Critique Style Guide: Improving Crowdsourced Design Feedback with a Natural Language Model

Markus Krause
ICSI, UC Berkeley
markus@icsi.berkeley.edu

Tom Garncarz
Carnegie Mellon University
tgarncarz@gmail.com

JiaoJiao Song
Huazhong University of Science and Technology
songjiaojiao1229@gmail.com

Steven P. Dow
Design Lab, UC San Diego
spdow@ucsd.edu

ABSTRACT

Designers have turned to online crowds for fast and affordable feedback. However online contributors may lack the motivation, context, and sensitivity to provide high-quality critique. Rubrics help critiquers, yet require domain experts to write them. This paper introduces automatic methods of extracting style-based language features to support feedback providers. Such style-based guides may benefit online feedback systems. In Study 1, 52 students across two design courses created artifacts and received feedback from 176 online feedback providers. Instructors, students, and crowd contributors rated the helpfulness of each point of feedback. From this data an algorithm extracted a set of natural language features (e.g., specificity, sentiment etc.) that correlate with helpfulness ratings. The features accurately predict helpfulness and remain stable across different raters and design artifacts. Based on these features, we designed a critique style guide with automatically picked examples for each feature that support critique providers to self-assess and edit their initial critique. Study 2 validates the guide with a between-subjects experiment (n=50). Participants gave feedback on design artifacts with or without the guide. Providers generated critiques with significantly higher perceived helpfulness when using style-based guidance.

ACM Classification Keywords

H.5.3. Information Interfaces and Presentation (e.g. HCI): Group and Organization Interfaces—*Computer-supported cooperative work*

Author Keywords

Design; critique; feedback; crowdsourcing; expertise; rubrics.

INTRODUCTION

IN SUBMISSION; PLEASE, DO NOT DISTRIBUTE.

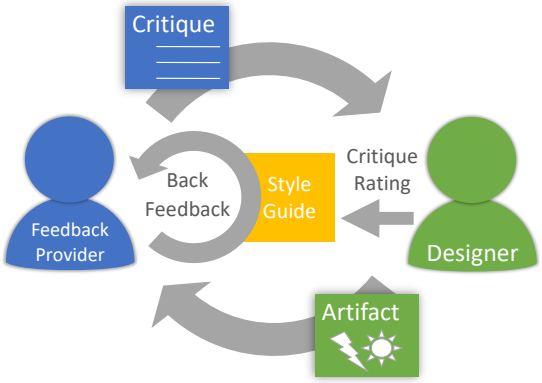

Figure 1. The critique style guide is based on a natural language model and assists critique providers to improve their initial feedback before the final iteration is send to the designer. The style guide consists of examples of high quality critiques that highlight specific stylistic aspects of good critiques. We developed a natural language model to mine high quality critiques and critiques with specific features.

Feedback helps designers gain an external perspective to improve their work [51]. Receiving feedback on design artifacts can happen through peers, mentors, or target users who provide comments and suggestions. With the rise of online labor markets, feedback can be obtained almost immediately [3]. Furthermore, with a growing demand for design education, designers in online as well as offline classes look into extending traditional methods of design feedback to personalized results [20, 51]. Crowdsourcing feedback is appealing due to its scalability, availability, and affordability. Researchers have explored various crowdsourcing methods to support designers [31, 50] and several online communities for crowd feedback exist, such as Forrst [53], Photosig [49], and Dribbble [32].

A concern regarding online community or crowd driven sources is that these approaches produce feedback of poor quality or low quantity [49]. The reasons for this are diverse contributors may lack the motivation [49], context [3], knowledge [31], and sensitivity [52] to provide high-quality design feedback. To address this, some crowd-based systems

break down feedback provision into micro tasks (e.g. [3, 50]) or provide rubrics to workers (e.g. [14, 31, 20]).

Recent research indicated the plausibility of obtaining relevant and rapid crowd feedback. However, this prior work requires experts to break down the task into more accessible sub tasks, or to write rubrics that embed key principles in a domain. Breaking down design feedback is difficult, especially when complex artifacts have to be evaluated. Personalized expert intervention is expensive. Such approaches place a high demand on experts which could limit the value of crowd-based feedback across different domains.

Our research introduces a more scalable and domain-independent method for design feedback. We present a natural language model that automatically extracts language features that correspond with student ratings of perceived helpfulness. Based on these language features we compiled a critique style guide that offers guidance to online feedback providers (figure 1 shows a sketch of the process).

In **study 1**, we collected student design artifacts from two different project-based university design classes and hired online workers (via Amazon Mechanical Turk and Upwork) to provide critiques. The students independently rated the helpfulness of each critique. We also collected perceived helpfulness ratings on a sub-sample of the critique's from design instructors (professors from three US universities) as well as from online workers hired via Mechanical Turk.

To identify the features that our three populations found most helpful, we conducted a linguistic analysis on the writing style of the collected critiques. We found evidence that critique length, emotional content, language specificity, grammatical mood and complexity of sentences, word complexity, and the presence of justifications, correlate with higher ratings. A random forest classifier trained with these features is able to predict the average perceived helpfulness with Krippendorffs alpha [28] levels close to or even higher than the inter rater reliability (IRR) of human raters.

Study 2 applies the findings of study 1 to improve perceived critique helpfulness. In a between subjects study we randomly split a pool of 90 online contributors into two groups. Both groups provided critiques to the same design artifacts from study 1 and were asked to improve their critique after submitting their initial draft. Contributors in one group received a style guide asking them to improve specific features of their writing based on our natural language model. The style guide consists of examples automatically extracted from results of the first study using our model. A control group received only general instructions. We found that the contributors in the first group significantly improved their average correlation with our natural language model as well as their critique ratings.

With these two studies this paper makes the following contributions.

1. Describing a set of natural language features that correlate with perceived critique helpfulness. **(study 1)**
2. Demonstrating that these correlations are stable across two different design tasks and three different rater populations. **(study 1)**
3. Demonstrating that these features allow predicting the perceived critique helpfulness. **(study 1)**
4. Illustrating that a style guide using automatically mined examples of these features can improve perceived critique helpfulness. **(study 2)**

RELATED WORK

Feedback and practice are key elements in developing new skills [39] and gaining insight to better understand how ones work is perceived by others [25]. In design, feedback supports designers in developing their next design iteration [15] and helps novices to better understand design principles [17]. Feedback can also help to explore and compare alternatives [12, 44].

Feedback Sources

Designers gather feedback from various sources. In educational settings, instructors provide feedback by writing comments on drafts or proposals and by grading assignments. It has been employed successfully in many contexts including design [11, 43, 29], programming [6], and essays [47]. Self-assessment has can achieve results comparable to external sources of feedback [14]. Scaling feedback in educational settings often involves peer reviews [35]. The benefit of peer reviews is that students learn from providing feedback to peers [36]. Critiquing work of peers helps students to practice their revision skills and strengthen their ability to find and solve problems [36]. Despite the positive effects there is scepticism whether students of all ability levels are capable of helping their peers [35].

Crowd feedback can be obtained through a variety of tools, methods, and platforms. Paid critique providers can be engaged on services such as CrowdFlower [10], Mechanical Turk (MTurk) [1], or UpWork [45]. Crowd feedback is particularly appealing due to its scalability and availability. Crowds are capable of contributing diverse perspectives that may be difficult to find within a classroom [13].

