The Delivery Matters: Examining Reactivity in Question Answering

Erin M. Buchanan1

Riley E. Foreman1

Jeffrey M. Pavlacic1

Rachel Swadley1

Stefan E. Schulenberg2

1Missouri State University

2University of Mississippi

Abstract

Scales that are psychometrically sound – meaning those that meet established standards regarding reliability and validity while measuring one or more constructs of interest – are customarily evaluated based on a set modality of delivery (i.e. via the Internet, handwritten, etc.) and administration (fixed item order). Deviating from an established administration profile could result in non-equivalent response patterns, indicating the possible evaluation of a dissimilar construct. Furthermore, item grouping may influence response patterns. Randomizing item administration may alter or eliminate these effects. Therefore, we examined the differences in scale relationships for computer versus handwritten scale delivery for two scales measuring meaning/purpose in life. These scales have questions about suicidality, depression, and life goals that may cause reactivity (i.e. a changed response to a second item based on the answer to the first item).

Keywords: *scales, reactivity, item answering*

The Delivery Matters: Examining Reactivity in Question Answering

* scale development
* scale comparisons
* previous work on scale comparisons

**Method**

**Participants**

The data analyzed for this study originated from two separate institutions. The sample population consisted of undergraduate students at either a large Midwestern or large Southern university, placing the approximate age of participants at around 18-22. Table XX includes the information about all datasets. Several datasets have been used previously for larger studies, and those citations are included in the table. Only two scales were used from each dataset, as described below. Participants were generally enrolled in an introductory psychology course that served as a general education requirement for the university. As part of the curriculum, the students were encouraged to participate in psychology research programs, resulting in their involvement in this study. These participants were given course credit for their participation.

**Talk about each sample separately**

**Materials**

Of the surveys included within each larger study, two questionnaires were utilized: the Purpose in Life Questionnaire (PIL; Crumbaugh & Maholick, 1964, 1969) and the Life Purpose Questionnaire (LPQ; Hablas & Hutzell, 1982; Hutzell, 1989).

The Purpose in Life Questionnaire. The PIL is a 20-item questionnaire that assesses perceived meaning and life purpose. Items are structured in a 7-point Likert type response format; however, each item has different anchoring points that focus on item content. Total scores are created by summing the items, resulting in a range of 20 to 140 for the overall score. The reliability for the scale is generally high, ranging from .70 to .90 (Schulenberg, 2004; Schulenberg & Melton, 2010). Previous work on validity for the PIL showed viable one- and two-factor models, albeit question loadings varied across publications (see Schulenberg & Melton, 2010 for a summary), and these fluctuating results lead to the development of a 4-item PIL short form (Schulenberg, Schnetzer, & Buchanan, 2012).

Life Purpose Questionnaire. The LPQ was modeled after the full 20-item PIL questionnaire, also measuring perceived meaning and purpose in life. The items are structured in a true/false response format, in contrast to the Likert response format found on the PIL. Each question is matched to the PIL with the same item content, altering the question to create binary answer format. A total score is created by summing questions, resulting in a range from 0 to 20. In both scales, higher scores indicated greater perceived meaning in life. Reliability for this scale is also correspondingly high, usually in the .80 range (Melton & Schulenberg, 2008; Schulenberg, 2004).

These two scales were selected because they contained the same item content with differing response formats, which would allow for cross comparisons between results for each scale.

**Procedure**

The form of administration was of interest to this study, and therefore, multiple formats were included: paper administration in a nonrandom question order, computerized administration in nonrandom order, and computerized administration with a randomized question order. Surveys completed on paper were administered in person in a classroom-type setting, and the questions were printed in a fixed order, consistent across all forms. This setting included a proctor to oversee the experiment and debrief the participants upon completion. Alternatively, computerized questionnaires were available for participants to access electronically, and they were allowed to complete the experiment without directly interacting with a proctor. To ensure participants were properly informed, both an introduction and a debriefing were included within the online form. One section of the computerized questionnaires followed the original scale question order that was never changed, consistent with the paper forms. A different group of participants were given each question in a randomized order. Once collected, the results were then amalgamated into a database for statistical analysis.

