You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks

Christopher G. Harris
Informatics Program
The University of Iowa
Iowa City, IA 52242

christopher-harris@uiowa.edu

ABSTRACT
Many human resource tasks, such as screening a large number of job candidates, are labor-intensive and rely on subjective evaluation, making them excellent candidates for crowdsourcing. We conduct several experiments using the Amazon Mechanical Turk platform to conduct resume reviews. We then apply several incentive-based models and examine their effects. Next, we assess the accuracy measures of our incentive models against a gold standard and ascertain which incentives provide the best results. We find that some incentives actually encourage quality if the task is designed appropriately.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software - *Performance Evaluation*

General Terms
Measurement, Design, Experimentation. Human Factors

Keywords
Relevance judgment, crowdsourcing, incentive models.

1. INTRODUCTION
One challenge companies constantly face is increasing worker productivity without substantially increasing costs. Recent technology has aided many of these productivity gains; however, the most elusive gains are those related to tasks that are repetitive, subjective, and not easy to define algorithmically.

A case in point is a company's human resources (HR) department, responsible for the new employee recruitment. The typical HR recruiter looks through an average of 200 resumes to fill a single mid-level position; for highly-desirable positions, they can receive ten times this number to review [11]. Technology can help with the search process to discover thousands of online resumes, but is yet unable to make the subjective assessment of which are adequate resumes for a job versus an inadequate one.

WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), Hong Kong, China, Feb. 9, 2011

Frequently the task of hiring mid-level employees and above is outsourced to executive search firms. These outside recruiters typically charge around one third of the annual base salary of a newly-hired employee. Therefore an inexpensive method of examining resumes can benefit employers or outside search firms cut costs substantially if this activity can be done effectively.

Crowdsourcing tools such as Amazon's Mechanical Turk[1] (AMT) show considerable promise in having simple yet tedious tasks executed rapidly. These platforms provide a legion of available Internet workers to complete HITs (Human Intelligence Tasks) in exchange for micro-payments – precisely the type of activity that can help HR recruiters narrow a pile of resumes to only those of interest. By dividing a tedious task among a large number of participants, a company can quickly and inexpensively execute tasks in a short timeframe, often within 24 hours.

Our objective is to examine how platforms such as AMT can do well with the types of subjective evaluations computers cannot perform well. Additionally, we wish to examine the role incentives play in aligning the worker's needs with those of the requester in this anonymous environment.

The remainder of this paper is organized as follows: In the next section, we briefly discuss the background of this emerging area. Next, we describe our experiments and the incentive-based variants of each model. In Section 4, we present our results. Finally, we summarize our findings and indicate our anticipated future work in this area.

2. BACKGROUND
The use of technology in the job search process is certainly not new. Companies such as Monster.com[2] and Jobing.com[3] make use of technology to aid in the indexing, searching and dissemination of resumes. More recently, online recruitment firms have made use of more advanced techniques such as semantic search. Some have even forayed into aspects of crowdsourcing. Previously, job search website TalentSpring[4] had job seekers rank 12 pairs of resumes in a specific professional niche, selecting which candidate is preferable [6]. This technique introduces potential bias – can job seekers be expected to fairly rate their anonymous competitors when a potential job is at risk?

[1] www.mturk.com
[2] www.monster.com
[3] www.jobing.com
[4] www.talentspring.com

Additionally, even if care was exercised to ensure they are not competitors for a specific position, what would encourage these job seekers to make an accurate assessment of another candidate?

The use of incentive models in behavioral economics has been well-studied, but none of this research has covered incentive models applied to anonymous workers that incorporate worker quality measures. With this in mind, we recognize that this resume selection task is a relevance judgment task and may be ideal for crowdsourcing. Recent research has demonstrated the many benefits of this approach to tasks such as annotating images [10], relevance judgments [2], tracking sentiment [1], and for translation tasks [9]. Likewise, the corporate world has embraced it for soliciting feedback and creative purposes [4], such as designing advertising campaigns, user studies, and or designing a corporate logo [1].

In contrast, there are also several well-discussed drawbacks, such those discussed in [1] [3] and [8] regarding poor or indifferent worker quality and potentially malicious worker intent. Moreover, when unqualified workers perform a judgment task, care must be taken to prevent noisy data as discussed in [5].

This tradeoff raises some important questions: First, can workers with little or no training be used to rate the resumes of job candidates effectively? Also, do some judgment models work better than others? Finally, is there a way to motivate workers through positive or negative incentives? We address these considerations through specifically-designed experiments in the next section.

