Hot or Not? Crowd-Sourced Evaluation of Computer-Generated Outfits

Lucy E. Dunne, Vivian Zhang, Loren Terveen
University of Minnesota, USA
Key words: Wardrobe, crowdsourcing, evaluation, methods

Introduction: The ritual of deciding what to wear is a decision-making process engaged in daily by most members of many societies. Our research focuses on the time- and resource-constrained daily dressing decision-making process, and seeks to alleviate the difficulty of this decision through the development of a smart system capable of providing outfit recommendations. To operate effectively, such a system must be capable of providing good recommendations.

The assessment of "goodness" or quality of an outfit currently has no standardized criteria that can be easily implemented in the development of a smart recommender system. Established theories of design principles and variables involved in affecting the quality of an outfit are in widespread use and are highly evident in advice literature (for example Davis, 1996). However, these principles are in general not empirically validated, and are complex to implement in assessment. The dominant assessment methodology is expert evaluation, which unfortunately is time-consuming and impractical to implement on a very large scale. Crowdsourcing, a method of using the human brainpower of everyday individuals engaged in short-term tasks of low complexity, offers a promising alternative to expert evaluation (Kittur, Chi, & Suh, 2008). Our research asks whether expert evaluation can be replicated from a crowd, and whether some crowds may be more capable than others of effective evaluation. In this research, we assess three crowd sources in evaluating computer-generated outfits: recruiting human evaluators via social networking within a University apparel department, Facebook advertising using "apparel" and "fashion" interests as filter terms, and Amazon Mechanical Turk (MT) using no pre-requisite filter. In addition, evaluations of three three "expert" apparel professionals were used as the comparison measure.

Table 1: Cost, duration, and ratings generated by three recruitment methods

	Apparel department social networking	Facebook	Mechanical Turk
Total Cost	$100.00	$162.42	$88.00
Total Time (Days)	9.5	7	0.7
Total raters	39	17	50
Total ratings	1165	637	2000

Method: A random sample of 975 outfits selected from the 491,185 possible combinations of the 137 garments in a single user's wardrobe was assessed using raters recruited from each of our three crowds (Table 1). Raters evaluated as many outfits as they were willing to on the following scale:

5: This is a great outfit; I can imagine someone looking good wearing exactly this.
4: This is an ok outfit. It might have some style problems, or it might be a little bland, but it's wearable.
3. This is a wearable outfit, but it has some problems. These garments could technically be worn together, but the outfit doesn't work very well.
2: This outfit has serious problems, it's hard to imagine someone wearing it but a few people might.
1: This outfit is not wearable; I can't imagine anyone wearing it in public.

For the 975-outfit sample, expert investigators had perfect agreement on 381 outfits (39.08%.) 767 outfits (78.67%) had agreement with a standard deviation less than one (equal to two investigators in agreement, and the third investigator differing by one value.) To eliminate outfits with less consensus, we removed the outfits with SD > 1. Once the oufits with poor expert consensus were removed from the evaluated sample, a total of 810 outfits were considered.

Results: Crowd-sourced raters completed a survey in addition to their ratings, which gathered information on their geographic location, consumer spectrum score (using questions derived from [REF]), and self-reported experience with apparel/fashion. Very little effect of location, experience, or consumer-spectrum score on accuracy of ratings (with respect to expert consensus ratings) was measured. We further explored two additional variables: crowd-sourced ratings of "difficult" outfits (outfits with poor consensus among expert raters), and extraction of a 10-rating sub-sample for use as a predictor of overall rating behavior. The former (difficult outfits) showed little relationship to accuracy of ratings. The latter showed a much more significant relationship. To assess the 10-rating sub-sample, we extracted the first 10 ratings from each rater who had rated more than 17 outfits (n=26). The average difference of these 10 ratings from the expert consensus for each rating was averaged, and that average was then compared to the rater's overall average difference (from expert consensus), and the correlation between 10-outfit average and overall average was computed, as well as the ratio of 10-rating sample to all ratings.

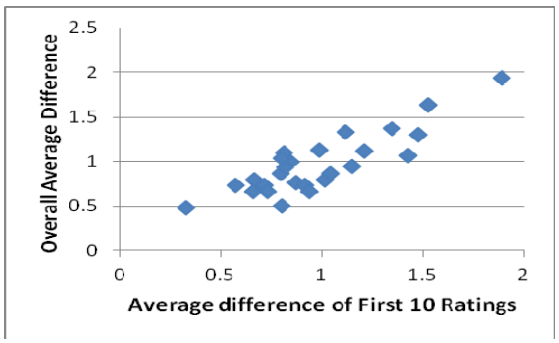

Figure 1: 10-rating sample vs. overall

A correlation with r-sq=.59 (p=0.00) and ratio of 1.02 was found between these variables (Figure 1). A subsequent second experiment was performed to expand this assessment to 64 MT raters and evaluate the prediction of their 10-rating sub-sample to a smaller total sample of 20 outfits. Results were confirmed, with a correlation of r-sq=0.67 and ratio of sub-sample/whole of 0.93.

Conclusions: We find that Amazon Mechanical Turk is the fastest, cheapest, and most direct recruitment tool for crowdsourced evaluators. Seeking a more selectively recruited sample using demographics and user interests was not effective. Not all crowd-sourced evaluators share the perspective of experts, but we find that using a 10-sample diagnostic sub-set to seek similarity in perspective is an extremely effective tool in recruiting a specialized crowd of evaluators. Further, we believe this method can easily be extended to use in filtering crowds for any expert perspective or specific type of subjective assessment perspective, and can significantly reduce the cost and difficulty of implementing expert assessment in qualitative tasks.

References

Davis, M. (1996). *Visual Design in Dress, 3rd Edition* (3rd ed.). Prentice Hall.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the 26th annual CHI conference on Human Factors in Computing Systems,* p. 453.