SUBJECTIVE INFORMATION IN HIRING DECISIONS

MARK AZIC AND DIEGO LAMÉ

Abstract. Information gathered throughout hiring processes has a great degree of subjectivity. Most job interviews include a cluster of questions that could be evaluated in very heterogeneous ways by different decision maker, even though its effects various aspects of the labor market are unknown. Using a simulated labor market experiment, we explore the effect of subjective information in hiring decisions. We find that subjective information changes the overall valuation managers assign to worker profiles, but it doesn't improve the hiring results. Furthermore we find that subjective information strongly reduces the well-known bias in favor of male workers when it comes to hiring for a stereotypically male tasks. This implies that subjective information is a potentially useful tool in combating discrimination. Finally, we find that experienced managers obtain worse results when subjective information is part of the workers' profile.

Introduction

Labor markets are characterized by a number of subjective processes: situations in which different agents share the same information set, the same technology, and the same objectives, yet reach different conclusions. Examples include interviewing potential hires, writing worker performance reviews, determining performance pay, and delegating tasks amongst workers. In this paper, we use a simulated labor market to study the effect of subjective information in the hiring decision. We focus on the value of subjective information in hiring the right worker; the way in which subjective information interacts with objective information; and whether subjective information has heterogeneous effects on hiring across demographic groups.

Understanding how subjective information is valued has the potential to help explain economic phenomena such as residual wage inequality (i.e. the different labor market outcomes between observationally equivalent workers) and the increasing returns to social skills in labor markets[1]. Our focus on hiring is due to its well-established importance in labor market outcomes (see, for example, (Topel and Ward, 1992); (Farber, 1994); (Altonji et al., 2013);

Date: November, 2018.

[1]For examples, see (Juhn et al., 1993) and (Deming, 2017), respectively.

and (Bagger et al., 2014)), as well as the long-run effects of early labor market outcomes[2]. That hiring matters so much to the outcomes of workers and firms is in sharp contrast to the lack of studies on the topic. In a recent Handbook of Labor Economics chapter, (Oyer and Schaefer, 2011) highlight the overall lack of empirical studies on hiring, calling the process a "black box". How employers use subjective information to evaluate potential hires is a major part of that black box, as many of the signals employers rely on — interviews, references, resumes — are inherently subjective.

In addition to being important to labor market outcomes and omnipresent in hiring, there is also good reason to believe that subjective information interacts with demographics in a meaningful way. For example, (Goldin and Rouse, 2000) show that female candidates for symphony orchestras have their music judged more favorably when auditions are gender-blind. Another example comes from an ongoing lawsuit brought against Harvard University by a group representing Asian-American applicants[3]. Like many universities, Harvard asks applicants to submit personal essays and letters of recommendation, in addition to submitting objective measures such as grade point average and standardized test scores. Amongst other complaints, the lawsuit argues that such subjective measures consistently harm Asian applicants. The interaction between subjective information and an applicant's demographics is relevant to the legal literature on adverse impact. Under Title VII of the 1964 Civil Rights Act, it is illegal for a firm to request information which has a disparate impact on minority groups — regardless of intent — unless that firm can demonstrate that the information is "reasonably related" to the job[4]. Therefore, if subjective information has heterogeneous effects across demographic groups is a concern for labor law.

In this paper we use an experimental labor market to explore subjective information in hiring. Our work has three main goals. First, we look at how hiring efficacy is affected by having both objective and subjective information on a job applicant. Second, we use treatments with and without subjective information to determine the trade-off between objective and subjective information in how managers value applicants. Third, we look at how the presence of subjective information affects gender discrimination in a setting where women are often discriminated against.

Our experiment consists of two types of sessions: worker sessions and manager sessions. In worker sessions, we collect worker profiles, which are akin to resumes. At the beginning of a worker session, subjects are asked to write a short statement about their ability to multiply

[2]For example, (Kahn, 2010) and (Oreopoulos et al., 2012) both show that graduating from college during a recession has negative and persistent effects on earnings, as graduates place in lower quality jobs initially.
[3]See *Students for Fair Admissions, Inc. v. President and Fellows of Harvard College.*
[4]For a famous case regarding adverse impact, see *Griggs v. Duke Power Co.*.

numbers together, as if they were being interviewed for a job that requires this as a labor input. Next, subjects are asked perform a multiplication task. After that, subjects engage in a related task, adding sets of numbers together. The short statement task provides us with a piece of subjective information for each worker profile, while the sum task provides us with objective information.

Manager sessions consist of two experiments. In the first one, we show subjects worker profiles we collected during the initial sessions, where we ask them to value each worker. In the second experiment, two worker profiles are placed side-by-side, and managers are asked to select the one they believe scored higher on the multiplication task. In each manager experiment there are two treatments: one in which the worker's subjective information is included in the worker's profile, and one in which it is not. The use of a price list gives us a detailed measure of valuation, which we use to explore the objective-subjective trade-off; the use of side-by-side comparisons allows us to compare similar worker profiles, and see what effect subjective information has on anti-female bias.

We find that subjective information significantly changes the overall valuation managers assign to worker profiles, although it doesn't lead to better or worse hiring results. Furthermore, these results are sensitive to manager gender. We also find that subjective information strongly reduces the bias in favor of male workers in our setting, but it causes experienced managers to make worse hiring decisions.