3. EXPERIMENTAL DESIGN

Our objective is to examine one of the most laborious steps of the hiring process – the resume review – and examine its fit to crowdsourcing. We begin with three actual management-level job descriptions, one for a Human Resources Manager in a financial services company, one for a National Sales Manager in a manufacturing company, and one for a Project Manager in a chemical company. These descriptions were provided by an executive search firm along with 16 applicant-submitted resumes for each of these positions.

After removing all contact information for each of the 48 candidates and anonymizing both the job descriptions and the resumes to alter any potentially-identifiable information, we then replaced all acronyms in the documents with the corresponding terms. We concentrated on management-level positions for three reasons: first, this data was available to us; second, there is more work experience and educational history provided by the candidates for evaluation, and third, this level of candidate represents the largest portion of a recruiter's workload and this is ripe for potential cost-savings through crowdsourcing.

For each of the following eight bundled HITs, we required a brief qualification process to ensure English ability and an AMT approval rating of at least 95%. Each HIT began with 100 participants who passed this initial qualification step. Participants were unable to participate in more than one HIT and had to complete all 48 ratings to have their answers considered for this study.

To ensure participants were not "gaming the ratings", or providing answers without careful consideration simply for compensation, as described in [3] and [8], we included some additional straightforward free-form questions about the job descriptions to ensure attention to detail. Our participants were prompted for basic information after the fifth and tenth rating for each of the three job descriptions, and the participants were not considered if the answers to these six questions indicated a participant had not read the job description carefully – a subjective assessment made by us based on their responses. We used the AMT (Amazon Mechanical Turk) platform for all of our experiments.

3.1 Resume Relevance HIT Design

Participants were asked to evaluate the fit of each resume to the job description on a five-point scale, from a score of 1 (non-relevant) to 5 (highly-relevant). The same anonymized information was provided to a HR Hiring Director with 14 years of experience in management-level executive search, who evaluated these resumes on the same five-point scale. These ratings were used as our gold standard.

3.1.1 Baseline Resume Relevance

Participants in this HIT were provided with 48 resumes to evaluate and were compensated $0.06 per question. No incentives were offered to participants based on their ratings.

3.1.2 Resume Relevance with Positive Incentive

Compensation in this HIT was set as $0.06 per rating; however, each participant was initially told that each resume had already been rated by an expert and if the participant's rating matched the expert's, they would receive a post-task *bonus* payment of $0.06, providing for the possibility of earning $0.12 per rating.

3.1.3 Resume Relevance with Negative Incentive

Compensation in this HIT was set as $0.06 per question; however, each participant was also told a previous expert rating had been made. If the participant's rating differed from the expert's, their compensation would be *reduced* to $0.03 for that rating.

3.1.4 Resume Relevance with Combined Incentives

Compensation for this HIT was set at $0.06 per rating. Participants were told a previous expert rating had been made. They were paid a *bonus* of $0.06 if it matched; however, if their rating differed from the expert's in more than half of the 48 resumes rated, compensation was *reduced* to $0.03 per rating for those which differed; therefore compensation could range from $1.44 (having all ratings differ) to $5.76 (having all ratings match our gold standard).

3.2 Resume Screening HIT Design

In this HIT, we wanted to examine the ability for crowdsourced workers to perform an initial screening of resumes. We included each of the three job descriptions and one resume for each of the 16 candidates for that position; the participant had to mark each resume in one of two ways: either as relevant or non-relevant. For our gold standard, we took the 17 resumes with ratings of 4 or 5 from our HR director as 'relevant'. Participants were unaware of the number of resumes that were determined relevant.

3.2.1 Baseline Resume Screening

Participants completing the HIT successfully were paid $0.06 per rating. No incentives were offered to participants.

3.2.2 *Resume Screening with Positive Incentive*

Compensation was set as $0.06 per rating. As with the Resume Relevance HIT, participants were notified about the potential of earning a *bonus* payment of $0.03 per rating if their rating matched the one made by our expert.

3.2.3 *Resume Screening with Negative Incentive*

Compensation in this HIT was set as $0.06 per question. Each participant was also told that if their rating differed from our expert's, compensation would be *reduced* to $0.03 for that rating.

3.2.4 *Resume Screening with Combined Incentive*

In this HIT, participants were paid $0.06 per rating, and told they could earn a *bonus* of $0.06 for each expert rating they matched; however, if their rating differed from the expert's in more than half of the 48 resumes rated, their compensation was *reduced* to $0.03 per rating for all ratings which differed.

Although in four of the HITs we clearly indicated to participants in advance that we would reduce their compensation if they failed to match the expert ratings, in actuality no participant compensation was reduced.