**Hypotheses and Data Analytic Plan**

Each hypothesis was tested using three dependent measures. The variance-covariance matrix for each type of delivery was estimated and compared to each other by using root mean squared error (RMSE; CITE). RMSE estimates the difference between covariance matrices and is often used in structural equation modeling to determine if models have good fit to the data. A criterion of < .06 for good fit, .06-.08 for acceptable fit, and > .10 for bad fit was used (Hu & Bentler, 1999?). This analysis was used to determine if the change in delivery changed the structure of the item relationships to each other. RMSE values were calculated using the *monomvn* package in *R* (Gramacy, CITE).

Next, item averages were calculated across all participants for each item. These 20 items were then compared in a matched dependent *t*-test (i.e. item 1 to item 1, item 2 to item 2) to determine if delivery changed the mean of the item. While covariance structure elucidates the varying relations between items, we may still find that item averages are pushed one direction or another by a change in delivery and still maintain the same correlation between items. If this test was significant, we examined the individual items across participants for large effect sizes, as large sample sizes would create significant *t*-test follow ups. Last, the total scores for each participant were compared across delivery type using an independent *t*-test. Item analyses allow a focus on specific items that may show changes, while total scores allow us to investigate if changes in delivery alter the overall score that is used in other analyses. In both the item and total score analyses, *d* values and Bayes Factors are provided to examine the size of effects for interpretation, instead of *p*-values that are biased by sample size.

Hypothesis 1. Paper forms were compared computerized non-random forms to examine the method of delivery on the relationships between items, item means, and total scores. We expected to find XX consistent with previous research by XX.

Hypothesis 2. Computer forms were then analyzed by randomized and nonrandomized groups to examine the impact of randomization on covariance structure, item means, and total scores. We expected to find that these forms would vary across covariance structure and item means, which would indicate differences in reactivity to questions (i.e. item 4 always has item 3 as a precursor on a nonrandom form, while item 4 may have a different set of answers when prefaced with other questions). We examined total scores; however, it was unclear if these values would change. A difference in item means may result in changes in total scores, but may also result in no change if some item means decrease, while others increase.

**Results**

**Data Screening**

Each dataset was analyzed separately by splitting on both scale, delivery, and randomization (i.e. PIL-Paper, PIL-Computer-Random, PIL-Computer-Nonrandom, etc.). First, all data was screened for accuracy and missing data. Participants with more than 5% missing data (i.e. 2 or more items) were excluded. For the PIL, we excluded 132 (*1P­, 83CR, 48CN*) and for the LPQ, we excluded 843 (*138P­, 555CR, 150CN*). Data was imputed using the *mice* package in *R* for participants with less than 5% of missing data. This procedure imputed 3,569 (*1,327P­, 1,378CR, 864CN*) participants on the PIL and 65 (*15P­, 15CR, 35CN*) participants on the LPQ. We examined each dataset for multivariate outliers using Mahalanobis distance, and 565 (*73P­, 455CR, 37CN*) PIL and 116 (*78P­, 23CR, 15CN*) LPQ participants were excluded for excessively high scores (i.e. *p* < .001, Tabachick & Fidell, 2012). Each dataset was then screened for multivariate assumptions of additivity, linearity, normality, homogeneity, and homoscedasticity.

**PIL Analyses**

Covariance Matrices Analysis. Covariance tables were examined to determine model fit for the different methods of deliverance. The covariance tables from paper forms and those questions delivered not randomly by a computer indicated poor model fit, RMSE = .39. Likewise, covariance tables from questions delivered not randomly on a computer fit poorly with those delivered randomly on a computer, RMSE = .55.

Item Means. Dependent *t*-tests were conducted on the item mean scores for each hypothesis. Questions delivered on paper were significantly different from those delivered not randomly on a computer, *t*(19) = -7.27, *p* < .001, *davg* = -1.62 These results indicated that paper scores (*M* = 5.45, *SD* = 0.46) have higher item averages than non-random computer scores (*M* = 5.15, *SD* = 0.49). However, questions delivered randomly on a computer were not different from those delivered not randomly on a computer, *t*(19) = 1.36, *p* = .184, *davg* = 0.31, indicating that computer questionnaires that presented the questions in a random format (*M* = 5.00, *SD* = 0.73) were similar to those presented in a non-random format (*M* = 5.15, *SD* = 0.49) on the computer. *d* effect sizes were calculated for each item between the paper delivery method and non-random computer delivery method to determine which items displayed the largest changes. Results are presented in Table XX.