Firms — in practice — solicit and interpret a diverse set of subjective measures on job applicants; therefore, the effects of subjective information are likely to be setting and type dependent. However, we do show that subjective information has important implications for hiring in our very generic setting. Therefore, we are confident in saying that our results show the need for firms to carefully consider how they use subjective information, as well as the need for researchers to carefully consider the presence of subjectivity when studying hiring.

As mentioned before, we know of no studies which explicitly address subjective information in hiring[5]. The difficulty of isolating subjective information is likely one reason for the dearth of literature on subjectivity. In survey data, firms employ unique screening and interview processes, leading to different subjective information collected across firms. Problems remain if we look within a firm, as different managers involved in a hiring decision might

[5]There are studies that consider the use of subjective information in compensation (see (Moers, 2005), (Bol, 2008), (Bol, 2011), and (Baker et al., 1994)). Empirical work in this field shows that subjectivity tends to compress compensation differences between workers.

view a potential hire at different times, have different levels of input on the hiring decision, or have competing motivations. Our experimental setup allows us to avoid these challenges.

The paper most closely related to our focus on information in hiring is (Hoffman et al., 2017). Hoffman et al. looks at the introduction of analytic job testing of applicants across fifteen firms, in which job applicants are given a score by a machine based algorithm. A signal of the applicant's aptitude (green, yellow or red) is provided to a hiring manager by a third party firm, who then uses traditional signals of worker quality (e.g. interview and resume) to determine whether to hire a worker. The authors find that managers who hire against the test recommendation do worse on average.

While Hoffman et al. obtain strong results regarding the efficacy of two different approaches to hiring, they're unable to characterize how managers made their hiring decisions. In particular, hiring managers don't know how the signal is constructed nor have any way to gage how informative it is ex-ante. The fact that the signal turned out to be, on average, more informative than the managers' own assessment is not relevant to our question. By contrast, our experimental setting allows us to carefully control what information a manager sees. This allows us to test the hiring efficacy under two different information settings but also, importantly, allows us to say something about *how* managers reach their conclusions[6].

In addition to contributing to a scant literature on information in hiring decisions, our work adds to a large literature using experimental labor markets to understand discrimination in hiring. Several of these papers study discrimination against women in math related tasks and consider ways to reduce bias. (Reuben et al., 2014) show that both male and female managers have a preference for men when hiring workers for a math task. They find that providing managers with a worker's previous score on the same exact task reduces, but does not eliminate, the bias. (Bohnet et al., 2015) show the same preference for male workers performing a math task, but show that it can be reduced when male and female workers are considered side-by-side. (Coffman et al., 2017) study the preference for male workers using math and sports quizzes for performance tasks. Using a clever partition of workers by birth month, they decompose the preference for men, finding results that are consistent with statistical discrimination. Our work differs from these papers, in that it allows workers to submit a second piece of information to managers: their written statement. That we find evidence that this reduces discrimination against women creates the possibility for another

[6]In addition, we also study a more common hiring setting than Hoffman et al. While it is true that more and more large employers have adopted analytic testing for job applicants, the majority of hiring is still done by small firms, employing less sophisticated approaches.

way to limit preference for men.

The rest of the paper is structured as follows: Section 2 gives our experimental design; Section 3 presents our results; and Section 4 concludes.

Experimental Design

Our experiment consisted of two types of sessions: Worker Sessions and Manager Sessions. In Worker Sessions, subjects performed incentivized tasks. These tasks comprised a worker's "profile". In Manager Sessions, subjects were shown worker profiles and asked to forecast worker productivity. We now describe the Worker and Manager Sessions in detail. We refer to subjects in the Worker Sessions as "workers" and subjects in the Manager Sessions as "managers" throughout.

Worker Sessions: Worker Sessions were conducted in the Pittsburgh Experimental Economics Laboratory using zTree ((Fischbacher, 2007)). Each session consisted of four tasks, one of which was randomly chosen for payment. We conducted three such sessions, with thirteen subjects in each. We now discuss the tasks in detail, before giving an explanation for the setup.

The first task prompted workers to write a short statement about their ability to multiply numbers. Workers were given the following instructions for the task:

> You're trying to convince another person of your ability to multiply numbers. It may help to think of this as an interview question, where the job you're applying for requires you to multiply numbers. Therefore, feel free to mention anything from your academic background, work experience, personal interests etc. which you think will help you convey your ability to multiply numbers.

Workers were given ten minutes to complete task one. There were no restrictions regarding word count or content.

In the second task, workers had three minutes to solve as many multiplication problems as they could, with problems consisting of a random two digit number times a random one digit number.

Task three was used to evaluate each worker's written statement from task one. Each worker was presented with the written statement of the other 12 workers on their screen,

one statement at a time. The worker was then asked to predict the other worker's performance on the multiplication task (task two) using **only** the statement from task one. After submitting a prediction, a worker was informed of the actual multiplication score associated with the statement they just evaluated. The order in which workers judge each other's subjective information was varied so as to avoid order effects.

In task four, workers had three minutes to solve as many sums of five, random two-digit numbers as they could ((Niederle and Vesterlund, 2007))[7].

Before performing any of the tasks, workers answered a short demographic questionnaire which included age, gender, year in school, and major.[8]

For tasks two and four, workers were paid a piece rate of \$0.50 and \$1.00 per correct answer, respectively. In task 3, workers received \$3.00 per exact prediction and \$1 for each close prediction (within 3 points of the actual multiplication score). Finally, a worker's earnings for task one were given by \$0.50 times the average prediction other workers gave based on the worker's statement during task three. Because payment on tasks one and three were interdependent, workers were given a brief overview of the experiment prior to beginning task one.

