Appendix A: Further analysis of the labor market test

In this section, we report more detail from the test on the online labor market. For much of this section, we analyze a subset of the jobs: some job covariate information is missing in what was given to us by the labor market. We have full covariate data for 100,438 jobs (out of 184,172).

A.1. Verifying randomization in allocation of clients

As noted in Section 3.3.2 of the main paper, there was a bug in the allocation code such that 1086 clients were assigned to different treatment cells upon submissions of different jobs. Since this could potentially create contamination between our cells, we disregard these clients in our analysis. Here we make sure that neither this bug nor any other affected experimental validity by checking the distribution of client covariates across the treatment cells. We do so as follows.

We have a set of *job* level covariates for a subset of the jobs: *hourly rate of job* (if applicable), *total cost of project if not hourly* (if applicable), *previous number of closed jobs by client at time of job*, *previous spend by client at time of job*, *value of the job* (4 options), *Tier 1 category* (12 options), *Tier 2 category* (88 options), and *expertise level* (3 options). The first four are continuous covariates, and the last 4 are categorical covariates.

For each client, we sample one of that client's jobs and associate the client with that job's covariates. Then we run tests of independence for the samples of each covariate across the treatment cells. Across a variety of tests and all covariates, the results are consistent with the randomization being valid.

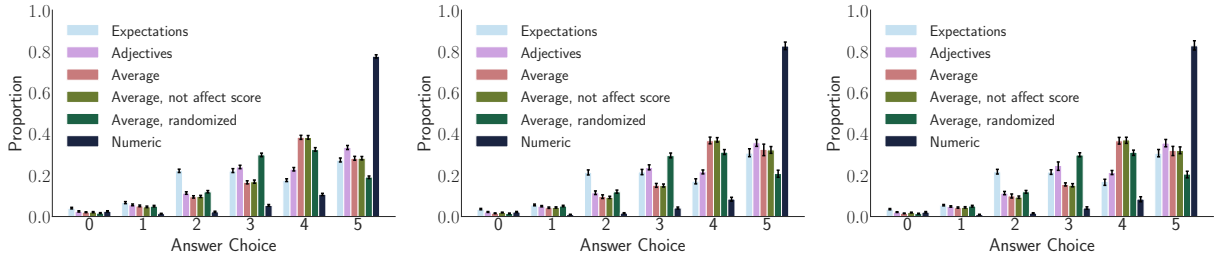
- For each continuous covariate, using the Kruskal-Wallis H-test for independent samples on all the treatment groups together, the null hypothesis that the population median of all of the groups are equal is not rejected, with $p > .9$.
- Similarly, for each continuous covariate, using the one way ANOVA F test, the null hypothesis that all the treatment groups have the same population mean is not rejected, with $p > .2$.
- For each categorical covariate, we run the chi-squared test of independence of variables in a contingency table, which tests whether the observed frequencies of values is independent of the treatment group. The null hypothesis is not rejected with $p > .1$, for each covariate.

These tests are consistent with fact that the allocation of *valid* clients we used for analysis across treatment cells was truly random. Note that these tests do not check whether the *invalid* clients (which we threw out) are similar to the *valid* clients. Invalid clients are more likely to be higher volume clients, as those who submitted many jobs during the test period provided more chances for the bug to manifest.

A.2. Robustness against high volume clients and allocation bug

Recall that in the main text we further threw out the 7 clients who submitted more than 200 jobs during the test period ("heavy users"). However, the following may still be the case: idiosyncratic rating behavior of medium-volume clients (over 50 or 100 jobs submitted) may be driving the difference in behavior between treatment cells. Here we show that this is not the case, as well as the fact that throwing out the 7 heavy users was not consequential. We further show that including the clients who were thrown out due to the allocation bug does not materially affect results.

In Figure 1, we plot the rating distributions when only sampling 1 job/client, including 7 clients excluded for submitting at least 200 jobs during the test period, and using all jobs and clients (even incorrectly allocated clients). The mean treatment responses are also included. Results are similar.



