Problem with Cross-Cultural Comparison of User-Generated Ratings on Mechanical Turk

Hao-Chuan Wang, Tau-Heng Yeo
National Tsing Hua University
Hsinchu, Taiwan
haochuan@cs.nthu.edu.tw,
yeosblue@gmail.com

Syavash Nobarany
University of British Columbia
Vancouver, BC, Canada
nobarany@gmail.com

Gary Hsieh
University of Washington
Seattle, WA, USA
garyhs@uw.edu

ABSTRACT

Many online services serve diverse populations spanning many countries and cultures. Some of these services rely on user-generated ratings to curate and filter information, or to inform other users. However, little is known about how various cultural biases and cross-cultural differences affect such ratings. We studied how Indian and American workers on Mechanical Turk differ in their response styles by asking them to rate three products. We also explored several dimensions of cultural differences including social orientation (individualism vs. collectivism), social desirability, and thinking style (holistic vs. analytic). We found that Indian workers tended to use higher ratings on all items, including both product ratings and the different survey instruments. We discussed the implications for collecting ratings from culturally diverse populations, and for cross-cultural studies on Mechanical Turk.

Author Keywords

User-generated ratings; cross-cultural response style

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

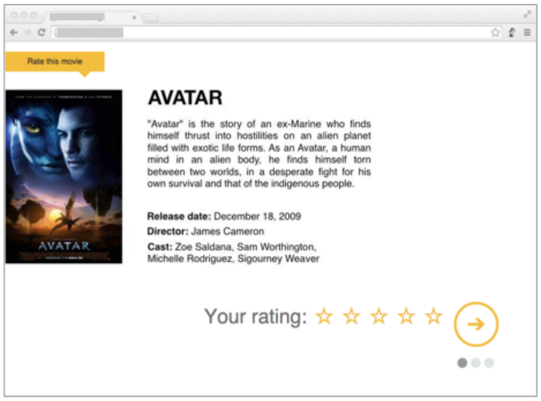
INTRODUCTION

User-generated rating is a core mechanism that supports social computing. Examples can be found in numerous applications, such as asking users to rate informational items (videos, posts, news etc.), objects and items (e.g., products) or services (e.g., restaurants). The role of user-generated ratings in social computing is versatile. It can serve as a window to understand users' experiences and perceptions, or as a source of information and a method for systems to recommend or filter out information for users.

As services leveraging user-generated ratings continue to expand to a global scale, both opportunities and challenges emerge. On the one hand, collecting user-generated ratings globally may help us benefit from using a more diverse and inclusive user pool, which can then help derive a more generalizable result on items to evaluate. On the other hand, collecting user-generated rating from a diverse global population may encounter a challenge originating from cross-cultural discrepancy in evaluations and responding behaviors. Individuals from different cultures may or may not share the same biases that shape how to evaluate an item as well as how to report their evaluation using a rating scale. In other words, it is worth asking: is a "five-star" rating in India equivalent to a "five-star" in North America?

Previous studies have found evidence of cultural differences in responding styles [4][6][9]. Common findings include that individuals with an Asian cultural background tend to be more moderate, using the midpoint of a rating scale more frequently than the Western counterparts [4][6]. Some studies also noted that East Asians tend to agree rather than disagree with the statement suggested by a rating item [9].

Response artifacts, if present, can pose threat to the validity of user-generated ratings and systems that operate upon these ratings. In social computing, user-generated ratings are often used as social traces to help systems make recommendations and to help users navigate the information space. Misrepresenting one's evaluation with



AVATAR

"Avatar" is the story of an ex-Marine who finds himself thrust into hostilities on an alien planet filled with exotic life forms. As an Avatar, a human mind in an alien body, he finds himself torn between two worlds, in a desperate fight for his own survival and that of the indigenous people.

Release date: December 18, 2009
Director: James Cameron
Cast: Zoe Saldana, Sam Worthington, Michelle Rodriguez, Sigourney Weaver

Your rating: ☆ ☆ ☆ ☆ ☆

Figure 1. The experimental rating site for collecting user-generated ratings on items.

