How Different are Men and Women in Their Personality? A Reanalysis of the IPIP Big-Five Questionnaire

In their study, Weisberg et al. (2011) examine gender differences in personality traits, using the Big Five Aspects Scale (BFAS). They find significant gender differences in Neuroticism, Agreeableness, and Extraversion, with women scoring higher on average. They further investigate the differences on the aspects level (two aspects for each trait) and find even more pronounced differences. Unfortunately, they do not provide the data publicly.

For this reason I will resort to publicly available data from the Open-Source Psychometrics Project. Instead of the BFAS, they use the 50-item Big-Five Factor Markers from the International Personality Item Pool (IPIP; Goldberg, 1992), providing information only on the trait level. The data were collected from the website by visitors to the website who completed the questionnaire themselves and then consented to the data being stored. In total, data from N = 19719 individuals are available, with the variables consisting of 50 Likert-rated statements, gender, age, race, native language, and country. On their website, the Open-Source Psychometrics Project claims that the data are of equal or even better quality than Amazon Mechanical Turk.

Published Analyses / Methods

What Methods Did They Use?

To analyse their data, Weisberg et al. (2011) assumed the Likert-type responses to the BFAS to be normally distributed. Consequently, they used a metric model and calculated the mean and standard deviation for each personality trait separately for men and women and the effect sizes (Cohen's d) of the differences. To evaluate the statistical significance of the differences, they performed a t-test on each trait.

New Data - Old Method

To be able to compare their analysis with my suggested improvement, I used their methods to analyse the open-source data from the IPIP. For computational reasons –which are important for the Bayesian modeling part below – I restricted the analysis to a subset of 250 persons, randomly drawn from the entire dataset. The replicated analysis is documented in Table 1.

Trait	Men		Women		Cohen's *d*
	M	SD	M	SD	
Extraversion (E)	2.88	1.36	3.09	1.34	**0.16**
Neuroticism (N)	2.88	1.27	3.23	1.23	**0.27**
Agreeableness (A)	3.65	1.17	3.88	1.19	**0.20**
Conscientiousness (C)	3.36	1.17	3.28	1.24	-0.06
Openness (O)	4.04	1.08	3.82	1.08	**-0.21**

Table 1

Means, Standard Deviations, and Cohen's d for Personality Variables by Gender. Bold Cohen's d values indicate statistically significant differences (p < 0.05).

Issues With the Metric Model

By using a metric model, Weisberg et al. (2011) assume that the intervals between the response options are equal. However, Likert-type responses are ordinal-scaled and we do not know the how large the intervals between the response options are on the latent variables. Liddell and Kruschke (2018) argue that using a metric model anyways can result in false alarms (Type I errors) where differences are detected when none truly exist, misses (Type II errors) where real differences are not detected, and even inversions where the direction of effects is reversed.

Ordinal Regression as an Alternative

Liddell and Kruschke (2018) show that ordinal regression models are a fitting alternative that recover the probabilities for each response option well. Instead of estimating the mean and standard deviation of the normal distribution, ordinal regression estimates so called cutpoints. Using the ordered-logit model (ordered-probit works equally well), the cutpoints are the cumulative probabilities of the every response option on the logistic function (going from 0 to 1).

Model equations

If we have the data in the long format, i.e. one row for every answer of a single individual to a single question, then our *input* into the model will look like the following:

$$\begin{aligned}
K &= \text{number of response categories} \\
N &= \text{number of observations} \\
nq &= \text{number of questions for this trait} \\
q_n &\in \{1,\ldots,nq\} \text{ for } n = 1,\ldots,N \qquad = \text{question} \\
y_n &\in \{1,\ldots,K\} \text{ for } n = 1,\ldots,N \qquad = \text{response category} \\
x_n &\in \{1,2\} \text{ for } n = 1,\ldots,N \qquad = \text{gender (male or female)}
\end{aligned} \tag{1}$$

Following are the *parameters* that we want to estimate. Notice that θ_k is only relevant in so far as it serves as a prior for the general cutpoints c_k.

$$\begin{aligned}
\theta_k &: \text{prior probability over all categories } k \\
c_k &: \text{overall cutpoints for each category } k\text{, serves also as the prior for } c_{q,k} \\
&= \text{logit}\left(\sum_{i=1}^{k} \theta_i\right) \text{ for } k = 1,\ldots,K-1 \\
c_{q,k} &: \text{cutpoints for each question } q \text{ and category } k \\
\beta &: \text{regression coefficient for predictor gender}
\end{aligned} \tag{2}$$

