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

Markus Krause
ICSI, University of California, Berkeley
markus@icsi.berkeley.edu

Tom Garncarz
Carnegie Mellon University
trg@andrew.cmu.edu

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

Elizabeth M. Gerber
Delta Lab
Northwestern University
egerber@northwestern.edu

Brian P. Bailey
University of Illinois
bpbailey@illinois.edu

Steven P. Dow
University of California, San Diego
spdow@ucsd.edu

ABSTRACT

Designers are increasingly leveraging online crowds; yet, online contributors may lack the expertise, context, and sensitivity to provide effective critique. Rubrics help feedback providers but require domain experts to write them and may not generalize across design domains. This paper introduces and tests a novel semi-automated method to support feedback providers by analysing feedback language. In our first study, 52 students from two design courses created design solutions and received feedback from 176 online providers. Instructors, students, and crowd contributors rated the helpfulness of each feedback response. From this data, an algorithm extracted a set of natural language features (e.g., specificity, sentiment etc.) that correlated with the ratings. The features accurately predicted the ratings and remained stable across different raters and design solutions. Based on these features, we produced a critique style guide with feedback examples —automatically selected for each feature —to help providers revise their feedback through self-assessment. In a second study, we tested the validity of the guide through a between-subjects experiment (n=50). Providers wrote feedback on design solutions with or without the guide. Providers generated feedback with higher perceived helpfulness when using our 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.

CHI 2017, May 06 – 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025883

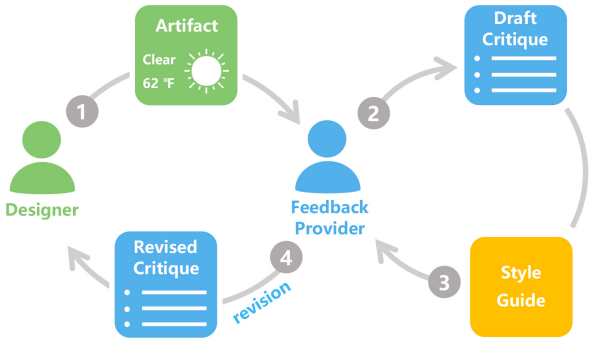

Figure 1. We developed a natural language model to mine the perceived helpfulness of feedback responses along specific linguistic features. This style guide assists feedback providers to improve their draft feedback before sending it to a designer. The style guide includes examples of feedback responses that highlight specific stylistic aspects of good feedback.

INTRODUCTION

Feedback helps designers gather external perspectives to improve their work [53]. Designers traditionally receive feedback on their in-progress designs from peers, mentors, or users who provide comments and suggestions. With the rise of online crowds, including task markets, online social networks, and online communities, feedback can be obtained fast and from a scalable and diverse audience [3, 33, 52]. Furthermore, with a growing demand for design education, instructors are looking to extend traditional methods to scale and personalize the feedback given to students [20, 53].

A challenge of crowdsourcing design feedback is that the results are often of low quality [52, 51]. The reasons for this may include diverse contributors who lack sufficient motivation [51], context [3], knowledge [33], and sensitivity [54] to provide effective feedback. To address these issues, researchers have sought to decompose feedback generation into micro tasks (e.g. [3, 52]) or provide rubrics to direct attention to key aspects of a design (e.g. [14, 33, 20]). However, this prior work requires experts to decide how to decompose feedback generation, or to write rubrics that embed key principles

in a domain. Expert intervention is expensive, and these approaches may become inflexible and may not generalize to different categories of designs or domains.

Our research introduces a domain-independent method to help providers write effective design feedback. We present a natural language model that automatically extracts language features that correspond with ratings of perceived helpfulness. Based on these language features, we compiled a *critique style guide* that offers feedback examples to guide providers (see figure 1). To develop the style guide, in study 1, we collected students' design artifacts from two project-based university design courses and hired online workers (via Amazon Mechanical Turk and Upwork) to provide feedback. The students independently rated the helpfulness of each feedback response. We also collected perceived helpfulness ratings on a sample of the feedback from design instructors (professors from three U.S. universities) as well as from a different set of online workers hired via Mechanical Turk.

To identify the features perceived as most helpful, we conducted a linguistic analysis of the writing style of the collected feedback. We found evidence that that various feedback features including feedback length, emotional content, language specificity, grammatical mood and complexity of sentences, word complexity, and the presence of justifications correlate with higher ratings. Among many alternatives tested, we found that a random forest classifier trained with these features was able to predict the average perceived helpfulness with Krippendorffs alpha [30] levels close to or higher than the inter-rater reliability of the instructors.

A second study was conducted to test how the guide produced from our first study affects a providers' ability to write helpful feedback. Using a between-subjects design, we randomly split 90 online feedback providers between a guided and control condition. Providers wrote feedback for the same design artifacts from study 1 and were asked to revise their feedback with (guided) or without (control) our style guide. The style guide consisted of examples of feedback selected semi-automatically from feedback collected in the first study. The control group received only general instructions. We found that providers in the guided condition improved their ratings of feedback helpfulness and their average correlation with our natural language model significantly more compared to the control condition.

