Proceedings of the ASME 2024 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2024
August 25-28, 2024, Washington, DC, USA

DETC2024-142788

EVALUATING DESIGN RATIONALE

Yakira Mirabito*
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, CA, USA
yakira@berkeley.edu

Xiaowen Liu
Cognitive Science Program
University of California, Berkeley
Berkeley, CA, USA
xiaowen_liu@berkeley.edu

Kosa Goucher-Lambert
Dept. of Mechanical Engineering
University of California, Berkeley
Berkeley, CA, USA
kosa@berkeley.edu

ABSTRACT

Design rationale captures the justification behind a design decision. Often, rationales vary in the content and depth of information, making the study and comparison of rationales challenging. This project aims to characterize design rationale and develop a computational approach to evaluate design rationale quality at scale. In total, 2250 rationales were machine-generated using GPT across two different representations, and a portion of the rationales (n = 512) were evaluated by two raters across five dimensions of quality. Rationales were then characterized using natural language processing techniques, resulting in 108 linguistic features for each rationale. The evaluations and linguistic features were used to build eight predictive models for each quality dimension. The main results show that structured rationales were rated higher than unstructured rationales across the five dimensions. Thus, the tested feature, specification, and evidence (FSE) framework was shown to be a worthwhile approach to represent the justification behind a design decision. Moreover, key linguistic features that correlate with higher quality ratings were identified. Future work will explore how design rationale quality and characteristics impact design decisions, particularly in a human-AI teaming context where generative design recommendations could benefit from including generative design rationales.

*Address all correspondence to this author.

1 INTRODUCTION

Design rationale is the justification behind a product component, often captured via written reports and oral presentations. Documenting design rationale is a critical aspect of the engineering design process; however, what to include and at what level of detail needs to be standardized in teaching or practice. Design rationale has many definitions, capture methods, and use cases [1]. We adopt Lee's definition of design rationale as "not only the reasons behind a design decision but also the justification for it, the other alternatives considered, the tradeoffs evaluated, and the argumentation that led to the decision" [2]. The scope of this research focuses on characterizing written design rationale.

The primary significance behind design rationale usage centers on the artifact's long-term success and financial implications. Rarely will a single individual be responsible for the entire design process; instead, engineers and designers need to work together and communicate with other members of a firm (e.g., supervisors, sales, marketing) and clients [3]. A team that spends hours developing an innovative product may see the product fail due to poor positioning or poor communication. Additionally, a firm may spend unnecessary resources repeating previous mistakes that were not documented or tracking down rationale from previously completed design iterations. Research in engineering design has also found that information and tone used to reason and explain decisions affect human behavior [4–6]. Thus, structuring design rationale in ways that enhance the design process is of high importance.

This paper aims to quantify the quality of design rationale using human evaluators and rubrics and then transfer those hu-

man evaluations into a computational tool to quickly evaluate new design rationales at scale. Human evaluators can provide a sense of whether the rationale provided is sufficient; however, doing so at scale is a time-consuming and unreasonable task. Existing critical thinking rubrics from technical writing domains were used to rate the quality of design rationale across multiple dimensions. Toward developing a tool to output a quality score, this work leverages natural language processes (NLP), which take in the written text and process the information into meaningful features that were selected and integrated into a model. In this knowledge transfer from human evaluators into a tool, we articulate the linguistic elements that correlate with higher ratings. The two specific research questions are:

RQ1: Does structuring rationale using the Feature, Specification, and Evidence framework result in higher-rated rationale than unstructured rationales?
RQ2: What linguistic features result in high-quality rationale?

The motivation behind developing an approach to evaluate design rationale quality at scale stems from the need to provide designers with actionable insights to enhance their decision-making processes and communication skills. Engineering documents tend to favor the technical aspects of the final solution but lack an explanation of the context of the process [7]. The potential impact of this research extends beyond individual skill development and could be integrated into generative artificial intelligence (AI) applications in design. For example, the tool can be valuable for evaluating and refining generative rationale approaches [8]. Overall, such a tool addresses the immediate gap of enhancing design rationale communication and holds promise for shaping future design support tools and methodologies.

2 BACKGROUND

In order to consistently differentiate rationale quality, an objective measure must be used. Currently, no commonly accepted measures exist. Thus, this research leverages human evaluators and rubrics to serve as the ground truth coupled with NLP feature extraction that helps explain the linguistic characteristics associated with higher-rated rationales. The following sections outline the importance of quality measures, the role of human raters and rubrics, and computational approaches to analyze language. This project uses similar pipelines to researchers who developed computational methods to automatically rate SAT essays based on standardized SAT rubrics [9, 10].

2.1 Measures to evaluate rationale quality

The first approach to evaluate design rationale leverages human raters and rubrics with defined scales. Selecting a rubric to evaluate design rationale first relies on characterizing what type of writing or processes occur in technical design documents. Writing rubrics tend to look holistically at an essay or book, while argumentation or critical thinking can be used in looking at smaller sections, such as design rationale. Moreover, the type of writing is technical, describing design methods, tools, and decisions compared to what might appear in an essay or novel. Technical writing centers on effectively communicating complex information, emphasizing clarity, precision, and adherence to established conventions.

2.2 Human evaluators and rubrics

To evaluate design rationale, the first approach leverages human raters and rubrics with defined scales. Selecting a rubric to evaluate design rationale first relied on characterizing what type of writing or processes occur in technical design documents. Writing rubrics tend to look holistically at an essay or book, while argumentation or critical thinking can be used in looking at smaller sections, such as design rationale. Moreover, the type of writing is technical in nature, describing design methods, tools, and decisions compared to what might appear in an essay or novel. Technical writing centers on effectively communicating complex information, emphasizing clarity, precision, and adherence to established conventions.

