Who Likes What? Comparing Personal Preferences with Group Predictions based on Gender and Extraversion Across Common Semantic Domains.

Simon De Deyne (simon.dedeyne@unimelb.edu.au) and Andrew Perfors (andrew.perfors@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Australia

Abstract

Some people like coffee while others prefer tea, but little is known about whether preferences like these are shared among groups and whether they vary systematically across many common semantic categories. This study addresses this gap by examining two major sources of variation – gender and extraversion – across twelve categories or domains, ranging from fruit and animals to sports and personal qualities. In Study 1, participants rated their own preferences for a set of 300 exemplars. Results showed significant preference differences between men and women for 40% of items spread across all categories, and smaller but reliable differences between introverts and extraverts for 11% of items concentrated in domains like personal qualities. Study 2 used an allocentric categorisation task where the same participants categorised items based on which they thought would be preferred by men vs women or introverts vs extraverts. Using the ratings from Study 1 to score accuracy, the judgments from Study 2 showed that participants were sensitive to even subtle differences in preference, although accuracy varied by the judge's gender and extraversion: women were more accurate than men across many categories and introverts more accurate than extraverts for a few categories. We also found incorrect but widely shared judgments for about 20% of items, suggestive of inaccurate stereotypes about group preferences. Together, these results suggest widespread and systematic variation by gender (and to a lesser extent extraversion) that can be accurately predicted by others, although with systematic biases. Our results have implications for theories of semantic representation and social cognition.

Keywords: concepts; preferences; semantic variation; gender, extraversion;

Introduction

Semantic representations are, by definition, predominantly shared: communication is only possible because we all think of approximately the same ideas when hearing words like *bicycle* or *beach*. While individuals may vary slightly in the connotations they give to these words, there is some indication that there is systematic semantic variation based on characteristics like gender, age, and geographical location (e.g., Capitani, Laiacona, & Barbarotto, 1999). A potential source of variation might stem from specific individual interests and preferences, which are highly subjective and grounded in each person's unique experience. As yet, we know little about how much variation in semantic representations due to preference differences exists in general, how (or whether) this varies for different kinds of categories, and how much is systematic (rooted in shared group characteristics) rather than idiosyncratic and individual.

The first goal of the current study is to investigate to what degree semantic cognition reflects systematic and consistent variation in personal preferences.

One possibility is that subjective factors like preference are highly idiosyncratic and only explain a negligible proportion of the variation in semantic representations across people. A second is that these factors are distinctive for some semantic domains or categories but not others. The third possibility is that preferences are substantially shared by people with the same personal or demographic characteristics, in which case those characteristics might explain semantic variation across very different domains. We evaluate these possibilities, focusing primarily on the essential step of precisely measuring the amount and nature of patterned variation that exists over a wide range of semantic domains as well as across two sources of systematic difference (gender and extraversion).

Our second goal is to investigate to what extent people have insight into systematic differences in other people's preferences. For instance, can they accurately predict whether more men prefer *bananas* or *strawberries*, or whether more introverts prefer *cats* or *dogs*? Are there regularities in which people are better than others, or which domains people find easiest to predict? How common is it for people to agree with each other but all be consistently wrong, as would happen if they share an incorrect stereotype? Although parts of some of these questions have been studied before, to our knowledge, this has not been investigated quantitatively or systematically for a wide range of items in multiple semantic domains.

This is important given that successful communication requires forming an accurate mental picture of other people's semantic representations. Deficiencies in this ability may be the root of a substantial misunderstanding, especially across demographic divides. It is important to be able to measure and document where these deficiencies lie, especially where the majority rely on incorrect stereotypes or where there are systematic blind spots.

Another reason this is important is society's increasing reliance on Large Language Models (LLMs). Because these models are trained on human-generated data but are in many ways "black boxes" with respect to the nature of their underlying representations, it is clear that they have internalised many human biases but not what those biases *are*. This is problematic not only because people are overly confident about the performance of LLMs (Steyvers et al., 2025) but also because biases with respect to social categories like gen-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

der (Caliskan, Bryson, & Narayanan, 2017) can have deleterious real-world effects. Part of the challenge of understanding the general semantic representational biases within LLMs is that we do not know exactly what *human* biases look like. Do LLMs capture the same preference structure as we observe empirically in people? How do they compare to humans in predicting preferences as a function of gender or extraversion? By robustly establishing a human baseline, our work sets the stage for future research that systematically evaluates LLMs.

