FEATURE ENGINEERING FOR DESIGN THINKING ASSESSMENT

Arlitt, Ryan (1); Khan, Sumbul (2); Blessing, Lucienne (2)

1: Technical University of Denmark; 2: Singapore University of Technology and Design

ABSTRACT

As design and design thinking become increasingly important competencies for a modern workforce, the burden of assessing these fuzzy skills creates a scalability bottleneck. Toward addressing this need, this paper presents an exploratory study into a scalable computational approach for design thinking assessment. In this study, student responses to a variety of contextualized design questions – gathered both before and after participation in a design thinking training course – are analyzed. Specifically, a variety of text features are engineered, tested, and interpreted within a design thinking framework in order to identify specific markers of design thinking skill acquisition. Key findings of this work include identification of text features that may enable scalable measurement of (1) user-centric language and (2) design thinking concept acquisition. These results contribute toward the creation of computational tools to ease the burden of providing feedback about design thinking skills to a wide audience.

Keywords: Design thinking, Design education, Machine learning, Semantic data processing

Contact:
Arlitt, Ryan Michael
Technical University of Denmark
Mechanical Engineering
Denmark
rmarl@mek.dtu.dk

Cite this article: Arlitt, R., Khan, S., Blessing, L. (2019) 'Feature Engineering for Design Thinking Assessment', in *Proceedings of the 22nd International Conference on Engineering Design (ICED19),* Delft, The Netherlands, 5-8 August 2019. DOI:10.1017/dsi.2019.396

1 INTRODUCTION

Design is an interdisciplinary domain that employs approaches, tools, and thinking skills that help designers devise more and better ideas toward creative solutions (Kelley and Kelley, 2013). The term "design thinking" refers to cognitive processes of design work**,** or the thinking skills and practices designers use to create new artefacts or ideas and solve problems (Cross, 2011). Design thinking is seen as a powerful approach for encouraging imagination and active problem solving in students in school (Carroll *et al*., 2010) as well as tertiary education (Wrigley and Straker, 2017) as it challenges students to find answers to complex and difficult problems and fosters students' ability to act as agents of change. It develops students' creative confidence by engaging them in hands-on projects that focus on building empathy, promoting a bias toward action, encouraging ideation, and fostering active problem solving (Carroll *et al*., 2010).

Design and design thinking are recognized as increasingly important skills in the future economy (DesignSingapore Council, 2016), and yet the burden of assessing these complex fuzzy skills creates a scalability bottleneck. Due to its subjective nature, assessment of design approach is a challenge, especially so when design teaching is carried out with many students. Design skills are conventionally assessed through labour-intensive examination of project artefacts such as individual journals, presentations, prototypes and reports. For larger cohorts of students this may be too time-consuming, necessitating multiple assessors and introducing a potential inconsistency in the assessment. And while automation is unlikely to replace subjective assessment of design students anytime soon, computational assistance can potentially reduce the complexity of this task.

The goal of this exploratory study is to investigate a computational approach to the assessment of design thinking competency and mindset that is quick and efficiently scalable. This objective is reflected in the high-level research questions of this work, "*Can easily-acquired textual information demonstrate that a design thinking approach has been adopted, and if so, which features of easily-collected textual information demonstrate a design thinking approach?"* In this paper, the research question is addressed by performing post hoc analysis on student responses to essay-type questions from pre- and post-conditions of a 4-12 month design program called Design Odyssey (SUTD-MIT International Design Centre, 2018). More specifically, feature engineering and feature selection are applied to this text data in order to find text features that meaningfully differentiate the pre- and post-project answers. Thus, the specific research subquestion addressed in this paper is, "*which textual features most strongly differentiate between responses to design methodology questions pre-Design Odyssey and post-Design Odyssey?"* The identification of differentiating features in this context – combined with the assumption that Design Odyssey has on average improved design thinking competency and mindset – provides a basis to address the more general goal of design thinking assessment. The main contributions of this paper are:

- We identify user-centric language as a feature that emerges more prevalently after participation in the Design Odyssey program.
- We provide evidence of design concept acquisition via an increase in the SMOG index (Mc Laughlin, 1969), and possible heightened incidence of "use."
- We provide evidence of several significant features without a clear interpretation, suggesting additional data representations that could meaningfully support assessment.
- Based on the above, we propose the use of feature engineering on text data from design teaching programs as a promising approach that can be explored for efficient assessment of design skills acquisition.