Discussion: The first task provides us with a piece of subjective information for each worker. This will be part of the worker profile shown during Manager Sessions, with the idea being that different managers can read the same response to a prompt and interpret it differently. The prompt we use is similar to a worker being asked: "why do you think you'd be a good fit for this job?" — a frequent question in job applications and interviews.

Task three performs the important role of quantifying the quality of a worker's subjective information. We want to control for the information a worker's statement provides when presented to managers, in order to draw inferences about the importance of subjectivity in manager's choices. But how to compare the similarity between written responses is unclear. Our approach, of having other workers attempt to predict a worker's multiplication score using only her subjective information, is an attempt to reduce all of the information in that worker's written statement to a single value. We comment more on how task three is used when discussing results.

Having workers complete multiplication and addition tasks was done in order to tie our results to the aforementioned literature showing statistical discrimination against women in

[7]In tasks two and four (multiplication and addition, respectively) subjects were allowed to use scratch paper, but not allowed to use a calculator. No electronic devices were permitted at any point in the experiment.
[8]The questionnaire was conducted at the start of the experiment to prime truthfulness in Task 1 statements, particularly regarding a worker's major. Whether a result of this or not, no worker's statement in task one mentioned a major different from what was indicated in the questionnaire, nor were there any other inconsistencies between the two.

math-related tasks. The decision to use multiplication *and* addition — rather than just one task twice — comes from job applicants rarely having experience or training which perfectly aligns with a position's needs.

Finally, the answers to the questionnaire at the beginning of each session are akin to the type of information a manager would obtain from a worker's CV.

Out of the 39 subjects in the Worker Sessions, we discarded 3 for deviating from the statement's original intent; writing as to a fellow experimental participant instead of to a manager[9]. Out of the remaining 36 workers we removed anyone with a Sum Task score below 4 or above 9. These could be considered under-qualified and over-qualified workers, which would rarely apply to the job in the first place, or they would be immediately discarded or hired. This leaves us with a pool of 27 workers for the managers to evaluate.

Manager Sessions: After all Worker Sessions were run, we conducted Manager Sessions using Amazon Mechanical Turk. We implemented a 2x2 experimental design, in which the main treatment condition was whether subjective information was included in a worker's profile or not, while the other dimension varied with respect to evaluation method.

Managers were randomly assigned to either a Partial Information or Full Information treatment. In the **Partial Information treatment**, the worker profiles a manager was shown consisted of the worker's age, gender, year in school, and major, as well as the worker's score on the addition task[10]. In the **Full Information treatment**, the worker profiles a manager was shown consisted of all same information as in the Partial treatment, as well as the worker's written statement. Therefore, the two information treatments only differed with respect to whether a worker's subjective information was included or not.

We used an Individual Valuation and a Side by Side Comparison to have managers evaluate worker profiles. 159 managers participated in the **Individual Valuation treatment**, where they were shown 18 worker profiles each, one at a time. For each worker profile, a manager is required to submit a valuation for the worker using a modified price list mechanism[11]. Each round's payoff was therefore either a random offer, or — if the worker was

[9]For example, one subject ended their statement with the following sentence: 'I guess since its a theoretical job and maybe you wouldn't see me every day and I'm just ok at math you wouldn't give me a very high prediction but maybe there is some compassion in this cruel world so I don't really see a downside to you giving me a high prediction thank you <3.'

[10]Although the effects of age, year in school and major are not our main interests, we include them so as to make gender less salient.

[11]Our modified price list had the simplicity of a typical price list, but did not allow subjects to select multiple switch points: once a price was chosen, the rest of the list auto-completed according to said selection.

hired — \$0.10 times the worker's performance in the multiplication task. Workers solved between 5 and 28 multiplication problems, leading to a price list ranging from \$0.50 to \$2.80, in ten-cent increments[12]. Once a valuation was submitted, the manager was informed of the worker's actual multiplication score, as well as what the manager's earnings for that round would be. The manager was then shown the next worker profile. Two randomly selected rounds, one from the first nine and one from the last nine, were chosen for payment.

In the **Side by Side Comparison**, each of the 496 managers saw five pairs of worker profiles, one at a time. For each pair, a manager had to try select the worker who performed better in the Multiplication Task. One of the five rounds was randomly selected for payment, and a correct selection in that round earned the manager \$2.00. Managers did not receive feedback after each round, and only learned whether they had made the correct selections after the experiment concluded. The pairs were selected such that the differential in the two workers' sum task scores was never more than 1. Furthermore, each manager saw: one pair consisting of two men; one pair consisting of two women; and three pairs consisting of a man and a woman. The three mixed gender pairs were constructed with possible score differentials on the sum task of +1 for the man, +1 for the woman, and even (zero differential in sum task score).

Discussion: In the Individual Valuation treatment we provide managers with feedback after each submitted valuation. This allowed us to check for learning effects between rounds. We did not provide feedback during the Side by Side Comparison; managers made just five binary selections in the Side by Side treatment, leaving us with insufficient power to test for learning.

