ASME Journal of Mechanical Design
Online journal at:
https://asmedigitalcollection.asme.org/mechanicaldesign

Characterizing Design Rationales Using Computational Linguistics and Human Evaluations

Yakira Mirabito[1]
Mem. ASME
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
e-mail: yakira.mirabito@berkeley.edu

Xiaowen Liu
Department of Computer Science,
University of California, Berkeley,
Berkeley, CA 94720
e-mail: xiaowen_liu@berkeley.edu

Kosa Goucher-Lambert
Mem. ASME
Department of Mechanical Engineering,
University of California, Berkeley,
Berkeley, CA 94720
e-mail: kosa@berkeley.edu

Design rationales capture the explicit justifications behind design decisions. Often, rationales vary in the content and depth of information, making the study and comparison of rationales challenging. This project aims to characterize design rationales and develop a standardized approach to assess design rationale quality at scale. In total, 2250 rationales were machine-generated across two different representations and 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 linguistic features were used to build predictive models for each quality dimension. The main results identify correlations between linguistic features and human evaluations, show that structured rationales were rated higher than unstructured rationales across the five dimensions, and present a model to predict rationale quality for new texts. These findings help inform strategies to improve human- and machine-generated rationales. The predictive tool offers a scalable method for evaluating different representations of rationale, thereby supporting more effective documentation practices. [DOI: 10.1115/1.4069232]

Keywords: design evaluation, design representation, design theory and methodology, generative design, machine learning, design process

1 Introduction

Design rationale captures the reasoning processes or justifications behind design decisions, often appearing in written reports or verbal presentations. While the importance of capturing design rationales in the design process is well established, there is a lack of clear guidelines on what constitutes effectively communicated design rationales (i.e., quality). Throughout the literature, design rationale has been defined in various ways, captured through different techniques, and applied in multiple contexts [1–3]. For this study, we use Lee's definition, which describes 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" [4]. For example, a rationale might state, "We selected aluminum for the bike frame to reduce weight while maintaining structural strength, based on a finite element analysis (FEA) and user preferences for portability." This research focuses on characterizing written design rationales and evaluating their quality.

Documenting design rationale is crucial due to its influence on a product's long-term success and financial viability. Design is inherently a collaborative process, requiring engineers and designers to work closely with various stakeholders within a firm (e.g., product managers, marketing, sales) and with clients [5]. Without effective communication and proper positioning, a team that invests substantial effort into creating an innovative product may see it fail. Researchers have found that engineering documents tend to favor the technical aspects of the final solution but often lack an explanation of the context of the process [6]. As a result, a firm might waste resources by repeating mistakes that were not adequately documented or by attempting to retrieve rationale from colleagues who worked on earlier iterations. Furthermore, even if documentation occurs, the information included and the tone used to explain design decisions can influence human behavior (e.g., willingness to make changes) [7–9]. Thus, structuring design rationale in ways that enhance design processes and outcomes is of high importance.

The motivation for this study stems from a need to provide designers with actionable insights to enhance design rationale documentation practices. While numerous capture and representation techniques exist [3], such as issue-based information systems (IBIS), [10], design rationale editor (DRed) [11], and free-text rationales [12], none have seen widespread adoption in practice. IBIS is a method that structures rationale around design issues, positions, and supporting arguments, while DRed is a software tool that graphically represents these elements. The limited adoption of these methods is likely due to the burden these methods place on designers, who are often expected to tag features or articulate decisions in ways that exceed what teams are willing (or resourced) to document. As a result, rationale is either captured inconsistently or omitted altogether, limiting the value of well-designed representations. Recent advances in large language models enable new forms of rationale generation or extraction, but their integration may also be limited without scalable evaluation methods. A

[1] Corresponding author.

Contributed by the Design Theory and Methodology Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received December 20, 2024; final manuscript received July 7, 2025; published online August 28, 2025. Assoc. Editor: Mark Fuge.

systematic approach to assessing rationale quality, regardless of how it is produced, may help bridge the gap between existing representations and practical adoption. Current methods rely on human evaluators to judge whether a rationale is sufficient; however, these approaches are time-consuming, nonstandardized, and challenging to scale.

This research aims to quantify the quality of design rationale using human evaluators and rubrics and then transfer those human evaluations into a predictive tool capable of assessing new design rationales at scale. While tools like Grammarly offer automated feedback on grammar and style, they are not designed to evaluate the reasoning or technical depth required in design rationales. The guiding research question is:

How do linguistic features and structural representations influence the quality of design rationales?

To address this overarching question, we explore the following three subquestions: (a) What linguistic features are most strongly associated with high-quality design rationales? (b) How does the structured feature, specification, and evidence (FSE) framework compare to unstructured approaches in producing high-quality design rationales? (c) How effectively can predictive models evaluate design rationale quality using linguistic features and human ratings? By addressing these questions, this work can help address the immediate gap in enhancing design rationale communication and holds the potential to shape future design support tools and methodologies.

In order to explore these questions in depth, a dataset of design rationales was collected. Rubrics and raters were used to assess the quality of design rationale, and natural language processing (NLP) was used to extract meaningful features from the written text. Section 2 provides context on rubrics considered, the NLP approach used, and the framework tested. Section 3 details the research process before presenting findings to each subquestion (Section 4) and discussing the implications of the work (Section 5).

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 adopts a similar pipeline to prior research that developed computational approaches for automatically scoring essays using standardized rubrics from the Scholastic Aptitude Test (SAT), a U.S. college admissions exam [13,14].

2.1 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 the types of writing or processes that 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, design rationales are technical in nature, describing design methods, processes, and decisions. This type of writing is clearly distinct from what might appear in an essay or novel. Technical writing centers on effectively communicating complex information, emphasizing clarity, precision, and adherence to established conventions.

The decision to use critical thinking as a measure of design rationale quality is rooted in its ability to assess, analyze, and synthesize information from various sources. Critical thinking involves cognitive skills that enable evaluators to make inferences and draw meaningful conclusions from complex information [15]. While technical writing is often evaluated at the report level, critical thinking is applied to smaller segments, such as paragraphs of design rationale. This distinction between the type of writing (technical versus nontechnical) and scope (report versus paragraph) in this context calls for carefully selecting a rubric to guide raters in assessing the depth of reasoning and decision-making embedded in design rationales.

The rubric selected for this study focuses on critical thinking across five dimensions (evaluating, analyzing, synthesizing, forming arguments-structure, and forming arguments-validity) [15]. 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 both their study and this one, none of the raters used a zero, and the meaning of a zero rating was not clearly defined. Descriptions of each dimension of the rubric are shown in Table 2.

2.2 Computational Approaches. This study draws inspiration from tools such as Grammarly, a cloud-based writing assistant that provides real-time feedback on spelling, grammar, clarity, and tone. While Grammarly is widely used and offers useful suggestions based on established writing conventions, it is not equipped to evaluate the reasoning quality or technical depth expected in design rationale. This work aims to assess deeper linguistic and argumentative features relevant to engineering contexts, using NLP to capture patterns in cohesion, coherence, and clarity that may signal rationale quality.