Meanwhile, critical thinking encompasses the capacity to assess, analyze, and synthesize information from various sources, requiring the cognitive skills to evaluate, make inferences, and draw meaningful conclusions [11]. Technical writing is evaluated at the report level, whereas critical thinking can be assessed in smaller quantities of data (i.e., condensed, paragraph-level format). As such, the distinction between the type of writing (technical vs. not) and scope (report vs. paragraph) in this context calls for carefully selecting a rubric annotators will use to provide the ground truth.

The rubric selected for this study focuses on critical thinking across five dimensions (evaluating, analyzing, synthesizing, forming arguments-structure, forming arguments-validity) [11]. The original scale from Reynders et al. uses a zero (worst) to five (best) scale, explaining what a 1-rating, 3-rating, and 5-rating should include. However, in their study, none of the raters used a zero, and what a zero rating is was not clearly defined. The same scale was used for this paper.

2.3 Computational approaches

The larger vision of this work is to use an automatic evaluation approach similar to Grammarly for design rationale. Grammarly is a cloud-based typing assistant that checks fundamental elements such as spelling, grammar, punctuation, and clarity. It also can assess more complex attributes of engagement, plagiarism, style, tone, and context-specific language. If errors are noted, Grammarly offers suggestions that you can accept. Grammarly, while helpful, focuses on established writing conventions,

whereas as a tool for engineering students and professionals, it should consider the overall content and norms of engineering.

Critical dimensions of natural language, including cohesion, clarity, coherence, and conciseness, play pivotal roles in determining the effectiveness and comprehensibility of a text. Cohesion pertains to the logical connection between sentences and paragraphs, ensuring that the text flows smoothly and transitions are seamless. Clarity focuses on the precision of language use, avoiding ambiguity, and using explicit, easily interpretable terms. Coherence addresses the overall logical structure of a text, verifying that the ideas and information are organized in a logical sequence and are mutually supportive. Furthermore, the evaluation tool should consider vocabulary diversity, grammatical accuracy, and adherence to established writing conventions.

Coh-metrix and TAACO (Tool for the Automatic Analysis of Cohesion) are two computational approaches used for feature extraction from a written text. Coh-metrix is rooted in discourse analysis principles and offers a framework for assessing text cohesion [12]. This feature extraction model examines sentence relationships, evaluating elements like references, conjunctions, and lexical choices. The 108 indices extracted from the Coh-metrix model are summarized into 11 categories noted in Figure 1 under featurization. The complete list and definitions can be seen [13]. On the other hand, TAACO calculates semantic overlap between sentences (local cohesion), paragraphs (global cohesion), and entire document (overall text cohesion) for nouns and verbs [10]. It assesses how well a text maintains consistency and logical connections with a central theme or topic. The Coh-metrix model was used in this study for feature extraction.

Work from the authors who developed the Coh-Metrix model highlighted characteristics associated with more cohesive texts, such as lexical diversity, connectivity, and word concreteness. Lexical diversity measures the variety of vocabulary in a text. Higher cohesion tends to correlate with lower lexical diversity due to repeated word usage. Connectivity measures the frequency of explicit linguistic devices (e.g., pronouns, conjunctions) used to link different parts of a text. Word concreteness assesses the syntactic clarity of expressions in a text. However, the researchers who developed the Coh-metrix model used a writing rubric from Breetvelt et al., including 15 dimensions (e.g., structure, thesis statement, evidential sentences) [12, 14]. Due to the nature of design rationale as technical writing rather than essays or novels that were used to develop the Coh-Metrix model, the rubric selected for this study focuses on critical thinking across five dimensions (evaluating, analyzing, synthesizing, forming arguments-structure, forming arguments-validity) [11].

2.4 Feature, Specification, and Evidence (FSE) Framework

The representation of design rationale is not standardized. Thus, the content and depth of information included in a design rationale varies. One framework that aims to enhance communication and documentation practices of design rationale is the feature, specification, and evidence framework [15]. The main content of the framework is described below. Key to this project, the framework was used to generate half the data to assess the effectiveness of the framework.

Feature (F) describes an artifact's design component or attribute that the rationale serves to justify, such as a brake pad, steering wheel, or tire. In general, the feature should meet a specification. The breakdown of which features to include in reporting can be best defined by the firm or industry. For example, the exact bolt material might need explicit rationale; however, features are more likely to serve as a critical component of the final solution.

Specification (S) describes the stated design requirement(s) the feature aims to address, defined in the early stages of the design process, such as slowing down the vehicle, steering the vehicle, or maintaining contact with the road. These specifications are noted early in the process, and existing tables may be referenced; however, authors of design rationale need to be explicit about which specification they are referring to rather than cite an entire table. A feature can address more than one specification, and multiple features can address a single specification.

Evidence (E) describes the relevant information from that design process that empowered the designer to select the final feature that meets the specification(s), such as interviews, background research, or product testing. For example, testing alternative braking mechanisms or brake pad materials is considered. The evidence is the meaningful output acquired using design methods. Thus, designers might include the tests used and the results from those tests influencing their decisions.

3 METHODS

Overall, this project quantifies the quality of design rationales as noted in Figure 1. In essence, what characterizes a 'good' design rationale or justification behind a design decision? To do so, a dataset of design rationales was generated for 25 consumer products (each rationale containing about 100 to 200 words). This dataset was then evaluated using human raters on five dimensions [11]. The raw text was processed using feature extraction by Coh-Metrix, resulting in 119 features [12]. Afterward, careful feature selection was performed, allowing for the most relevant features to be used in building a predictive model. A predictive model would enable designers to evaluate new design rationales at scale. The tool would support design communication and serve as a tool to assess AI-generated rationales.