(a) Sampling 1 job per client (b) All valid clients and jobs (c) All clients and jobs

Figure 1 **Rating distributions for different client sampling techniques. As in the main text, the confidence intervals are 95% bootstrapped confidence intervals, with bootstrapped sampling at the client level.**

Data sampling policy:	From main text	One job per client	With outlier clients	All clients, (inc. incorrectly allocated)
Expectations	3.339	3.243	3.354	3.350
Adjectives	3.650	3.597	3.650	3.651
Average	3.763	3.687	3.788	3.774
Average, not affect score	3.777	3.693	3.777	3.771
Average, Randomized	3.465	3.438	3.463	3.458
Numeric	3.594	4.534	4.635	4.639

Table 1 Average treatment responses under different data policies

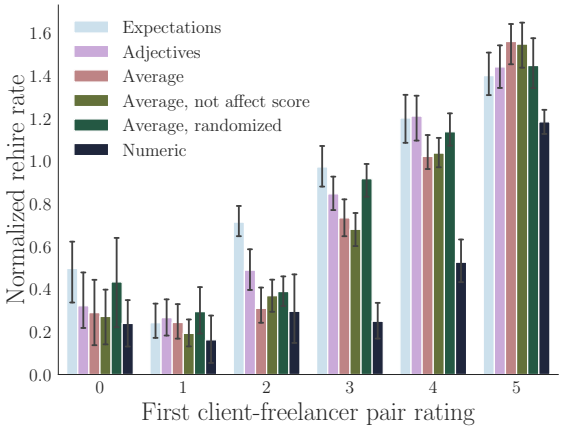


Figure 2 Likelihood that a client will rehire a freelancer during the time period of the test, given just the first rating the client gives that freelancer during the test period. Values are normalized by the overall mean rehire rate. Confidence intervals are 95% intervals with bootstrapped sampling done at the client level. Note that the overall rehire rates are similar for each treatment; that the verbal cells appear to be higher conditional on each rating is an example of Simpson's Paradox: in the numeric cell, about 80% of experiences are represented by the "5," and this value is lower for the verbal cells, cf. Figure 1.

A.3. Regressing treatment response with treatment cell and other covariates

We regress the treatment response with treatment cell and all of our job covariates (except tier 2 category, which had 88 unique values and is a more granular version of tier 1 category). (Note: to maintain full rank, each categorical covariate is encoded such that one of the levels is missing, except for treatment cell, and there is no intercept. As a result, the treatment cell coefficients cannot be interpreted as treatment means — they are the treatment means conditional on a specific value of each of the categorical covariates and of 0 for the continuous variables). Further note that for simplicity, we only include one set of interaction terms: treatment cell vs. the number of previous treatment responses. Finally, note that the displayed standard errors are cluster-robust standard errors where each client is a cluster, to take into account that ratings given by the same client are correlated. We learn several things from this regression, displayed in Table 2:

- There is some heterogeneity in ratings across the job covariates, but on the order of .1 points on the average rating. This heterogeneity is dwarfed by the differences between the treatment cells, especially the numeric vs. non-numeric treatments. This relative lack of heterogeneity further supports that the differences between the mean treatment responses are not due to randomness caused by some types of jobs being more present in some treatment groups than others.
- We can directly measure the effect of the number of previous jobs during that testing period a given client has submitted, i.e., estimate the inflation that will result over time as clients submit additional jobs. From the table below, each additional job a client has submitted raises the treatment response for the *Expectations* and the *Averages* treatments, on the order of .008 to .014 points per previous response. At this rate, these coefficients suggest that only after giving 100 ratings would a client inflate ratings by an average of between .8 and 1.4 points. The *Numeric* treatment cell does not further inflate substantially.

A.4. More on re-hire rates and treatment responses

Table 3 shows the mean response rate for each treatment overall, along with – for each first client-freelancer interaction during our test – the mean responses conditional on whether the freelancer was rehired. It also shows the in-sample RMSE for predicting rehires given just the treatment response in each cell.

Dep. Variable:	treatment-response	**R-squared:**	0.128
Model:	OLS	**Adj. R-squared:**	0.128
Method:	Least Squares	**Log-Likelihood:**	-1.6001e+05
No. Observations:	100438	**AIC:**	3.201e+05
Df Residuals:	100406	**BIC:**	3.204e+05
Df Model:	31		