We address our main questions in two separate tasks.

Study 1 is a *category preference rating task* that overcomes some limitations of previous studies that look at the related construct of emotional valence (Warriner, Kuperman, & Brysbaert, 2013). It is different from them in two key ways. First, it directly measures actual (egocentric) preferences by asking participants about what *they themselves* like. Second, items are presented contextually (e.g., asking about *apples* as an instance of *fruits* rather than just in general); this lets us avoid the ambiguity that arises from polysemy (Reijnierse, Burgers, Bolognesi, & Krennmayr, 2019).

Study 2 is an *allocentric categorisation task* in which people are presented with (e.g., *banana*) and asked which group (either males/females or extraverts/introverts) prefers that item the most. By comparing their choices with the results from Study 1, we can measure how accurate people are in real life and whether there are systematic deviations from accuracy.

Study 1: Preference Rating Task

Participants. The same 563 native-English speaking undergraduates participated in both Study 1 and Study 2 in a single hour-long session, with Study 1 occurring first. Three people were removed because they gave the same response to all questions on the BFI, six because they were non-binary, and 49 because their ratings either correlated less than .1 with the average ratings, or they did not know more than 50 exemplars. This left 505 in the final sample (310 female, 195 male), aged between 17 and 63 (mean: 19.8, 98% younger than 30).

Materials and Measures. Our stimuli consisted of 300 words covering 12 different semantic domains (25 in each domain).[1] There were six concrete domains (*Four-footed Animals*, *Fruit*, *Vegetables*, *Colours*, *Vehicles*, *Musical Instruments*) and six abstract/social ones (*Sports*, *Academic Subjects*, *Professions*, *Personality Traits*, *Values*, and *Pastimes*).

Personality was measured using the 60-item Big Five Inventory-2 (Soto & John, 2017). Given our focus on extraversion, we also included the 20 extraversion items from the Big Five Aspects Scale (John, Naumann, & Soto, 2008).

Procedure. Each person completed Study 1 and 2, followed by the personality questionnaires. Each of the 12 trials in Study 1 corresponded to one of the 12 semantic categories,

[1]Materials and scripts: `https://osf.io/p65ta/`.

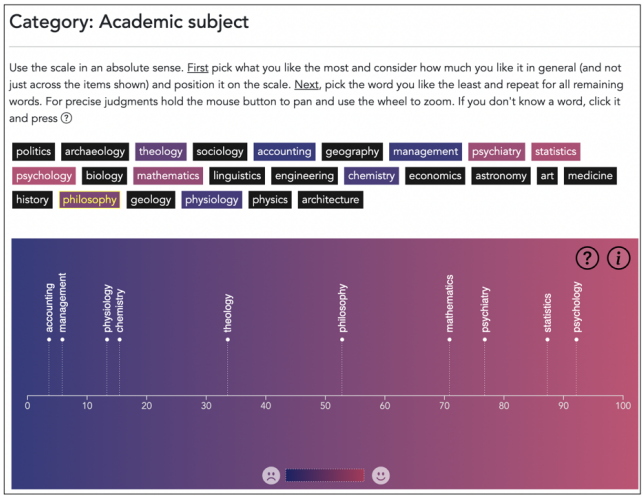

Figure 1: Screenshot of the preference rating task (Study 1) for one of the 12 categories (*Academic Subject*). For each category, people were asked to place each of the 25 exemplars on an axis from 0 to 100, where lower numbers indicated dislike. In this example, the participant has rated 10 of the 25 exemplars so far, and likes *psychology* best and *accounting* least.

as shown in Figure 1. On each trial, people saw a dynamic sliding scale underneath a list of all 25 exemplars in a random order. They were told to drag and drop the exemplars on the scale, in which 1 signified a strong degree of dislike and 100 a strong degree of like. People could skip words they did not know by tagging them accordingly, and they were not permitted to continue to the next trial until all 25 exemplars were placed. Study 1 took 24 minutes on average.