For this study, the independent variables of interest are the design rationales. The key dependent variable is quality, which needs to be defined via a set of features in a model. The main components of this project entail data collection, data labeling, featurization, model building, and evaluation. The exact nuances

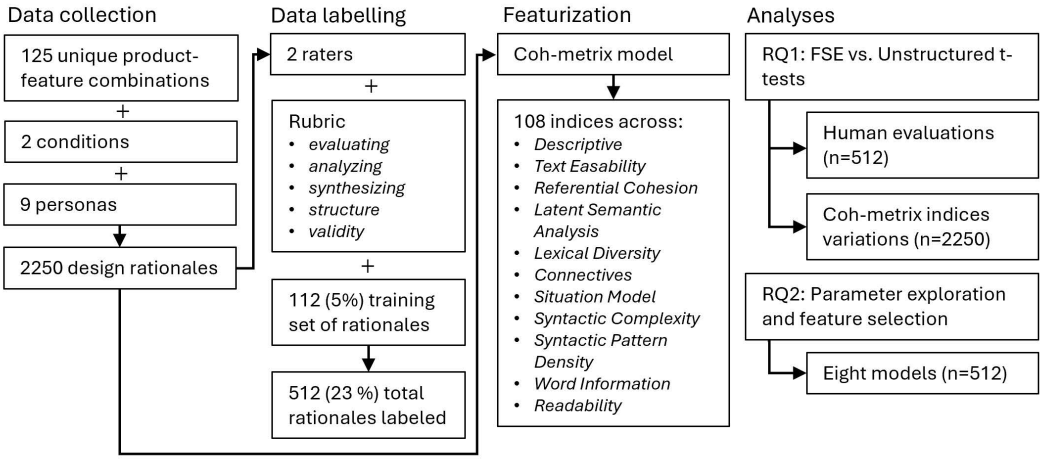

FIGURE 1. Overview of research design.

and plans for each are described in the following sections.

3.1 Dataset

A total of 2250 design rationales were generated for 25 products (Table 1) with five features under two conditions with nine personas. Representative rationales for both the *Unstructured* and *FSE* conditions are shown.

25 products x 5 features x 2 conditions x 9 personas = 2250 rationales

Unstructured rationale - The walking belt on a treadmill serves a crucial purpose in effectively simulating the natural process of walking or running. As an engineer, when designing such a system, there are two primary goals: firstly, to create a user-friendly device, and secondly, to replicate the actual physical experience as closely as possible. The belt provides a continuously moving platform that imitates the ground movement one would experience while walking or running outside. This allows the users to stay in one place, making it safer and more convenient. Additionally, the belt's surface presents a consistent, predictable terrain, reducing the risk of injury. Advanced treadmills also adjust the belt's speed in real-time based on the user's pace, ensuring a more customized workout. The integration of cutting-edge technology in the belt system—like cushioning for shock absorption, enhanced grip for stability, and digital control for speed adjustments—improves the overall functionality of the treadmill, making it an indispensable tool for indoor fitness routines.

FSE structured rationale - The cover on an electric pressure cooker (F) is designed to ensure safety and pressure regulation, both crucial facets of using the device. The specification (S) states that the product must safely contain high pressure and temperature, avoid accidental opening, and allow an effective release of pressure when necessary. The cover is equipped with safety features including a locking mechanism to prevent accidental opening and a pressure release valve for controlled steam venting. The selection of these features is backed by substantial evidence (E). In user studies and interviews, safety was consis-

TABLE 1. List of 25 consumer products.

Product
electric toothbrush
coffee maker
road bike
microwave
blender
toaster
electric kettle
hearing aid
electric shavor
hair dryer
treadmill
electric standing desk
ceiling fan
stand mixer
electric pressure cooker
drill
table saw
random orbital sander
magnetic rowing machine
seated leg press
segway
projector
gas weed eater
gas leaf blower
manual blood pressure monitor

tently emphasized as a top priority for users when engaging with pressure cookers. Moreover, background research on cooker related accidents revealed many incidents occurred due to improper pressure management or accidental opening of the cooker under pressure. Historical product testing also showed that a well-designed cover can significantly reduce these risks. Therefore, the cover has been designed to meet these user-specific needs and safety standards of the industry, leading to a better, more intuitive user experience with the device.

3.2 Data collection

3.2.1 Materials GPT 4.0 from OpenAI was used to generate the rationales, which required an API key and Python terminal. Considering GPT is known to produce similar responses, nine personas were used to increase variability in the responses produced. Prior research has shown that including personas within GPT prompts has improved response variability to various prompt engineering tasks [16]. The personas are based on role titles that human subject participants of experimental studies might hold, including mechanical engineer and industrial designer. Gender-neutral names and pronouns were used when providing GPT with a description of each persona. The personas varied on two dimensions (form-function and experience) and were used to increase variability. Form-function sought to capture information related to domain expertise, while the experience was captured by role title (e.g., entry-level, senior). A full list of titles and experience levels can be seen in the Appendix.

3.2.2 Procedure One hundred twenty-five unique prompts were asked from 25 consumer products, each containing five features. For example, "What is the rationale behind the walking belt on a treadmill?" Prompt engineering was required to output the desired response from GPT, which focused on limiting the number of words GPT used. To answer RQ1, which evaluates if the FSE framing is better, two conditions were used (unstructured and FSE structure) as noted in Table 2.

3.3 Data labeling

Annotators were required to have prior experience assessing student or employee reports in an engineering or design context (i.e., education or practice). Two raters in total used a five-dimension rubric [11] that assesses critical thinking (evaluating, analyzing, synthesizing, forming arguments (structure), forming arguments (validity) on a one (worst) to five (best) scale, using whole integers. Annotators were provided the list of rationales as a CSV in which they were trained on a fraction of the data (approx. 112 or 5% of the total dataset). Annotators were blind to the conditions.