	coef	std err	z	P>|z|	[0.025	0.975]
treatment_cell[1]	3.0596	0.062	49.052	0.000	2.937	3.182
treatment_cell[2]	3.3965	0.063	53.862	0.000	3.273	3.520
treatment_cell[3]	3.4516	0.062	55.353	0.000	3.329	3.574
treatment_cell[4]	3.4414	0.062	55.796	0.000	3.321	3.562
treatment_cell[5]	3.1887	0.062	51.379	0.000	3.067	3.310
treatment_cell[6]	4.3745	0.062	70.044	0.000	4.252	4.497
value_group[T.lv]	0.1031	0.034	2.998	0.003	0.036	0.170
value_group[T.mv]	0.0206	0.034	0.601	0.548	-0.047	0.088
value_group[T.vlv]	0.2920	0.032	9.061	0.000	0.229	0.355
category_group[T.Admin Support]	-0.0591	0.046	-1.281	0.200	-0.150	0.031
category_group[T.Customer Service]	-0.1070	0.081	-1.320	0.187	-0.266	0.052
category_group[T.Data Science & Analytics]	0.1177	0.050	2.354	0.019	0.020	0.216
category_group[T.Design & Creative]	0.1077	0.042	2.581	0.010	0.026	0.189
category_group[T.Engineering & Architecture]	0.1235	0.058	2.122	0.034	0.009	0.238
category_group[T.IT & Networking]	0.1277	0.049	2.595	0.009	0.031	0.224
category_group[T.Legal]	0.0643	0.061	1.047	0.295	-0.056	0.185
category_group[T.Sales & Marketing]	-0.0869	0.045	-1.920	0.055	-0.176	0.002
category_group[T.Translation]	0.0405	0.060	0.676	0.499	-0.077	0.158
category_group[T.Web, Mobile & Software Dev]	0.0940	0.042	2.256	0.024	0.012	0.176
category_group[T.Writing]	-0.1158	0.044	-2.638	0.008	-0.202	-0.030
expertise_tier[T.Expert/Expensive]	0.1465	0.020	7.276	0.000	0.107	0.186
expertise_tier[T.Intermediate]	0.0582	0.018	3.306	0.001	0.024	0.093
hr_charge	1.376e-05	2.16e-06	6.376	0.000	9.53e-06	1.8e-05
fp_charge	3.64e-05	6.73e-06	5.409	0.000	2.32e-05	4.96e-05
log(1 +client_prev_spend)	-0.0069	0.006	-1.156	0.248	-0.018	0.005
log(1 +num_prev_asg)	-0.0177	0.010	-1.769	0.077	-0.037	0.002
treatment_cell[1]:# prev. treatment resps. by client	0.0080	0.004	2.042	0.041	0.000	0.016
treatment_cell[2]:# prev. treatment resps. by client	-0.0043	0.006	-0.675	0.500	-0.017	0.008
treatment_cell[3]:# prev. treatment resps. by client	0.0085	0.003	2.850	0.004	0.003	0.014
treatment_cell[4]:# prev. treatment resps. by client	0.0141	0.003	5.468	0.000	0.009	0.019
treatment_cell[5]:# prev. treatment resps. by client	0.0024	0.005	0.485	0.628	-0.007	0.012
treatment_cell[6]:# prev. treatment resps. by client	0.0010	0.004	0.246	0.806	-0.007	0.009

Omnibus:	11064.189	**Durbin-Watson:**	1.911
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	15421.891
Skew:	-0.876	**Prob(JB):**	0.00
Kurtosis:	3.785	**Cond. No.**	9.60e+04

Table 2 OLS Regression Results with covariate for previous number of treatment responses

	Mean treatment response			Re-hired prediction RMSE	
Condition	All responses	Not Re-hired	Re-hired	In-sample	Out-of-sample
Expectation	3.34	3.16	3.71	0.335	0.336
Adjectives	3.65	3.44	4.03	0.337	0.335
Average	3.76	3.53	4.14	0.328	0.331
Average, not affect score	3.78	3.54	4.16	0.329	0.330
Average, randomized	3.46	3.28	3.76	0.328	0.323
Numeric	4.59	4.45	4.85	0.334	0.346

Table 3 Mean treatment response in each cell, and in-sample and out-of-sample RMSE for predicting rehires given just the treatment response in each cell. For out-of-sample RMSE, 80% of clients are used to calculate the re-hire rates and the remaining 20% are used to test.

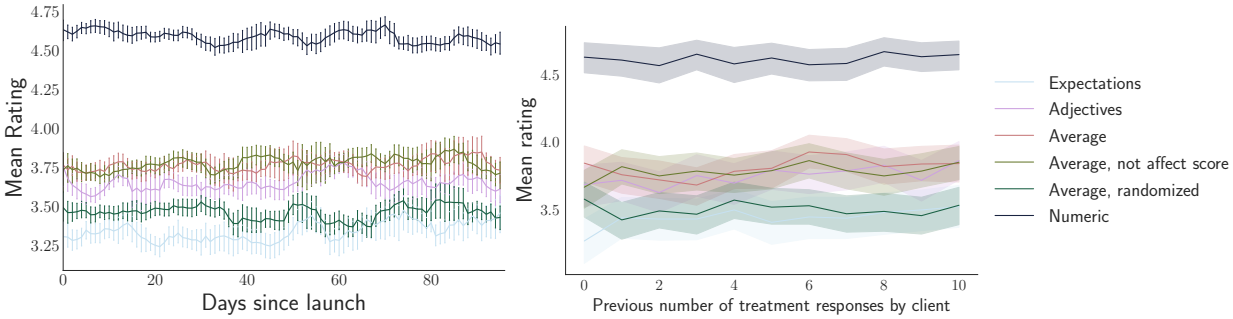
Figure 2 shows the rehire rate, conditional on the rating given for each treatment cell.