Chinese CHI '15, April 18 - 19, 2015, Seoul, Republic of Korea
Copyright 2015 ACM 978-1-4503-3464-8/15/04···$15.00
http://dx.doi.org/10.1145/2739999.2740001

stylistically modified ratings thus can be a source of misinterpretation and system failure.

In this paper, we explore possible cross-cultural differences on user-generated ratings between workers from United States and India using Amazon's Mechanical Turk. We asked workers to rate three items from a controlled set of pre-selected items on a website we built (see Figure 1). After rating the items, we asked workers to respond to a set of questionnaires designed for measuring individualism-collectivism, thinking style, and social desirability.

We found that Indian workers gave higher ratings on items than US workers, regardless of item type and their experience with the item. Furthermore, Indian workers give significantly higher ratings on numerous self-reported survey measures. Most notably, Indian workers rated higher on *both* collectivism and individualism scales. These results suggest that Indians crowdsourcing workers in general, provide higher rating than workers based in the U.S. We discussed the problem of collecting user-generated ratings cross-culturally on Mechanical Turk.

CULTURAL DIMENSIONS

Two important aspects of how individuals from various cultures differ are how individuals think of themselves in relation to other people (individualism-collectivism) [16] and how they process information and reason (analytic versus holistic thinking) [5].

Collectivism refers to the social orientation in which one values group harmony over individual goals and interests [16]. It can be a country-level property [8] or internalized as a property of individuals [16]. Individuals with high collectivism tend to consider other people's opinions and behaviors when making their decisions [11]. In contrast, individuals with high individualism tend to focus more on themselves, considering relatively less about others' thoughts or behavior. According to the most recent result of Hofstede's cultural indices, US has a higher individualism score than India at the country level [7].

Another way cultures tend to differ is how individuals perceive and process information. Numerous studies using various methods have found cultural differences in thinking styles, of which Asians tend to be more holistic while Westerners tend to be more analytic [5][12]. Holistic thinking style refers to the tendency to reason based on the relations and interactions between objects, and to distribute attention between focal and peripheral information. In contrast, analytic thinking style refers to the tendency to reason based on rules, and to pay attention solely to focal objects or issues.

Theories and studies suggest that cultural thinking style and social orientation are highly related. Collectivism and holistic thinking style tend to co-occur [17]. Based on the background offered by previous research, we expect that Indian workers tend to be more collectivistic, less individualistic, and of higher tendency towards holistic thinking than US workers. So we expect Indian workers to be behaviorally more moderate, and to give less extreme ratings (e.g., neural or median rating) than US workers.

METHOD

In this study, we recruited Indian and U.S. MTurk workers to rate a set of three items in a 1-to-5 star scale using an experimental rating site we built. After the rating task, we also asked workers to complete a number of surveys to collect measures of cultural dimensions and demographics pertinent to our exploratory research question, concerning cross-cultural differences on user-generated ratings.

Design of the Study

After entering our site, three items were presented in three consecutive pages, one per page. For each item, the page shows a description of the item, including its name, properties and an image of the item (see Figure 1).

Our item pool included a set of eight different items: four high-grossing movies, two cameras and two books. For each participant, a subset of three items were sampled and presented for rating. Considering that participants may or may not have previous experience with the items, after rating each item, we also asked them whether they have ever experienced or used the item they just rated. In our statistical analysis, we use this information to account for the influence of prior experience with the items.

After completing all the three ratings, participants were asked to complete a post-experiment survey consisting of several measures of cultural dimensions and demographics.

We compiled and posted the item-rating task using the website we created on Amazon's Mechanical Turk. We recruited a total of 55 workers residing either in US or India. Among the workers, 32 (58.2%) reported that they are living in the US, and the rest 23 (41.8%) reported that they are living in India. In terms of gender, 33 (60%) of them are males, and 22 (40%) are females.

To determine if a worker was paying attention, we included a question: "*If you are reading this, select 2 as the response to this question.*" on a 7-point Likert scale in the middle of the survey. Workers who chose an answer other than a 2 were removed from our analyses. Only 2 of the workers were removed due to this.

Measures

We collected two types of measures in the study: rating on items, and self-reports of cultural values and attitudes, including: individualism-collectivism [1], holistic thinking style [5], and social desirability [14]. Collecting these measures allow us to explore more deeply how culture may affect user-generated ratings.

