BOTTLED WATER VERSUS TAP:

DOES A LABEL MATTER

**Abstract**

In April of 2016, a group from our statistics class performed an experiment to analyze the tastes of bottled and tap water and to see whether or not labeling of these waters affected their perceived taste. The sample was drawn from 38 students who volunteered to be a part of the experiment. These students were told to sample four different cups of water in a randomized order assigned to them. Two of the cups were incorrectly labeled (bottled water labeled as tap water and tap water labeled as pure water) and two were correctly labeled (bottled water labeled as pure water, and tap water labeled as tap water). Students then ranked these samples of water in order of prefered taste. After collecting and analyzing the data, it was found that students did indeed rank the taste of water higher if the water sample was labeled “pure”.

**Background and Significance**

It is common to hear students complain about the taste of the water on campus. As a group, we were interested to see if these complaints were founded on facts, or if students actually noticed a difference between the water on campus and bottled water. Our experiment combined the placebo effect with a blind taste test to yield results.

An article from Harvard Magazine discusses the effects of what is known as the “placebo phenomenon.” In other words, indicators such as care or mental capacities can influence the effect of certain drugs or procedures (Feinberg). Ted Kaptchuk of Harvard Medical School says that while placebos will not shrink tumors or eradicate viruses, placebos “can stimulate real physiological responses, from changes in heart rate and blood pressure to chemical activity in the brain, in cases involving pain, depression, anxiety, fatigue, and even some symptoms of Parkinson’s” (Feinberg). A similar article from Boston University discusses the trend of bottled water. Students at Boston University conducted a survey to test the preference of individuals when given various kinds of bottled water and a sample of tap water. They found that a large percentage of the student population could not taste the difference between bottled and tap water (Friday).

In our experimental design, we used labels as our placebos. Using the premise found by the study at Boston University, we decided to test whether or not labeling affects ranking of water taste given the fact that most individuals cannot tell the difference between bottled and tap water. We labeled bottled water in cups as tap water and also labeled similar cups of tap water as bottled water. These falsely labeled cups were placed next to cups of properly labeled bottled and tap water. With this design, we hoped to determine the effect that specific labeling had on taste perception for various individuals. Our hypothesis was that individuals would rank the two waters labeled ‘bottled’ above the water cups labeled ‘tap,’ demonstrating that perception of taste was determined by the label on the cup rather than the actual taste of the water.

**Methods**

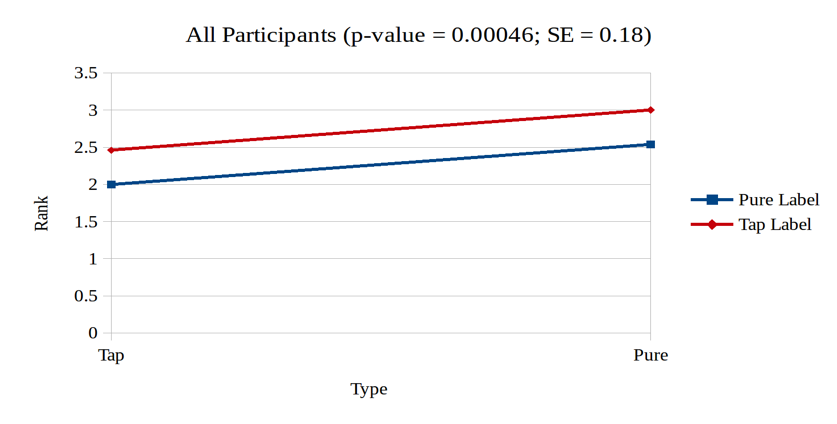
The experiment was designed to see if labeling a sample of water would have a significant effect on people’s perception of its quality of taste. It was a 2x2 full factorial design with the label of the sample and the actual type of water sampled as the variables. Each was a binary categorical variable where the label was either “pure” or “tap” and the actual type of water was either from a single water fountain on the campus (tap) or was a particular brand purchased from a grocery store (pure). This means that altogether there were the following four samples which we will refer to as combinations: tap water labeled “tap”, tap water labeled “pure”, pure water labeled “tap”, and pure water labeled “pure”.