A.5. More on inflation over time

We find that the rating scales do not inflate substantially. Figure 3a shows the average rating per treatment group over the 90 days after the launch of the test, in a sliding window of 7 days. There is no discernible inflation over time. It is instructive to compare the (lack of) inflationary trend to the inflation after the launch of a new numeric scale on the same platform in 2007, as reported by Filippas et al. (2019): average ratings inflated from about 3.8 stars to about 4.1 stars in the first three months after the system launched. (In contrast, introducing a new *Numeric* question in 2018 yields immediately inflated responses, suggesting that current platform users have been conditioned to the norm of inflated ratings.)

We also repeat the regression in Table 2, limiting the analysis to those clients who have more than ten treatment responses during the test period (all of which have the job covariates). The table is omitted; the coefficients for inflation over time are largely the same. To further help visualize (the relative lack of) inflation over the number of submitted ratings, Figure 3b shows the mean ratings for each treatment cell by

the number of previous treatment responses given during the test period. As the plot has no covariate data, we use the first ten responses for all 2145 clients who submitted at least 10 ratings during the test period. Clients are not substantially more likely to give more positive ratings on their tenth rating during the test than they give on their first rating.



(a) Mean ratings in 7 day sliding windows

(b) Mean ratings by number of previous treatment responses in the test period.

Figure 3 **More information for inflation during the test period. Error bands are bootstrapped 95% confidence intervals.**

A.6. Analysis of cell with randomized order of answer choices

The *Average, Randomized* contained the same question and answer choices as the *Average* condition, but the choices were presented in a random order. If the raters read all the answer choices and pick the most applicable one, then this condition would have returned a rating distribution identical to that of the *Average* condition. However, it does not. Furthermore, the *location* of the chosen choice would be distributed uniformly, i.e., the rater should pick the choice presented first no more frequently than the other choices. We find this not to be the case: the first answer choice presented to the rater is picked $6806/26978 = 25.2\%$ of the time. The second through sixth answer choices are picked $17.3\%, 14.7\%, 14.3\%, 13.9\%$, and 14.5% of the time each, respectively.

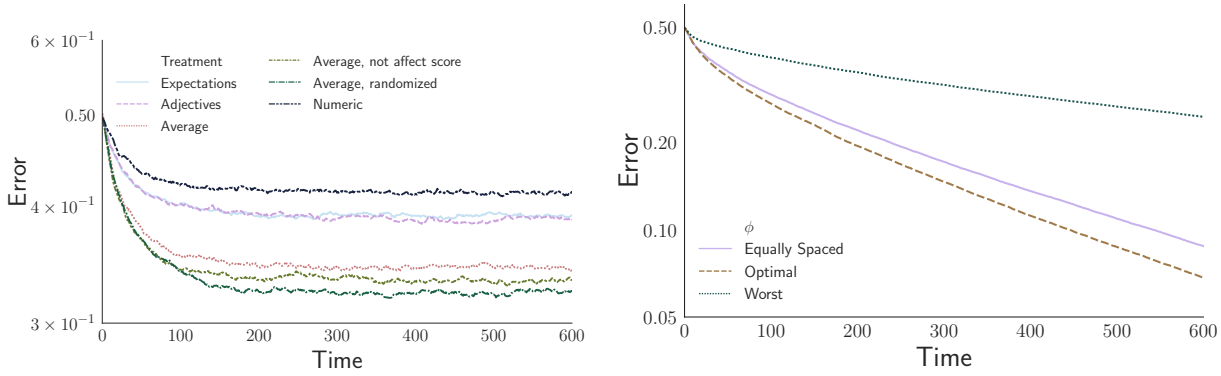
This phenomenon suggests that (a) a small percentage (up to $10 - 13\%$) of raters do not read the answer choices at all and simply select the first answer choice, and (b) many raters start reading from the first presented choice and select the first one that approximately describes their experience. Our test design cannot disambiguate between these (or other plausible) explanations. Nevertheless, this effect is second-order relative to the overall finding that more descriptive scales are substantially more informative than numeric scales, and the *Average, Randomized* treatment results are comparable to those of other verbal scales.

A.7. Design approach using labor market data

A.7.1. Supplementary information for design approach application Table 4 and Figure 4 contain information for our design approach applied to the labor market data, as described in the main text.