This paper makes the following contributions.

1. Describes a set of natural language features that correlate with perceived feedback helpfulness. **(Study 1)**
2. Illustrates that these correlations are stable across two different design tasks and three different rater populations. **(Study 1)**
3. Demonstrates that these features can predict the perceived helpfulness of the feedback. **(Study 1)**
4. Provides evidence that feedback providers can improve the perceived helpfulness of their feedback when using our NLP-informed style guide. **(Study 2)**

RELATED WORK

In design work, feedback helps designers iterate on their in-progress solutions [15], compare alternatives [12, 45], learn about the problem, and refine their process. Feedback can also help novices better understand design principles [17].

Measuring Feedback Quality

Measuring feedback quality is challenging and prior work uses a range of measurements. Luther et al. compares differences between design iterations [33, 53], while others contrast critiques with feedback produced by experts [33, 31]. Measuring post-feedback design quality [12] and collecting designer ratings on the helpfulness of the feedback [7] are other viable methods to measure feedback quality.

Various definitions exist that describe qualities of effective feedback. Sadler [40] argues that effective feedback is able to establish an understanding of the underlying patterns and concepts (conceptual), can communicate the difference of the work to this standard (specific), and lays out steps that reduces this gap (actionable).

Rucker et al. hypothesised that perceived helpfulness captures the value of feedback for its recipient [39]. 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 contains positive or encouraging remarks. In our work, we also use perceived helpfulness as the measure of feedback quality.

Xiong and Litman [50] investigated how natural language features correlate with perceived helpfulness. Their work on peer feedback for history papers found that lexical features regarding transitions and opinions can predict helpfulness. We apply a similar approach, but we use a different and more generalizable set of features and show how these features can be used to select examples that aid the composition of feedback.

Sources of Feedback

Designers gather feedback from various sources. In educational settings, instructors provide feedback by commenting on and grading assignments. This type of feedback provision has been used successfully in design [11, 44, 31] and other contexts such as computer programming [6] and writing [48]. A key limitation is that instructor-led feedback struggles to keep pace with the increasing scale of design education.

Self-assessment can achieve results comparable to external sources of feedback [14], but may not always yield the same insights possible from a diverse external audience. Scaling feedback in educational settings often involves peer review [37]. The benefit of peer review is that students learn from both providing and receiving feedback [38]. Critiquing the work of peers helps students to practice their revision skills and strengthen their ability to find and solve problems [38]. Despite the positive effects, there is skepticism as to whether students of all ability levels are capable of helping their peers [37]. Our approach of producing a style guide for design feedback could also be applied to improve the quality of peer review across domains.

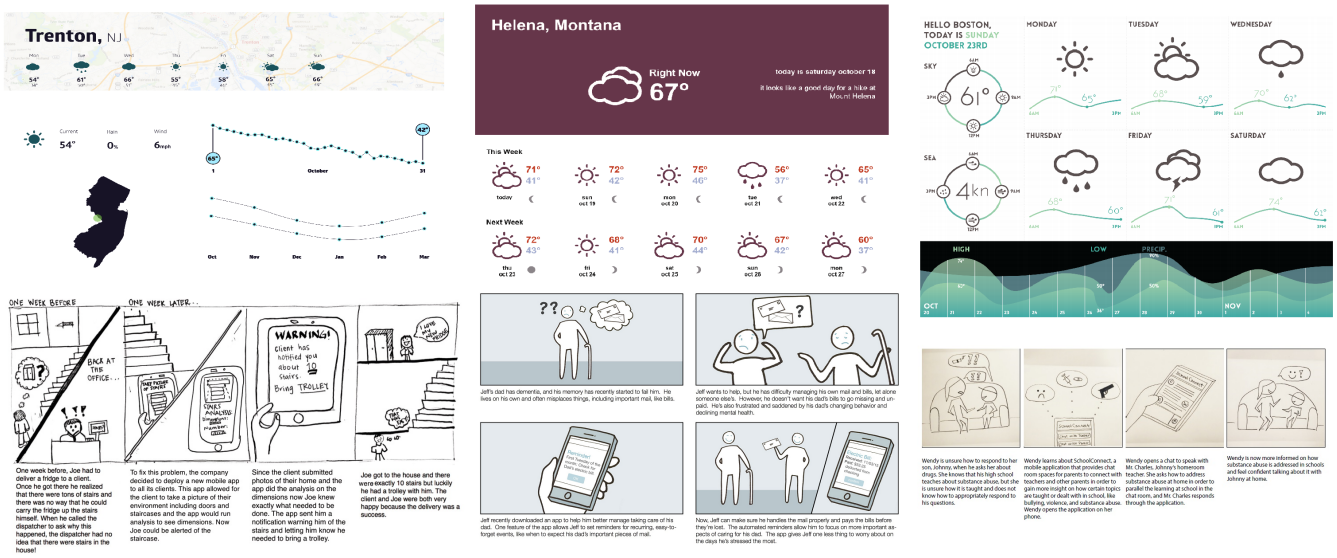

Figure 2. The top three designs were submitted by students working individually for a course assignment which involved creating a user interface for a weather dashboard. The lower three examples are storyboards designed by students working in teams to create mobile apps in a service design course.

