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

Caleb Z. Marshall

Dr. Erin M. Buchanan

Dr. Melissa Fallone

Missouri State UniversityAbstract

This study utilized Latent Semantic Analysis to evaluate the relationship between personality and participants’ written personal account of an ideal romantic partner. From participants’ writing samples, we calculated thematic cosines (a measure of relatedness) between each male and female participant. Participants completed the Big Five Personality Questionnaire Short Form (Morizet 2014), a personality inventory which evaluated participants’ Openness, Extraversion, Agreeableness, Conscientiousness and Emotional Stability. We examined if similarity among two participants’ personality traits predicted similar romantic writing cosines. Extraversion, Agreeableness and Conscientiousness were related to cosine, which suggested small to medium relationships from similiar personality traits to matching writing styles about mate preference. These pairwise differences are observed beyond sex differences in personality and thematic cosines. We concluded by comparing our cosine measure with existing survey methods in the literature and suggest that Latent Semantic Analysis constructs a continuous measure for modeling participants’ mate preferences.

The Interaction of Personality, Interest and Romantic 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. However, romantic preference, which as an individual’s abstract set of sexual and socially favorable traits in a mate, is rarely aligned with concrete mating choices. Still, romantic preference defines many cultural phenomena. From beauty standards to courtship, sexual interaction and long-term relationships, our aggregate sexual preferences shape our environment in innumerable ways. Romantic preference varies from person-to-person, however, using meta-analyses, research can elucidate common trends across people. Feingold observed unique differences between American men and women on responses to mate selection questionnaires by aggregating across 28 studies available at the time. For example, the literature suggested that men value attractiveness more than women on questionnaires, but that attractiveness and mate selection were equally high for both sexes in concrete dating scenarios.

A seminal experiment on mate preference was conducted by David Buss in the late 1980’s. In a cross-sectional study, Buss examined sex differences in mate preference across 37 samples from 33 distinct cultural paradigms. (1989) His work found consistent sex differences in mate preference across all cultures. Examples include: higher preference among women for fiscally stable partners; and higher preference among men for younger female partners.

Buss also carefully checked census data from each country to determine how mate preference influenced eventual mate choices. For example, across every culture, an age-gap of approximately three years was found between older men and younger women. Yet, age differences are easily-measured, external variables. Moreover, as stated by Buss, age differences were the most statistically reliable findings in his study, which suggests that perhaps this result is more either a stronger effect between men and women or is more easily measured in his samples.

Often the connection between mate preference and mate choice is convoluted. Castro, Hattori, and Lopez (2012) studied differences between mate preference and mate choice using a survey based in-part on Buss’ (1989) original work. In general, Castro et al.’s results agreed with Buss’s on specific sex differences in mate preference, such as males valuing physical attractiveness more than women. However, they found participant’s preferences in non-physical traits (i.e., humor, intelligence) did not necessarily correlate with concrete perceptions of current or recent mates. This result suggests that intersex romantic preferences differ, but do not necessarily predict an individual’s perception of a real-world romantic partners (Castro et al., 2012).

Castro’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 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 found that certain personality traits were unappealing. These included disagreeableness, emotional instability, and non-equal Intellect-Openness between partners. This result was even more distinct when the personality differences were observed in existing long-term relationships .

Together, Botwin’s et al. (1997) and Feingold’s (1990) results portray a conversation between romantic preference and mate choice, both on an individual or societal scale and across 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 (2012) suggests plays a lesser role in abstract romantic preference, especially among males. Back et al. (not on reference list) also observed that, in short-term sociosexual 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 interaction among gender, personality, and common interests by examining participants’ written romantic preferences and personal hobbies. We hypothesized that, like previous non-linguistic research, similarity in participants’ personality scores would predict similar romantic preference as demonstrated by a written prompt. We also examined the effect of within-gender and between-gender comparison and participants’ personal interests, again collected through a writing prompt.

Quantitative data were derived from participants’ writing samples using Latent Semantic Analysis (LSA), a technique from linear algebra which generates a vector space, where each vector represents a direct connection between individual terms or documents. Cosines between these connections quantify thematic similarities between two participants’ written responses, and greater cosines indicate similarity in themes among two writing samples. (Landauer et al., 1998) In this sense, one can think of LSA as a quasi-correlational method of quantifying linguistic data for inferential statistical analysis.