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 [16]. 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, as noted in Fig. 1 under featurization. The complete list and definitions can be seen [17]. On the other hand, TAACO calculates semantic overlap between sentences (local cohesion), paragraphs (global cohesion), and the entire document (overall text cohesion) for nouns and verbs [14]. 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.

Prior research from McNamara et al., who developed the Coh-Metrix model, highlighted characteristics associated with more cohesive texts, such as lexical diversity, connectivity, and word concreteness [13,18]. 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 and 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, and evidential sentences) [16,19]. 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

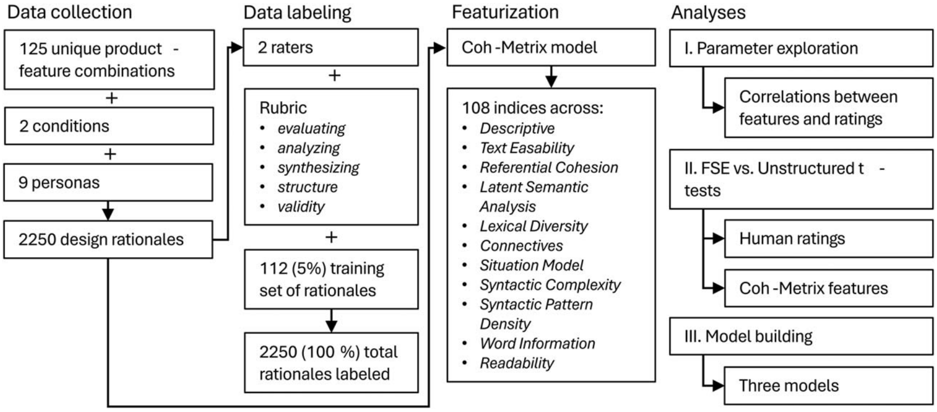


Fig. 1 Overview of the research design

across five dimensions (evaluating, analyzing, synthesizing, forming arguments-structure, and forming arguments-validity) [15].

2.3 Feature, Specification, and Evidence Framework.

Design rationales can take multiple forms depending on the phase of the design process, ranging from requirement-focused rationales to detailed justifications tied to specific design features or decisions (i.e., feature-based) [1,3]. This study focuses on retrospective feature-based rationales that appear in final design documentation in educational and industry contexts. These rationales articulate why a concept or feature was selected while drawing connections to earlier specifications and supporting evidence.

The representation of design rationale lacks standardization, leading to variability in both the content and depth of information included in different rationales [1]. One framework that aims to enhance communication and documentation practices of design rationale is the feature, specification, and evidence framework [20]. The FSE framework was developed through an empirical analysis of engineering technical reports, revealing how designers in unprompted settings articulate rationales in natural language by linking final product features to relevant specifications and supporting evidence. The main content of the framework is described below, and it was used to generate data for one of the conditions used in this study.

- Feature (F) refers to a specific component or attribute of a product that the rationale justifies, such as a brake pad, steering wheel, or tire. Generally, each feature should align with a relevant specification. The decision of which features to include in a report is often defined by the firm or industry standards. For example, the material selection for a bolt may need a detailed justification based on the scale of the system, whereas another firm may only want to document key features that impact the overall functionality of the final design.
- Specification (S) outlines the design requirement(s) that the feature is intended to meet, typically established early in the design process. These may include objectives like reducing stopping distance or ensuring traction under specific weather conditions. If specification or requirement tables are already available, it is essential that designers explicitly cite the requirement addressed, rather than cite an entire table or separate document. In some cases, a single feature may address multiple requirements, or multiple features may combine to satisfy one specification.
- Evidence (E) includes the relevant data or insights from the design process that informed the decision to select a particular feature that meets the specification(s). This could include background research, user interviews, or product testing. For example, when testing various braking mechanisms or brake pad materials, the evidence would detail the outcomes from these tests and how they influenced the final design choice.

3 Methods

Overall, this project quantifies the quality of design rationales as noted in Fig. 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 100–200 words). This dataset was then labeled using human raters on five dimensions [15]. The raw text was processed using the computational tool Coh-Metrix, resulting in 108 features [16]. Afterward, analyses identified correlations between linguistic features and quality dimensions and compared the two representations. The most relevant features were explored and integrated into the predictive model.

3.1 Data.

A total of 2250 design rationales were generated for 25 products (Table 6) with five features under two conditions with nine personas. Representative rationales for both the *unstructured* and *FSE* conditions are shown.

$$
\begin{aligned}
&25 \text{ products} \times 5 \text{ features} \times 2 \text{ conditions} \times 9 \text{ personas}\\
&\quad = 2250 \text{ rationales}
\end{aligned}
$$

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: first, to create a user-friendly device, and second, 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 consistently emphasized as a top priority for users when engaging with pressure cookers. Moreover, background research on cooker-related accidents revealed that 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. Generative Pre-trained Transformer (GPT) 4.0 from OpenAI was used to generate the rationales, which required an API key and PYTHON terminal. To increase variability in GPT-generated responses, nine personas were used. Prior research has shown that including personas within GPT prompts has improved response diversity in prompt-engineering tasks [21]. 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). The list of titles and experience levels can be seen in Table 7 in the Appendix.

One hundred twenty-five unique prompts were asked about 25 consumer products, each containing five features. For example, "What is the rationale behind the walking belt on a treadmill?" The prompt instructions were refined to limit response length, as initial outputs from GPT were overly verbose (e.g., exceeding 500 words). To assess different representations of rationale, two conditions were used, as noted in Table 1. All rationales were generated using zero-shot prompting, meaning the model was given no context (e.g., rubric and prior solutions) from which it could infer what would lead to higher ratings.

3.3 Data Labeling. To assess the quality of design rationales, two raters evaluated each rationale along five dimensions using a structured rubric adapted from Ref. [15]. The rubric assesses critical thinking across evaluating, analyzing, synthesizing, forming arguments (structure), and forming arguments (validity). Each dimension was rated on a zero (worst) to five (best) integer scale. Raters were required to have prior experience assessing student or employee reports in an engineering or design context (i.e., education or practice). The two raters were senior PhD students in mechanical engineering who hold BS and MSc degrees in engineering. Descriptions of each rubric dimension are shown in Table 2. The modified rubric and scale (without training notes) shown to raters are presented in Fig. 6 of the Appendix.

Raters were trained on a subset of data (approximately 112 or 5% of the total dataset) using three calibration sessions. The first session involved jointly rating five rationales. The second and third sessions involved comparing individual ratings on 25 and then 112 rationales, respectively, to align interpretations of the rubric. After calibration, the remaining rationales were divided between the two raters. While raters were blind to the conditions, differences in the response structure may have made the conditions inferable.

Inter-rater reliability was calculated using intraclass correlation coefficients (ICCs) with a mean rating ($k = 2$), consistency-agreement, 2-way random-effects model, implemented in RStudio using the *irr* package. ICC scores for the 112 jointly rated rationales were: evaluating (0.94), analyzing (0.79), synthesizing (0.83), structure (0.79), and validity (0.85). Based on established benchmarks, 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 [22].