Involvement in communities such as Behance [2], Forrst [53], and Dribbble [32] are ways for experienced designers to give and receive design feedback. These platforms require a certain level of commitment and expertise to fully experience their potential. This somehow limits their accessibility [8]. Participants tend to be motivated to develop their own skills and status [49]. Novices in such communities experience evaluation apprehension and may be hesitant to share preliminary work [32] making these communities an option for advanced designers rather than novices.

Measuring Feedback Quality

Measuring feedback quality is challenging and prior work uses a range of measurements. Luther et al. compares differences between design iterations [31, 51]. While others contrast critiques with feedback produced by experts [31, 29]. Measuring post-feedback design quality [12], and collecting

designer ratings on the helpfulness of feedback [7] are other viable methods to measure feedback quality.

Various definitions exist that describe qualities of effective feedback. Sadler [39] argues that effective feedback help to understand the concept of a standard (conceptual), compare the actual level of performance against this standard (specific), and engage in action that reduces this gap (actionable). Cho et al. [7] examined the perceived helpfulness of feedback in the context of writing psychology papers and found that students find feedback more helpful when it suggests a specific change and when it contains positive or encouraging remarks. Xiong and Litman [48] investigate peer feedback for history papers and constructed models using natural language processing to predict perceived helpfulness. They found that lexical features regarding transitions and opinions can predict helpfulness.

This study uses perceived helpfulness as the measure of critique quality. Perceived helpfulness captures the value of feedback for its recipient and mediates the interaction between feedback and later revisions [38].

Improving Feedback

The main challenge with all crowd-based methods such as crowdsourcing, communities, or peer feedback is that crowd driven sources often produce feedback of poor quality or low quantity [49]. The reasons for this are manifold. Contributors may lack the motivation [49], context [3], knowledge [31], and sensitivity [52] to provide high-quality design feedback. Prior work has contributed screening processes to disqualify workers that provide constantly low quality responses [16] as well as other mechanisms such as the Bayesian Truth Serum [41] to increase work quality.

Voyant and CrowdCrit structure design feedback tasks for online crowds. Both systems are motivated by the goal of producing higher quality feedback from inexperienced workers. Recent studies compare the *characteristics* of feedback produced by these structured systems against both open-ended feedback and expert feedback with promising results [31, 51, 20]. Soylent [3] is a crowdsourcing tool for efficient proofreading tool that incorporates paid crowdsourced workers in a structured way avoiding the need for a user to have expertise in constructing human interface tasks.

Kulkarni et al. [30] reports that peer grades correlated highly with staff-assigned grades in two observed MOOCs although students had a tendency to rate their work higher than staff. To counter act this grade inflation Piech et al. [37] uses a calibrated peer assessment.

Automated feedback has been applied in various contexts such as essay grading [21, 46, 40, 7], kitchen design [19]. Especially for MOOCs automated essay grading is an interesting solution due to its scalability. Complex artifacts however might not be automatically gradable. Our approach overcomes this challenge letting humans analyse the artifact. The generated critique can be analysed using automated systems. Similar in its structure to an essay [26] computational models might be able to analyse critiques effectively [42].

RESEARCH QUESTIONS AND HYPOTHESES

This paper explores which style-based natural language features correlate with perceived helpfulness of critiques and how a style guide derived from these features affect the way people provide design critique. With this in mind, we investigate the following research questions:

RQ1: Which stylistic natural language features correlate with perceived critique helpfulness?

RQ2: Are these correlations stable across different populations and tasks?

RQ3: How can these features improve perceived critique helpfulness?

We hypothesize that valuable critiques incorporates the qualities suggested by Sadler [39], Cho et al. [7], Yuan et al. [52] and Krause [26, 27]. That is, a valuable critique is conceptual in that it incorporates design domain knowledge, specific in that it presents a clear issue, actionable in that it provides guidance on how to resolve the issue, and positive in that it also encourages the recipient. All studies presented so far used human annotators to extract language features in contrast we use an automated system that offers a better scalability.

Although some features of good critiques are known it can be difficult to find good examples for each feature especially in an automated and scalable way. Yet good examples are necessary to teach a feedback provider how to write good critiques. Our approach enables mining existing critiques to find critiques that highlight linguistic features associated with high quality. We use our language model to generate a style guide that we suspect to significantly increase the frequency of these features and enhance the perceived helpfulness ratings of critiques.

STUDY 1: PREDICTING PERCEIVED HELPFULNESS

In the first study we collected design artifacts from students in two design classes. Online contributors recruited on Mechanical Turk provided critiques on these design artifacts. Students, instructors, and online contributors rated these critiques. As a final step we used our natural language model to extract feature vectors from the collected critiques. We estimated the correlation between the features and perceived helpfulness and predicted the average perceived helpfulness with our model.

Apparatus

We use 2 different tools to 1) let online contributors from Mechanical Turk critique student design artifacts and 2) to allow students, instructors and a other online contributors to rate these critiques.

Figure 2 shows the interface that was used to collect design critiques from online contributors recruited through MTurk. Contributors were presented with 3 design artifacts (one at a time) and asked to write a critique for each artifact. The figure shows the complete interface including the style guide (sub figure b). In study 1 critique providers only used first part of the interface to write critiques (sub figure a).

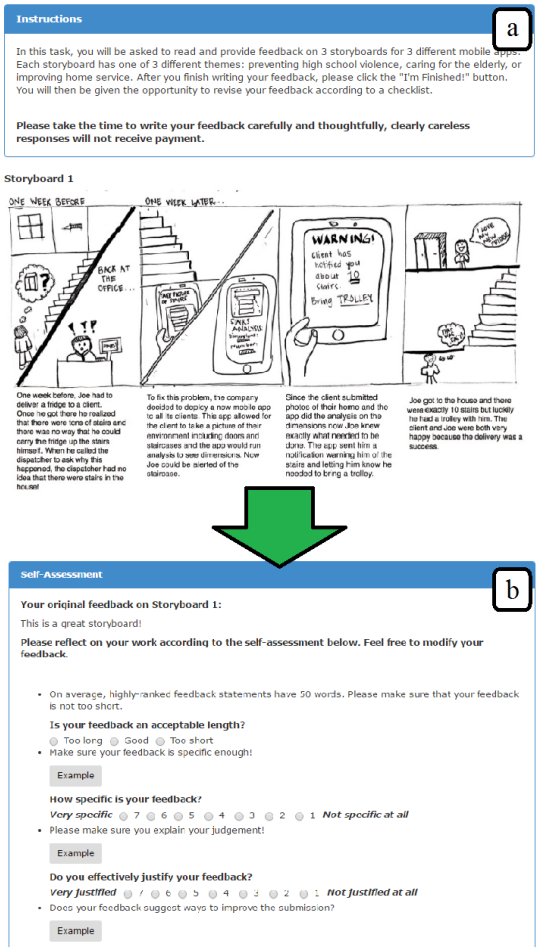

Figure 2. The interface used to generate critiques. Critique provider recruited via Mechanical Turk were asked to write a number of critiques (called feedback in the figure). In study 1 critique provider only saw the first part of the interface (sub figure a). In study 2 critique provider in the guided condition were also asked to revise their feedback according to a style guide (sub figure b). Critique provider in the control condition of study 2 were asked to revise their critique but without the style guide.