Total Scores. Independent *t*-tests conducted on total scores revealed that questions delivered on paper were significantly different from those delivered not randomly on a computer, *t*(2079) = 8.31, *p* < .001, *d* = 0.37. On average, paper total scores (*M* = 108.96, *SD* = 14.14) are higher than the computer not random scores (*M* = 103.04, *SD* = 18.24). Questions delivered not randomly on a computer also were significantly higher from those delivered not randomly on a computer, *M* = 100.10, *SD* = 15.42, *t*(1747) = 3.66, *p* < .001, *d* = 0.17. However, very small effect sizes were observed for both analyses.

**LPQ Analyses**

Covariance Matrices Analysis. The covariance tables from paper forms and those questions delivered not randomly by a computer indicated excellent model fit, RMSE = .04. Likewise, covariance tables from questions delivered not randomly on a computer fit excellently with those delivered randomly on a computer, RMSE = .02. These RMSE values indicate that the covariance tables were nearly identical.

Item Means.Dependent *t*-tests were conducted on the mean scores showed that questionnaires delivered on paper were significantly different from those delivered not randomly on a computer, *t*(19) = -10.05, *p* < .001, *davg* = 0.62. However, questions delivered randomly on a computer were not different from those delivered not randomly on paper, *t*(19) = 0.22, *p* = .83, *davg* = 0.01. These results indicate that paper scores (*M* = 0.80, *SD* = 0.16) have higher item averages than non-random computer scores (*M* = 0.71, *SD* = 0.15), but computer questionnaires that presented the questions in a random format (*M* = 0.71, *SD* = 0.18) were similar to those presented in a non-random (*M* = 0.71, *SD* = 0.15) format on the computer. *d* effect sizes were again calculated for each item between the paper delivery method and non-random computer delivery method. Results are presented in Table XX. All items indicated either a small or negligible effect size. Specifically, items 1, 5, 7, 15, 18, and 19 indicated a negligible effect size. Items 2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 16, 17, and 20 indicated a small effect size. These results suggest that computer questions presented in a random were not different from those presented in a non-random format on the computer.

Total Scores. Questions delivered on paper were significantly different from those delivered not randomly on a computer, *t*(1795) = -11.05, *p* < .001, *d* = 0.53. On average, paper total scores (*M* = 16.09, *SD* = 3.10) are higher than the computer not random scores (*M* = 14.19, *SD* = 4.22). Questions delivered randomly on a computer were not different from those delivered not randomly on a computer, *t*(1631) = 0.19, *p* = .85, *d* = 0.01. Questionnaires delivered randomly on the computer (*M* = 14.15, *SD* = 4.00) and those not delivered randomly (*M* = 14.19, *SD* = 4.22) are statistically similar.

**Discussion**

Write this by hypothesis … need to figure out the bayes factors to make sure we can strengthen the it’s significant but not important.

