ONLINE APPENDIX FOR "PAYING TO PROGRAM? ENGINEERING BRAND AND HIGH-TECH WAGES"
APPENDIX A: GENERATION OF TEXT BASED MEASURES OF IT USE AND EMPLOYER CHARACTERISTICS

I. IT MEASURES FROM RESUMES

To develop measures of technologies used on the job, we text mined the resumes that workers submitted to the job board. In their resumes, job seekers provide details about prior employment including information related to programming languages or software packages used on the job. Table A1 provides an illustrative example.

TABLE A1. SAMPLE RESUME WITH PROJECT DESCRIPTIONS

EXPERIENCE
E. xxx Corporation Long Island City. N.Y
11/2005-31/2006 Help Desk Support/Analyst
Tested Proprietary Database System
Gathered User Requirement
Assisted System Design(Design Patterns)
Coded in Java, C#

Sxxx Magazine New York, N.Y
1/2003-6/2003 Web Designer
Designed Front End/ Back End
Assisted System Analyst
Headed 4 Team members in the Project

We wrote software scripts that extracted the following fields:

1) The technologies mentioned by job seekers
2) The beginning and end years of the jobs in which they used the technology

For each year *t* and skill *s*, we computed the fraction of IT workers starting a job in that year who reported using that skill, (i.e. the ratio between 1) n_{it}: IT workers starting a job and reporting the skill and 2) n_i: IT workers starting a job):

$$share(s)_t = \frac{n_{st}}{n_t}.$$

We measure skill growth as changes in this skill share:

$$Growth(s) = Average_{t \in T}(Share(s)_t - Share(s)_{t-1})$$

We then compute whether that skill has a growth rate that is greater than or equal to the median value for all skills in that year. When workers list multiple skills in a job, we select the newest skill used by the worker in that job.

Measurement error: Because our measures are based on changes in skill counts, whether workers accurately report skills and whether the skills possessed by a worker influence her decision to post a resume can both affect measurement error. Because reporting high-growth skills may increase the likelihood of future employment, workers may over-report these skills. On the other hand, workers using emerging technologies may have less incentive to use job boards if the market is tighter for these skills. In this case, we may underestimate the growth of higher growth skills and overestimate the growth of low growth skills. When we discretize this measure, skills growing at the median rate are most vulnerable to being misclassified, which means that we may identify some new but slower growth skills as emerging skills. This source of error can lead to underestimates of the IT coefficient because it associates a low difference between a target wage and current wage to a skill mislabeled as emerging. For robustness, we used two alternative measures: 1) whether the growth is in the top 25 percentile and 2) the standardized growth value.

II. GLASSDOOR MEASURES OF EMPLOYER ATTRIBUTES

We used the Glassdoor.com data to quantify the value workers assign to employer features by text-mining the online reviews. We focused on the content in the "pros" section of the reviews, which limits the text to those attributes that employees find valuable. We used 1) automated and 2) semi-automated approaches to create different measures of employer features that we used for two separate analyses. The key distinction between these approaches was whether the set of employer features was pre-specified before starting the text mining task.

1. Automated generation of employer features: First, we used a procedure that does not rely on pre-specified employer features. This approach replicates the procedure used to mine product features in Archak et al (2011) and Ghose et al (2012). First, we assigned part-of-speech tags to keywords, "chunked"[1] sentences based on grammar rules, and identified noun phrases as candidates for employer keywords. These were implemented using the Python NLTK package. We dropped infrequent noun phrases by (1) using the algorithm from Liu and Hu (2004) and (2) removing adjectives that did not introduce new concepts. Our final feature candidates included approximately 3,000 unique lemmatized nouns or noun phrases.

Next, we clustered noun phrases based on their context, where context is defined as the five-word window preceding and succeeding each feature. The intuition behind this approach is to cluster phrases that are conceptually similar to one another using association patterns that appeared around these phrases. Each noun phrase context was treated as a separate document. K-means clustering was used to group the noun-phrases. We varied the number of clusters from five to fifteen, and finally settled on eight clusters because each feature set represented a unique employer attribute and the correlations between clusters suggest that they are reasonably orthogonal to one another.

For each cluster, we generated a list of the most frequent phrases and phrases with the top TF-IDF scores (TF-IDF in the feature context documents)[2] and assigned labels to them. In Table B.1, we list the clusters and the top TF-IDF phrases for each cluster. The labels for each topic were assigned using our own judgment, which is standard for many unsupervised machine learning tasks.