	Response Score					
Condition	**0**	**1**	**2**	**3**	**4**	**5**
Expectations	1.16	3.63	3.71	3.72	3.99	5.0
Adjectives	0.292	2.21	2.26	4.07	4.15	5.0
Average	1.61	2.14	3.62	3.71	4.91	5.0
Average, not affect score	1.78	1.96	2.03	3.23	3.53	5.0
Average, randomized	1.93	3.05	4.22	4.54	4.54	5.0
Numeric	1.31	1.33	1.44	1.45	3.49	5.0

Table 4 **Optimal scores ϕ for each treatment, where the score of the top position is normalized to 5, using freelancer quality as in the main text.**



(a) With optimal ϕ and prob. of exit of 0.01. (b) *Average* treatment with different ϕ

Figure 4 **Simulated performance over time with various other configurations. The "Worst" scoring rule corresponds to the rule ϕ with the smallest learning rate found for each treatment.**

	Learning rates		Response Score					
Condition	**Naive ϕ**	**Optimal ϕ**	**0**	**1**	**2**	**3**	**4**	**5**
Expectations	0.022	0.024	1.22	1.22	2.28	3.74	4.38	5.00
Adjectives	0.021	0.027	1.47	1.55	1.63	3.22	4.97	5.00
Average	0.023	0.026	1.80	1.84	1.88	2.53	3.83	5.00
Average, not affect score	0.013	0.014	0.89	1.57	1.59	3.32	4.04	5.00
Average, randomized	0.014	0.019	0.72	2.41	2.63	4.18	4.30	5.00
Numeric	0.009	0.009	0.50	1.20	1.98	2.88	3.45	5.00

Table 5 **We apply and test our design approach using experimental data from our online labor market, using freelancer quality as in Section A.7.2.**

A.7.2. Alternate ways to define freelancer quality – mean treatment responses in *other* cells

We now repeat the approach using an alternate way to define freelancer quality. Results are broadly similar, with all scales outperforming the numeric one. However, differences suggest that the full approach strongly depends on the true quality estimates.

Recall that in our experiment design, a client is only in a single treatment cell throughout the test period, and freelancers may complete jobs with and receive ratings from clients across treatment cells. Thus the ratings any freelancer receives in different treatment cells are independent (given by different clients who do not see each others' ratings). We can leverage this independence to construct approximate joint distributions of freelancer quality and ratings in each treatment cell as follows. For a given treatment cell, we consider all freelancers who received at least three ratings in the *other* treatment cells, and we estimate a freelancer's quality via a simple average of these ratings. For each given treatment cell, these estimates of quality are exogenous to the ratings received in that cell. For each treatment cell, we then construct a joint distribution over freelancers of the rating received in that cell, and the estimated quality of that freelancer.

We consider all freelancers who received at least three ratings in the *other* treatment cells, and estimate a freelancer's quality via a simple average of these ratings. For each treatment cell, these estimates of quality are exogenous to the ratings received in that cell. For each treatment cell, we then construct a joint distribution over freelancers of the rating received in that cell, and the estimated quality of that freelancer. Then, *Low*, *Medium*, and *High* quality sellers refer to those with other cell average ratings in [0,2),[2.5,3.5) and [4.5,5], respectively. Table 5 shows the results from applying our design approach to the resulting qualities.

Appendix B: Amazon Mechanical Turk synthetic experiment

In this section, we deploy an experiment on Amazon Mechanical Turk ("MTurk") to repeat and analyze our design approach, in a synthetic setting where we have expert (external) quality information on items. We note that this section is not a replication of the behavioral components of our results, as the MTurk and online labor market settings are too different to meaningfully compare. Furthermore, one should be aware of limitations of using MTurk convenience samples in research (Landers and Behrend 2015); such limitations

	Learning rates	
Condition	**Equally spaced ϕ**	**Optimal ϕ**
Every Other	0.058	0.069
Extremes	0.077	0.079
Negative-skewed	0.051	0.059
Positive-skewed	0.034	0.043
Close to Every Other	0.043	0.044

Table 6 Large deviations learning rates for each treatment in the Mturk experiment, calculated using Equation (4) and the joint distributions generated using the training data plotted in Figure 5. Optimal for each treatment corresponds to the highest learning rate among many random score functions tested.

mean that there will be behavioral biases that differ from those on other platforms. For these reasons, this section should be seen as a synthetic, example application of our overall comparison and design methodology to other domains, and in particular will show how our methods are useful not just to counter rating inflation but also other types of biases. This appendix section is organized as follows. In B.1 we describe the task, and in B.2 we repeat our analysis from the main text, including: showing the resulting marginal and joint distributions of ratings and quality, and testing designs on new, unseen data.

