Do Jobseekers Value Diversity Information? Evidence From a Field Experiment and Human Capital Disclosures

Step-By-Step Data Description

Choi, Pacelli, Rennekamp, and Tomar

January 2023

Step-By-Step Data Description

Given our project is based in large part on a proprietary/confidential dataset (from the career advice agency Zippia), we are complying with the "Data and Code Sharing Policy for the *Journal of Accounting Research*" by providing "a detailed step-by-step description of the code or the relevant parts of the code."

This document describes the detailed data steps behind each the tables and figures depicted in the paper. Please note that Table 1 (Description of Treatment Conditions), Figure 1 (Experimental Design), and Figure 2 (Zippia's User Interaction) are qualitatively descriptive in nature and are not supported by any underlying code; as such, we do not describe them below.

Dataset Construction

Tables 2 through 7 pull from the following datasets. We describe what each dataset represents and how each one was constructed.

mainUserProfile

This dataset describes the user demographics for all Zippia users. The raw data come from: Data/userProfile.csv (which provides user_id, user_group, and education_level) and Data/userProfile_gender.csv (which contains more complete, updated info from Zippia regarding user gender; possible values include "Male," "Female," "Unknown", and "NA"). We merge the two raw data files together, by user_id and user_group. We create a binary variable for a user having or not having a bachelor's degree (i.e., if a user's education level contains "bachelors", "masters," or "doctorate," then the variable "hasbach" is set to 1; otherwise, 0). We also clean the data by setting any blank cells to NA for education_level. We also use the raw data file: Data/userProfile_preferredLevel.csv, which describes users' preferred job level. We merge the userProfile data to the preferred level data, by user_id, so that for each user in the dataset, we know their maximum preferred job level.

In addition, we also use the raw data file: Data/userProfile_locations.csv, which describes users' locations (city and state). We remove duplicates and take the first locational entry. We then create a table of states and their corresponding abbreviations, assigning regions to each of the states with R's "state.region" function (i.e. assigning users to "South", "West," "North Central," "Northeast", and "Puerto Rico," based on their state). We first merge the de-duplicated user location data with the state data, by state, before merging this intermediary dataset with the userProfile dataset, by user_id. Furthermore, we also use the raw data file: Data/userProfile_ethnicityDistribution.csv, which contains user ethnicity information. In this file, Zippia provides numeric scores, by user, for the following ethnicities: Asian, Black, Hispanic, White, and Unknown. We first find users' max

ethnicity (i.e. we assign users' to their highest-scored ethnicity). We then create a binary variable for poc/white ("poc" is set to 1 if the max ethnicity variable is non-missing, not unknown, and non-white; it is set to 0 if the max ethnicity variable is non-missing, not unknown, and white; the variable is set to NA otherwise). We then merge this ethnicity data with the userProfile data, by user_id.

Finally, we merge the BLM data* (Data/blm_events.csv) with the userProfile data, by state. This file contains the number of BLM events, 2020 state population, and BLM events per capita at the state level, and has 51 lines (50 states + D.C.), with an "events_per_capita_100k" generated for each of the 51 states. We generated "median_blm" which is the median of the "events_per_capita_100k" variable (i.e., each state is equally weighted in its contribution toward the median, since it takes the median of the 51 lines). We generated an "events_above_median" binary variable in the userProfile dataset, which is set to 1 if events_per_capita_100k is greater than the median_blm, and 0 otherwise. The final dataset contains Zippia user characteristics for analysis, including information on users' id, group, gender, ethnicity, preferred job level, education level, state, region, and BLM events per capita by users' state.