Finally, to construct employer values for each cluster measure, we computed whether the fraction of workers referencing each of these measures is greater than or equal to the median value when compared with all organizations in the sample. This transformation absorbs some of the differences in employer size. Moreover, larger companies receive more reviews, but systematic error is introduced into our measures only if workers from larger companies disproportionally favor one feature over another.

2. Manually generated labeling of the SKILLS measure: For another analysis, we manually pre-specified phrases referring to "skill development/training" (SKILL).[3] We analyzed the 2,000 most frequent keywords in the Pros section of the reviews contributed by IT workers. For each keyword, we identified phrases that referred to skills and learning at work. Then, we limited the list to bigrams that appeared at least thirty times in the corpus. In Table B.2, we provide a list of illustrative bigram terms. We coded each review as referring to SKILL if at least one of these bigrams appears in the review. For each employer, we computed the fraction of IT workers referencing SKILL terms in their reviews.

---

[1] Phrase chunking is a natural language process that separates and segments a sentence into its sub constituents, such as noun, verb, and prepositional phrases.

[2] TF-IDF is a value that increases proportionally to the number of times a word appears in a document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

[3] This approach is similar to an "Axial coding" paradigm, in which researchers use pre-defined constructs to guide how they analyze interview questions. Because of the size of the review text, we are not able to code each review. We therefore deconstruct the reviews into the smallest meaningful unit of text.

TABLE A2. CLUSTERS OF EMPLOYER FEATURES

LEARNING values, helpful colleague, learn environment, high turnover, tech support, flexibility in work schedule, high salary, pros, learn experience, summer hour, exciting, high tech, family, much stress, technical work, coworker, work schedule, near future, free train, positive
TECHNOLOGY concept, industry leader, interesting product, various project, date technology, supply, project, advance technology, different technology, engineering, software, many location, aerospace industry, automotive industry, agile development, project work, database, employee development, various technology, innovative product
FLEXIBILITY week vacation, hour shift, comp time, flexible work, remote, hours, hour day, short period of time, work time, free time, flexible work time, generous vacation, ample time, flexible work hour, vacation day, time management, amount of time, time-off policy, work hour, other week
COWORKER group of people, team of people, decent people, sense of humor, many people, knowledgeable people, diverse group of people, creative people, netapp, trench, experience people, kind people, bright, work environment, awesome people, helpful people, competent people, quality people, smart people, cool people
COMPENSATION bonus structure, ok salary, meal, medical, free access, solid benefit, federal government, balance between work, care of employee, fitness, excellent benefits, region, pet, package, company car, competitive compensation, life/work balance, company vehicle, stock purchase plan, work life balance
CAREER different opportunity, fast growth, opportunities, job opportunity, growth potential, advancement opportunity, career development, same company, opportunity for career advancement, growth opportunity, company with lot, mobility, promotions, company growth, own career, own career path, educational opportunity, room for growth, carrier, opportunity for career growth
LOCATION drama, cube, perfect place, beautiful location, beautiful place, san diego, easy place, san francisco, beach, nyc, amenity, weather, valley, small office, stable place, private office, proximity, street, other place, seattle
INDUSTRY everything, military, animal, professionalism, red tape, proud, let, paperwork, workload, competitor, face, vast majority, style, age, thats, show, income, recession, personnel, alot
Table notes: We show the top keywords in this table for each employer feature. These keywords were selected based on their TF-IDF score.

TABLE A3. BIGRAMS USED TO CONSTRUCT SKILLS MEASURE

learn lot, cutting edge, great training, great experience, opportunity learn, good training, place learn, training program, learn new, learning experience, good experience, opportunities learn, gain experience, great learning, work experience, training opportunities, learning opportunities, learn grow, experience working, get experience, edge technology, excellent training, new skills, training good, new technologies, training programs, training great, good learning, experience great, new technology, hands experience, job training, skill set, learning environment, experience good, work learn, lot experience, great technology, people learn, learning new, get learn, lots training, sales training, paid training, able learn, leading edge, want learn, learn something
Table notes: This table lists the bigrams used to create the SKILL measure used in our empirical analysis.

APPENDIX B: COMPARISONS OF SAMPLES USED IN ANALYSIS

TABLE B1: DESCRIPTIVE STATISTICS FOR SAMPLES USED IN ANALYSIS

	(1)	(2)	(3)
Worker sample	**Full regression sample**	**Employer FE sample**	**Glassdoor sample**
N	**50,628**	**11,070**	**2,414**
Current wage	73,901	72,292	72,843
Target wage	77,775	75,802	76,618
Experience	13.48	13.17	13.63
Job tenure (Years)	2.23	2.22	2.35
Emerging	0.22	0.22	0.20

Table notes: This table compares descriptive statistics of key variables for (1) full regression sample, (2) observations included in employer fixed effects regressions and (3) observations included in HRM practices regressions. Wage statistics are rounded to the nearest dollar.