We had volunteers first state whether they prefered bottled water or tap water[[1]](#footnote-1), and then we asked them to taste each of the four combinations and rank them from 1 (best) to 4 (worst) so that each sample was given a distinct rank. In this way, no participant was allowed to give two different samples the same rank. Samples were lined up by combination and each combination was given a letter ‘A’ through ‘D’ so that we could easily keep track of ranking. Participants were told that we were trying to rate four different water sources--two bottled water brands and water from two sources on campus.

In order to prevent the possibility of confounding due to the order in which participants sampled the four combinations, we used randomization so that each of the first 24 participants would sample the combinations in a unique order, then the next 24 would each have a unique order and so on. In conducting the experiment in this manner, it would be possible to block on the sampling order if there were enough participants. Overall, 38 people participated in the experiment, so blocking was not a viable option.

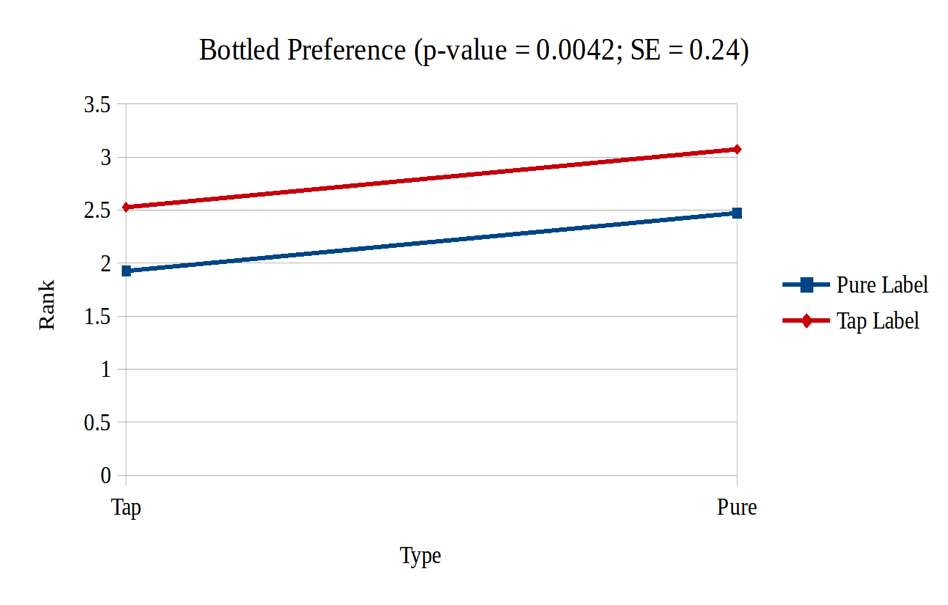
To begin, an analysis of variance (AOV) test was performed on three variables and their interactions: label (“pure” or “tap”), type (bottled or tap), and stated preference (bottled, campus tap, filtered, or no preference). The interaction term associated with stated preference and type showed borderline significance (p = 0.07) in the initial AOV test. Because of this we chose to proceed with the analysis by trying to generate separate models for each stated preferences. This served as a compromise to either leaving it in the model as a factor or neglecting it altogether. Proceeding, we separated the collected data into four groups--one for each of the stated preference options[[2]](#footnote-2). For each of these groups we performed an AOV test on the label and type variables.

**Results**

As mentioned, the stated preference was borderline significant in the initial AOV test, so in order to generate a linear model for the whole sample set (one that should be indicative of the campus as a whole) the stated preference was removed as a variable in that model. This left the label and the type, but not the interaction between them, as significant. This model (Figure 1) shows that regardless of the type of water sampled, the mean rank decreased by 0.46 when labeled as pure as opposed to being labeled as tap. Because samples that participants preferred

were given ranks that were lower in magnitude, this means that participants tended to prefer both types of water more when they believed that they were pure.

