Online Appendixes

A Jobs postings that ended with vs. without a successful hire

The observational study presented in the main manuscript focuses on jobs that ended with successful hires after the employer posted the job and waited for freelancers to submit their applications (approx. 25% of the jobs posted on the platform). In Appendix B, we also analyze cases where employers directly extended invitations to specific freelancers (approx. 25% of the jobs on the platform, with each invitation considered as a job; thus, if an employer invites ten freelancers to apply for the job, it is considered as ten jobs by the platform). In other cases, jobs are canceled after being posted (approx. 15% of the jobs posted on the platform) or expired without a successful hire (approx. 35% of the jobs posted on the platform). In the following, we explore possible reasons as to why some job postings end with a successful hire while others do not.

One possible explanation is that the employers or the jobs are systematically different. Table A1 shows that this argument seems plausible: Jobs postings that end without a successful hire are far more likely to be from employers with zero reviews, who neither verify their payment method nor make a deposit, and are four times more likely to have a budget above $1,000, a rather unusual amount for typical jobs on the platform. Such differences suggest that these employers might be simply testing the platform without serious intention to hire a freelancer (especially considering that jobs can be posted for free), or they might also consider other outside options (e.g., for jobs with large budgets).

Another possible explanation for why some job postings end with a successful hire while others do not, which could potentially confound our key finding, is that the pool of applicants that these two types of jobs receive is systematically different, particularly around the "look the part" variable. Table A1 shows there are some significant differences between applicants in the two groups: the former receives a smaller and "better" pool of applicants than the

latter (e.g., with a higher number of reviews, more likely to have certifications, and higher perceived job fit score from profile pictures). Despite these small differences, jobs that ended without a hire still received a pool of applicants with "good" attributes, and we conjecture that such differences are not the main reason for ending the job postings without a successful hire.

Table A1: Comparing characteristics of jobs postings that ended with and without a successful hire

	Avg. among jobs listings that end:	
	With a successful hire	Without a successful hire
Employer characteristics:		
Employer has reviews	0.950	0.366
Employer verified payment	0.905	0.408
Employer made deposit	0.981	0.437
Employer completed his/her profile	0.335	0.284
Job characteristics:		
Budget is above $1,000	0.025	0.113
Job posting is potentially duplicated	0.014	0.075
Job description word count	59.380	51.765
Applicants characteristics:		
Number of applications received	31.637	32.281
% Applicants with zero reviews	0.179	0.272
Applicants' Avg. log(1+reviews)	3.487	3.135
Applicants' Avg. rating	4.839	4.830
Applicants' Avg. perceived job fit	0.616	0.592
% Applicants with certification	0.106	0.092

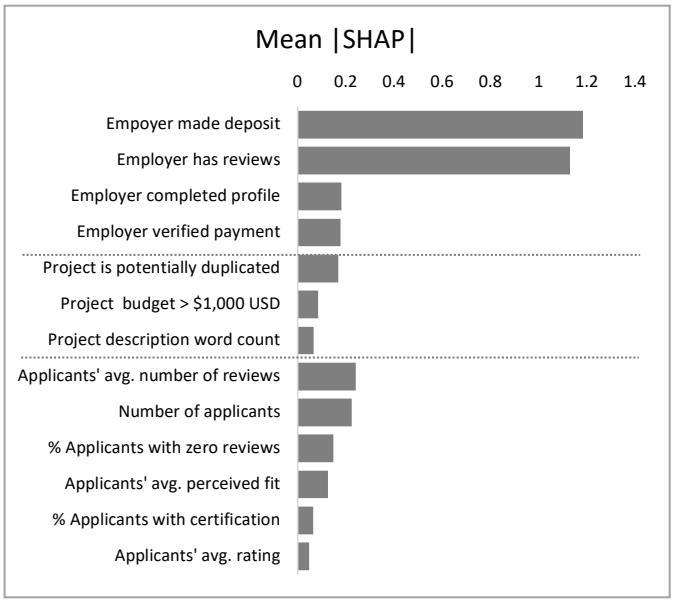
Note: We consider a job posting is potentially duplicated if it was posted by the same employer on the same day and has a similar title (more than 70% characters in common). All differences are statistically different from zero (p-value of t-test<0.01).

To formally explore which of the variables discussed above seems to be more likely to explain why jobs postings end without a successful hire, we use the XGBoost classifier to predict final job status (i.e., ended with or without a successful hire) using three groups of variables:[1] characteristics of (i) the employer, (ii) the job, and (iii) the pool of applicants the

[1]We fine-tune the parameters using random search. We split the sample in 80% training and 20% validation and set the following hyperparameters to train the model: max. number of boosting iterations = 50, max. depth of a tree = 6, min. number of nodes in a leaf = 2, fraction of features used to build each tree = 1, fraction of observations used to build each tree = 0.6, min. loss reduction to further partition on

job receives. In Figure A1, we illustrate the importance of each variable based on the Shapley values (SHAP). Our findings confirm our conjecture that employers' and jobs' characteristics are more important than those of applicants in predicting whether a job posting ends with or without a successful hire. More importantly, our main variable of interest, perceived job fit from profile pictures, contributes very little to this prediction compared to other variables. Therefore, we believe that focusing on jobs that end with a successful hire does not invalidate the findings presented in the main manuscript.

Figure A1: Feature importance in predicting whether a job listing ends with or without a successful hire



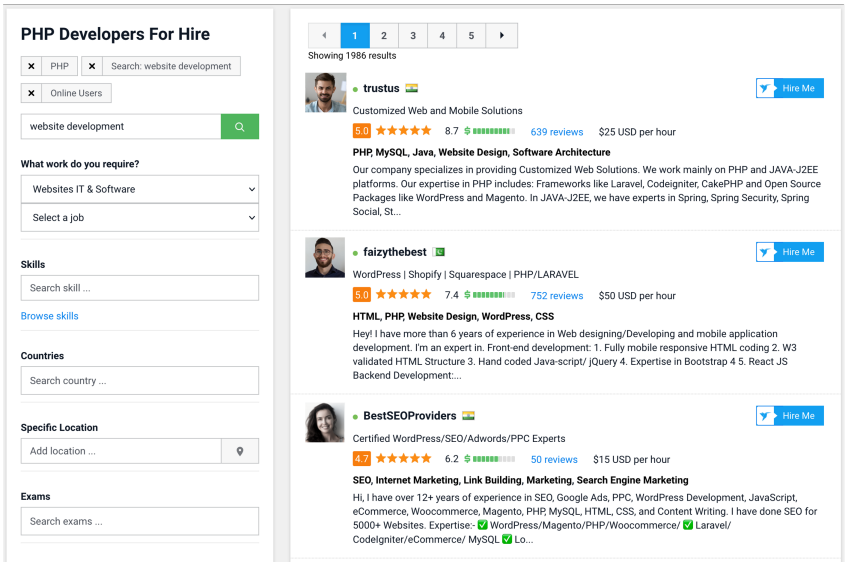
B Perceived job fit and hires from direct invitations

In this section, we analyze the outcomes from an alternative recruiting option available on the platform, which allows employers to search for freelancers using different criteria (job category, skills, hourly rate, etc.) and directly invite them for the focal job. As illustrated in Figure A2, when using this recruiting option, employers can also see a summary of each

a leaf node = 0.1, learning rate = 0.3, and regularization term on weights = 0.0. As a result, we obtained a 88.211% in-sample ROC-AUC and 89.520% out-of-sample ROC-AUC.

freelancer's reputation variables (number of reviews and average rating) along with his/her profile picture, and can see more information by clicking on their profile (certifications, review history, etc.). For these cases, we cannot observe how employers interact with the platform (i.e., which freelancers the platform displays to each employer, how many freelancers employers inspect, etc.). We can, however, observe the number of invitations each freelancer in the sample receives and the percentage that translates into hires.

Figure A2: Example of alternative recruiting option: Search for and invite freelancers



Note: We replace the original pictures with licensed images purchased from an online stock photography company called Shutterstock for illustration purposes. Note that the perceived job fit based on these pictures exhibits much less variation from actual freelancer profile pictures available at Freelancer.com.

