemDis program produces several different models estimating semantic distance, and even in the original publication not all of them were equal in their ability to approximate human rating scores [1]. We found similar evidence of discrepancies across individual models in the form of lower inter-rater reliability for models generated by responses in one of the language groups, a heterogenous sample of multilingual students at Penn State University. Unlike human raters, who undergo a training regimen aimed at producing similar ratings across potential creative responses, the underlying SemDis models are mathematically different and not expected to necessarily generate identical scores (and even human raters are not consistent when evaluating the same semantic associative responses twice, [16]). It is hard to know why the SemDis models were in less agreement for the Penn State multilingual students – perhaps some word choices resulted in discrepancies in semantic distance calculations that were not present in the other participant samples. Given that the average of the automated SemDis models correlated strongly with human raters across all language populations, the original recommendation to use the average of the SemDis models is found to be sound advice for those interested in using this application.

The present results in combination with the prior research demonstrating the utility of SemDis as a proxy measure of response creativity are encouraging for its broader use in educational and research settings. Notably, the SemDis tool is freely available on a public website, allowing for open use of the product without needing to provide funding for a license. The most obvious benefit of the tool is in the saving of labor and time. A secondary, and important, benefit of SemDis may also be in the removal of human subjective judgment. That is, when evaluating the creative thinking of linguistically and culturally diverse students, human raters may notice the language being used and base their judgement on surface-level grammar use or culturally specific phrasing choices instead of conceptual content.

There is also the possibility that human raters fail to appreciate the creativity of described alternate uses outside their life experience if the participant is of a different gender or sub-culture. Semantic association models have no such conflicts in recognizing that the words used in a description of a use is semantically distant from the name of the object. Thus, using this type of automated software has the additional benefit of being unbiased in this regard. A limitation, however, is that unlike human raters, if an idea is nonsensical, SemDis has no ability to recognize this feature of a response. Therefore, a hybrid approach of using SemDis while also having humans evaluate responses for compliance with the task may be an ideal arrangement.

The importance of creativity for engineers, and thus enhancing creative potential of engineering students have been well-documented in the engineering education literature (e.g., [17]-[19]). Investigations of engineering creativity requires use of language to identify and name objects (engineering or lay), and their relationships. The empirical evidence presented here validates the use of SemDis for a wide range of speakers, and offers a reliable and objective alternative replacing the time-consuming work of human raters for creativity research, and will also accelerate research outputs in this domain. Given that the SemDis tool is free and easily accessible, creativity lectures by faculty can be complemented by AUT-based assignments where students can also evaluate their responses – as a team or individually. Similar in-class assignments or creativity tasks in laboratory conditions can also be rapidly and robustly

evaluated, accelerating research outputs in this domain. As shown herein, SemDis results consistently mimic human's comprehension, identification and rating of distance across words. Although further studies are needed to enhance the sample size, and to further diversify participant groups and the set of object names used (engineering vs. common or lay), the present study provides evidence that warrants the continued development of SemDis and supports the use of such automated methods for evaluating creative thinking outcomes in research studies as well as in in class settings as part of interventions designed to enhance creativity.

Acknowledgments

This study has been supported by NSF grants DUE 1561660 to Zahed Siddique and Janet van Hell; NSF DUE IUSE-1726811 to Janet van Hell, NSF DUE IUSE-1726358 to Zahed Siddique, and NSF DUE IUSE-1726884 to Gul Okudan-Kremer.

References

[1] R. E. Beaty and D. R. Johnson. "Automating creativity assessment with SemDis: An open platform for computing semantic distance," *Behav. Res.*, vol 53, no. 2, pp. 757-780, Apr. 2021. https://doi.org/10.3758/s13428-020-01453-w

[2] Y. Fan, H. C. Lane, and O. Delialioğlu. "Open-Ended Tasks Promote Creativity in Minecraft," *Ed. Technol. Soc.*, vol. 25, no. 2, pp. 105-116, Apr. 2022. Available: https://www.jstor.org/stable/10.2307/48660127

[3] S. M. Ritter and N. Mostert. "Enhancement of Creative Thinking Skills Using a Cognitive-Based Creativity Training," *J. Cogn. Enhanc.*, vol. 1, pp. 243–253, Oct. 2017. https://doi.org/10.1007/s41465-016-0002-3

[4] N. Vargas Hernandez, L. Schmidt, and G. E. Okudan. "Systematic Ideation Effectiveness Study of TRIZ," *J. Mech. Des.*, vol. 135, no. 10, 101009, Oct. 2013. https://doi.org/10.1115/1.4024976

[5] C. B. Masters, M. Schuurman, G. Okudan, and S. T. Hunter. "An investigation of gaps in design process learning: Is there a missing link between breadth and depth*?" presented at the 115th annual Am. Soc. Eng. Educ. Annu. Conf.*, Pittsburgh, PA, June 22-25, 2008. https://doi.org/10.18260/1-2--3438

[6] J. P. Guilford. *The nature of human intelligence*. New York: McGraw-Hill, 1967.