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

3.4 Featurization. To better characterize the design rationales, the Coh-Metrix model was used for feature extraction. This process produced numerical values for linguistic and discourse representations of each rationale. The output was provided as a comma separated values (CSV) file containing data for all 2250 rationales. Coh-Metrix generated 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 Ref. [17]. These NLP analyses run within the Coh-Metrix tool helped characterize each rationale with information such as word count and lexical diversity. Thus, the resulting indices were explored and selected in the model-building phase.

3.5 Model Building. A broader goal of this work is to predict the quality of design rationale. To do so, a combination of human ratings for five dimensions [15] and linguistic features from Coh-Metrix was combined. Trends between linguistic features and quality measures were first identified before feature selection was performed. Considering that a portion of the data was double-coded by each rater, an average rating for each dimension of rationale quality was used. Thus, the scale or output of the prediction task results in a continuous variable 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, 108 linguistic features were generated for all 2250 design rationales. The data were standardized. Next, model selection considered the task's regression nature, the dataset's size (2250 samples), and the desire for a model that balances simplicity with interpretability. Therefore, simplicity guided the feature selection process. While multiple models were explored in preliminary results (i.e., random forest, gradient boosting machine, and Bidirectional Encoder Representations from Transformers (BERT)), ultimately, linear regression models without Principal Component Analysis (PCA) were used. Root mean square error (RMSE) was used to assess error rates, while Akaike information criterion (AIC) and Bayesian information criterion (BIC) were explored for complexity evaluation. The models used were:

Table 1 Prompt instructions for two conditions

Condition	Instructions
Unstructured	Please write your rationale (approximately 100–200 words) in a single paragraph format
FSE	Please write your rationale (approximately 100–200 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.

Table 2 Descriptions of rubric dimensions based on Ref. [15]

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
Structure	Degree in which their decision, evidence, and reasoning are explicitly stated and clearly linked
Validity	Degree to which their claim, evidence, and reasoning are consistent with accepted disciplinary ideas and practices

- ➢Dummy regressor—serves as a baseline model for comparison. This model gives predicted values based on a simple mean strategy that disregards input data, as noted in Eq. (2).
- ➢Linear regression—selected for its simplicity and interpretability, linear regression is an effective model. It offers straightforward insights through model weights (Eq. (3)) and can utilize all 108 indices. A key challenge in this work is simplifying the model to include only the most important features that can be discussed in terms of text characteristics that writers can improve.
- ➢Lasso regression—the linear model was refined by incorporating regularization. In essence, hyperparameter tuning using Lasso regression (via Eq. 4) and cross-validation optimize the parameters.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2}$$

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \tag{3}$$

$$\underset{\beta}{\text{minimize}}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right) \tag{4}$$

where [x_{ij}] is the value of the jth feature for the ith sample (lasso), [p] is the number of features, and [λ] is the regularization parameter controlling the L1 penalty.

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 assessment metric and utilized grid search cross-validation to fine-tune the hyperparameters. For model comparison, AIC (Eq. (6)) and BIC (Eq. (7)) values, alongside RMSE, were used to gauge their performance comprehensively. This work leveraged the work from Wilson and Collins on suggested steps for computational modeling [23]. The coefficients and intercept from the best model were then presented.

$$\hat{L} = -\frac{1}{2}n\left(\ln(2\pi) + \ln(\sigma^2) + 1\right) \tag{5}$$

$$\text{AIC} = -2\ln\left(\hat{L}\right) + 2k \tag{6}$$

$$\text{BIC} = -2\ln\left(\hat{L}\right) + k\ln(n) \tag{7}$$

where [σ^2] is the variance of the residuals, [n] is the number of samples, and [k] is the number of estimated parameters in the model.

4 Results

Design rationales ($n = 2250$) were collected, characterized using natural language processing, and evaluated by human raters. The following sections outline the findings in three parts. The first identifies NLP features correlated with increased design rationale quality across the entire dataset. The second finding compares the two representations of design rationale (FSE structured and unstructured) regarding their human evaluations of quality and NLP characteristics. Lastly, the third finding presents a predictive model to assess design rationale quality.

4.1 The Majority of Linguistic Features Correlated With Rationale Quality.

The associations across the 108 linguistic features and five quality dimensions were explored, starting with the "evaluating" dimension. Figure 2 visualizes the six strongest correlations between linguistic features and the "evaluating" dimension rating. The correlation coefficients (Spearman's ρ) and p-values are shown. Word count (DESWC), hypernymy for nouns and verbs (WRDHYPnv), and noun incidence (WRDNOUN) were positively correlated with increased rationale quality. Conversely, temporality (PCTEMPp), adverb incidence (WRDADV), and causal verb incidence (SMCAUSv) resulted in negative correlations with rationale quality. Only six linguistic features were visualized, despite a total of 75 statistically significant correlations. Scatter plots of the correlations for the remaining four quality dimensions are not shown.

Furthermore, Spearman's correlation coefficients were calculated between the 108 linguistic features and each quality dimension to identify significant relationships. The Benjamini–Hochberg correction was used to adjust the p-values for multiple testing and filters for statistically significant correlations (adjusted p-value <0.05). The evaluating dimension had 75 significant correlations, and analyzing and synthesizing each had 65 correlations, structure had 71, and validity had 69. Figure 3 visualizes the same six linguistic features across the "evaluating" dimension and the corresponding correlation values for the four other dimensions (analyzing, synthesizing, structure, and validity). The correlation coefficients visualized range from weak (0.1–0.3) to moderate (0.3–0.5) in strength.

4.2 Feature, Specification, and Evidence Structured Rationales Were Rated Higher Quality Compared to Unstructured Rationales.

To compare the rated rationales of FSE and unstructured, t-tests comparing the means of each condition were conducted. Analyses for all five dimensions were statistically significant, meaning the two conditions differed. 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. Specifically, the FSE condition scored higher on *evaluating* ($M = 4.45$) compared to the unstructured condition ($M = 1.41$), $t(1250) = 108$, $p < 0.001$. Similar patterns were observed for *analyzing* ($M_{\text{FSE}} = 4.05$, $M_{\text{Unstr}} = 2.47$, $t(1250) = 41.9$, $p < 0.001$), *synthesizing* ($M_{\text{FSE}} = 4.11$, $M_{\text{Unstr}} = 2.56$, $t(1250) = 42.5, p < 0.001$), *structure* ($M_{\text{FSE}} = 4.32$, $M_{\text{Unstr}} = 2.37$, $t(1250) = 55.3, p < 0.001$), and *validity* ($M_{\text{FSE}} = 4.04$, $M_{\text{Unstr}} = 2.04$, $t(1250) = 53.4, p < 0.001$). Figure 4 visualizes the mean rating for each condition per quality dimension.