To explore the role of perceived job fit in this alternative recruitment option, we run two regressions using the percentage of times that a direct invitation received by the focal freelancer turns into a hire as the dependent variable. We present our results in Table A2, with the two columns corresponding to a different set of controls. The results suggest that perceived job fit is positively and significantly associated with the percentage of direct invitations freelancers receive that translate into hires. For instance, the estimates from column 2 suggest that a 1 SD increase in perceived job fit (0.35 points) is associated with a 1.439

percentage point (0.351×0.041) increase in the percentage of direct invitations that translate into hires. These findings are consistent with the idea that freelancers who "look the part" are more likely to be hired.

Table A2: Estimating hires from direct invitations

	(1)	(2)
Profile Pictures Variables:		
Perceived Job Fit	0.056^{***}	0.041^{***}
Has Picture	0.000	-0.004
Reputation Variables:		
No Reviews Yet	-0.200^{***}	-0.198^{***}
Log (1 + N. Reviews)	0.017^{***}	0.015^{***}
Avg. Rating	0.015^{**}	0.015^{**}
Additional Variables:		
Certifications	✓	✓
Human, Demographics, and Beauty		✓

Note: OLS estimates. The dependent variable is the percentage of direct invitations from employers that turned into successful hires for freelancer j.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

C Validating API labels

To validate the quality of the outputs from the Cloud Vision APIs, we use a subsample of 100 images and recruit human raters through Amazon Mechanical Turk to provide the same labels. Raters were presented with 20 images selected at random. Given that this classification task is relatively straightforward, each image was labeled by three different raters. In case of disagreement, we label the image based on the majority of votes.

In Table A3, we present the percentage of the agreement between the categorical gender and race labels provided by the Cloud Vision APIs and those provided by MTurkers.

We followed a similar procedure to validate the continuous beauty scores, and obtained a 0.721 correlation between the scores provided by the Cloud Vision APIs and those provided by MTurkers.

Table A3: Agreement between image labels provided by Cloud Vision APIs and MTurkers

Variable	Labels	Percentage of agreement
Human	Yes, No	94.62%
Gender	Female, Male	83.33%
Race	Far East Asian, Black, Indian, White	63.88%

D Labeling perceived job fit: Details on the deep learning image classifier

Below we provide additional details on how we create our training dataset, including a description of the architecture, hyperparameters used for our classification tasks (i.e., perceived job fit as a programmer and as a graphic designer), and additional metrics on the performance of our classifiers (i.e., precision, specificity, and recall). We also summarize the results of two additional checks that explore the robustness of our perceived job fit measure.

Training data To create our training dataset, we drew a random sample of 3,000 profile pictures and asked raters with experience in each job category to score these images based on their perceptions of the freelancer's fit for a job in each category. We recruit raters with experience in the programming job category through Amazon Mechanical Turk using premium qualifications to target respondents employed in the "Information Technology" industry. Premium qualifications reflect self-reported information about Mturker workers and provide a venue to screen workers with specific qualifications (in our case, specific employment industry). For the graphic designer job category, we recruited graphic designers from Upwork because we could not find a premium qualification that accurately reflects this job category in Amazon Mechanical Turk.

Raters were asked to use a 5-point Likert scale, wherein 1 is the lowest perceived job fit and 5 is the highest perceived job fit, and each image was rated by three to five independent raters for each job category. Following standard practices in the literature (Zhang et al. 2015, Liu et al. 2019, Zhang et al. 2021, Zhang and Luo 2022), we convert the 5-point Likert

scale to binary levels (low and high) to mitigate potential noise in the training data. More specifically, as in Zhang et al. (2021), for each image i we take the mean $score_i$ averaged across the raters, and define two thresholds $\theta_1 = \overline{score} - gap/2$ and $\theta_2 = \overline{score} + gap/2$, where $\overline{score}$ is the average score for all the images in the subsample, and $gap = 0.8$. Finally, a profile picture i is labeled as *low perceived job fit* if $score_i < \theta_1$ or as *high perceived job fit* if $score_i > \theta_2$. We discard images with $score_i \in (\theta_1, \theta_2)$ from the training sample, which leaves us with 1,521 training samples for the perceived job fit as a designer task (46.942% labeled as high), and 2,174 samples for the perceived job fit as a programmer task (55.198% labeled as high).

Architecture We use transfer learning to fine-tune various Convolution Neural Networks (CNN), including VGG-16, ResNet, and Inception (Canziani et al. 2016). We focus our discussion around VGG-16, which provided the most accurate and stable results in our setting.

Our final architecture is the modified version of the VGG-16 CNN as in Hartmann et al. (2021). We *freeze* the first four convolutional blocks, because their layers extract generic information or low-level features, such as contours, textures, and colors, that can serve for a wide range of classification tasks (Hartmann et al. 2021, Zhang et al. 2021). We initialize the model weights with pre-trained weights and fine-tune the parameters of the last convolutional block, which consists of three convolutional layers followed by a max-pooling layer. We then add three two connected layers, where the last layer is the output layer.

We illustrate the final architecture in Table A4. Note that, by freezing the first convolutional blocks we are training a smaller proportion (57%) of the total number of parameters in the original VGG architecture. Moreover, the majority of the trainable parameters (69%) are fine-tuned with pre-trained weights.

Hyper-parameters We use the Adadelta algorithm for optimization, a method that dynamically adapts learning rates and has been shown to be robust to noisy gradient informa-

Table A4: VGG architecture (modified and tuned)

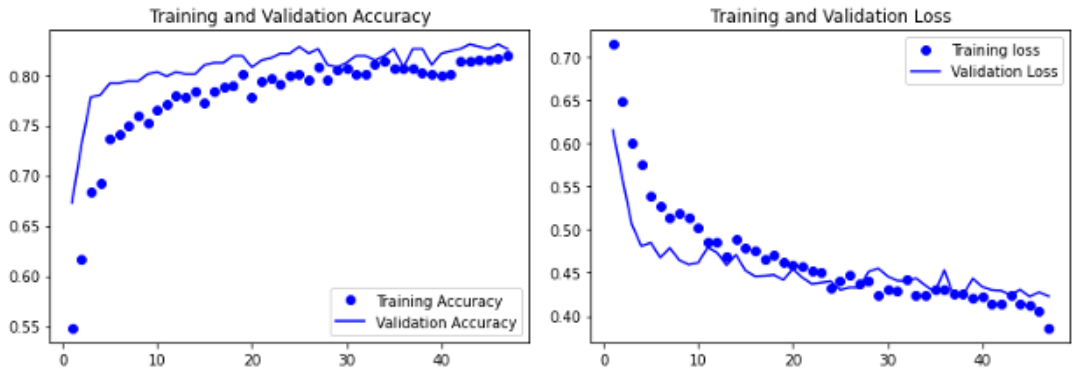
		Number of Parameters	
Layer	Output Shape	Total	Trainable
Input	(224, 224, 3)	0	0
Convolutional Block 1			
Convolutional Layer 1.1	(224, 224, 3)	1,792	0 (frozen)
Convolutional Layer 1.2	(224, 224, 3)	36,928	0 (frozen)
MaxPooling Layer 1	(112, 112, 64)	0	0
Convolutional Block 2			
Convolutional Layer 2.1	(112, 112, 128)	73,856	0 (frozen)
Convolutional Layer 2.2	(112, 112, 128)	147,584	0 (frozen)
MaxPooling Layer 2	(56, 56, 128)	0	0
Convolutional Block 3			
Convolutional Layer 3.1	(56, 56, 256)	295,168	0 (frozen)
Convolutional Layer 3.2	(56, 56, 256)	590,080	0 (frozen)
Convolutional Layer 3.3	(56, 56, 256)	590,080	0 (frozen)
MaxPooling Layer 3	(28, 28, 256)	0	0 (frozen)
Convolutional Block 4			
Convolutional Layer 4.1	(28, 28, 512)	1,180,160	0 (frozen)
Convolutional Layer 4.2	(28, 28, 512)	2,359,808	0 (frozen)
Convolutional Layer 4.3	(28, 28, 512)	2,359,808	0 (frozen)
MaxPooling Layer 4	(14, 14, 512)	0	0
Convolutional Block 5			
Convolutional Layer 5.1	(14, 14, 512)	2,359,808	2,359,808 (fine-tuned)
Convolutional Layer 5.1	(14, 14, 512)	2,359,808	2,359,808 (fine-tuned)
Convolutional Layer 5.3	(14, 14, 512)	2,359,808	2,359,808 (fine-tuned)
MaxPooling Layer 1	(7, 7, 512)	0	0
Flatten Layer	25,088	0	0
Droput Layer	25,088	0	0
Fully Connected Layer	128	3,211,392	3,211,392
Output Layer (Prediction)	1	129	129
Total	-	17,926,209	10,290,945

tion and selection of hyper-parameters (Zeiler 2012). We use batch size equal to 16, number of epochs equal to 100 with early stopping, and a binary cross-entropy loss function. We use ReLU activation for our convolutional layers, and SoftMax activation for the output layer. For data augmentation, we allow for horizontal flips, zoom range, width range, and height range equal to 0.2, and rotation range equal to 15 degrees.