Results

The average inter-individual correlation between exemplar ratings was .49 (*SD*=.14), and the Spearman-Brown split-half reliability was very high ($r_{sb} \geq .98$ for each category). To assess any effects of gender imbalance, we also calculated reliability for males and females separately and again very high values ($r_{sb} \geq .97$ for each category).

Gender. Given the negatively skewed distribution of preference ratings, we use a two-sample Mann–Whitney U test to compare the ratings of men and women, controlling for false discovery rate by adjusting the p-values via the Benjamini-Hochberg method (Benjamini & Hochberg, 1995). As Table 1 reveals, multiple items in every domain showed a significant difference in preferences between genders (from 24% for *Personal Qualities* to 76% for *Pastimes*). Overall, 122 of the 300 exemplars (40.7%) were significantly different across gender. While there are several measures to express effect size for the Mann-Whitney U test, we chose Vargha and Delaney's A (VDA), which captures the probability that a value from one group is greater than a value from the other group (Vargha & Delaney, 1998). Overall, 97 (32.3%) of the items were significant and *also* had non-negligible effect sizes; details and indicative examples are shown in Table 1.

Table 1: **Study 1 significant differences.** Across 12 categories (top six concrete, bottom six abstract), the results show the number of items with significant differences in preference between men and women (left side) or introverts and extraverts (right side). The % sign indicates the percentage of items that were significantly different. The subsequent columns report how many exemplars (out of 25) had small (S), medium (M), or large (L) effect sizes. The two with the largest effect size are listed in the appropriate column. For instance, 52% of colours differed significantly by gender, with women preferring *pink* and *lilac* more than men, and men preferring *black* and *orange* more than women.

	Gender						Extraversion				
Category	%	S	M	L	**Female Prefer**	**Male Prefer**	%	S	M	**Introvert Prefer**	**Extravert Prefer**
Colour	52	9	1	1	*pink*, *lilac*	*black*, *orange*	0	0	0		
Animal	48	5	4	1	*pony*, *deer*	*gorilla*, *crocodile*	12	3	0	*cat*, *deer*	
Fruit	44	7	0	0	*cherries*, *strawberries*	*banana*, *apple*	4	1	0	*melon*	
Instrument	40	4	0	0	*harp*, *didgeridoo*	*trumpet*, *synthesizer*	12	3	0	*violin*	*didgeridoo*, *banjo*
Vegetable	32	5	0	0	*cucumber*, *zucchini*	*onion*, *spinach*	0	0	0		
Vehicle	52	6	2	1	*limo*, *scooter*	*tank*, *spaceship*	0	0	0		
Academic	52	9	3	0	*art*, *sociology*	*physics*, *engineering*	4	1	0		*management*
Pastime	76	8	4	4	*knitting*, *baking*	*video games*, *sports*	24	5	1	*video games*, *reading*	*socialising*, *beach*
Personal	24	4	0	0	*empathetic*, *bubbly*	*cold*, *shy*	56	14	0	*quiet*, *introvert*	*extravert*, *social*
Profession	40	6	2	0	*secretary*, *writer*	*carpenter*, *astronaut*	8	2	0		*lawyer*, *politician*
Sports	56	8	3	1	*dance*, *gymnastics*	*fishing*, *chess*	12	3	0	*badminton*, *archery*	*football*
Values	32	3	1	0	*equality*, *security*	*prestige*, *pleasure*	0	0	0		

Extraversion. The average extraversion score on the BFI-2 for women was 3.13 (SD = 0.67) and for men was 3.10 (SD = 3.10), slightly lower than the 3.31 and 3.20 scores (respectively) reported for US students (Soto & John, 2017). The correlation between the extraversion scores for BFI-2 and BFAS was .87 for women and .85 for men. We obtained an overall extraversion score for each person by averaging the scores from the two questionnaires and classified a person as an introvert if their score was below the mean.

Because extraversion is continuous while gender is binary, effect sizes were calculated using Kendall's tau (r_τ) and interpreted following Schober, Boer, and Schwarte (2018). As Table 1 shows, a much smaller percentage (11%) of items showed significant extraversion differences, with only one item having a medium effect and the other 32 having a small effect size. Most significant effects were concentrated in two categories: *Pastimes* and *Personal Quality*.