Therefore, this paper has value, at one level, for the assessment of large numbers of students, and at another level, for the development of effective quantitative assessment methods. While these contributions are made in the context of design thinking assessment, other terminology (e.g., "designerly ways of thinking") may also be appropriate; "design thinking" is chosen here because the training intervention used in this study is based on design thinking. Conversely, this assessment approach ignores discipline-specific technical competencies, and is thus not appropriate for the broader category of design assessment.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

2 BACKGROUND

2.1 Assessment techniques in design education

Assessment is useful for evaluating student achievement, for identifying strategies to improve student learning, and provides evidence of capabilities in students (Davis *et al*., 2002). However, there are challenges in the assessment of design skills acquisition. Design problems are generally open ended, and require students to develop solutions that meet specific criteria and conflicting constraints that could change along the problem solving process (Jonassen *et al*., 2006). Ill-structured and loosely constrained design problems provide situations where students define their own problems and establish their own criteria and constraints, thus, allowing many possible routes to success (Bartholomew, 2017). An added complexity is that multiple instructors may be evaluating student work (Diefes-Dux *et al*., 2010). Finally, the students in a cohort may be working on different problem statements.

Assessment tools have been developed for design in the context of engineering education (Davis *et al*., 2002) and in capstone courses (Beyerlein *et al*., 2006). For the assessment of students in open ended design problems, studies have employed techniques such as rubrics (Diefes-Dux *et al*., 2010) and comparative judgement (Bartholomew, 2017; Kimbell, 2012). A number of studies focus on the assessment of creativity in design learners (Demirkan and Afacan, 2012; McLaren and Stables, 2008), employing the use of rating scales and portfolio assessment (Doppelt, 2009).

By comparison, there are few studies that research the evaluation of design thinking skills in students (Aflatoony *et al*., 2018). While one study employed exploratory factor analyses of survey based data for assessment of design thinking traits (Blizzard *et al*., 2015), others have employed manual document analysis techniques (Aflatoony *et al*., 2018; Christensen *et al*., 2016) and participant observation (Aflatoony *et al*., 2018). For the evaluation of online courses of design thinking, a review reveals that most online courses rely on traditional assessment methods based on project or assignment outcomes, self-assessment questions, reflections, peer reviews, quizzes, and exams (Wrigley *et al*., 2018).

While these works exemplify the issue that assessing fuzzy design skills remains challenging and time-consuming, feature engineering techniques from automated text assessment may reduce this burden.

2.2 Feature engineering for text assessment

The process of feature engineering – transforming raw data into a new set of descriptors that better models the data's underlying structure – is of key importance in machine learning tasks such as text classification (Domingos, 2012). The text features that support this task fall into several categories. The simplest of these is to count surface features of the text (e.g., number of spelling errors). More complex measures use the bag of words representation – the words themselves are considered without regard for syntactic or semantic structure. Features using this representation can involve (1) simply counting each word, (2) weighting word importance according to their commonness as in Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck Jones, 1972) and Latent Semantic Analysis (LSA) (Deerwester *et al*., 1990), or (3) building topics as statistical mixtures of these words as in Latent Dirichlet Allocation (LDA) (Blei *et al*., 2003). More sophisticated and more challenging to interpret, embedding-based representations such as word2vec (Mikolov *et al*., 2013) capture structure implicitly. This technique represents words in a many-dimensional embedding space where a word's position tends to correspond to its meaning (as defined by its neighbours). Doc2vec (Le and Mikolov, 2014) applies a similar technique to capture structure in a collections of documents.

Feature engineering on text data is not new in assessment. Automated Essay Scoring (AES) algorithms are trained using engineered features that are predictive of how humans score those essays (Balfour, 2013). For example, LSA has been shown to be effective when the goal is to assess acquisition of new vocabulary (e.g., from reading a passage) (Miller, 2003). More generally, a commercial AES example – the e-rater system – demonstrates that using a variety of feature types is an effective strategy for differentiating essays. This system captures elements of grammar, structure, word relevance, and more (Shermis and Burstein, 2013). Similarly, in this work we explore features from a range of representations including counting statistics, TF-IDF, LDA, and doc2vec embeddings in order to discover features that meaningfully describe text in the context of written design approaches.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

2.3 Need for scalable assessment techniques for design thinking

While there has been significant attention given to scaling text assessment outside the design context (e.g., AES systems), and increasing attention has been given to the question of assessing the design thinking aspects of a design approach, the topic of scaling design thinking assessment is relatively young. To the authors' knowledge there have been no data-driven investigations into how we might automatically infer the extent to which a written design strategy reflects a design thinking mindset. This exploratory study addresses this issue by generating and evaluating text features from student responses to direct questions about design methods. Specifically sought are generalizable features which can not only (1) capture the differences between responses of students who participated in a design training program and those who have not, but (2) can also be explained in terms of the design thinking framework. Such features would support creation of a scalable means to automatically assess design skills in students.