For each condition (FSE and unstructured), the Coh-Metrix indices were averaged and tested to determine which linguistic features were statistically different between the two groups using Kruskal–Wallis tests. Eighty-one of the 108 Coh-Matrix indices had p-values less than 0.05, 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 Fig. 1,

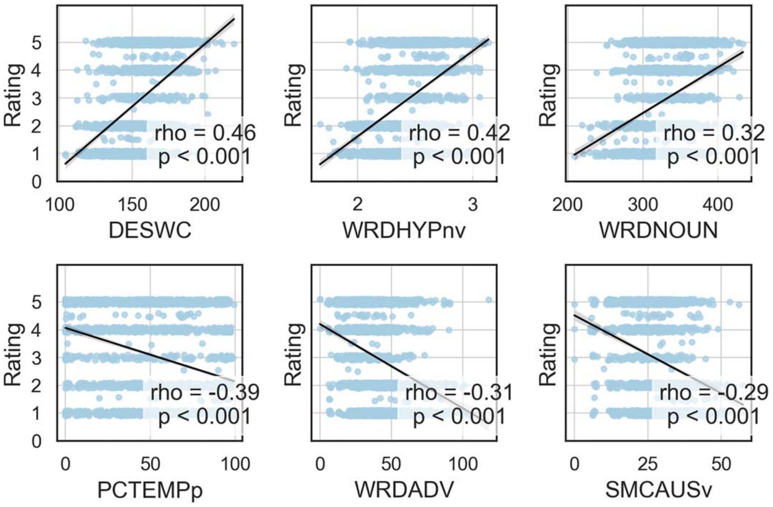


Fig. 2 Scatter plots for the top six Coh-Metrix indices plotted against the "evaluating" dimension ratings. Each data point represents one design rationale ($n = 2250$).

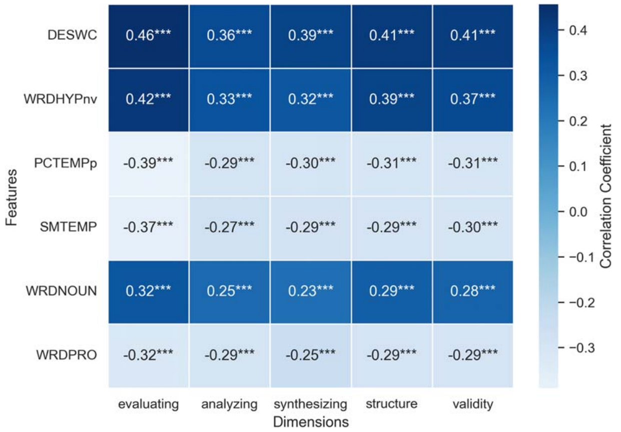


Fig. 3 Correlation matrix between top six Coh-Metrix features and five quality dimensions (*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$)

although the full list and definitions are detailed in Ref. [17]. The indices from Coh-Metrix with the largest H-statistic, meaning the most significant difference of normalized means across the two conditions, are shown in Table 3.

4.3 Combining Linguistic Characterizations and Human Ratings to Predict Rationale Quality. To work toward building a model that automatically evaluates the quality of design rationales, the Coh-Metrix indices and human ratings were combined. Three models were built and compared using RMSE, AIC, and BIC. The dummy regressor model generates predictions without considering the input features, serving as the baseline for comparison. The linear regression model uses a reduced set of features, comprising 82 of the 108 indices (24 were removed due to high collinearity with correlations over 0.85). The lasso model was iterated on the linear regression model, starting with 82 features that were then adjusted to between 7 and 14 features depending on the quality dimension at hand.

Figure 5 visualizes the optimal lambda (and corresponding number of features), which was determined using the bias-variance tradeoff. Meanwhile, Table 4 shows the coefficients and statistical significance for the resulting linguistic features per dimension. Word count (DESWC) and temporality (PCTEMPz) were two indices that were statistically significant across the five dimensions.

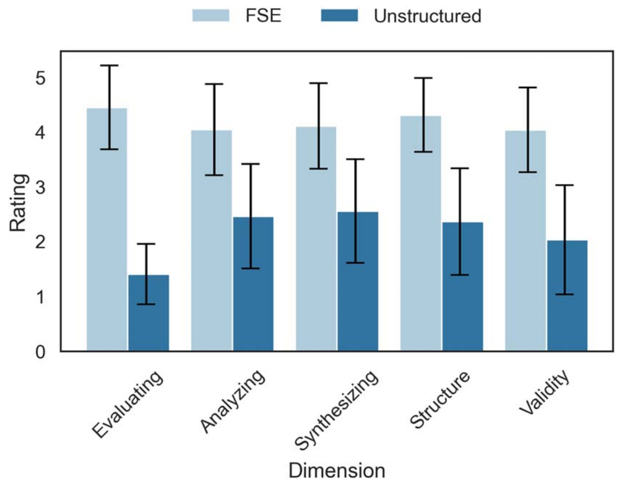


Fig. 4 Average ratings for FSE and unstructured for each of the five dimensions of rationale quality: evaluating, analyzing, synthesizing, structure, and validity

Table 3 Means, H-statistic, and *p*-values for the six most distinct Coh-Metrix indices across the FSE and unstructured conditions

Coh-Metrix index	M_{FSE}	M_{Unstr}	H-stat	*p*-Value
Word count (DESWC)	163	147	531	<0.001
Hypernymy for nouns and verbs (WRDHYPnv)	2.53	2.33	457	<0.001
Temporality (PCTEMPp)	45.5	71.9	414	<0.001
Temporal cohesion (SMTEMP)	0.831	0.926	395	<0.001
Adverb incidence (WRDADV)	35.8	48.0	287	<0.001
Pronoun incidence (WRDPRO)	11.4	18.2	274	<0.001

Table 5 shows the RMSE, AIC, and BIC for the three resulting models. The RMSE measures the average difference between predicted and actual values. A lower RMSE is generally better, indicating less error between predicted and actual values. All models were assessed similarly using AIC and BIC values. Lower AIC and BIC values are better. Note that AIC does not penalize the number of parameters included in the model as much as BIC. Thus, models

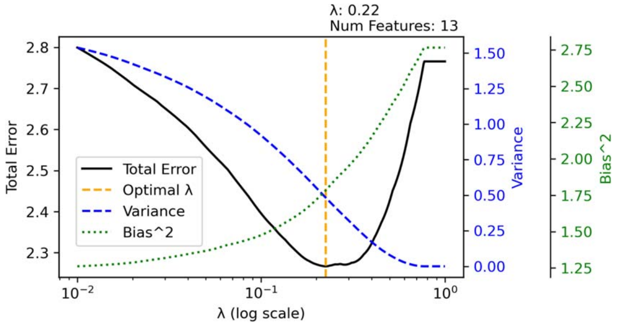


Fig. 5 Bias-variance tradeoff for lasso regression for "evaluating" dimension. Optimal lambda and the resulting number of features used in the model.

with lower scores are more parsimonious, avoiding overfitting and reducing the capture of irrelevant features (i.e., noise) in the data. Of the models explored, the Lasso regression obtained the lowest RMSE for all dimensions except "structure," where the linear regression model performed slightly better. The BIC results for four of the five dimensions suggest that the Lasso regression model is the better choice.

5 Discussion

This study explored the relationship between linguistic features and the quality of design rationales by analyzing 108 linguistic features from Coh-Metrix and assessing quality through human ratings across five dimensions (evaluating, analyzing, synthesizing, structure, and validity) [15]. The overarching research question—*How do linguistic features and structural representations influence the quality of design rationales?*—was addressed through three key findings. First, we identified linguistic features that were associated with higher-quality rationales, addressing subquestion (a). Second, we found that rationales structured using the FSE framework were rated significantly higher than unstructured rationales across all five dimensions, addressing subquestion (b). Third, we developed a preliminary predictive model using linguistic features and human ratings, providing a foundation for scalable evaluation and addressing subquestion (c). The following sections interpret these findings in detail and offer insights for improving design rationale quality and communication.

