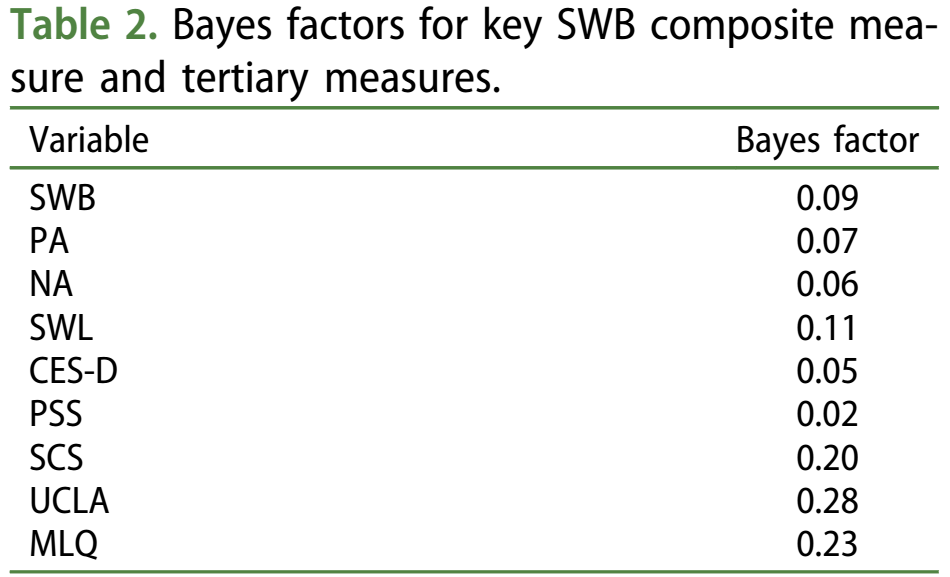
**Examining non-significant results with Bayes factors and equivalence tests.**

*In this blog, I’ll compare two ways of interpreting non-significant effects: Bayes factors and TOST equivalence tests. I’ll explain why reporting more than only Bayes factors makes sense, and highlight some benefits of equivalence testing over Bayes factors. I’d like to say a big thank you to Bill (Lihan) Chen and Victoria Savalei for helping me out super-quickly with my questions as I was re-analyzing their data.*

Does volunteering improve well being? A recent [article](http://sci-hub.bz/10.1080/23743603.2016.1273647) by Ashley Whillans, Scott Seider, Lihan Chen, Ryan Dwyer, Sarah Novick, Kathryn Gramigna, Brittany Mitchell, Victoria Savalei, Sally Dickerson & Elizabeth W. Dunn suggests the answer is: Not so much. The study was published in [Comprehensive Results in Social Psychology](http://explore.tandfonline.com/cfp/beh/comprehensive-results-in-social-psychology), one of the highest quality journals in social psychology, which peer-reviews pre-registrations of studies before they are performed.

People were randomly assigned to a volunteering program for 6 months, or to a control condition. Before and after, a wide range of well-being measures were collected. Bayes factors support the null for all measures. The main results (and indeed, except for some manipulation checks, the only results – not even means, or standard deviations) are communicated in the form of Bayes factors in Table 2.



The Bayes factors were calculated using the [Bayes factor calculator](http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm) by Zoltan Dienes, who has a great open access [paper in Frontiers](http://journal.frontiersin.org/article/10.3389/fpsyg.2014.00781/full), cited more than 200 times since 2014, on how to use Bayes to get most out of non-significant results. I won’t try to explain in detail how these Bayes factors are calculated – too many Bayesians on Twitter have told me I am too stupid to understand the math behind Bayes factors, and how I should have taken calculus in high school. They are right on both accounts, so just read Dienes (2014) for an explanation.