The following sections describe how we address this challenge by generating a wide variety of text features from students' descriptions of their design process approach, testing these features for significance, and interpreting their meanings in the context of design thinking.

3 METHOD

The method for finding text features to assess design thinking skills using written questions starts from two premises. First, it is assumed that participation in an experiential design training program called Design Odyssey changes the way students approach design. Second, it is assumed that if one's approach to design changes, then the way one discusses one's approach to design also changes. Based on these premises, we analyse student responses to six questions about general design process approaches both before and after participation in Design Odyssey. More specifically, text features are engineered and evaluated based on these responses. If any features are found to be predictive and generalizable across randomized splits of the data, it suggests both (1) an impact of Design Odyssey and (2) a reusable textual identifier of a design thinking approach.

3.1 Design Odyssey program

The Design Odyssey program (SUTD-MIT International Design Centre, 2018) is an initiative of the SUTD-MIT International Design Centre based at the Singapore University of Technology and Design. The program started in 2016, and now caters to students from secondary schools to university students, with the aim to nurture human-centred leaders, as change agents for innovating the future for Singapore and beyond. In the Design Odyssey program, students work in teams of two to five, on a social innovation project of their choice. The program aims to deepen the knowledge, adaptation, and practice of Design Innovation (a variety of design thinking using the UK Design Council double diamond, see (Camburn *et al*., 2017)).

3.2 Data collection

This paper reports on a post hoc analysis of data from the second run of the Design Odyssey program, called Design Odyssey 2.0, that ran from Sept 2017 to Aug 2018. Assessment was carried out by the Design Odyssey program managers, independent of the research team. It was done in two stages: (1) pre-Design Odyssey assessment, carried out two weeks before the program commenced, and (2) post-Design Odyssey assessment, carried out four weeks before the program ended. For both stages the assessment instrument was a google form with text questions that was designed to be completed in 30 minutes; a link was emailed to each student. The instrument described a locally relevant design issue in pictures and words, followed by questions about developing a solution for the issue. For the pre-Design Odyssey assessment, the design issue pertained to elderly cleaners in Singaporean hawker centres and food courts. In the post-Design Odyssey assessment, the issue pertained to bike sharing services in Singapore. Students were also asked demographic questions (school, gender, birth year, and nationality).

Assessment data was collected from four Design Odyssey cohorts – three situated in Singaporean polytechnic schools and one situated in the Singapore University of Technology and Design. Due to collection methodology inconsistencies, only data from one polytechnic and the university are analysed in this study. In this data subset there are in total 81 student responses – 24 from the polytechnic and 57 from the university. Of these, the polytechnic group generated 18 pre-Design Odyssey and 6 post-Design Odyssey responses, while the university group generated 41 pre- and 16 post- responses.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

3.3 Data analysis

Of the data collected from these two school cohorts, all responses to all six questions about design methods were selected for analysis (summarized in Table 1). Additionally, demographic data were retained for use as features; all other data were discarded.

Table 1. Summary of Questions asked before and after Design Odyssey

ID	Question
Q0	Describe how you would you lead a great team to join you and what the qualities of a great team are. [sic]
Q1	How will you arrive at a clear and well-scoped problem statement?
Q2	How would you prepare a killer pitch to investors or your clients?
Q3	What do you think is a good prototype and how will you test your prototype with your users?
Q4	What methods will you use to generate quality ideas and concepts for the problem you have identified?
Q5	What will you do to identify your users and understand your user needs?

Each question was analysed separately using the procedure described in the subsequent sections, summarized in Figure 1. All features were generated as a batch in 7.10 seconds in a single unoptimized Jupyter notebook (Kluyver *et al*., 2016) on a 2.7 GHz Intel Core i7 MacBook Pro with 16 GB of RAM. The remaining steps were performed on a question-wise basis where each analysis was completed in approximately 9 seconds (also unoptimized). Human intervention is not required at any point in the analysis except for interpreting the outcome of feature reconciliation.