Performance metrics and loss curve In Table A5, we show different performance metrics on the validation sample for both prediction tasks, namely, perceived job fit as a programmer and as a designer (columns 1 and 4, respectively). Without loss of generality, images labeled "low perceived job fit" are used as the positive examples and those labeled "high perceived job fit" are used as the negative examples during training. Overall, the classifiers provide reasonably accurate and balanced performance.

In Figures A3 and A4, we show the accuracy and loss curve across training iterations. Visually, we see no concerning signs of over-fitting (i.e., the training curves are not better than the validation curves).

Figure A3: Training and validation accuracy and loss curves for programmer labels



In Figure A5, we provide a histogram of the predicted perceived job fit scores for all the freelancers in our sample.

Additional robustness checks Our first robustness check aims to further mitigate over-fitting concerns by reducing the number of trainable parameters in the model. We implement

Figure A4: Training and validation accuracy and loss curves for graphic designer labels

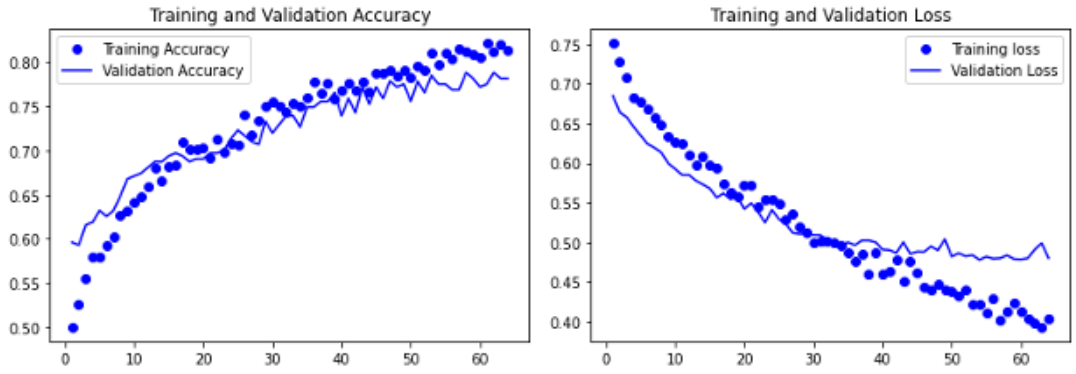
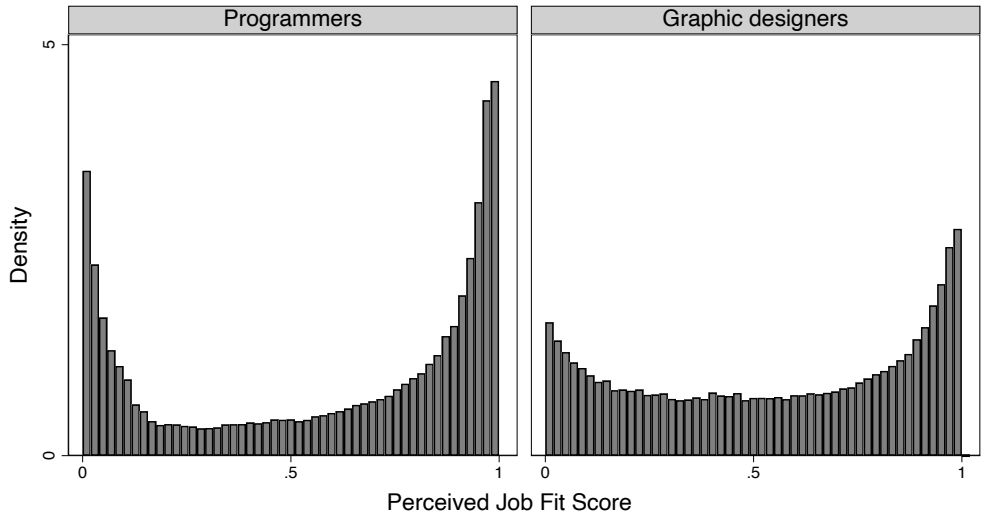


Figure A5: Histogram of *perceived job fit scores* for all the freelancers in the sample



an alternative variation of our focal approach (summarized in Table A4) in which we freeze the fifth convolutional block. Thus, we train only 18% of the total parameters in the model, corresponding to the parameters from the last fully connected layer and output layer.

Note that the perceived job fit scores provided by this alternative approach are highly correlated with the scores obtained by our focal approach (0.976 for the perceived job fit as a programmer score and 0.894 for that as a graphic designer). Nevertheless, as shown in columns 2 and 5 of Table A5, this alternative approach performs slightly worse (lower

ROC-AUC) than our original approach. Therefore, we decided to use the perceived job fit score as described in our main paper as the focal measure throughout our analyses.

Table A5: Performance metrics on validation sample

	Programmers			Designers		
	Focal	Alternative 1	Alternative 2	Focal	Alternative 1	Alternative 2
	(1)	(2)	(3)	(4)	(5)	(6)
ROC-AUC	0.889	0.858	0.880	0.898	0.821	0.849
Accuracy	0.889	0.858	0.812	0.822	0.717	0.772
Precision	0.740	0.798	0.810	0.814	0.747	0.796
Specificity	0.853	0.778	0.855	0.855	0.729	0.778
Recall	0.896	0.773	0.760	0.781	0.706	0.767

Note: Columns 1 and 4 show the metrics for the focal perceived job fit score, i.e., obtained by fine-tuning the last convolutional block. Columns 2 and 4 show the metrics obtained by freezing the last convolutional block. Columns 3 and 6 shows the metrics obtained by removing the bias term (constant term) of the classification layer.

Our second robustness check aims to address concerns regarding the definition of the perceived job fit score, i.e., the predicted probability that an image is perceived as a high fit for a job in a certain category. As an alternative, we define the perceived job fit score as the input to the classification layer removing the bias term (constant term). Note that while our focal score is a positive number between 0 and 1, this alternative score is a non-bounded real number.

Despite the obvious differences in the scale of the original perceived job fit scores and this alternative definition, we observe a high correlation between them (0.949 for the perceived job fit as a programmer and 0.924 for that as a graphic designer). As shown in columns 3 and 6 of Table A5, this alternative approach also performs slightly worse (lower ROC-AUC) than our original approach. Therefore, we decided to use the perceived job fit score as described in our main paper as the focal measure throughout our analyses.

E What makes a profile picture "look the part?" Details on the interpretability analysis

To provide insights into what makes an image "look the part," we start by collecting interpretable image-related features using different computer vision APIs. We categorize these features into four groups, and motivate their choices as follows:

- *Demographic variables:* Whether there is a human and, if so, the apparent gender, race, and age of the freelancer. Stereotypical beliefs about demographic groups might influence perceptions of fit (e.g., the stereotypical White or Asian male computer/software engineer).
- *Facial features:* Smile, beauty, face pose (roll, pitch, and way), face prominence, beard. The beauty premium or facial expressions that generally elicit positive responses, such as smiling (Fagerstrøm et al. 2017), might also influence perceptions of fit.
- *Image quality:* Blurriness, exposure, noise, technical quality, and aesthetic quality.[2] The quality of the image could serve as a signal of a graphic designer's skills and taste, which could influence perceptions of fit.
- *Accessories and background:* Reading glasses, sunglasses, formal dress, casual dress, computer, artistic,[3] portrait, indoors, outdoors. These variables could serve as a proxy for perceived intelligence (e.g., wearing glasses, Wei and Stillwell 2017) or creativity (e.g., artistic, outdoors), which could also influence perceptions of job fit.