Study 2: Allocentric Categorisation Task

Study 2 investigates how accurately participants can take an allocentric perspective when judging the relative preferences of people of different genders or levels of extraversion.

Participants Participants from Study 1 were randomly assigned to the GENDER condition (147 female, 101 male) or EXTRAVERSION condition (163 female, 94 male).

Materials and Procedure. The 300 exemplars were identical to those in Study 1. People were shown the exemplars in 12 randomly ordered blocks corresponding to the 12 semantic domains. In each block, they saw the exemplars from that domain appear in random order, one-by-one, in the centre of the screen. People in the GENDER condition indicated which gender preferred each item, pressing *m* for male and *f* for female. Those in the EXTRAVERSION condition made the same classification but about introverts vs extraverts (*i* for introvert and *e* for extravert). People pressed *x* for unfamiliar words, and the task took an average of 11 minutes.

Results

After removing the words marked as unfamiliar (0.7% in GENDER and 0.6% in EXTRAVERSION), we coded the `accuracy` of each response as a binary variable based on the results from Study 1. To do so, Study 1 ratings were first standardised by z-transforming each person's scores to ensure baseline differences across categories did not affect results. We then calculated the preference difference across genders by subtracting men's average preference for that exemplar from women's average preference. As an example, since the results from Study 1 indicated a female preference for *pink*, all participants in Study 2 who said women preferred *pink* obtained an `accuracy` of 1 (correct) on that item.

Gender. Both female and male participants were more accurate than chance (50%), with women obtaining an average accuracy of 66.4% (range: 52.9% to 74.2%) and men an average of 64% (range: 51.1% to 75.4%). There was also a slight bias in both groups to respond in a way that was congruent with their own gender: women responded "female" on 51.9% of trials, and men responded "male" on 58.7%.

To investigate accuracy across categories, we used the R package brms (Bürkner, 2017) to fit a Bayesian logistic regression model in which the binary outcome variable was `accuracy` and the predictor variables were `gender` and `category`, with an interaction term and a random intercept for participant. All models used a Bernoulli likelihood and logit link function, and were estimated using 2000 iterations across four chains. Noninformative priors were used for all parameters, and convergence was confirmed (Rhat < 1.01).

The odds ratios (OR) and surrounding credible intervals in all twelve categories excluded 1, meaning that performance was above chance on all of them. However, the size of the