4. DATA ANALYSIS

Approximately 87% of all task participants passed our qualification exercise (i.e., they supplied coherent answers to our six free-form questions). This passing percentage was fairly consistent across all eight examined HITs. As expected, the average time taken for each Resume Relevance rating was significantly higher than the Resume Screening rating.

4.1 Resume Relevance

The distribution of ratings in all four Resume Relevance HITs was roughly normal as shown in Figure 1. Like our gold standard, the positive incentive model showed a positive bias (skewed right). The negative incentive model was much tighter around the mean (smaller variance). This may indicate that participants with positive incentives may rate job candidates more highly, whereas those with negative incentives take a far more conservative approach. The combined incentive model showed a mix of these effects (positive bias but with a smaller variance).

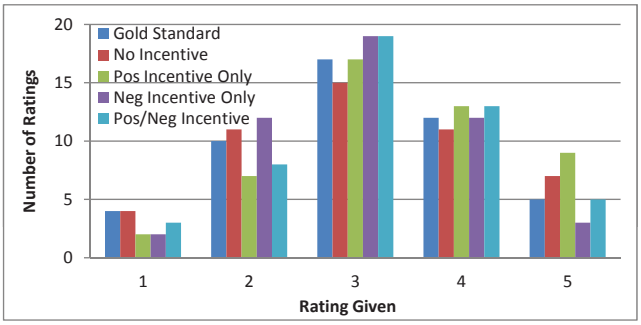

Figure 1. Rating distribution of for the Resume Review HIT

A more important issue was the degree to which the participant's ratings matched our gold standard. As observed in Figure 2, the best matches to our gold standard were the positive model and combined incentive models. Since the granularity of a five-point scale may be too fine, we divide the judgments into two resume judgment groups: scores of 4 or 5 to be 'accepts' and 3 or less to be 'rejections' and compare this with our gold standard. We can then calculate the recall, precision and F-scores for each model (provided in Table 1). Again we find all three incentive models are an improvement over the baseline, with the positive and combined incentive models performing best.

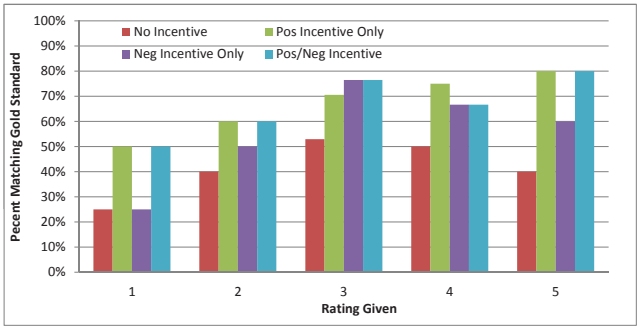

Figure 2. Distribution of Resume Review Ratings Matching the Gold Standard

Since the granularity of a five-point scale may be too fine, we divide the judgments into two resume judgment groups: scores of 4 or 5 to be 'accepts' and 3 or less to be 'rejections' and compare this with our gold standard. We can then calculate the recall, precision and F-scores for each model (provided in Table 1). Again we find all three incentive models improve upon our baseline and the best performers were the positive and combined incentive models.

Table 1. Accuracy Measures for the Resume Review HIT

Incentive Model	Recall	Precision	F-Score
None	0.32	0.47	0.38
Pos	0.54	0.76	0.63
Neg	0.48	0.65	0.55
Pos/Neg	0.55	0.71	0.62

In all HITs, the 48 resumes to be ranked were roughly the same length. By examining the time taken to rank them, we can ascertain a rough metric on each model's encouragement for attention to detail. We were surprised to see the difference in magnitude our incentive models had on each participant's time to complete each rating. Figure 3 illustrates this difference, showing the HIT response time (y-axis) varies as the participant moves through a group of 16 resumes matching a single job description (x-axis).

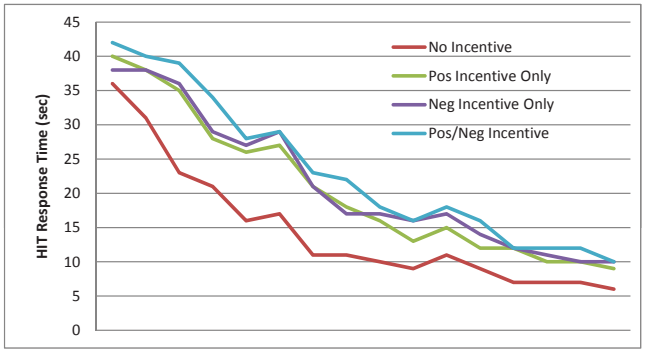

Figure 3. Time taken per rating in the Resume Review HIT.

