The UX Factor: Using Comparative Peer Review to Evaluate Designs through User Preferences

SARAH BAWABE*, Brown University, USA
LAURA WILSON*, Brown University, USA
TONGYU ZHOU*, Brown University, USA
EZRA MARKS, Brown University, USA
JEFF HUANG, Brown University, USA

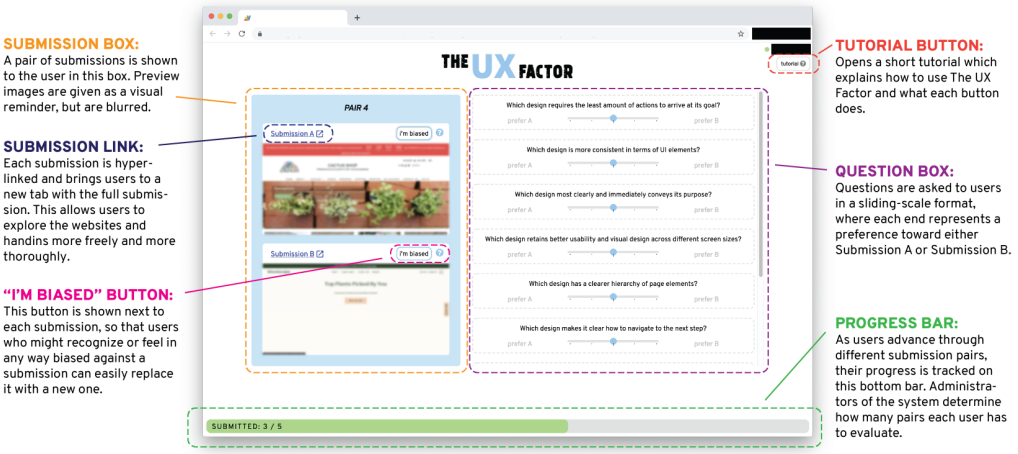

Fig. 1. Screenshot of a sample pair being shown in the UX Factor application, along with the comparison questions that were asked of these two submissions.

Peer review has been used in both online and offline classrooms to inspire creativity, gather feedback, and lessen instructor grading loads, especially for design-based tasks without definitive rubrics. To explore the nuances and quality of peer feedback, we developed UX Factor, a peer grading platform that aims to characterize the behavior of peer reviews and the consistency of the ranking models used to aggregate these reviews. This system harnesses the power of pairwise comparisons to minimize bias and encourage context-driven analysis. We adopted UX Factor in a user interface course of 133 students and teaching assistants (TAs) across 3 different individual design projects over a semester and found that the system was effective in eliciting high-quality

*These authors contributed equally to the research.

Authors' addresses: Sarah Bawabe, Brown University, Providence, Rhode Island, USA; Laura Wilson, Brown University, Providence, Rhode Island, USA; Tongyu Zhou, Brown University, Providence, Rhode Island, USA; Ezra Marks, Brown University, Providence, Rhode Island, USA; Jeff Huang, Brown University, Providence, Rhode Island, USA.

2573-0142/2021/10-ART476 $15.00
https://doi.org/10.1145/3479863

feedback. We saw that raters have higher agreement than random preferences, and with at least 15 ratings per submission, a simple average of ratings produced rankings that were consistent to both the raw ratings and other more complex models. These rankings were robust to disagreeable raters and changing class sizes, demonstrating the potential of comparative peer review to match the quality of expert feedback at scale.

CCS Concepts: • **Human-centered computing** → **Open source software**; • **Social and professional topics** → **Student assessment**.

Additional Key Words and Phrases: peer assessment; comparative peer review; design critique

ACM Reference Format:
Sarah Bawabe, Laura Wilson, Tongyu Zhou, Ezra Marks, and Jeff Huang. 2021. The UX Factor: Using Comparative Peer Review to Evaluate Designs through User Preferences. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 476 (October 2021), 23 pages. https://doi.org/10.1145/3479863

1 INTRODUCTION

In physical and digital classrooms alike, peer evaluation has provided students with invaluable feedback on both individual and group-based assignments [12] from introductory to advanced courses [11]. In comparison to traditional feedback methods, it can greatly improve learning [4, 28] by enhancing motivation and agency [9, 32] and incite iterative improvement [18]. In fields such as art and design without a notion of "correctness," peer review from other students who worked on the same assignment and followed similar thought processes can reveal additional insights that are not immediately obvious to an instructor.

While many peer review frameworks operate on a *cardinal* system, where students rate the submissions on an absolute scale, these evaluations are seldom made in a vacuum. That is, human judgment is comparative by nature [25], so when a student makes the assessment that "Submission A is 4/5," they may mean different things depending on their personal rubrics. For example, if they consider "3" as the average, then they are saying Submission A is better than their perceived average; conversely, if their perceived average is a "4," then they are judging the submission to be about the same as their average. Left alone, this definition of average may differ significantly across individuals, leading to discrepancies in evaluation which can inflate the inherent subjectivity of ratings [38] and thus decrease the overall integrity of peer reviews. The alternative *ordinal*, or comparative, approach, although still subjective, does not suffer from the same ambiguities and was demonstrated to be more robust to differences in student skill levels [36] and comparable to TA feedback in terms of accuracy [33]. By viewing different design solutions to the same problem at once, raters can evaluate submissions more critically and better identify issues in the designs [38] by focusing on gaps. Using paired submissions, the comparative ratings can attribute anchors to subjectivity and thus allow for more *objective assessments* of the subjective ratings in comparison to absolute ratings.

We introduce UX Factor, an Apps Scripts based comparative peer review platform for design work that randomly assigns pairs of anonymous submissions to each submitter after the submission period. The submitter then specifies which of the two submissions they prefer within each pair along a number of preset rubric questions. Once the review period is over, UX Factor uses a ranking model that takes the pairwise judgments and produces an ordered ranking of the submissions.

UX Factor was deployed in a user interface and user experience (UI/UX) class of 120 students. In this paper, we explore both the practical experience of this deployment, as well as quantitative measures of consistency: whether students can provide reliable preferences and the robustness of four different ranking models across different classroom sizes and settings. In practice, we also took the resulting ranking and anchored them to project grades based on the instructor's judgment,

but the focus in this research will be the relative ordering of the submissions that were produced, rather than the actual grades.

In this paper, we make the following contributions: (1) an analysis of the reliability of student raters and strategies to maintain rating consistency through ranking model choices and (2) the open-source comparative peer feedback system and design considerations that can be used to elicit feedback via pairwise preferences. These contributions together can guide a peer review experience that can be effective as educational tools and in reducing instructor effort for larger courses.

2 RELATED WORK

2.1 Systems for Comparative Peer Review

While cardinal grading platforms like Calibrated Peer Review (CPR) [7], Organic Peer Assessment [21], and EduPCR4 [44] demonstrated success in providing good feedback, comparison-based evaluations have been shown to promote deeper thinking [2]. Recent systems have thus focused on comparative judgement [10, 20], where raters compare two of their peers' answers side-by-side and select an overall "winner," to elicit overall higher quality feedback, albeit with different embellishments. ComPAIR [32] adds adaptivity by dynamically generating increasingly similar submission pairs, following principles of Adaptive Comparative Judgement (ACJ) [31], and found that this framework contributed to greater learning in a pilot study with English, physics, and math courses. PeerStudio [23] employs comparative review in open-ended writing assignments, and found that rapid early peer feedback improves the quality of the final submission. Like ComPAIR, Juxtapeer [6] also presents two submission pairs side-by-side to elicit the best feedback, but additionally scaffolds comparison to one submission so that raters see two submissions but only grade one each time. Studies across music, typography, and teaching courses for Juxtapeer revealed that contrasting cases are especially effective for visual submissions, as raters can better identify details they otherwise would have overlooked. While UX Factor similarly employs a comparative framework like PeerStudio and Juxtapeer, it specifically focuses on the review of design–something entirely visual. While existing systems employ binary questions and short response comments, we also provide a more comprehensive framework containing Likert scale, checklist, and slider-based rubric questions, supplied by the instructor, to allow the rater to convey more nuanced, although still guided, judgments of visual design. In comparison to these previous systems, UX Factor is also more accessible; the platform is open source and educators can manipulate rubrics and assignments within the familiarity of Google Forms.

