**Supplementary Appendix to
“Fine Water: A Blind Taste Test”**

**[NOT FOR PUBLICATION]**

This appendix reports additional results for the three experiments discussed in “Fine Water: A Blind Taste Test.”

**A.1. Additional Results for Study 1 (Sensory Discrimination Experiment)**

*A.1.1 No significant differences between genders in the number of correctly identified singletons*

Tasting abilities might differ by gender, nationality, or other dimensions, so the regression results presented in our paper and in this appendix control for subject-specific fixed effects. For other results that lack such controls, we examined whether the results differed along those dimensions. Examining that for the first experiment, we found the average (standard deviation) of the number of singletons correctly identified by women and men were 1.9 (1.1) and 1.6 (0.9), respectively, although that difference was not statistically significant (t=1.17, *p*-value=0.24) and it does not seem substantively significant, either. Women were therefore perhaps slightly better than men at correctly identifying the singletons, but we did not find a significant difference based on gender.

*A.1.2. No significant differences based on the position of a singleton*

To explore why some of the bottled waters were easier to distinguish in the triangle tests, we ran the binary logit regressions reported in Appendix Table A. For each regression, the dependent variable was whether a given subject correctly identified the singleton in a given triangle test. Fixed effects for each subject were included to control for the fact that some participants seemed to be better than others at distinguishing the singletons. Dummy variables for the position of the singleton in a participant’s triangle test were also included to see whether the position—on the left, the right, or in the middle in a row—biased the results in any way. As seen in the table, there is weak evidence that a singleton was more likely to be correctly identified if it was in the middle, but none of the positional dummies were statistically significant—either individually or jointly—at conventional levels in any of our regressions, which suggests that the position of the singleton did not significantly affect its likelihood of being correctly identified.

The first regression in Table A also includes dummy variables for each triangle test. The regression suggests that the singleton in the second, third, and fourth triangle tests were more likely to be correctly identified than the one in the first triangle test, although only statistically significantly so at the 10% level for the second triangle test. Those results are consistent with the ones reported in Table 3 of our paper.

The waters in the triangle test differed in terms of their TDS levels and the amounts of minerals that make up the total amount of dissolved solids. The second regression in Table A therefore includes the absolute value of the difference between the TDS level of the singleton, on the one hand, and the TDS level of the twin, on the other hand. The slight negative correlation between that TDS difference and the likelihood of correctly identifying the singleton can be attributed to our subjects doing somewhat worse on the last triangle test with the biggest difference in TDS levels. The poor performance on the last test could reflect fatigue, perhaps, but as noted in the text, the results of Dietrich and Gallagher (2013) suggest there may not be a monotonic relationship between TDS differences and the percentage of subjects who correctly identify a singleton. Without randomizing over the order of our four triangle tests, no clear conclusions can be drawn.

 The final regression in Table A includes the differences in the sodium levels and the hardness of the water. We see the singleton is more likely to be correctly identified when there is a bigger sodium difference. Thus, Fiji water may have been easier to distinguish from either Acqua Panna or Hildon because of its relatively higher sodium level.

The hardness differences does not have a statistically significant relationship, but again, that finding can be attributed to the relatively poor performance on the last triangle test.

*A.1.3. More details on similar results from previous triangle tests with wine*

The text of our paper claims that the results of our triangle tests with water are similar to results of previous triangle tests with wine because Ashton (2014) and Weil (2001, 2005, 2007) found that tasters generally identified the singleton no more than around half of the time. We can be elaborate upon that claim as follows. Out of six triangle tests conducted by Weil (2001) with different vintages of otherwise similar wines, the singleton was correctly identified as frequently as 55% of the time (when comparing a 1991 and 1994 Bordeaux) and as infrequently as 32% of the time (when comparing a 1996 and 1997 Tuscan wine; Weil 2001, Table 3). Likewise, out of 15 triangle tests conducted by Weil (2005) with reserve and regular bottlings of otherwise similar wines, the singleton was identified as frequently as 55% of the time (with Robert Mondavi Cabernet Sauvignons) and as infrequently as 23% of the time (with Andrew Will Cabernets; Weil 2005, Table 2). And out of 10 tests conducted by Weil (2007) with wines that were similar except for the tasting descriptions given to them by experts, the singleton was identified as frequently as 100% of the time (when comparing a Chablis and a Sauvignon Blanc) and infrequently as 36% of the time (when comparing a premier cru Burgundy wine and a village-named wine made in the same year by the same producer; Weil 2007, Table 1; Authors’ calculations). Finally, out of two triangle tests conducted by Ashton (2014) with wines that were similar except some were from California and others were from New Jersey, the singleton was correctly identified 33% of the time in both tests (Ashton 2014, p. 312).