As mentioned, the worker pairs in the Side by Side treatment were all within 1 point in their sum task score. The reason for this is that actual hiring decisions are often narrowed down to a choice between two competitive, similar applicants. If we allowed for pairs to consist of workers who scored very differently in the sum task, managers would overwhelmingly choose the candidate with the higher sum score; that choices are binary would make it impossible for us to see what effect subjective information had on preferences over workers.

[12]Since real life hiring managers should have an idea of what range of worker performance they can expect, managers were provided with the overall max and min of worker performance on the multiplication and sum tasks.

Results

Workers:

Table 1 shows summary statistics for the Worker Sessions. Subjects solved 14.49 multiplications and 5.82 sums on average. In Task 3 of our worker sessions, subjects were asked to use the statements of other workers to predict how many multiplications those workers solved. We use these predictions as our measure of the "quality" of a worker's statement. The average prediction the workers received from their peers based on the statement they provided in Task 1 was 14.96, with substantial heterogeneity in how statements were interpreted; for example the average predicted scores ranged from 10.08 to 21.92. This shows that worker statements were judged to be of varied quality.

Table 1. Worker Summary Statistics

Variable	**Mean**	**S.D.**	**Min**	**Max**
Multiplication Score	14.49	6.49	4	32
Sum Score	5.82	2.55	1	12
Words in Statement	139.00	53.71	16	229
Relevant Major	0.31	0.47		
Relevant Classes	0.49	0.51		
Avg. Predicted Score	14.96	2.72	10.08	21.92
N	39			

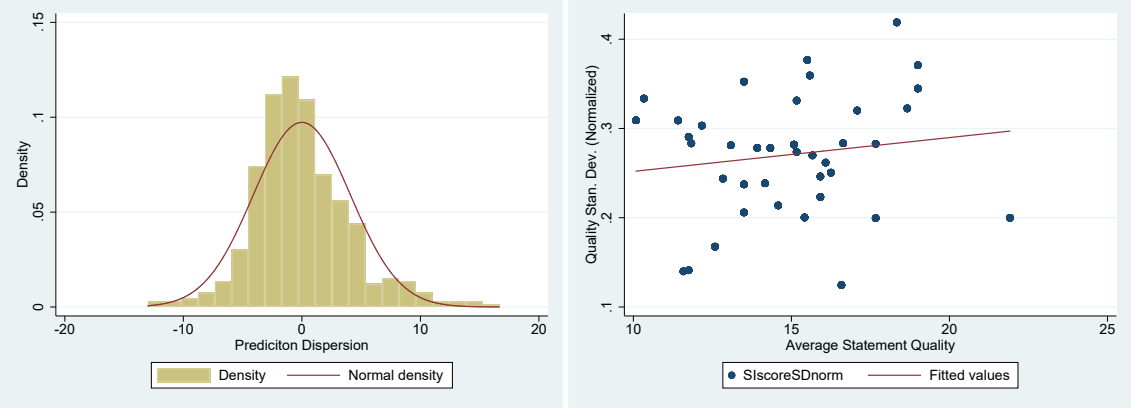
Note: Relevant Major is a binary variable that indicates whether a worker mentions that they are pursuing a major typically regarded as math intensive. Similarly, Relevant Classes indicates the mention of a worker having taken math intensive classes. Average Predicted Score is the average Multiplication Score prediction the workers received from their peers based on their Task 1 statement.

Moreover, there was substantial disagreement on the quality of each statement. Let i denote a worker, and $s_{i,j}$ denote the quality of worker i's statement, according to worker j. Therefore, if worker j reads worker i's statement, and predicts that i would correctly solve 10 multiplication problems, $s_{i,j} = 10$. We're interested in characterizing the distribution of $s_{i,j}$ in our worker sessions.

We begin by looking at the overall distribution of statement quality disagreement. For each statement, we calculate the mean quality score, and subtract it from the quality score given by a worker, i.e. $s_{i,j} - \bar{s_i}$ where $\bar{s_i} = \sum_{i \neq j} s_{i,j}$. We compute this differences for each of the 27 workers included in our manager sessions, giving us a measure of the disagreement

over a worker's statement quality, net of the statement's mean quality. With each worker's statement being judged by 12 other workers, we compute a total of 324 such differences, and plot them in Figure 1 (left). The histogram shows that the distribution of subjectivity looks approximately normal[13], with mean 0 (by construction) and a standard deviation of 4.1. Superimposed on the histogram is the probability density function for a Normal distribution with the same mean and standard deviation. The Normal distribution fits our data well.

FIGURE 1. Statement Quality Disagreement



The histogram (left) shows the frequency distribution for our disagreement measure: $s_{i,i'} - \bar{s}_i$, across all worker statements. Superimposed on the histogram is the probability density function for a Normal distribution with the same mean and standard deviation as our disagreement measure ($\mu = 0$, $\sigma = 4.10$). In the scatter plot (right), we plot each statement's average quality against the standard deviation of its quality, where the standard deviation is normalized by the statement's average quality. The correlation between the two is 0.148

We now look at whether subjectivity differed across statements: were statements equally likely to generate disagreement, or were some more contentious than others? To do this, we calculate the standard deviation in quality for each statement, normalizing it by the statement's average quality. In Figure 1 (right) we plot these normalized standard deviations against average statement quality. The graph does not appear to show any relationship between the two (corr=0.148); there was as much disagreement for low quality statements as for high quality ones.