We then use interpretable machine learning methods to predict the perceived job fit labels provided by human raters (training set described in Section 3.2.2) as a function of the features listed above. Specifically, we adopt the Xgboost classifer (Chen et al. 2015) to predict whether an image is labeled as a "high perceived fit" for each category, and

[2]The last two are obtained using the implementation of the Neural Image Assessment method by Lennan et al. (2018), which predicts how humans would rate the technical quality of an image (i.e., pixel-level degradation), and the aesthetic quality of an image (i.e., semantic level characteristics associated with emotions and beauty in images). The implementation is available at `https://github.com/idealo/image-quality-assessment`.

[3]Artistic images are drawings or pictures with filters to cartoonize or give some artistic style touch (watercolor, pop art, pointillism, etc.).

use Shapley values (Lundberg et al. 2020) to explore what image-related features are more important to predict such labels.[4]

Figure A6 shows the SHAP values of the top 20 most important features for the programmer label, and Figure A7 shows the SHAP values of the top 20 most important features for the graphic designer label. The figures illustrate interesting differences across categories. For instance, for programmers, mid-range and high beauty scores have positive and similar SHAP values, while, for designers, only high beauty scores have positive SHAP values. We also see important differences in the impact of gender, e.g., the female level has a negative SHAP value for programmers but a positive SHAP value for designers.

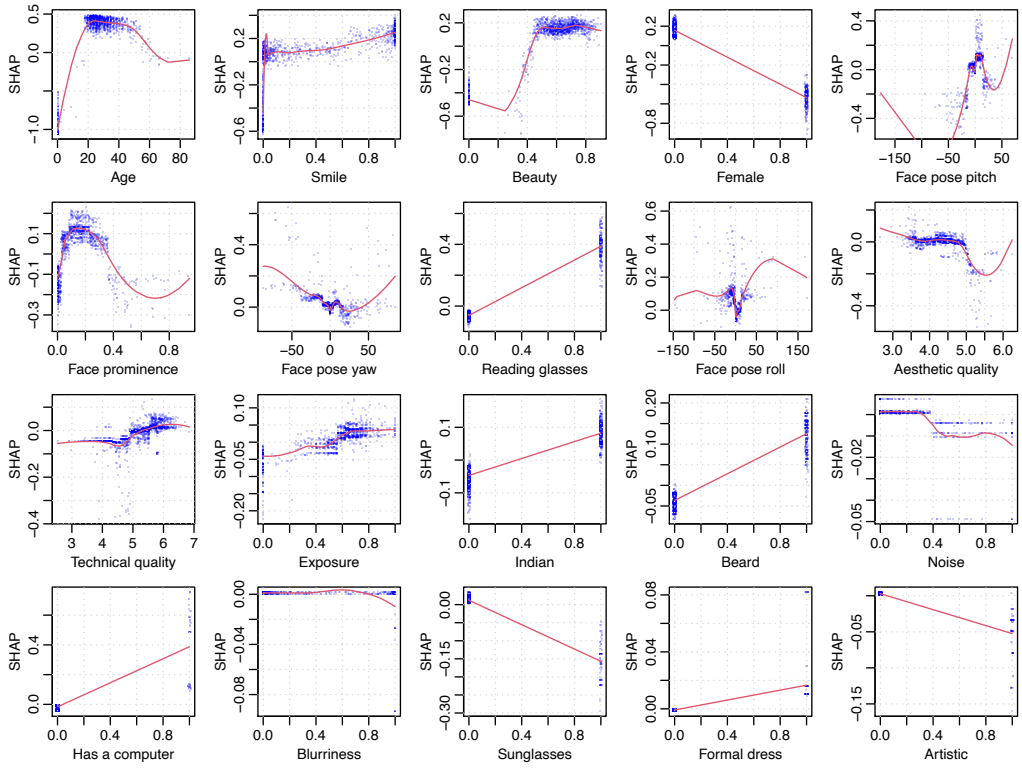
F Manipulating profile pictures to elicit different perceptions of job fit

Our results on what makes a profile picture "look the part" suggest that perceptions of job fit can be partially but not entirely explained by a freelancer's gender and race. These results may imply that, after holding gender and race constant, variations in other image-related variables such as background and accessories can elicit different perceptions of job fit.

To formally test this possibility, we run the following study. We select a sample of ten profile pictures (including at least one freelancer per gender and racial group) and manipulate their backgrounds and accessories to test whether such modifications can make the same freelancer elicit different perceptions of job fit. We focus on perceptions of job fit as

[4]To train the classifier, we fine-tune the hyperparameters of the model using random search, and select the values that give higher ROC-AUC in the validation sample. For the "high perceived job fit as a programmer" prediction task, the selected values are: max. number of boosting iterations = 30, max. tree depth = 4, min. number of nodes in a leaf = 1, fraction of features used to build each tree = 1, fraction of observations used to build each tree = 1, min. loss reduction to further partition on a leaf node = 0.1, learning rate = 0.1, and regularization term on weights = 0.0. As a result, we obtained a 95.165% in-sample ROC-AUC and 90.148% out-of-sample ROC-AUC. For the "high perceived job fit as a designer" prediction task, the selected values are: max. number of boosting iterations = 30, max. tree depth = 4, min. number of nodes in a leaf = 2, fraction of features used to build each tree = 0.6, fraction of observations used to build each tree = 1, min. loss reduction to further partition on a leaf node = 0.0, learning rate = 0.2, and regularization term on weights = 0.0. As a result, we obtained a 93.022% in-sample ROC-AUC and 81.134% out-of-sample ROC-AUC.

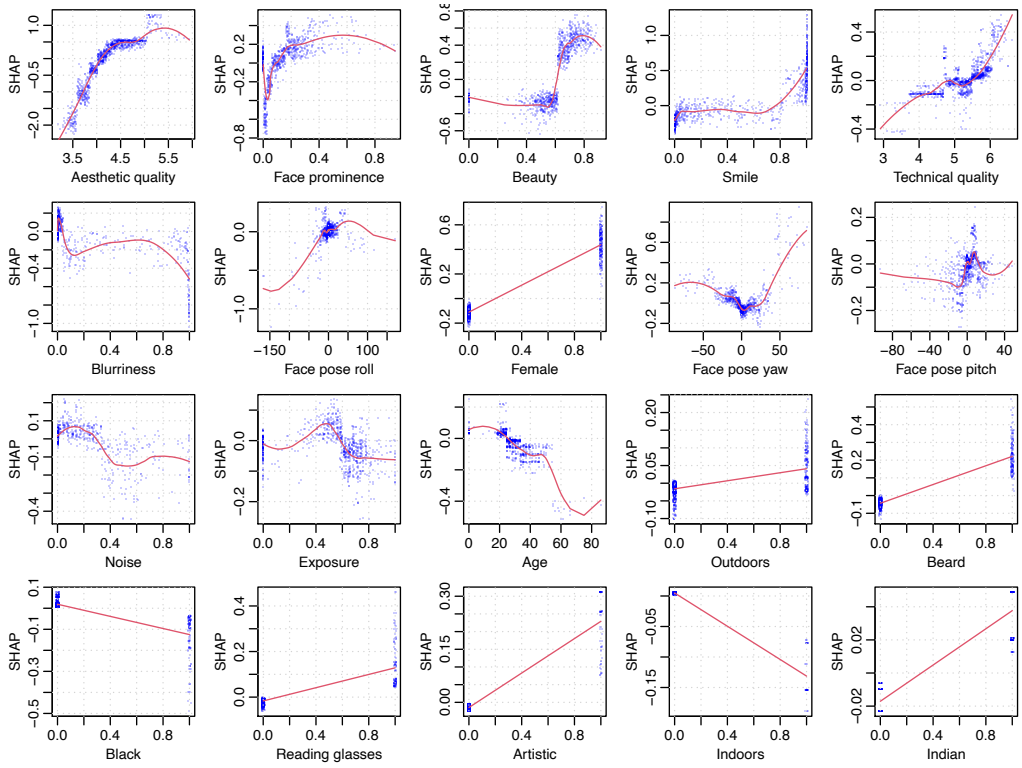
Figure A6: SHAP values for the programmer labels (top 20 features are displayed only)



a programmer, the largest category in the platform, as it allows us to create a more conservative test (recall that gender and race are important predictors of perceptions of fit as a programmer but are less so for perceptions of fit as a graphic designer). We hire a professional photo editor to create two versions of each original picture: (i) a version with an outdoor background and no glasses; and (ii) a version with a computer in view and wearing glasses. Motivated by our findings in Appendix E, we refer to the former group as the profile pictures where the freelancer "looks the part," and the latter as the group of images where the freelancer "does not look the part."

