Exploring the Value of Personality in Predicting Rating Behaviors: A Study of Category Preferences on MovieLens

Raghav Pavan Karumur, Tien T. Nguyen, Joseph A. Konstan
GroupLens Research
University of Minnesota
Minneapolis, MN 55455 USA
+1 – (612) 626 1831
{raghav, tien, konstan}@cs.umn.edu

ABSTRACT
Prior work relevant to incorporating personality into recommender systems falls into two categories: social science studies and algorithmic ones. Social science studies of preference have found only small relationships between personality and category preferences, whereas, algorithmic approaches found a little improvement when incorporating personality into recommendations. As a result, despite good reasons to believe personality assessments should be useful in recommenders, we are left with no substantial demonstrated impact. In this work, we start with user data from a live recommender system, but study category-by-category variations in preference (both rating levels and distribution) across different personality types. By doing this, we hope to isolate specific areas where personality is most likely to provide value in recommender systems, while also modeling an analytic process that can be used in other domains. After controlling for the family-wise error rate, we find that High Agreeableness users rate at least 0.5 stars higher on a 5-star scale compared to low Agreeableness users. We also find differences in consumption in four different personality types between people who manifested high and low levels of that personality.

1. INTRODUCTION
Prior research has shown that personality influences the way humans behave, and that it is related to taste preferences and interests [17]. However, much of prior social science work examining relationship between personality and preferences reports no moderate-to-large relationships [3, 10, 13, 14, 17, 18, 23], and often, is limited in statistical approaches [1, 3, 10, 23].

More recent work incorporated personality into their recommender algorithms to address challenges surrounding cold start [7, 9, 19, 20] and showed that personality-encompassing models have slightly better performance in terms of lower MAE, more precision, and more recall than prior techniques that computed user-user similarity purely based on ratings. So, despite good reasons to believe personality assessments should be useful in recommenders, we are left with no substantial demonstrated impact. If research in improving personalization services and enhancing user experience is to proceed in the direction of incorporating personality, it seems essential to know in what categories personality is likely to have value.

We use data from a live movie recommender system called MovieLens for this purpose. We take an approach different from prior research and study category-by-category variations in preference (both rating levels and distribution) across different personality types. Our goal is to complement, and find the potential utility of the existing research in this area, by throwing light on the size of the effects on specific categories in isolation. Simultaneously, we also model an analytic process that can be used in other domains.

1.1 Research Questions
In the context of MovieLens (movielens.org), a movie recommender system, taking movie genres as categories, we seek to understand answers to the following two questions:

RQ1. Does the magnitude of ratings across categories vary by personality type? If so, how?

RQ2. Do the proportions of items consumed across categories vary by personality type? If so, how?

We find differences in *consumption* in four different personality types between people who manifested high and low levels of that personality, and we find differences in *rating level* in only one personality type, Agreeableness.

2. BACKGROUND AND RELATED WORK
One of the most influential models in psychology for studies encompassing personality and human behavior is the Five Factor Model (FFM), which characterizes Personality in terms of the five dimensions Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism [5, 6, 11, 12, 15, 22]. Table 1 gives an overview of characteristics associated with each individual personality type [1, 5, 11, 12, 15, 21, 22].

Personality was found to explain individual differences in preferences for media [14], arts such as music and paintings [17, 18], types of activities [12, 14] and for specific categories in media and entertainment [3, 10, 12, 13, 14, 17, 18, 23]. Some investigated preferences of users across movie genres [3, 23]. Others examined individual differences in genre preferences in a variety of entertainment domains including music, TV shows, movies, books and magazines and attending museums and concerts [1, 14, 17, 18]. However, most of these found small correlations and some were limited to very specific populations and cultures. While in this work, we set out to explore the size of the effects (which is different from any of the above investigations), MovieLens has the added advantage that its users come from all over the world.