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 [46]. Crowd feedback is particularly appealing due to its scalability and availability, and the ability to acquire diverse perspectives not possible in an isolated classroom [13]. Involvement in communities such as Behance [2], Forrst [55], and Dribbble [34] are ways for experienced designers to exchange design feedback. These platforms require a certain level of commitment and expertise to fully experience their potential, limiting their accessibility [8]. Participants tend to be motivated to develop their own skills and status [51]. Novices in such communities may experience evaluation apprehension and be hesitant to share preliminary work [34].

Improving Crowdsourced Feedback

Crowd-based methods can yield feedback that is of low quality or quantity [51] because contributors may lack the motivation [51], context [3], knowledge [33], or sensitivity [54] needed to provide high-quality design feedback. Prior work has explored screening processes to disqualify workers that regularly provide low quality responses [16] as well as other mechanisms such as the Bayesian Truth Serum [42] to increase work quality.

CrowdCrit [33] and Voyant [52] structure design feedback tasks for online crowds. Both systems are motivated by the goal of producing higher quality feedback from non-experts. Recent studies compare the *characteristics* of feedback produced by these structured approaches against both open-ended feedback and expert feedback with promising results [33, 53, 20].

Automated Feedback

Automated feedback has been applied in various contexts such as essay grading [21, 47, 41, 7] and kitchen design [19]. The scalability of these approaches make them particularly attractive. However, it is difficult to automate in-depth assessment of design artifacts because they are often subject to diverse interpretations. Our approach overcomes this challenge by providing a style guide that can help providers, especially non-experts, write more effective feedback. Computational language models are able to analyse written feedback [43, 24] since their structure is similar to an essay [26].

RESEARCH QUESTIONS AND HYPOTHESES

This paper explores how style-based natural language features correlate with perceived helpfulness of design feedback and evaluates whether a style guide derived from these features improves the quality of the design feedback provided. Specifically, our work investigates the following research questions:

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

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

RQ3: How can a style guide informed by these natural language features improve perceived helpfulness?

We argue that valuable feedback incorporates the qualities suggested by Sadler [40], Cho et al. [7], Yuan et al. [54] and Krause [26, 27, 28, 29]. That is, valuable feedback 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 encourages the recipient. Prior work used human annotators to extract language features. This paper develops an algorithmic approach that offers better scalability and domain independence.

Although some features of good feedback are known, it can be difficult to find good examples for each feature especially

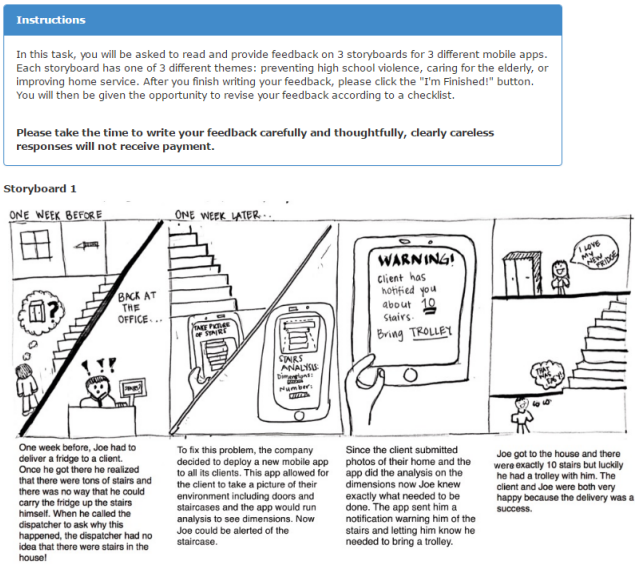

Figure 3. The user interface used to write design feedback. Contributors entered their feedback in a text field positioned below the design artifact (not shown). Only one artifact was visible at a time.

in an automated and scalable way. Yet good examples are necessary to teach a feedback provider how to write good feedback [31]. Our approach enables mining existing sets of feedback to find specific responses that highlight linguistic features associated with high quality. We use our language model to semi-automatically generate a style guide that we hypothesize will significantly increase the frequency of these features and enhance the perceived helpfulness ratings of the feedback.

STUDY 1: PREDICTING PERCEIVED HELPFULNESS

We collected design artifacts from students in two university design courses. Online contributors recruited on Mechanical Turk (MTurk) provided feedback on these design artifacts. Students, instructors, and a different set of online contributors then rated the feedback. We then created a natural language model to extract feature vectors from the collected feedback. We estimated the correlation between the features and perceived helpfulness and predicted the average perceived helpfulness with our model.

Procedure

We collected design artifacts from two non-overlapping student populations in two independent design courses. In one course, students created storyboard artifacts for a team assignment focused on mobile phone applications. In the second course, students individually designed a dashboard to display weather forecasts.