* Rachel’s PIL talk: The covariance analyses indicated that the item relationships differed dependent upon the questionnaire delivery method. Additionally, the questionnaires delivered on paper had significantly different item scores than questionnaires delivered non-randomly on the computer. After investigating differences in scores for each item delivered on paper versus each item delivered non-randomly on the computer, items 10, 16, and 5 were found to have the largest effect size, although all effect sizes were small or negligible. These results indicate that the item differences created a cumulative effect leading to significant item difference scores between the paper delivery method and non-random computer delivery method. Total scores also appeared to differ between the paper delivery method and the non-random delivery method. While the item scores did not appear to significantly differ between the random and non-random computer delivery method, there were significant total score differences between these two methods. However, total score differences had small effect sizes for both the paper versus non-random computer delivery analysis and the random versus non-random computer delivery analysis.
* Jeff’s: Overall, we see that covariance structures between delivery types indicate excellent model fit, which indicates that they are similar. However, results indicated a large difference in item scores for paper questionnaires versus those delivered not randomly by computer for both mean scores and total scores. T-tests were conducted by item to examine possible differences in specific questions. Although no specific question was significantly different based on delivery method, questions appeared to be slightly different, which accumulated over each question. This resulted in the differing item scores for the questionnaires delivered by paper and those delivered not randomly on a computer. Bayes’ Factors and Tests of Equivalence (citation) were utilized to determine whether or not differences in question means were actually significant. Results are presented in Table XX. Specifically, for Tests of Equivalence, significant results suggests support for the null hypothesis. For Bayes’ Factor tests, a number > 100 suggests very strong support for the alternative hypothesis in reference in the null. A score of 1-3 suggests barely any evidence for the alternative hypothesis. A score of 3-10 suggests positive evidence for the alternative hypothesis, while a Bayes’ Factor score of 10-100 suggests strong support for the alternative hypothesis.

Also discuss the comparison between the sets of results.

Table 1.

*Sample Size and Demographic Information for Each Dataset*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PIL | | | | LPQ | | |  |  |
|  | *N* | *M* Age (*SD*) | % Female | % White | *M* Age (*SD*) | % Female (assuming female = 2) | % White (assuming white = 1) | Citation | Admin Type |
| MSU IRT Data | 2531 |  |  |  | 19.59 (*3.23)* | 55.16% (assuming 10 = female) | This one showed up weird in R so might need some help. Guess it was bc we had people enter it manually? |  | Computer |
| Week 1 file  Lindsay thesis | 269 | 19.14 | 75.75% | 76.03% | 19.14 (*2.05)* | 75.46% | 75.46% |  | Paper |
| MSU data | 201 | NA | NA | NA | 19.96 (*3.82)* | 53.73% | 68.66% |  | Computer |
| Meaning working file | 298 | 19.67 | 63.36% | 71.92% | 19.67 (*2.27)* | 62.08% | 70.47% |  | Paper |
| Logotherapy data factor analysis study | 341 | 19.50 | 66.17% | 83.57% | 19.50 (*1.60)* | 65.98% | 83.58% |  | Paper |
| Mike and mike | 265 | 18.84 | 43.40% | 79.92% | NA | NA | NA |  | Paper |

Table XX.

