Love, Language, and Linear Algebra: Linguistic Modeling of Personality and Mate Preference

Abstract

This study utilized Latent Semantic Analysis to determine whether similarities in personality predicted similarities in responses to a romantic writing prompt. (Landauer & Dumais, 1997). From participants’ writing samples, we calculated thematic cosines (a measure of relatedness) between each male and female participant. Participants also completed the Big Five Personality Questionnaire Short Form (Morizet, 2014). Extraversion, agreeableness and conscientiousness were related to cosines, which suggested medium-small relationships from personality traits to written responses. This relationship is consistent with previous studies on mate preference, which suggests that Latent Semantic Analysis may be useful in quantifying mate preference, especially when alongside traditional survey methods. We conclude with a discussion of the compatibility of ordinal measures (survey data) and continuous measures in examining complex phenomena in the Behavioral Sciences, such as mate preference.

*Keywords*: mate preference, mate choice, Five-Factor Personality Model, Latent Semantic Analysis, thematic cosines,

Love, Language, and Linear Algebra: Linguistic Modeling of Personality and Mate Preference

Sexual and romantic desirability are vital in forming a basic unit of human culture, the mated pair. Through natural selection, general preference for certain traits, such as intelligence and physique, lead to our evolution as a species. Romantic preference, an individual’s abstract set of desirable traits in a mate, defines many cultural phenomena. In many ways, our similarities in individual mate preference shape our social environment. For example, the literature suggested that men value attractiveness more than women in survey based research paradigms. This specific sex difference was observed by Feinstein (1990) in a meta-analysis of 28 separate samples of American females and males. However, Feinstein also found similar differences in personal ads and billboards targeting males and females, which suggests that survey based results often correlate realistically with real-world sex differences in mate preference.

In a far-reaching cross-sectional study, Buss (1989) examined sex differences in mate preference across 37 samples from 33 distinct cultural paradigms. To compare mate preference and sex differences across cultures, Buss administered a three-part survey. This survey asked for participants’ demographic information (age, sex, religious and familial background). The second portion of the survey asked participants for their ideal age to marry, their preferred age difference to a potential spouse, and how many children they desired. The final section asked participants to rate 18 characteristics (i.e. sociability, intelligence, chastity) on how important they were in determining a potential romantic partner. Incredibly, Buss found that sex differences in mate preference were almost entirely homogeneous across all cultures. Examples included higher preference among women for fiscally stable partners, and higher preference among men for younger female partners.

Within the same study, Buss (1989) also carefully checked census data from each country to determine how mate preference influenced mate choices. As an example, in every culture studied, an age-gap of approximately three years was found between older men and younger women in census data. This dovetailed neatly with the second survey section which assessed participants’ ideal age difference between a potential mate. Yet, age differences are easily-measured, external variables. Moreover, as stated by Buss, age differences were the most statistically reliable findings in his study, while other variables, such as previous sexual experience, showed weaker effects across different cultures.

Buss’s (1989) and Feingold’s (1990) research suggests that mate preference is a valid cognitive construct in multiple cultures and paradigms. Moreover, certain sex differences in preference, such as physical attractiveness and age, are apparent in census and environmental data. Yet, the relationship of other traits, such as personality or intelligence, to concrete mate choice is more complex. In survey-based research of Brazilian college students, Castro, Hattori, and Lopez (2012) found that preferences in non-physical traits (i.e., humor, intelligence) did not always correlate with concrete perceptions of current or recent mates. Their results show how mate preference may significantly differ across sex within a sample without necessarily predicting individuals’ perceptions of real-world romantic partners (Castro et al., 2012).

Castro et al.’s findings highlight the difference between our abstract romantic preferences and our concrete sexual selection process. These results imply that people often choose mates which do not fit their stated preferences. This discrepancy necessitates research into the intricacies of romantic preference and its role in evolutionary psychology and human cognition. Of course, while an individual’s romantic preferences may fail to predict their mate choices, certain social phenomena can be explained as a function of observed gender differences in romantic preference.