Artifact Collection

In the first classroom study, we recruited 15 students from an undergraduate design course. Each student submitted one design from a course assignment which involved creating a user interface for a weather dashboard. Figure 2 shows three examples of the designs. To generate feedback, we recruited

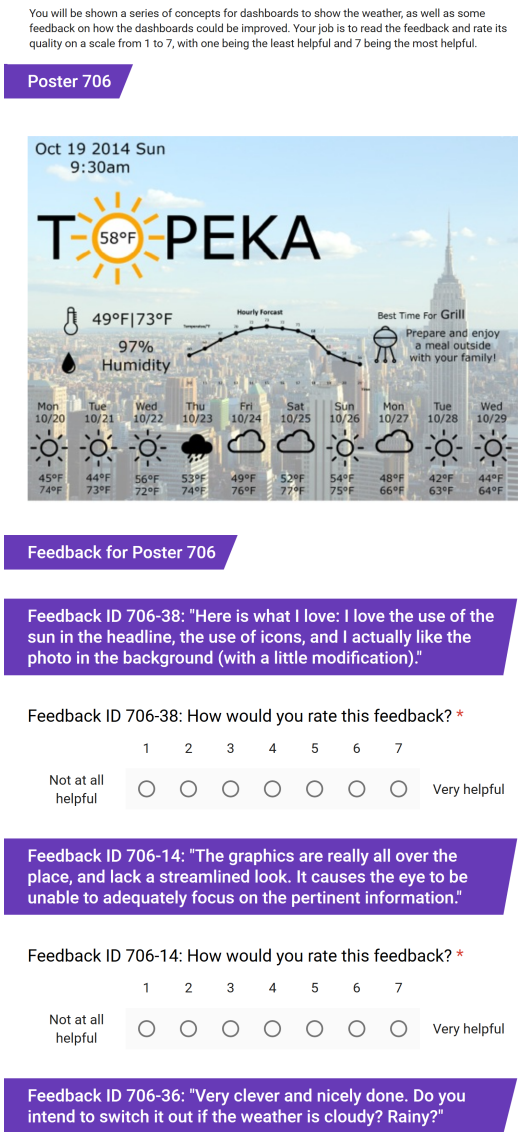

Figure 4. Raters saw each design artifact and its associated feedback on one page. They rated each feedback response on a scale from 1 to 7 (very helpful). The ratings could be revised if desired.

24 feedback providers via MTurk and collected 615 pieces of feedback.

In our second classroom study, 37 undergraduate students worked in groups (up to four) to create storyboards for mobile apps in the domains of service design, high school violence prevention, and elder care. Figure 2 shows three examples. Independent feedback providers (N=71) recruited from MTurk wrote feedback for the storyboards. A feedback provider wrote feedback for 3 storyboards of a student group and worked on storyboards of up to 5 groups. Each design received at least three pieces of feedback. We collected 568 pieces of feedback in total.

Figure 3 shows the user interface that was used to collect the design feedback from online contributors. Providers were presented with three design artifacts, one at a time, and asked

to write feedback for each one. To normalize the population's language skill, the task on MTurk was configured to only accept workers based in the U.S. We priced tasks so that contributors earned the equivalent of $10 an hour.

Feedback Ratings from Students and Instructors

After all of the feedback was collected, the student designers rated the helpfulness of each piece of feedback they received. Students were shown one response at a time in random order, and rated its helpfulness on a scale from 1 (Not helpful) to 7 (Very helpful). Students rated only the feedback that was given to their group's designs or their individual work. This means we collected multiple ratings per feedback response for the storyboard (group) assignment and only one rating per response for the dashboard (individual) assignment.

A subset of these feedback responses from both courses were also rated by three design instructors. Only one of the three instructors rated the feedback responses from both courses. Instructors rated 141 pieces of feedback for the storyboard designs and 60 pieces for the dashboard designs, following the same procedure as the students. Some instructors rated more responses than others. Instructors took one hour to complete the rating task. Figure 4 shows the user interface used to rate the feedback.

Feedback Ratings from Online Contributors

To ensure a diverse set of ratings, we conducted a survey on MTurk (n=60) to identify workers with and without significant experience in writing and receiving design feedback. 14 out of these 60 workers had at least one year of design education. The average years of education among participants (n=60) was 0.9 years (SD: 2.4). From this pool, we selected twelve workers (six for each course) to rate the feedback.

We selected three experienced and three less experienced workers for each course based on the self-reported experience in design education and professional design work. We considered a worker to be experienced if they had at least five years of professional design experience or taught design for at least five years. These participants used the same procedure and rated the same set of feedback as the instructors. The task took participants approximately one hour to complete, and they were compensated $10 for their participation.