We recruited 100 participants on Amazon Mechanical Turk and showed them one randomly selected version of each freelancer's profile picture. For each profile picture, we asked raters to use a 7-point Likert scale to indicate: (i) their perception of the freelancer-job fit as a programmer, and (ii) their likelihood to hire the freelancer in the picture to build them a website. As shown in Table A6, we find that when participants see the version in which

Figure A7: SHAP values for the graphic designer labels (top 20 features are displayed only)



the freelancer "looks the part," they perceive the freelancer to be a higher fit and are more likely to hire her/him for a programming task.

G Observational study: Variables description and summary statistics

We describe the variables used in our observational study in Table A7, along with a summary statistics of the continuous and discrete variables in Tables A8a and A8b, respectively. These summary statistics are obtained at the application level (each row in our data), whereas the statistics presented in Section 3.2.1 are obtained at the user level (each freelancer in our data).

In the following paragraphs, we provide additional details on the creation of two text-related variables: Application Similarity and Application Prototypicality.

Table A6: Manipulating profile pictures to elicit different perceptions of fit

	Perceived Job Fit (1)	Likelihood to Hire (2)
Look the part version	1.335***	1.413***
Intercept	3.882***	3.860***
Controls:		
Freelancer FE	Yes	Yes
Observations	494	494
Adjusted R^2	0.260	0.246

Note: OLS estimates. In column 1, the dependent variable is the respondent perception of the freelancer-fit for the job as a programmer. In column 2, the dependent variable is the respondent likelihood to hire the freelancer to build a website. Both regressions include freelancer specific fixed-effects.

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

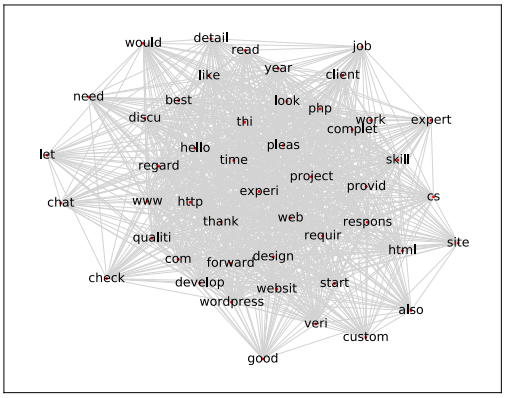
Application similarity: To create this variable, we first create a dictionary using all the job descriptions provided by employers, removing words that appeared in less than 5 of those documents. Because our goal is to capture whether a freelancer is a good match for a certain job, we also removed common words that appeared in more than 75% of the job descriptions. Using this vocabulary, we create a document-term matrix v_j to represent the job description (provided by the employer) and a document-term matrix w_j to represent the application description (provided by the freelancer). Finally, we define *application similarity* as the cosine similarity between v_j and w_j.

Distance to the prototypical application: To create this variable, we follow the semantic network-based approach proposed by Toubia and Netzer (2017). Using all the applications in our sample, we first build a semantic network for each job category, where nodes are word stems and edge weights are based on word stems co-occurrence. We illustrate the resulting semantic networks in Figure A8. Next, we define the prototypical edge weight distribution as the average distribution among all the job applications submitted to each job category. Finally, we measure the distance between each application's weight distribution

and the prototypical edge weight distribution (Kolmogorov–Smirnov statistic). We illustrate the prototypical edge weight distribution in Figure A9, and the distribution of the job application *prototypicality* in Figure A10.

Figure A8: Semantic networks of job application descriptions

(a) Programmers



(b) Graphic Designers

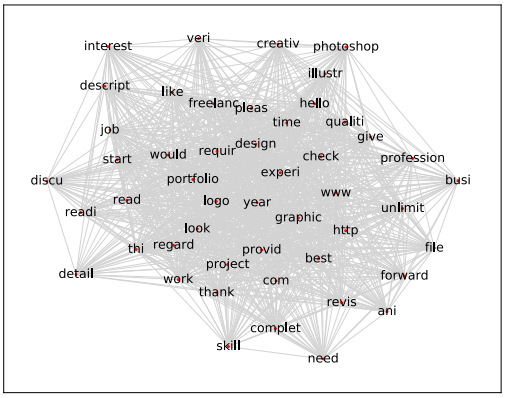
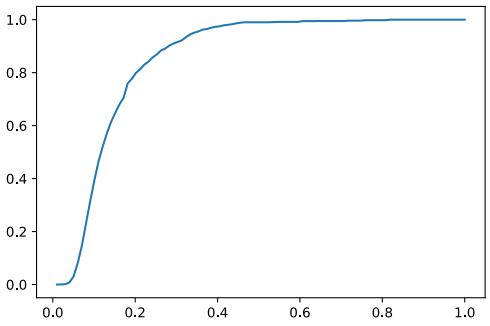
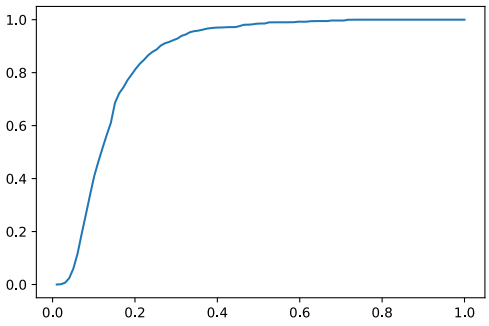


Figure A9: Prototypical edge weight distribution

(a) Programmers



(b) Graphic Designers



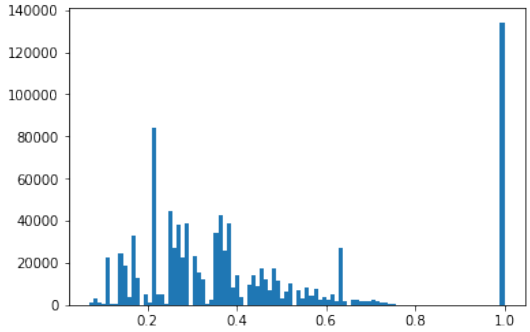
H Observational study: Robustness checks

H.1 Robustness of the perceived job fit coefficient

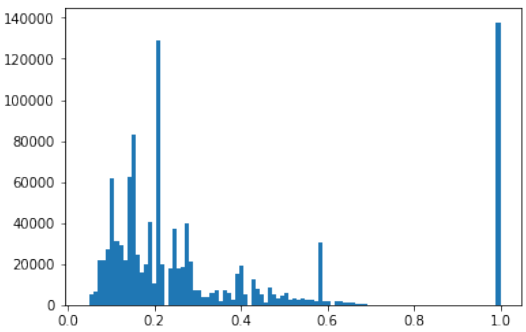
We present two analyses aimed at examining the robustness of the estimated perceived job fit coefficient. As a first robustness check, we separately estimate the model for each of the

Figure A10: Distribution of *prototypicality* of the job applications

(a) Programmers



(b) Graphic Designers



Note: The peaks on values around 1 correspond to approx. 13% of the applications that contain very few words. The average word count in applications with prototypicality ¡ 1 is 67.562, while the average word count in applications with prototypicality = 1 is 1.533.

two job categories and report our results in Table A9. We note that perceived job fit has a positive and significant coefficient for both job categories. For programmers, the effect is no longer significant once we control for demographics and beauty, which we believe is driven by supply-side factors in our observational data (i.e., perceived job fit for programmers is highly correlated with gender, race, and age), which make it harder to separate their effects. For designers, we observe that the perceived job fit coefficient changes very little when controlling for demographics and beauty, consistent with findings that these variables are less important in explaining perceptions of job fit.