5.1 Explaining Design Rationale Quality Through Linguistic Features. Section 4.1 visualizes six of the 75 statistically significant correlations between the linguistic features and "evaluating" rating (Fig. 2) and then shows similar patterns for the other dimensions (Fig. 3). The "evaluating" dimension reflects a designer's ability to assess the relevance and reliability of information. Rationales rated highly on this dimension typically reference reliable data sources, such as explicit user feedback or product testing results. For example, one of the rationales for a blender's container component came "from numerous user tests and feedback," illustrating how higher-rated rationales incorporate reliable evidence to justify design decisions.

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 to support decisions. 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 (i.e., engineering design).

Looking closer at the positive correlations of word count (DESWC), hypernymy for nouns and verbs (WRDHYPnv), and noun incidence (WRDNOUN) for the "evaluating" dimension, they demonstrate which linguistic features reflect more effective communication and reasoning. A higher word count generally indicates more detailed information, which could explain why higher-rated rationales can provide a deeper understanding of the thought process behind design decisions. Hypernymy refers to the use of general and abstract terms, indicating a higher-level conceptual understanding. Higher-rated rationales may have higher hypernymy usage since they connect specific design decisions to overarching goals or principles. A higher noun incidence indicates that a text contains more references to concrete entities, like design features, which results in a more focused argument or higher writing clarity.

Meanwhile, the negative correlations of temporality (PCTEMPp), adverb incidence (WRDADV), and causal verb incidence (SMCAUSv) can be explained by how these linguistic features might reflect less effective communication or reasoning. Temporality measures cues on time-related language. While some temporal language is necessary, an overemphasis could be explained by focusing on procedural or chronological details instead of discussing "why" decisions were made. A high frequency of adverbs, which modify verbs, adjectives, or other adverbs, can signal vagueness in the text. Adverbs such as "possibly" introduce

Table 4 Resulting Coh-Metrix indices from the Lasso regression at the optimal lambda value

Coh-Metrix index	Evaluating ($n = 13$)	Analyzing ($n = 7$)	Synthesizing ($n = 14$)	Structure ($n = 9$)	Validity ($n = 13$)
DESWC (word count)	0.401***	0.207***	0.264***	0.276***	0.294***
DESSLd (sentence length)	0.021		0.013		0.015
DESWLltd (word length)	0.007				
PCTEMPz (temporality *z*-score)	−0.294***	−0.094***	−0.134***	−0.118***	−0.147***
CRFANP1 (anaphor overlap)	−0.016		−0.021		−0.007
CNCTempx (expanded temporal connectives incidence)	0.050		0.047	0.021	0.057*
SMCAUSvp (causal verbs and causal particles incidence)	−0.107*	−0.002	−0.037	−0.052	−0.055
SYNSTRUTt (syntactic structure similarity all)	−0.015		−0.011		
WRDNOUN (noun incidence)	0.122***	0.063*	0.056	0.106***	0.088**
WRDADV (adverb incidence)	−0.089*		−0.005	−0.015	−0.033
WRDPRO (pronoun incidence)	−0.058	−0.074*	−0.042	−0.058	−0.067*
WRDFRQa (average word frequency)	−0.040	−0.037	−0.038	−0.030	−0.044
WRDAOAc (age of acquisition for content words)	0.106**				0.004
LDVOCD (lexical diversity VOCD)		0.011	0.047		0.002
CNCADC (adversative/contrastive connectives incidence)			−0.006		
SMCAUSv (causal verb incidence)			−0.005	−0.022	−0.012

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$.

Table 5 Root mean squared error (RMSE) for the five rubric dimensions across three models

Dimension	Model	RMSE	AIC	BIC
Evaluating	Model 1: dummy regressor	1.66	6936	6941
	Model 2: linear regression	1.35	5365	5821
	Model 3: Lasso regression ($n = 13$)	**1.34**	6037	6114
Analyzing	Model 1: dummy regressor	1.18	5762	5767
	Model 2: linear regression	1.22	5105	5561
	Model 3: Lasso regression ($n = 7$)	**1.09**	5404	5448
Synthesizing	Model 1: dummy regressor	1.15	5659	5665
	Model 2: linear regression	1.24	5084	5540
	Model 3: Lasso regression ($n = 14$)	**1.01**	5212	5294
Structure	Model 1: dummy regressor	1.28	6012	6018
	Model 2: linear regression	**1.08**	5192	5648
	Model 3: Lasso regression ($n = 9$)	1.13	5510	5535
Validity	Model 1: dummy regressor	1.37	6156	6161
	Model 2: linear regression	1.35	5275	5731
	Model 3: Lasso regression ($n = 13$)	**1.21**	5611	5688

The bolded value represents the model with the lowest RMSE for a given dimension.

Model 1: dummy regressor, model 2: linear regression (reduced for collinearity at 0.85 resulting in $n = 82$), and model 3: Lasso regression ($n =$ number of Coh-Metrix indices used for optimal lambda).

ambiguity that could weaken the strength of the argument. Causal verbs like "lead to" emphasize relationships between actions and outcomes. Overusing causal verbs might reflect oversimplified reasoning that neglects multiple tradeoffs considered. These features are just a subset of the linguistic features that help explain the depth, clarity, and quality of design rationales.

Combining insights from human evaluations and linguistic features informs guidelines to improve design rationale communication and aids in the development of a computational tool to assess rationale quality. Each Coh-Metrix feature is well-documented, offering actionable insights for refining writing (e.g., increasing noun incidence) to enhance rationale quality. Prior work from Lei et al. explored the potential of using a computational linguistic tool (Coh-Metrix) to analyze and improve technical essay writing [24]. The research described the patterns in linguistic features across students and alluded to a future goal of developing an automated essay assessment system that educators can use to curate learning activities (i.e., not for grading). Their work highlights the potential of computational linguistic tools to evaluate and improve technical writing, reinforcing the motivation for this study.

5.2 Using the Feature, Specification, and Evidence Framework Led to Higher-Quality Rationale. Section 4.2 revealed that FSE-structured rationales consistently outperformed unstructured rationales across all quality dimensions (Fig. 4), demonstrating the value of applying the FSE framework. FSE rationales incorporated detailed and relevant information from background research and user interviews to support their argument (evaluating). These rationales accurately extracted relevant evidence, such as key failure modes and safety concerns (analyzing). Accurately connecting features, specifications, and evidence helped form logical claims (synthesizing). The claim was well supported with multiple pieces of evidence (structure). The claim, evidence, and reasoning were consistent with accepted engineering standards (validity). Section 3.1 and Table 8 show examples of both rationale conditions.