B.1. Experiment description

B.1.1. Task Information We asked subjects to rate the English proficiency of 10 paragraphs which are modified TOEFL (Test of English as a Foreign Language) essays with known scores as determined by experts and reported in a TOEFL study guide (Educational Testing Service 2005); these are our true quality types for each essay. Expert scores range from 1 through 5, with two paragraphs with each score. Essays are shortened to a single paragraph of just a few sentences, and the top rated paragraphs are improved and the worst ones are made worse; this is largely to ensure the quality could be sufficiently distinguished between paragraphs despite having shortened them. In other words, for each topic, we improved the language of the best rated paragraph and further degraded the language of the worst one. In principle, our editing of these paragraphs may remove the validity of the expert ratings. However, the estimated $R(\theta, y|Y)$ indicates that this does not substantially occur, suggesting our editing of the paragraphs preserved the quality ordering of the paragraphs per the expert ratings.

Subjects were given one of five possible verbal scales, where the scales were designed using a list of adjectives, {*Abysmal, Awful, Bad, Poor, Mediocre, Fair, Good, Great, Excellent, Phenomenal*}, compiled by Hicks et al. (2000). Each scale had five options. The scales are:

- **Every Other**: *Awful, Poor, Fair, Great, Phenomenal*
- **Close to Every Other**: *Abysmal, Poor, Mediocre, Good, Phenomenal*
- **Extremes**: *Abysmal, Awful, Bad, Excellent, Phenomenal*
- **Negative-skewed**: *Abysmal, Awful, Bad, Poor, Mediocre*
- **Positive-skewed**: *Fair, Good, Great, Excellent, Phenomenal*

We note that it is not a priori clear which of these scales will perform well in this setting, or what the optimal scoring mapping should be.

Raters (i.e., mTurk workers) were shown each of the ten paragraphs. The instructions were: "Please rate on English proficiency (grammar, spelling, sentence structure) and coherence of the argument, but not on whether you agree with the substance of the text." The specific question then asked was: "How does the following rate on English proficiency and argument coherence?" One paragraph was shown per page; returning to modify a previous answer was not allowed; and paragraphs were presented in a random order. Each rater was shown one of the scales picked at random, and the same scale was used for all paragraphs for that rater. There were approximately 500 raters overall across the 5 treatment cells, with between 97 and 104 raters in each cell. For each cell, we divide the raters (randomly) into train (75%) and test (25%). We design optimal scoring rules using the training data, and then test performance on the test data.

B.1.2. Rater logistics We did not exclude any data, and all raters were paid $1.50. Instructions advised raters to spend no more than a minute per question, though this was not enforced. The median rater spent 325 seconds, corresponding to a median wage of $16.61/hr. About 80% of raters spent 8 minutes or less.

B.2. Results

We now repeat the design and test procedure from the main text, for this setting. All plots, figures, and scoring rules are generated exactly as in the main text, with the following exceptions: (1) we have true expert

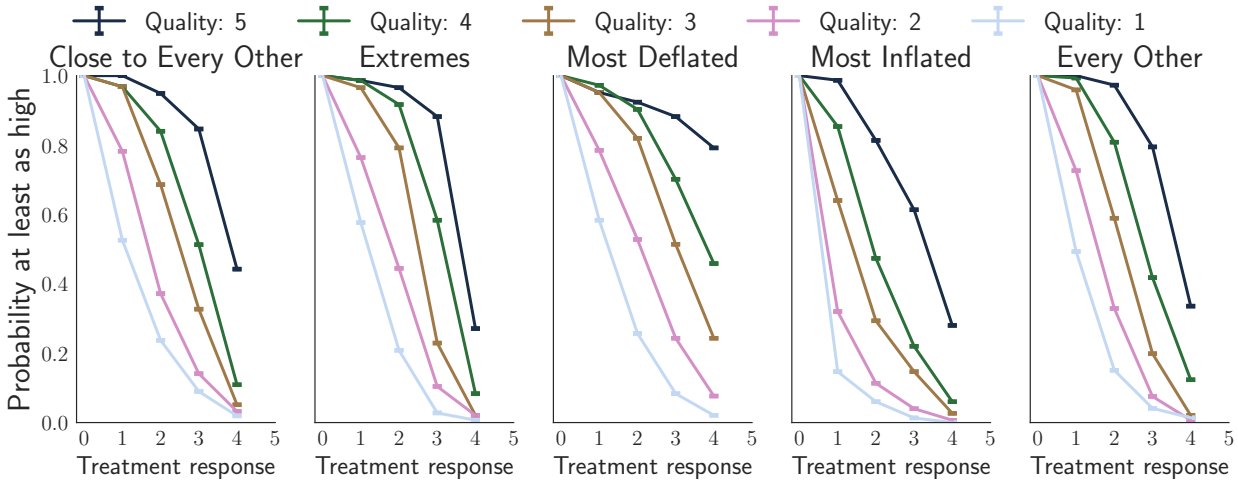
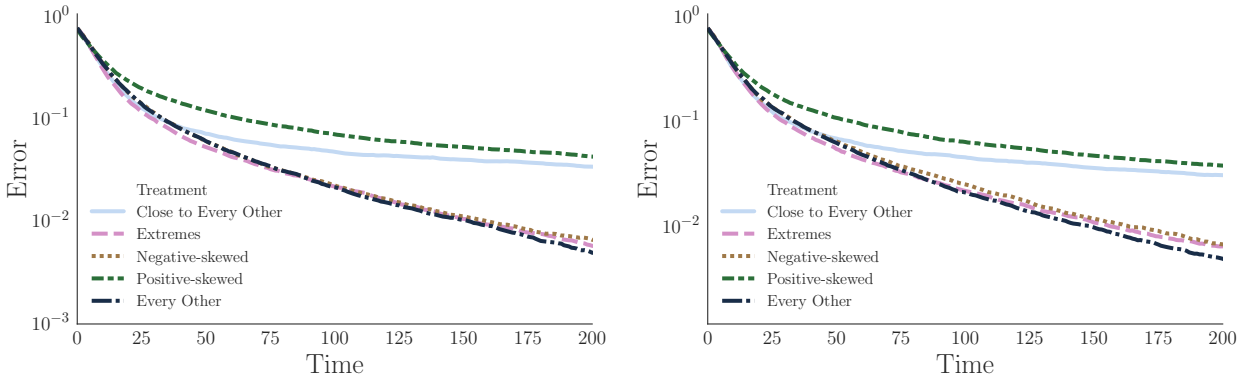