Intraclass correlation (ICC) estimates and their 95% confidence intervals were calculated using RStudio package version (irr) based on a mean-rating (k = 2), consistency-agreement, 2-way random-effects model. The resulting values for 112 rationales were: evaluating (.94), analyzing (.79), synthesizing (.83), structure (.79), and validity (.85). Values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability [17].

TABLE 2. Prompt instructions for two conditions

Condition	Instructions
Unstructured	Please write your rationale (approx. 100 to 200 words) in a single paragraph format.
FSE	Please write your rationale (approx. 100 to 200 words) in a single paragraph format using the FSE framework. Feature (F) describes an artifact's design component or attribute that the rationale serves to justify. In general, the feature should meet a specification. Specification (S) describes the stated design requirement(s) the feature aims to address, defined in the early stages of the design process. Evidence (E) describes the relevant information from that design process that empowered the designer to select the final feature that meets the specification(s), such as interviews, background research, or product testing.

$$ICC(3,k) = \frac{MS_R - MS_E}{MS_R} \tag{1}$$

3.4 Featurization

The NLP approach used the Coh-Metrix model to produce numerical values for linguistic and discourse representations of each rationale. Coh-metrix was provided an input CSV of the 2250 rationales, and the output produced numerical values for 108 indices across 11 categories (descriptive, text easability, referential cohesion, latent semantic analysis, lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information, and readability). For the complete list of linguistic indices the model generated and their corresponding definitions, see [13]. These NLP analyses run within the Coh-Metrix tool helped characterize each rationale with information such as word count or lexical diversity. Thus, the resulting indices serve as features that were then explored and selected in the model-building phase.

3.5 Model building

This project aimed to predict the quality of design rationale on five dimensions [11]. Additionally, trends between linguistic features and quality measures were identified. Due to the immense burden on human raters to evaluate the rationale solution set, the trained experts hand-coded a random sampling of 512 out of the 2250 total rationales on an ordinal scale from 1 to 5. Considering that a portion of the data was double-coded by each annotator, an average rating for each dimension of rationale quality was used. Thus, the output of the prediction task results in a continuous variable also ranging from 1 to 5 (e.g., 3.45).

Identifying and selecting relevant features is crucial to tackling this regression problem. Using the Coh-Metrix model, one hundred-eight linguistic features were generated for all 2250 design rationales. To reduce data dimensionality, the data was standardized before running PCA and employing KneeLocator to identify the elbow point. This point indicates the optimal number of features for maximal variance explainability (n=31).

Next, model selection considered several factors: the task's regression nature, the dataset's size (2250 samples), and the desire for a model that balances simplicity with explanatory power. Therefore, simplicity guided the model selection process, and RMSE was used to assess error rates, while AIC and BIC were explored for complexity evaluation. The models used were:

Dummy Regressor: Serves as a baseline model for comparison. This model gives predicted values based on simple strategies disregarding input data.

Linear Regression: Selected for its simplicity and interpretability, linear regression is an effective model. It offers straightforward insights through model weights. The model was refined by incorporating PCA and regularization, using cross-validation to optimize regularization parameters.

Random Forest: Selected for its robustness to outliers and ability to capture non-linear relationships. Random Forest constructs multiple decision trees on data and feature subsets, minimizing the impact of outliers and enhancing model resilience.

Gradient Boosting Machines (GBM): GBM, similar to Random Forest, effectively captures non-linear relationships and surpasses linear regression in predictive accuracy. GBM incrementally corrects errors from weak prediction models (e.g., decision trees) and systematically improves accuracy.

The dataset was split into training and validation sets for model training and evaluation. This approach allowed us to train models on the training data and assess accuracy using the validation set. We used RMSE as the primary accuracy metric and utilized grid search cross-validation to fine-tune hyperparameters.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

4 RESULTS

Design rationales were collected (n = 2250), and a subset (n=512) was evaluated by human raters and used in building a model that could predict design rationale quality for new rationales. The following sections outline the findings of the two research questions. The first focuses on comparisons across FSE structured and unstructured rationales for both human ratings and Coh-metrix indices. The second finding focuses on model development that aims to predict values for each of the five rubric dimensions.

4.1 FSE structured rationales were rated higher across the five dimensions (n=512)

To compare the rated rationales of FSE and unstructured, t-tests comparing the means of each grouping were shown in Table 3. Results for all five attributes were statistically significant, meaning that the two groups are not the same, and thus, we reject the null hypothesis that they are the same. Across the five rubric dimensions, FSE structured rationales were, on average, higher rated than unstructured rationales. Figure 2 visualizes the mean rating for each condition per quality measure.

The 'evaluating' dimension assesses a designer's ability to determine the relevance and reliability of information. Rationales with high 'evaluating' scores identified information derived from product testing or user interviews. For example, the rationale behind the container in a blender, "to support the design of the blender's container, comes from numerous user tests and feedback." Rationales with high 'analyzing' scores indicated an ability to extract patterns from data that could be used as evidence, while 'synthesizing' is the ability to connect information. For example, "Users remarked on the importance of seeing their food as it's blended, ensuring proper texture and consistency - suggesting transparent material." Meanwhile, 'structure' assesses holistically the degree to which evidence and reasoning are clearly linked. 'Validity' identifies the degree to which a claim, evidence, and reasoning are consistent with disciplinary standards (in this case, engineering design). Validity relies on domain expertise and experience in design.

4.2 The majority of linguistic features across FSE and unstructured differed (n=2250)

For each condition (unstructured and FSE), the Coh-metrix indices were averaged and tested to determine which linguistic features were statistically different between the two groups

TABLE 3. Means, t-statistic, and p-values for the five dimensions.