As a second robustness check, we use the bootstrap procedure to obtain confidence intervals for the coefficients of interest (main effect of perceived job fit and its interactions in Tables 1 and 3). For each model, we use 100 bootstrap replications, each replication randomly sampling jobs (and their respective applications) with replacement. The results are consistent with those reported in the main manuscript.

Table A7: Variables included in the conditional logit model

Variable name	Description
Profile Picture Variables	
Perceived Job Fit	Perceived job fit score as predicted by the VGG-16 classifier
Has Picture?	Whether the freelancer has a profile picture
Human	Whether a human face is detected in the profile picture, obtained from Face++ API
Gender	Female or Male, obtained from Face++ API
Race	White, Black, Asian, or Indian, obtained from Face++ API
Age	Age, obtained from Face++ API
Beauty	Beauty score, obtained from Face++ API
Reputation Variables	
No Reviews Yet	Whether the freelancer has zero reviews
Log(1 + N. of Reviews)	Log of the cumulative number of reviews the freelancer has at the time of the application
Avg. Rating	Cumulative average rating of the reviews the freelancer has at the time of the application
Avg. Sentiment Score	Cumulative average sentiment valence score (-1 to 1) of the text of the reviews the freelancer has at the time of the application, obtained with Google Natural Language API
Avg. Sentiment Magnitude	Cumulative average sentiment magnitude or strength regardless of valence (non-negative number) of the text of the reviews the freelancer has at the time of the application, obtained with Google Natural Language API
Application Variables	
Price	Price the freelancer requests to complete the job, normalized within the job
Number of Days	Number of days offered by the freelancer to complete the job, normalized within the job
Application log word count	Log of the number of words in the application description submitted by the freelancer
Application-description similarity	Cosine similarity between the application description submitted by the freelancer and the job description posted by the employer. More details in the paragraphs after the tables.
Distance to prototypical application	Distance to the prototypical application in the job category, constructed using the semantic network-based approach proposed by Toubia and Netzer (2017). More details in the paragraphs after the tables.

Continued on next page

Table A7 (Continued): Variables included in the conditional logit model

Variable name	Description
Application Variables	
Recommended Freelancer?	Whether the freelancer is recommended by the platform and highlighted in the top position of the list of applicants, as seen by the employer.[1] The platform recommends one freelancer per job.
Log Application Position	Position of the application relative to the entire list of applications, as seen by the employer
Performance Variables	
Earning score	Total earning score (ranging from 0 to 10) from previous projects that required similar skills and were successfully completed
Percentage of jobs on time	Percentage of previous jobs delivered on time
Percentage of jobs on budget	Percentage of previous jobs delivered on budget
Additional controls	
Preferred Freelancer Certification?	Whether the freelancer has the preferred freelancer badge, which can obtain after meeting certain criteria (e.g., rank high in one or more skill tests, have a verified profile, etc.)[2]
Exam on required skill?	Whether the freelancer passed an exam on a skill required by the employer (e.g., Word-Press)[3]
From Developed Country	Whether the freelancer is from a developed country
From Employers' Country	Whether the freelancer and the employer are from the same country
Freelancer Region	Region of residence of the freelancer (e.g., North America)
Previously Reviewed?	Whether the freelancer has a review from the same employer. We use this as a proxy for whether the freelancer was hired by the same employer in the past.
Number of items on Portfolio	Number of items in the portfolio of the freelancer profile
Membership Category	Membership category of the freelance: Free, Intro, Basic, Plus, Professional, Premier[4]
Profile Completed	Whether the freelancer completed his/her user profile

[1] The choice of the recommended freelancer is based on his/her reviews and previous experience. Specific details of the criteria used by the platform are unknown.
[2] For more information, see: `https://www.freelancer.com/support/General/what-are-preferred-freelancers`.
[3] These exams are implemented by the platform. See: `https://www.freelancer.com/exam/exams/`
[4] For more information, see: `https://www.freelancer.com/membership/index.php`.

Table A8a: Summary Statistics for continuous variables

Variable	Mean, Std. Dev.	25th, 50th, 75th Percentiles	Min, Max
Profile Picture Variables			
Perceived Job Fit	0.616, 0.351	0.301, 0.740, 0.934	0, 1
Has Picture	0.991, 0.096	1, 1, 1	0, 1
Human	0.723, 0.447	0, 1, 1	0, 1
Age*	30.237, 7.992	24, 29, 34	1, 90
Beauty*	0.650, 0.120	0.562, 0.653, 0.743	0.202, 0.962
Reputation Variables			
No Reviews Yet	0.154, 0.361	0, 0, 0	0, 1
Log(1 + Number of Reviews)**	3.985, 1.689	2.708, 4.127, 5.187	0.693, 8.462
Average Rating**	4.806, 0.459	4.797, 4.895, 4.970	0, 5
Avg. Sentiment Score**	0.691, 0.206	0.664, 0.748, 0.803	-0.9, 0.9
Avg. Sentiment Magnitude**	1.484, 0.514	1.325, 1.493, 1.672	0, 14.9
Application Variables			
Price (Normalized)	0.277, 0.270	0.062, 0.196, 0.419	0, 1
Number of days (Normalized)	0.285, 0.284	0.100, 0.200, 0.375	0, 1
Log(1 + Application Word Count)	3.467, 1.517	3.219, 3.912, 4.431	0, 5.333
Application-description similarity	0.199, 0.140	0.089, 0.199, 0.296	0, 1
Distance to prototypical application	0.373, 0.279	0.182, 0.273, 0.455	0.040, 1
Recommended Freelancer?	0.032, 0.176	0, 0, 0	0, 1
Log (Application Position)	2.831, 1.072	2.197, 2.996, 3.638	0, 4.605
Performance Variables			
Earning Score on job category	4.635, 2.650	2.996, 5.335, 6.644	0, 10
Percentage of jobs on time	0.781, 0.355	0.843, 0.945, 0.991	0, 1
Percentage of jobs on budget	0.792, 0.358	0.872, 0.959, 0.995	0, 1
Additional Controls			
Preferred Freelancer?	0.115, 0.319	0, 0, 0	0, 1
Exam on required skill?	0.170, 0.538	0, 0, 0	0, 1
From Developed Country	0.092, 0.288	0, 0, 0	0, 1
From Employers Country	0.054, 0.225	0, 0, 0	0, 1
Previously Reviewed?	0.005, 0.069	0, 0, 0	0, 1
Number of items on Portfolio	24.156, 40.551	5, 13, 27	0, 1782
Profile Complete	0.993, 0.082	1, 1, 1	0, 1

* Conditional on Human = 1. Note that the min and max of age label are both outliers with either babies or a very old individual in the picture.

** Conditional on No Reviews Yet = 0.

Table A8b: Summary Statistics for discrete variables

	Distribution
Profile Picture Variables	
Gender*	26.652% Female, 73.348% Male
Race*	9.206% Black, 58.719% Indian, 8.992% Far East Asian, 23.084% White
Additional Controls	
Freelancer Region (23 in total)	72.492% Southern Asia, 4.232% Northern America, 3.815% Eastern Europe, 3.396% Eastern Asia, 2.616% South-Eastern Asia, 13.449% Other
Membership Category	32.316% Free, 5.846% Intro, 3.376% Basic, 14.241% Plus, 44.222% Premium

* Conditional on Human = 1.