*This section further elaborates on how the "Data/blm_events.csv" file was constructed. This file contains BLM events per capita by state as well as state population. We later merged this with Zippia user characteristics (as described above). The raw external ACLED data lives in Data/acled/2018-05-21-2021-05-30-United_States.csv. The data is pulled from https://acleddata.com/#/dashboard. We also used ACLED's codebook (https://www.acleddata.com/wp-content/uploads/2015/01/ACLED_Codebook_2015.pdf) to better understand this data. The ACLED data contains information on BLM-related events in 2020 and 2021, as well as information on location and affiliation of the main actors. We also made use of the external BEA Population Data (https://www.bea.gov/data/income-saving/personal-income-by-state), which contains information on state populations from 1929-2020. The raw BEA data lives at: Data/bea/SAINC4__ALL_AREAS_1929_2020.csv. First, to understand if an event was "pro_blm" or not, we generated a "pro_blm_list" comprising the following groups: Black Panthers, Black Lives Matter (BLM), Black Student Union (BSU), African American Group, National Association for the Advancement of Colored People (NAACP). We also generated an "anti_blm" list, comprising actors such as Back the Blue and QAnon (there are several more). If the associated actors in an event were anti_blm == F & pro_blm == T & pro_missing == F, then we assigned the event to the category "pro_blm." Next, we filtered for just these "pro_blm" events and looked at pro_blm events by state. Then, we explored state-level pro_blm events per 100k population. To measure the state population, we leveraged the BEA population data for the year 2020. We took the number of pro_blm events in a state and divided it by the state population in 2020, multiplied by 100,000. This is how we arrived at the number of pro blm_events per 100,000 people by state, i.e. the variable "events_per_capita_100k."

emailsSent_by_id

This dataset describes the number of emails sent by Zippia to users of the platform, by user. There is a binary variable included in the dataset that is set to one if a user was sent at least one email. The raw data comes from: Data/userProfile_emailsSent.csv. It contains the time stamp and email campaign id by user. We merged the raw data with user demographics (the mainUserProfile dataset, described above), by user_id and user_group. If the time stamp is not missing, we set a binary variable called "sent" equal to 1; otherwise, 0. Finally, we created a "num_sent" variable that is the sum of the "sent" variable by user_id and user_group. The final dataset contains user_id, user_group, num_sent, and "bin_sent" which is a binary variable set to 1 if num_sent is greater than 0.

emailsOpened_by_id

This dataset describes the number of emails opened by users of the platform, by user. There is a binary variable included in the dataset that is set to one if a user opened at least one email. The raw data comes from: Data/userProfile_emailsOpened.csv. It contains the time stamp and email campaign id by user. We merged the raw data with user demographics (the mainUserProfile dataset, described above), by user_id and user_group. If the time stamp is not missing, we set a binary variable called "opened" equal to 1; otherwise, 0. Finally, we created a "num_opened" variable that is the sum of the "opened" variable by user_id and user_group. The final dataset contains user_id, user_group, num_opened, and "bin_opened" which is a binary variable set to 1 if num_opened is greater than 0.

emailsClicked_by_id

This dataset describes the number of jobs clicked by users of the platform, by user. There is a binary variable included in the dataset that is set to one if a user clicked on at least one job. The raw data comes from: Data/userProfile_emailsClicked.csv. It contains the time stamp and email campaign id, company name and id, and position of the company on the email, by user. We merged the raw data with user demographics (the mainUserProfile dataset, described above), by user_id and user_group. If the time stamp is not missing, we set a binary variable called "clicked" equal to 1; otherwise, 0. Finally, we created a "num_clicked" variable that is the sum of the "clicked" variable by user_id and user_group. The final dataset contains user_id, user_group, num_clicked, and "bin_clicked" which is a binary variable set to 1 if num_clicked is greater than 0.

mainCompanyDataResults

This dataset contains cleaned company information for all Zippia companies. The raw data comes from: Data/companyDataResults.csv. We transform the relevant Zippia-provided diversity and salary scores, creating (for non-missing, non-zero values) logged diversity scores ("log_diversity_score"), logged salary scores ("log_salary_score"), and logged median salaries ("log_median_salary"). The final dataset contains 107,810 observations, corresponding to 107,810 unique companies, and involves such information as company_id, company_name,

company_employee_size, company_revenue, company_type (industry), diversity score, logged diversity score, salary score, logged salary score, logged median employee salary, etc.

mainClickedData

This dataset describes the merged clicked data, i.e., user and company information for clicked job postings. The raw data comes from: Data/userProfile_emailsClicked.csv. First, we merged this raw data file with the userProfile dataset, maintaining multiple clicks and only keeping users who clicked on a posting. More specifically, we merged the emailsClicked dataset with the userProfile dataset, by user_id, only maintaining those users who clicked on job postings. Next, we merged this intermediary dataset with the company information dataset (mainCompanyDataResults), by company_id.

recommended_companies

This dataset contains user and company information for both clicked or recommended job postings. The relevant input datasets used to construct the sample are as follows:

- User data: Data/mainUserProfile.csv
- Company data: Data/mainCompanyDataResults.csv (refer above)
- Emails data: Data/userProfile_emailsClicked.csv (contains observations with time stamps; we create the binary "clicked" variable below).
- Ratings data: Data/userProfile_companyRatings.csv (contains user_id, company_id, and company rating).