Existing systems handle the two major challenges of peer review platforms, namely reviewer credibility and validity [43], in different ways: each student may (1) undergo a "calibration period" where they receive active feedback for their evaluations [7, 37], (2) be presented with example exemplar submissions [23], or (3) receive gradual training through structured comparisons [6]. However, (1) and (2) are time-consuming and require significant additional work from the instructor for each new project, while (3) requires double the amount of total comparisons since only one submission is evaluated at each iteration while the other is used for scaffolding. UX Factor adopts a different approach to galvanize raters; after all ratings are submitted, it additionally assigns each rater a score informing them of how close their own ratings were to the final ranking a submission received. Oftentimes, however, instructor and TA grading are still considered the gold standard for subjective feedback; peer grades are often compared against these "expert" grades to assess accuracy [6, 27, 33, 41]. Nevertheless, even experienced instructors have expressed that grading effectively remains difficult, especially in fields such as design without objective grading criteria [1]. Complete reliance on expert subjective feedback can thus become dangerous, as disparities between popular appeal and expert critique of art indicate differences in criteria that should be

further investigated [16]. Studies comparing novice and expert subjective feedback revealed that their judgments can converge if both were previously involved in the discussion of criteria [9, 11]. In spirit of and to synthesize the critiques of both parties, UX Factor encourages TAs and instructors to create rubrics that are "peer reviewed" by the students. Before submitting peer evaluations, students can view and directly provide feedback on the rubric itself. Shifting the responsibility of rubric determination away from sole experts in this way can afford raters more autonomy in their own learning process.

2.2 Impacts on Learning and Student Satisfaction

When peer review systems are deployed in the wild, their effects on student satisfaction and learning have been mixed. In one instance of a digital humanities massive open online course (MOOC) where students were tasked with peer grading a series of blog posts, transparency in the grading criteria, knowledge that their work will be seen by peers, and actual exposure to peer work led to an increase in submission quality [18]. Conversely, another MOOC that involved the peer grading of student essays saw that exposure to excellent submissions discouraged students and incited them to drop the course [34]. In attempts to mitigate these divergent individual responses when encountering peer submissions, UX Factor provides the option of an "opt-in share club." In this framework, students can choose whether they are notified of other submissions that were determined to be better than theirs in the final feedback report.

Previous peer grading platforms also identified bias and anti-reciprocity in the way students graded others. Even after rubric-guided training, students in a middle-school science class still awarded lower grades to higher performing peers [35]. Similarly, a study with PeerStudio revealed that over time, students who received better reviews on their work write worse reviews in the future [22]. These observations suggest a gradual diffusion of responsibility [39] after repeatedly viewing high quality reviews; the student may believe their own high quality reviews are less necessary. Failure to appropriately address these issues may lead to gradual decrease in overall feedback quality and consequently loss of trust in the system [19]. One approach to handle grader bias and reliability is to account for them quantitatively in the computational model [30]. However, algorithmic solutions require some notion of "ground truth," typically provided by instructors. Although some studies argue that expert evaluations need to be present in all formal courses [17], these experts may not exist or scale properly for large amounts of subjective assessments. Instead, motivated by the idea that learning occurs through both direct evaluation of the self (by creating the submission) and vicarious evaluation of others (by judging other submissions) [27, 28], UX Factor presents a numeric measure of how well the student rates others as part of the feedback report. The biases of poor reviews are also inherently mitigated by pairwise-comparison structure, which asks students to score submissions relative to each other.

After all peer feedback is submitted, a rank aggregation model is typically used to synthesize the comparative judgment scores into a relative ordering of submissions. While using the simple median was demonstrated to be sufficient for rank computations in cardinal systems [23], most existing ordinal systems rely on ranking algorithms that are harder to explain to students [6, 30, 33]. Existing methods for rank aggregation include PageRank [29], Bradley-Terry (BT) [5], Spring Rank [3], and Rank Centrality [26]. When systems incorporate these models for their own grade computations, they adapt them specifically to comparative peer review, resulting in algorithms such as Crowd-BT [8], which improves BT by additionally accounting for the quality of the reviewer. As these ranking algorithms grow more niche and complex, they gain greater accuracy at the expense of understandability. Since people's lay concepts of algorithmic decision making can directly characterize their experience with algorithmic technology [24], we also use UX Factor

to conduct a comparative analysis of different ranking models to identify the tradeoffs between accuracy and understandability.

3 UX FACTOR SYSTEM DESIGN

3.1 Design Philosophy

Peer assessment involves the intimate dynamics between users who submit work to be evaluated, users who evaluate that work, and the work itself. As there is no right or wrong answers to the pairwise preferences, we use terms like **agreement** and **consistency** to express matching preferences, rather than accuracy or correctness. We use the following terms in describing these dynamics in the context of the UX Factor:

Raters: the users who are making pairwise comparisons to judge the work that has been submitted. Note that in the UX Factor, this includes both students and teaching assistants.
Submissions: the work handed in by students that is being assessed in peer comparisons by the raters.
Projects: an assignment for which students submitted their work to the UX Factor. Each project was evaluated by raters using a distinct set of questions.
Models: the ranking strategy used to convert pairwise raw rater preferences into rankings.
Agreement: between-rater similarity; the extent to which rater preferences correspond with each other.
Consistency: between-model similarity; the extent to which rankings models correspond with either themselves or with the raw rater preferences.

As most users in UX Factor adopt a dual student-rater role, participating in the peer review process may incite meaningful self-reflection. While previous systems mainly emphasized different strategies to provide high quality feedback to users who submitted work [6, 10, 20, 30], we also wish to foster learning for the reviewers. Gaining practice in judging others' work allows students to consider their own submissions through a more critical eye. As designs in the wild for mass consumption are typically "graded" by mass appeal, we were especially interested in how well a rater's preferences agree with everyone else's preferences for each pair of submissions. Understanding the overall reliability of raters can also help students better trust the feedback they receive. Using UX Factor, we quantify this between-rater similarity and trace its development as raters compare more submissions.

In the spirit of prior work that encourages transparent [18] peer review systems, we were concerned with both complexity and accuracy when determining how to create robust models. Less complex models are easier to understand and consequently to judge, but may not capture latent properties of data as precisely as more complex ones that are harder to judge. These latent properties may lead to higher model accuracy; balancing the two thus becomes a tradeoff. To determine the right balance of complexity and consistency, we compare various rank aggregation models and derive recommendations for future comparative peer review systems.

The resultant UX Factor offers a scalable platform for comparative assessments of design. To provide users with a fluid experience that aids the learning process, we explore the presentation of submissions (via thumbnails, links), types of responses to the rubric questions (discrete versus continuous-scaled slider answers, short responses), overall layout of the interface, ways to pair submissions for comparison, and a mechanism known as the "Share Club," which are all detailed in the sections below. UX Factor is also publicly available[1] along with the source code for instructors to download and use in their own online classrooms. Using our system, we sought to investigate the following research questions:

[1] https://uxfactor.cs.brown.edu/

Q1: How agreeable are peer raters across both duplicate and transitive submissions in the comparative grading of design? Are there specific rubric questions that produce higher agreement?
Q2: Are there relationships between a submission's agreement and its final ranking?
Q3: Which ranking models for pairwise comparisons are the most consistent? Are there tradeoffs between consistency and understandability for these models?
Q4: How sensitive are the ranking model consistencies to different class sizes and random noise?

3.2 Iterative Design of the User Interface

To design an interface that is easy to navigate and can help students better engage with comparative evaluation, we followed an iterative design process to identify key features to include.

3.2.1 UI Design. The user interface of UX Factor (see Figure 1) features two main modals. The first panel on the left contains a set of pairwise submissions submitted by two students, arranged vertically. Each submission is accompanied by a blurred preview image and a hyperlink that opens up a new tab containing the full submission. Accounting for the fact that students who are rating each other's work in a real classroom setting may have previous connections with one another, we also included a way for students to skip over submissions that they would be biased against for any reason via the "I'm Biased" button. Clicking on this button would replace a submission with a new one while flagging the old submission so that it could not be randomly shown to the user again. This allowed for more comfort amongst users when evaluating submissions, as they would never have to evaluate a submission that they recognized, that belonged to a friend, or that they for any reason felt biased in rating.

The second panel on the right contains a list of rubric questions determined by the instructor. By changing the value of each slider, the rater can provide peer comparisons by answering the questions on a scale demonstrating preference towards either Submission A or Submission B. Below the rubric questions, we also add text boxes for the two submissions so that raters can optionally provide free-form qualitative feedback to each submission. Finally, on the bottom of the interface, we also include a progress bar so that raters can track how many submission pairs they have already evaluated.

3.2.2 Usability Testing. To identify interface elements that are most conducive to an effective comparative peer grading platform, we conducted two rounds of pilot studies on a group of 11 graduate and undergraduate researchers.

In the first study, we determined that utilizing a Likert scale from -2 to +2 for the rubric questions was the most effective in creating variation in data, but not so much that extreme values would be rarely chosen by the users. We also considered using a continuous instead of a discrete scale, but users reported that with the former, they felt that their scores were arbitrary and reported a lack of understanding in what the score "meant." One user specifically mentioned, "I think having a continuous slider made me grade more harshly towards the extremes." To reduce ambiguities and bias, we added demarcations between submissions, discrete tick marks for the Likert sliders, and clearer wording of buttons and question labels. These clarifications are in line with previous studies that improve feedback quality by discretizing rubric-based tasks [15].