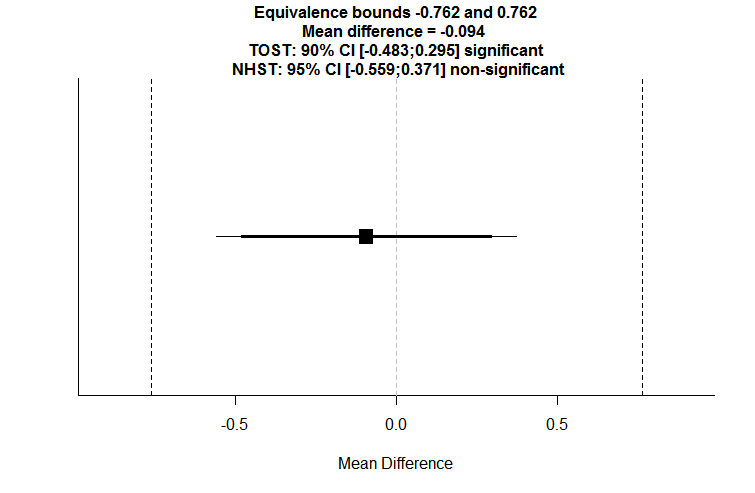
But even though I am not smart enough to understand the math behind Bayes factors, I know that the Bayes factor is a number based on a comparison of two models, one for the null hypothesis and one for the alternative hypothesis. Bayes factors will tell you whether the data is more likely under one model or the other, and we can interpret Bayes factors as relative evidence. For example, you might see a platypus, and quantify whether it is more likely to be born from a family of beavers, than from a family of ducks. In this case, neither model is true, but depending on whether you look at the platypus from the front or the back, you might judge one data-generating model more likely than the other, and calculate a Bayes factor that tells you how much you should rationally update your belief in which family birthed a platypus. As you can see from this example, the better your models are, the more useful Bayes factors are (and vice versa).

As Dienes (2014) discusses, you can *also* interpret non-significant results using Frequentist statistics, In a TOST equivalence test, which consists of two simple one-sided *t*-tests, you determine whether an effect falls between equivalence bounds set to the smallest effect size you care about (for an introduction, see [Lakens, 2017](https://osf.io/preprints/psyarxiv/97gpc/)). Dienes (2014) says it can be difficult to determine what this smallest effect size of interest is, but for me, if anything, it is easier to determine a smallest effect size of interest than to specify an alternative model in Bayesian statistics.

The authors examined whether well-being was improved by volunteering, and specified an alternative model (what would a true effect of improved well-being look like?) as follows (page 9): “*Because our goal was to contrast the null hypothesis to an alternative hypothesis that the effect is moderate in size, we used a normal distribution prior with a mean of 0.50 and a standard deviation of 0.15 for the standardized effect size (e.g. the difference score between standardized T2 and T1 measures).*”

It is interesting to see the authors wanted to specify their alternative in terms of a ‘standardized effect size’. I fully agree that using standardized effect sizes is currently the easiest way to think about the alternative hypothesis, and it is the reason my spreadsheet and R package “TOSTER” allow you to specify equivalence bounds in standardized effect sizes when performing an equivalence test.

In equivalence testing, we can test whether the observed data is surprisingly smaller than anything we would expect, if the true effect size was *d* = 0.5. We can just fill in the means, standard deviations, and sample sizes from both groups, and test against the equivalence bound of *d* = 0.5 (see the code at the bottom of the post). Note that the authors perform a two-sided test (even though they have a one-sided hypothesis, as indicated in the title “Does volunteering improve well-being?”, but following the authors, I will test whether the effect is statistically smaller than *d* = 0.5, *and* statistically larger than *d* = -0.5, instead of only testing whether the effect is smaller than *d* = 0.5. The most important results are summarized in the Figure below:



Testing the effect for WSB, one of the well-being measures, we could write: Using a TOST procedure to test the data against equivalence bounds of *d* = -0.5 and *d* = 0.5, the observed results were statistically equivalent to zero, *t*(78.24) = -2.86, *p* = 0.003. The mean difference was -0.094, 90% CI[-0.483; 0.295].

The standardized effect size of 0.5 equals a raw effect of 0.762 in scale points on the original measure. Because the 90% confidence interval around the mean difference does not contain -0.762 or 0.762, the observed data is surprising (a.k.a statistically significant), if there was a true effect of *d* = -0.5 or *d* = 0.5 (see Lakens, 2017, for a detailed explanation). We can reject the hypothesis that *d* = -0.5 or *d* = 0.5, and if we do this, given our alpha of 0.05, we would be wrong a maximum of 5% of the time, in the long run.