We merged the emails data with the company ratings data by user_id, company_id, and company_code. We then created a binary "clicked" variable in this newly merged dataset that we set to 1 if there was a time stamp, and 0 if the time stamp was missing (i.e., we consider a user to have clicked on the job posting if there is a time stamp). We then merged this dataset with the user data by user_id. Finally, we merged in the company data by company_id. The final dataset contains both clicked and unclicked (clicked == 1 & clicked == 0) companies and companies with ratings (as a proxy for potentially recommended companies).

<u>Tables and Figures</u>

Table 2: Descriptive Statistics

Table 2 uses six different raw datasets: mainUserProfile, emailsSent_by_id, emailsOpened_by_id, emailsClicked_by_id, mainCompanyDataResults, and mainClickedData. Please refer above for how these datasets were constructed.

- In Panel A, we use the mainUserProfile dataset to tabulate descriptive sample statistics, limiting the sample to users assigned to the baseline and diversity conditions (N = 178,862). We use this dataset to provide the number of users and percentage of the sample falling into each category of the variables for gender, education level, preferred job level, and region.
- In Panel B, we use the emailsSent_by_id, emailsOpened_by_id, and emailsClicked_by_id datasets to tabulate descriptive statistics of user engagement, limiting the sample to users assigned to the baseline and diversity conditions (N = 178,862).
- In Panel C, we use the mainCompanyDataResults dataset (N = 107,810) to tabulate descriptive statistics for all firms in the sample, whether clicked on or not by users. 66,694 observations in the dataset have a non-missing diversity score.
- In Panel D, we use the mainClickedData dataset, limiting the sample to users assigned to the baseline and diversity conditions, to tabulate descriptive statistics for only those firms in the sample that were clicked on by users. We limit the sample to exclude missing company_id and discover that the number of unique company_ids is 7,871; of these companies, 7,304 have a non-missing diversity score.

Table 3: Covariate Balance across Treatment Conditions

Table 3 uses four different raw datasets: mainUserProfile, emailsSent_by_id, emailsOpened_by_id, and emailsClicked_by_id. Please refer above for how these datasets were constructed.

- In Panel A, we use the mainUserProfile dataset, limiting the sample to users assigned to the baseline and diversity conditions. We then determine what proportion of the sample falls into the associated user categories (e.g., female, entry-level preferred job level, etc.) for both the baseline and diversity conditions, and run t-tests to compare the means of covariates across treatment conditions.
- In Panel B, we use the emailsSent_by_id, emailsOpened_by_id, and emailsClicked_by_id datasets, limiting the sample to users assigned to the baseline and diversity conditions. We then determine the mean number of emails received, emails opened, and jobs clicked upon for both the baseline and diversity conditions and run t-tests to compare the means of user engagement across treatment conditions.

Table 4: Diversity Information and Job Search

All panels in this table use the mainClickedData dataset, which represents clicks along with user and company characteristics. Please refer above for how this dataset was constructed.

- In Panel A, we limit the sample to users assigned to the baseline and diversity conditions and exclude observations with missing diversity scores. We then run a t-test of means of the Zippia-provided diversity score (variable: "zippia_unified_scores_unified_work_env") across the treatment conditions.
- In Panel B, an "early interaction" analysis, we first order the data by user_id and time_stamp, and then create a "firstCampaign" variable which lists the first email campaign by user. We then keep

only the observations corresponding to each user's first email campaign and run the same analysis as above; that is, we limit the sample to users assigned to the baseline and diversity conditions and exclude missing diversity scores. We then run a t-test of means of the Zippia-provided diversity score across the treatment conditions.