As an example, Feingold’s (1990) meta-analysis explored the types of empirical methodologies used to study romantic preference and mate choice. He also compared this meta-data with linguistic analyses of advertisements and billboards targeted towards men or women specifically. Interestingly, he noted that advertisements targeting men focus on attractive female partners more than advertisements for women, a conclusion that mirrored Buss’ (1989) findings and meta-data collected from survey-based research in romantic preference. That advertisements dovetail with observed research shows the direct applicability of empirical research in romantic preference. It also reveals the influence of romantic preference in shaping our understanding of desirability across two distinct genders.

Of course, that romantic preference influences society suggests it also motivates individuals and influences their actions. Botwin, Buss, and Shackelford (1997) found that individuals from both sexes prefer romantic partners whose personality traits mirror their own. Long-term partners were likely to exhibit similar personality traits, showing a distinct connection between personality preferences in romantic partners and successful long-term romantic relationships. Even more, among all participants, Botwin et al. (1997) found that certain personality traits were unappealing. These included low agreeableness, low emotional stability, and non-equal openness to experience between partners. Here, low agreeableness is defined as hostility or wariness towards others; low emotional stability is defined as a tendency to experience negative emotions quickly; and openness to experience is defined as creativity and willingness to enter unfamiliar situations. These are as defined in the Five Factor Model (see McCrae and John [1992]). In relationships which had lasted longer than a year, personality differences were even stronger predictors of dissatisfaction (Botwin et al., 1997).

Botwin et al.’s (1997) results suggest a relationship between mate preference and mate choice which is consistent across several physical and personality traits. In observed concrete mate choices, similar personality scores are strong indicators of relational satisfaction. Yet, personality is a factor which Castro et al. (2012) suggests plays a lesser role in abstract romantic preference, especially among males. Back, Penke, Schmukle, and Asendorpf (2011) also observed that, in short-term socio-sexual interactions (i.e., speed dating scenarios), an individual’s agreeableness not only predicted desirability but also significantly correlated with participants’ ability to predict their desirability among fellow participants.

This study examined the effect of personality differences on mate preference among males and females. However, unlike previously mentioned research, we measured participants’ mate preference through written responses to a prompt. We hypothesized that, like previous non-linguistic research, similarity in participants’ personality scores would predict mate preference as recorded through responses to a written prompt. To incorporate linguistic data, we utilized Latent Semantic Analysis (LSA), an algebraic technique which converts word frequency and co-occurrence into thematic cosines, which behave like correlations (Landauer, Folt, & Laham, 1998). These thematic cosines allowed us to compare similarity from one participant’s written response to another.

**Method**

**Participants**

A sample of undergraduate students (*N* = 105) was recruited from a large Midwestern university. All participants were enrolled in an introductory psychology course and received two research-participation credits for completing the study. Relatively even samples of male (*N* = 53) and female (*N* = 52) participants were recruited. The average age of the participant was around 19 years of age (*M* = 18.75, *SD* = 1.60), and the majority were white (96.15%) with the remainder not answering (3.85%). Sample collection occurred over a two-month period from October through early-December.

**Materials and Procedure**

All participants received online survey materials through Qualtrics, an internet survey platform. After reporting demographic information (e.g., gender, age, academic major, ethnicity), participants completed the Big Five Personality Trait Short Questionnaire (Morizot, 2014), which assessed openness, extraversion, agreeableness, conscientiousness and emotional stability. Finally, in random order, participants responded to a pair of writing prompts. One concerned their interests and hobbies (“Describe your interests and/or hobbies”), while the other asked them to describe their ideal romantic partner (“Describe an ideal date with your perfect romantic partner”). The order of prompts was counterbalanced, and responses had to exceed a minimum of 2200 characters to move on with the study. This requirement was to ensure enough information density in the writing samples to guarantee usable latent semantic data. For this specific study, we did not utilize the interests-and-hobbies written data. In the future, we may analyze whether similarities in writing on other dimensions (i.e., interests, personal statements, etc.) moderates the relationships of personality and romantic writing. However, in this study, we only tested the relationship between similarity across each personality measure with romantic writing.

**Results**

Data analysis was conducted in two major steps: Latent Semantic Analysis to create the dependent thematic cosine variable, and several multilevel models (MLM) examining the influence of individual participants’ personality differences on romantic writing similarity as measured by thematic cosines.