RecSys '16, September 15-19, 2016, Boston, MA, USA

DOI: http://dx.doi.org/10.1145/2959100.2959140

Table 1. Five Personality Traits and associated characteristics

Personality Type	Characteristics
Openness	Appreciation for novelty or variety in experiences, diversity in interests.
Conscientiousness	Organized, consistent, cautious and dutiful, less creative.
Extroversion	Appreciation for environments with higher levels of stimulation, high energy, more activity and social life.
Agreeableness	Cooperative, Adaptable, submissive, tolerant, generous, modest and trusting.
Neuroticism	High susceptibility to anger, frustration, insecurity, pessimism, anxiety and negative emotions.

One of the earliest attempts involving incorporating personality into recommendations was in the context of an image recommender system [19]. They computed user similarity based on the five factor scores and found this approach more accurate than a standard ratings-based one. Hu et al [9] computed user similarity based on personality vectors and found that the personality-based algorithm resulted in better MAE, Recall and Specificity compared to the ratings-based one. Tkalcic and Chen [21] found that the personality-based approach generates more accurate recommendations than traditional ratings-based approach on a music dataset. Elahi et al [7] incorporated personality data, and found that their approach performed better than the random baseline method and the *log*(popularity)*entropy method in terms of MAE. However, the improvements reported in these works are small and nearly equal to the ratings-based techniques. We hope to expand on this line of research, by throwing light on the size of the effects in specific categories to understand where this value is coming from. So, we research on a set of questions related to personality and specific categories of consumption on a movie recommender system.

3. DATASET AND METHODOLOGY

For this study, we use 38,675 movies of MovieLens classified under one or more of the 17 genres Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Horror, Musical, Mystery, Romance, Science Fiction, Thriller, War and Western.

MovieLens users primarily consume movies by rating them on a five-star scale in half-star increments. We administer a questionnaire based on [8] to MovieLens users to obtain their personality information and gather responses from 1840 users. Based on the responses, we compute a personality score for each user for each of the five personality traits on a scale ranging from 1-7. We also obtain the de-identified data for 985,918 movie ratings provided by these users and this formed our dataset.

In order to increase the sensitivity of analyses and ensure comparability of results some of the prior works divide the scores into thirds [3, 5] and compare users in the higher and lower thirds against each other. We similarly classify users scoring <=2 on any personality trait as belonging to the "low" type and >=6 as belonging to the "high" type. For Openness however, we increase the lower threshold to 3.5 since we have too few users in the "low Openness" category to make a statistically significant comparison. Note that this will still convey the same meaning, since 4 on the Likert scale corresponds to "Neither Agree Nor Disagree". The numbers of users in the low and high personality type categories based on the above approach are reported in Table 2.

Table 2. Number of users with low and high personality types

Personality trait	# low users	# high users
Openness	119	730
Conscientiousness	58	417
Extroversion	439	148
Agreeableness	65	203
Neuroticism	100	430

4. RESULTS AND DISCUSSION

RQ1. *Does the magnitude of ratings across categories vary by personality type? If so, how?*

We hypothesize that users' ratings might be different for movies in the two categories: popular and less popular. Also, some movies generally get high ratings (>=4 on a 5-star rating scale), some get moderate ratings (3-3.5) and some get very low ratings (0.5-2.5). We feel that combining movies across all these categories and analyzing them might average out effects. So we separately analyze the ratings across the three rating level categories and the two popularity categories mentioned above. We thus have six conditions for each of the 17 genres leading to a total of 102 pairwise comparisons for each of the 5 personality types between the 'low' and 'high' categories. In all, we have 510 tests comparing magnitude of ratings provided by the high and low personality types. An example of a comparison looks like: "difference in magnitude of rating between the high and low Openness users for *popular* Action movies that are rated *low*". Since a user can produce more than one rating, the data points in each group are not independent. We therefore add a random effect for each user to take care of variation due to individual user differences. For each of the 510 tests, we come up with a mixed model that does not include a term for personality (the null model), and another that includes the personality term (the full model) and compare the full model against the null model.