- In Panel C, we limit the sample to users assigned to the baseline and salary conditions and exclude missing salary scores. We then run a t-test of means of the Zippia-provided salary score (variable: "zippia_unified_scores_unified_salary") across the treatment conditions.

Table 5: Relative Sensitivity to Diversity Information (RSDI)

Table 5 uses the recommended_companies dataset, which represents both clicked jobs and jobs recommended to users that were not clicked on. Please refer above for how this dataset was constructed.

We limit the sample to exclude observations that are missing an education level, a preferred job level, and a Zippia-provided diversity score (variable: "zippia_unified_scores_unified_work_env"), and then estimate our probit model, clustering standard errors by user.

Table 6: Diversity Information and Job Search Activities across Demographic Groups

All panels in this table use the mainClickedData dataset, which represents clicks along with user and company characteristics. Please refer above for how this dataset was constructed.

We limit the sample to users assigned to the baseline and diversity conditions. We then generate the mean Zippia-provided diversity score for the baseline condition, diversity condition, and both conditions along the dimensions relevant to each panel (e.g., in Panel C: Preferred Job Level, we generate the mean diversity score for the entry, junior, mid, senior, management, and executive levels for both treatment conditions). We also use this dataset to run our ANOVA tests and generate the depicted ANOVA output for each panel. For the sample pertinent to each ANOVA panel, the relevant variable cannot be missing (e.g., in Panel C, the dataset excludes observations where "max_preferred_level" is missing). In Panel A, we use the variable "female" to code gender; in Panel B, we use the variable "hasbach" to code education level; in Panel C, we use the variable "max_preferred_level" to code preferred job level; and in Panels D, we use the variable "events_above_median" to code Above or Below Median Pro-BLM Events Per Capita.

Table 7: Follow-up Survey Results

All panels in Table 7 use the raw data file, "FormattedSurveyResponses.xlsx." Our research team wrote the survey questions in Qualtrics. We provided the survey link to our data partner, Zippia, who made the link available to users of the platform. Whenever Zippia users answered survey

questions, the responses were automatically saved to our Qualtrics server. After the survey period concluded, we pulled the survey data out of Qualtrics and uploaded it to the SAS statistical software, John's Macintosh Project (more commonly known as JMP) for data analysis, using the application's established functionality. The raw Excel data file referred to above is pulled from JMP. As such, there is no supporting code; Panels A, B, and C simply summarize the count of participant responses (and percentage of the relevant sample) in response to our survey questions about the usefulness of diversity information.

Human Capital Disclosure Data

By hand, we collect the full sample of Human Capital Disclosure (HCD) sections in firms' 2020 10-Ks that were available for download from the SEC's EDGAR portal by the end of March 2021. We then create the indicator variable "disc" as follows. We extract sentences that contain one or more demographic groups and a percentage or fraction. The demographic groups are: persons of color or other minority groups; ethnically or racially diverse; ethnically diverse; people of color; persons with disabilities; racial minority; persons of color; professionals of color; racially and/or ethnically diverse; racial or ethnic minorities; chosen not to identify; racial and ethnic diversity; racial minority groups; racially and ethnically diverse; black / african american; black/african american; hispanic / latino; black or african american; hispanic/latinx; hispanic / latinx; native american and pacific islander; native hawaiian/pacific islander; native hawaiian or other pacific islander; american indian or alaska native; american indian/alaskan native; native american; unknown or undeclared; additional groups; two or more races; underrepresented minorities; under-represented groups; underrepresented groups; underrepresented minorities; asian american; hispanic or latino; pacific islander; racially diverse; gender diverse; caucasian/white; veterans; non-white; other than white; non-minority; hawaiian; veterans; male; female; asian; hispanic; latinx; latino; black or african; african american; african-american; african; black; white; veteran; minorities; minority; minorities; women; men ; multiracial.

We then read through each sentence and focus on those where a company discloses the number of women (any of the gender-based categories listed above) or people of color (any of the race-based categories listed above) employed during fiscal year 2020. Our process covers disclosures in figures and tables. The disclosures can vary in scope. For example, the region could be the US, outside the US, unspecified, etc. The function could be all roles, middle management, upper management, etc. We focus on disclosures that consider all roles (including when firms reference 'associates'). These disclosures can still vary in geographic scope.