Table A9: Estimating hiring decisions by job category

	Websites IT & Software			Design Media & Architecture		
	(1)	(2)	(3)	(4)	(5)	(6)
Profile Pictures Variables:						
Perceived Job Fit Score		0.065*	−0.007		0.084***	0.089***
Has Picture		0.281**	0.278**		0.256**	0.258**
Reputation Variables:						
No Reviews Yet	−1.140***	−1.142***	−1.163***	0.487	0.425	0.357
Log(1 + N. Reviews)	0.389***	0.389***	0.388***	0.489***	0.492***	0.490***
Avg. Rating	0.207***	0.207***	0.204***	0.581***	0.566***	0.555***
Application Variables:						
Offered Price	−1.910***	−1.909***	−1.909***	−1.811***	−1.816***	−1.819***
Log(1 + Application WC)	0.163***	0.163***	0.164***	0.163***	0.164***	0.165***
Application-Description Similarity	1.664***	1.663***	1.659***	0.851***	0.845***	0.844***
Distance to Prototypical Application	0.837***	0.839***	0.843***	0.488***	0.489***	0.493***
Additional Variables:						
Performance Variables	✓	✓	✓	✓	✓	✓
Other Application Variables	✓	✓	✓	✓	✓	✓
Control Variables	✓	✓	✓	✓	✓	✓
Human		✓	✓		✓	✓
Demographics (Gender, Race, Age)			✓			✓
Beauty			✓			✓
N	936,141	936,141	936,141	1,092,623	1,092,623	1,092,623
LL	-75,056	-75,050	-75,008	-79,420	-79,400	-79,375
AIC	150,207	150,202	150,130	158,937	158,902	158,863
BIC	150,771	150,801	150,799	159,508	159,509	159,542

Note: Conditional logit estimates with standard errors clustered at the job level. The dependent variable is whether employer i hired freelancer j from the pool of applicants for job t. In columns 1 and 4, we estimate the model controlling for everything except for profile picture related variables. In columns 2 and 5, we add the variable of interest, perceived job fit score. In columns 3 and 6, we incorporate additional profile picture related control variables including demographics and beauty.

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table A10: Bootstrap standard errors and confidence intervals.

Model (1)	Variable (2)	Observed Coefficient (3)	Bias (4)	SE (5)	95% Confidence Interval (6)	
Main effect	Perceived job fit	0.089	−0.003	0.016	[0.058, 0.121]	(N)
					[0.054, 0.114]	(P)
					[0.062, 0.119]	(BC)
With demographics and beauty controls	Perceived job fit	0.075	−0.002	0.016	[0.044, 0.106]	(N)
					[0.043, 0.104]	(P)
					[0.043, 0.104]	(BC)
Interaction with dispersion in Avg. Rating	Perceived job fit	0.076	−0.001	0.017	[0.042, 0.110]	(N)
					[0.036, 0.115]	(P)
					[0.049, 0.121]	(BC)
	⋯ × Dispersion in Avg. Rating	0.010	−0.001	0.007	[-0.003, 0.023]	(N)
					[-0.005, 0.021]	(P)
					[-0.005, 0.021]	(BC)
Interaction with dispersion in N. Reviews	Perceived job fit	0.133	−0.006	0.022	[0.090, 0.176]	(N)
					[0.079, 0.167]	(P)
					[0.090, 0.175]	(BC)
	⋯ × Dispersion in N. Reviews	−0.031	0.002	0.011	[-0.054, -0.009]	(N)
					[-0.057, -0.006]	(P)
					[-0.060, -0.008]	(BC)
Interaction with distance to the next best candidate	Perceived job fit	0.212	0.002	0.024	[0.165, 0.259]	(N)
					[0.159, 0.249]	(P)
					[0.145, 0.244]	(BC)
	⋯ × Distance to next best candidate	−0.153	−0.002	0.017	[-0.186, -0.120]	(N)
					[-0.185, -0.119]	(P)
					[-0.184, -0.109]	(BC)
Interaction with avg. distance to best 5 next candidates	Perceived job fit	0.261	0.002	0.027	[0.20, 0.314]	(N)
					[0.199, 0.307]	(P)
					[0.180, 0.299]	(BC)
	⋯ × Avg. Distance to best 5 next candidates	−0.138	−0.001	0.015	[-0.167, -0.110]	(N)
					[-0.166, -0.106]	(P)
					[-0.163, -0.101]	(BC)

Note: Columns 1, 2, and 3 indicate the model, main variables of interest, and their observed coefficients in the main manuscript (Table 1 and Table 3), respectively. Column 4 shows the bias, equal to the difference between the observed coefficient and the average of the bootstrapped estimates. Column 5 shows the estimated bootstrap standard error. Column 6 shows the 95% confidence intervals based on the following approaches: normal (N), percentile (P), and biased corrected (BC).

H.2 Out-of-sample prediction of hiring choices in observational study

We measure the predictive accuracy of the conditional choice model in our observational study under different model specifications using 10-fold cross-validation. In each fold, we randomly take 80% of jobs and their respective applications to calibrate the model and use the estimated parameters to predict which freelancer will be hired in each of the remaining 20% jobs. We are especially interested in the out-of-sample hit rates, i.e., the percentage of times the model correctly predicts the winner.[5] We start from a full model specification (column 2 of Table 1 in the main manuscript) and then remove one variable at a time to measure its impact on predictive accuracy.

Our results, reported in Table A11, suggest that removing the perceived job fit score in the model doesn't significantly decrease predictive accuracy. Interestingly, removing other arguably important variables such as the average rating and whether an application is recommended by the platform do not decrease predictive accuracy either. Moreover, removing the number of reviews and price in the model (separately) also leads to arguably quite small decreases of 0.712 and 2.532 percentage points (respectively) from a baseline of 26.535%.

Table A11: Out-of-sample hit rates under different model specifications (Observational Study)

Model specification	Avg. Hit Rate	Diff. with full model	p-value
Full model	26.535%		
Full model (-) Perceived Job Fit	26.472%	−0.064%	0.391
Full model (-) Avg. Rating	26.417%	−0.118%	0.141
Full model (-) Log (1 + N.Reviews)	25.823%	−0.712%	0.000
Full model (-) Price	24.004%	−2.532%	0.000
Full model (-) Recommended by Platform	26.471%	−0.065%	0.187

Overall, these results suggest that improving the prediction accuracy of the conditional logit model is challenging within the setting of our observational data. We believe that

[5]We focus on hit rates because only one freelancer per job gets hired, and each job corresponds to a unique set and number of alternatives.

this is most likely driven by the complexity of our field data. In our context, each choice set comprises a unique set of 30+ freelancers, many of whom are highly similar strong applicants with desirable attributes (high reputation, competitive price, etc.) As such, pinpointing which freelancer which be hired based on any single variable in the conditional choice model can be quite difficult. Moreover, we do not have enough repeated observations per employer to account for employer-level heterogeneity in hiring preferences. Such characteristics of the observational data have made it extremely challenging to predict which freelancer will be hired in the out-of-sample jobs.

We also would like to point out that, in our experimental study 1, incorporating the perceived job fit variable in the model indeed improves the out-of-sample hit rate by 3 percentage points when the reputation system is less diagnostic (more similar to the diagnosticity in the secondary data). We believe that, in this case, the incremental gains in out-of-sample predictive accuracy can be explained by the following characteristics of the experimental data: (i) having fewer (10 rather than 30+) and more differentiated candidates (from the orthogonal design) in each choice set than in the secondary data; and (ii) having repeated observations per respondent which allow us to better gauge the respondent-level preferences which can be leveraged to more accurately predict the same respondent's choices on the holdout tasks.

H.3 Perceptions of professionalism and competence

In this section, we examine how perceived job fit relates to other general social attributions, such as perceptions of freelancer professionalism and competence. In our view, while perceptions of job fit can be a function of these general attributions, it differs from them because it is more domain-specific. This idea is similar to that in Olivola et al. (2012), who explore the link between having a "Republican-looking face," an attribution inherently tied to the political domain and election outcomes.

To further explore the relationship between perceived job fit, professionalism, and competence, we use a procedure similar to that described in Section 3.2.2 of the main manuscript to label profile pictures based on perceived professionalism and competence. Thus, we use human raters to label a subsample of profile pictures and leverage the VGG16 image classifier to predict the labels for the remaining samples in our dataset.