The three incentive models show a markedly higher response time compared with the baseline model. We believe that the higher rating accuracy for the incentive models and greater response times is likely due to participants with incentives intentionally making more careful decisions.

4.2 Resume Screening

Our Resume Screening HIT was a simple binary judgment and therefore our interest was to investigate which of our models best matched the gold standard. As observed with the Resume Review HIT, the combined and positive incentive models perform best, followed by the negative incentive model. Table 2 illustrates the summary recall, precision and F-score for each model.

Table 2. Accuracy Measures for the Resume Screening HIT

Incentive Model	Recall	Precision	F-Score
None	0.33	0.47	0.39
Pos	0.67	0.82	0.74
Neg	0.54	0.68	0.60
Pos/Neg	0.78	0.82	0.80

All three incentive models performed significantly better than our baseline, non-incentive model, and are similar to those obtained in our Resume Review HIT. We note that the recall measure – arguably more important than precision for our relevance judgment task – is significantly higher for both the positive and the combined incentive models. This further demonstrates the strength of incentives, even when used for simple binary judgments.

The time to complete the Resume Screening HIT showed a similar gap between the three incentive models and the non-incentive model, although the gap was not as pronounced. As with the Resume Review HIT, this likely indicates a higher attention to detail relative to the non-incentive model.

5. CONCLUSION

This preliminary study examined the use of crowdsourcing in resume review and examined the effects of incentives on participant's accuracy in rating resumes. We observe that these platforms, when the correct incentives are offered, can provide a method of classifying resumes. We also discover that incentives encourage participants to make more accurate judgments.

Although none of the examined incentive models perfectly matched the gold standard in our resume rating assessments, we observe that incentives in general have promise in crowdsourcing activities. Positive and combined incentives are best to encourage more careful consideration of tasks compared with no incentives. These observations applied equally to the five-point ratings in our Resume Review and our binary Resume Screening task.

6. FUTURE WORK

We plan to explore other relevance judgment methods, such as pair-wise preference, respond to incentive models. Additionally, we plan to examine if the size and frequency of the incentive offered has an impact on our results. We also plan to extend the size of our study to incorporate additional raters, examine some of the demographic aspects of our participants, investigate how the clarity of instructions affect participant performance, and examine what is the appropriate task length to achieve the best results. We also plan to examine how crowdsourcing can compare with many machine learning methods. Finally, we plan to examine methods to limit the amount of noisy data in our results.

7. REFERENCES

[1] Alonso, O. 2009. Guidelines for designing crowdsourcing-based relevance evaluation. In *ACM SIGIR*, July 2009.

[2] Alonso, O., Rose, D. E., and Stewart, B. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9-15, 2008.

[3] Donmez P, Carbonell J.G., and Schneider J. 2010. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In: *Proceedings of the SIAM Conference on Data Mining (SDM 2010)*, 826–837

[4] Howe, J. The rise of crowdsourcing. *Wired Magazine 14*, 6 (2006).

[5] Hsueh, P., Melville, P., and Sindhwani, V. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing* (HLT '09). Association for Computational Linguistics, Morristown, NJ, USA, 27-35.

[6] John Cook's Venture Blog. http://blog.seattlepi.com/venture/archives/115549.asp. Retrieved on Nov 16, 2010.

[7] Kelly, P.G. 2010. Conducting usable privacy & security studies with amazon's mechanical turk. In *SOUPS '10: Proceeding on Symposium on User Privacy and Security.* Redmond, WA. July 14-16, 2010.

[8] Kittur, A, Chi, E. H. and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, 453-456.

[9] Negri, M. and Mehdad, Y. 2010. Creating a bi-lingual entailment corpus through translations with Mechanical Turk: $100 for a 10-day rush. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (CSLDAMT '10). Association of Computational Linguistics, Morristown, NJ, USA, 212-216.

[10] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J. 2010. Collecting image annotations using Amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (CSLDAMT '10). ACL, Morristown, NJ, USA. 139-147.

[11] Recruiting Blogs. http://www.recruitingblogs.com/profiles/blogs/talentspring-secures-16. Retrieved on Nov 16, 2010.

[12] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. 2008. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference for Empirical Methods in Natural Language Processing* Conference (EMNLP '08). Association of Computational Linguistics, Morristown, NJ, USA, 254-263.

[13] Yang, J., Adamic, L.A., and Ackerman, M.S. 2008. Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. In *Proceedings of the 9th ACM conference on Electronic commerce* (EC '08). ACM, New York, NY, USA, 246-255.