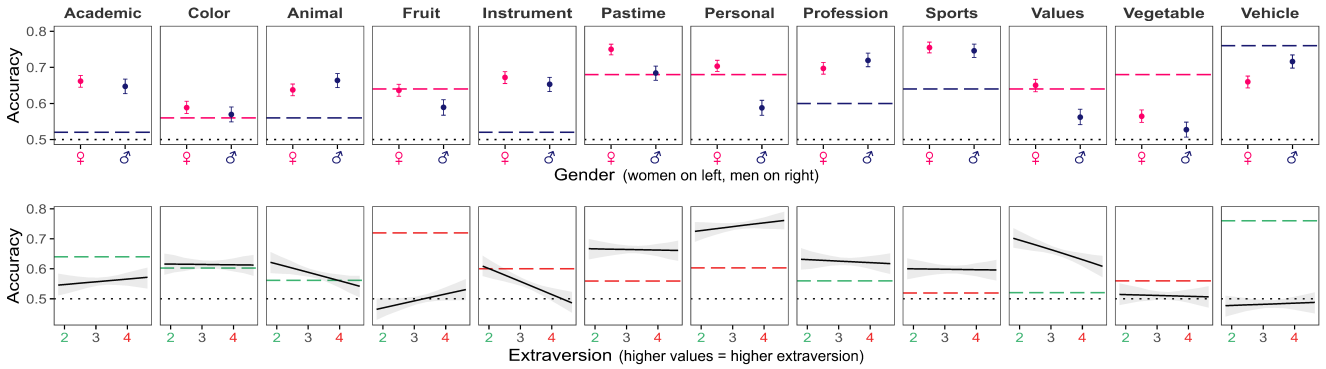

Figure 2: **Study 2 performance across 12 domains.** Each panel shows accuracy effects in predicting exemplar preferences from Study 1 by either GENDER (top) or EXTRAVERSION (bottom). For GENDER, dots with error bars indicate the estimated accuracy (95% CI) for women (pink) and men (blue). For EXTRAVERSION, which is continuous, the black line shows the estimated slope and 95% CI. Coloured horizontal lines reflect the baseline accuracy achieved by always picking the dominant response for that category; the colour of the line indicates which response is dominant (pink = female, blue = male, green = extraverts, red = introverts). For example, for *Academic Subject* with GENDER, the blue baseline shows that most items were thought to be preferred by men, although this baseline was close to 0.5. Women were slightly more accurate than men (0.68 vs 0.66). With EXTRAVERSION, the green category baseline shows that people thought introverts preferred most academic subjects, and the positive slope suggests that extraverts were slightly more accurate.

OR varied significantly by category, with people performing most accurately for *Sports*, *Pastimes*, *Professions*, and *Vehicles*, all of which had an OR > 2 and significant 95% CIs. This indicates that people were generally good at predicting which items were preferred by which gender for all categories, but were more accurate for some categories than others.

We also ask whether women or men were more accurate at predicting the preferences of others (and, if so, which domains each gender found easiest). Table 2 shows that women were more accurate than men for *Fruit*, *Vegetables*, *Pastimes*, *Personal Qualities*, *Values*. Men were more accurate for *Vehicles*, and there were no differences in other domains.

The top row of Figure 2 shows similar findings, but with the data transformed so that proportion correct is indicated on the *y* axis. None of the credible intervals overlap with the chance baseline of 0.5, indicating that people could predict preferences by gender in all 12 categories. Dashed horizontal bars show the baseline accuracy achieved if all exemplars received the modal response (blue if it is "male", pink if it is "female"). This corresponds to a situation where participants believe, for instance, that women have a higher preference for fruits in general. People perform at or above this category baseline in all but two categories (*Vegetables* and *Vehicles*).

Extraversion. Since EXTRAVERSION is continuous, we dichotomised the ground truth from Study 1 based on the mean ratings for each exemplar (with lower ratings corresponding to introvert and higher corresponding to extravert) to mirror the GENDER condition's binary comparison. This resulted in 151 items being coded as extravert and 149 as introvert. We also coded each participant as an introvert or extravert themselves on the basis of their extraversion score (lower scores indicating introversion).[2]

[2]We also considered performance for EXTRAVERSION split by GENDER. While there were some differences, these were limited to specific exemplars and did not affect the overall pattern of results.

Both extraverts and introverts were more accurate than chance, with extraverts obtaining an average accuracy of 59% (range: 49.7% to 67.0%) and introverts an average of 59.7% (range: 47.0% to 68.0%). Both groups had a slight baseline assumption that extraverts would prefer more items (for introverts, 52.3%; for extraverts, 52.4%).

Next, we fitted a Bayesian logistic regression with `extraversion` in place of `gender` and a binary outcome corresponding to accuracy on the EXTRAVERSION trials. After converting the estimated marginal means from different categories to odds ratios, we found that nine categories had ORs and CIs greater than 1; one category (*Vehicles*) was less than

Table 2: **Accuracy differences in Study 2.** Each row corresponds to a category, showing the odds ratios and 95%CI for men vs women (left) and introverts vs extraverts (right) in their accuracy in predicting the true preferences from Study 1. For GENDER, OR < 1 (in blue) means men were more accurate (CIs do not overlap with 1); OR > 1 (in pink) means women were more accurate. For EXTRAVERSION, OR < 1 (in red) indicates introverts were more accurate. Performance varies by domain, with women generally outperforming men, and introverts sometimes outperforming extraverts.