**Benefits of equivalence tests compared to Bayes factors.**

If we perform equivalence tests, we see that we can conclude statistical equivalence for all nine measures. You might wonder about whether we need to correct for the fact that we perform nine tests for all the different well-being measures. Would we conclude that volunteering has a positive effect on well-being, if any single one of these tests showed a significant effect? If so, we should indeed correct for multiple comparisons to control our overall Type 1 error rate, and you can do this in equivalence testing. There is no easy way to control error rates in Bayesian statistics. Some Bayesians simply don’t care about error control, and I don’t exactly know what Bayesian who care about error control do. I care about error control, and the attention *p*-hacking is getting suggests I am not alone. In equivalence testing, you can control the Type 1 error rate simply by adjusting the alpha level, which is one benefit of equivalence testing over Bayes factors.

To calculate a Bayes factor, you need to specify your prior by providing the mean and standard deviation of the alternative. Bayes factors are quite sensitive to how you specify these priors, and for this reason, not every Bayesian statistician would recommend the use of Bayes factors. Andrew Gelman, a widely known Bayesian statistician, recently co-authored [a paper](http://rsos.royalsocietypublishing.org/content/4/1/160426) in which Bayes factors were used as one of three Bayesian approaches to re-analyze data. In footnote 3 it is written: “*Andrew Gelman wishes to state that he hates Bayes factors*” – mainly because of this sensitivity to priors. So not everyone likes Bayes factors (just like not everyone likes *p*-values!). You can discuss the sensitivity to priors in a *sensitivity analysis*, which would mean plotting Bayes factors for alternative models with a range of means and standard deviations and different distributions, but I rarely see this done in practice. Equivalence tests also depend on the choice of the equivalence bounds. But it is very easy to see the effect of different equivalence bounds on the test result – you can just check if the equivalence bound you would have chosen falls within the 90% confidence interval. So that is a second benefit of equivalence testing.

A final benefit I’d like to discuss concerns the assumptions of statistical tests. You should not perform tests when their assumptions are violated. The authors in the paper examining the effect of volunteering on well-being correctly report Welch’s *t*-tests, because they have unequal sample sizes in each group, and the equal variances assumption is violated. This is excellent practice. I don’t know how Bayes factors deal with unequal variances (I think they don’t, and simply assume equal variances, but I’m sure the answer will appear in the comments, if there is one). My TOST equivalence test spreadsheet and R code use [Welch’s *t*-test by default](http://daniellakens.blogspot.nl/2015/01/always-use-welchs-t-test-instead-of.html) (just as R does), so unequal variances is no longer a problem. The equal variances assumption is not very plausible in many research questions in psychology (Delacre, Lakens, & Leys, under review), so not having to assume equal variances is another benefit of equivalence testing compared to Bayes factors.

**Conclusion**

Only reporting Bayes factors seems, to me, an incomplete description of the data. To me, it makes sense to report an effect size, the mean difference, and the confidence interval around it. And if you do that, and have determined a smallest effect size of interest, then performing the TOST equivalence testing procedure is nothing more than checking and reporting whether the *p*-value for the TOST procedure is smaller than your alpha level to conclude the effect is statistically equivalent. And you can still add a Bayes factor, if you want.

All approaches to statistical inferences have strengths and weaknesses. In most situations, both Bayes factors and equivalence tests lead to conclusions that have the same practical consequences. Whenever they do not, it is never the case that one approach is correct, and one is wrong – the answers differ because the tests have different assumptions, and you will have to think about your data more, which is never a bad thing. In the end, as long as you share the data of your paper online, like the current authors did, anyone can calculate the statistics they like. But only reporting Bayes factors is not really enough to describe your data. You might want to at least report means and standard deviations, so that people who want to include the effect size in a meta-analysis don’t need to re-analyze your data. And you might want to try out equivalence tests next time you interpret null results.