LSA transforms word co-occurrence and frequency into quasi-correlational data, which can be used to measure similarities in thematics and contextual semantics across distinct writing samples. LSA is an application of singular values decomposition, a technique from linear algebra which factorizes a real or complex matrix into three component matrices. Geometrically, these component matrices correspond to a rotation, scaling and final rotation which allows an inner product (in this case, cosine) to be calculated between any members of the vector space generated by the original linguistic corpora. We utilized LSA as a method of data-transformation for our participants’ written responses, with thematic cosines being the unique score connecting each participant to one 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 under 21 years of age, and the majority were white. 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. Initially, participants were informed of the potential risks, incentives, and requirements of the study. After reporting demographic information, 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: LSA 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* and *SnowballC* packages. 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 frequent versus infrequent used words. 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: https://osf.io/5qw67/.

**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 in order 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. Only one participant-pair fell outside the Mahalanobis cutoff score and was excluded. Data were then screened for accuracy, additivity, normality, linearity and heteroscedasticity. The data were slightly right-skewed and had issues with linearity in the extremities of the data.

**Analysis**

Following outlier analysis, descriptive statistics were calculated for romantic cosines and personality measures across both males and females. The average romantic cosine (*M* = 0.19, *SD* = 0.17) 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 Multilevel Model; however, for convenience, table 1 shows personality means, standard deviations and Cohen’s *d* across both males and females.

In our analysis, each personality variable was analyzed in a separate Multilevel Model. We chose this design to streamline our analysis while still controlling for the correlated error introduced by our thematic cosine measure. As is standard, within each MLM, we compared an intercept-only model, a random-intercepts model without predictors, and a random-intercept model with personality differences as a predictor. Except for the MLM examining Openness, the random-intercepts model with predictors was the best fit for our data in each MLM. Individual model’s degrees of freedom, intercepts, as well as significance among all models can be found in Table 2. (Note that the intercept-only model and random-intercept model is identical in each MLM, and hence is only listed once.)

Within our statistically reliable random-intercepts models, we found that differences in Extraversion (*b* = -0.002, *t* = -3.16, *p* = 0.002), Agreeableness (*b* = -0.003, *t* = -3.92, *p <* .0001) and Conscientiousness (*b* = -0.002, *t* = -4.37, *p* < .0001) were all significant predictors of similarities in thematic cosines across romantic writing. With negative slopes, this suggests that smaller differences in personality predicted larger thematic cosines. Difference in Emotional Stability (*b* < .001, *t* = 0.47, *p* = 0.64) was not a significant predictor similarity in thematic cosines. For convenience, see Table 3 for predictors, intercepts, significance levels and standard errors of each predictor.

**Discussion**

Our results show that similarity in Extraversion, Agreeableness and Conscientiousness are significant predictors for similarities in thematic cosines. With the largest *b*-value, the significance of Agreeableness as a predictor aligns neatly 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. These results mirror Botwin’s findings that, alongside Agreeableness, similarity in Extraversion and Conscientiousness were significant predictors of desirability and marital satisfaction. Yet, this study extends beyond a simple replication of previous findings.

Our findings show linguistic modelling of the relationship of personality to mate preference returns similar conclusions as traditional survey methods. This suggests that thematic cosines are a valid measurement of mate preference with respect to the existing literature. So, what justifies and necessitates the use of a new measure if existing survey methods return similar results?

If we examine the structure of our measure versus traditional survey data, we begin to see profound differences. By construction, traditional surveys examining mate preference utilize prefabricated questions centered around strong predictors for mate preference in the existing literature. Naturally, responses are usually rated on a Likert-scale, which restricts a construct to analysis through an ordinal measure. Often to avoid complications, we assume Likert-style data behaves similarly to continuous measures. This is to satisfy the statistical assumptions of the parametric statistical tests used to convert our data into publication-worthy results. Usually, this far-reaching assumption rests on the hypothesized continuity of the underlying construct.

However, applied statisticians in several fields, such as Jamieson (2004) in Oncology or Grice, Barrett, Schlimgen and Abramson (2012) in Behavioral Science reject the validity of parametric statistical tests with ordinal measures. Among critiques of these methods are the misapplication of standard deviation from group means as a measure of central tendency or the incorrect pairing of discrete measures with statistical tests which depend on continuous variances to construct a normal distribution.