Conversely, unstructured rationales often focused on describing the product's form and function but lacked depth in articulating the reasoning or justifications behind design decisions. The Coh-Metrix analyses reveal key linguistic differences between the two conditions in 81 out of 108 features (Table 3). Higher-rated rationales contained more words (word count) and were more specific (hypernymy for nouns and verbs), while lower-rated rationales had more temporal cues (temporality) or pronouns (pronoun incidence) that reduced the clarity of the text. These rationales resemble those commonly found in student reports or patent data [20]. Often, these descriptions are associated with the product's function, behavior, or structure [25]. While such descriptions are useful, they do not address critical questions about why a design decision was made.

Designers understand the importance of documenting design rationale, what information to include, and at what level of detail is not standardized in teaching or practice [2]. This ambiguity stems from the varying uses and capture mechanisms of rationale [3], reinforcing the need for structured representations. This article 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 the issue-based information system (IBIS) [10] and questions, options, criteria (QOC) [26]. Both are more expressive and laborious than the FSE framework, often requiring graphical network software [27]. These alternative structured representations should be explored in future work using human evaluations or the computational approaches presented in this article. Future work could also examine how different structured formats compare to FSE to determine whether specific types of scaffolding better support reasoning and evidence generation. Such comparisons could clarify which structured elements are most effective to include in rationales for improving quality.

5.3 Predicting Rationale Quality Based on Linguistic Features and Human Ratings. This study sought to evaluate whether linguistic features could be used to predict design rationale quality. Three models were compared (dummy regressor, linear regression, and Lasso regression) using RMSE, AIC, and BIC as performance metrics. The Lasso model, which selected between seven and 14 features depending on the quality dimension, outperformed the dummy regressor and achieved the lowest RMSE for four of the five dimensions. Word count (DESWC) and temporality (PCTEMPz) were statistically significant across all five dimensions, suggesting that more detailed rationales with less emphasis on procedural language tend to be rated more highly. Although the Lasso models outperformed the baseline, RMSE values remained above 1.0 on a zero-to-five scale, indicating predictions deviated by approximately one point on average. These results suggest that while linguistic features capture meaningful patterns in rationale quality, additional information is needed for more precise predictions.

Future work should build on the preliminary models presented in this study by exploring interaction effects and refining feature selection to prioritize statistically significant coefficients. While the predictive models highlight linguistic features associated with higher-rated rationales, we do not claim causal relationships. Establishing causality would require controlled experiments that manipulate specific linguistic features and isolate their effects, a worthwhile direction for future work. Another remaining question includes whether rationale quality should have these five dimensions or be condensed into a holistic measure. For example, the dimensions could be averaged, multiplied, or weighted differently to produce an overall measure such as the innovation measure [28]. These findings offer a preliminary response to the third research subquestion, demonstrating the potential for using linguistic features and human ratings to support scalable, predictive evaluation of rationale quality.

5.4 Broader Implications. Documentation is often an afterthought despite its central role in capturing design decisions and supporting knowledge transfer. This study reinforces the importance of high-quality rationale documentation and offers a systematic, scalable method to evaluate and improve best practices.

Engineering design research commonly relies on verbal and written data from engineers and designers to understand underlying processes and decision-making. Occasionally, this information is documented in reports; however, more often than not, this information is shared verbally via presentations or follow-up conversations with colleagues when design changes are needed. There is a clear gap in documentation practices regarding the content and quality of design rationale. This work makes the following contributions: identifies linguistic trends to help improve writing quality, applies the FSE framework to rationales to achieve higher-quality rationales, and develops a computational tool to assess the quality of new rationales (human and machine-generated).

5.4.1 Educators and Practitioners. Based on the findings from this work, educators and practitioners can readily implement the use of the Coh-Metrix model to gain insight into the linguistic features present in the texts of interest (e.g., student writing or technical reports). The initial analysis using Coh-Metrix serves as a baseline, and revisions to those texts can be made that integrate the linguistic trends from this study, like increasing word count and specificity of the rationale while decreasing the emphasis on chronological processes. Those strategies should help address common issues in technical writing, which include focusing too much on the design itself (form and function) or on the design processes without extracting meaningful insight or connecting the process to the decisions [6,20].

Educators and practitioners can also incorporate the FSE structure into reporting standards for their classrooms or teams. The feature, specification, and evidence framework was previously developed using student and industry reports to help identify what information is contained within design rationales. Definitions of each element are noted in Sec. 2.3 and in the prior publication, which also includes sample instructions and example outputs in the Appendix [20]. FSE is just one representation of design rationale. Alternative representations, such as issue-based information system (IBIS) [10] and QOC [26], or design rationale management systems could also be incorporated into education and practice.

Additionally, the predictive tool from this study can provide a feedback mechanism to help generate better rationales. The intent is not to be used as a grading tool but as a feedback mechanism to support learning and development. The tool can provide quantifiable feedback between iterations as students try documenting design rationales using different representations or focusing on more concrete language. Thus, students can better understand how minor adjustments in content and structure can improve their writing and communication abilities.

5.4.2 Design Rationale Researchers. This study revealed the capability of GPT models to generate adequate rationales with the potential for future refinement. In the dataset, there were clear variations in the quality of the rationales generated, which were explained by linguistic differences. Future iterations could refine the prompt instructions (shown in Table 1) to more closely align with Lee's definition of design rationale, which includes alternatives considered and tradeoffs evaluated [4]. Different representations, such as IBIS or QOC [10,26], could also be explored as part of the prompt-engineering process. The strategies for improving human-generated rationales (linguistic trends and FSE) could also be applied to machine-generated rationales.

Aside from integrating the insights from this study, researchers can directly use the predictive tool to assess the rationale quality of generative or extraction approaches. Currently, machines can assist designers by generating design suggestions, often based on human behavior (e.g., alternative CAD designs) or complex algorithms. Such suggestions often lack explanations [29], a gap that generative rationale or rationale extraction techniques could address. Recent work by Yue et al. proposes another capture method that extracts information, such as artifacts, issues, intentions, and arguments, from technical literature into a design rationale management system [30]. In addition to existing design rationale management systems (e.g., SIBYL [31], ExplainIT [32], RATionale [33], and DRed [34]), these approaches could benefit from quantitatively assessing the quality or effectiveness of the rationales, as outlined in this article. Standardized assessment enables systematic comparisons across capture mechanisms and helps establish benchmarks for high-quality rationale—strengthening support for industry-wide adoption, a challenge faced by many design rationale capture approaches.

5.5 Limitations and Future Work. While these findings provide valuable contributions, unresolved questions remain, highlighting the need for further research to address the limitations of the dataset, rubric, and features. Since the rationales were machine-generated, GPT-generated rationales raise concerns about response variability and empirical validity. Distributions of linguistic features showed that the responses did not cover the entire realm of possibilities, potentially indicating that machine-generated responses may 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, diagrams, and words). The decision to use GPT-generated rationales removed the guesswork for researchers in identifying what is and is not rationale in a technical report. Adequately identifying rationale content within technical reports is another challenge that needs to be addressed when using human-generated rationales.