Figure 3 shows the interface used to rate each critique collected. A critique rater sees a design artifact and a list of critiques to this interface. A Likert scale ranging between 1–7 is shown below each critique and critique provider can see all artifacts and critiques on one page.

Procedure

We collected design artifacts from 2 non overlapping student populations in 2 independent design classes. In the first setting, students created storyboard artifacts for a team assignment focused on mobile phone applications. In the second setting, students individually designed a dashboard to display weather forecasts.

Storyboard Artifact Collection

In our first classroom study 37 undergraduate students created 42 storyboards in groups of up to 4 students in a course on mobile service design. They designed mobile phone applications in the domains of home service, high school violence

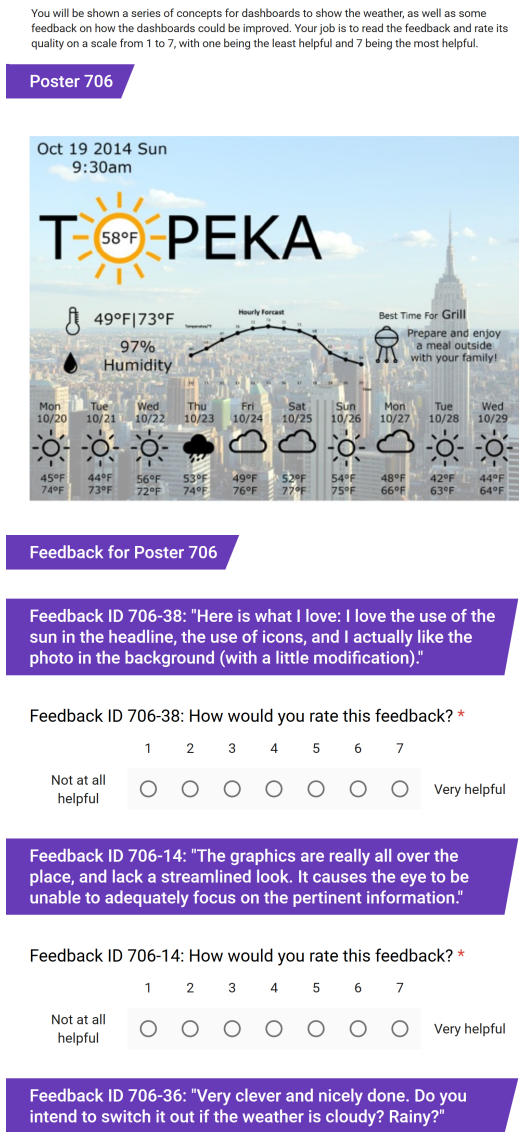

Figure 3. Critique raters see all design artifacts and critiques on one page. They are asked to provide a rating for each critique and are allowed to freely scroll on the page. A Likert scale ranging between 1–7 is shown below each critique.

prevention, and elder care (figure 4 shows 3 examples). 71 independent critique providers recruited from Amazons Mechanical Turk evaluated and rated the storyboards.

To normalize the population's language skill, we accepted only US-based contributors. Contributors critiqued 3 storyboards from one group each and were compensated $3 to match the expected pay rate of US minimum wage. These numbers ensured that each design received critiques from at least 3 workers. We collected a total of 568 critiques.

Dashboard Artifact Collection

We recruited 15 students from an undergraduate-level design course. Each student submitted 1 design from a course assignment which involved creating a weather UI dashboard. Figure 4 shows 3 samples of the submitted designs. To gener-

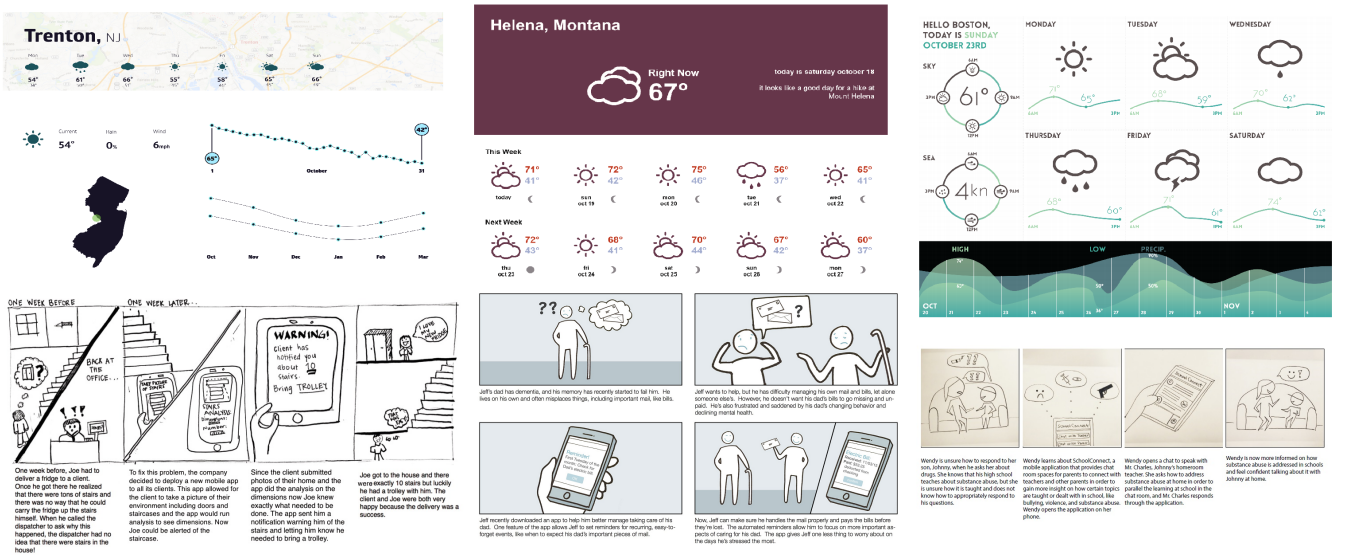

Figure 4. The top 3 elements show storyboard examples designed by students for mobile phone applications in a design class. The lower 3 elements are designs submitted by students for a course assignment which involved creating a weather UI dashboard.

ate critiques, we recruited 24 contributors from Amazon Mechanical Turk (MTurk). To help normalize the population's language skill, we restricted both pools of workers to consist of US-based workers only. Contributors critiqued 4 designs each and were compensated $3 to match the expected pay rate of US minimum wage. These numbers ensured that each design received critiques from at least 3 workers. We collected a total of 615 critiques.

Critique Ratings from Students and Instructors

After all critiques in a setting were submitted, the student designers rated the helpfulness of each critiques they received. Critiques were shown one critique at a time in random order, and students rated helpfulness on a 1-7 Likert scale (7=best) for each critique. Students rated only critiques given to their groups or respectively their own work. Which means we collected multiple ratings per critique for the storyboard (group assignment) setting yet not for the dashboard setting (individual assignment). A subset of these critiques in both settings were rated by 3 instructors. Only one instructor was rating critiques from both settings. The instructor are professors of assistant and associate levels from 3 US universities.

Instructors rated 141 critiques in the storyboard setting and 60 critiques in the dashboard setting. Following the same procedure as students. Some instructors rated more critiques than others. Participants were asked to read each critique, and provide a rating of the helpfulness of the critique. The task took our instructors approximately 1 hour to complete.