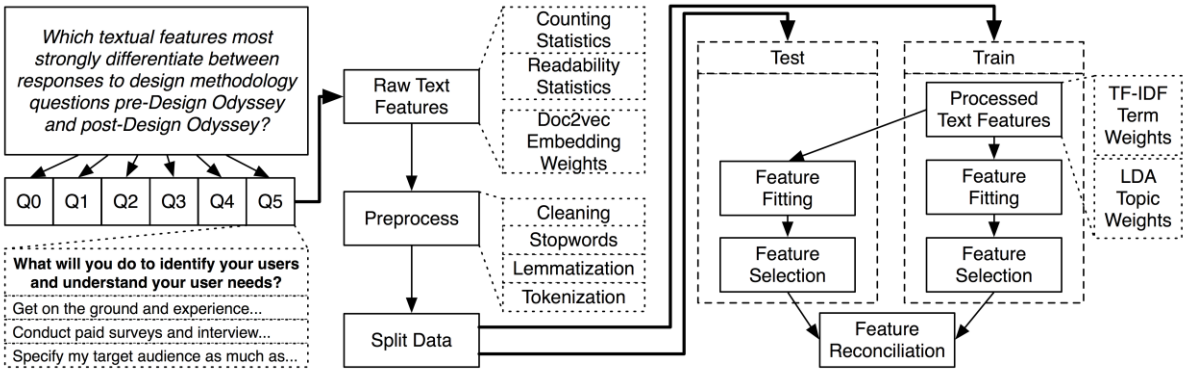

Figure 1. Analysis Process Summary

3.3.1 Raw text features

After organizing the data, a variety of features were generated from the responses to each question. First, simple text statistics were generated using the Textacy python library ("Textacy", 2018). These include counting statistics (e.g., the number of sentences) as well as a variety of readability statistics (e.g., the SMOG index (Mc Laughlin, 1969), which predicts the reading grade level based on number of polysyllabic words and number of sentences).

Next, following basic tokenization, the gensim (Řehůřek and Sojka, 2010) implementation of doc2vec (Le and Mikolov, 2014) was used to generate an embedding vector for each response; each element becomes a new feature. Essentially, doc2vec places each response into an embedding space, similar in principle to principal component analysis (PCA) (Abdi and Williams, 2010). Also like PCA, each component of a doc2vec embedding does not have a straightforward interpretation. In this experiment, all responses for all questions were represented by a 15 dimensional doc2vec embedding vector. The doc2vec parameters – including number of dimensions – did not undergo extensive tuning.

3.3.2 Preprocess text

The rest of the features require additional text preprocessing, but are also more interpretable. In this step, each statement underwent a variety of cleaning and normalization processes including tokenization; removal of formatting symbols, digits, and English stopwords; and lemmatization.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

3.3.3 Split data

Responses for each question were then split into a training set and a test set. The remaining features were engineered using only the training set data; the test set was used to check for generalizability as a mitigation against multiple hypothesis testing. Responses were assigned randomly into one of the two sets, and this sampling was stratified according to whether the response was given pre-Design Odyssey or post-Design Odyssey.

3.3.4 Processed text features

Next, two types of content features were generated from the preprocessed text. The first of these was a set of term frequency-inverse document frequency (TF-IDF) features, which were generated for only the training set. These features were generated using the scikit-learn (Pedregosa *et al*., 2011) TfidfVectorizer model with options for sublinear term frequency and L2 normalization to control for the length of each response. Additionally, the set of all terms for this model included both unigrams and bigrams, all terms that appear fewer than twice were ignored, and inverse document frequency weighting was used. Each response was represented by the full set of these features – one feature for each word in the collective vocabulary of all responses. Each feature is roughly interpretable as the extent to which a given term is used in that response, scaled up by the uniqueness of that term across all responses.