Next, we look at whether written statements are informative of worker multiplication scores. We do this in Table 2. Subjects in the worker sessions seem to be able to correctly use the subjective information to make better than average predictions of others' performance. In an attempt to understand this, we look at the effect of mentioning a major stereotypically

[13]We excluded one outlier from our data. One subject predicted that one of their peers solved 45 multiplication in 3 minutes, generating a disagreement of 26 points with respect to the mean prediction for that worker, and 13 points higher than the maximum score attained.

considered as math intensive (STEM, Economics, or Business), mentioning having taken math intensive courses, and the length of the statement. This can be considered as objective information included in the statement. Once we control for these variables, we find that the statement is no longer informative of worker multiplication score. In particular, the mention of a math intensive major is the only variable that is significantly predictive of a higher score in the multiplication task.

TABLE 2. Predictiveness of Subjective information Score

Variable	**(1)**	**(2)**
Predicted Score	0.735*	0.291
	(0.374)	(0.398)
Relevant Major		4.680*
		(2.475)
Relevant Classes		-0.685
		(2.041)
Words in Statement		-0.0196
		(0.0207)
Constant	3.493	11.74*
	(5.686)	(6.678)
Observations	39	39

Standard errors in parentheses [* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$]
Both specifications are ordinary least squares regressions. The dependent variable is the score in the Multiplication task.

In the manager sessions, all worker profiles include their major. Therefore we can safely use our 'statement quality' measure in the following section since any mention of a relevant major in the statement should be irrelevant for the managers.

Managers:

We first look at the Individual Valuation treatment. In this setup, we obtain a precise valuation for worker profiles. This allows us to study the trade-off between objective and subjective information, as well as any differential valuation due to a worker's gender. We begin by looking at manager valuation errors, and whether subjective information helps reduce valuation errors.

We find that manager predictive ability is indistinguishable across information treatments. We define a manager's error to be the absolute difference between a worker's actual multiplication score and the manager's predicted score for that worker, where the manager's predicted score is given from the price list valuation. In the partial information treatment, manager's make an average error of 5.36, while average error in the full information treatment is 5.40, with the difference being statistically insignificant (p-value 0.79). While average errors under either information treatment are indistinguishable, managers in both treatments would have done significantly better by choosing the midpoint of the price list for every worker (average error 4.86).

TABLE 3. Information effects in worker valuation

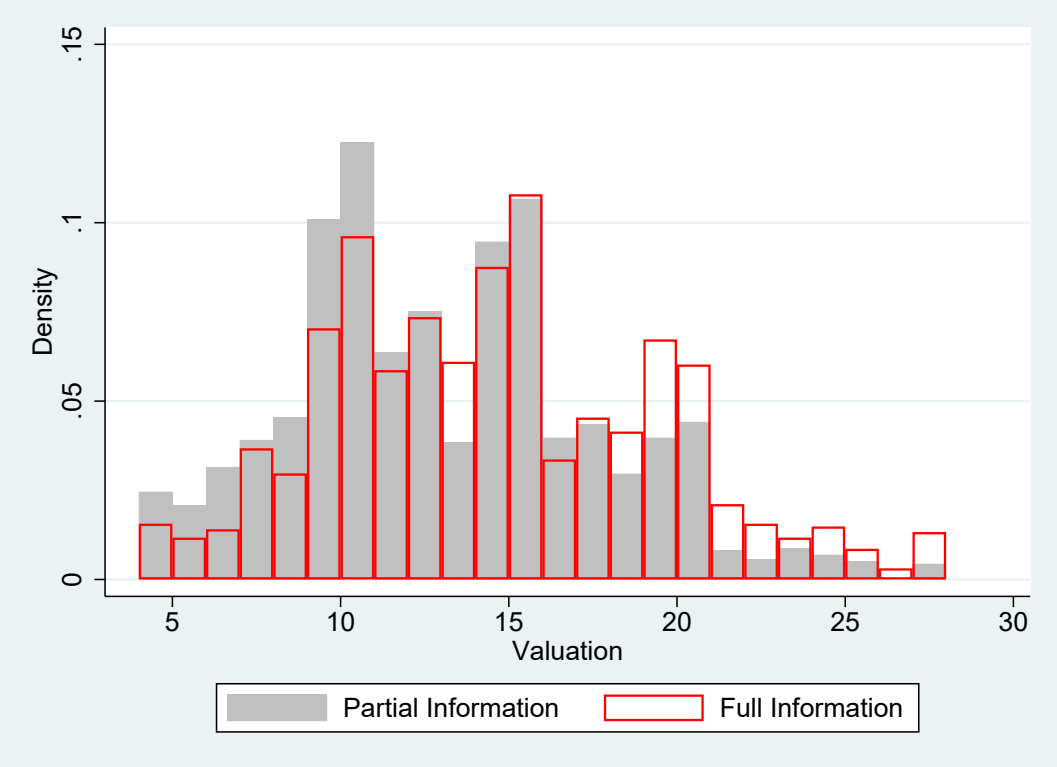
Variable	**(1)**	**(2)**	**(3)**
Sum Task Score	1.287***	1.226***	1.060***
	(0.0590)	(0.0610)	(0.0464)
Full Information	3.763***	3.735***	1.615***
	(0.712)	(0.708)	(0.464)
Sum Score * Full Info	-0.371***	-0.367***	
	(0.0882)	(0.0880)	
Worker is Male		0.601***	0.860***
		(0.154)	(0.202)
Male * Full Info			-0.561*
			(0.294)
Constant	5.178***	5.400***	6.361***
	(0.477)	(0.478)	(0.414)
Observations	2862	2862	2862

Robust Standard Errors in parentheses [* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$]
The results are from a Generalized Least Squares regression where the dependent variable is the worker valuation given by the managers. Full Information a binary variable indicating if a manager saw the subjective information as part of the worker profiles, and this is interacted with the other two variables of interest in the regressions.