TABLE B2: DEMOGRAPHIC VARIABLES FOR SAMPLES USED IN ANALYSIS

	(1)	(2)	(3)	(4)	(5)	(6)
Worker sample	**Full regression sample**		**Sample used for Employer FE**		**Glassdoor sample**	
Number of observations	**50,628**		**11,070**		**2,414**	
	N	%	N	%	N	%
A. By Gender						
Female	5,623	11.1	1,245	11.3	250	10.4
Male	23,816	47.0	5,010	45.3	1,086	45.0
Gender not reported	21,189	41.9	4,815	43.5	1,078	44.7
B. By Educational Degree						
Other	7,582	15.0	1,392	12.6	276	11.4
High School & Below	1,217	2.3	268	2.4	38	1.6
Vocational School	1,163	2.4	128	1.2	17	0.7
Two-Year Degree	5,963	11.8	1,388	12.5	310	12.8
Four-Year Degree	22,643	44.7	5,191	46.9	1,124	46.6
Graduate Degree	11,591	22.9	2,662	24.1	649	26.9
Doctorate Degree	469	0.9	41	0.4	0	0.0
C. By Ethnicity						
African American/Black	3,665	7.2	915	8.3	194	8.0
Asian	4,830	9.5	964	8.7	220	9.1
White	25,247	33.4	3,596	32.5	777	32.2
Other	16,886	49.9	5,595	50.5	1,223	50.7

Table notes: This table reports sample comparisons of gender, educational degree and ethnicity groups for (1) full regression sample, (2) observations included in employer fixed effects regressions and (3) observations included in HRM practices regressions.

APPENDIX C: SUPPLEMENTARY TESTS

TABLE C.1: LABOR MARKET FACTORS, EMERGING TECHNOLOGIES, AND WAGES

Table notes. This table analyzes how labor market factors that influence a worker's outside options can affect the worker's willingness to exchange wages for the opportunity to use more interesting IT. These cross-market comparisons should be interpreted with some caution, because the extent to which a worker and employer split the costs of various types of "training" (and therefore the sizes of the effects we observe) are, in part, a function of tightness in the labor market in which the worker is participating (e.g. Acemoglu and Pischke 1999). In this table, we test whether the patterns we observe are consistent with the idea that workers are more likely to exchange wages for skills in markets with better outside options, with the caveat that we cannot directly account for differences in supply and demand conditions across markets that might influence these estimates. The table reports worker-level regression estimates that analyze correlations between the workers' IT use, characteristics of the labor markets in which they are searching, and the difference between the target and current wage. In addition to the variables shown, all regressions control for wages, job tenure, experience, race, gender, and education. Robust standard errors are shown in parentheses; Standard errors are clustered at the geographic level in columns (1), (3), and (5); *** p<0.01, ** p<0.05, * p<0.1.

To characterize labor markets, we measured a) the fraction of local employers hiring workers with emerging IT skills and b) the ratio of the number of employers on the hiring side of the market to the number of workers with these skills. We use these two measures in different sets of regressions. In each, markets are classified according to whether they are above the median value for these measures. We created a 2 x 2 matrix between this market measure and whether or not the focal worker uses emerging IT.

Column (1) uses measures of the fraction of employers in a market hiring workers with emerging IT skills. The effects of workers' emerging IT use on target wages are statistically significant only in markets where more than the median fraction of employers hire for these skills (t=5.75). The effects in smaller markets or where workers are using mature IT systems, are insignificant. Column (2) includes fixed-effects, which compares IT workers who have the same employer and use the same category of IT system, but who are located in different markets. The interaction effect when workers have emerging IT skills and employees are hiring these skills remains positive (t=3.67), but the estimate on emerging IT in small markets turns negative. This may indicate that for smaller market establishments, wages are bid up due to a constrained supply of workers. Columns (3) and (4) use our alternative measure, the ratio of the number of employers using emerging IT systems (i.e. employers with at least one worker using the system) to the number of IT workers using emerging IT systems, which captures relative differences in the worker's bargaining power. The pattern of estimates is similar to those we observed in columns (1) and (2).