In addition to TF-IDF features, topic features were generated via Latent Dirichlet Allocation (LDA) (Blei *et al*., 2003) using the scikit-learn LatentDirichletAllocation model. Five topics were generated for each question. The number of topics was selected empirically to maximize the number of topics while minimizing term overlap between topics, but this parameter did not undergo extensive tuning. In LDA, topics are generated as probabilistic mixtures of many words that tend to capture the distinct topics across a collection of documents. Thus, each LDA topic feature is interpreted here as the collection of words that have the highest probability in that topic. While these TF-IDF and LDA features were generated based entirely on the training set, the same features were applied to describe both the training and test sets.

3.3.5 Feature selection and reconciliation

After re-representing the training and test data in terms of these features, feature selection was performed to identify which are predictive of the difference between pre-Design Odyssey responses and post-Design Odyssey responses. ANOVA F statistics and accompanying p-values were generated for all features in both data sets. The effect size of each feature was also calculated as the difference between the mean of all pre-Design Odyssey values of that feature and the mean of all post-Design Odyssey values of that feature. These effect sizes retain their units (i.e., they are not normalized) and are only comparable to other features of the same type.

A feature that is fitted on the training data and strongly predictive on both training and test data has the best chance of generalizing to other datasets and is thus promising as a repeatable assessment metric. In the reconciliation step, features that have low p-values and large effect sizes in both training and test sets were identified as potentially generalizable features for differentiating between pre-Design Odyssey and post-Design Odyssey responses to questions about design methods.

4 RESULTS & DISCUSSION

A summary of the most predictive reconciled features for each question is shown in Table 2. Several of these features represent potentially meaningful evidence of design thinking skill acquisition. Because this is an exploratory study, the discussion will entertain a higher chance of false positives than the traditional 0.05 significance cut off for p-values. In addition to features that meet this cut off, this section also examines several features that rank highly across both training and test groups (i.e., those with best-in-class F statistics).

Without discussion, these results directly address the most granular of our research questions, "*which textual features most strongly differentiate between responses to design methodology questions pre-Design Odyssey and post-Design Odyssey?"* The following discussion will elaborate on and contextualize these results in an attempt to address the more general research question, "*which features of easily-collected textual information demonstrate a design thinking approach?"*

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

Table 2. Most predictive features for each question

ID	Feature Method	Feature Description	Train			Test		
			F Score	*p* Value	Effect Size	*F* Score	*p* Value	Effect Size
Q0	TF-IDF	solve problem	2.937	0.095	0.036	2.854	0.099	0.037
Q1	doc2vec	12	5.886	0.020*	-0.167	3.587	0.066	-0.266
Q2	doc2vec	6	3.193	0.082	-0.080	7.758	0.008*	-0.098
Q3	basic	SMOG index	3.037	0.089	2.032	3.837	0.057	2.119
	TF-IDF	use	2.752	0.105	0.066	3.774	0.059	0.108
Q4	doc2vec	1	4.024	0.052	0.186	3.050	0.089	-0.105
	TF-IDF	use	3.232	0.080	0.056	9.543	0.004*	0.149
Q5	TF-IDF	bike	15.102	0.000*	0.171	21.101	0.000*	0.240
	TF-IDF	user	6.208	0.017*	0.114	12.994	0.001*	0.159
	doc2vec	9	5.543	0.024*	0.325	5.434	0.025*	0.154
	LDA	survey, bike, user, ask, ground, shoe, people, understand, problem, work	4.321	0.044*	0.226	4.251	0.046*	0.161
	doc2vec	12	4.256	0.046*	-0.275	8.110	0.007*	-0.246
	LDA	problem, face, elderly, interview, food, problem face, centre, interview elderly, research, hawker	4.135	0.049*	-0.198	6.498	0.015*	-0.218
	TF-IDF	use	2.980	0.092	0.066	16.883	0.000*	0.180

* $p < 0.05$.

4.1 User-centric vocabulary

Analysis of Q5: "What will you do to identify your users and understand your user needs?," resulted in a relatively large number of significant features as shown in Table 2. Unfortunately, most of these significant results capture problem-specific terminology and thus cannot be generalized for assessment in other contexts. More precisely, the TF-IDF feature "bike," the two doc2vec features, and both LDA features in the table clearly reflect the differences between the assessment prompts. This effect also shows that these features are capable of capturing meaningful vocabulary differences in this context, which serves as a positive ad hoc verification of the method.