In addition to submitting comparative scores for submission pairs, we also gave raters the option to provide written feedback for each submission. In our second study, 5 out of 11 (45%) of users elected to take this option, citing reasons such as "I wanted to justify my rating" and "It gives the grader more specificity if they're not sure of the slider inputs." These motives are consistent with earlier findings that in comparisons, reviewers tend to explain what they don't like about the

work they rated lower [6]. Thus, although less than half of the users wrote qualitative feedback, we wanted to still keep the option open, even if under-used. 4 out of 11 (36%) of users reported that they sometimes forget which submission was A and which one was B," so we added preview images for each submission so that users could be both visually reminded of each and better able to delineate the two. While this led to more confidence amongst users in differentiating the submissions, it simultaneously created a false sense of confidence that the preview image was sufficient information for comparison, occasionally leading users to not bother to click on the submission links. To alleviate this issue, we decided to slightly blur each preview image, so that they would be a useful distinction between submissions without eliminating the need to click on the link to the full submission.

3.3 Backend Design & Implementation

The backend of the UX Factor system focuses on sending, receiving, and organizing all necessary data for the platform, largely through the use of Google Sheets in order to create a familiar and manageable interface for administrators of the app. The choice of a Sheets backend further eases the process of data analysis, as it provides a quick means of computation using its built-in Google Scripts features, a simple way to download data in alternate formats (e.g. CSV files), and ease-of-use for instructors not familiar with programming.

3.3.1 Project Creation and Hand-in. When creating a new project to be hosted on UX Factor, administrators must fill out a Google Form specifying the questions to be asked to users, the types of questions asked, and other project-specific information. Given that projects can vary greatly in both concept and content, the freedom to allow administrators to customize each project is critical to the user experience, generalizability, and extensibility of the platform, and, most importantly, to gathering meaningful relationships and relative rankings between submissions. If "poor" questions are asked to users, evaluations become less confident, have less variance, and even become inconsistent—as we will explore more in Section 6.

Submissions are also handed in via a Google Form, which is linked to a Google Sheet. UX Factor then reads off of this sheet to obtain the data it needs in order to display each submission. This includes the submission's URL, preview image, and associated project.

The choice of Google Forms for project creation and submission hand-in provides several benefits. As a tool commonly used by students and instructors alike, it provides familiarity and thus a very low learning curve, if any. Additionally, submitting via Google Form provides response receipts that can be kept for record purposes.

3.3.2 Submission Pair Generation. The backend not only displays each anonymized submission, but also tracks that students are not shown duplicates during their session. Though submissions are anonymized before being shown to users, the backend tracks the author data behind the scenes, so that the rater and author can be cross-checked beforehand to avoid users being shown their own submission.

To ensure that submissions received about the same number of evaluations, a pseudo-randomized process selects the pair of submissions that each rater would see, so that submissions with less evaluations could be given a higher priority. To avoid duplicate pairs, which frequently arose when we opted to show simply the two least-viewed submissions, we decided to pair **one** least-viewed submission with one random submission. This allowed for better control and certainty that submissions would receive similar amounts of evaluations, while also ensuring that pairings would be sufficiently different such that meaningful rankings of the data could be generated while minimizing duplicate pairs.

3.3.3 Opt-in "Share Club". Previous studies have shown that viewing other people's stellar work may discourage users by causing them to believe they can never attain similar levels of high performance [34]. However, we still want to provide opportunities for students to learn how to improve their own submissions through viewing high quality work. Thus, to balance the two, we created the notion of an opt-in "Share Club" where students who participate can view submissions that were ranked higher than their own. In addition, only the submissions of participants may be shared, so unwilling students can also prevent their own work from being viewed.

4 RANKING MODELS

A ranking model takes in all the ratings given to each submission and computes a ranked ordering of the submissions. In UX Factor, a rating ranges from 2 if the rater strongly preferred it, to -2 if the rater strongly preferred the submission it was compared against. To convert these raw ratings into rankings, we explore 4 different ranking models and their tradeoffs as implications for UX Factor. The following ranking models were chosen due to their simplicity, error-minimization, compatibility with pairwise preferences, and robustness to duplicate comparisons.

4.1 Simple-average

Inspired by previous findings that simple strategies like the median [23] suffices as an accurate grading scheme, we started with a simple baseline which computes a standard mean of all the ratings a submission has received. These means are then sorted in decreasing order to construct the final ranking. The simple-average method is easily explainable to students and computationally efficient. The nature of averaging is robust to outliers, so submissions that receive multiple ratings of 2 and a few ratings of -2 can still receive a high rank.

However, a simple-average makes the assumption that all ratings have equal weight, and doesn't consider the quality of the submission that each submission was being compared to. Thus, this method would require that submissions are compared enough times such that the probability that it encountered both worse and better submissions is high. Due to the narrow range of the [-2, 2] Likert scale, the resultant range of averages is also small; differences in the ranking scores may not reflect the actual difference in preference of the submissions.

4.2 Additive

There are instances, however, where attributing equal weights to every rating has detrimental effects. When submissions are repeatedly compared against another same submission, generating *duplicate comparisons*, extreme duplicate ratings may heavily skew submission scores with a naive simple-average model. Thus, we also consider an additive approach, outlined in Algorithm 1, which aims to reduce this skew by accounting for the impact of duplicate comparisons when computing the submission average. Specifically, for each rubric question a submission is rated for, the additive model divides the sum of its ratings by (1 + the number of duplicates), thereby damping the effect of potential repeated extreme ratings. To expand the range of the rankings to allow for greater differences between submissions, we mapped the ranking scores to a numeric mean of 85. Note that this standardization is optional, however, as the relative ranking remains unaltered.

While the additive ranking model attempts to improve the equal weight and narrow range problems of the simple-average model, it still suffers from other assumptions. For example, it assumes that submission pair generation is fair. That is, if perchance one submission is only compared against other submissions that are strictly "worse" than itself, its ranking score will be artificially inflated.

Algorithm 1 Additive ranking model. $sScore_i$ is the ranking score for submission s_i.

$l \leftarrow$ new Array
for submission s_i where $i = 1, ..., N$ **do**
    $pos, neg \leftarrow 0, 0$
    for each rubric question q_j where $j = 1, ..., M$ **do**
        $s_iq_j \leftarrow$ (sum of all ratings s_i received for q_j) / (1 + # of duplicate comparisons)
        if $s_iq_j > 0$ **then** pos += s_iq_j
        else neg += $|s_iq_j|$
        end if
    end for
    if $pos + neg == 0$ **then** $avg \leftarrow 0$
    else $avg \leftarrow ((pos - neg)/(pos + neg) \times 50) + 50$
    end if
    l.append(avg)
end for
$l_{mean} \leftarrow$ mean(l)
$l_{sd} \leftarrow$ sd(l)
for $sAvg_i$ in l where $i = 1, ..., N$ **do**
    $standardized \leftarrow (sAvg_i - l_{mean})/l_{sd}$
    $sScore_i \leftarrow (standardized * ((l_{sd}/l_{mean}) \times 15)) + 85$
end for

4.3 Bradley-Terry

To mitigate the assumptions of fair pair generation, we also consider the Bradley-Terry (BT) model [5], which assumes that in any comparison between submissions A and B, the probability that submission A is better than B can be represented by α_A/α_B. We can model this as a logit:

$$\text{logit}[P(A > B)] = \log \alpha_A - \log \alpha_B, \tag{1}$$

where some α_X can be approximated using maximum likelihood (MLE) for any submission X. In fitting this model, we convert all ratings into pairs of binomial frequencies of "wins" between two submissions [40]. For example, if a rating $r \in [-2, 2]$ was provided for a comparison between A and B, the corresponding frequency pair is $(2 + r, 2 - r)$. With the binomial frequencies, we then fit a generalized linear model to find the α values for each submission.

Since BT finds α values to best satisfy α_A/α_B for any pair A and B and does so for *all* submission pairs, it can effectively output relative rankings that maximize adherence to the transitive relationships of the input comparisons. This property also allows the model to more effectively rank submissions that receive relatively fewer and less extreme comparisons, since the model can account for both how the submission fared in that one comparison and the quality of the submission it was paired against. For example, if a submission received only one rating of 0 when paired against another "strong" submission, we can assume it is roughly as strong as that submission (whereas in the simple-average and additive models, it would receive a more modest ranking score).

4.4 PageRank

Alternatively, another strategy that can effectively capture transitive relationships to establish rankings is based on PageRank, a method originally designed to rank web pages [29]. In this algorithm, the final ranking score for a submission s reflects the probability that randomly following

sequences of submissions that were preferred to previous ones will lead one to s. In other words, this can be thought of as a random walk on a directed graph [13].

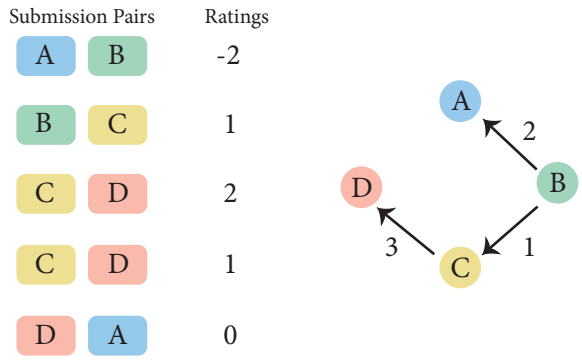

Fig. 2. To compute PageRank rankings, ratings of submission pairs are converted into a directed graph where the nodes denote submissions and the weighted edges denote preference towards a particular submission. Duplicate comparisons occupy the same edge.