*A.1.4. More details on negative correlation between self-reported and observed ability to distinguish bottled waters*

As discussed in the text, there is a statistically significant negative correlation between subjects’ self-reported ability to distinguish bottled waters (as reported in preliminary demographic questions) and their observed ability (as measured by the number of singletons they correctly identified in the triangle tests). That negative correlation is driven by the fact that, for those who self-reported an above-average, average, or below-average ability to distinguish different bottled waters, the average number of singletons they correctly identified was about 1.5, 1.7, and 2.0, respectively. That negative correlation was somewhat stronger among the women in our study (⍴=−0.24, p-value=0.04) than the men (⍴=−0.06, p-value=0.73).

*A.1.5. More details on preferences expressed by confident discerners*

Our paper informally argues that, for each triangle test, the confident discerners expressed little or no preference between the bottled waters they tasted. More formal tests yield the same conclusion. According to the nonparametric sign test discussed by Olkin et al. (2015, pp. 23–24), the only preference difference that is statistically significant at conventional levels is the preference for Acqua Panna over Fiji (C=6.25, *p*-value=0.01). Even in that case, only slightly more than half of the confident discerners (51%) said they preferred or strongly preferred Acqua Panna, while 20% preferred or strongly preferred Fiji, and the rest were indifferent.

**A.2. Additional Results for Study 2 (Preference Rating Experiment)**

*A.2.1. More details on the lack of agreement in subjects’ rankings*

Kendall’s coefficient of concordance, which is discussed by Olkin et al. (2016, pp. 18–19), provides another measure of the agreement (or lack thereof) among our subjects’ rankings of the five waters served to them in the second experiment. That coefficient would be unity if all the subjects ranked the waters in the same way, and it would approach zero if they do not. For the rankings by our subjects, the coefficient is essentially zero (W=0.03).

*A.2.2. No significant differences between French and everyone else in rating tap water*

The French subjects in our study could potentially rate the local tap water we served them differently than non-French subjects, so we examined whether that was the case. We found the average (standard deviation) of the ratings given to tap water by subjects who identified French as a nationality and subjects who did not was about 9.2 (3.7) and 8.9 (3.4), respectively. That difference was not statistically significant, however (t=0.35, *p*-value=0.73). The size of that difference is also less than half a point on the 14-point scale.

**A.3. Additional Results for Study 3 (Description Matching Experiment)**

*A.3.1. More details on correlations between self-reported and observed abilities*

Unlike in the first experiment in which there was a negative correlation between self-reported and observed abilities to distinguish different bottled waters, there was a slight positive correlation between our subjects’ self-reported ability to distinguish tap from bottled water, on the one hand, and whether they correctly identified the tap water in the third experiment, on the other hand (⍴=0.11, *p*-value=0.23). Our subjects therefore seemed to be somewhat better at evaluating their ability to distinguish tap from bottled water.

*A.3.2. No correlation between performance in first and third experiments*

There was also no correlation between the number of singletons that a subject correctly identified in the first experiment and the number of descriptions they correctly matched in the third experiment (⍴=0.04, *p*-value=0.70).

*Appendix Table A*

**Binary Logit Results for Correctly Identified Singletons**

|  |  |  |  |
| --- | --- | --- | --- |
| *Dependent variable:* Correct?  | *(I)* | *(II)* | *(III)* |
| *Independent variables* | *Marginal effect* | *(Std. error)* | *Marginal effect* | *(Std. error)* | *Marginal effect* | *(Std. error)* |
| Triangle tests |  |  |  |  |  |  |
|  Second test | 0.13\* | (0.08) |  |  |  |  |
|  Third test | 0.12 | (0.07) |  |  |  |  |
|  Last test | 0.05 | (0.07) |  |  |  |  |
| Water-related variables |  |  |  |  |  |  |
|  TDS difference |  |  | −4E-4 | (0.00) |  |  |
|  Sodium difference |  |  |  |  | 0.01\* | (0.01) |
|  Hardness difference |  |  |  |  | 1E-4 | (0.00) |
| Singleton position |  |  |  |  |  |  |
|  First position | −0.05 | (0.09) | −0.06 | (0.09) | −0.05 | (0.09) |
|  Last position | −0.06 | (0.07) | −0.05 | (0.08) | −0.05 | (0.07) |
| Subject fixed effects | Yes |  | Yes |  | Yes |  |
| Observations | 404 |  | 404 |  | 404 |  |
| Degrees of freedom | 299 |  | 301 |  | 300 |  |
| −Log-likelihood | 244 |  | 246 |  | 244 |  |
| Pseudo *R*-squared | 0.25 |  | 0.24 |  | 0.25 |  |

*Note:* This table reports the results of binary logit regressions for whether a subject correctly identified the singleton in a triangle test. Fixed effects were used for each subject, and standard errors were clustered by subject. The number of observations corresponds to only 101 subject because 5 subjects were correct and 12 incorrect for all four triangle tests; the observations associated with those participants would be perfectly predicted by the corresponding fixed effects. The differences in mineral content between the singleton and twin were all in absolute value. The two positional dummies are not jointly significant for the first (χ2=0.66, *p*-value=0.72), second (χ2=0.52, *p*-value=0.77), or third (χ2=0.52, *p*-value=0.77) regression. The pseudo R-squared is McFadden’s.

\* p<0.10, \*\* p<0.05, \*\*\* p<0.01