This leaves the TF-IDF features "user" and "use," both of which showed significant increases in the post-Design Odyssey condition and are not clearly attributable to the differences between prompts. It is possible that this increased incidence of user-specific terminology indicates a more user-centric and thus empathic mindset – a key aspect of design thinking (Kelley and Kelley, 2013). For example, Table 3 shows both a high scoring and low scoring response for the TF-IDF feature "user." While the low-scoring approach details a more comprehensive investigation, the high-scoring approach has a singular focus on direct and even empathic methods of user research (e.g., "using these apps myself").

Table 3. Examples of statements about user analysis and corresponding tfidf_user scores

What will you do to identify your users and understand your user needs?	**Preprocessed Text**	**tfidf_user score**
"I would examine the experience of using bike-sharing apps by interviewing users and also by using these apps myself. I would see patterns in user behavior and user types, to form personas about the main and extreme users"	examine experience use bike sharing app interview user use app pattern user behavior user type form persona main extreme user	0.408
"I will need to do needs analysis with all the relevant stakeholders (if possible) through interviews. research and observation. It includes going down to the environment where the users are comfortable with and having a chat/ observing with them, interviewing relevant authorities/ specialist and doing online research on how other countries are handling the situation."	need need analysis relevant stakeholder possible interview research observation include environment user comfortable chat/ observe interview relevant authorities/ specialist online research country handle situation	0.117

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

It is also noteworthy that the TF-IDF feature "use" appears as a highly ranked feature with borderline significance for the three questions regarding prototyping, idea generation, and user research. In all three cases this feature increases in the post-Design Odyssey condition. Explanation is challenging due to the versatility of the word "use"; while in some cases it is used to communicate an experiential approach (e.g., "using these apps myself"), it can also communicate knowledge of a method without explaining how it works (e.g., "use brainstorming"). As a consequence, this feature may capture a combination of user-centric mindset and knowledge of design methods.

4.2 SMOG index

From the analysis of Q3: "What do you think is a good prototype and how will you test your prototype with your users?," the results suggest that the SMOG index (Mc Laughlin, 1969) is predictive. This index is proportional to the number of polysyllabic words (with 3 or more syllables), inversely proportional to the number of sentences, and is adjusted by a variety of constant factors. The results show that the SMOG index increases in descriptions of prototyping after Design Odyssey – the number of polysyllabic words increases and the number of sentences decreases. The lack of significance for either of these separate counting features (number of polysyllable words and number of sentences) indicates that the combination of both factors is meaningful.

A possible explanation is that the acquisition of design terminology (i.e., more polysyllabic words) enables more efficient communication of more complex ideas (i.e., fewer sentences). This interpretation is supported by the findings of (Aflatoony *et al*., 2018), that students gained design-specific terminology from a design thinking curriculum.

Table 4 shows several examples of responses and their corresponding SMOG indices. The lowest of these is brief, but only the word "prototype" has three or more syllables. In contrast the highest-scoring response is much longer (which would lower the SMOG index), but also includes more polysyllabic terms: prototype, demonstrate, idealized, experience, explaining, intended, creating, obtaining, functionality, and improvement. Many of these terms reflect goals similar to those taught in a design thinking framework.

Table 4. Examples of statements about prototyping and corresponding SMOG indices

What do you think is a good prototype and how will you test your prototype with your users?	**SMOG Index**
"A good prototype must be useful to user. It is a need to users instead of want."	7.2
"A good prototype will demonstrate the basic idea of the product well. I will invite several of my target audience to try my product and get feedback from them."	10.1
"I think a good prototype ought to be able to demonstrate the basic functions of the idealized end-product. I would try to approach people who fit my target user profile, and let them experience using the prototype for a few days at least, before explaining to them my intended purpose behind creating the product and obtaining their feedback on its functionality and the areas for improvement."	15.9

4.3 Doc2vec embeddings

Doc2vec embedding features approach statistical significance for multiple questions, most notably for Q1 with *p*-values of 0.020 and 0.066 for training and test groups respectively. This suggests a deeper underlying structure for discriminating between pre- and post- Design Odyssey conditions, but the more easily interpretable features in this work did not describe this effect. On a practical level, these results suggest that a crude design thinking assessment classifier built to distinguish between pre- and post- design thinking training would benefit from using this style of embeddings. On a theory level, it suggests there is more to learn about using this type of text data in this assessment context.