Critique Ratings from Online Contributors

To ensure a diverse set of contributors we conducted a preliminary demographic survey on Mechanical Turk (n=60) to find contributors with more and less experience in writing and receiving design critiques. From this pool, we selected 12 workers (6 for each setting) to rate critiques. We selected 3 experienced and 3 less experienced contributors for each classroom setting based on the self-reported experience in design education and professional design work. We considered a contributor to be experienced when they worked at least 5 years as a professional designer or teached design for at least five years. Contributors used the same procedure and rated the same subset of critiques as instructors. The task took participants approximately 1 hour to complete, and they were compensated $10.00 for their participation.

Measures

In our Study we consider 2 main responses expressed in 2 dependent variables, four independently variable factors, and eight covariates. We describe the properties and operationalizations of these variables in the following sections.

Dependent Variables

Our dependent variable is the subjective measure of critique helpfulness. We asked all populations to provide a rating ranging from 1 to 7 on how helpful they consider the given critique to be (**helpfulness**).

Independent Variables

The independent variables of the study are 2 factors with 2 and 3 levels. We encode if a rating was obtained in one of the 2 classroom settings by the variable **experiment** with the levels dashboard and storyboard. The second independent variable **population** encodes if a rating was given by a online contributor crowd, a student, or the instructor.

Covariants

All language features explained in the language model section below are covariates for the analysis. For the covariate analysis we aggregate our results so that each critique provider corresponds to one observation, resulting in 176 individual observations.

Language Model

Feature	Low	High
complexity	*The blue/gray color palette is great but adding a third, possibly complementary colors could help highlight areas and potentially give viewers a pathway through the display.* **helpfulness = 7.0**	*Images are too small to be seen. Need to be blown up to larger sizes.* **helpfulness = 3.3**
rarity	*(...) it would almost be like a sense of privacy being invaded for the person they are catching up on (...)* **helpfulness = 6.0**	*I thought this was clear and easy to understand.* **helpfulness = 3.0**
specificity	*This seems like a good way to keep a dementia patient safe without physically being with them.* **helpfulness = 6.0**	*I like the first one the best.* **helpfulness = 3.0**
justification	*When you move these to the center, increase the size as to promote them as the most important area of the design.* **helpfulness = 7.0**	*(...) The first one is the easiest to implement and more promising, while the last one needs a lot more clarification and support to backup the idea* **helpfulness = 4.3**
actionable	*Days of week font color could be difference (navy blue or same orange as "Today") to make optics clearer.* **helpfulness = 6.3**	*The handwriting is small and the pictures are kinda blurry (...)* **helpfulness = 5.0**
sentiment	*Excellent idea!!! Are the water drops a representation of precipitation?* **helpfulness = 4.3**	*aaaaaaaagh jesus that sounds awful for all involved (...)* **helpfulness = 6.0**
subjective	*This almost seems like it could be used AS a form of bullying - a popular student could start a rumor and tell people to "like" or "vote up" the story. I feel giving the tools to create AND use the crowd could easily be abused.* **helpfulness = 5.3**	*The app is a one-stop-shop which lessens the load on the caregiver. However, he should confirm with his mother that she is alright with being videotaped to maintain autonomy.* **helpfulness = 5.5**

Table 1. The left most column gives the feature name. For a comprehensive explanation of each feature please see the language section of study 1. The second column (low) gives an example taken from the collected data that is ranked below the 25th percentile for the given feature. The last column (high) shows an example of a critique ranked above the 85th percentile for the given feature. The bold text below the example gives the average rating of the critique across all populations.

Our natural language model extracts 257 features in 12 categories. We used 8 of these categories in our study (see table 1). We left out features such as character frequency and part of speech frequency as those features tend to be predictive only for very large data sets. The feature extractor of our model is written in python and uses the Natural Language Toolkit (NLTK [4]) and the *pattern.en* package for processing critiques.

We preprocessed all critiques with the NLTK part-of-speech (POS) tagger [4] and filtered stop words and words not in Wordnet [18]. Wordnet is a natural language tool that provides linguistic information on more than 170,000 words in the English language. We also lemmatized the remaining words to account for different inflections.

The most basic feature we examined is critique **length** operationalized as number of characters. We counted every alphanumeric character including punctuation and special characters but not spaces. The average critique length is M = 123 (SD = 128).

The second feature text **complexity** is operationalized as the automated readability index (ARI [24]) the higher the value the more complex the text. Another very similar metric is word **rarity** which is operationalized as the term frequency.

The **specificity** feature measures how deep each word appears in the Wordnet structure [18]. This feature is not yet well explored but preliminary research indicates that it is a strong predictor for text quality in various scenarios [26, 27]. Words that are closer to the root are more general (e.g. dog) and words deeper in the Wordnet structure are more specific (e.g. Labrador). Word depth ranges from 1 to 20 (20 = most specific).

Previous studies predicted that the amount of **justifications** may correlate with positive ratings. These studies used human annotators to extract this feature [23]. We operationalized this feature with a bag of words approach counting words that indicate a justification (e.g. because).

A feature also found to be predictive of perceived helpfulness is how **actionable** the provided critique is. As argued by Sadler [39] effective feedback help to engage in an action that reduces the gap between a given standard and the actual level of performance against this standard. We operationalize this feature with the grammatical mood of sentences in each critique. Each sentence was classified as either indicative (written as if stating a fact), imperative (expressing a command or suggestion), or subjunctive (exploring hypothetical situations). The feature, which we refer to as actionable, correspond to the ratio of non-indicative sentences in a critique, with values falling between 0 and 1 (1 means all sentences are non-indicative or active). We used pattern.en to extract the sentence mood. As the perception of critiques is subjective we include also features to measure sentiment and subjectivity. For both features we used classifiers provided by the pattern.en tool kit. The values for these features fall in the range between 0 (low) and 10 (high).

The next 2 features we looked at were **sentiment** and **subjectivity**. Yuan et al. [52] illustrated that a positive sentiment is a predictive for perceived helpfulness. The sentiment refers to whether a critique is positive or negative. A value of 0 is a strongly negative sentiment, 5 is a neutral sentiment, and 10 is a strongly positive critique. Subjectivity refers to whether a critique uses an emotional language or has a more objective tone. The feature value ranged from 0 to 9 (9 highly subjective). We used *pattern.en*, a tool based on *NLTK*, to extract sentiment and subjectivity. A list of examples for each feature from the collected critiques can be found in Table1.

RESULTS

The first study investigates the correlations between our language model and perceived helpfulness. We also aim to demonstrate that our language model is predictive for perceived helpfulness.

Features Correlate Non-linearly with Helpfulness

As figure 5 illustrates the observed features do not linearly correlated with perceived helpfulness but all features show a nonlinear correlation.

To estimate the nonlinear correlation we use a method called local polynomial regression fitting. The method is described in detail by Cleveland et al. [22]. The model creates a polynomial surface. With this surface we predict perceived helpfulness from the fitted language feature. We calculate the correlation between this prediction and the actual perceived helpfulness using Pearson product moment correlation.

In accordance with [34] we choose Pearson correlation. Alternative methods such as Spearman correlation yield inaccurate p-values with ties. Do to the relatively high sample size these ties occur frequently within the data. Table 2 shows correlations and p-values for each feature. Confidence intervals are obtained through bootstrapping using 10K bootstrap samples.