As depicted in Figure 2, we consider each submission as a node and the ratings made by each rater as directed edges. Specifically, we direct edges towards the submission which was rated higher in a pairwise evaluation and assign it an edge weight based on how much better it scored quantitatively. Ratings from duplicate comparisons occupy the same edge and adjust that particular edge weight. The subsequent PageRank scores from this graph using a damping factor of 0.85 were then ordered to construct the final ranking. One caveat with this strategy, however, is that it suffers from a size bias; submissions that were compared more often will receive more ratings and thus a higher PageRank. And alternatively, submissions that were in more duplicate pairs would have fewer edges due to the concatenation of those ratings into one edge, leading to a seemingly lower number of ratings and thus lower PageRank scores.

5 DEPLOYMENT IN A UI/UX CLASS

UX Factor was deployed in a university computer science course on designing user interfaces and user experiences. 120 students and 13 teaching assistants used the platform for comparative peer grading across 4 major projects that spanned different aspects of UI/UX. All projects used in the analysis were individual assignments. These projects are listed below in chronological order with a description of what the project was about.

Project A: Personas & Storyboarding. Observe and interview users interacting with an interface, create personas based on these users, and illustrate a storyboard for one of the personas.

Project B: Redesign. Identify flaws in an existing interface, create low-fidelity and high-fidelity prototypes for various screen sizes, and build a responsive website based on those prototypes.

Project C: Portfolio. Build a personal portfolio website featuring projects from the class and beyond.

For each project, the students created publicly-accessible anonymous websites and submitted links to these websites via a Google hand-in form. After the submission deadline, the students were able to login to UX Factor to view and compare their peers' submissions across 4–5 rubric

Proj	Q#	Question
A	1	Which submission provides more objective user observations that directly inform their personas? Consider: Does the author avoid assumptions? Do these observations align with their personas and storyboard?
A	2	Which submission had interview questions that were not leading questions in themselves, but led to more informative responses?
A	3	Using only what's in the empathy map, which persona could you more confidently act out in different scenarios?
A	4	Which submission's storyboard more clearly depicts the entire interaction with the interface (from start to finish)? Consider the three metrics from the UX planet reading (authenticity, simplicity, emotion).
A	5	Which submission's storyboard better reflects the goals and characteristics of its matching persona?
B	6	Which submission's low-fidelity wireframes have more intuitive layouts for their respective devices?
B	7	Which wireframes (with annotations) better address the usability issues listed by the author in the "Identifying Usability Problems" section of the assignment?
B	8	Which submission's high-fidelity prototypes (with annotations) would be more readily made into a working website?
B	9	Which submission's high-fidelity prototypes better uses visual design principles (color, typography, text hierarchy, etc.) to show hierarchy and consistency?
B	10	Which visual design style guide better highlights the different base states, interaction states, and other visual elements used on the actual website?
C	11	Which portfolio does a better job explaining the premise of each project? Consider how someone who was not familiar with the projects would understand the premise of each project.
C	12	Which portfolio presents the projects in a more concise and engaging way? Consider the portfolio-ready examples: interesting lessons learned, a surprising finding, relates to the reader, etc.
C	13	Which portfolio better portrays the overall character of the student? Consider the theme, expertise, and goals of the portfolio as a whole.
C	14	Which portfolio better follows the usability principles we've learned in class? Consider principles such as navigation, text hierarchy, visual design, interaction behavior, etc.

Table 1. Rubric questions used in the UX Factor, for each of the three projects, labelled in chronological order.

questions. Although we had a total of 133 potential raters, only 126 raters participated in rating for Project A, 131 for Project B, and 131 for Project C. Each rater was asked to complete a minimum of 10 comparisons for Project A, 8 for Project B, and 8 for Project C. In practice, this request led to slightly different numbers of paired judgments for Project A ($\bar{x} = 10.3$, $sd = 1.8$), Project B ($\bar{x} = 9.3$, $sd = 1.4$), and Project C ($\bar{x} = 7.2$, $sd = 1.2$). We observe that although the recommended quota did not change between Projects B and C, raters provided fewer pairwise comparisons in general. The number of submissions assessed is roughly double the number of comparisons. We also record different numbers of assessments provided to each submission[1] for Project A ($\bar{x} = 22.2$, $sd = 4.6$), Project B ($\bar{x} = 20.8$, $sd = 2.4$), and Project C ($\bar{x} = 15.8$, $sd = 1.9$).

[1]Clarification: Although the sum of submissions assessed is double the sum of comparisons, we have more raters than submissions rated due to the inclusion of TAs. Thus, the mean of paired judgements = (sum paired judgements ÷ number of raters) and the mean of assessments for each submission = (sum assessments ÷ number of submissions).

Informed by prior literature that suggest rubrics with parallel sentence structure, clear wording, and specificity led to lower grading variance [23], we designed rubric questions (see Table 1) with these ideas in mind. Qualitative observational studies have also shown that greater transparency of grading criteria improves submission quality [18], so rubric questions guided the ratings and were presented to the students when each project was initially assigned. The students were able to evaluate each question and suggest modifications prior the UX Factor assessment period.

Before using the system to submit ratings, each student familiarized themselves with the platform by navigating through a pilot project with example submissions provided by the instructor. After this "calibration" process, students were given 1 week per project to complete their peer comparisons. At the end of the course and before the research was underway, all peer evaluations were anonymized by replacing their original names with pseudonyms, and the name mapping was permanently deleted.

For the analysis in this paper, the only data used were the pairwise preferences [-2, 2] and pseudonyms. None of the grading, or the rankings generated during the class, were made available for research. Our human subjects review office reviewed the procedure, and determined that it did not fall under the definition of human subjects research and opted not to review any further.

6 AGREEMENT BETWEEN STUDENT RATERS

Rankings produced from the preferences of the masses depend on some level of rater agreement. If raters largely disagreed about which submissions are better, the resulting ranking would be muddled with inconsistencies. Higher agreement among raters indicates similarity in beliefs and produces a robust ranking. Therefore, analyzing the agreement between raters is important to evaluate whether students can be a reliable source of ratings.

6.1 Students' evaluations have higher agreement than random.

Though the intention of a pseudo-random pairing system is that submissions will be evaluated against n different submissions, it is still possible for submissions to be compared against the same submission multiple times. This apparent weakness actually provides us valuable insight over transitive analysis, as it allows us to directly examine the agreement between raters of the same pair. The number of duplicates (comparisons where another unique comparison for the same pair already exists) varied per project ($A = 122, B = 560, C = 58$). Note that due to a bug in our randomization algorithm for project B, there were significantly more duplicates for that project than projects A or C. To compensate for the number of duplicates in project B, TAs were asked to complete extra comparisons until each submission had at least 9 unique comparisons, comparable to the minimums of the other two projects ($A : Min = 10, C : Min = 12$).

To calculate agreement, we analyze the comparisons of all the pairs that have at least one duplicate. This is the superset of the duplicates and the first comparisons of those pairs with duplicates. We refer to these as "duplicate comparisons." The number of duplicate comparisons for each project is sufficient for analysis ($A = 238, B = 733, C = 133$).

Using these duplicate comparisons, we can investigate whether students have a stronger sense of agreement than random preferences. To do so, we calculate inter-rater *disagreement* by finding the average standard deviation of the preferences given to duplicate pairs for each question. To mitigate the effect of students' different tendencies to rate with extreme values, we only consider the direction of the preference, and not the magnitude, thus collapsing the rating scale from [-2, 2] to [-1, 1]. We also perform these calculations on a set of 1,000,000 randomly generated preferences for comparison. We found that all 14 questions yielded disagreement values lower than the disagreement calculated from the random preferences. Additionally, the percentage of duplicate comparisons where more than 50% of raters agreed was 42%, 63%, and 57% for projects A, B, and C,

	Duplicates*	Duplicate Comparisons**	Total Comparisons
Project A	122 (9.1%)	238 (17.7%)	1,345
Project B	560 (46.0%)	733 (60.2%)	1,218
Project C	58 (6.4%)	113 (12.5%)	901

Table 2. The number of duplicates and duplicate comparisons per project. Note that Project B had an especially high number of duplicates due to a small bug in our randomization pairing system, which was fixed about halfway through the UX Factor rating window. To compensate, TAs were asked to complete extra comparisons until each submission had a similar number of unique comparisons to the prior project. *Comparisons where another unique comparison for the same pair already exists. **Duplicates *and* the first comparisons for the pairs with duplicates.

respectively. These values are higher than the expected percentage for random preferences (33%), suggesting that the agreement among raters is greater than the agreement due to random chance.