Table 3. Significant rating differences between High and Low Agreeableness users for various categories

No.	Category/genre	Number of stars higher on a 0-5 star scale	Statistical significance
1	High-rated less popular Adventure	0.5 ($\pm$ 0.1)	χ^2 :22.3 *p*-adj: 0.001
2	High-rated popular Adventure	0.4 ($\pm$ 0.09)	χ^2 :15.3 *p*-adj: 0.046
3	High-rated popular Animation	0.5 ($\pm$ 0.1)	χ^2 :17.6 *p*-adj: 0.014
4	High-rated less popular Children	0.6 ($\pm$ 0.14)	χ^2 : 15.5 *p*-adj: 0.042
5	High-rated popular Children	0.5 ($\pm$ 0.1)	χ^2 : 23.5 *p*-adj: 0.0006
6	High-rated less popular Fantasy	0.6 ($\pm$ 0.1)	χ^2 : 20.8 *p*-adj: 0.003
7	High-rated less popular Romance	0.5 ($\pm$ 0.1)	χ^2 : 16.7 *p*-adj: 0.02
8	Low-rated popular Romance	0.6 ($\pm$ 0.1)	χ^2 : 20.8 *p*-adj: 0.003

If the difference between the likelihoods of these two models is significant, then the fixed effect personality term is significant. However, the large number of simultaneous analyses conducted increases the likelihood of Type I error. So, we employ the Holm's method to control for family-wise error rate at 0.05 by doing a sequential adjustment of *p*-values. In the end, we obtain significant effect sizes for one personality trait Agreeableness (shown in Table 3).

All results stated in Table 3 are for high Agreeableness users in comparison to low Agreeableness users. The χ^2 reported, is for the likelihood ratio test comparing the full model with the null model. We find, for instance, that high Agreeableness users rate high-rated popular Children movies by about 0.5 stars higher compared to low Agreeableness users. Others in the table follow similar interpretation. The adjusted *p*-value indicates the statistical significance after controlling for family-wise error rate.

Because users high in Agreeableness are tolerant, they might be expected to show higher magnitude of ratings compared to low Agreeableness users. We test if High Agreeableness users in general have high ratings over all categories on the aggregate compared to low Agreeableness users. While we find a statistically significant result, the effect size (magnitude by which they rate higher) is only 0.08. This not only is too small, but also masks the fact that there is no significant difference in some categories. Chamorro-Premuzic et al [2] propose that high agreeableness users are likely to watch movies for fun. Their finding justifies the effect sizes we find in the categories we report above. Furthermore, Cosley et al [4] find that recommendation interfaces can affect how users rate by 0.23 stars and Nguyen et al [16] suggest that improving recommendation interfaces can mitigate such natural differences in ratings. In our case, by anchoring at one personality trait (Agreeableness), we already observe a difference of 0.5 stars (10% on a 5-star scale) (about twice as much) in how users rate. This observation signifies that personality trait per category does contribute to substantial differences in ratings and that it cannot be ignored.

RQ2. *Do the proportions of items consumed across categories vary by personality type? If so, how?*

To answer this question, for each user, we first obtain the proportion of movies consumed across each of the 17 genres. Since a single movie can belong to multiple genres, and we do not want to triple-count a movie with three genres (compared with a movie labeled with only one genre), we count each movie as $1/G$ consumptions for each of the G genres the movie is associated with. For example, Star Wars (1977) is labeled "Adventure", "Action" and "Science Fiction" on MovieLens. So, we assess a user who rates (consumed) this movie as having consumed 1/3 of the movie under each of the three genres, with the total consumption being one movie. This formulation also has the property that the genre proportions sum to one. Based on this formulation, for each of the five personality types, we test the difference in median proportions of movies consumed for each of the 17 categories between the high types and low types by employing a Wilcoxon-Mann-Whitney test. For this, we pick an equal number of samples (100 each for Openness, 50 each for Conscientiousness, 120 each for Extroversion, 60 each for Agreeableness, and 100 each for Neuroticism) from both comparison groups. We have 85 such tests. So we correct for the family-wise error rate and obtain the adjusted *p*-values for each of the tests. Significant effects based on these adjusted *p*-values are reported in Table 4. In Table 5 we provide distributions of consumptions across each category by MovieLens users as a guide to gain insight into the value of effect sizes in Table 4. For instance, the highlighted cell in Table 4 indicates that high Conscientiousness users consume 2% higher proportion of movies in the Romance category compared with low Conscientiousness users. The highlighted row in Table 5 indicates that users can consume anywhere between no Romance movies to a maximum of 29% of Romance movies out of all the movies that they rate. A 2% higher magnitude of consumption therefore indicates a shift in the magnitude of user's consumption from 1st quartile to mean or from median to 3rd quartile consumption, showing substantial value. Others follow similar interpretations.