Measures

In our study, we consider one dependent variable, two independent variables and eight covariates. Our dependent variable is the subjective rating of helpfulness. We asked all populations to rate feedback helpfulness on a scale from 1 to 7 (**helpfulness**). There were two independent variables in the study, each with two and three levels, respectively. We encode if a rating was obtained in one of the two courses by the variable **experiment** with the levels storyboard and dashboard. The second independent variable **population** encodes if a rating was given by an online contributor, a student, or an instructor.

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 feedback provider

Feature	High	Low
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 model" section of Study 1. The middle column (High) shows an example of a feedback response ranked above the 85th percentile for the given feature. The right column (Low) gives an example taken from the collected data that is ranked below the 25th percentile for the given feature. The bold text below each example gives the average rating of the feedback.

corresponds to one observation, resulting in 176 individual observations.

Language Model

Our natural language model extracts eight feature categories (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. We preprocessed all of the feedback 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 feedback **length** operationalized as the number of characters. We counted every alphanumeric character including punctuation and special characters but not spaces. The average length of a piece of feedback was M = 123 (SD = 128).

The second feature, text **complexity**, is operationalized as the automated readability index (ARI [25]). The index is calculated from the number of characters, words, and sentences. The higher the value the more complex the text. Another 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 initial 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 by counting words that indicate a justification (e.g. because).

We expect the extent to which a piece of feedback is **actionable** to be predictive of perceived helpfulness as argued by Sadler [40]. We operationalize this feature with the grammatical mood of the sentences. Each sentence was classified as either indicative (written as if stating a fact), imperative (expressing a command or suggestion), or subjunctive (explor-

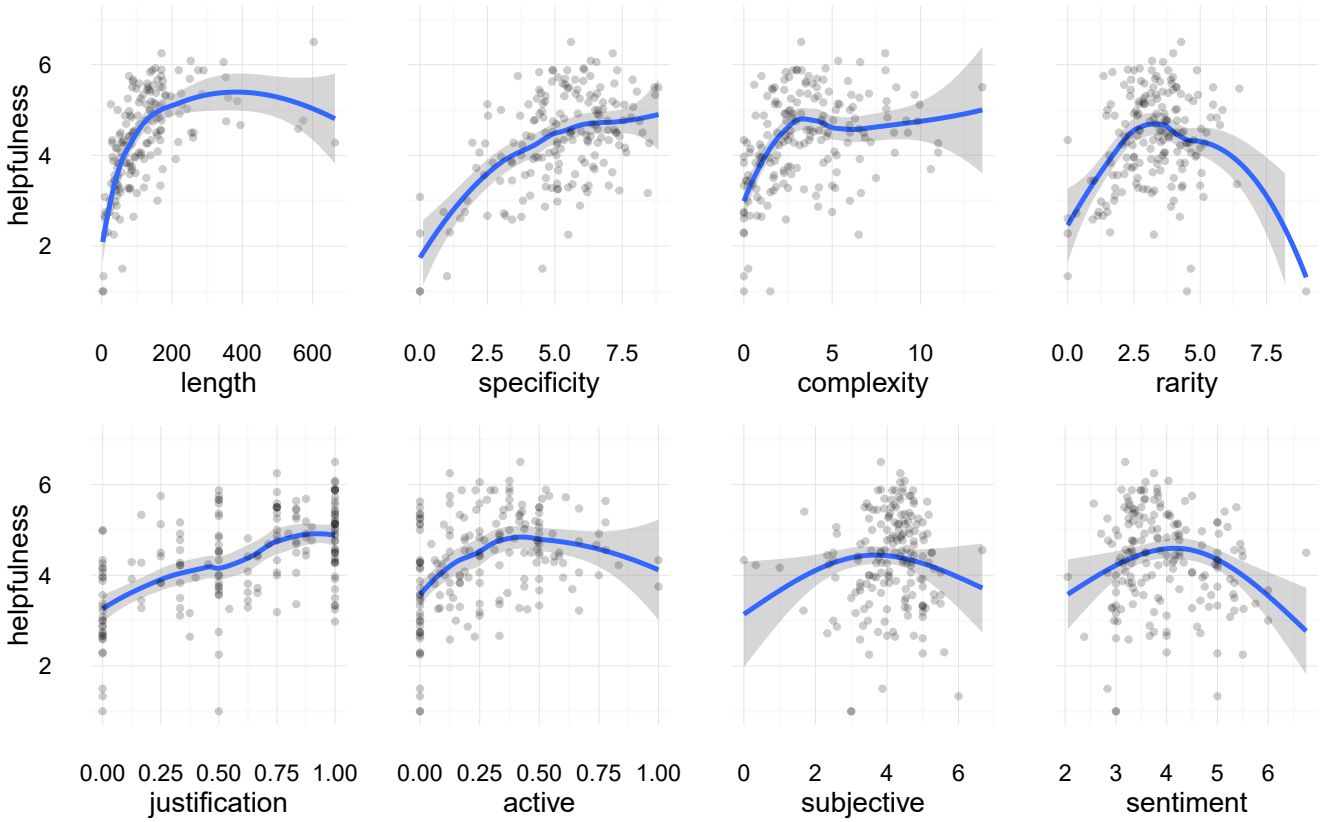

Figure 5. Correlation between the eight observed language features and the perceived helpfulness across all populations (instructors, students, and online contributors). Each point is the aggregated average for one feedback provider on the given feature. The blue line is calculated using local polynomial regression fitting [22]. The correlations for these surfaces can be found in Table 2. The figure shows the relation between a feature and perceived helpfulness. The length of a feedback for instance is predictive for helpfulness, yet plateaus at around 250 characters. Other features such as sentiment have a peak and the perceived helpfulness increases before this peak but declines afterwards.