It is obvious that all ten questions on any scale (e.g. Extraversion) are not independent of each other. For this reason it makes sense model the cutpoints hierarchically. We draw the prior for the over all cutpoints c_k from a dirichlet distribution, giving moderate response categories a higher prior probability than more extreme categories (see Figure 1). This is a so-called weakly-informative prior. In turn, c_k serves as a prior for the lower-level cutpoints $c_{q,k}$, i.e. cutpoints for each question separately. The predictor β has a flat prior as we have no valuable information beforehand. Here are the *prior* distributions:

$$\begin{aligned}\theta &\sim \text{Dirichlet}(4,5,6,5,4)\\ c_{s,k} &\sim \text{Normal}(c_k,2) \text{ for } s=1,\dots,nq \text{ and } k=1,\dots,K-1 \qquad (3)\\ \beta &\sim \text{Uniform}(-\infty,\infty)\end{aligned}$$

This leads us to our *likelihood*. The responses are drawn from an ordered-logit distribution, with gender as a predictor and c_{q_n} as the cutpoints:

$$y_n \sim \text{OrderedLogistic}(x_n\beta, c_{q_n}) \text{ for } n=1,\dots,N \qquad (4)$$

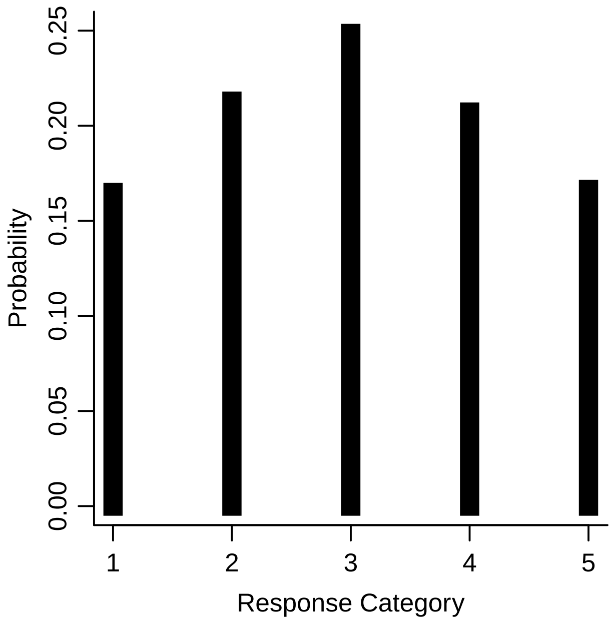

Figure 1

Prior predictive check for the over-all cutpoints c_k. Central (moderate) responses are more likely than marginal (extreme) responses.

I implemented the model with Stan (Carpenter et al., 2017).

Results

Cutpoints

The estimated cutpoints and corresponsing 95 % credibility intervals are documented in Table 2. In order to get from the estimated cutpoints to estimated probabilities we need to apply the inverse-logit function and translate from cumulative probabilities to probabilities. The resulting probabilities are illustrated in Figure 2.

Trait	c_1 [95% CI]	c_2 [95% CI]	c_3 [95% CI]	c_4 [95% CI]
Extraversion (E)	-1.40	-0.29	0.77	2.02
	[-2.20, -0.69]	[-0.93, 0.35]	[0.12, 1.47]	[1.22, 2.88]
Neuroticism (N)	-1.36	-0.20	0.87	2.13
	[-2.18, -0.61]	[-0.81, 0.42]	[0.24, 1.53]	[1.34, 2.97]
Agreeableness (A)	-2.08	-0.87	0.20	1.37
	[-2.95, -1.24]	[-1.55, -0.21]	[-0.45, 0.84]	[0.63, 2.20]
Conscientiousness (C)	-2.20	-0.94	0.17	1.34
	[-3.08, -1.37]	[-1.60, -0.29]	[-0.46, 0.81]	[0.61, 2.16]
Openness (O)	-3.42	-1.86	-0.60	0.58
	[-4.45, -2.47]	[-2.60, -1.15]	[-1.27, 0.05]	[-0.15, 1.38]

Table 2

Cutpoints (c) and their 95% Credibility Intervals for Big Five Personality Traits

Gender Effect

The research question I claim to provide evidence for is if men and women differ in their personalities as measured by the Big-Five factors. If they differed, then the β-coefficients must not be zero. The estimated parameters are documented in Table 3.

Comparison

When comparing the results of the metric analysis with the results of the ordered regression, we find no particular difference in our conclusion. In the metric analysis Extraversion, Neuroticism, Agreeableness, and Openness show a statistically significant difference for men and women. For the same traits we can be at least 95 % confident that gender as a predictor is non-zero.

Trait	β	2.5%	97.5%
Extraversion (E)	**0.29**	0.15	0.42
Neuroticism (N)	**0.47**	0.33	0.61
Agreeableness (A)	**0.35**	0.21	0.49
Conscientiousness (C)	-0.12	-0.26	0.02
Openness (O)	**-0.41**	-0.55	-0.26

Table 3

Beta coefficients and 95% credibility intervals for Big Five personality traits. Bold coefficients indicate credibility intervals that do not include 0.

Discussion and Contextualization

Sensitivity Analysis

I performed the analysis again for the trait Extraversion and used a flat prior (∼Dirichlet(1, 1, 1, 1, 1)). This did not impact the estimate of the gender effect. It did, however, change the shape of the probability distribution over the response options. Even though I expected my initial prior distribution to be only weakly informative, it had a visible impact on the posterior distribution.