Table 4. A summary of proportions of consumptions across various categories. ($p < 0.001$: *** $p < 0.01$ ** $p < 0.05$ *)

	Openness	**Conscientiousness**	**Extroversion**	**Agreeableness**	**Neuroticism**
Action		low > high (+2%) *			low > high (+2%)**
Adventure	low > high (+1%) *				low > high (+1%) *
Comedy					high > low (+2%) *
Drama	high > low (+4%) **				
Fantasy	low > high (+1%) ***				low > high (+1%) *
Romance	high > low (+1%) **	high > low (+2%) **	low > high (+1%)*		high > low (+1%) *
Thriller	low > high (+1%) *	low > high (+2%) *			low > high (+1%) *

Table 5. Dist. of consumption across selected categories

	Min	**1Q**	**Med**	**Mean**	**3Q**	**Max**
Action	0%	7%	10%	10.2%	13%	36%
Adventure	0%	6%	8%	8%	10%	29%
Comedy	0%	1%	14%	14.3%	18%	44%
Drama	2%	15%	20%	20.6%	25%	72%
Fantasy	0%	3%	4%	4%	5%	25%
Romance	0%	4%	5%	6%	7%	29%
Thriller	1%	7 %	9 %	9 %	11 %	27 %

5. CONCLUSION

Despite good reasons to believe personality assessments should be useful in recommenders, prior work shows no substantial demonstrated impact. In the hope of complementing prior research, in this work, we set out to explore if and by how much personality affects (i) the magnitude of ratings and (ii) the proportion of items consumed across various categories in a live recommender system. Our work shows substantial effects in multiple categories for various personality types and demonstrates that further research along the lines of incorporating personality is promising.

Our findings suggest that recommendations based purely on aggregate ratings of all users may not be a good idea. Instead, for a user with a certain personality type, recommendation models may consider showing two kinds of ratings - one produced by users of the same personality, such as "*Users similar to you rated* *xx stars*", and the other, an aggregate rating by all users, to improve consumption. We leave this for future work. Also, a documentation of stable relationships between individual differences and category preferences can throw light on who is likely to consume or rate across various categories. This is important not only for recommending the right categories to elicit ratings from users, but also in contexts such as providing better cold-start recommendations, and personalizing lists of novel, diverse and serendipitous movies to users at different times.

The nature of the effects we find suggests that certain algorithm types may be more able to productively incorporate personality data. The fact that the effects tend to be localized in certain item categories suggests that correlational algorithms may be less able to exploit personality data than dimensionality-reduction (or condition-probability) algorithms. This remains as future work.

A chief limitation of this work is that although we have data from a live recommender, we have personality data for only 1840 users and the personality data is not evenly distributed. We also do not have data on users' demographics to perform interaction analyses. Further work should look at classifying users in other domains and with additional information. Also, in this work, we looked at both the categories of consumption and the personality traits in isolation. Future work should look at combinations of personalities and the combinations of categories which these personalities might be interested in consuming.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant IIS-1319382. Additionally, we thank the MovieLens users who took the Personality survey.

7. REFERENCES

[1] Cantador, I., Fernández-Tobías, I., & Bellogín, A. 2013. Relating personality types with user preferences in multiple entertainment domains. In *CEUR Workshop Proceedings*, Shlomo Berkovsky.