*PIL Item, Total Averages, and Effect Size of the Differences*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Item | Paper | Non-Random  Computer | Random  Computer | *d* Paper v  Non-Random | *BF* | *Tost* Paper v Non-Random |
| 1 | 4.97 (1.00) | 4.81 (1.28) | 4.83 (1.31) | -0.15 | 0.05639996 | *t*(2079) = -3.36,  *p* < .001 |
| 2 | 4.98 (1.15) | 4.61 (1.45) | 4.93 (1.42) | -0.29 | 86753055 | *t*(2079) = -0.12,  *p = .*453 |
| 3 | 5.89 (0.98) | 5.73 (1.10) | 5.80 (1.15) | -0.15 | 17.21983 | *t*(2079) = -3.26,  *p* < .001 |
| 4 | 6.03 (1.02) | 5.66 (1.28) | 5.66 (1.31) | -0.33 | 1053547959 | *t*(2079) = 7.29,  *p* < .001 |
| 5 | 4.92 (1.27) | 4.41 (1.49) | 4.65 (1.52) | -0.37 | 36629016917803 | *t*(2079) = 1.67,  *p = .*953 |
| 6 | 5.70 (1.15) | 5.34 (1.40) | 5.41 (1.34) | -0.29 | 34895351 | *t*(2079) = -0.26,  *p = .*396 |
| 7 | 6.35 (0.97) | 6.08 (1.37) | 6.15 (1.23) | -0.24 | 57723.32 | *t*(2079) = -1.38,  *p = .*084 |
| 8 | 5.23 (1.00) | 5.01 (1.14) | 5.03 (1.09) | -0.21 | 2860.367 | *t*(2079) = -1.99,  *p = .*023 |
| 9 | 5.53 (1.00) | 5.33 (1.20) | 5.35 (1.17) | -0.18 | 244.5024 | *t*(2079) = -2.55,  *p =* .005 |
| 10 | 5.71 (1.23) | 5.16 (1.54) | 5.19 (1.50) | -0.40 | 5648400076465052 | *t*(2079) = 2.28,  *p = .*989 |
| 11 | 5.57 (1.38) | 5.17 (1.61) | 5.20 (1.64) | -0.27 | 3124784 | *t*(2079) = -0.66,  *p = .*255 |
| 12 | 4.87 (1.45) | 4.53 (1.60) | 4.50 (1.58) | -0.23 | 10037.4 | *t*(2079) = -1.72,  *p = .*042 |
| 13 | 5.59 (1.25) | 5.73 (1.22) | 5.68 (1.26) | 0.11 | 1.939122 | *t*(2079) = 3.98,  *p* < .001 |
| 14 | 5.51 (1.28) | 5.24 (1.55) | 5.39 (1.40) | -0.19 | 428.613 | *t*(2079) = -2.42,  *p =* .008 |
| 15 | 4.48 (1.84) | 4.15 (1.88) | 3.39 (2.07) | -0.18 | 110.0661 | *t*(2079) = -2.75,  *p =* .003 |
| 16 | 5.93 (1.62) | 5.26 (1.86) | 3.73 (2.08) | -0.39 | 435869038730940 | *t*(2079) = 1.98,  *p = .*976 |
| 17 | 5.71 (1.11) | 5.40 (1.41) | 5.38 (1.46) | -0.26 | 396080.4 | *t*(2079) = -1.02,  *p = .*154 |
| 18 | 5.34 (1.51) | 5.30 (1.59) | 5.07 (1.47) | -0.03 | 0.05929815 | *t*(2079) = -6.12,  *p* < .001 |
| 19 | 5.03 (1.18) | 4.91 (1.46) | 3.60 (1.73) | -0.10 | 0.4510323 | *t*(2079) = 2.10,  *p = .*036 |
| 20 | 5.60 (1.07) | 5.21 (1.29) | 5.10 (1.43) | -0.34 | 41369659723 | *t*(2079) = 0.78,  *p = .*784 |

Here’s the PIL information from Rachel for the table:

Hey Dr. B!

This is just a reminder about the letter for grad school travel funding!

I have also provided percentages of demos below!

Gender

> table(week1.file.04.29.10$gender)

  1   2

 65 203

268 total

1 = 24.3%

2 = 75.7%

ethnicity

> table(week1.file.04.29.10$ethnicit)

  1   2   3   4   6

203  46   6   5   7

267 total

1 = 76.0%

2 = 17.2%

3 = 2.2%

4 = 1.9%

5 = 2.6%

> table(meaning.working.file.5.10.10$gender)

  1   2

107 185

292 total

1 = 36.6%

2 = 63.4%

> table(meaning.working.file.5.10.10$ethnic)

  1   2   3   4   5   6

210  61   5  10   1   5

292 total

1 = 71.9%

2 = 20.9%

3 = 1.7%

4 = 3.4%

5 = 0.3%

6 = 1.7%

> table(SES.Mike.and.Mike.Data$SEX)

  0   1

150 115

265 total

1 = 56.6%

2 = 43.4%

> table(SES.Mike.and.Mike.Data$RaceEthnicity)

  1   2   3   4

207  43   8   1

259 total

1 = 79.9%

2 = 16.6%

3 = 3.1%

4 = 0.4%

> table(logotherapy\_data\_factor\_analysis\_study$sex)

  0   1   2

  1 115 225

Is the zero an accident or coded?