ing hypothetical situations). The feature, which we refer to as actionable, corresponds to the ratio of non-indicative and indicative sentences in a piece of feedback. Non-indicative sentences are either imperative, conditional, or subjunctive. Values fall between 0 and 1 (all sentences are non-indicative or active). We used pattern.en to extract the sentence mood. As all feedback must be interpreted, we also include features to measure sentiment and subjectivity. For both features we used classifiers provided by the pattern.en toolkit. The values for these features fall in the 0 (low) and 10 (high) range.

Next, we looked at the **sentiment** feature. Yuan et al. [54] found that a positive sentiment is predictive for perceived helpfulness. Sentiment refers to whether a feedback response is positive or negative. A value of 0 is strongly negative and 9 strongly positive. Subjectivity refers to whether the feedback uses 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 set of feedback can be found in Table1.

Results

The first study investigates the correlations between our language model and perceived helpfulness and demonstrates the prediction quality of our language model.

Features Correlate Non-linearly with Helpfulness

As figure 5 illustrates, the observed features do not linearly correlate 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 Norman [36], we choose Pearson correlation. Alternative methods such as Spearman correlation yield inaccurate p-values with ties. Due 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 two sets of design artifacts as well as within all of 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.

The Language Model can Predict Perceived Helpfulness

As previous research has indicated, natural language models can predict essay grades with high accuracy, sometimes even outperforming human raters [43]. We were interested in whether 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 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. The reason for this is that some raters have a strong bias for some features; our algorithm is able to correct for this and predict the most likely average.

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 three columns give the correlations for the online contributors, 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 feedback rating given by each population of raters. 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 piece of feedback will receive. The column Pred. Avg. gives the Krippenforff's alpha between the prediction of the average and the observed average rating of the feedback. Rows split the results based on rater populations. High and low columns give the lower and upper bounds of the 95% CI.

STUDY 2: CRITIQUE STYLE GUIDE INTERVENTION

The second study investigates the effect of the language model on the perceived helpfulness of feedback. We conducted a between-subjects study with two conditions (guided and control). Participants in the guided condition received a critique style guide to revise their initial feedback while participants in the control condition received only general instructions to improve their work. We analyzed the style guide, the language features, and the perceived helpfulness ratings.

Critique Style Guide

The style guide provides five comments and examples of highly rated feedback drawn from the first study (see Table 4 for an overview). We selected feedback examples that scored high according to our language model and were rated high by the human raters. We excluded some features that we felt

Comment	Feedback Example
On average, highly-ranked feedback statements have 50 words. Please make sure that your feedback is not too short.	We did not provide a specific example for long feedback.
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 feedback 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 feedback providers and the second column gives the automatically retrieved feedback examples. We mined the feedback set from the previous experiment using our language model to find highly rated examples that also highlight a specific feature.

would be hard to explain to feedback providers, including grammatical complexity and word rarity. The user interface for writing feedback is similar to the user interface for Study 1 and can be seen in Figure 6.

Procedure

From the first study, we selected a sample of design artifacts to be critiqued again. The three selected designs are shown in the bottom row of Figure 2. We selected one artifact from each of the three domains; home service, high school violence prevention, and elder care.

We recruited 90 feedback providers from MTurk. Recruitment was restricted to the U.S. to reduce language bias. Half (45) of the providers worked in the guided condition and the other half (45) in the control condition. Each provider wrote feedback for each storyboard, and were then instructed to revise their feedback. Providers in the guided condition revised their feedback using the style guide as shown in Figure 6.

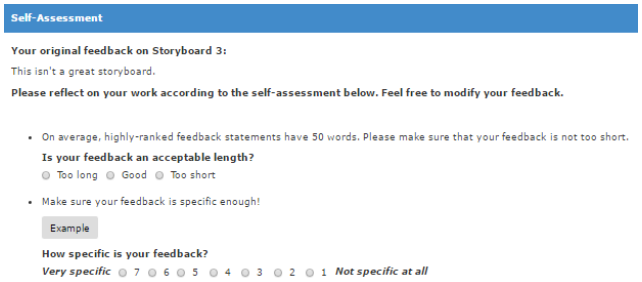

Figure 6. The user interface of the critique style guide. Feedback providers were asked to self reflect on how well their feedback follows the guidelines on a scale from 1 to 7 (closest). The figure shows only two elements of the style guide. The full list of examples can be found in Table 4.