	M_{FSE}	M_{Unstr}	T-stat	P-Value
Evaluating	4.23	1.28	44.9	2e-179
Analyzing	4.16	2.54	19.1	8e-62
Synthesizing	4.27	2.70	19.7	8e-65
Structure	4.56	2.91	21.1	2e-71
Validity	4.11	2.05	25.8	1e-94

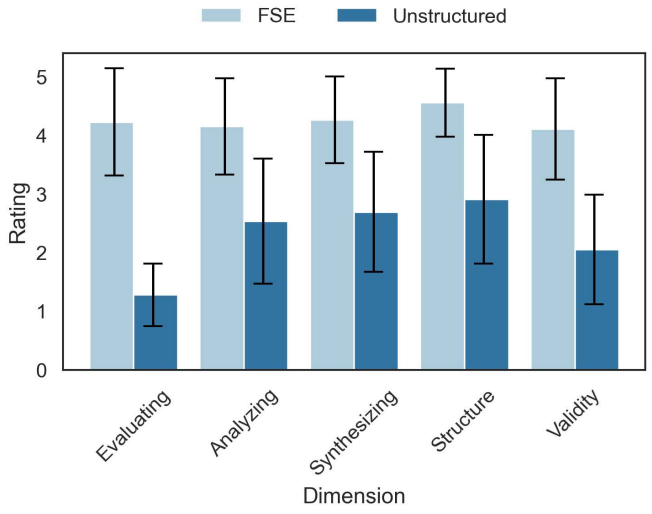

FIGURE 2. Average ratings for FSE and Unstructured for each of the five dimensions of rationale quality: a) evaluating, b) analyzing, c) synthesizing, d) structure, and e) validity.

using Kruskal-Wallis tests. Eighty-one of the 108 Coh-matrix indices had p-values less than .05, which means these linguistic features were different across the two conditions, which was unlikely due to chance. The condensed list of 11 categories is shown in Figure 1 although the detailed list and definitions are detailed in [13]. The indices from Coh-metrix with the largest H-statistic, meaning the most significant difference of normalized means, are shown in Table 4.

The analysis of six Coh-Metrix indices at a high level revealed intriguing insights into the characteristics of design rationales and their quality (Figure 3). Firstly, the indices DESWC (Word count), WRDHYPnv (Hypernymy for nouns and verbs), and PCTEMPp (Temporal cues and consistency) were positively correlated with increased rationale quality. A higher word count may indicate more detailed information, offering deeper insights into the design decision-making process. Similarly, greater hypernymy for nouns and verbs suggests a higher level of abstraction and conceptual clarity within the rationale, contributing to its coherence and effectiveness. Moreover, texts with more cues about temporality and consistent temporality are easier to process, aligning with prior literature suggesting that temporal cohesion enhances the reader's comprehension and situational understanding of textual events.

TABLE 4. Means, H-statistic, and p-values for the six most distinct Coh-metric indices.

	M_{FSE}	M_{Unstr}	H-stat	P-Value
Word count (DESWC)	163	147	531	2e-117
Hypernymy for nouns and verbs (WRDHYPnv)	2.53	2.33	457	1e-101
Temporal cues and consistency (PCTEMPp)	45.5	71.9	414	6e-92
Temporal cohesion (SMTEMP)	.831	.926	395	8e-88
Adverb incidence (WRDADV)	35.8	48.0	287	2e-64
Pronoun incidence (WRDADV)	11.4	18.2	274	1e-61

Conversely, SMTEMP (Temporal cohesion), WRDADV (Adverb incidence), and WRDPRO (Pronoun incidence) resulted in negative correlations with rationale quality. Decreased temporal cohesion, characterized by tense and aspect repetition, may indicate a lack of clarity or coherence in the timeline of events described within the rationale. Moreover, higher incidence scores of adverbs and pronouns could potentially signify verbosity or ambiguity within the rationale, hindering its clarity and effectiveness in communicating design decisions. These findings underscore the importance of coherence, clarity, and conciseness in design rationales, as reflected by their specific linguistic features extracted through the Coh-Metrix analysis. Further exploration of the other linguistic features may offer valuable insights into enhancing design rationales' quality and communicative efficacy in various contexts, ultimately contributing to more informed design decision-making processes. Aside from enhancing design communication for human designers, computational agents or generative rationale tools could use these indices to guide the writing characteristics that good explanations or rationale should contain.

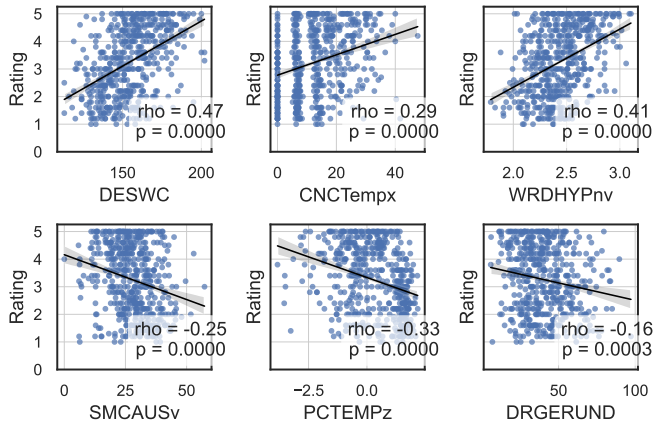

FIGURE 3. Scatter plots for the top six Coh-metrix indices plotted against the average rating for the five dimensions of quality. Each data point represents one design rationale (n=512).

4.3 Towards model development combining ratings and indices

To work towards building a model to automatically evaluate the quality of design rationales, both human evaluations and Coh-metrix indices were combined. Figure 3 plots the top six indices with the average ratings across the five dimensions of quality (evaluating, analyzing, synthesizing, structure, and validity) as part of the feature selection process. Their correlation coefficient (spearman's rho) and p-values are shown.