341 total

0 = .3%

1 = 33.7%

2 = 66.0%

OR

340 total

1 = 33.8%

2 = 66.2%

> table(logotherapy\_data\_factor\_analysis\_study$ethnicit)

  1   2   3   4   6

285  48   3   2   3

341 total

1 = 83.6%

2 = 14.1%

3 = 0.9%

4 = 0.6%

5 = 0%

6 = 0.9%

Bayes for not random and paper

> ttestBF(x = finalnotrandom$total, y = finalpaper$total, paired = FALSE)

Bayes factor analysis

--------------

[1] Alt., r=0.707 : 19019086793721 ±0%

Against denominator:

Null, mu1-mu2 = 0

---

Bayes factor type: BFindepSample, JZS

Bayes for not random and random

> ttestBF(x = finalnotrandom$total, y = finalrandom$total, paired = FALSE)

Bayes factor analysis

--------------

[1] Alt., r=0.707 : 46.94438 ±0%

Against denominator:

Null, mu1-mu2 = 0

---

Bayes factor type: BFindepSample, JZS

Tost for paper and not random totals

> TOSTtwo(m1 = mean(finalpaper$total), m2 = mean(finalnotrandom$total),

+ sd1 = sd(finalpaper$total), sd2 = sd(finalnotrandom$total),

+ n1 = length(na.omit(finalpaper$total)), n2 = length(na.omit(finalnotrandom$total)),

+ low\_eqbound\_d = -.3, high\_eqbound\_d = .3,

+ var.equal = T, alpha = .05)

Using alpha = 0.05 Student's t-test was significant, t(2079) = 8.287848, p = 0.0000000000000002041748

Using alpha = 0.05 the equivalence test based on Student's t-test was non-significant, t(2079) = 1.590749, p = 0.9440909

TOST results:

t-value 1 p-value 1 t-value 2

1 14.98495 0.000000000000000000000000000000000000000000000001390655 1.590749

p-value 2 df

1 0.9440909 2079

Equivalence bounds (Cohen's d):

low bound d high bound d

1 -0.3 0.3

Equivalence bounds (raw scores):

low bound raw high bound raw

1 -4.771563 4.771563

TOST confidence interval:

Lower Limit 90% CI raw Upper Limit 90% CI raw

1 4.732492 7.077394

Tost for rondom and not random totals

> TOSTtwo(m1 = mean(finalrandom$total), m2 = mean(finalnotrandom$total),

+ sd1 = sd(finalrandom$total), sd2 = sd(finalnotrandom$total),

+ n1 = length(na.omit(finalrandom$total)), n2 = length(na.omit(finalnotrandom$total)),

+ low\_eqbound\_d = -.3, high\_eqbound\_d = .3,

+ var.equal = T, alpha = .05)

Using alpha = 0.05 Student's t-test was significant, t(1748) = -3.702516, p = 0.0002200948

Using alpha = 0.05 the equivalence test based on Student's t-test was significant, t(1748) = 2.562985, p = 0.005230357

TOST results:

t-value 1 p-value 1 t-value 2 p-value 2 df

1 2.562985 0.005230357 -9.968018 0.00000000000000000000004215658 1748

Equivalence bounds (Cohen's d):

low bound d high bound d

1 -0.3 0.3

Equivalence bounds (raw scores):

low bound raw high bound raw

1 -5.047334 5.047334

TOST confidence interval:

Lower Limit 90% CI raw Upper Limit 90% CI raw

1 -4.308412 -1.6569

Table XX.