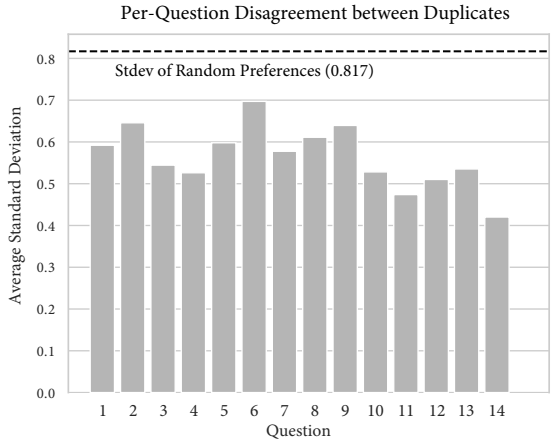

Fig. 3. Column chart comparing the average inter-rater disagreement (standard deviation) between preferences of duplicate pairs for each question. All 14 questions yielded lower disagreement than that of the set of 1,000,000 randomly generated preferences, as indicated by the dotted line. Since lower disagreement indicates that raters agreed more on which submissions were better, this suggests that agreement among students is stronger than random chance.

Differences in agreement appeared to vary per question, with the highest being Question 14 and the lowest being Question 6. Due to the small number of questions, it's difficult to determine which aspects contributed to a question's agreement. Though it may be natural to assume questions with more explanation would result in higher understanding and therefore agreement, no statistically significant correlation was found between question length and agreement ($r_s = 0.37$, $p = 0.2$).

6.2 Ranking models are not biased against submissions with high disagreement.

A concern of peer grading systems is that ranking models do not fairly rank "disagreeable" submissions (submissions which yield high disagreement among their raters). Analyzing the relationship between per-submission agreement and scores from the ranking models can provide insight into this claim. To investigate this, we analyze duplicate comparisons to find the average disagreement

(standard deviations) between raters for each submission. We suspected that submissions with high disagreement would obtain "mediocre" scores. However, plotting disagreement against ranking model scores revealed no significant relationship, as seen in Figure 4. "Mediocre" submissions yielded similar disagreement to strong- and weak-performing submissions. This suggests that these ranking models are not biased against disagreeable submissions.

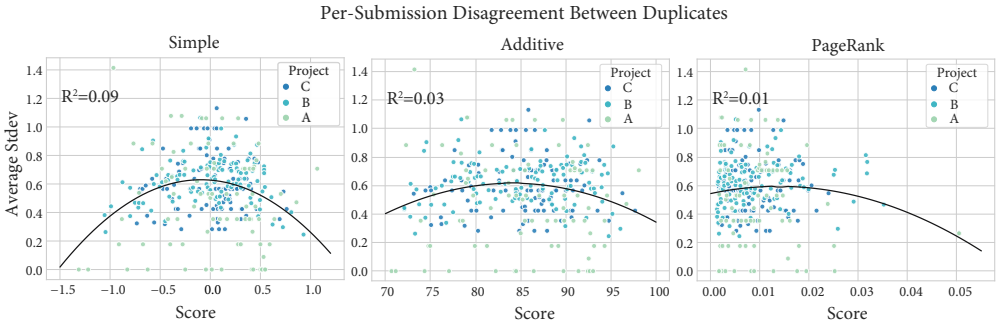

Fig. 4. Scatter plots comparing the per-submission disagreement (average standard deviation) of ratings against the scores produced by three ranking models. No correlation was found between a submission's disagreement and its model score. The Bradley-Terry model was omitted for brevity as it also demonstrated a very low correlation. *Left*: Simple model ($R^2 = 0.09$) *Middle*: Additive model ($R^2 = 0.03$) *Right*: PageRank ($R^2 = 0.01$)

6.3 Students are also agreeable across transitive comparisons, but for different questions.

While looking at duplicate comparisons provides a more direct comparison of inter-rater agreement, we can follow a transitive approach to gather a more holistic view of the general agreement between raters. We define *transitive* as comparing multiple submission ratings indirectly; i.e. if A is preferred over B and B is preferred over C, then we can conclude that A would be preferred over C. Since the nature of directed graphs captures the transitive relationships between pair-wise comparisons, we used the PageRank model scores to assess transitive agreement. The plot showing the transitive scores of submissions where all questions were utilized in the edge weights is shown in Figure 5. We see that there is a moderately strong correlation between our additive model scores and PageRank model scores, ranging from $r_s = 0.61$ to $r_s = 0.76$ depending on the project. This indicates that the students agreed fairly well with one another on a broader, more holistic level.

This graph-based approach can also be utilized in the analysis of per-question consistency, in that we can create graphs where only evaluations from one question are taken into account in creating edges, thereby yielding transitive comparisons for only that question. The PageRank scores each submission received on a per-question basis were plotted against the overall PageRank scores that each submission received when all questions were taken into account, and the strength and slope of these correlations are shown in Figure 6. The strength of these correlations can be used to determine how well raters agreed with one another in their evaluations, while the slope of the correlation indicates the strength of agreement with the final PageRank scores. It was found that Questions 14, 10, 4, and 5 have the strongest slopes, indicating that these questions agree the most with the ultimate PageRank scores. These questions also have the largest correlations, indicating high inter-rater agreement across transitive raters.

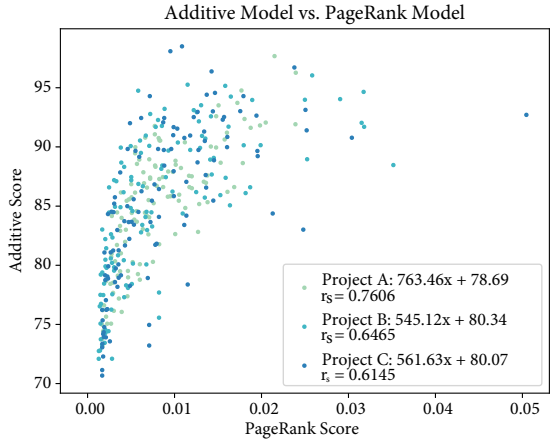

Fig. 5. Scatter plot of computed PageRank scores using the transitive graph-based approach versus the scores that the submission received from our additive model calculations. A fairly strong correlation was found between these values, ranging from $r_s = 0.61$ to $r_s = 0.76$ depending on the project.

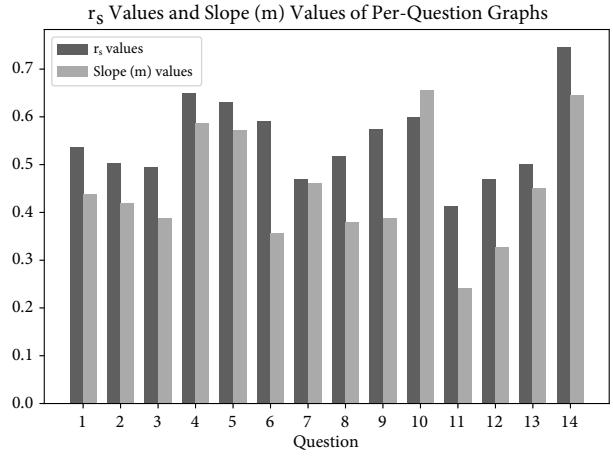

Fig. 6. A bar chart depicting the r_s values and slope (m) values for the correlations between each question graph's PageRank scores against the project's PageRank scores. We see that Questions 14, 10, 4, and 5 have the strongest correlations and the largest slopes, indicating that these questions have the highest agreement with the final PageRank scores as well as the highest agreement between across transitive raters.

7 CONSISTENCY OF MODEL RANKINGS

Previous peer assessment systems have used various rank aggregation methods to generate rankings from raw ratings, the accuracy of which was determined by comparing the model rankings to TA, instructor, or expert rankings [6, 21, 23, 27, 33, 41], despite other meta-analyses suggesting that agreement between peers and teachers are not a good measure of validity [11]. Thus, we instead gauge ranking validity and robustness based on *consistency* to other models and to the raw raters' preferences, as previously defined in Section 3.1.

7.1 Models of varying complexity output similar final rankings.

To gauge between-model consistencies, we compute Pearson's correlations (r_s) between the rankings provided by each model. On average across the three projects, the simple-average vs. additive models demonstrated the strongest consistency ($r_s = 0.97$), followed by simple-average vs. BT ($r_s = 0.95$), additive vs. BT ($r_s = 0.95$), PageRank vs. BT ($r_s = 0.70$), PageRank vs. additive ($r_s = 0.67$), and PageRank vs. simple-average ($r_s = 0.63$). A finer breakdown of the final rankings comparison per project is provided in Figure 7. Following the x and y axes for each pair of models provides scatterplots (bottom-left) and correlations (top-right) between each model pair. The diagonals depict histograms of the model ranking scores. We note that the histograms corresponding to the simple, additive, and Bradley-Terry models generally have a normal distribution of scores while PageRank exhibits a slight right skew.