**Latent Semantic Analysis**

Raw written data were marked with a participant number, gender, and prompt number. LSA was conducted in *R* using the *lsa* (Wild, 2015) package. Initially, LSA encodes the word frequency and co-occurrence of each participant’s written response in a text-frequency matrix. This matrix was normalized using log weighting to control for the sparsity/skew of text frequencies, that is, the differences in number of very frequently versus infrequently used words. We also removed common English stop words (e.g., “the”) to reduce the number of meaningless co-occurrences across writing samples (see Rajaraman and Ullman [2011] for justification). LSA was then performed, which created a matrix of concepts by documents with values in this matrix representing the relationship of each concept to a document. Cosine values between each male-female participant combination were calculated, and therefore, the final dependent variable dataset included 5485 cosine values (i.e., male participant one to female participant one, two, etc.). The complete scripts and data set can be found at: [removed for blinding].

**Data Screening**

Next, the independent variables were added to the cosine values. Difference scores were calculated by subtracting our male participant’s score from our female participant’s score across each personality variable. Following this, we took an absolute value to normalize the order effects of subtraction on our personality measure. Next, the data were analyzed for assumptions of parametric regression. Mahalanobis distance was calculated on the cosine scores and personality responses (Tabachnick & Fidell, 2012). Only one participant-pair fell outside the Mahalanobis cutoff score (χ2(6)*p*<.001= 22.46) and was excluded. Data were then screened for accuracy, additivity, normality, linearity and heteroscedasticity with all necessary assumptions being met before analysis.

**Multilevel Model Analysis**

Following data screening, descriptive statistics were calculated for romantic cosines and personality measures across both males and females. The average romantic cosine (*M* = .19, *SD* = .17) was relatively small and showed a comparatively large standard deviation. Personality scores ranged from 10-50 on an interval scale, although we utilized a difference score in our MLM. However, for convenience, Table 1 shows personality means, standard deviations, and Cohen’s *ds* (Lakens, 2013) across both males and females.

In our analysis, each personality variable was analyzed in a separate MLM. We chose this design to control for the correlated error introduced by examining each participant paired with every other opposite gender participant (i.e., therefore, controlling for male participant one being represented in the data multiple times across female participants). We compared three distinct models: an intercept-only model, which estimates the *y*-intercept as the same across all participants; a random-intercept model, which allows estimation of the *y*-intercept controlling for multiple instances of the same participant, thus handling correlated error; and a random-intercept model with personality differences as a predictor, which controls for repeated measures for each participant and estimates the relationship between the IV and the DV(Field, Miles, & Field, 2012).

Except for the MLM examining openness, the random-intercept model with predictors was the best fit for our data in each MLM. However, due to the repeated measures of the data, we included all models from the random-intercepts main effects, as we wished to control for correlated error. Model significance was evaluated using a chi-square difference test where each model is compared to the previous model to determine how adding random slopes or predictors improves the model; however, in order to determine the best-fit for our data, we utilized the Aikake Information Criterion (AIC). A lower AIC corresponds to less information lost, and hence, models with lower AIC scores correspond to better fits for our data. Individual model’s degrees of freedom, intercepts, as well as significance among all models can be found in Table 2.

We found that differences in extraversion, agreeableness, and conscientiousness were predictors of similarities in thematic cosines across romantic writing. With negative slopes, this finding suggests that smaller differences in personality predicted larger thematic cosines. Therefore, as personality scores were more similar (small differences, closer to zero), the larger the overlap between the romantic writing provided for participants. Difference in emotional stability and openness were not predictors of similarity in thematic cosines. For convenience, see Table 3 for predictors, intercepts, standard errors, and *p*-values for each predictor.

**Discussion**

Our results show that similarity in extraversion, agreeableness, and conscientiousness predicted similarity in writing about a romantic partner. With the largest predictor *b*-value, agreeableness as a predictor aligns with existing findings by Back et al. (2011) and Botwin et al. (1997), who suggested that agreeableness was the strongest personality predictor for high mate value and relational satisfaction in concrete mate choices. Since our study examined mate preference specifically, we cannot draw conclusions related to mate choice. However, our results show that similar levels of agreeableness predict similarities in written responses. This finding suggests that further research in mate preference and personality may uncover a similar relationship of agreeableness to mate preference as in Back et al.’s and Botwin et al.’s studies on mate choice.