Figure 5 **Joint distributions of rating and expert score on the MTurk training set, by treatment condition.**



(a) Training data (b) Test data

Figure 6 **Simulated performance of each rating scale with Equally Spaced scores.**

scores for the paragraph qualities and so do not use the procedure where we estimate such qualities from the other treatment cells, and (2) we split the rater responses into test and training sets. We show the joint distributions $R(\cdot|Y)$ and optimal scores calculated from the training set, and then we evaluate performance on simulations using the test set.

Figure 5 shows the joint distributions of rating and expert score on the MTurk training set, for each treatment condition.Table 6 shows the training set learning rates for each treatment using equally spaced scores, as well as the best performing scores, respectively.

Finally, we simulate the performance of the designs (generated using the training data), following the same simulation technique as outlined in the main text. We also evaluate performance on the test data, in order to demonstrate how a platform would use our design approach. Figure 6 shows the resulting errors over time with Equally Spaced scores. Errors with optimal scores are qualitatively similar.

Appendix C: Proofs

LEMMA 1.

$$\lim_{k\to\infty} -\frac{1}{k}\log\left[\mu((x_k(\theta_1)-x_k(\theta_2))\leq 0|\theta_1,\theta_2)\right] = \inf_{a\in\mathcal{R}}\left\{g(\theta_1)I(a|\theta_1)+g(\theta_2)I(a|\theta_2)\right\}$$

where $I(a|\ell)=\sup_z\{za-\Lambda(z|\theta)\}$, $\Lambda(z|\theta)$ *is the log moment generating function of a single sample from* $x(\theta_1)$, *and* $g(\theta)$ *is the sampling rate.*

Proof. $\lim_{k\to\infty} -\frac{1}{k}\log\left[\mu((x_k(\theta_1)-x_k(\theta_2))\le 0|\theta_1,\theta_2)\right]$

$$= \lim_{k\to\infty} -\frac{1}{k}\log\left[\int_{a\in\mathcal{R}} \mu((x_k(\theta_1)=a|\theta_1)\mu(x_k(\theta_2)\ge a|\theta_2)da\right] \tag{1}$$

$$= \lim_{k\to\infty} -\frac{1}{k}\log\left[\int_{a\in\mathcal{R}} e^{-kg(\theta_1)I(a|\theta_1)}e^{-kg(\theta_2)I(a|\theta_2)}da\right] \tag{2}$$

$$= \inf_{a\in\mathcal{R}}\{g(\theta_1)I(a|\theta_1)+g(\theta_2)I(a|\theta_2)\} \qquad \text{Laplace principle} \tag{3}$$

Where (2) is a basic result from large deviations, and $kg(\theta_i)$ is the number of samples item of quality θ_i has received. □

This lemma also appears in Glynn and Juneja (2004), which uses the Gartner-Ellis Theorem in the proof. Our proof is conceptually similar but instead uses Laplace's principle.

We can now establish the rate function for $P_k(\theta_1,\theta_2)$.