Root Mean Squared Error (RMSE) is the primary means for model comparison. Table 5 shows the RSME results for each of the five rubric dimensions. The RMSE represents the root mean squared difference between the predicted and actual values. A lower RMSE is generally better, indicating less error between predicted and actual values. Seven models were compared against the baseline model (dummy regressor). The DummyRegressor generates predictions without considering the input features, serving as a reference point for comparison against more complex regression models. For three of the five dimensions, linear regression with PCA and regularization showed the lowest RMSE (i.e., evaluating, synthesizing, and validity). A gradient-boosting machine model without PCA best predicted the analyzing dimension. Random forest models performed best when predicting structure values.

5 DISCUSSION

The study has three main findings: first, that FSE structured rationales were rater higher than unstructured rationales across the five dimensions; second, identification of the most distinctive linguistic features across FSE and unstructured conditions using an NLP approach (Coh-metrix); and lastly, preliminary results for a predictive model that evaluated design rationale quality. The following section discusses the findings, implications for design research, and limitations.

The statistical tests between FSE and unstructured were statistically significant across each of the five dimensions (n=512) [11], meaning that the rating differences were not due to chance. The comparison between unstructured and structured design rationales resulted in statistically significant differences in perceived quality and effectiveness. On average, unstructured rationales received lower ratings, showcasing the impact that structured representations of design rationale can have. In firsthand analyses of the generated rationales, unstructured rationales were noted to often focus on describing the form and function of the product without delving into the underlying decision-making processes. These lower-rated rationales are aligned with what might appear in student reports or patent data, reinforcing the need to improve the norms in education and practice [15]. Often, these descriptions are associated with the product's function, structure, or behavior [18]. While the FBS information is helpful in the broader design process, the information does not provide reasoning or justifications for a design decision.

In contrast, FSE structured rationales connected their design decisions with evidence and information obtained throughout the design process. This finding suggests that structuring design rationales according to the FSE framework facilitates the integration of critical thinking elements, such as evidence-based reasoning and justification, into the explanatory nature of rationales. For example, in Section 3.1, evidence supporting the designer's decision stems from several reliable and valid methods that a designer would use, such as interviews, background research, and product testing.

Moreover, of the 108 total linguistic features extracted from the Coh-metrix analysis from the total rationale dataset (n=2250), 81 were statistically different. Meaning that these indices compared across both conditions were not due to chance. Thus, there are fundamental differences linguistically across the two conditions. These features, coupled with human ratings, enabled feature selection to be performed, resulting in seven ML models that were compared and evaluated. The features were selected based on their predictive power and interpretability.

Each extracted Coh-metrix feature contains sufficient documentation on approaches to increasing or decreasing those ratings from a writing standpoint. Lei et al. have shown the relevance of using Coh-metrix to assist technical students in improving their writing [19]. Future work will build on the models in this study. The remaining questions include whether rationale quality should have these five dimensions or be condensed into a holistic measure. For example, the dimensions could be averaged, multiplied, or contain differing weights to produce an overall measure like the innovation measure [20].

TABLE 5. Root Mean Squared Error (RMSE) for the five rubric dimensions across eight models. The bolded value represents the model that had the lowest RMSE for a given dimension.

	evaluating	analyzing	synthesizing	structure	validity
Model 1: Dummy Regressor	1.61	1.28	1.23	1.26	1.39
Model 2: Linear Regression without PCA	7.20	11.1	9.99	10.6	12.0
Model 3: Linear Regression with PCA	1.20	1.20	1.16	1.23	1.19
Model 4: Linear Regression with PCA and Regularization	**1.19**	1.19	**1.15**	1.22	**1.19**
Model 5: Random Forest without PCA	1.24	1.20	1.17	**1.19**	1.22
Model 6: Random Forest with PCA	1.32	1.20	1.17	1.21	1.22
Model 7: GBM without PCA	1.30	**1.18**	1.17	1.26	1.21
Model 8: Best GBM with PCA	1.25	1.21	1.16	1.23	1.21

5.1 Design rationale contents and structure

This paper adopted Lee's definition of design rationale, including not only the justification for a design decision but also the consideration of alternatives, tradeoffs, and the argumentation process [2]. In reality, most design rationales—whether human or machine-generated— often fall short of including all these elements. Across the machine-generated rationales and prior work with human-generated rationales [15], the content frequently included descriptions of a product feature. Better-rated rationales included argumentation linking how the feature meets the desired specifications. For instance, regarding the rationale behind a battery in a hearing aid, "considering the target user group, comprising primarily of the elderly, the battery design adheres to easy handling principles. It ensures users can comfortably change the battery themselves without encountering challenges linked with fine motor skills decline." This excerpt says the design ensures comfortable changing of the battery, but empirical evidence from the design process is absent. An improved version might include the alternatives considered and more information on how they were evaluated (tradeoffs).

In situating the rationales generated in this study within the broader design rationale research, it is important to acknowledge other characteristics of design rationale that could impact perceptions of quality, such as the type of logical reasoning used (deductive, inductive, abductive) [21]. Deductive reasoning might use ergonomic principles to justify which design is better without actually testing with users. Inductive reasoning might use information from product testing of two designs to justify which is better at meeting user needs. Abductive reasoning might justify the decision with some of the least amount of information available, an educated guess. Dong et al. showed that the logical framing structure significantly influenced design decisions [5]. In user-centered design, usability testing methods mainly rely on inductive reasoning by observing user interactions and collecting data across different contexts [22].