Correlations are Stable Across Tasks and Populations

We found that correlations are stable across the 2 artifact collections as well as within all our populations. We calculated non linear correlations based on a decision surface obtained with a local polynomial regression fit (ρ). All obtained p-values for this table are below the 0.01 alpha level. We interpret correlations over 0.3 as weak, above 0.5 as moderate, and over 0.7 as strong. Values below 0.3 are considered uncorrelated. Table 2 shows all calculated correlations.

Feature	Avg.	Dash.	Story.	Crowd	Student	Instr.
length	0.73	0.76	0.73	0.62	0.63	0.75
justification	0.57	0.68	0.56	0.34	0.46	0.59
specificity	0.55	0.70	0.61	0.40	0.54	0.56
complexity	0.52	0.45	0.56	0.50	0.46	0.50
rarity	0.47	0.34	0.54	0.48	0.43	0.40
active	0.45	0.51	0.44	0.30	0.38	0.54
subjective	0.40	0.21	0.51	0.30	0.34	0.36
sentiment	0.34	0.52	0.42	0.35	0.34	0.26

Table 2. Most of the features in our language model correlate non-linearly with perceived helpfulness. We calculated ρ based on a decision surface obtained with a local polynomial regression fit. All p-values are below the 0.01 alpha level. We interpret correlations over 0.3 as weak, above 0.5 as moderate, and over 0.7 as strong. Values below 0.3 are considered uncorrelated. The columns *Dash.* and *Story.* give the correlations for the dashboard and the storyboard artifact collection. The last there columns give the correlations for the online crowd, students, and instructors.

Population	Mean	SD	IRR	Pred. Avg.	low	high
Combined	4.7	1.7	0.25	0.39	0.23	0.58
Instructor	3.9	2.0	0.67	0.59	0.46	0.71
Student	4.8	1.7	0.35	0.42	0.28	0.68
Crowd	4.9	1.5	0.32	0.41	0.24	0.65

Table 3. Mean and SD rows indicate the average critique rating a population gave to critiques. The IRR column gives the inter rater agreement among human raters (Krippendorff's alpha). Our language model can be used to predict the average rating a critique will receive. The column Pred. Avg. gives the Krippenforff's alpha between the prediction of the average and the observed average rating of critiques. Rows split the results based on rater populations. High and low columns give the lower and upper bounds of the 95% CI.

The Language Model can Predict Perceived Helpfulness

As previous research has indicated natural language models can predict essay grades with a high accuracy sometimes even outperforming human raters [42]. We were interested in the question if our model is equally capable of predicting average helpfulness ratings in our data set. We used a random forest regressor generating 500 random trees and used gini impurity as the split criterion [5]. We found that our model is capable of predicting of predicting average helpfulness ratings. Table 3 shows the Krippendorff's alpha values calculated comparing the true average and the prediction made by the regressor. When the inter rater reliability in a group of human raters is low the regressor gives better predictions of the average rating of human raters.

STUDY 2: CRITIQUE STYLE GUIDE INTERVENTION

The second study investigates the effect of the language model on the perceived helpfulness of critiques. We conducted a randomized control study with 2 conditions (guided, control). Participants in the guided group received a critique style guide to revise their initial critique while participants in the control group received only general instructions to improve their work. We analyzed the effect the style guide had on the language features and perceived helpfulness scores.

Critique Style Guide

The style guide provides five comments and examples of high rated critiques (see Table 4 for an overview). The examples were selected using our natural language model. We selected

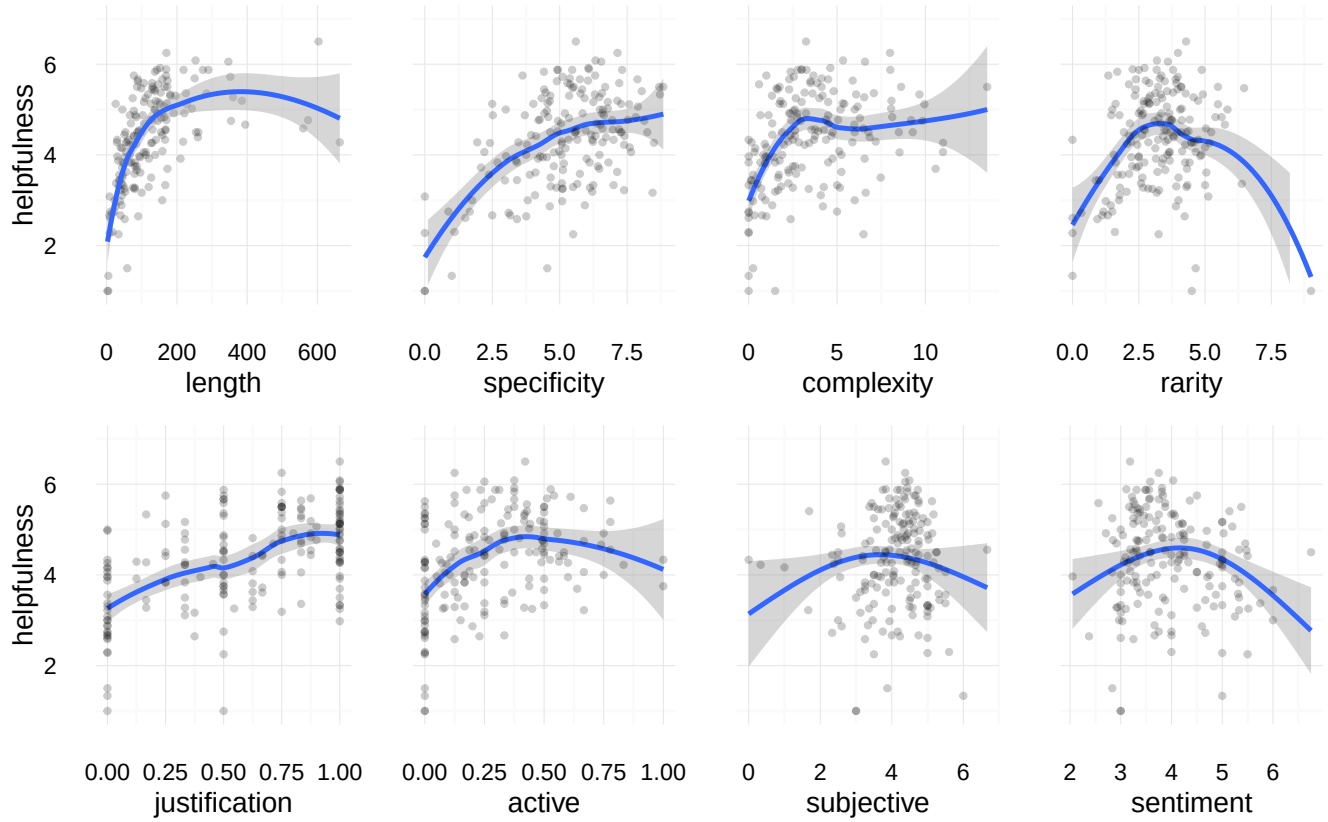

Figure 5. Correlation between the eight observed language features and the perceived helpfulness across all populations (instructors, students, and crowd). Each point is the aggregated average for one critique provider on the given feature. The blue line is calculated using local polynomial regression fitting [22]. The correlations for these surfaces can be found in Table 2

highly rated critiques that also scored high for the given feedback. We excluded some features from this process that were hard to explain to critique providers. We excluded grammatical complexity and word rarity. The interface is similar to the interface for study 1 and can be seen in Figure 2.