Providers in the control condition did not use the style guide but were asked to revise and improve their feedback. All participants received a bonus of $1.00 if their revision improved the feedback. This bonus was paid regardless of quality after the task. We also asked all providers to answer two questions 1) the editing process helped to improve my work and 2) I liked the editing process. All questions asked were measured on a Likert scale from 1 (do not agree) to 7 (strongly agree). Additionally, we asked an open ended question on how the editing process affected their feedback.

To rate the collected feedback and estimate the improvement, we recruited twenty raters from MTurk. Each rater rated 75 pieces of feedback following the same procedure as described in Study 1 (see Figure 4 for reference). The feedback was ordered randomly and each rater rated each piece of feedback. For this study, we only collected ratings from online contributors.

Measures

The study investigates two independent variables, three dependent variables, and the covariates derived from the natural language model. The manipulated independent variables in this study are **edited** and **condition**. The edited variable has two levels. All initial feedback responses are labeled **before** while the revised responses are labeled **after**. The condition variable had two levels: **guided** if the style guide was used to revise the feedback and **control** if the feedback was revised without the style guide. The main dependent variable is again perceived **helpfulness**. We also measured the **helpfulness** of the intervention and how much the feedback provider **liked** the back feedback process. All variables were measured on a 7-point scale. As in the previous experiment, we extracted eight natural language features from the collected feedback following the same process as in Study 1.

Results

The second study tests whether the automatically extracted feedback responses can be used as examples in a style guide and lead to feedback with improved perceived helpfulness.

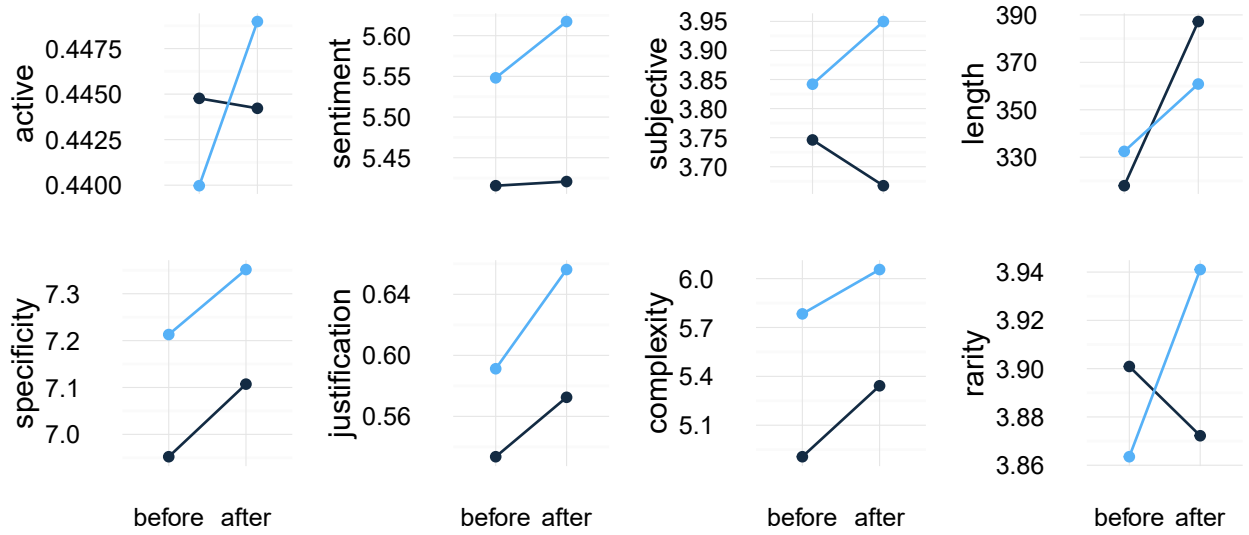

Figure 7. The presence of the language features before and after the revision round. The features increase more when providers use the style guide. As the features are not linearly correlated with helpfulness an increase by itself is not necessarily desirable. Yet most contributors in both conditions where on the lower end of most features. Light blue line = style guide condition; black line = control condition.

Feature Presence Increases with Style Guide Use

To estimate if the presence of features in our model increases significantly more when feedback providers 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. [35]. The MANOVA showed a significant multivariate interaction between condition and editing $F(7, 183)=3.413$, $p=0.04$. Figure 7 shows the changes for each feature.

The Guided Intervention is Perceived More Helpful

We asked feedback providers in both conditions how they perceived the intervention. We found that the feedback provider 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 ($M = 4.95$, $SD = 0.91$). The providers also commented that they 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...*
>
> **control**

The Guided Intervention Improves Feedback More

The final question of this study is whether using the style guide improves the perceived helpfulness of feedback more than the control condition. We analyzed the results using a two-way ANOVA and found a significant interaction between the two variables condition and edited ($F=(3,596)=4.09$, $p=0.01$). The increase in perceived feedback 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 both studies.