Comparing the Approaches

Even though my prior distribution on the cutpoints influenced the posterior distribution to look more like a normal distribution (more likely in the middle and less likely on the margins), Figure 2 visualizes how different both approaches estimate the probability of a particular response option. For every single trait the metric model underestimates the probability of response categories 1 and 5 but works reasonably well for categories 2, 3, and 4.

My analysis supports the conclusion of the original paper that gender differences exist in all five personality traits except for Openness.

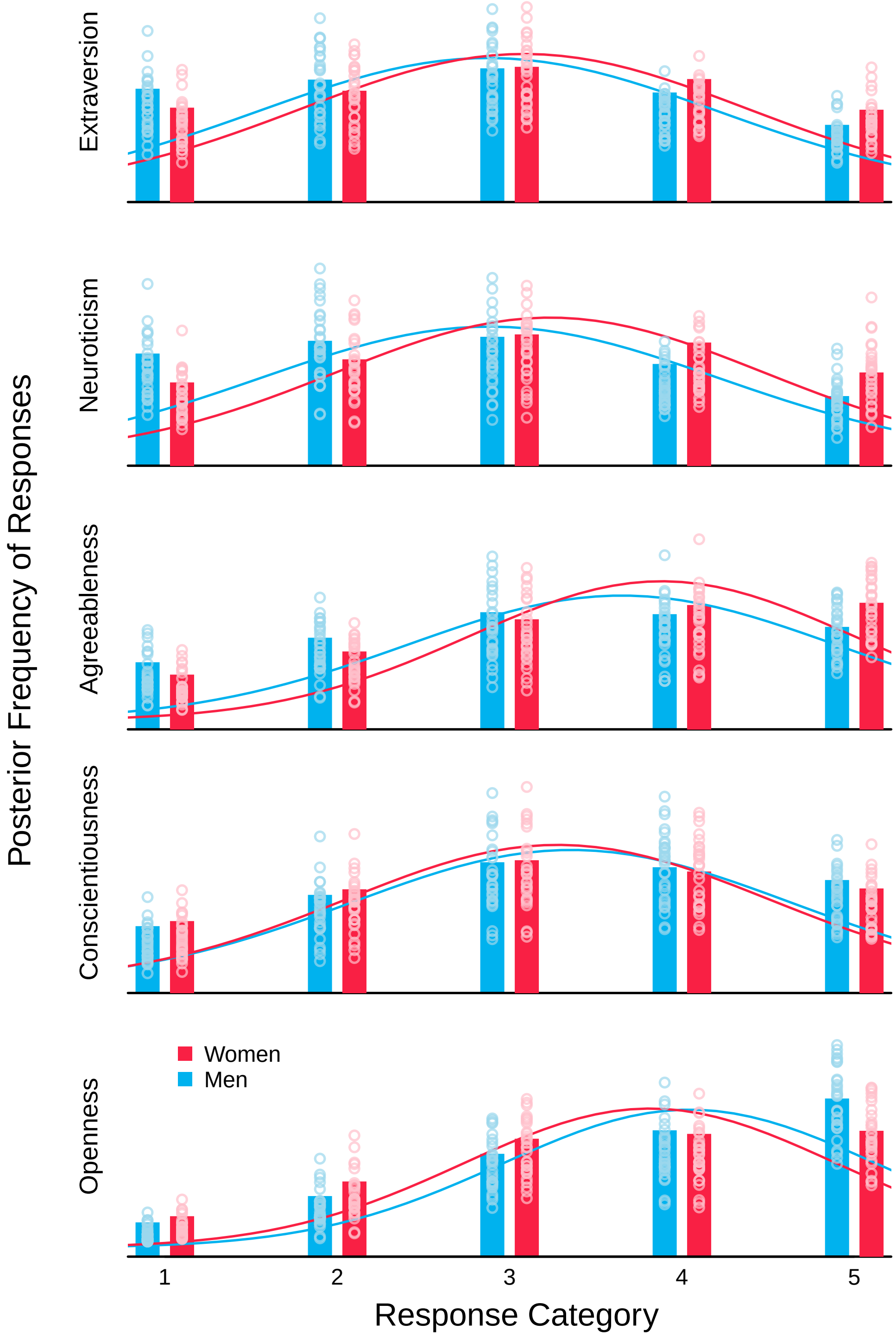

Figure 2

Posterior distribution of frequency (or probability) of the response categories for each personality trait. Bars represent mean frequency, circles are 30 random draws from the posterior distribution to visualize uncertainty. Normal distributions represent metric model. Separate for men and women.

References

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment*, *4*(1), 26.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/https://doi.org/10.1016/j.jesp.2018.08.009

Weisberg, Y. J., DeYoung, C. G., & Hirsh, J. B. (2011). Gender differences in personality across the ten aspects of the big five. *Frontiers in Psychology*, *2*, 11757.