That both information treatments produced similarly large errors is not due to disregard for the information in worker profiles. Table 3 regresses worker profile valuation on profile components. Columns 1 and 2 show that an extra point in the Sum Task score lead managers to increase their prediction of worker performance by 1.2 points, which is significant. The table also shows that managers valued workers more highly in the full information treatment,

and that the presence of subjective information lead to managers putting less weight on objective information (Sum Task score). That workers were given higher valuations under full information is robust even at the individual worker level: 25 out of the 27 workers received higher average predictions when subjective information was provided[14]. Figure 2 plots the distribution of values across information treatments. It shows that the presence of subjective information pushes the entire distribution to the right.

FIGURE 2. Valuation by Information Treatment



We also find a small but significant gender effect in favor of male workers (Table 3 Column 2). After controlling for their Sum Task score, men are predicted to solve 0.60 more multiplications than women (4.3% of the mean valuation). This discrimination is severely mitigated when providing subjective information (Worker is Male - Male*Full Info): the valuation-gap (i.e. wage-gap) decreases by 65%.

This mitigating effect of subjective information on gender discrimination could be caused by women writing better statements than men. We check for this by comparing the results from Task 3 of our worker sessions. In Task 3, workers were asked to predict other workers'

[14]These differences were significant at the 10% level in 15 out of the 25 cases and to the 5% level for 11 of those.

multiplication scores using only their written statements. We find that the predictions male and female workers receive are indistinguishable. Male workers are predicted to score 14.38 in the Multiplication Task, while female workers are predicted to score 15.46 (p=0.94)[15]. Given that workers didn't see the gender of who they were evaluating, this shows that the shrinking of the gender gap cannot be attributed to women writing better statements.

To further understand these results, we look at manager gender. We find that female managers, on average, increase their valuation by 2.49 points (p-value 0.001) when the subjective statement is included in a worker's profile, while men increased their valuation by just 0.58 (p-value 0.363).

TABLE 4. Information use by Manager Gender

Variable	**Female Man. Partial Info**	**Male Man. Partial Info**	**Female Man. Full Info**	**Male Man. Full Info**
Sum Task Score	0.972***	1.439***	0.598***	1.019***
	(0.0866)	(0.0785)	(0.132)	(0.0855)
Worker is Male	0.477*	0.881***	0.582	0.803***
	(0.286)	(0.260)	(0.446)	(0.301)
Subj. Info Score			0.0254	0.194***
			(0.0852)	(0.0567)
Constant	6.518***	4.448***	10.94***	5.018***
	(0.654)	(0.626)	(1.535)	(1.024)
Observations	756	828	522	756

Robust Standard Errors in parentheses [* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$]

These results are from Generalized Least Squares regressions where the dependent variable is the worker valuation given by the managers. The first two columns include only the managers that did not see the subjective statement as part of the worker profiles, while the last two correspond to the Full Information treatment.

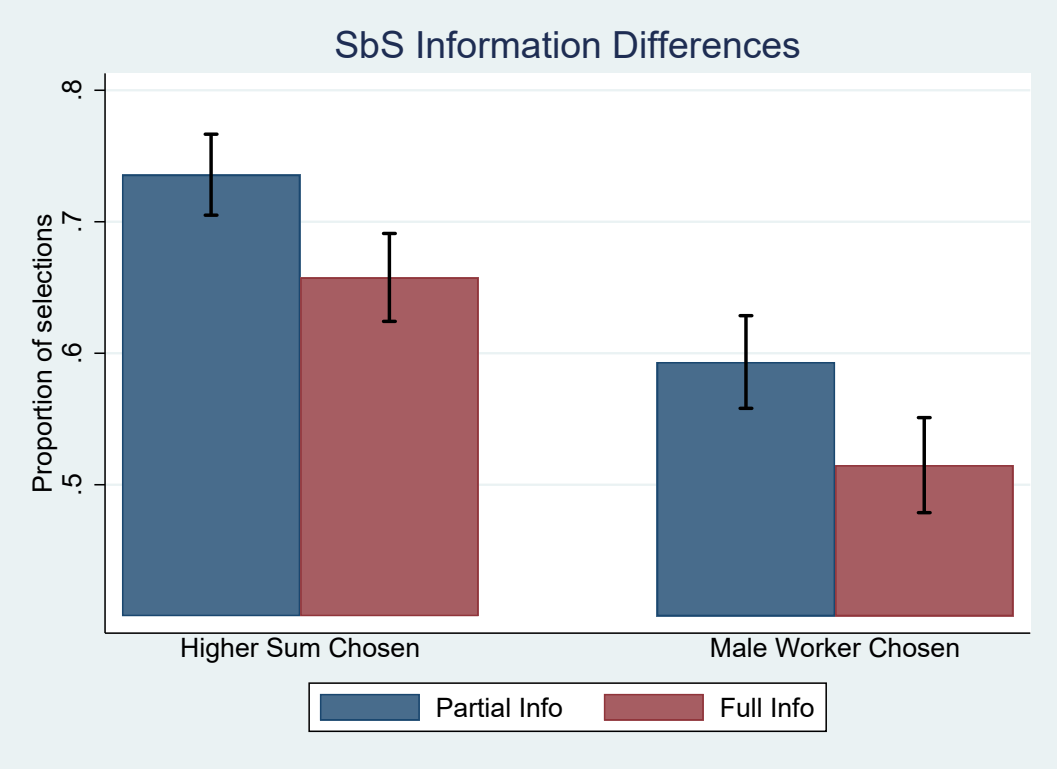
Table 4 refines the analysis of Table 3 to account for manager gender. For female managers, the increase in valuation under full information is irrespective of statement quality; the subjective information score we constructed from the worker sessions doesn't significantly explain how female managers value workers (row three in Table 4). We do find that male managers positively react to a statement's quality. Table 4 also shows that female managers put less weight on the sum task score than their male counterparts, while both genders