Documenting design rationale is essential, but what information to include and at what level of detail is not standardized in teaching or practice [23]. The lack of specificity of what should be included in a design rationale can be attributed to the differing use cases and representations [24]. This paper generated rationales using the feature, specification, and evidence framework. However, several alternative representations of design rationale exist in the broader literature. Two dominant process-based representations include issue-based information system (IBIS) [25] and Questions, Options, Criteria (QOC) [26]. Both are more expressive and laborious than the FSE framework, often requiring graphical network software [27]. These alternative representations could be evaluated in future work using human or machine evaluations outlined in this paper.

5.2 Implications

5.2.1 Design research Engineering design research commonly relies on verbal and written data from engineers and designers to understand underlying processes and decision-making, often captured in project reports or think-aloud protocols. This reliance assumes that engineers and designers are good communicators, and thus, their rationales and explanations are ground truth. How can we accurately compare rationales if no standards or methods exist to evaluate these rationales or explanations? This research shows the FSE framework as a means to improve rationale quality. Moreover, the preliminary models presented provide a tool for evaluating the rationale quality. The design research community can leverage this tool to characterize and evaluate both human-generated rationales and machine-generated rationales. Educators could implement the FSE structure in reporting standards for design classes or use the computational tool to evaluate rationale quality to help students iterate on their rationale writing abilities.

5.2.2 Human-AI teaming Machines can assist designers by making design suggestions often based on human behavior (e.g., alternative CAD designs). Such suggestions often mimic human behavior but can not adequately explain the rationale behind the design recommendation. Therefore, there is a need to supplement automated design recommendations with design rationale [28]. Overall, this research aims to address this research gap by providing standards for 'higher-quality' design rationale to justify the design recommendations being made by computational agents. Building on prior research that demonstrates human explanations have a large variance in quality [15], our findings advocate for the integration of the FSE framework and computational tools to improve generative rationales.

This research highlights the capability of GPT models to generate adequate rationale, with potential for future refinement. Future iterations could refine the prompt instructions (2)to more closely align with Lee's definition of design rationale, which includes alternatives considered and tradeoffs evaluated. Additionally, the computational tool can evaluate the quality of rationale generated throughout the prompt-engineering process. This approach can characterize and assess different representations of design rationale, such as IBIS or QOC [25, 26]. The tool enhances existing methods for capturing design rationale and holds promise for shaping new approaches for generative rationale.

5.3 Limitations and future work

The study, while able to differentiate differences between the two groups, has limitations regarding the generation of the dataset, the rubric, and the quantity of data points. Since they are machine-generated, GPT-generated rationales pose concerns about response variability and empirical validity. Distributions of linguistic features showed that the responses did not cover the entire realm of possibilities, meaning that machine-generated responses are likely to have less variability than human-generated rationales. In future work, human-generated rationale collection must address potential confounding variables, such as incomplete information and variability in design rationale representations (e.g., images, words). The decision to use GPT-generated rationales removed the guesswork from researchers in trying to identify what is and is not rationale in a report.

Moreover, rubric selection poses limitations since alternative rubrics could have been utilized. For example, a simplified holistic rubric could be utilized similar to that of the SAT [29] or a more granular rubric with more dimensions or a more extensive range (i.e., 0 to 100). A rubric uses ordinal data; however, since we used the mean of two raters, the ratings were treated as continuous variables. Considering the study also tested the structuring of design rationale using the feature, specification, and evidence framework, careful consideration was given regarding feature selection so as not to bias the model with features tied to the FSE components.

Due to the textual nature of the data, interpretable and explainable NLP and ML approaches were considered. Early analyses explored BERT embeddings, TF-IDF, NER, and POS approaches, which generated enormous amounts of features that, while they may be significant in this dataset, raised concerns regarding interpretability and generalizability. Lastly, only a portion of the 2250 rationales were evaluated for this conference paper (n=512) in which preliminary results were shown. Future work will finish the evaluation across the entire dataset (n = 2250) and iterate on the model. The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) will be used in addition to RSME for model comparison. Models with lower AIC and BIC scores are more parsimonious and thus avoid overfitting and reduce capturing irrelevant features or noise in the data. Therefore, iterating on the model should produce a computational approach to measure rationale quality.

6 CONCLUSION

Design rationales are the justifications behind a product or feature. The structure and content included in rationales impact their quality. A dataset of 2250 design rationales was collected, and a randomly selected subset of the data was evaluated by two raters (n=512) on five dimensions of quality. Rationales were then characterized using natural language processing techniques. Results show that rationales represented in a feature, specification, and evidence format were rated higher than unstructured rationales on the five dimensions of evaluating, analyzing, synthesizing, structure, and validity. Parameter fitting and model comparison were used to identify the most relevant and distinguishable features to build a preliminary model that can computationally evaluate design rationales.

ACKNOWLEDGMENT

This work has been supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2146752, the University of California Regents, and partially supported by the National Science Foundation under Grant No. DGE-1633740. The findings presented in this work represent the views of the authors and not necessarily those of the sponsors. We want to express our gratitude to the following undergraduate student researchers for their valuable contributions to this study: Xiaowen Liu (second author), Lily Barbagelata, and Adam Ousherovitch.

REFERENCES

[1] Moran, T. P., and Carroll, J. M., 2020. *Design rationale: Concepts, techniques, and use*. CRC Press.

[2] Lee, J., 1997. "Design rationale systems: understanding the issues". *IEEE expert,* ***12***(3), pp. 78–85.

[3] Hirsch, P. L., Shwom, B. L., Yarnoff, C., Anderson, J. C., Kelso, D. M., Olson, G. B., and Colgate, J. E., 2001. "Engineering design and communication: The case for interdisciplinary collaboration". *International Journal of Engineering Education,* ***17***(4/5), pp. 343–348.