5 CONCLUSION

This paper presents exploratory work toward the objective of developing a computational approach to assessing design thinking skills. By engineering a variety of text features from responses to questions about design approach, it was found that increases in features related to user-centric language and method-specific concept acquisition were significantly predictive of participation in Design Odyssey. Specifically, these features are (1) the appearance of "user" and "use" in lemmatized and TF-IDF weighted text and (2)

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

the SMOG index. Additionally, several significant document embedding features suggest the presence of additional undetected features.

This work has several limitations. First, the limited sample size on a relatively limited population scope makes it more challenging to detect generalizable effects. Second, the conclusions are based on the assumption that the differences between the pre-Design Odyssey data and post-Design Odyssey data are attributable to an intended learning effect. Instead these differences could be caused by a systemic noise factor (e.g., a selection bias in who answers the assessment questions). With respect to analysis, the feature creation algorithms can be sensitive to the preprocessing and tuning parameters that go into their creation. While an effort was made to follow best practices in this regard, a full sensitivity analysis could reveal different combinations of parameters that lead to more predictive features.

In order to address these limitations and extend the contributions, there exist several avenues for future work beyond simply collecting additional data and refining algorithm parameters. The first is to conduct a more detailed investigation of the significant features found here, to investigate refinement, and to confirm the interpretations proposed in this paper. Similarly, further exploration of feature combinations and transformations may yield stronger or more varied predictors. Finally, the application of a scoring rubric to evaluate each statement according to its representativeness of design thinking principles such as empathy, collaboration, and experimentation (Aflatoony *et al*., 2018; Blizzard *et al*., 2015) would enable more granular exploration of evaluation methods.

This study highlights a significant issue in design and design thinking teaching: the need for efficiently scalable assessment techniques. As opposed to fully qualitative approaches for design thinking assessment that require manual content coding, the strength of a computational approach is that it can reduce the burden of design thinking assessment by providing relevant information to support assessor judgment. Hence, this framework can be valuable in the assessment of large cohorts of students as well as online courses of design thinking. Moving ahead, additional work is needed in order to understand the feature requirements of design thinking assessment at the same level as those of more automated text assessment methods, but studies like this one can potentially change the way assessment is performed in large design courses.

REFERENCES

Abdi, H. and Williams, L.J. (2010), "Principal Component Analysis", *WIREs Comput. Stat.*, Vol. 2 No. 4, pp. 433–459, John Wiley & Sons, Inc., New York, NY, USA.

Aflatoony, L., Wakkary, R. and Neustaedter, C. (2018), "Becoming a Design Thinker: Assessing the Learning Process of Students in a Secondary Level Design Thinking Course", *International Journal of Art & Design Education*, Vol. 37 No. 3, pp. 438–453, Wiley/Blackwell (10.1111).

Balfour, S.P. (2013), ""Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™.", *Research & Practice in Assessment*", *ERIC*, Vol. 8, pp. 40–48.

Bartholomew, S.R. (2017), "Assessing open-ended design problems", *Technology and Engineering Teacher*, Vol. 76 No. 6, pp. 13–17.

Beyerlein, S., Davis, D., Trevisan, M., Harrison, K. and Thompson, L. (2006), "Assessment Framework for Capstone Design Courses", *Proceedings of American Society for Engineering Education Annual Conference*.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *Journal of Machine Learning Research : JMLR*, Vol. 3, pp. 993–1022.

Blizzard, J., Klotz, L., Potvin, G., Hazari, Z., Cribbs, J. and Godwin, A. (2015), "Using survey questions to identify and learn more about those who exhibit design thinking traits", *Design Studies*, Vol. 38, pp. 92–110. Elsevier Ltd.

Camburn, B.A., Auernhammer, J.M., Sng, K.H.E., Mignone, P.J., Arlitt, R.M., Perez, K.B., Huang, Z., et al. (2017), "Design Innovation: A Study of Integrated Practice", *Volume 7: 29th International Conference on Design Theory and Methodology*, pp. 1–10.

Carroll, M., Goldman, S., Britos, L., Koh, J., Royalty, A. and Hornstein, M. (2010), "Destination, imagination and the fires within: Design thinking in a middle school classroom", *International Journal of Art and Design Education*, Available at: https://doi.org/10.1111/j.1476-8070.2010.01632.x.

Christensen, K.S., Hjorth, M., Iversen, O.S. and Blikstein, P. (2016), "Towards a formal assessment of design literacy: Analyzing K-12 students' stance towards inquiry", *Design Studies*, Vol. 46, pp. 125–151.