In this study, linguistic modelling of mate preference returned similar results as traditional survey methods. Many of the surveys examining mate preference utilize questions constructed around observed constructs related to mate preference (such as socio-economic status or personality), and usually measure these variable on a Likert-style scale (see Buss, 1989; Castro et al., 2012). This method of analysis has several benefits, including: generalizability of results from study-to-study, ease of drawing meaningful conclusions from data, and simplification of replicability. What, then, justifies the future use of linguistic modelling if our conclusions seem to agree?

In the context of measuring mate preference, linguistic modelling has several valuable assets. The principal strength of modelling writing is that it allows participants to respond freely to writing prompts before any data transformation takes place. Latent Semantic Analysis transforms a truly continuous measurement (writing) into a continuous variable (thematic cosines). The individuality of each participant’s written response reflects the uniqueness of their own innate set of mate preferences in this study. Theoretically, these thematic cosines capture more of the individual variance among our participants’ mate preference when compared to a discrete measure, such as selecting a 4 on a Likert-style scale. However, Latent Semantic Analysis presents several challenges, both theoretical and pragmatic.

Foremost among these is the interpretability of results. Often when working with ordinal measurements, such as age (measured in years) or Likert-scales, descriptive statistics of a sample are easily interpreted and explained. That does not mean a specific sample’s mean is the correct or ideal measurement of central tendency. However, it is easier to understand a statement such as, “Our sample had a mean age of 23 with a standard deviation of 2.5 years,” than one like, “Our sample had a mean thematic cosine of .35 with a standard deviation of .25.” Mathematically, thematic cosines may be more difficult to interpret than a standard correlation, such as Pearson’s *r* (1896). This difficulty is because, while thematic cosines and correlations both measure similarity, there are no traditional small, medium, or large score-markers for thematic cosines. However, the direction and magnitude interpretations for correlations and cosines are the same.

Thematic cosines can provide a continuous measurement of mate preference using this this of writing study. This variable is incredibly valuable, as continuity leads to a broader understanding of variance in a sample while avoiding common statistical problems associated with ordinal measurements, such as issues with Type I and Type II errors with small (e.g., 4-5 item) scale data in parametric statistical tests (Gregoire & Driver, 1987). Variable selection is a complicated issue, with many professional psychologists disagreeing on the use of Likert-style data in parametric statistical tests (see Rasmussen [1987] for a contrasting opinion to Gregoire & Driver [1987]). Instead, we see Latent Semantic Analysis as complementary to traditional survey methods in modelling mate preference. Moreover, in situations where ordinal data is either statistically inappropriate or cumbersome, Latent Semantic Analysis provides a more all-encompassing and continuous measure for parametric statistical tests.

Of course, in an ideal situation, every hypothesis would be measured with several unique and contrasting measures. Since we ourselves only utilized Latent Semantic Analysis in this study, and did not present any complementary surveys, we naturally understand that resources and time are usually limited. Fortunately, Latent Semantic Analysis is relatively time-and-cost effective and can be executed using the *lsa* package (Wild, 2015) in *R*. For those interested in trying Latent Semantic Analysis for their next project (or just for fun), feel free to download our scripts and data utilized in this study from our OSF page: [removed for blinding]. In conclusion, we look forward to seeing the unique insight Latent Semantic Analysis can provide in many diverse research areas, both in Evolutionary Psychology specifically, and throughout all of Behavioral Science.

References

Back, M. D., Penke, L., Schmukle, S. C., & Asendorpf, J. B. (2011). Knowing Your Own Mate Value. Psychological Science, 22(8), 984–989. doi:10.1177/0956797611414725

Botwin, M.D., Buss, D.M., & Shackelford, T.K. (1993). Personality and mate preference: Five factors in mate selection and marital satisfaction. Journal of Personality, 65(1), 107-136. doi: 10.1111/j.1467-6494.1997.tb00531.x

Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. Behavioral and Brain Sciences, 12, 1-49. doi:10.1017/S0140525X00023992

Castro, F.P., Hattori, W.T., & Lopes, F.A. (2012) Relationship maintenance or preference satisfaction? Male and female strategies in romantic partner choice. Journal of Social, Evolutionary, and Cultural Psychology, 6(2), 217-226. doi: 10.1037/0099213