DV: Log(Target wage)	**(1)**	**(2)**	**(3)**	**(4)**
Model	**OLS**	**FE**	**OLS**	**FE**
Emerging & Above median fraction of employers use emerging IT	0.023***	0.033***		
	(0.004)	(0.009)		
Emerging & Below median fraction of employers use emerging IT	0.011	-0.092***		
	(0.007)	(0.015)		
Not emerging & Above median fraction of employers use emerging IT	-0.000	-0.003		
	(0.003)	(0.006)		
Emerging & Above median ratio of emerging IT employers to workers			0.020***	0.038***
			(0.004)	(0.009)
Emerging & Below median ratio of emerging IT employers to workers			0.016***	-0.075***
			(0.006)	(0.014)
Not emerging & Above median ratio of emerging IT employers to workers			-0.004	0.001
			(0.002)	(0.006)
Log(Wage)	0.763***	0.763***	0.763***	0.761***
	(0.005)	(0.007)	(0.005)	(0.007)
Employer FE		✓		✓
Robust SE	✓		✓	
Demographic controls	✓	✓	✓	✓
R-squared	0.736	0.686	0.736	0.685
Observations	50,628	11,070	50,628	11,070

TABLE C.2: PAIRWISE CORRELATIONS AMONG EMPLOYER CULTURE ATTRIBUTES

Table notes: This table reports the full set of pairwise correlations between the key workplace dimensions constructed from the Glassdoor review text. Correlations are reported at the employer level for 473 employers that appear both in our sample of workers and in the Glassdoor data. Details about how we produce these workplace dimensions from the online review text are provided in Appendix A, Part II. The strongest correlations in the sample are between *LEARNING* and *TECHNOLOGY*, between *LEARNING* and *CAREER,* and between *TECHNOLOGY* and *CAREER*, which supports the argument that IT access is highly valued by workers for new skills and future opportunities. *** p<0.01, ** p<0.05, * p<0.1.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) LEARNING	1						
(2) TECHNOLOGY	.139***	1					
(3) COWORKERS	-.025	.025	1				
(4) FLEXIBILITY	.085*	.037	.084*	1			
(5) COMPENSATION	.037	-.052	-.047	.100**	1		
(6) CAREER	.153***	.234***	.064	-.013	.032	1	
(7) INDUSTRY	.080*	.122***	.051	.048	.018	.103**	1
(8) LOCATION	.086*	.117**	.081*	.076*	-.079*	.003	.094**

TABLE C.3: WORKPLACE SKILLS DEVELOPMENT, EMERGING IT USE, AND JOB SEARCH

Table notes. This table tests interactions between IT systems and whether the employer is known as a good place to learn new skills, as indicated by the Glassdoor reviews. To develop a measure of whether an employer is known to be a good place to learn skills (*SKILL*), we manually selected phrases that indicate skill development at work.[4] Details of this process are available in Appendix A, Part II. The dependent variable in columns (1) and (2) is a binary variable indicating whether they used emerging IT systems or older IT systems, respectively. The dependent variable in columns (3) through (5) is the logged target wage and the regression is specified similarly to what was used in earlier tables. In addition to the variables shown in the regressions, all regressions control for wages, job tenure, experience, race, gender, and education. Standard errors are shown in parentheses; *** p<0.01, ** p<0.05, * p<0.1.

Columns (1) and (2) indicate correlations between emerging IT use and the workplace *SKILL* measure. The estimates suggest that the employer measure of *SKILL* is positively correlated with the use of emerging IT (t=1.97), but negatively correlated with the use of mature IT systems (t=-2.57). Columns (3) through (5) test our target wage specification with the *SKILL* measures. T he estimates in column (3) indicate that more is required to induce IT workers to leave firms that are higher on the *SKILL* measure, which is consistent with the view that opportunities for skill development are valuable to IT workers (t=2.00). The estimates in (4) indicate that this relationship holds only when the use of emerging IT and an emphasis on skill development are present together, although the estimate on the interaction term is significant only at the 10% level. The relationship does not hold when either of these two factors is present in isolation. In fact, the estimated coefficient on emerging IT when present alone is negative, which is consistent with firms paying a premium for workers who are already skilled in that technology. The overall pattern of estimates is similar in column (5) after adding a measure controlling for the extent to which workers note longer hours (*HOURS*) in their reviews.[5] We include this measure because some employers that are known as being good places to learn also require long hours from their workers. The estimate on *HOURS* is negative and significant (t=-2.13). The coefficient estimates on Emerging IT, SKILLS, and the interaction terms in columns (4) and (5) suggest that workers who use emerging IT and are employed at firms that are known as good places to acquire skills set their target wages about 5% higher.