Procedure

From the first study we selected a sample of artifacts to be critiqued again. The 3 selected designs are shown in the top row of Figure 4. We selected 1 artifact from each of the 3 domains (home service, high school violence prevention, and elder care).

Collecting Critiques

We recruited 90 contributors from MTurk. Contributors where recruited only from US to reduce language bias. 45 contributors worked in the treatment condition and 45 in the control condition. Each contributor provided 1 critique for each storyboard. After providing all 3 critiques participants were asked to revise their critiques. Contributors in the guided condition were asked to edit their critiques using the style guide as seen in Figure 2 directly below the green arrow.

Contributors in the control condition did not use the style guide but were asked to revise and improve their critique. Both populations received a bonus of $1.00 if their revision improved the critique. This bonus payment was payed regardless of quality but after the task in both conditions was finished. We asked all critique providers to give feedback on 2 questions 1) does the editing process help improve my work and 2)do you like the editing process. All questions asked were measured on a Likert scale from 1–7 (absolutely agree). Additionally we asked an open ended question on how the editing process affected their critique.

Rating Critiques

To rate the collected critiques and estimate the improvement we recruited another group of 20 contributors from MTurk. Each contributor rated 75 critiques following the same procedure as described in study 1 (see Figure 3 for reference). Critiques were ordered randomly and each contributor received rated the same critiques. In contrast to study 1 we collected critique ratings only from online contributors.

Measures

The study investigates 2 independent variables, 4 dependent variables, and the covariants derived from the natural language model.

Independent Variables

The manipulated independent variables in this study are **edited** and **condition**. The edited variable has 2 levels. All initial critiques are labeled before while revised critiques are

Comment	Example
On average, highly-ranked feedback statements have 50 words. Please make sure that your feedback is not too short.	We did not provided a specific example for a long critique.
Make sure your feedback is specific enough!	*This seems like a good way to keep dementia patients safe without physically being with them.*
Please make sure you explain your judgement!	*I think the solution presented in the storyboard is a good idea, but there are a few issues. The first 1 is that the solution seems to only pertain to this specific situation. Many people don't have a home service system nor a home security monitor. Secondly, regardless of how she let the service man in (because the door is broken, hidden key, unlocked back door, etc.), not everyone would feel comfortable with leaving that accessible.*
Does your feedback suggest ways to improve the submission?	*I like that this shows how responsive the app can be and how it can prevent future problems. I would like to see how it integrates with the other aspects, though (get notified of a problem at work, use the app to find a service man, turn on the security camera and allow him access when he gets there, all through one app)*
The highest rated feedbacks are generally slightly positive. Make sure your feedback isn't too negative.	*I think this is a good starting point. I would like to see how this app would react when it looses it's internet connection as I think it is important to notify the user that he or she is no longer protected.*

Table 4. The style guide contains comments and examples for five features. The first column gives a comment provided to critique providers and the second column the automatically retrieved critique. We mined the critiques from the previous experiment using our language model to find highly rated critiques that also highlight a specific feature.

labeled after. The condition variable has two levels guided if the style guide was used to revise the critique and control without the style guide.

Dependent Variables

The main dependent variable is again perceived **helpfulness**. Additionally we investigate how critique provider perceive the intervention. We measured the **helpfulness** of the intervention, how much they **liked** the back feedback process. All variables were meassured on a Likert scale from 1–7.

Covariates

As in the previous experiment we extracted 8 natural language features from the collected critiques following the same process as in study 1.

Results

The second study aims at illustrating the automatically extracted high rated critiques can be used as examples in a style guide and improve percevied helpulness.

Feature Presence Increases with Style Guide Use

To estimate if the presence of features in our model increases significantly more when critiquers use the style guide we conducted a multivariate analysis of variance (MANOVA) [9]. Prior to conducting the MANOVA, we ensured that our data meets the necessary requirements as described by Meyers et al. [33]. The MANOVA showed a significant multivariate interaction between condition and editing $F(7, 183)=3.413$, $p=0.04$. Figure 6 shows the changes for each feature.

The Guided Intervention is Perceived More Helpful

We asked critique provider in both conditions how they perceived the intervention. We found that critiquer liked the guided intervention ($M = 4.26$, $SD = 0.91$) significantly ($t(89)=2.13$, $p=0.03$) more than the control condition ($M = 5.01$, $SD = 0.96$) they also perceived the guided editing process to be significantly ($t(89)=2.52$, $p=0.01$) more helpful ($M = 4.06$, $SD = 0.96$) than the control ($M = 4.95$, $SD = 0.91$). We also received various comments that critiquer followed the style suggestions.

> *I had to take a different approach. Initially I focused more on the visual aspects of the storyboards. I was also too wordy and not concrete enough in my feedback. I tried to fix this as best I could.*
>
> **guided**

> *A lot, I realized that my feedback could be more efficient with the examples and guidelines.*
>
> **guided**

> *It made me feel like I had to change things, but I'm not sure that any of my changes were improvements, at best they were lateral moves, and they very well could have been worse. What's the point of editing without feedback on which to base the edits?*
>
> **control**

The Guided Intervention Improves Critiques More

The final question of this study is whether using the style guide improves the perceived helpfulness of critiques more than the control condition. We analyzed the results using a 2 way ANOVA and found a significant interaction between the 2 variables condition and edited $F=(3,596)=4.09$, $p=0.01$). The increase in perceived critique quality using the style guide is 8% higher compared to the increase without the guide.

DISCUSSION

We now revisit our original research questions and discuss our findings from the results.

RQ 1: Which stylistic natural language features do correlate with perceived critique helpfulness

We found that all features discussed in this paper do correlate significantly with helpfulness. It is however important to accept that these correlations are not linear and have to be

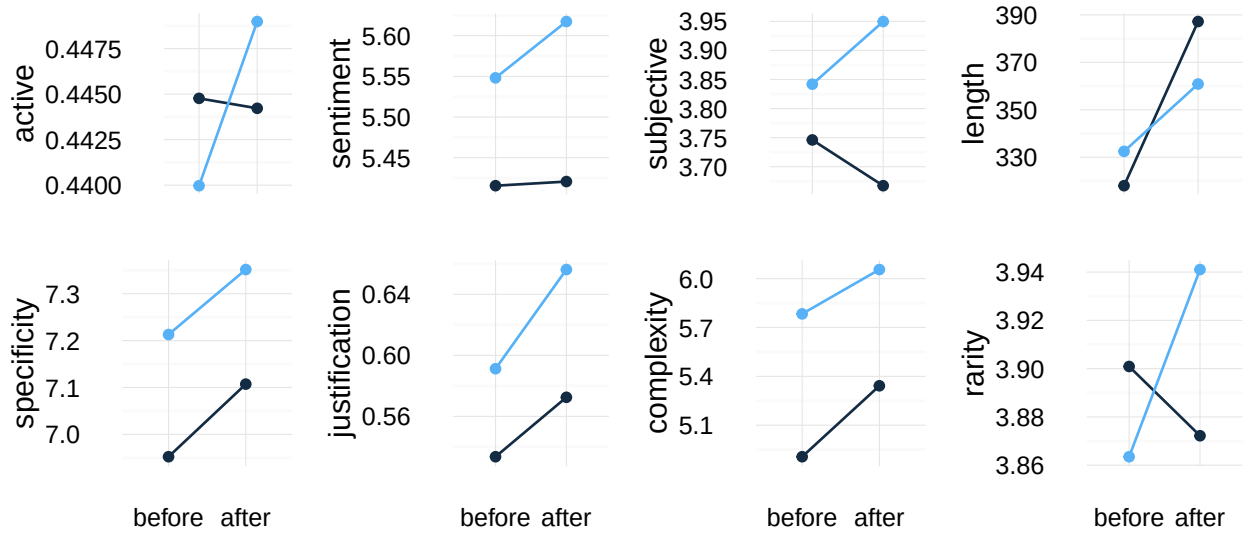

Figure 6. The presence of the language feature increases more when critique provider use the style guide. The light blue line shows the increase with the style guide and the black line without style guide.