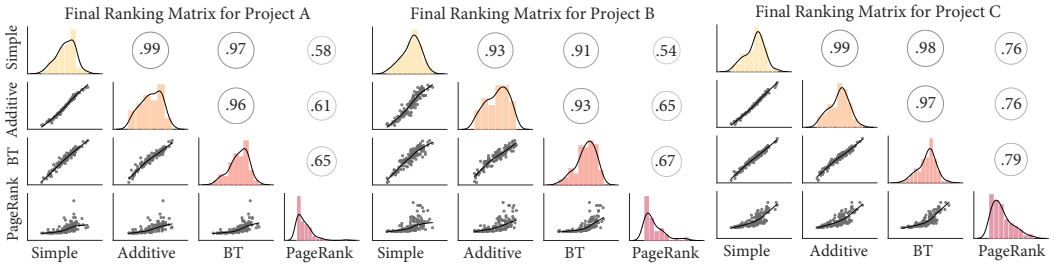

Fig. 7. Matrix plots demonstrating that the outputs of the simple-average, additive, Bradley-Terry (BT), and PageRank algorithm across each project produce similar final rankings. PageRank is less consistent to the other models and exhibits a slight right skew in its distribution. *Bottom left*: scatter plots with fitted curves, *diagonal*: histograms, *top right*: Pearson's correlation.

These correlations indicate that the simple-average, additive, and BT models generate relative rankings that are almost identical, despite the BT model involving more complex computations than that of a simple-average. Across all three projects, we also observe that PageRank exhibits the lowest consistency with the other models (see the last columns and rows in each plot of Figure 7). A closer inspection into PageRank score outliers reveals that this discrepancy results from different ranking ideologies. Specifically, there was one submission that received the highest possible PageRank score but only a slightly higher than average additive score. In practice with the additive model, this submission was actually preferred to 4 other submissions that scored above it, but lost to 3 other submissions that scored below it. This pattern can be seen in other submissions who received high PageRank scores but comparatively lower additive scores, indicating that the additive model emphasizes the *number* of other submissions a student submission was preferred over, while PageRank places greater weight on the *rankings* of the submissions it was preferred over.

Comparing the between-model correlations additionally reveals lower overall r_s values for Project B, which contains at least 3 times more duplicate comparisons per submission. This indicates that without controlling random submission pair generation to minimize duplicate pairs, greater care needs to be taken in the model selection process as each model may handle disagreements between raters for duplicate pairs differently.

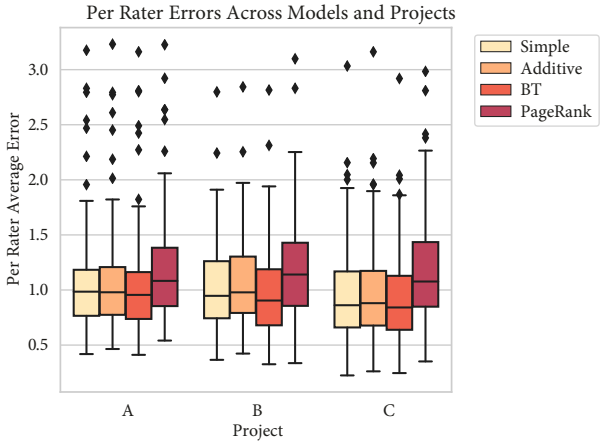

Fig. 8. Boxplots of the per-rater errors for the four ranking models investigated across three separate projects. Diamonds indicate outliers. PageRank exhibits the greatest error, followed by the additive, simple-average, and BT models.

7.2 The complexity of the model does not affect its consistency with raw rater preferences.

We quantify model-rater consistency, or the extent to which ranking models correspond with raw ratings, using squared errors (denoted ϵ). This error represents the squared differences between the observed preference (provided by raters) and the aggregated preference (provided by a ranking model) between pairwise submissions. This indicates that the higher ϵ is, the less apt the model is at capturing the overall ranking of all submissions from the raw preferences. Specifically, given two submissions s_A and s_B, a rater rating $r(s_A, s_B) \in [-c, c]$, a map $R_m(X)$ denoting the ranking score of some submission X in a ranking model R_m, and $R_m(X, Y)$ denoting the difference between $R_m(X)$ and $R_m(Y)$ scaled to $[-c, c]$, the error of that rating ϵ_r is

$$\epsilon_r = \big(r(s_A, s_B) - R_m(A, B)\big)^2 \tag{2}$$

$$R_m(A, B) = \frac{R_m(B) - R_m(A)}{\max R_m(X) - \min R_m(X)} \cdot c \tag{3}$$

In Equation 3, we specifically divide by $(\max R_m(X) - \min R_m(X))$ and multiply by c to map $R_m(A, B)$ to $[-c, c]$. We also set $c = 2$ to match the Likert scale range of our rater preferences. We then compute the average ϵ_r across all ratings provided by each rater, as displayed in Figure 8.

Overall, across the 3 projects, per rater errors of each model were significantly different. Specifically, paired t-tests revealed that the PageRank model had higher error than the additive model ($t = 11.92, p < 0.001$), the additive model had higher error than the simple-average model ($t = 4.61, p < 0.001$), and the simple-average model had higher error than the BT model ($t = 9.22, p < 0.001$). Ranked from highest to lowest error, we thus have PageRank ($\bar{\epsilon}_r = 1.18$), additive ($\bar{\epsilon}_r = 1.03$), simple-average ($\bar{\epsilon}_r = 1.01$), and BT ($\bar{\epsilon}_r = 0.98$). However this error order does not agree with model complexity, as a simpler algorithm such as the simple-average model generated lower errors in comparison to the PageRank model.

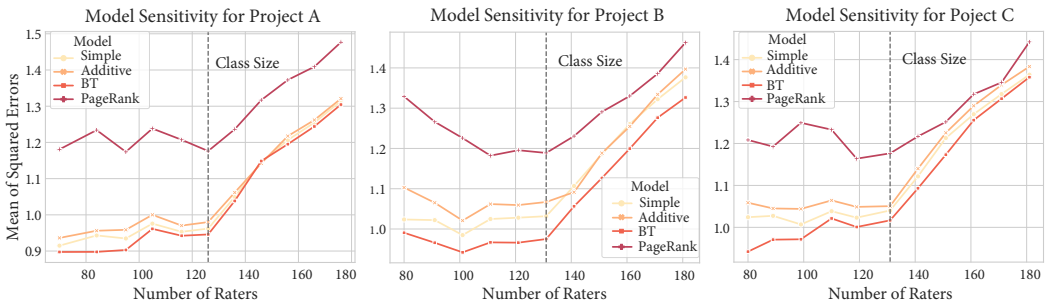

Fig. 9. Line plots of MSE against the total number of raters. The vertical dashed line, class size, denotes how many raters there actually were ($A = 126, B = 131, C = 131$). From these plots, we can see that (1) more noisy raters increases mean error, (2) PageRank generally has the slowest **rate** of error increase with noise, and (3) the amount of error with fewer raters is increased most drastically in Project B.

7.3 Model sensitivity depends on the class size and level of noise.

While we evaluated model-rater consistency based on the premise that each rater provided genuine preferences, this assumption may not necessarily hold in all online classrooms, where there exists a possibility of "bad" ratings generated by less careful raters. We were also interested in seeing whether our observations for model-rater consistency persisted with a smaller classroom size. Thus, we investigated how the models performed with different numbers of raters. Noise was generated by incrementally adding 10 raters, each providing random ratings $\in [-2, 2]$ for 8–10 randomly sampled existing submissions. We also incrementally removed 10 randomly sampled raters and all the corresponding ratings submitted by these raters. The effects of these rater modifications can be seen in Figure 9. As expected, mean rater error increases when raters provide random ratings. However, for Project B and more notably Project C, we can observe that this rate of increase is slower for PageRank in comparison to the other models. Fitting a regression line for the Project C errors resulting from random raters (from 131 raters to 181 raters) produces the following order of models in terms of slope increase: PageRank ($m = 0.0051$), simple-average ($m = 0.0065$), additive ($m = 0.0067$), and BT ($m = 0.0069$). Interestingly, BT generated the greatest slope in error increase with noise despite it exhibiting the lowest error with "proper" raters.

The mean error appears to remain relatively constant, even occasionally dropping, with fewer raters for Projects A and C. However, this is not the case for Project B, especially with PageRank and to some extent the additive and BT models where errors starts to increase when the number of raters falls below 100. The only differences between Project B and the two other projects are that Project B has significantly more duplicate comparisons (60.2%, as opposed to 17.7% for Project A and 12.5% for Project C) and that its rubric contains a set of questions tailored specifically for assessing the redesign of low-fidelity and high-fidelity prototypes. Thus, we can attribute this increase in error rates for fewer raters to either *disagreeable duplicates* or *project-specificity*, that (1) models are less effective at capturing ranking with more duplicate submissions as they are statistically more likely to conflict or (2) models are less effective at aggregating ratings resultant from assessing redesigns. Further user studies should be conducted to confirm these suspected relationships. However, we do note that in Project B, the simple-average model exhibited the least increase in mean error under 100 raters, and would serve as a good model choice in this circumstance.