Gender disclosures typically do not specify a region or specify "global/worldwide" as the region. We therefore consider a firm to be a "gender-discloser" if it makes a location non-specific or global/worldwide disclosure of the percentage of women employed. To facilitate a consistent analysis, we treat firms making only a non-US gender disclosure as a gender non-discloser. We drop firms making only a US-based gender disclosure from the sample. People of color disclosures typically do not specify a region or specify the US as the region. We, therefore, consider a firm to

be a 'people of color-discloser' if it makes a location non-specific or US-based disclosure of the percentage of people of color employed. To facilitate a consistent analysis, we treat firms making a non-US people of color disclosure as a people of color non-discloser. We drop firms making only a global/worldwide people of color disclosure from the sample. Finally, we consider a firm to be a diversity-discloser if it is a gender-discloser and/or a people of color-discloser, as described above.

We merge in variables from Compustat (SIC code, market value, EBIT, total assets, state of incorporation). We also merge in the BLM data described above by state, retaining the number of events per capita. Using a combination of fuzzy-string matching and hand-matching, we merge in the Zippia company data contained in the file "companyDataResults." This produces some duplicates (by firm). In these cases, we keep the observation with the highest value of gender_total to capture the observation that better represents the company's workforce because it contains the most Zippia user input in terms of workforce demographics. We limit the sample to HCD observations that we can match with the Zippia company data. We then merge in the field experiment data contained in "mainClickedData," described above. By 2-digit SIC code and user_group (experimental treatment), we compute the difference in mean diversity score of the firms clicked in the Baseline and Diversity Conditions. We call this "experiment_delta."

Table 8: Human Capital Disclosure Analysis

Panel A of Table 8 presents descriptive statistics for the variables used in the regression analysis in Panel B. Standard errors are heteroskedasticity-consistent.

Figure 3: Human Capital Disclosure Topics

Figure 3 considers the firms used in the Table 8 Panel B regression tests. In Panel A, we perform textual analysis; first, we write a Python function that checks whether a word exists in a given list of words. We then write a function that counts how many rows contain all the words in a list of words. In doing so, we are able to count the frequency of HCD subsection names. In Panel B, we write a function that allows us to perform topical analysis on the contents of firms' HCD subsections. We count the frequency of 10-K HCD sections that include a topic within their subsections. The function is written such that topics with multiple words in their title are only "counted" when all of the words in the title appear in a firm's HCD subsection.

Additional Table – Data Linking & Roadmap

Description of Inputs	Raw Data - Inputs	Merged Datasets – Final Output
User demographics Users' gender Users' preferred job level Users' location Users' ethnicity BLM events per capita, by state ACLED data on BLM events BEA data on state populations	Data/userProfile.csv Data/userProfile_gender.csv Data/userProfile_preferredLevel.csv Data/userProfile_locations.csv Data/userProfile_ethnicityDistribution.csv Data/blm_events.csv Data/acled/2018-05-21-2021-05-30-United_States.csv Data/bea/ SAINC4__ALL_AREAS_1929_2020.csv	mainUserProfile
Time stamp and email campaign id by user User demographics (merged)	Data/userProfile_emailsSent.csv Data/mainUserProfile	emailsSent_by_id
Time stamp and email campaign id by user User demographics	Data/userProfile_emailsOpened.csv Data/mainUserProfile	emailsOpened_by_id
Time stamp and email campaign id by user User demographics	Data/userProfile_emailsClicked.csv Data/mainUserProfile	emailsClicked_by_id
Company information	Data/companyDataResults.csv	mainCompanyDataResults
Time stamp and email campaign id by user User demographics Company information	Data/userProfile_emailsClicked.csv Data/userProfile.csv Data/mainCompanyDataResults.csv	mainClickedData
User demographics (merged) Company information (merged) Time stamp and email campaign id by user Company ratings	Data/mainUserProfile.csv Data/mainCompanyDataResults.csv Data/userProfile_emailsClicked.csv Data/userProfile_companyRatings.csv	recommended_companies