Of the four groups of potential models, two were shown to be statistically significant: the bottled water preference group and the filtered water preference group. For the bottled water preference group (n = 20) the results were very similar to the model for the whole sample set. This is not totally surprising because more than half of the participants in the sample set were in the bottled water preference group, thus the data in this group weighs heavily in the model for the whole sample. In the model for bottled water preference (Figure 2) it can be seen that participants’ perception of the taste of the water improved with the pure label just as in the previous model. However, the result is stronger here with a rank decrease of 0.6, showing again that people tended to prefer a sample more if it was labeled as pure.



The other group that resulted in a model that was statistically significant was the filtered water preference group (n = 2). It should be noted that the small sample size for this group means that we should be cautious not to extrapolate these results to all those who state a preference for filtered water. To emphasize this, no chart will be given here showing that model. Nonetheless, the fact that within this group the type of water was significant while the label was not is interesting and could be explored further[[3]](#footnote-3).

Because none of the variables were shown to be significant based on an AOV test for both the tap water preference group (n = 5) and the no preference group (n = 5), there is not enough evidence to reject the null model, and thus no charts for these data will be presented here. This is to say that we cannot rule out the possibility that for these two groups of participants, neither the label nor the type of water affected their preference. However, as with the filtered water preference group, the sample size is quite small, so it may not be representative of all those who state a preference to tap water or who state no preference. This is another area for possible future study.

Further exploratory analysis was done in order to determine if statistically significant results could be found by regrouping participants. One attempt was to put everyone who did not state a preference for bottled water into a group, but this did not give significant results. A second attempt was to put those who stated a preference for filtered water, and those that stated no preference into a group representing those that did not respond with one of the two preferences we were looking for. However, this did not yield significant results either.

**Discussion and Conclusions**

The project had potential limitations in some areas of the experimental protocols. The tap water sourced for the project came from a nearby drinking fountain that possessed a filtering system. This system could have made the tap water better tasting to the subjects. In addition to this, the drinking fountain, or tap, water was colder in temperature. There is potential that the subjects found colder water more refreshing and therefore more tasteful. But while this limits any conclusions we can draw about whether the campus actually prefers tap or bottled water, it does not limit conclusions about the effects of labeling. This is due to the full factorial nature of the experiment and the results obtained. Analysis showed that taste perception increased regardless of the water type sampled, thus it did not matter that the campus tap water was filtered nor did it matter that it was colder.

The labeling of the cups could have also been a limitation. In this experiment, cups were labeled as ‘pure’ rather than ‘bottled’. Subjects could perceive these labels in different ways, providing a potentially different taste rating in the experiment.

Because the experiment was a full factorial design, there were 24 different combinations of tasting order for the participants. We had a total of 38 participants to taste the water, and we therefore could not very effectively block for tasting order. There were only one or two participants per tasting order, limiting our ability to block for this variable. Furthermore, one participant’s responses had to removed due to incorrect tasting procedure.

This pilot study provided evidence that labels had a significant impact upon a person’s perceived taste rating of water. We can apply these findings in a wide variety of areas, particularly that of product marketing. Marketing strategies and brand development have always been of utmost importance when selling or promoting a product, and here we reaffirm how significant this area of business is. Labeling is an important factor for consumers when they consider taste of a product.

Additionally, we saw participants who possessed a preference for filtered water rank pure water better tasting than tap water regardless of the label. There is opportunity here for further study to determine perhaps what causes the filtered water preference group to rank taste regardless of the label provided.

**References**

Feinberg, Cara. "The Placebo Phenomenon." *Harvard Magazine*. Harvard Magazine, Jan.-Feb. 2013. Web. 02 May 2016.

Friday, Leslie. "Bottled vs. Tap: Which Tastes Better?" *BU Today*. Boston University, 24 Mar. 2011. Web. 02 May 2016.