RQ 1: Which stylistic natural language features do correlate with perceived helpfulness of design feedback

We found that all features discussed in this paper do correlate significantly with perceived helpfulness. It is however important that these correlations are not linear and have to be investigated with advanced statistical models. We observe two distinct relations. We found that three features – text length, grammatical complexity, and specificity – reach a plateau. A second group of features – activeness, word rarity, subjectivity, and sentiment – follows an inverted U-shape. Where the positive influence of a feature reaches a peak and decreases afterwards. Finally the justification feature is linearly correlated but limited between 0 and 1.

Some correlations are also relatively weak. One reason for lower levels of predictive power in some features is the accuracy of the feature extraction method used. Yet the main challenge in predicting helpfulness is the high variance between and within rater populations. Nonetheless our language model is able to predict the average for individual populations with Krippendorffs alpha values close to the inter-rater 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 population

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 low. This might indicate a personal as well as a task-specific component to the importance and shape of individual features. Instructors strongly correlated 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 helpfulness of feedback

Our model is able to predict and extract high quality feedback that highlights specific stylistic features. A style guide can incorporate the selected feedback as examples to help providers reflect on their work and write feedback that is perceived to be more helpful by a designer.

LIMITATIONS AND FUTURE WORK

This work has illustrated that a style guide that uses automatically retrieved examples can provide support for feedback providers. The presented results are promising and give way for many future investigations.

Extending the Language Model

This study analyzed the relationship of eight natural language features with perceived helpfulness and used five of these features in the 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 such as automated essay grading [28], stylometric authorship prediction and plagiarism detection [26], and predicting sales e-mail responses [32].

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 the literature suggests a large impact on feedback quality (e.g. sentiment [49]). 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 only loosely maps the high-level concepts of "good feedback" discussed in the literature to operationalized language features. Future research should try to find better models to identify these high-level concepts in design feedback and further investigate how well these features represent these concepts.

Interactive Feedback

Future work should also explore systems that structure the feedback task to improve style more dynamically and more selectively. Our style guide contained hints on all five features. A more advanced system could predict the perceived value of a piece of feedback while it is written and then provide stylistic guidelines on only those features that need improvement. For example, if the feedback is written with a neutral tone, the system could suggest to the provider to make it clearer whether he or she is criticizing or praising the design.

Personalized Feedback

This paper demonstrated that features do not linearly correlate with perceived helpfulness. In fact it might be possible that the "sweet spot" for individual features is different from person-to-person. Systems could be trained to identify feedback responses that fit the personal preferences of the designer. Furthermore such a system could mine existing collections of feedback to provide examples to providers that reflect the preferences of the designer. Although we manually chose examples retrieved using our language model, our results illustrate the feasibility to automatically select high-quality examples. A future system can use a clustering algorithm to find groups of similar feedback. Experts can provide guidelines for how to improve low-quality samples in each cluster. The system could compare new feedback to these clusters and return an expert guideline that explains how to improve a feedback similar to the new feedback.

CONCLUSION

Designers depend on online crowds for fast and affordable feedback. However online contributors may lack the expertise, context, and sensitivity to provide high-quality feedback. In this paper, we presented two studies. Study 1 demonstrates that our natural language model correlates with perceived feedback 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 helpfulness rating of feedback.

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

ACKNOWLEDGEMENTS

Financial support was provided by the German Academic Exchange Service through the FIT Worldwide program and the National Science Foundation under awards 1530837, 1530818, 1122320, 1122206, and 1217225. Special thanks go to the instructors, students, and online contributors who participated in our study.

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. Niklas Kilian, Markus Krause, Nina Runge, and Jan Smeddinck. 2012. Predicting Crowd-based Translation Quality with Language-independent Feature Vectors. In *HComp'12 Proceedings of the AAAI Workshop on Human Computation*. AAAI Press, Toronto, ON, Canada, 114–115. http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/viewPDFInterstitial/5237/5611

25. 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

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. 2015a. 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. Markus Krause. 2015b. Bull-O-Meter : Predicting the Quality of Natural Language Responses. In *HComp'15 Proceedings of the AAAI Conference on Human Computation: Works in Progress and Demonstration Abstracts*. San Diego, CA, USA.

29. Markus Krause. 2015c. Stylometry-based Fraud and Plagiarism Detection for Learning at Scale. In *Proceeding of the KSS Workshop'15*. Karlsruhe, Germany.

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

31. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. 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

32. Chinmay E. Kulkarni, Michael S. Bernstein, and Scott R. Klemmer. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15*. ACM Press, New York, New York, USA, 75–84. DOI: http://dx.doi.org/10.1145/2724660.2724670

33. 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

34. 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

35. 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

36. 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

37. 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

38. 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

39. 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.

40. 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

41. 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

42. 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
43. 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.
44. 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
45. 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
46. UpWork. UpWork. (2016). https://www.upwork.com/
47. 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
48. 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
49. Wenting Xiong, Diane Litman, and Christian D Schunn. 2012. Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research* 4, 2 (2012), 155–176.
50. 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
51. 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
52. 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
53. 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
54. 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.
55. ZURB. Forrst. (2015). http://zurb.com/forrst