[2] Chamorro-Premuzic, T., Kallias, A., & Hsu, A. 2013. Understanding individual differences in film preferences and uses: a psychographic approach. *The Social Science of Cinema*, 87.

[3] Chausson, O. 2010 Who watches what?: assessing the impact of gender and personality on film preferences. *Paper published online on the MyPersonality project website* http://mypersonality.org/wiki/doku.php

[4] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '03)*, ACM, New York, NY, USA, 585-592.

[5] Costa Jr, P.T. and McCrae, R.R. 1992. Neo personality inventory–revised (neo-pi-r) and neo five-factor inventory (neo-ffi) professional manual. *Odessa, FL: Psychological Assessment Resources*.

[6] Costa, P., and McCrae, R. 1996. Toward a new generation of personality theories: Theoretical contexts for the five-factor model. *The Five-Factor Model of Personality: Theoretical Perspectives* (1996) 51-87.

[7] Ealhi, M., Braunhofer, M., Ricci, F., & Tkalcic, M. 2013. Personality-based active learning for collaborative filtering recommender systems. *AAAI 2013*.

[8] Gosling, S. D., Rentfrow, P.J., and Swann, W.B. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (Dec 2003), 504–528.

[9] Hu, R., and Pu, P. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems* (RecSys '11). ACM, New York, NY, USA, 197-204.

[10] Hu, R., and Pu, P. 2013. Exploring Relations between Personality and User Rating Behaviors. In *UMAP Workshops*.

[11] John, O.P. The Big Five factor taxonomy. Dimensions of personality in the natural language and in questionnaires. 1990. *Handbook of personality: Theory and research*, 14 (1990) 66-100.

[12] Karumur, R., and Konstan, J.A. 2016. Relating Newcomer Personality to Survival and Activity in Recommender Systems. In *Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '16)*. In Press.

[13] Kraaykamp, G. 2001. Parents, personality and media preferences. *Communications* 26, 1 (2001) 15-38.

[14] Kraaykamp, G. and Van Eijck, K., 2005. Personality, media preferences, and cultural participation. *Personality and individual differences*, 38, 7 (May 2005), 1675-1688.

[15] McCrea, R. and John, O. An introduction to the five-factor model and its applications. 1992. *Journal of personality*, 60, 2 (Jun 1992), 175-215.

[16] Nguyen, T. T., Kluver, D., Wang, T. Y., Hui, P. M., Ekstrand, M. D., Willemsen, M. C., & Riedl, J. 2013. Rating support interfaces to improve user experience and recommender accuracy. In *Proceedings of the 7th ACM conference on Recommender systems* (RecSys '13). ACM, New York, NY, USA, 149-156.

[17] Rentfrow, P.J., and Gosling, S.D. 2003. The do re mi's of everyday life: the structure and personality correlates of music preferences. *J Pers Soc Psychol*. 84, 6, (Jun 2003), 1236-1256.

[18] Rentfrow, P. J., Lewis R. G, and Ran Z. Listening, watching, and reading: The structure and correlates of entertainment preferences. *Journal of personality* 79, 2 (2011) 223-258.

[19] Tkalcic, M., Kunaver, M., Tasic, J., & Košir, A. 2009. Personality based user similarity measure for a collaborative recommender system. In *Proceedings of the 5th Workshop on Emotion in Human-Computer Interaction-Real world challenges*, 30-37.

[20] Tkalcic, M., and Kunaver, M. 2011. Addressing the new user problem with a personality based user similarity measure. *DEMRA 2011*.

[21] Tkalcic, M., and Chen, L. 2015. Personality and Recommender Systems. In *Recommender Systems Handbook,* Ricci, F., Rokach, L., and Shapira, B, Ed. Springer US, 715-739.

[22] Tupes, E.C., and Christal, R.E. 1992. Recurrent personality factors based on trait ratings. *Journal of Personality*, 60, 2, (April 2006) 225-251.

[23] Wen, W., and Chen, L. Implicit acquisition of user personality for augmenting movie recommendations. 2015. User Modeling, Adaptation and Personalization. Springer International Publishing, 302-314.