**Appendix**

In order to randomize the drinking order for the participants, the following python script was used. The script generates a list of orderings for up to 240 participants (each printed four times which is an implementation detail). It does so by randomly ordering the 24 possible permutations 10 times. In this way, at any given time in the experiment there are no two orderings that differ in their number of uses by more than one. This serves also as a type of blocking that could have been used if the sample size was large enough.

import random

import itertools

#get a randomly ordered list of all possible permutations of base

def getBlock(base):

perms = [x for x in itertools.permutations(base)]

random.shuffle(perms)

return perms

#print out 4 times each a random permutation of "ABCD"

def main():

base = 'ABCD'

block = getBlock(base)

for x in block:

string = ""

for e in x:

string = string + e

for i in range(4):

print string

#run main 10 times (enough for 240 people)

for i in range(10):

main()

This code generated the list that can be seen in the 'water' variable in the included dataset. However, the ninth sequence has been removed from the dataset because the ninth volunteer accidentally drank one sample type twice and another type not at all.

The R code below shows the script that was run to generate and test various models with the data collected.

#Water Preference Analysis

water = read.csv(file.choose()) ### Water Experiment Data Complete.csv

colnames(water)

[colnames(water)=="Do.you.generally.prefer.campus.water.or.bottled.water"] = "Q"

usable = water[water$Rank!="x",­1]

bottled = usable[usable$Q=="bottled",]

campus = usable[usable$Q=="campus",]

filtered = usable[usable$Q=="filtered",]

nopref = usable[usable$Q=="nopref",]

fil\_nop = usable[usable$Q %in% c("filtered", "nopref"),]

not\_bottled = usable[usable$Q %in% c("filtered", "nopref", "campus"),]

#Some reference values

Q\_sizes = as.table(cbind(c(length(bottled[,1]), length(campus[,1]),

length(filtered[,1]), length(nopref[,1]))/4))

colnames(Q\_sizes) = c("number of people with stated preference")

rownames(Q\_sizes) = c("bottled", "campus", "filtered", "no pref")

sample\_by\_rank = as.table(cbind(c(length(usable[usable$Code=="A" & usable$Rank==1,]

[,1]),length(usable[usable$Code=="B" & usable$Rank==1,]

[,1]),length(usable[usable$Code=="C" & usable$Rank==1,]

[,1]),length(usable[usable$Code=="D" & usable$Rank==1,][,1]))))

colnames(sample\_by\_rank) = c("# times ranked 1st")

bottled\_by\_rank = as.table(cbind(c(length(bottled[bottled$Code=="A" &

bottled$Rank==1,][,1]),length(bottled[bottled$Code=="B" & bottled$Rank==1,]

[,1]),length(bottled[bottled$Code=="C" & bottled$Rank==1,]

[,1]),length(bottled[bottled$Code=="D" & bottled$Rank==1,][,1]))))

colnames(bottled\_by\_rank) = c("# times ranked 1st by those who stated bottled

preference")

campus\_by\_rank = as.table(cbind(c(length(campus[campus$Code=="A" & campus$Rank==1,]

[,1]),length(campus[campus$Code=="B" & campus$Rank==1,]

[,1]),length(campus[campus$Code=="C" & campus$Rank==1,]

[,1]),length(campus[campus$Code=="D" & campus$Rank==1,][,1]))))

colnames(campus\_by\_rank) = c("# times ranked 1st by those who stated campus

preference")

filtered\_by\_rank = as.table(cbind(c(length(filtered[filtered$Code=="A" &

filtered$Rank==1,][,1]),length(filtered[filtered$Code=="B" & filtered$Rank==1,]

[,1]),length(filtered[filtered$Code=="C" & filtered$Rank==1,]

[,1]),length(filtered[filtered$Code=="D" & filtered$Rank==1,][,1]))))

colnames(filtered\_by\_rank) = c("# times ranked 1st by those who stated filtered

preference")

nopref\_by\_rank = as.table(cbind(c(length(nopref[nopref$Code=="A" & nopref$Rank==1,]

[,1]),length(nopref[nopref$Code=="B" & nopref$Rank==1,]

[,1]),length(nopref[nopref$Code=="C" & nopref$Rank==1,]