investigated with advanced statistical models. Some correlations are also relatively weak. One reason for lower levels of predictive power in some features is the accuracy of the used feature extraction method. Yet the main challenge in predicting helpfulness is the high variance between and within rater populations. Nonetheless the used language model is able to predict the average for individual populations with Krippendorffs alpha values close to the interrater reliability of the population. In cases with very low IRR the model is even more predictive for the average helpfulness than individuals in the population.

RQ 2: Are these correlations stable across populations

We found that the features in our model show significant correlations across all populations and tasks. The prevalence of individual features however shift between populations and tasks. Furthermore, the inter rater reliability in some populations is very low. This might indicate that their is a personal as well as a task specific component to the importance and shape of individual features. Instructors had a stronger correlation with most of the features compared to students and online contributors. This might indicate that it requires a certain expertise, training, and awareness of these features to value them.

RQ 3: How can these features improve perceived critique helpfulness

Our model is able to predict and find high quality critiques that highlight specific stylistic features. A style guide using these critiques as example to help critique providers to reflect on their work is effective in improving the perceived helpfulness in a back feedback process.

LIMITATIONS AND FUTURE WORK

This work has illustrated that a style guide can provide support for critique provider. We think that the results of our work are promising and give way for many future investigations.

Extending the Language Model

This study analyzed 8 features and used 5 of these features in the actual style guide. Future work should extend this feature space. A possible avenue could be to extend the feature space by mining n-grams of highly rated feedback and thereby collect a vocabulary of relevant words and phrases for a specific domain. Similar approaches have been successful in a variety of tasks so far. Another interesting question is how the accuracy of a language model influences the performance of this method. Some features showed a relatively low correlation although literature suggest a profound impact on feedback quality (e.g. sentiment). One reason for lower levels of predictive power might be that features are extracted with a relatively low accuracy. A more accurate language model might therefore lead to better predictions.

Connection Between Features and Theoretical Concepts

Our language model although informed by the literature on linguistic features of effective feedback only loosely connects the high level concepts discusses in the literature with operationalized language features. Future research should try to find better models to identify these high level concepts in critiques and further investigate how well these features represent these concepts.

Interactive Feedback

One avenue to explore are systems that structure the feedback task to explicitly improve style more dynamically and more selective. The provided style guide always contained hints on all 5 features. A more advanced system could predict the perceived value of a critique while it is written and then provide stylistic guidelines on only those features that need improvement. For example, if the critique is written with a neutral tone, the system could suggest to the worker to make it clearer whether he or she is criticizing or praising the design.

Personalized Feedback

This paper demonstrated the the features do not linearly correlate with perceived helpfulness. In fact it might be possible that the 'sweet spot" for individual features is different for the person receiving the critique. It is a valuable question if a system can be trained to identify critiques that fit to the personal preferences of a critique receiver. Furthermore such a system could mine existing critiques to provide examples to critique providers that reflect the preferences of the receiver.

CONCLUSION

Designers use online crowds for fast and affordable feedback. However online contributors may lack the motivation, context, and sensitivity to provide high-quality critiques. In this paper we presented two studies. Study 1 demonstrated that our natural language model correlates with with perceived critique helpfulness and that these correlations are stable across populations and different design artifacts. Furthermore, we demonstrated that the model can be used to predict the average rating of critiques.

In a second study we used the language model to mine the critiques collected in the first study for high quality examples that also highlight specific stylistic language features that are correlated with critique ratings. We used the retrieved examples to create a style guide that supports critique providers to self-assess and edit their initial critique. We validated the guide with a between-subjects experiment and found that participants using the guide generated critiques with significantly higher perceived helpfulness compared to the control condition. These findings motivate further investigation into how feedback systems can use natural language models to improve critique quality.

REFERENCES

1. Amazon. Mechanical Turk. (2016). https://www.mturk.com
2. Behance. Behance. (2016). https://www.behance.net/
3. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: A Word Processor with a Crowd Inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*. ACM, New York, NY, USA, 313–322. DOI: http://dx.doi.org/10.1145/1866029.1866078
4. Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. DOI:http://dx.doi.org/10.1097/00004770-200204000-00018
5. Leo Breiman. 2001. Random Forests. *Machine learning* 45, 1 (2001), 5–32.
6. Donald Chinn. 2005. Peer Assessment in the Algorithms Course. In *Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education (ITiCSE '05)*. ACM, New York, NY, USA, 69–73. DOI: http://dx.doi.org/10.1145/1067445.1067468
7. Kwangsu Cho, Christian D. Schunn, and Davida Charney. 2006. Commenting on Writing Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts. *Written Communication* 23, 3 (July 2006), 260–294. DOI: http://dx.doi.org/10.1177/0741088306289261
8. Eric Cook, Stephanie D. Teasley, and Mark S. Ackerman. 2009. Contribution, commercialization & audience. In *Proceedinfs of the ACM 2009 international conference on Supporting group work - GROUP '09*. ACM Press, New York, New York, USA, 41. DOI: http://dx.doi.org/10.1145/1531674.1531681
9. E. M. Cramer and R. D. Bock. 1966. Multivariate Analysis. *Review of Educational Research* 36 (1966), 604–617. http://library.wur.nl/WebQuery/clc/1809603
10. Crowdflower. Crowdflower. (2016). https://crowdflower.com
11. Barbara De La Harpe, J. Fiona Peterson, Noel Frankham, Robert Zehner, Douglas Neale, Elizabeth Musgrave, and Ruth McDermott. 2009. Assessment Focus in Studio: What is Most Prominent in Architecture, Art and Design? *International Journal of Art & Design Education* 28, 1 (Feb. 2009), 37–51. DOI: http://dx.doi.org/10.1111/j.1476-8070.2009.01591.x
12. Steven Dow, Julie Fortuna, Dan Schwartz, Beth Altringer, Daniel Schwartz, and Scott Klemmer. 2011. Prototyping Dynamics: Sharing Multiple Designs Improves Exploration, Group Rapport, and Results. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2807–2816. DOI: http://dx.doi.org/10.1145/1978942.1979359
13. Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A Pilot Study of Using Crowds in the Classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 227–236. DOI: http://dx.doi.org/10.1145/2470654.2470686
14. Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. DOI:http://dx.doi.org/10.1145/2145204.2145355
15. Steven P. Dow, Kate Heddleston, and Scott R. Klemmer. 2009. The Efficacy of Prototyping Under Time Constraints. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition (C&C '09)*. ACM, New York, NY, USA, 165–174. DOI: http://dx.doi.org/10.1145/1640233.1640260

16. Julie S. Downs, Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. Are Your Participants Gaming the System?: Screening Mechanical Turk Workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 2399–2402. DOI: http://dx.doi.org/10.1145/1753326.1753688