	Gender		**Extraversion**	
Category	OR	CI	OR	CI
Colour	1.08	[0.97, 1.21]	1.00	[0.90, 1.10]
Animal	0.89	[0.79, 0.99]	0.89	[0.81, 0.99]
Fruit	1.22	[1.08, 1.37]	1.10	[1.00, 1.21]
Instruments	1.09	[0.96, 1.21]	0.84	[0.77, 0.93]
Vegetable	1.16	[1.04, 1.29]	0.99	[0.89,1.09]
Vehicle	0.77	[0.69, 0.87]	1.02	[0.93, 1.11]
Academic	1.07	[0.95, 1.20]	1.04	[0.95, 1.14]
Pastimes	1.39	[1.23, 1.56]	0.99	[0.90, 1.10]
Personal	1.66	[1.47, 1.85]	1.07	[0.96, 1.19]
Profession	0.90	[0.79, 1.01]	0.98	[0.89, 1.08]
Sports	1.05	[0.92, 1.18]	1.00	[0.91, 1.10]
Values	1.45	[1.30, 1.62]	0.86	[0.78, 0.95]

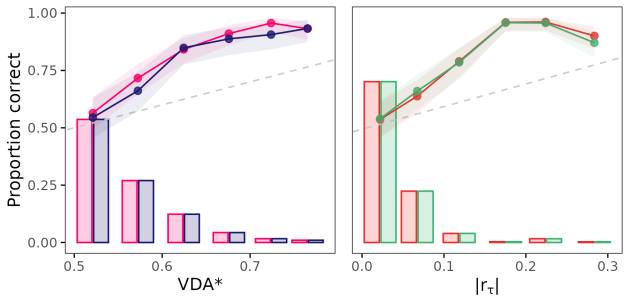

Figure 3: **Relationship between accuracy and effect size.** Within each condition (GENDER on the left, EXTRAVERSION on the right), the lines show the proportion correct (y axis), coloured by the demographic of participants (pink for women, blue for men, green for extraverts, red for introverts). The x axis shows the effect size, broken down into equal-sized bins covering the full range (the height of the bars corresponds to the number of observations at that effect size). This figure shows that people are most accurate when effect sizes are large, but accuracy is high even for relatively small effects.

1, indicating below-chance performance. Two domains, *Vegetables* and *Fruit*, were consistent with chance (CIs overlapping 1). Finally, we explored whether the participants performed differently depending on their level of extraversion. As shown in Table 2 and Figure 2, the effects were small and limited to three domains (*Animals*, *Instruments*, and *Values*), all of which introverts were more accurate for.

Exemplar Accuracy and Stereotypes. To investigate how accuracy varied for each item and compare it with the effect sizes of Study 1, we fitted a new logistic regression that was similar to the previous ones but with a term for `item` and an interaction between it and the binary condition term (`gender` or `extraversion`). We then derived significance by counting the number of items where the CIs did not overlap with 1.

To understand how accuracy calculated over exemplars relates to the size of the difference found in Study 1, we grouped the range of effect size values (as measured by VDA) into six equal-sized bins and calculated the proportion of correct items within each bin. As Figure 3 shows, people in the GENDER condition were accurate even for relatively small effect sizes, especially when there were differences in accuracy across gender. People in the EXTRAVERSION condition showed a similar pattern, achieving above-chance accuracy even when the effect size (measured by r_τ) was small (though with no differences in accuracy by extraversion levels).

Exemplars with low accuracy might reflect random responding or disagreement, but might also reflect incorrect stereotypes – situations where people agree with each other about what preferences they *expect*, but are incorrect about the actual preferences people *have*. Figure 4 shows that in both the GENDER and EXTRAVERSION conditions, not all exemplar predictions were significant, regardless of accuracy. The pink rectangles correspond to the exemplars for which we observed incorrect stereotypes; these were relatively common among all groups of participants. To explore them further, we selected the subset of these items for which the effect size in Study 1 was either negligible or in the opposite direction.

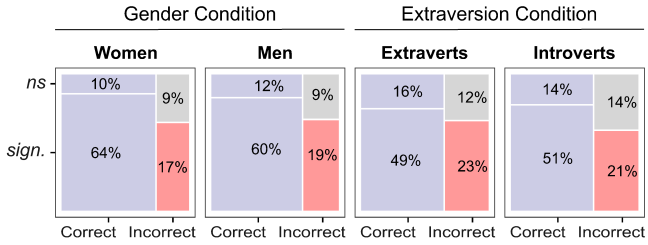

Figure 4: Study 2 exemplar accuracy split by significance with insignificant effects (*ns*) referring to an OR of 1 within the 95% CI) for GENDER (left two panels) and EXTRAVERSION (right two panels).