[,1]),length(nopref[nopref$Code=="D" & nopref$Rank==1,][,1]))))

colnames(nopref\_by\_rank) = c("# times ranked 1st by those who stated no preference")

writeLines("­­­­­Some reference data­­­­­")

cat("\n")

cat(c("n =", length(usable[,1])/4))

cat("\n\n")

print(Q\_sizes)

cat("\n")

print(sample\_by\_rank)

cat("\n")

print(bottled\_by\_rank)

cat("\n")

print(campus\_by\_rank)

cat("\n")

print(filtered\_by\_rank)

cat("\n")

print(nopref\_by\_rank)

#Models for the entire sample set

#see if a model including labeling, water type, and stated water preference

#with full interaction terms is significant

writeLines("\n\n\n­­­­­Full Interactive Model:­­­­­")

print(summary(aov(Rank~Label\*Actual\*factor(Q), data=usable)))

#stated preference was not significant, so remove it in the next model

writeLines("\n­­­­­Interactive Model without Q:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=usable)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Model without Q:­­­­­")

print(summary(aov(Rank~Label+Actual, data=usable)))

#all terms in the previous model are significant, so use that as the model

writeLines("\n­­­­­Final Linear Model For Total Sample:­­­­­")

print(summary(lm(Rank~Label+Actual, data=usable)))

#Models for just those who stated a preference for bottled water

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive Bottle Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=bottled)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Bottle Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=bottled)))

#all terms in the previous model are significant, so use that as the model

writeLines("\n­­­­­Final Linear Bottle Model:­­­­­")

print(summary(lm(Rank~Label+Actual, data=bottled)))

#Models for just those who stated a preference for campus water

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive Campus Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=campus)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Campus Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=campus)))

#all terms in the previous model are insignificant, so there is no evidence against

the null model

writeLines("\n­­­­­Final Linear Campus Model:­­­­­")

writeLines("there is no evidence against the null model")

#Models for just those who stated a preference for filtered water

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive Filtered Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=filtered)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Filtered Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=filtered)))

#the 'label' term was not significant, so remove it in the next model

writeLines("\n­­­­­Single Term Filtered Model:­­­­­")

print(summary(aov(Rank~Actual, data=filtered)))

#all terms in the previous model are at least borderline significant, so use that in

the model

writeLines("\n­­­­­Final Linear Filtered Model:­­­­­")

print(summary(lm(Rank~Actual, data=filtered)))

#Models for just those who stated no water preference

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive No Preference Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=nopref)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative No Preference Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=nopref)))

#all terms in the previous model are insignificant, so there is no evidence against

the null model

writeLines("\n­­­­­Final Linear No Preference Model:­­­­­")

writeLines("there is no evidence against the null model")

#Models for those who stated no water preference or filtered water preference

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive Nop\_Fil Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=fil\_nop)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Nop\_Fil Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=fil\_nop)))

#all terms in the previous model are insignificant, so there is no evidence against

the null model

writeLines("\n­­­­­Final Linear Nop\_Fil Model:­­­­­")

writeLines("there is no evidence against the null model")

#Models for not bottled

#see if interactive model is significant

writeLines("\n\n\n­­­­­Interactive Not\_Bottled Model:­­­­­")

print(summary(aov(Rank~Label\*Actual, data=not\_bottled)))

#the interaction term was not significant, so remove it in the next model

writeLines("\n­­­­­Addative Not\_Bottled Model:­­­­­")

print(summary(aov(Rank~Label+Actual, data=not\_bottled)))

#all terms in the previous model are insignificant, so there is no evidence against

the null model

writeLines("\n­­­­­Final Linear Not\_Bottled Model:­­­­­")

writeLines("there is no evidence against the null model")

The tables printed in the reference values section were used to construct plots to see if the frequency with which each taste sample was ranked first provided a different model than the average rank of each sample. However, the results were very similar in nature to using the mean rank of each sample, and because it is easier to create models based on the means, that is the route that was taken.

1. Some participants struggled to answer this question by either posing that they had no preference or by stating that they filtered their own water, so we allowed for these two responses to be acceptable. See the limitations section for more information. [↑](#footnote-ref-1)
2. There were a five participants from whom we did not collect data regarding stated preference, and thus they were not included in any of the four stated preference groups. See the limitations section for more information. [↑](#footnote-ref-2)
3. See the implications section for further details. [↑](#footnote-ref-3)