[15]These values correspond to the 27 worker statements that were shown to the managers. We obtain similar results on the entire worker sample.

decrease this weight when subjective information is provided. The results also show that gender-wage gap from Table 3 is primarily the result of male managers.

We now turn to results from the Side by Side comparison (SbS). We begin by looking at manager predictive ability and how it's affected by information treatment. Overall, we find that managers don't do better than chance, picking the best worker just over 50% of the time: 50.7% with subjective information and 50.1% without it (p-value 0.746). However, as in the individual valuation treatment, this isn't due manager inattention. We find that managers choose the worker profile with the higher sum task score 69.7% of the time, and that this is sensitive to the information treatment. As shown in Figure 3, when subjective information is provided, the profile with the higher sum score is chosen 65.7% of the time, compared with 73.6% of the time when no subjective information is provided (p-value 0.001).

FIGURE 3. Information effects in Side by Side comparison



The figure shows the proportion of selections of the worker with the higher Sum Task Score (left) and the male worker (right). Vertical segments represent 95% confidence intervals.

The small gender discrimination effect we found in the Individual Valuation treatment translates to more drastic results when workers are compared side by side.

Overall, male workers were chosen 55.4% of the time in male-female pairings. This bias is especially strong when men have the higher sum task score: male workers are chosen 48% of the time when both candidates have the same sum task score, but 81% of the time when they have the higher sum score. By comparison, women are only chosen 62% of the time when they have the better objective credentials.

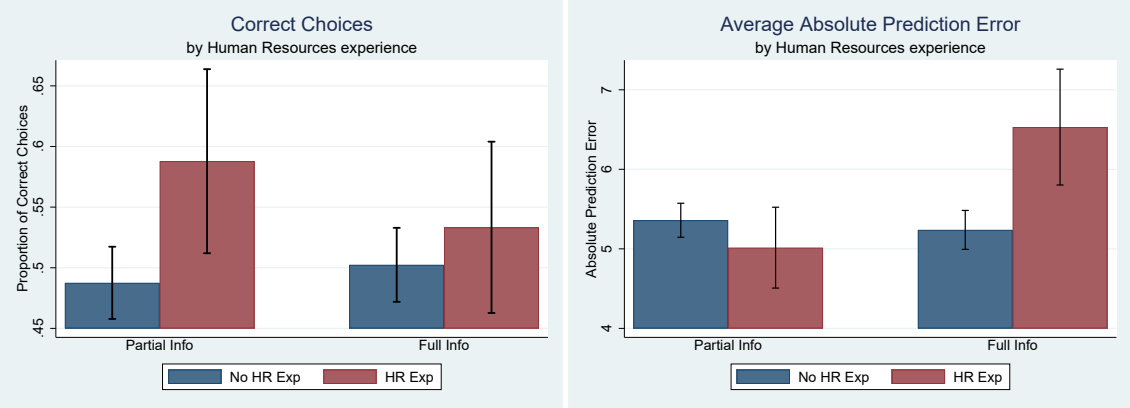
This preference for male workers is completely removed when subjective information is provided. Under full information, male workers are chosen just 51.5% of the time, which is not statistically different from 50% (p-value 0.418). By comparison, male workers are chosen 59.3% of the time when subjective information is not included. The pairings in which male workers have a higher sum score is responsible for removing all of preference for male profiles when subjective information is added; males go from being chosen 91.6% of the time under partial information to being chosen just 69.5% of the time under full information. As stated before, this cannot be attributed to women writing better statements than men.

Our experiment is also idoneous to test the hypothesis proposed in (Bohnet et al., 2015). In that paper, they find evidence that discrimination against a group perceived to have a disadvantage in a certain task is mitigated when workers are compared side by side. They use a mathematical task as the stereotypically male task, and find that a side by side comparison of worker profiles mitigates the preference for male workers they find when workers were valued individually. Considering only the 30 pairs of workers used in our side by side treatment, we constructed a counterfactual side by side choice using the managers' predictions from our individual valuation treatment; assuming that if presented with a pair, each manager would have selected the worker to which they assigned the higher valuation. In contrast to Bohnet et al. (2016). Overall, we find no significant difference in the proportions of male workers chosen from male-female pairs between the actual and the counterfactual (55.4% vs 57.2%, p=0.40). Nevertheless, we do find some evidence for said effect in our Full Information Treatment; while male workers are selected 51.5% of the time when compared Side by Side, they would have been chosen 56.3% of the time according to their valuations in the IV treatment (p-value 0.107).