Feingold, A. (1990). Gender differences in effects of physical attractiveness on romantic attraction: A comparison across five research paradigms. Journal of Personality and Social Psychology, 59(5), 981-993. doi: [10.1037/0022-3514.59.5.981](http://psycnet.apa.org/doi/10.1037/0022-3514.59.5.981" \t "_blank)

Gregoire, T.G., & Driver, B.L. (1987) Analysis of ordinal data to detect population differences. Psychological Bulletin, 101(1), 159-165. doi: [10.1037/0033-2909.101.1.159](http://dx.doi.org/10.1037/0033-2909.101.1.159" \t "_blank)

Haufe, C. (2008). Sexual selection and mate choice in evolutionary psychology. Biological Philosophy, 23, 115-128. doi: 10.1007/s10539-007-9071-0

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology, 4, 863. doi: 10.3389/fpsyg.2013.00863

Landuaer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Landauer, T. K., Folt, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25(2), 259–284. doi: 10.1080/01638539809545028

McCrae, R. and John, O. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2), pp.175-215.

Morizot, J. (2014). Construct validity of adolescents’ self-reported big five personality traits: Importance of conceptual breadth and initial validation of a short measure. Assessment, 21(5), 580-606. doi: 10.1177/1073191114524015

Pearson, K. (1896). Mathematical contributions to the theory of evolution: Regression, heredity, and panmixia. Philosophical Transactions of the Royal Society, 187, 253-318. doi: [10.1098/rsta.1896.0007](https://doi.org/10.1098/rsta.1896.0007)

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. Psychological Bulletin, 105(1), 167-170. doi: [10.1037/0033-2909.105.1.167](http://psycnet.apa.org/doi/10.1037/0033-2909.105.1.167" \t "_blank)

Schwarz, S., & Hassebrauck, M. (2012). Sex and age difference in mate-selection preferences. Human Nature, 23, 447-466. doi: 10.1007/s12110-012-9152-x

Tabachnick, B.G., & Fidell, L.S. (2012). Using Multivariate Statistics. Boston, MA: Allyn and Bacon.

Wild, F. (2015). lsa: Latent Semantic Analysis. R package, version 0.73.1.

Table 1

*Means, Standard Deviations, and Effect Size for Personality and Thematic Cosines across Sex*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Personality Measures | *M*Female | *SD*Female | *M*Male | *SD*Male | *ds* |
| Openness | 36.83 | 6.27 | 36.89 | 6.05 | 0.01 |
| Extraversion | 39.60 | 6.96 | 37.15 | 7.63 | 0.34 |
| Agreeableness | 37.65 | 7.09 | 34.91 | 6.03 | 0.42 |
| Conscientiousness | 37.65 | 7.09 | 34.64 | 6.03 | 0.46 |
| Emotional Stability | 26.61 | 7.51 | 32.00 | 7.94 | 0.70 |

Table 2

*Model statistics for MLM analyses*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Models | *df* | AIC | BIC | Log. Lik. | χ2 | *p* |
| Intercept-Only | 2 | -1755.52 | -1744.29 | 879.76 |  |  |
| Random-Intercept | 3 | -2168.23 | -2151.40 | 1087.12 | 414.72 | < .001 |
| Openness | 4 | -2168.37 | -2145.91 | 1088.18 | 2.13 | .14 |
| Extraversion | 4 | -2176.17 | -2153.72 | 1092.10 | 9.93 | .001 |
| Agreeableness | 4 | -2181.52 | -2159.10 | 1094.76 | 15.28 | < .001 |
| Conscientiousness | 4 | -2185.26 | -2162.81 | 1096.63 | 19.03 | < .001 |
| Emotional Stability | 4 | -2166.46 | -2144.01 | 1087.23 | 0.22 | .64 |

*Note*. The intercept-only model and random-intercept model is identical for each IV, and hence is only listed once.

Table 3

*Individual predictors included in the third and final random-intercepts model.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | *b* | *SE* | *t* | *p* |
| Openness | 0.001 | < 0.001 | 1.460 | .145 |
| Extraversion | -0.002 | < 0.001 | -3.156 | .002 |
| Agreeableness | -0.003 | 0.001 | -3.915 | < .001 |
| Conscientiousness | -0.002 | 0.001 | -4.371 | < .001 |
| Emotional Stability | < 0.001 | 0.001 | 0.474 | .636 |

Note: *df* = 1979.