Moreover, while the rubric was carefully selected, alternative rubrics could also be explored in future work. For example, a simplified holistic rubric could have been used similar to that of the SAT [35] or a more granular rubric with more dimensions or a more extensive range (e.g., 0 to 100). The selected rubric was designed for use with ordinal data; however, since the mean of two raters in this study was used, the ratings were treated as continuous variables. While strong inter-rater reliability was achieved, as with any human-labeled data, bias is a risk. Rather than increasing the number of raters, we suggest that future work prioritize validating and refining the rubric to reflect shared standards for high-quality rationale. One should note that the linguistic trends across the five dimensions (Fig. 3) were similar in direction and magnitude, pointing to the correlations between the dimensions. These similarities might occur based on the rubric or rationales that are not diverse enough. Considering that 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. We also acknowledge that this study examined only one structured rationale format (FSE), limiting the generalizability of our findings to other structured representations. Future studies should explore how rationale quality varies across multiple structured formats to identify best practices.

Due to the textual nature of the data, interpretable and explainable NLP and machine-learning approaches were prioritized. Early analyses explored BERT embeddings, term frequency-inverse document frequency, named entity recognition, and parts-of-speech approaches, which generated enormous amounts of features that, while significant in this dataset, raised concerns regarding interpretability and generalizability. Future work should consider iterating further on the model to reduce model complexity (i.e., number of features) or exploring a holistic measure for quality that integrates multiple dimensions. Therefore, iterating on the model should produce an improved computational approach to measure rationale quality that avoids overfitting or capturing irrelevant features in the data.

6 Conclusion

Design rationales articulate the justifications behind a design decision. The structure and content of these rationales impact their perceived quality and usefulness. This study examined the relationship between linguistic features and human-rated rationale quality across five dimensions: evaluating, analyzing, synthesizing, structure, and validity. A dataset of 2250 design rationales was evaluated by trained raters and characterized using natural language processing techniques. Results identified correlations between linguistic features and human ratings of quality, showed that FSE-structured rationales outperformed unstructured rationales, and built a predictive model that can assess the design rationale quality of new texts. This work contributes a scalable, data-driven approach to evaluating design rationales and provides a foundation for improving both human- and machine-generated rationale quality in design practice.

Acknowledgment

This work is based on preliminary research presented at the 2024 ASME International Design Engineering Technical Conferences (DETC2024-142788). 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.

Conflict of Interest

There are no conflicts of interest.

Data Availability Statement

The datasets generated and supporting the findings of this article are obtainable from the corresponding author upon reasonable request.

Appendix

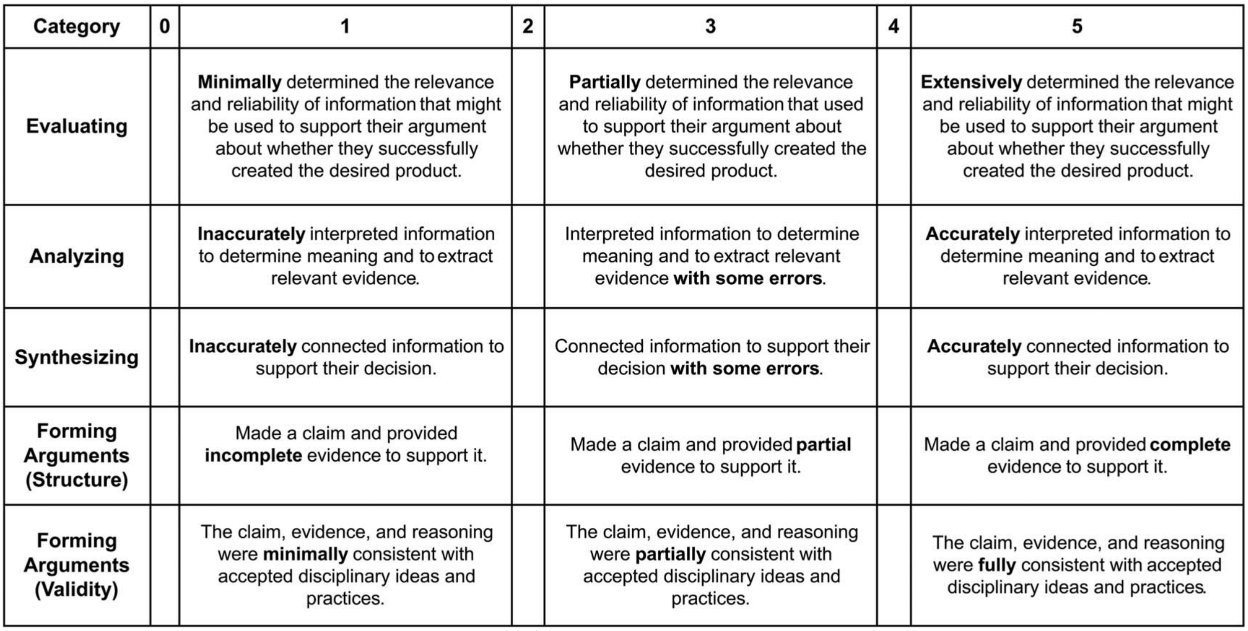
Table 6 List of 25 consumer products

Products
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

Table 7 List of personas

Title (experience level)
Mechanical engineer (entry-level)
Industrial designer (senior level)
Automotive engineer (mid-level)
User experience 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)

Full descriptions are not shown.



Category	0	1	2	3	4	5
Evaluating		**Minimally** determined the relevance and reliability of information that might be used to support their argument about whether they successfully created the desired product.		**Partially** determined the relevance and reliability of information that used to support their argument about whether they successfully created the desired product.		**Extensively** determined the relevance and reliability of information that might be used to support their argument about whether they successfully created the desired product.
Analyzing		**Inaccurately** interpreted information to determine meaning and to extract relevant evidence.		Interpreted information to determine meaning and to extract relevant evidence **with some errors**.		**Accurately** interpreted information to determine meaning and to extract relevant evidence.
Synthesizing		**Inaccurately** connected information to support their decision.		Connected information to support their decision **with some errors**.		**Accurately** connected information to support their decision.
Forming Arguments (Structure)		Made a claim and provided **incomplete** evidence to support it.		Made a claim and provided **partial** evidence to support it.		Made a claim and provided **complete** evidence to support it.
Forming Arguments (Validity)		The claim, evidence, and reasoning were **minimally** consistent with accepted disciplinary ideas and practices.		The claim, evidence, and reasoning were **partially** consistent with accepted disciplinary ideas and practices.		The claim, evidence, and reasoning were **fully** consistent with accepted disciplinary ideas and practices.