Cross, N. (2011), *Design Thinking : Understanding How Designers Think and Work*, Berg.

Davis, D.C., Gentili, K.L., Trevisan, M.S. and Calkins, D.E. (2002), "Engineering design assessment processes and scoring scales for program improvement and accountability", *Journal of Engineering Education*, Vol. 91 No. 2, pp. 211–221, Blackwell Publishing Ltd.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 391–407, Wiley Online Library.
Demirkan, H. and Afacan, Y. (2012), "Assessing creativity in design education: Analysis of creativity factors in the first-year design studio", *Design Studies*, Vol. 33 No. 3, pp. 262–278, Elsevier Ltd.
DesignSingapore Council (2016), *Design 2025*.
Diefes-Dux, H.A., Zawojewski, J.S. and Hjalmarson, M.A. (2010), "Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems", *International Journal of Engineering Education*, Vol. 26 No. 4, pp. 807–819, SI.
Domingos, P. (2012), ""A few useful things to know about machine learning", *Communications of the ACM*", *ACM*, Vol. 55 No. 10, pp. 78–87.
Doppelt, Y. (2009), "Assessing creative thinking in design-based learning", *International Journal of Technology and Design Education*, Vol. 19 No. 1, pp. 55–65.
Jonassen, D., Strobel, J. and Lee, C.B. (2006), "Everyday problem solving in engineering: Lessons for engineering educators", *Journal of Engineering Education*, Available at: https://doi.org/10.1002/j.2168-9830.2006.tb00885.x.
Kelley, T. and Kelley, D. (2013), "Creative confidence: Unleashing the creative potential within us all", *The Business Source*, Available at: https://doi.org/10.4103/0255-0857.93014.
Kimbell, R. (2012), "Evolving project e-scape for national assessment", *International Journal of Technology and Design Education*, Vol. 22 No. 2, pp. 135–155, Springer Netherlands.
Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., et al. (2016), "Jupyter Notebooks-a publishing format for reproducible computational workflows", *ELPUB*, pp. 87–90.
Le, Q.V. and Mikolov, T. (2014), "Distributed Representations of Sentences and Documents", *CoRR*, abs/1405.4, Available at: http://arxiv.org/abs/1405.4053.
Mc Laughlin, G.H. (1969), "SMOG grading-a new readability formula", *Journal of Reading*, Vol. 12 No. 8, pp. 639–646, JSTOR.
McLaren, S.V. and Stables, K. (2008), "Exploring key discriminators of progression: relationships between attitude, meta-cognition and performance of novice designers at a time of transition", *Design Studies*, Vol. 29 No. 2, pp. 181–201.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013), "Distributed representations of words and phrases and their compositionality", *Advances in Neural Information Processing Systems*, pp. 3111–3119.
Miller, T. (2003), "Essay assessment with latent semantic analysis", *Journal of Educational Computing Research*, Vol. 29 No. 4, pp. 495–512, SAGE Publications Sage CA: Los Angeles, CA.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011), "Scikit-learn: Machine Learning in {P}ython", *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830.
Řehůřek, R. and Sojka, P. (2010), "Software Framework for Topic Modelling with Large Corpora", *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, Valletta, Malta, pp. 45–50.
Shermis, M.D. and Burstein, J. (2013), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, Routledge.
Sparck Jones, K. (1972), "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, MCB UP Ltd, Vol. 28 No. 1, pp. 11–21.
SUTD-MIT International Design Centre (2018), *Design Odyssey*, Available at: https://idc.sutd.edu.sg/programmes/design-odyssey/.
"Textacy" (2018), *Chartbeat Labs*, Available at: https://github.com/chartbeat-labs/textacy.
Wrigley, C., Mosely, G. and Tomitsch, M. (2018), "Design Thinking Education: A Comparison of Massive Open Online Courses", *She Ji: The Journal of Design, Economics, and Innovation*, Elsevier, Vol. 4 No. 3, pp. 275–292.
Wrigley, C. and Straker, K. (2017), "Design Thinking pedagogy: the Educational Design Ladder", *Innovations in Education and Teaching International*, Available at:https://doi.org/10.1080/14703297.2015.1108214.

ACKNOWLEDGMENTS

This study was supported by SUTD-MIT International Design Centre based at the Singapore University of Technology and Design.

https://doi.org/10.1017/dsi.2019.396 Published online by Cambridge University Press