	Simple-average	Additive	Bradley-Terry	PageRank
Complexity	**low**	medium	high	medium
Model-rater consistency (avg. rater error)	$\bar{\epsilon}_r = 1.01$	$\bar{\epsilon}_r = 1.03$	$\boldsymbol{\bar{\epsilon}_r = 0.98}$	$\bar{\epsilon}_r = 1.18$
Sensitivity to noise (slope of avg. rater error)	$m = 0.0065$	$m = 0.0067$	$m = 0.0069$	$\boldsymbol{m = 0.0051}$

Table 3. Comparison of different ranking models: simple-average is the least complex, Bradley-Terry exhibits the lowest error overall, and PageRank is the least susceptible to inconsistent raters.

8 DISCUSSION

Overall, data collected across 3 projects in a semester-long UI/UX course revealed that a combination of 133 student and TA raters are agreeable both when they evaluate duplicate pairs of submissions and across transitive pairs of submissions. This rater reliability allows us to trust the source of pairwise ratings, which was subsequently used to create a ranking of all the submissions. We found that each of the models investigated generated similar rankings with each other, and that their consistency to raw ratings did not correlate with model complexity. Other findings are summarized in Table 3. When considering the average mean rater error of each model with respect to the raw ratings, the BT model performed the best, even with a randomly sampled smaller class size. However, when we simulated a scenario when raters provided random comparisons, the PageRank model exhibited the lowest increase in error rate. With these findings, we can return to the research questions previously outlined in Section 3.1.

Q1. We observed higher than random agreeability between peer raters across both duplicate and transitive comparisons, albeit with different trends. Specifically, across duplicate comparisons, agreement was highest in Project B and lowest in Project A. Conversely, across transitive comparisons, agreement was highest in Project A and lowest in Project B. These differences indicate that the *type* of project may not be associated with how well raters agree with each other. However, across rubric questions in both duplicate and transitive instances, agreement was highest for question 14, which asks, "Which portfolio better follows the usability principles we've learned in class? Consider principles such as navigation, text hierarchy, visual design, interaction behavior, etc." This question differs from others in that it asks about the adherence to an objective property (usability principles) and provides examples of that property. The only similar question is 9, "which yielded average agreement," although this question additionally asks how the objective property is used to achieve "hierarchy and consistency," which could be more subjective.

Q2. We initially suspected that disagreeable submissions would receive "mediocre" ratings. Since there was no consensus that a submission was universally "good" or universally "bad," it would fall on the middle of the spectrum. However, plotting disagreement against their ranking grades revealed no such correlation. Regardless of how agreeable a particular submission was, they received both very low and very high rankings. This finding indicates that even though raters may disagree when using peer grading systems, the rankings returned will not be biased due to greater disagreement or agreement.

Q3. Comparing the four examined ranking models, we identified the following order in terms of decreasing consistency: BT, simple-average, additive, and PageRank. Interesting, BT and PageRank are less understandable in comparison to simple-average and additive, but fell on opposite ends of the consistency spectrum, indicating that the complexity of a ranking model does not affect its

consistency with raw rater preferences. The simple-average model performs marginally worse than BT in terms of consistency, but is significantly easier to explain to students.

Q4. While all four models increased in inconsistency with the addition of more rater noise, the PageRank model demonstrated the least rate of increase, indicating that it is the least sensitive to noise. Interestingly, PageRank performs the worst out of all the models at the tested class size, but has merits in settings with "bad" ratings. With smaller class sizes, consistency remained relatively constant in Projects A and C, but increased slightly in Project B. In Project B, the simple-average ranking model appeared the least sensitive to inconsistency increase.

8.1 Student Perception

Initially, there was some concern among students that peer grading was the primary determinant of their performance for the class. Their concern may have been motivated by the misconception that peer grading is an unreliable and bias-ridden process. In response, we fully explained the grade calculation process and ensured that the instructor would closely monitor all rankings before mapping them to scores. Distrust of the peer grading process seemed to diminish over the course of the semester. When students had the choice of requesting TA grading versus UX Factor grading for one of the projects, 80% opted to be graded through UX Factor. This change may be attributed to the students' growing familiarity with the system, gradually acquired through both rating others and receiving their own ratings, which lends to greater trust [14] in comparison to unfamiliar TA rating methods. Since the students have repeated this process for some time, they would also know what to expect. We also believe that the increase in trust and subsequently usage is associated with the transparency of UX Factor's decision-making process, which was previously demonstrated to have positive effects on the self-calibration of trust [42]. Incorporating students as raters may also increase empathy for the grading process.

Some students felt that UX Factor contributed to a competitive atmosphere in the class. They disliked that rating one submission highly required rating the other poorly, contributing to the perception that the submissions are pitted against each other to vie for the positive rating. Relative grading is a phenomenon that can occur in TA grading as well, but being transparent about the relative nature of grading was a surprise for some students. With this in mind, it may comfort students to limit the contribution of UX Factor on their project grade to a small portion, perhaps 10–25%. In the UI/UX classroom described in this paper, the UX Factor rankings were mapped to scores that contributed to 50% of the students' project grade, leading to some competitive stress.

Anecdotally, TAs found grading both with and without UX Factor required some level of relative comparison. In both instances, they would compare their submissions and assign grades relatively. However they felt less pressure contributing their preferences on UX Factor than assigning grades outright based on their preferences.

Some students expressed that seeing each other's work was educational and even inspiring. Students also appreciated the detailed feedback they received from each project, which included a breakdown of their ranking for each question, specific submissions they lost against (if they opted in to the "share club"), and detailed written feedback from their raters (including at least one TA). This level of feedback was higher than past years, where in the past only one individual TA provided a few sentences about their thoughts. Even though written feedback was not required in the UX Factor, it was provided alongside 88.1% of ratings on average across the three projects. However, some students noted that feedback from different raters were conflicting, so they could receive a relatively lower score even when some comments were positive.

8.2 Recommendations for Peer Review

UX Factor is not the final iteration of comparative peer review platforms. To guide the creation of future systems that can generate robust grades and augment learning, while minimizing student bias, we pose the following recommendations.

- **With enough random pairings rated, a simple average of rater preferences suffices as an explainable and accurate ranking model.** Across all of the projects we investigated, the simple-average model produced rankings that were highly correlated with more complex ranking schemes. While other models like Bradley-Terry can exhibit higher overall consistency with raw rater preferences and PageRank can account for more nuances like how well a submission's competitors performed, the simple-average is easier to understand but sacrifices a small amount of performance. However, we do note that this observation was made with at least 15 ratings per submission (duplicates are allowed) and randomly generated submission pairs.
- **In situations where students are likely to grade inconsistently, consider the PageRank model.** We hope that every time a rater submits their peer reviews, they do so diligently. However, this rigor cannot be guaranteed, as raters may gradually provide less attentive preference judgments when they see others already putting in the effort [22, 39]. Thus, in situations where the ratings become increasingly more inconsistent, models that are based on between-submission connections, and thus are more resistant to noise, such as PageRank can be used instead.
- **Write rubric questions that will differentiate submissions.** We found that the impact of questions on the model scores reflected the diversity in quality of the relevant project features. For example, asking students to evaluate a feature that most submissions completed with similar quality resulted in neutral comparisons and thus yielded little impact on the final ranking. Questions should aim to differentiate between submissions as much as possible. Similarly, questions should be clear and explicit to avoid misinterpretation and thus inconsistent ratings.

9 CONCLUSION

This work explores the relationships between student submissions, raters, and ranking models in comparative peer review. We deploy the UX Factor system in a semester-long UI/UX class and found that students are fairly agreeable as peer raters. Using UX Factor produced more feedback on student work, and the diverse opinions of the class provided varying comments that were not limited to the subjective views of one expert grader. We found that the optimal ranking model depends on the class size and level of noise, where "optimal" is defined as the balance of consistency and sensitivity. We suggest using a simple average for small classes (under 100 students), the Bradley-Terry for larger classes (over 100 students), and PageRank for noisy classes with inconsistent raters.

Overall, UX Factor can provide a consistent way to rank designs and distribute peer feedback in educational contexts as part of grade computations. We had originally set out to find a more meaningful, reliable alternative to TA grading. In the end, UX Factor provided students more feedback on their projects, exposed students to each other's work, explored the trade-offs of different ranking models, and produced scores that were not a reflection of just one undergraduate TA, but of a maturing cohort of students in the field. This work explores the ability of novice designers to consistently judge submissions, showing that pairwise preferences can be used educationally in a large group.

ACKNOWLEDGMENTS

This work was partly funded by NSF grant IIS-1552663, a gift from Figma, and the Brown University Zern Endowment.