Fig. 6 Modified rubric to a design context based on Ref. [15]

Table 8 Sample design rationales and corresponding ratings across evaluation dimensions from both raters

Rationale	Dimension: R1R2
The protective cover for an electric shaver (F) was conceived and designed as an essential component to offer protection and enhance the product's longevity. The chief specifications (S) were put forth during the design phase that the shaver should have an added layer of security to prevent accidental damage to its blades and the sensitive shaving elements when not in use or during transportation. This necessity is underscored to maintain the product's functionality and efficiency over time, in turn, heightening user satisfaction and assurance. Empirical evidence (E) gathered from user interviews, product testing, and market research indicated users' preferences for a shaver that maintains its sharpness and effectiveness over prolonged use and can resist potential harm from drops or jostling. This evidence made a strong case for the incorporation of a protective cover as an intrinsic feature, which not only meets the functional specification but also raises the aesthetic appeal, safety, and overall user experience of the product.	E: 5 / 5 A: 3 / 5 Sy: 4 / 4 St: 5 / 5 V: 5 / 4
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: first, to create a user-friendly device, and second, 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.	E: 1 / 2 A: 1 / 3 Sy: 1 / 3 St: 2 / 3 V: 1 / 3
The cover on an electric pressure cooker serves a variety of crucial functions that contribute to the overall performance of the system. First, the cover creates a sealed environment which, along with the heat produced by the cooker, generates pressure that speeds up the cooking time substantially. This pressure environment enables heat to penetrate food quickly and cook evenly, resulting in better retention of nutrients and flavors. Second, the cover also plays a critical safety role. If the pressure builds too much, most modern cookers have a pressure-release valve under the cover that can safely vent the excess pressure. Furthermore, most covers include a locking mechanism, preventing the cover from being opened when the pressure inside is too high which could lead to a hazardous situation. Ultimately, the inclusion of cover in the design of electric pressure cooker enables efficient operation, enhances safety, and ensures optimal cooking results.	E: 1 / 1 A: 2 / 1 Sy: 2 / 3 St: 2 / 3 V: 1 / 1
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 consistently 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.	E: 5 / 5 A: 5 / 4 Sy: 5 / 4 St: 5 / 5 V: 5 / 4

References

[1] Moran, T. P., and Carroll, J. M., 2020, *Design Rationale: Concepts, Techniques, and Use*, CRC Press, Boca Raton, FL.

[2] 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**(3), pp. 211–231.

[3] 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**(3–4), pp. 209–235.

[4] Lee, J., 1997, "Design Rationale Systems: Understanding the Issues," IEEE Expert, **12**(3), pp. 78–85.

[5] 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," Int. J. Eng. Educ., **17**(4/5), pp. 343–348.

[6] Hertzum, M., and Pejtersen, A. M., 2000, "The Information-Seeking Practices of Engineers: Searching for Documents as Well as for People," Inform. Process. Manage., **36**(5), pp. 761–778.

[7] Das, D., and Chernova, S., 2020, "Leveraging Rationales to Improve Human Task Performance," Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, Mar. 17–20, pp. 510–518.

[8] Dong, A., Lovallo, D., and Mounarath, R., 2015, "The Effect of Abductive Reasoning on Concept Selection Decisions," Des. Stud., **37**, pp. 37–58.

[9] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F., 2019, "Human Evaluation of Models Built for Interpretability," Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019), Stevenson, WA, Oct. 28–30, Vol. 7, No. 1, pp. 59–67.

[10] Conklin, E. J., and Yakemovic, K. C. B., 1991, "A Process-Oriented Approach to Design Rationale," Hum. Comput. Interact., **6**(3–4), pp. 357–391.

[11] Bracewell, R., Wallace, K., Moss, M., and Knott, D., 2009, "Capturing Design Rationale," Comput. Aided Des., **41**(3), pp. 173–186.

[12] Chen, H., Brahman, F., Ren, X., Ji, Y., Choi, Y., and Swayamdipta, S., 2023, "REV: Information-Theoretic Evaluation of Free-Text Rationales," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Toronto, Canada, July 9–14, pp. 2007–2030.

[13] McNamara, D. S., Crossley, S. A., and McCarthy, P. M., 2010, "Linguistic Features of Writing Quality," Written Commun., **27**(1), pp. 57–86.

[14] 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," Behav. Res. Methods, **51**(1), pp. 14–27.

[15] 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," Int. J. STEM Educ., **7**(1), pp. 1–15.

[16] Crossley, S., and McNamara, D., 2010, "Cohesion, Coherence, and Expert Evaluations of Writing Proficiency," Proceedings of the 32nd Annual Meeting of the Cognitive Science Society, Portland, OR, Aug. 11–14, pp. 984–989.

[17] McNamara, D., and Graesser, A., n.d., "Coh-Metrix Version 3.0 Indices," Coh-Metrix Documentation.

[18] McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C., 2010, "Coh-metrix: Capturing Linguistic Features of Cohesion," Discourse Process., **47**(4), pp. 292–330.

[19] Breetvelt, I., Van den Bergh, H., and Rijlaarsdam, G., 1994, "Relations Between Writing Processes and Text Quality: When and How?" Cogn. Instruct., **12**(2), pp. 103–123.

[20] Mirabito, Y., Tchatchouang Kayo, M. A., and Goucher-Lambert, K., 2024, "Feature, Specification and Evidence Framework for Communicating Design Rationale," Des. Sci. **10**(20).

[21] 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," Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing (AIS 2024), Sydney, Australia, June 22–23, pp. 63–81.

[22] Koo, T. K., and Li, M. Y., 2016, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," J. Chiropractic Med., **15**(2), pp. 155–163.

[23] Wilson, R. C., and Collins, A. G., 2019, "Ten Simple Rules for the Computational Modeling of Behavioral Data," Elife, **8**, p. e49547.

[24] Lei, C.-U., Man, K. L., and Ting, T. O., 2014, "Using Coh-Metrix to Analyse Writing Skills of Students: A Case Study in a Technological Common Core Curriculum Course," Proceedings of the International Multiconference of Engineers and Computer Scientists, Vol. 2, Hong Kong, Mar. 12–14, pp. 823–825.

[25] Gero, J. S., and Kannengiesser, U., 2004, "The Situated Function–Behaviour–Structure Framework," Des. Stud., **25**(4), pp. 373–391.

[26] MacLean, A., Young, R. M., Bellotti, V. M. E., and Moran, T. P., 1991, "Questions, Options, and Criteria: Elements of Design Space Analysis," Hum. Comput. Interact., **6**(3–4), pp. 201–250.

[27] Lee, J., and Lai, K.-Y., 1991, "What's in Design Rationale?" Hum. Comput. Interact., **6**(3–4), pp. 251–280.

[28] 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," ASME J. Mech. Des., **142**(9), p. 091401.

[29] Raina, A., McComb, C., and Cagan, J., 2019, "Learning to Design From Humans: Imitating Human Designers Through Deep Learning," ASME J. Mech. Des., **141**(11), p. 111102.

[30] Yue, G., Liu, J., and Zhang, W., 2025, "Extracting Design Rationale in Technical Literature to Provide Inspirational Design Stimuli," ASME J. Mech. Des., **147**(7), pp. 1–30.

[31] Lee, J., 1990, "SIBYL: A Tool for Managing Group Design Rationale," Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW'90), Los Angeles, CA, Oct. 7–10, Association for Computing Machinery, pp. 79–92.

[32] Stahovich, T. F., and Raghavan, A., 2000, "Computing Design Rationales by Interpreting Simulations," ASME J. Mech. Des., **122**(1), pp. 77–82.

[33] Burge, J. E., and Brown, D. C., 2008, "Software Engineering Using Rationale," J. Syst. Soft., **81**(3), pp. 395–413.

[34] Eng, N., Aurisicchio, M., and Bracewell, R., 2017, "Mapping Software Augments Engineering Design Thinking," ASME J. Mech. Des., **139**(5), p. 051103.

[35] College Board, n.d., "Sat Essay Scoring," SAT Essay Scoring—SAT Suite—College Board.