	(1)	(2)	(3)	(4)	(5)
DV	**Emerging**	**Old High-growth**	**Log (Target)**	**Log (Target)**	**Log (Target)**
Model	**Logit, FE**	**Logit, FE**	**FE**	**FE**	**FE**
Emerging				-0.059**	-0.066**
				(0.028)	(0.032)
SKILLS	0.371***	-0.388**	0.016**	0.011	0.017**
	(0.119)	(0.151)	(0.008)	(0.008)	(0.008)
HOURS					-0.003
					(0.008)
Emerging x SKILLS				0.095*	0.096*
				(0.053)	(0.050)
Log (Wage)	0.183	-0.440**	0.807***	0.806***	0.831***
	(0.169)	(0.188)	(0.010)	(0.010)	(0.011)
Location FE	✓	✓	✓	✓	✓
Demographic controls	✓	✓	✓	✓	✓
R-squared			0.761	0.761	0.777
Observations	2,414	2,414	2,414	2,414	2,414

[4] We also tested variations that account for phrase negation in the review text with minimal change in coefficients.

[5] We thank an anonymous reviewer for recommending this test.

TABLE C.4: EMPLOYER CHARACTERISTICS OF WORKERS' SOURCE AND DESTINATION FIRMS

Table notes. One potential source of bias is related to the unobserved non-wage characteristics of the job for which the worker is searching. If workers who use emerging IT systems tend to target employers who offer jobs with undesirable characteristics, such as restrictive work practices or the use of old technologies, it can bias our estimates upwards. Alternatively, if they target jobs offering more desirable characteristics, it will exert a downward bias on our estimates. The fact that the job board does not permit workers to set targets for non-wage characteristics should discourage job switchers who are looking for a significant change in work context, but some workers may be looking for such changes. We cannot directly control for the non-wage component of the worker's target job, but we can probe the importance of this source of bias using the resume data to assess patterns of employment transitions between firms for which we have Glassdoor data, which we show in this table.

We use IT workers' career histories through 2011 to identify, for job transitions, characteristics of source and destination firms. For each employer in our Glassdoor sample, we compute the share of workers using emerging IT (*Emerging%)* and the value for each employer work attribute, and we label these as being High if they are greater than or equal to the median value for that feature for all firms. Then, we report the fraction of worker transitions that fall into each of the following four categories: moving from an employer with a high value of a given feature to another employer with a high value for that feature (high to high), and using similar notation, high to low, low to high, and low to low.

Panel A reports these fractions based on workers in the sample who left a *high-emerging IT* employer; Panel B is based on workers who left a *low-emerging IT* employer. For instance, the value in the first row and first column indicates that of works leaving high emerging-IT employers, 71.3% went from an employer where the value of LEARNING was high to another firm where this value was also high. 7.1% of these workers went from an employer where the value of LEARNING was high to a firm where this value was low.

The results suggest that i) workers tend to move between firms that are similar in terms of the non-wage compensation they offer and ii) the patterns are similar for workers whether or not they are using emerging IT systems. The transition patterns in the table are inconsistent with the argument there is a large pool of workers who use emerging IT systems and are seeking to move to jobs with undesirable non-wage characteristics.

	(1)	(2)	(3)	(4)
	High to High	**High to Low**	**Low to High**	**Low to Low**
A. Percentage of workers leaving *High-Emerging IT* Employers, N=86,219				
LEARNING	71.3	7.1	2.5	19.2
TECHNOLOGY	61.9	6.2	4.0	27.9
COWORKERS	52.4	5.6	4.2	37.8
FLEXIBILITY	38.3	4.5	5.5	51.8
COMPENSATION	36.9	4.9	5.1	53.1
CAREER	61.0	6.7	3.4	28.9
INDUSTRY	61.8	13.4	3.7	21.2
LOCATION	45.0	4.9	5.1	45.0
B. Percentage of workers leaving *Low-Emerging IT* Employers, N=80,942				
LEARNING	26.6	5.8	9.9	57.7
TECHNOLOGY	35.2	6.4	8.1	50.3
COWORKERS	43.1	5.6	8.3	43.0
FLEXIBILITY	57.4	8.2	7.3	27.2
COMPENSATION	52.6	6.9	7.6	33.0
CAREER	30.2	5.1	8.1	56.6
INDUSTRY	25.8	5.4	9.4	59.5
LOCATION	48.3	7.7	8.0	35.9

REFERENCES

Acemoglu, D., & Pischke, J. S. (1999). Beyond Becker: training in imperfect labour markets. *The Economic Journal*, *109*(453), 112-142.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1,485-1,509.

Ghose, A., P. Ipeirotis, and B. Li. "Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content." *Marketing Science* 31, no. 3 (2012): 493-520.

Liu, B. and Hu, M. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.