Rating

Rating is the number of stars that a worker chooses and applies to an item to rate.

Individualism and Collectivism

To assess workers' social orientation we used an individualism-collectivism survey adapted from [1]. There are eleven items used in our survey, six for assessing individualism (e.g., *"The most important thing in my life is to make myself happy."*), and five for assessing collectivism (e.g., *"What I look for in a job is a friendly group of co-workers."*). Cronbach's alpha reliabilities for individualism and collectivism were .56 and .66 respectively.

Holistic Thinking Style

To measure thinking styles, we used an eight-item questionnaire sampled from the holistic thinking style instrument developed by [5]. Sample items include *"It is more desirable to take the middle ground than go to extremes"* and *"The whole is greater than the sum of its parts."* The reliability Cronbach's alpha is .78.

Social Desirability

We also measured social desirability, the tendency for one to over-report socially desirable attitude/behavior or under-report socially undesirable ones, using the common Marlow-Crowne social desirability scale [14].

RESULTS

Rating

To analyze the rating data, we used a linear mixed model ANOVA that controlled for type of item and previous experience with the item, as well as random effects of users to account for the influence of repeated measuring.

Surprisingly, we found that Indian participants gave significantly higher ratings (mean=3.85) than US participants (mean=3.45), $F(1,58.2)=3.96$, $p=.05$.

Individualism-Collectivism, Thinking Style and Social Desirability

In terms of individualism-collectivism orientation, Indian participants' level of collectivism (mean=5.44) is higher than US counterparts' (mean=4.84), $F(1,53)=6.06$, $p<.05$. However, Indian participants' level of individualism (mean=5.12) is also higher than US participants (mean=4.46, $F(1,53)=9.61$, $p<.01$). Our analysis suggests that workers living in India are more collectivistic and more individualistic than those living in U.S. at the same time.

There is also a significant correlation between collectivism and individualism for Indians ($r=.62$, $p<.05$), but not for Americans ($r=.07$, *n.s.*).

As for thinking style, Indians are more holistic than US participants, $F(1,53)=3.68$, $p=.06$.

We did not detect a significant difference between Indian and US participants on social desirability, $F(1,53)=1.24$, *n.s.* although Indians tended to score higher (mean=6.26) than Americans (mean=5.38).

	Statistics	Avg. score/rating	
		US-based workers	India-based workers
User-generate ratings*	$F(1,58.2)=3.96$, $p=.05$	3.45	3.85
Collectivism**	$F(1,53)=6.06$, $p<.05$	4.84	5.44
Individualism***	$F(1,53)=9.61$, $p<.01$	4.46	5.12
Holistic thinking*	$F(1,53)=3.68$, $p=.06$	4.76	5.29
Social desirability	$F(1,53)=1.24$, $p=.27$	5.38	6.26

*$p<.1$ **$p<.05$ ***$p<.01$

Table 1. Summary of scales responded by participants from US and India in the study.

DISCUSSION

In this research, we aim to examine whether and how user-generated online ratings may be different across cultures in an online marketplace for work – Mechanical Turk. Prior research suggested some different reasons why ratings between Americans and Indians would be different. For example, prior research on cultural differences in collectivism/individualism [4], and holistic thinking [6] would suggest that Indian workers give more neutral ratings compared to American workers who would be more willing to give more extreme ratings.

Surprisingly, what we found does not seem to be predicted by prior research. We found that Indians gave higher movie and product ratings, in addition to giving higher ratings across a number of survey scales used in the study (see Table 1 for a summary). Perhaps most notable is that the Indian workers gave higher ratings for *both* the collectivism scale and the individualism scale, while prior work suggest that Indians are more collectivistic than Americans (and not more individualistic). There's also an unexpected high correlation between individualism and collectivism scores for Indian workers. This suggests a general tendency for Indians to give higher ratings.