[4] Das, D., and Chernova, S., 2020. "Leveraging rationales to improve human task performance". In Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 510–518.

[5] Dong, A., Lovallo, D., and Mounarath, R., 2015. "The effect of abductive reasoning on concept selection decisions". *Design studies,* ***37***, pp. 37–58.

[6] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F., 2018. "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation". *arXiv preprint arXiv:1802.00682*.

[7] Hertzum, M., and Pejtersen, A. M., 2000. "The information-seeking practices of engineers: searching for documents as well as for people". *Information Processing & Management,* ***36***(5), pp. 761–778.

[8] Gruber, T. R., and Russell, D. M., 2020. "Generative design rationale: Beyond the record and replay paradigm". In *Design Rationale*. CRC Press, pp. 323–349.

[9] McNamara, D. S., Crossley, S. A., and McCarthy, P. M., 2010. "Linguistic features of writing quality". *Written communication,* ***27***(1), pp. 57–86.

[10] Crossley, S. A., Kyle, K., and Dascalu, M., 2019. "The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap". *Behavior research methods,* ***51***, pp. 14–27.

[11] Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., and Cole, R. S., 2020. "Rubrics to assess critical thinking and information processing in undergraduate stem courses". *International Journal of STEM Education,* ***7***, pp. 1–15.

[12] Crossley, S., and McNamara, D., 2010. "Cohesion, coherence, and expert evaluations of writing proficiency". In Proceedings of the Annual Meeting of the Cognitive Science Society, Vol. 32.

[13] McNamara, D., and Graesser, A., n.d.. Coh-metrix version 3.0 indices. Coh-Metrix Documentation.

[14] Breetvelt, I., Van den Bergh, H., and Rijlaarsdam, G., 1994. "Relations between writing processes and text quality: When and how?". *Cognition and instruction,* ***12***(2), pp. 103–123.

[15] Mirabito, Y., and Goucher-Lambert, K., 2022. "Investigating how engineers and designers communicate design rationale". In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Vol. 86267, American Society of Mechanical Engineers, p. V006T06A033.

[16] Olea, C., Tucker, H., Phelan, J., Pattison, C., Zhang, S., Lieb, M., Schmidt, D., and White, J., 2024. "Evaluating persona prompting for question answering tasks". In 10th International Conference on Artificial Intelligence and Soft Computing, pp. 1–18.

[17] Koo, T. K., and Li, M. Y., 2016. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research". *Journal of chiropractic medicine,* ***15***(2), pp. 155–163.

[18] Gero, J. S., and Kannengiesser, U., 2004. "The situated function–behaviour–structure framework". *Design studies,* ***25***(4), pp. 373–391.

[19] Lei, C.-U., Man, K., and Ting, T., 2014. "Using coh-metrix to analyse writing skills of students: A case study in a technological common core curriculum course". *Lecture Notes in Engineering and Computer Science*.

[20] Goucher-Lambert, K., Gyory, J. T., Kotovsky, K., and Cagan, J., 2020. "Adaptive inspirational design stimuli: using design output to computationally search for stimuli that impact concept generation". *Journal of Mechanical Design,* ***142***(9), p. 091401.

[21] Hernandez, A., 2018. "An engineering reasoning-based course on research methodologies for systems engineers". In ECRM 2018 17th European Conference on Research Methods in Business and Management, Academic Conferences and publishing limited, p. 172.

[22] Schön, D. A., 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[23] Sagoo, J., Tiwari, A., and Alcock, J., 2014. "Reviewing the state-of-the-art design rationale definitions, representations and capabilities". *Int. J. Eng. Educ.,* ***5***, p. 211.

[24] Regli, W. C., Hu, X., Atwood, M., and Sun, W., 2000. "A survey of design rationale systems: Approaches, representation, capture and retrieval". *Eng. Comput.,* ***16***, pp. 209–235.

[25] Conklin, E. J., and Yakemovic, K. C. B., 1991. "A process-oriented approach to design rationale". *Human–Computer Interaction,* ***6***, pp. 357–391.

[26] MacLean, A., Young, R. M., Bellotti, V. M. E., and Moran, T. P., 1991. "Questions, options, and criteria: Elements of design space analysis". *Human–Computer Interaction,* ***6***, pp. 201–250.

[27] Lee, J., and Lai, K.-Y., 1991. "What's in design rationale?". *Human–Computer Interaction,* ***6***, pp. 251–280.

[28] Raina, A., McComb, C., and Cagan, J., 2019. "Learning to design from humans: Imitating human designers through deep learning". *Journal of Mechanical Design,* ***141***(11), p. 111102.

[29] College Board, n.d.. Sat essay scoring. SAT Essay Scoring – SAT Suite — College Board.

Appendix A: Supplemental Tables

TABLE 6. List of personas. Full descriptions are not shown.

Title (Experience Level)
Mechanical Engineer (Entry-Level)
Industrial Designer (Senior Level)
Automotive Engineer (Mid-Level)
User Experience (UX) Designer (Mid-Level)
Product Development Engineer (Senior Level)
Systems Engineer (Entry-Level)
Sustainable Design Specialist (Mid-Level)
Research and Development Engineer (Senior Level)
Product Designer (Entry-Level)

TABLE 7. Rubric dimensions.

Category	Description
Evaluating	Ability to determine relevance and reliability of information to support an argument(i.e., whether they successfully created the desired product.
Analyzing	Ability to extract patterns from data/information that could be used as evidence for their claims.
Synthesizing	Ability to connect information to make a claim.
Forming Arguments (Structure)	Degree in which their decision, evidence, and reasoning are explicitly stated and clearly linked.
Forming Arguments (Validity)	Degree to which their claim, evidence, and reasoning are consistent with accepted disciplinary ideas and practices.