[7] T. H. Gollan, R. I. Montoya, C. Cera, and T. C. Sandoval. "More use almost always a means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis," *J. Mem. Lang.*, vol. 58, no. 3, pp. 787-814, Apr. 2008. https://doi.org/10.1016/j.jml.2007.07.001

[8] K. Borodkin, Y. N. Kenett, M. Faust, and N. Mashal. "When pumpkin is closer to onion than to squash: The structure of the second language lexicon," *Cognit.*, vol. 156, pp. 60-70, Nov. 2016. https://doi.org/10.1016/j.cognition.2016.07.014

[9] K. V. Lange, E. W. M. Hopman, J. C. Zemla, and J. L. Austerweil. "Evidence against a relation between bilingualism and creativity," *PLoS ONE*, vol. 15, no. 6, e0234928, June 2020. https://doi.org/10.1371/journal.pone.0234928

[10] D. S. Dickson, R. Jończyk, Y. Liu, G. Kremer, Z. Siddique, and J. G. van Hell. "The impact of social interventions on neural correlates of creative performance," *presented at the 28th Annu. Meet. Cog. Neuro. Soc.*, virtual, March 13-16, 2021.

[11] American Society for Engineering Education, "Engineering and Engineering Technology by the Numbers 2019", Washington, DC, 2020. Available: https://ira.asee.org/by-the-numbers.
[12] US Census Bureau, 2015-2019. Available: https://www.census.gov/quickfacts/fact/table/US/POP815219.
[13] K. Lemhöfer and M. Broersma. "Introducing LexTALE: A quick and valid lexical test for advanced learners of English," *Beh. Res.*, vol. 44, no. 2, pp. 325-343, Sept. 2020. https://doi.org/10.3758/s13428-011-0146-0
[14] M. Gamer, J. Lemon, I. Fellows, and P. Singh. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. Accessed: Nov. 2021. [Online]. Available: https://CRAN.R-project.org/package=irr
[15] J. J. Bartko. "The Intraclass Correlation Coefficient as a Measure of Reliability," *Psych. Rep.*, vol. 19, no. 1, pp. 3–11, Aug. 1966. https://doi:10.2466/pr0.1966.19.1.3
[16] J. G. van Hell and A. M. B. De Groot. "Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association," *Biling. Lang. Cognit.*, vol. 1 no. 3, 193-211, Dec. 1998. https://doi.org/10.1017/S1366728998000352
[17] K. H. Kim. "The Creativity Crisis: The Decrease in Creative Thinking Scores on the Torrance Tests of Creative Thinking," *Creativity Res. J.*, vol. 23, no. 4, pp. 285–295, Nov. 2011. https://doi.org/10.1080/10400419.2011.627805
[18] K. Kazerounian and S. Foley. "Barriers to Creativity in Engineering Education: A Study of Instructors and Students Perceptions," *J. Mech. Des.*, vol. 129, no. 7, pp. 761-768, July 2007. https://doi:10.1115/1.2739569
[19] D. H. Cropley. "Promoting creativity and innovation in engineering education," *Psych. Aesthet. Creativity Arts*, vol. 9, no. 2, pp. 161-171, May 2015. https://doi.org/10.1037/aca0000008

Appendix 1. Frequently Asked Questions

Q. If the participant repeats essentially the same idea multiple times, should each one receive the same score, or should the highest score go only to the first instance of the idea, and the rest get 0, because they're not new ideas?

A. For the repeat or near-repeat responses, I would only score the best of the ideas and then give the rest a 0 because it's not a new/creative use after one time it's mentioned. There are some participants where they would list an action ("cleaning", "building"), and then list a series of "cleaning"/"building" examples as part of the same single idea. It's really just one answer being elaborated on, so only one of the answers should receive a score.

Q. Sometimes they suggest ideas for novel uses of objects that are definitely unique and no one else is saying it, but also the idea doesn't make that much sense and I can't figure out why they think that would be a use for the object.

A. It's okay if their idea doesn't seem practical or likely, but there is a fine line between a nonsense idea and an extremely novel/creative idea, where something tips over into nonsense if it really doesn't make any sense.

Q. A participant described the items either physically or what their main purpose is instead of coming up with novel ideas for them (e.g., "can be strong depending on how you fold it", "if you make a big toothbrush company you can make money", "you can avoid going to the dentist if you brush your teeth often"). Does describing the object count as a use?

A. This type of response is a 0, it's not providing an idea of a new use for the object at all.

Q. The participant described doing something to an object – is that a use?

A. The way these genre of "do something to the item" answers work is if they say what they would use the object for after they did the thing to it. For example, if they said "melt it" alone, that's not a use, but if they said, "melt it and then reshape it into a statue for your desk" that *is* a novel use. "Burn it" alone is not good enough for a novel use for an object, but "burn it and use the ashes to make some artwork" would be. With only the first part of the idea, it's like they got partway through the thinking process but didn't complete it. So, we can't complete their ideas for them - these are 0 responses unless they can unpack the use for the item after something's been done to it.

Q. What if a participant describes a scenario and then provides a use for the object given this scenario that might have happened ("if someone threw [object] in the lake you could catch it with your fishing device")?

A. For this I think depending on how creative their thinking is it could get a higher score or a kind of mediocre one (a 0 if it doesn't make sense). They're basically describing the item as an object to be fished, which it could be given the scenario they concocted. I would give that a 1, and it would be a 2-3 if they had elaborated more on it and if no one else said anything like it. In that specific example, they left it up for the reader to infer why someone might want to do that (are they practicing how to fish? playing?) – so, as a rater, there's only so much you can infer about what a participant might have intended (can't rate things they never said).

Q. The word for the object has multiple meanings and the participant started talking about the secondary meaning of the word. Does that count?

A. The items were presented as words on the screen without a picture, and participants knew they were supposed to be talking about common objects. Alternate meanings of a word isn't a novel use of the object – it's not the task – so those type of responses get 0.