REFERENCES

[1] Alexandra C Achen and Paul N Courant. 2009. What Are Grades Made Of? *The Journal of Economic Perspectives : A Journal of the American Economic Association* 23, 3 (2009), 77–92. https://doi.org/10.1257/jep.23.3.77

[2] Louis Alfieri, Timothy J Nokes-Malach, and Christian D Schunn. 2013. Learning Through Case Comparisons: A Meta-Analytic Review. *Educational Psychologist* 48, 2 (April 2013), 87–113. https://doi.org/10.1080/00461520.2013.775712

[3] Caterina De Bacco, Daniel B Larremore, and Cristopher Moore. 2018. A physical model for efficient ranking in networks. *Science Advances* 4, 7 (July 2018), 31 pages. https://doi.org/10.1126/sciadv.aar8260

[4] Stephen P Balfour. 2013. Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment* 8 (2013), 40–48. https://eric.ed.gov/?id=EJ1062843

[5] Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39, 3/4 (1952), 324–345. https://doi.org/10.2307/2334029

[6] Julia Cambre, Scott Klemmer, and Chinmay Kulkarni. 2018. Juxtapeer: Comparative Peer Review Yields Higher Quality Feedback and Promotes Deeper Reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, NY, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173868

[7] Orville L Chapman. 1999. Calibrated Peer Review®. http://cpr.molsci.ucla.edu/Home

[8] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. 2013. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM International Conference on Web search and data mining (WSDM '13)*. ACM, NY, NY, USA, 193–202. https://doi.org/10.1145/2433396.2433420

[9] Susan J Deeley and Catherine Bovill. 2017. Staff student partnership in assessment: enhancing assessment literacy through democratic practices. *Assessment & Evaluation in Higher Education* 42, 3 (April 2017), 463–477. https://doi.org/10.1080/02602938.2015.1126551

[10] Constantinos Demonacos, Steven Ellis, and Jill Barber. 2019. Student Peer Assessment Using Adaptive Comparative Judgment: Grading Accuracy versus Quality of Feedback. *Practitioner Research in Higher Education* 12, 1 (2019), 50–59. https://eric.ed.gov/?id=EJ1212981

[11] Nancy Falchikov and Judy Goldfinch. 2000. Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research* 70, 3 (Sept. 2000), 287–322. https://doi.org/10.3102/00346543070003287

[12] Leslie Feigenbaum and Nancy Holland. 1997. Using Peer Evaluations to Assign Grades on Group Projects. http://ascpro0.ascweb.org/archives/1997/feigenbaum97.htm

[13] Santo Fortunato and Alessandro Flammini. 2007. Random walks on directed networks: the case of pagerank. *International Journal of Bifurcation and Chaos* 17, 07 (July 2007), 2343–2353. https://doi.org/10.1142/S0218127407018439

[14] David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (Dec. 2000), 725–737. https://doi.org/10.1016/S0305-0483(00)00021-9

[15] 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*. ACM, NY, NY, USA, 458–469. https://doi.org/10.1145/2858036.2858195

[16] Morris B Holbrook. 1999. Popular Appeal Versus Expert Judgments of Motion Pictures. *Journal of Consumer Research* 26, 2 (1999), 144–155. https://doi.org/10.1086/209556

[17] David A Joyner. 2017. Scaling Expert Feedback: Two Case Studies. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. ACM, NY, NY, USA, 71–80. https://doi.org/10.1145/3051457.3051459

[18] Frédéric Kaplan and Cyril Bornet. 2014. A Preparatory Analysis of Peer-Grading for a Digital Humanities MOOC. In *Digital Humanities 2014 : Book of Abstracts*. Alliance of Digital Humanities Organizations, Lausanne, Switzerland, 227–229. https://infoscience.epfl.ch/record/200911

[19] Julia H Kaufman and Christian D Schunn. 2011. Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science* 39, 3 (May 2011), 387–406. https://doi.org/10.1007/s11251-010-9133-6

[20] Pushkar Kolhe, Michael L Littman, and Charles L Isbell. 2016. Peer Reviewing Short Answers using Comparative Judgement. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale (L@S '16)*. ACM, NY, NY, USA, 241–244. https://doi.org/10.1145/2876034.2893424

[21] Steven Komarov and Krzysztof Z Gajos. 2014. Organic Peer Assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*. ACM, NY, NY, USA, 6 pages.

[22] Yasmine Kotturi, Andrew Du, Scott Klemmer, and Chinmay Kulkarni. 2017. Long-Term Peer Reviewing Effort is Anti-Reciprocal. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale (L@S '17)*. ACM, NY, NY, USA, 279–282. https://doi.org/10.1145/3051457.3054004

[23] 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 Transactions on Computer-Human Interaction* 20, 6 (Dec. 2013), 33:1–33:31. https://doi.org/10.1145/2505057

[24] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (Jan. 2018), 1–16. https://doi.org/10.1177/2053951718756684
[25] Thomas Mussweiler. 2003. Comparison processes in social judgment: mechanisms and consequences. *Psychological Review* 110, 3 (July 2003), 472–489. https://doi.org/10.1037/0033-295x.110.3.472
[26] Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2015. Rank Centrality: Ranking from Pair-wise Comparisons. arXiv:1209.1688 [cs.LG]
[27] Ha Nguyen, June Ahn, William Young, and Fabio Campos. 2020. Where's the Learning in Education Crowdsourcing?. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. ACM, NY, NY, USA, 305–308. https://doi.org/10.1145/3386527.3406734
[28] David Nicol, Avril Thomson, and Caroline Breslin. 2014. Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education* 39, 1 (Jan. 2014), 102–122. https://doi.org/10.1080/02602938.2013.795518
[29] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/
[30] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. arXiv:1307.2579 [cs.LG]
[31] Alastair Pollitt. 2012. The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice* 19, 3 (Aug. 2012), 281–300. https://doi.org/10.1080/0969594X.2012.665354
[32] Tiffany Potter, Letitia Englund, James Charbonneau, Mark Thompson MacLean, Jonathan Newell, and Ido Roll. 2017. ComPAIR: A New Online Tool Using Adaptive Comparative Judgement to Support Learning with Peer Feedback. *Teaching & Learning Inquiry* 5, 2 (2017), 89–113. https://eric.ed.gov/?id=EJ1156350
[33] Karthik Raman and Thorsten Joachims. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and data mining (KDD '14)*. ACM, NY, NY, USA, 1037–1046. https://doi.org/10.1145/2623330.2623654
[34] Todd Rogers and Avi Feller. 2016. Discouraged by Peer Excellence. *Psychological Science* 27, 3 (2016), 365–374. https://doi.org/10.1177/0956797615623770
[35] Philip M Sadler and Eddie Good. 2006. The Impact of Self- and Peer-Grading on Student Learning. *Educational Assessment* 11, 1 (Feb. 2006), 1–31. https://doi.org/10.1207/s15326977ea1101_1
[36] Nihar B Shah, Joseph K Bradley, Abhay Parekh, Martin Wainwright, and Kannan Ramchandran. 2013. A Case for Ordinal Peer-evaluation in MOOCs. In *NIPS Workshop on Data Driven Education*. Curran Associates, Lake Tahoe, Nevada, USA, 8 pages.
[37] Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. 2016. Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale (L@S '16)*. ACM, NY, NY, USA, 389–398. https://doi.org/10.1145/2876034.2876043
[38] 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*. ACM, NY, NY, USA, 1243–1252. https://doi.org/10.1145/1124772.1124960
[39] Milena Tsvetkova and Michael W Macy. 2014. The Social Contagion of Generosity. *PLOS ONE* 9, 2 (Feb. 2014), e87275. https://doi.org/10.1371/journal.pone.0087275
[40] Heather Turner and David Firth. 2012. Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software* 48, 1 (May 2012), 1–21. https://doi.org/10.18637/jss.v048.i09
[41] Maya Usher and Miri Barak. 2018. Peer assessment in a project-based engineering course: comparing between on-campus and online learning environments. *Assessment & Evaluation in Higher Education* 43, 5 (July 2018), 745–759. https://doi.org/10.1080/02602938.2017.1405238
[42] Ewart J Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. 2014. A Design Methodology for Trust Cue Calibration in Cognitive Agents. In *Proceedings, Part I, of the 6th International Conference on Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments - Volume 8525*. Springer-Verlag, Berlin, Heidelberg, 251–262. https://doi.org/10.1007/978-3-319-07458-0_24
[43] Usman Wahid, Mohamed Amine Chatti, and Ulrik Schroeder. 2016. A Systematic Analysis of Peer Assessment in the MOOC Era and Future Perspectives. In *Proceedings of the Eighth International Conference on Mobile, Hybrid, and On-line Learning*. International Academy, Research, and Industry Association, Venice, Italy, 64–69.
[44] Yanqing Wang, Wenguo Ai, Yaowen Liang, and Ying Liu. 2015. Toward Motivating Participants to Assess Peers' Work More Fairly. *Journal of Educational Computing Research* 52, 2 (Apr 2015), 180–198. https://doi.org/10.1177/0735633115571303

Received April 2021; accepted July 2021