Figure 5 shows these items by plotting them against the effect size from Study 1. We observed stereotypes in multiple domains. For instance, women in Study 2 thought *apple* would be strongly preferred by women, but Study 1 actually showed that men preferred it; men in Study 2 strongly assumed that women would prefer *humility*, but in actuality, the effect was negligible and leaned towards being preferred by men. Figure 5 is also revealing about the overlap in stereotypes – which ones are believed by both genders as opposed to only one. 63.6% of the stereotypes about women and 58.3% of the ones about men were shared; as an example, both genders thought that *magenta* would be preferred by women, even though men actually preferred it. Conversely, some stereotypes are gender-specific: for instance, women incorrectly thought women would prefer *philosophy* more, and men incorrectly thought women would prefer *papayas* more.

Extraversion stereotypes were also found in all domains (Figure 5). For example, introverts thought that *running* would be preferred by introverts, but Study 1 ratings showed that extraverts actually preferred it. There were also many stereotypes that are shared by both introverts and extraverts alike: e.g., *nurse* was thought by everyone to be preferred by extraverts, when in reality it was evenly preferred by both.

Discussion

Two studies demonstrated systematic preference differences across twelve common semantic domains depending on a person's gender and extraversion. In Study 1, where participants were asked about their own (egocentric) preferences, we observed a high degree of overall agreement between individuals, as well as systematic differences between demographic groups. Individual variation was most pronounced for gender: there were significant differences (after correcting for multiple comparisons) between the preferences of men and women on 40% of items in all domains. There was less variation based on personality, where effect sizes were small and significant for only 11% of items and a subset of domains.

In Study 2, the same participants judged whether exemplars were preferred by men/women or introverts/extraverts. The average choice proportions derived from this allocentric categorisation task were reasonably accurate compared to the actual preferences measured in Study 1, even when these

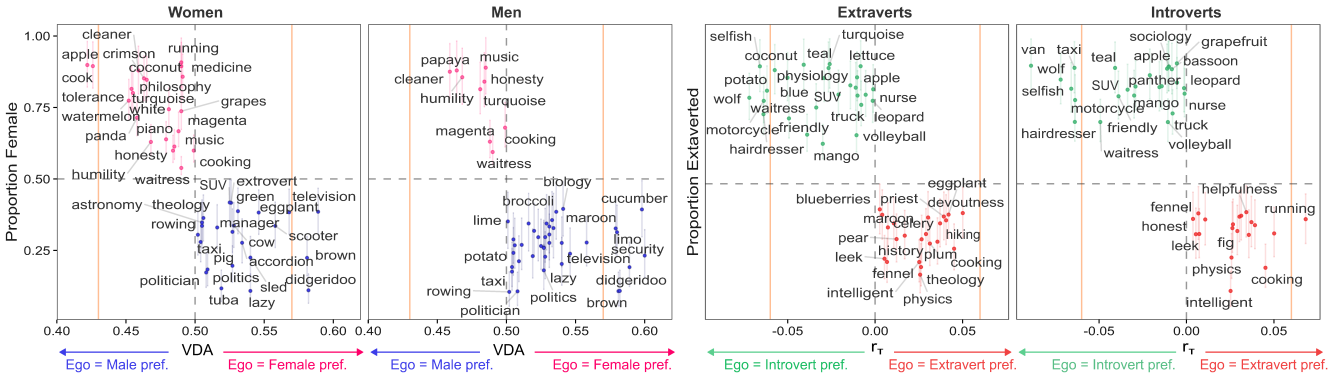

Figure 5: **Stereotypes by GENDER and EXTRAVERSION**. Each panel shows items that participants in Study 2 incorrectly believed were preferred by a given group, contrary to Study 1's egocentric ratings. The left two panels show incorrect gender stereotypes by men and women, while the right two panels show incorrect extraversion stereotypes by extraverts and introverts. The *y*-axis indicates the assumed preferences in Study 2, with the *x*-axis indicating the actual effect size observed in Study 1. For GENDER, items in the upper left were thought to be preferred by women but were actually preferred by men (in pink to indicate that it is a stereotype about women); for EXTRAVERSION, items in the upper left were thought to be preferred by extraverts but were actually preferred by introverts (in green to indicate that it is a stereotype about extraverts). The solid vertical lines delineate the difference between small and negligible effects.

were subtle. Accuracy was somewhat better for predicting gender differences, but this was not pronounced, given that the effect sizes in Study 1 were much larger for gender. We also found that women were better at taking an allocentric perspective, making more accurate predictions than men on five of the 12 categories. Introverts were also slightly better at judging preferences according to extraversion, but this finding was limited to only three categories.