We then include the perceived professionalism and competence metrics directly into the choice model. Our results, reported in Table A12, show that the perceived job fit coefficient is still positive and significant after including these variables in the model. Interestingly, we also find that the perceived professionalism coefficient is negative and significant. Our conjecture is that attributes generally affecting perceptions of professionalism (i.e., wearing a suit, having a plain passport-like picture) might be less relevant in our context, particularly for programmers and graphic designers who are often known for their inclination to dress casually rather than formally. We also find that the perceived competence coefficient is not always significant. Overall, these findings are consistent with our claim that domain-specific perceptions of job fit might be more important than general perceptions of professionalism and competence on these types of platforms.

H.4 Are freelancers who "look the part" better at their jobs?

In the following, we provide additional details on how we explore the relationship between perceived job fit and a series of outcome variables. Specifically, we ran several regressions to examine if "looking the part" is associated with outcomes such as whether the hired freelancer received a review after being hired and completing the job, and if so, the rating received, as well as other performance metrics (whether the job was completed on time/budget). We report our results in Table A13. Note that each row in the table corresponds to a different outcome used as the dependent variable, and each column displays the estimates for the perceived job fit coefficient when using a different set of independent variables.

Table A12: Controlling for perceived professionalism and perceived competence

	(1)	(2)	(3)
Profile Pictures Variables:			
Perceived Job Fit Score	0.103***	0.086***	0.100***
Perceived Professionalism	−0.059***		−0.079***
Perceived Competence		0.023	0.048**
Has Picture	0.317***	0.277***	0.314***
Reputation Variables:			
No Reviews Yet	−0.682***	−0.685***	−0.684***
Log(1 + N. Reviews)	0.438***	0.438***	0.439***
Avg. Rating	0.309***	0.308***	0.309***
Application Variables:			
Offered Price	−1.868***	−1.868***	−1.869***
Log(1 + Application WC)	0.161***	0.162***	0.161***
Application-Description Similarity	1.289***	1.289***	1.288***
Distance to Prototypical Application	0.679***	0.682***	0.679***
Recommended by the Platform	0.298***	0.297***	0.298***
Additional Variables:			
Performance Variables	✓	✓	✓
Other Application Variables	✓	✓	✓
Control Variables	✓	✓	✓
Human	✓	✓	✓
N	2,028,764	2,028,764	2,028,764
LL	-154,836	-154,838	-154,833
AIC	309,775	309,781	309,772
BIC	310,427	310,432	310,436

Note: Conditional logit estimates with standard errors clustered at the job level. The dependent variable is whether employer i hired freelancer j from the pool of applicants for job t.
Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

When using perceived job fit as the only independent variable (results in column 1 of Table A13), we observe that it correlates positively with the probabilities that (i) the employer writes a review for the freelancer (row 1 of Table A13), and (ii) the freelancer completes the job on time (row 2 of Table A13). Nevertheless, after controlling for all information employers observe when hiring (results in column 2 of Table A13), we find no significant correlation between perceived job fit and any outcome metric (with the only exception that freelancers who look the part are slightly less likely to complete the job on budget). Taken

Table A13: Exploring the relationship between perceived job fit and post job completion outcomes

Specification: / Outcome:	Perceived job fit only	Perceived job fit with controls	Observations
Receives a review	0.010**	0.003	62,936
Completed on time	0.008**	−0.002	52,581
Completed on budget	0.001	−0.005*	52,581
Rating overall	0.006	0.000	52,581
Rating quality	0.007	0.001	52,581
Rating expertise	0.008	0.003	52,581
Rating communication	0.008	0.002	52,581
Rating professionalism	0.005	0.000	52,581
Rating would hire again	0.001	−0.005	52,581
Sentiment Score	0.003	0.000	52,581
Sentiment Magnitude	0.007	0.014	52,581

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Note: Each row corresponds to a different outcome used as the dependent variable. Columns 1 and 2 show the perceived job fit coefficient estimated under different model specifications, that is, using perceived job fit as the only independent variable and using additional controls, respectively. Column 3 shows the number of observations available for the analysis. For the results in the first row, the number of observations is smaller than the number of jobs in the sample (62,936 vs. 63,014) because employers can review freelancers only after the payment is processed, and not all payments were yet completed at the time of the data collection. For the results in the remaining rows, the number of observations decreases because these values are only observed conditional on receiving a review.

at face value, these findings could suggest that perceptions of fit are a rather noisy signal of quality that add little additional information to reputation and performance metrics.

Table A14: All coefficients for regression with additional control variables

	Receives a review (1)	Completed on time (2)	Completed on budget (3)	Rating overall (4)	Rating communication (5)	Rating expertise (6)
Perceived Job Fit Score	0.003	−0.002	−0.005*	0.000	0.001	0.003
Prior Avg. Rating Overall	0.138***	−0.008	−0.010*	0.686***	0.689***	0.703***
Prior Log(N. Reviews)	0.015***	−0.002	−0.005***	−0.007***	−0.006**	−0.011***
Prior Earning Score	−0.026***	−0.002	−0.000	0.002	0.003	0.008***
Prior % of jobs on time	0.014	0.758***	−0.029**	−0.001	−0.028	−0.025
Prior % of jobs on budget	0.001	−0.015	0.779***	0.036	0.043*	0.010
Preferred Freelancer Certification	0.018***	0.003	0.001	0.016***	0.010*	0.011**
Qualification Score	0.000	0.000	0.000	0.000	−0.000	0.000
Qualification Number	0.002***	0.000	0.001*	0.002**	0.002**	0.002***
Intercept	0.242***	0.289***	0.294***	1.503***	1.502***	1.454***
N	62,936	52,581	52,581	52,581	52,581	52,581
Adjusted R2	0.055	0.089	0.087	0.101	0.076	0.091

Note: OLS coefficients. In column 1, the dependent variable is whether the freelancer receives a review after completing the job. The number of observations is smaller than the number of jobs in the sample (62,936 vs. 63,014) because employers can review freelancers only after the payment is processed, and not all payments were yet completed at the time of the data collection. In columns 2 to 6, the dependent variables correspond to different characteristics of the review. The number of observations decreases from column 1, because these values are only observed conditional on receiving a review.

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

Table A14 (Continued): All coefficients for regression with additional control variables

	Rating would hire again (7)	Rating quality (8)	Rating professionalism (9)	Sentiment score (10)	Sentiment magnitude (11)
Perceived Job Fit Score	0.002	0.000	−0.005	0.000	0.014
Prior Avg. Rating Overall	0.625***	0.654***	0.760***	0.205***	0.124***
Prior Log(N. Reviews)	−0.002	−0.004	−0.013***	0.004**	−0.107***
Prior Earning Score	−0.004	0.000	0.005	−0.003**	0.073***
Prior % of jobs on time	0.036	−0.012	0.022	0.072***	0.065
Prior % of jobs on budget	0.035	0.060**	0.033	0.016	0.263***
Preferred Freelancer Certification	0.018***	0.020***	0.020***	0.009***	0.039***
Qualification Score	0.000	0.000	0.000	0.000	−0.000
Qualification Number	0.001*	0.002**	0.002**	−0.000	0.006***
Intercept	1.778***	1.659***	1.122***	−0.290***	0.653***
N	52,581	52,581	52,581	52,581	52,581
Adjusted R2	0.085	0.089	0.084	0.037	0.009

Note: OLS coefficients. In columns 7 to 11, the dependent variables correspond to different characteristics of the review. These values are only observed conditional on receiving a review.

Significance levels: * p<0.1, ** p<0.05, *** p<0.01.

I Pretesting diagnosticity conditions in experimental study 1

For this pretest, we recruited 200 Mturkers and showed them two freelancers with different attribute levels and no profile pictures. We randomly assigned participants to see the two freelancers under either the low or high diagnosticity condition and asked them to rate their similarity using a 7-point Likert scale. Participants in the less diagnostic condition find freelancers more similar than those in the more diagnostic condition (4.885 vs. 4.204, $p < 0.01$).