Recall $P_k(\theta_1,\theta_2)=\mu_k(x_k(\theta_1)>x_k(\theta_2)|\theta_1,\theta_2)-\mu_k(x_k(\theta_1)<x_k(\theta_2)|\theta_1,\theta_2)$. Then, we have

LEMMA 2. *Given θ_1,θ_2, let $\overline{P}_k(\theta_1,\theta_2)=1-P_k(\theta_1,\theta_2)$. Then:*

$$-\lim_{k\to\infty}\frac{1}{k}\log\overline{P}_k(\theta_1,\theta_2)=\inf_{a\in\mathcal{R}}\{g(\theta_1)I(a|\theta_1)+g(\theta_2)I(a|\theta_2)\}, \tag{4}$$

where $I(a|\theta)=\sup_z\{za-\Lambda(z|\theta)\}$, and $\Lambda(z|\theta)$ is the log moment generating function of a single rating given to seller of type θ:

$$\Lambda(z|\theta)=\log\sum_{y\in Y}\rho(\theta,y|Y)\exp(z\phi(y)).$$

Proof. Follows directly from Lemma 1.

$$-\lim_{k\to\infty}\frac{1}{k}\log\overline{P}_k(\theta_1,\theta_2|\beta)$$

$$= \lim_{k\to\infty} -\frac{1}{k}\log\left[1+\mu_k(x_k(\theta_1)-x_k(\theta_2)<0|\theta_1,\theta_2)-\mu_k(x_k(\theta_1)-x_k(\theta_2)>0|\theta_1,\theta_2)\right]$$

$$= \lim_{k\to\infty} -\frac{1}{k}\log\left[2\mu_k(x_k(\theta_1)-x_k(\theta_2)<0|\theta_1,\theta_2)+\mu_k(x_k(\theta_1)-x_k(\theta_2)=0|\theta_1,\theta_2)\right]$$

$$= \inf_{a\in\mathcal{R}}\{g(\theta_1)I(a|\theta_1)+g(\theta_2)I(a|\theta_2)\} \qquad \text{Lemma 1}$$

□

Now we show that this rate function transfers to a rate function for W_k.

Proof of Theorem 1

$$r\triangleq -\lim_{k\to\infty}\frac{1}{k}\log(1-W_k)=\min_{0\le i<M}\inf_{a\in\mathbb{R}}\{g(\theta_{i+1})I(a|\theta_{i+1})+g(\theta_i)I(a|\theta_i)\} \tag{5}$$

where $I(a|\theta)=\sup_z\{za-\Lambda(z|\theta)\}$, and $\Lambda(z|\theta)=\log\sum_{y\in Y}\rho(\theta,y|Y)\exp(z\phi(y))$ is the log moment generating function of a single rating given to seller of type θ.

Proof.

$$-\lim_{k\to\infty}\frac{1}{k}\log(1-W_k) = -\lim_{k\to\infty}\frac{1}{k}\log\left(1-\frac{2}{M(M-1)}\sum_{\theta_1>\theta_2\in\Theta}P_k(\theta_1,\theta_2)\right) \tag{6}$$

$$= -\lim_{k\to\infty}\frac{1}{k}\log\frac{2}{M(M-1)}\sum_{0\le i<j\le M}\overline{P}_k(\theta_j,\theta_i) \tag{7}$$

$$= -\max_{0\le i<j\le M}\left(\lim_{k\to\infty}\frac{1}{k}\log\left(\overline{P}_k(\theta_j,\theta_i)\right)\right)=\min_{0\le i<j\le M}\left(-\lim_{k\to\infty}\frac{1}{k}\left[\log\left(\overline{P}_k(\theta_j,\theta_i)\right)\right]\right) \tag{8}$$

$$= \min_{0\le i<j\le M}\inf_{a\in\mathcal{R}}\{g(\theta_j)I(a|\theta_j)+g(\theta_i)I(a|\theta_i)\} \tag{9}$$

$$= \min_{0\le i<M}\inf_{a\in\mathcal{R}}\{g(\theta_{i+1})I(a|\theta_{i+1})+g(\theta_i)I(a|\theta_i)\} \tag{10}$$

Where the last line follows from adjacent θ_i,θ_{i+1} dominating the rate due to properties of R. Line (8) follows from: $\forall a_i^\epsilon\ge 0$, $\limsup_{\epsilon\to 0}\left[\epsilon\log\left(\sum_i^N a_i^\epsilon\right)\right]=\max_i^N\limsup_{\epsilon\to 0}\epsilon\log(a_i^\epsilon)$. See, e.g., Lemma 1.2.15 in Dembo and Zeitouni (2010) for a proof of this property.

□