17. Edmund Burke Feldman. 1994. *Practical Art Criticism*. Pearson, Englewood Cliffs, N.J.

18. Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

19. Gerhard Fischer, Kumiyo Nakakoji, Jonathan Ostwald, Gerry Stahl, and Tamara Sumner. 1993. Embedding Computer-based Critics in the Contexts of Design. In *Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems (INTERCHI '93)*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 157–164. http://dl.acm.org/citation.cfm?id=164632.164891

20. Michael D. Greenberg, Matthew W. Easterday, and Elizabeth M. Gerber. 2015. Critiki: A Scaffolded Approach to Gathering Design Feedback from Paid Crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition (C&C '15)*. ACM, New York, NY, USA, 235–244. DOI: http://dx.doi.org/10.1145/2757226.2757249

21. M.A. Hearst. 2000. The debate on automated essay grading. *IEEE Intelligent Systems and their Applications* 15, 5 (Sept. 2000), 22–37. DOI: http://dx.doi.org/10.1109/5254.889104

22. Tim Hesterberg, John M. Chambers, and Trevor J. Hastie. 1993. Statistical Models in S. *Technometrics* 35, 2 (may 1993), 227. DOI: http://dx.doi.org/10.2307/1269676

23. Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback : Choosing Review Environment Features that Support High Quality Peer Assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, New York, New York, USA, 458–469. DOI: http://dx.doi.org/10.1145/2858036.2858195

24. J Peter Kincaid, Robert P Fishburne, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* Technical Report. Naval Technical Training Command, Naval Air Station Memphis-Millington, TN, USA. http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED108134

25. Scott R. Klemmer, Björn Hartmann, and Leila Takayama. 2006. How Bodies Matter: Five Themes for Interaction Design. In *Proceedings of the 6th Conference on Designing Interactive Systems (DIS '06)*. ACM, New York, NY, USA, 140–149. DOI: http://dx.doi.org/10.1145/1142405.1142429

26. Markus Krause. 2014. A behavioral biometrics based authentication method for MOOC's that is robust against imitation attempts. In *Proceedings of the first ACM conference on Learning @ scale conference - L@S '14*. ACM Press, Atlanta, GA, USA, 201–202. DOI: http://dx.doi.org/10.1145/2556325.2567881

27. Markus Krause. 2015. A Method to automatically choose Suggestions to Improve Perceived Quality of Peer Reviews based on Linguistic Features. In *HComp'15 Proceedings of the AAAI Conference on Human Computation: Works in Progress and Demonstration Abstracts*. San Diego, CA, USA.

28. Klaus Krippendorff. 1970. Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement* 30, 61 (1970), 61–70.

29. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013a. Peer and Self Assessment in Massive Online Classes. *ACM Trans. Comput.-Hum. Interact.* 20, 6 (Dec. 2013), 33:1–33:31. DOI: http://dx.doi.org/10.1145/2505057

30. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013b. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction* 20, 6 (dec 2013), 1–31. DOI: http://dx.doi.org/10.1145/2505057

31. Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 473–485. DOI: http://dx.doi.org/10.1145/2675133.2675283

32. Jennifer Marlow and Laura Dabbish. 2014. From Rookie to All-star: Professional Development in a Graphic Design Social Networking Site. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 922–933. DOI: http://dx.doi.org/10.1145/2531602.2531651

33. LS Meyers, G Gamst, and AJ Guarino. 2006. *Applied multivariate research: Design and interpretation*. Sage Publishers, Thousand Oaks, CA, USA. http://scholar.google.com/scholar?hl=en

34. Geoff Norman. 2010. Likert scales, levels of measurement and the laws of statistics. *Advances in Health Sciences Education* 15, 5 (dec 2010), 625–632. DOI: http://dx.doi.org/10.1007/s10459-010-9222-y

35. Melissa M. Patchan, Brandi Hawk, Christopher A. Stevens, and Christian D. Schunn. 2013. The effects of skill diversity on commenting and revisions. *Instructional Science* 41, 2 (mar 2013), 381–405. DOI: http://dx.doi.org/10.1007/s11251-012-9236-3

36. Melissa M. Patchan and Christian D. Schunn. 2015. Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instructional Science* 43, 5 (sep 2015), 591–614. DOI: http://dx.doi.org/10.1007/s11251-015-9353-x

37. Chris Piech, J Huang, Zhenghao Chen, C Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM'13)*. Memphis, TN, USA, 153–160.

38. Mary L. Rucker and Stephanie Thomson. 2003. Assessing Student Learning Outcomes: An Investigation of the Relationship among Feedback Measures. *College Student Journal* 37, 3 (Sept. 2003), 400.

39. D. Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, 2 (June 1989), 119–144. DOI: http://dx.doi.org/10.1007/BF00117714

40. Christian Schunn, Amanda Godley, and Sara DeMartino. 2016. The Reliability and Validity of Peer Review of Writing in High School AP English Classes. *Journal of Adolescent & Adult Literacy* 60, 1 (jul 2016), 13–23. DOI: http://dx.doi.org/10.1002/jaal.525

41. Aaron D. Shaw, John J. Horton, and Daniel L. Chen. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 275–284. DOI: http://dx.doi.org/10.1145/1958824.1958865

42. Mark D Shermis and Ben Hamner. 2013. Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. *Handbook of Automated Essay Evaluation: Current Applications and New Directions* (2013), 313–346.

43. David Tinapple, Loren Olson, and John Sadauskas. 2013. CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology* 15, 1 (2013), 29. http://www.ieeetclt.org/issues/january2013/Tinapple.pdf

44. Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the Right Design and the Design Right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 1243–1252. DOI: http://dx.doi.org/10.1145/1124772.1124960

45. UpWork. UpWork. (2016). https://www.upwork.com/

46. Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education* 2 (2003). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.5757

47. Anne Venables and Raymond Summit. 2003. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* 40, 3 (Aug. 2003), 281–290. DOI: http://dx.doi.org/10.1080/1470329032000103816

48. Wenting Xiong and Diane J. Litman. 2011. Understanding Differences in Perceived Peer-Review Helpfulness using Natural Language Processing. In *IUNLPBEA '11 Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Stroudsburg, PA, USA, 10–19. http://dl.acm.org/citation.cfm?id=2043132&picked=prox

49. Anbang Xu and Brian Bailey. 2012. What Do You Think?: A Case Study of Benefit, Expectation, and Interaction in a Large Online Critique Community. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 295–304. DOI: http://dx.doi.org/10.1145/2145204.2145252

50. Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-experts. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1433–1444. DOI: http://dx.doi.org/10.1145/2531602.2531604

51. Anbang Xu, Huaming Rao, Steven P. Dow, and Brian P. Bailey. 2015. A Classroom Study of Using Crowd Feedback in the Iterative Design Process. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 1637–1648. DOI: http://dx.doi.org/10.1145/2675133.2675140

52. Alvin Yuan, Kurt Luther, Markus Krause, Sophie Vennix, Björn Hartmann, and Steven P. Dow. 2016. Almost an Expert: The Effects of Rubrics and Expertise on Perceived Value of Crowdsourced Design Critiques. In *The 19th ACM conference on Computer-Supported Cooperative Work and Social Computing (CSCW'16)*. to appear.

53. ZURB. Forrst. (2015). http://zurb.com/forrst