A curious result arises from one question we asked our managers: 'Do you have any work experience in Human Resources?'. Though only a small number of managers answered this question in the affirmative[16], we find significant differences between 'experienced' and 'inexperienced' managers in their predictive accuracy and in how they judge subjective information.

[16] 20 of 159 in IV and 72 of 496 in SbS

FIGURE 4. Valuation Error by Experience



The left graph shows the proportion of correct choices by manager experience and Information in the Side by Side treatment. On the right, we see the average prediction error in the Valuation Treatment for the same populations. Vertical segments represent 95% confidence intervals.

Managers with HR experience do significantly better overall at picking the right candidate in the Side by Side treatment, but they are only more successful than their inexperienced counterparts when judging profiles without subjective information. When evaluating workers individually, HR and non-HR managers are equally accurate overall, but subjective information makes experienced managers' predictions significantly worse as shown in Figure 4.

Conclusion

In this paper we consider the role subjective information can play in a hiring setting. Using an experimental labor market, we're able to isolate the information channels that make studying subjectivity so difficult in field data. Our work is motivated by an interest in the efficacy of subjective information, as well as how it affects the weight managers give to objective measures of ability and to gender.

What we find is that subjective information changes the overall valuation managers assign to worker profiles, although it doesn't lead to better or worse hiring results on average. Furthermore, these results are sensitive to manager gender. Our strongest findings come from the effect of subjective information on gender discrimination. There we find that subjective information strongly reduces the well-known bias in favor of male workers when it comes to hiring for math-related tasks and it's unrelated to any potential gender differences in verbal ability. This implies that subjective information — and the type of subjective information a hiring manager collects — are potentially useful tools in combating discrimination. It also implies that hiring tasks are sensitive to differences in subjective information provided — something that researchers should keep in mind in future work. Finally, we find that experienced workers do much worse when the worker profiles they are evaluating include

subjective information, suggesting that collecting and evaluating this type of signals may be most prejudicial where it matters most.

References

Altonji, Joseph G, Anthony A Smith Jr, and Ivan Vidangos, "Modeling earnings dynamics," *Econometrica*, 2013, *81* (4), 1395–1454.

Bagger, Jesper, Fran Fontaine, Fabien Postel-Vinay, and Jean-Marc Robin, "Tenure, experience, human capital, and wages: A tractable equilibrium search model of wage dynamics," *American Economic Review*, 2014, *104* (6), 1551–96.

Baker, George, Robert Gibbons, and Kevin J Murphy, "Subjective performance measures in optimal incentive contracts," *The Quarterly Journal of Economics*, 1994, *109* (4), 1125–1156.

Bohnet, Iris, Alexandra Van Geen, and Max Bazerman, "When performance trumps gender bias: Joint vs. separate evaluation," *Management Science*, 2015, *62* (5), 1225–1234.

Bol, Jasmijn C, "Subjectivity in Compensation Contracting," *Journal of Accounting Literature*, 2008, *27*, 1.

_____, "The determinants and performance effects of managers' performance evaluation biases," *The Accounting Review*, 2011, *86* (5), 1549–1575.

Coffman, Katherine B, Christine L Exley, and Muriel Niederle, "When gender discrimination is not about gender," Technical Report, Harvard Business School Working Paper 2017.

Deming, David J, "The growing importance of social skills in the labor market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

Farber, Henry S, "The analysis of interfirm worker mobility," *Journal of Labor Economics*, 1994, *12* (4), 554–593.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental economics*, 2007, *10* (2), 171–178.

Goldin, Claudia and Cecilia Rouse, "Orchestrating impartiality: The impact of" blind" auditions on female musicians," *American economic review*, 2000, *90* (4), 715–741.

Hoffman, Mitchell, Lisa B Kahn, and Danielle Li, "Discretion in hiring," *The Quarterly Journal of Economics*, 2017, *133* (2), 765–800.

Juhn, Chinhui, Kevin M Murphy, and Brooks Pierce, "Wage inequality and the rise in returns to skill," *Journal of political Economy*, 1993, *101* (3), 410–442.

Kahn, Lisa B, "The long-term labor market consequences of graduating from college in a bad economy," *Labour Economics*, 2010, *17* (2), 303–316.

Moers, Frank, "Discretion and bias in performance evaluation: the impact of diversity and subjectivity," *Accounting, Organizations and Society*, 2005, *30* (1), 67–80.

Niederle, Muriel and Lise Vesterlund, "Do women shy away from competition? Do men compete too much?," *The quarterly journal of economics*, 2007, *122* (3), 1067–1101.

Oreopoulos, Philip, Till Von Wachter, and Andrew Heisz, "The short-and long-term career effects of graduating in a recession," *American Economic Journal: Applied Economics*, 2012, *4* (1), 1–29.

Oyer, Paul and Schaefer, "Handbook of labor economics," *Volume 4b, Chapter Personnel Economics: Hiring and Incentives*, 2011, pp. 1769–1823.

Reuben, Ernesto, Paola Sapienza, and Luigi Zingales, "How stereotypes impair women careers in science," *Proceedings of the National Academy of Sciences*, 2014, p. 201314788.

Topel, Robert H and Michael P Ward, "Job mobility and the careers of young men," *The Quarterly Journal of Economics*, 1992, *107* (2), 439–479.