These results cannot be justified by social desirability bias, for a number of reasons [1]. First, social desirability bias is "the tendency of people to deny socially undesirable traits or qualities and to admit to socially desirable ones"[13]. In many of the scales we used, such as collectivism/ individualism, holistic thinking, etc., a higher rating would not necessarily be considered more "socially desirable." Second, social desirability bias usually affects ratings about self, but in this case, we found that Indian workers' ratings about movies and products are also higher.

Prior work has found that Indian workers generally tend to do lower quality work [15]. One alternative explanation to our finding may be that the Indian workers have simply adapted to selecting higher scores with these online studies and surveys, without carefully reading through the prompts. However, our analyses of time spent on study actually showed that Indian workers spent about 50% more time on task (7 additional minutes). This time spent on task does not lend support to this potential explanation, although we should note that our analysis does not control for Internet speed and page load time, which may be worse in India.

Our findings are important for a number of reasons. First, it raises a critical issue of using Mechanical Turk to collect users' subjective ratings. We show that there are systematic differences. Not only is this a problem for questions about self, like suggested in prior studies of social desirability on Mechanical Turk, but it also affects general ratings on movies and products. Requesters on Mechanical Turk should be aware of these systematic biases to be able to better interpret the ratings they receive. The biases we observed could occur more broadly in online services that rely on user-generated ratings. Future research needs to study design of more effective mechanisms for eliciting and aggregating user-generated ratings that are produced by a diverse global population.

More immediately, our results highlight a critical challenge for researchers conducting cross-cultural studies. Sites like Mechanical Turk have quickly become a popular human research subject pool [2]. On one hand these sites offer researchers quick access to workers from across the world and greatly facilitate cross-cultural studies. On the other hand, our results indicate that the responses across cultures are systematically different. This can greatly undermine the validity of these cross-cultural studies. More research is needed to understand the underlying mechanisms for the observed differences to help find out ways to control for the systematic differences.

ACKNOWLEDGEMENT

We thank the College of EECS, National Tsing Hua University and the Prosocial Computing Lab at University of Washington for supporting this study. We also thank the reviewers for providing their useful comments and thoughts.

REFERENCES

1. Antin, J., & Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. *Proc. of CHI 2012.*
2. Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data?. *Perspectives on Psychological Science*, *6*(1), 3-5.
3. Chan, D. K. S. (1994). Colindex: A refinement of three collectivism measures. In U. Kim, H. C. Triandis, C. Kagitcibasi, S. C. Choi & G. Yoon (Eds.), *Individualism and collectivism: Theory, method, and applications* (pp. 200-210). Thousand Oaks, CA: Sage Publications.
4. Cheng, C., Lee, S-Y., & Stevenson H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science,* 6, 170-175.
5. Choi, I., Koo, M., & Choi, J. A. (2007). Individual differences in analytic versus holistic thinking. *Personality and Social Psychology Bulletin*, 33, 691-705.
6. Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences*, *44*, 932-942.
7. Hofstede, G. (2014). Retrieved Sep 20th, 2014, from http://geert-hofstede.com/india.html.
8. Hofstede, G. (1983). Dimensions of national cultures in fifty countries and three regions. In J. Deregowski, S. Dzuirawiec & R. Annis (Eds.), *Explications in Cross-Cultural Psychology.*
9. Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology*, *36* (2), 264-277.
10. Kim, S. H., & Kim, S. (2013). National culture and social desirability bias in measuring public service motivation. *Administration & Society.* DOI: 10.1177/0095399713498749
11. Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224-253.
12. Nisbett, R. E., Peng K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*, 291-310.
13. Phillips, D.L., & Clancy, K.J. Some effects of "social desirability" in survey studies. *The American Journal of Sociology*, *77*(5):921–940, Mar. 1972.
14. Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlow-Crowne social desirability scale. *Journal of Clinical Psychology*, *38*, 119-125.
15. Shaw, A. D., Horton, J. J., & Chen, D. L. (2011). Designing incentives for inexpert human raters. *Proc. of CSCW 2011..*
16. Triandis, H. C. (1995). *Individualism and Collectivism*. Boulder, CO: Westview.
17. Varnum, M. E. W., Grossmann, I., Kitayama, S., & Nisbett, R. (2010). The origin of cultural differences in

cognition: The social orientation hypothesis. *Current Directions in Psychological Science*, 19, 9-13.