While Grice et al. (2012) propose Observation Oriented Modeling as one of many elegant alternatives to null hypothesis testing, many talented researchers across the Behavioral Sciences depend on traditional parametric statistics to conduct their work. So, as a complement to these non-parametric methods of analysis, we suggest constructing measures which better suit the Mathematical underpinnings of null hypothesis testing. In our study, by utilizing writing as a measure of mate preference, participants interacted with a continuous measure instead of the traditional ordinal survey measures. Since mate preference is observed to be stable across culture (Buss, 1989) and age (Schwarz & Hassbrauck, 2012), it is logical to assume that the distribution of mate preference across the global population is a uniformly normal distribution.

This necessitates the construction of continuous measures (such as thematic cosines) to examine mate preference through null hypothesis testing. That does not mean writing is a perfect measure for quantifying mate preference, nor are we guaranteed that the distribution of thematic cosines is identical to the uniformly normal distribution of mate preference. However, we do know that the statistical assumptions and screenings of our Multilevel Model are conducted appropriately with respect to a continuous, ratio dependent variable.

That our findings mirror the previous literature on mate preference suggests that Latent Semantic Analysis is a fitting complement to survey measures in this domain. Not only does it provide appropriately continuous data for parametric statistical tests, it allows for measurement of the construct in a flexible, unrestricted paradigm. Further research is needed to understand the validity of Latent Semantic Analysis in this context. However, in a scientific community centered on ethical and diverse statistical research methods, we hope measures of this type will provide our colleagues with the freedom and reassurance to pursue complex research questions.

References

Axler, S. (2015). Polar Decomposition and Singular Values Decomposition. In (3rd ed.) *Linear Algebra Done Right (*pp. 233-239). New York, NY: Springer Publications.

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)

Grice, J.W., Barrett, P.T., Schlimgen, L.A., & Abramson, C.I. (2012). Toward a brighter future for psychology as an observation oriented sciences. *Behavioral Sciences*, *2*(1), 1-22. doi: 10.3390/bs2010001.

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

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, *38*(12), 1217–1218. doi:10.1111/j.1365-2929.2004.02012.x

Kenrick, D. T., Sadalla, E. K., Groth, G., & Trost, M. R. (1990). Evolution, traits, and the stages of human courtship: Qualifying the parental investment model. *Journal of Personality*, *58*, 97-116. doi:10.1111/j.1467-6494.1990.tb00909.x

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

Miller G.F. (2000). *The mating mind: how sexual choice shaped the evolution of human nature*. New York City, New York: Doubleday.

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

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

Table 1

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Personality Measures | Female M | Female SD | Male M | Male SD | Cohen’s *d* |
| 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 |
| Conscient. (\*\*) | 37.65 | 7.09 | 34.64 | 6.03 | 0.46 |
| Emotional Stability | 26.61 | 7.51 | 32.00 | 7.94 | 0.70 |

Table 2

*Multilevel Model of participants’ thematic cosines with individual model’s Akaike Information Criterion and significance values.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Models | | *df* | | AIC | BIC | Log. Lik. | Ratio | *p* |
| Simple Regression | | 2 | | -1755.52 | -1744.29 | 879.76 |  |  |
| Random Slopes | | 3 | | -2168.23 | -2151.40 | 1087.12 | 414.72 | < .0001 |
| Openness | | 4 | | -2168.37 | -2145.912 | 1088.18 | 2.13 | .14 |
| Extraversion | | 4 | | -2176.166 | -2153.715 | 1092.0832 | 9.93 | .002 |
| Agreeableness | | 4 | | -2181.519 | -2159.067 | 1094.7593 | 15.28 | .0001 |
| Conscientiousness | | 4 | | -2185.26 | -2162.81 | 1096.63 | 19.03 | < .0001 |
| Emotional Stab. | | 4 | | -2166.46 | -2144.01 | 1087.23 | 0.22 | .64 |

Table 3

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



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | *b* | SE | *t* | *p* |
| Openness | 0.001 | 0.0001 | 1.46 | .14 |
| Extraversion | -0.002 | 0.0001 | -3.16 | .002 |
| Agreeableness | -0.003 | 0.001 | -3.92 | < .0001 |
| Conscientiousness | -0.002 | 0.001 | -4.37 | < .0001 |
| Emotional Stability | <0.001 | 0.001 | 0.47 | .64 |