The allocentric categorisation task in Study 2 also provided evidence of stereotypes for a substantial subset of exemplars. We found high levels of agreement about predicted preferences in about 20% of items, where the actual differences were negligible or in the opposite direction. We might ask why people in Study 2 exhibited stereotypes despite being exposed to considerable training by making judgments themselves in Study 1. One possibility is that participants used a relative domains baseline, which means they discounted the base rate of preference in a domain because they expected preferences to be more equally split between groups. For example, if participants believed that men in general preferred vehicles, they still thought about which *specific* vehicles had a higher preference among women.

A second possibility is that participants in the allocentric categorisation task adopted an associative strategy for at least a subset of harder items, where they considered how *related* items were to either group when they were responding. To explore this explanation, we compared our findings with previous research where participants were asked to judge the degree to which a word is associated with men and women (Scott, Keitel, Becirspahic, Yao, & Sereno, 2019). Using the gender association norms from Scott et al. (2019), we identified 197 words that overlapped with the current study. The correlation between the proportion of male judgments and gender association was r = .91 (males) and r = .86 (females), and r = .70 for the VDA derived from rated preference (Study 1). This suggests more overlap between allocentric judgments than egocentric judgments, which tentatively supports the proposal that associative relations might bias the judgments in Study 2. In social psychology, this idea is at the basis of the implicit association test and used as a measurement of stereotypes and biases in humans, but also word embeddings and LLMs (Caliskan et al., 2017; Lewis & Lupyan, 2020). The weaker relation with Study 1 points towards an alternative based on egocentric judgments is less prone to such associative biases.

Future directions. This study focused on extraversion as we considered it a construct that most participants would understand well (participants were mainly psychology undergraduates). While extraversion differences were observed for domains like *Pastime activities* or *Personal Quality*, where we expected them to occur, the effect of extraversion was observed for other concrete noun categories (*Fruit*, *Animals*, *Vegetable*, and *Vehicles*) as well. A priori, it is unsurprising that categories like *Pastime activities* or *Personal Qualities* would provide evidence for preference differences among gender and extraversion. However, our work extends these findings to categories where we might not have expected to see systematic significant differences. Similar to previous work (e.g., Caliskan et al., 2017), we expect that factual information such as occupation statistics or consumer studies (e.g., consumption of fruit and vegetables) would corroborate the external validity of our findings.

Finally, further insights can be gained by investigating the interplay between personality and demographic differences as they affect preferences and semantic variation. While the current Australian student sample is relatively homogenous, it is likely that there are other factors, such as cultural differences or language background also play a role. A systematic exploration of these questions will be the subject of future work. Similarly, it is possible that other factors aside from extraversion also provide a basis for preference variation. For example, preliminary results suggest that Openness could account for more category variation where extraversion effects were small.

Acknowledgments

This work was supported by Discovery Project DP240101873 funded by the Australian Research Council. We are grateful to Zawi Hoodbhoy for the design and data collection support. We also express our gratitude to Luke Smillie, Steven Verheyen, and three anonymous reviewers for valuable suggestions on earlier drafts of this manuscript.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289–300.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Capitani, E., Laiacona, M., & Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex*, *35*(2), 273–278.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm Shift to the Integrative Big Five Trait Taxonomy. *Handbook of Personality: Theory and research*, *3*(2), 114–158.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*, 1021–1028.

Reijnierse, W. G., Burgers, C., Bolognesi, M., & Krennmayr, T. (2019). How polysemy affects concreteness ratings: The case of metaphor. *Cognitive Science*, *43*, e12779.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, *126*, 1763–1768.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*, 1258–1270.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., … Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 1–11.

Vargha, A., & Delaney, H. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23*, 170-192.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207.