*LPQ Item, Total Averages, and Effect Size of the Differences*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Item | Paper | Non-Random Computer | Random Computer | *d* Paper v Non-Random | BF10 Paper v Non-Random | TOST Paper v Non-Random | Decision |
| 1 | .70 (.46) | .61 (.49) | .57 (.50) | -0.18 | 44.89 | *t*(1796) = -3.69,  *p* < .001 | Null |
| 2 | .88 (.32) | .76 (.43) | .75 (.43) | -0.33 | 582817225 | *t*(1796) = -6.87,  *p* =.72 | Alt |
| 3 | .92 (.27) | .84 (.36) | .86 (.34) | -0.26 | 120140 | *t*(1796) = -.82,  *p* =.21 | Alt |
| 4 | .97 (.16) | .87 (.34) | .91 (.29) | -0.42 | 3.487143 x 10^14 | *t*(1796) = 2.38,  *p* =.99 | Alt |
| 5 | .55 (.50) | .51 (.50) | .42 (.49) | -0.09 | 0.32 | *t*(1796) = -4.37,  *p* < .001 | Null |
| 6 | .68 (.47) | .58 (.49) | .63 (.48) | -0.21 | 465.54 | *t*(1796) = -1.99,  *p* = .02 | Null |
| 7 | .82 (.39) | .81 (.39) | .77 (.42) | -0.01 | 0.06 | *t*(1796) = -5.94,  *p* < .001 | Null |
| 8 | .64 (.48) | .47 (.50) | .48 (.50) | -0.35 | 6312931530 | *t*(1796) = 0.95,  *p* = .83 | Alt |
| 9 | .92 (.28) | .78 (.41) | .81 (.39) | -0.39 | 8.122311 x 10^12 | *t*(1796) = 1.91,  *p* = .97 | Alt |
| 10 | .79 (.41) | .65 (.48) | .64 (.48) | -0.32 | 175744354 | *t*(1796) = .41,  *p* = .66 | Alt |
| 11 | .87 (.34) | .76 (.43) | .73 (.45) | -0.29 | 2441504 | *t*(1796) = -.28,  *p* = .39 | Alt |
| 12 | .90 (.30) | .75 (.43) | .79 (.41) | -0.42 | 1.540006 x 10^14 | *t*(1796) = 2.28,  *p* = .99 | Alt |
| 13 | .96 (.19) | .91 (.29) | .96 (.18) | -0.22 | 1381.01 | *t*(1796) = -1.73,  *p* = .04 | Null |
| 14 | .88 (.32) | .77 (.42) | .76 (.43) | -0.31 | 93292986 | *t*(1796) = .32,  *p* = .62 | Alt |
| 15 | .44 (.50) | .39 (.49) | .32 (.47) | -0.10 | 0.64 | *t*(1796) = -4.03,  *p* < .001 | Null |
| 16 | .98 (.15) | .87 (.33) | .86 (.34) | -0.42 | 2.2222385 x 10^15 | *t*(1796) = 2.60,  *p* < 1.00 | Alt |
| 17 | .89 (.31) | .81 (.39) | .85 (.36) | -0.24 | 10266.8 | *t*(1796) = -1.30,  *p* = .10 | Alt |
| 18 | .89 (.32) | .83 (.38) | .83 (.38) | -0.18 | 43.88 | *t*(1796) = -2.59,  *p* = .004 | Null |
| 19 | .55 (.50) | .50 (.50) | .46 (.50) | -0.10 | 0.38 | *t*(1796) = -4.28,  *p* < .001 | Null |
| 20 | .86 (.35) | .71 (.45) | .72 (.45) | -0.38 | 491515415658 | *t*(1796) = 1.55,  *p* = .94 | Alt |

Interpretation of Bayes/equivalence:

Bayes (10) = 1-3 suggests barely any evidence for the alternative, 3-10 suggests positive evidence for the alternative, 10-100 suggests strong evidence, > 100 very strong, < 1 suggests no evidence for anything (keep the null)

TOST – anything significant suggests support for the null hypothesis

|  |
| --- |
| TOST (random vs. not random) |
| *t*(1632) = -4.14, *p* < .001 |
| *t*(1632) = -5.72, *p* < .001 |
| *t*(1632) = -4.81, *p* < .001 |
| *t*(1632) = -3.52, *p* < .001 |
| *t*(1632) = -2.42, *p* = .01 |
| *t*(1632) = -3.75, *p* < .001 |
| *t*(1632) = -4.27, *p* < .001 |
| *t*(1632) = -5.38, *p* < .001 |
| *t*(1632) = -4.55, *p* < .001 |
| *t*(1632) = -5.58, *p* < .001 |
| *t*(1632) = -4.49, *p* < .001 |
| *t*(1632) = -4.34, *p* < .001 |
| *t*(1632) = -1.32, *p* = .09 |
| *t*(1632) = -5.66, *p* < .001 |
| *t*(1632) = -3.00, *p* = .001 |
| *t*(1632) = -5.45, *p* < .001 |
| *t*(1632) = -4.15, *p* < .001 |
| *t*(1632) = -5.98, *p* < .001 |
| *t*(1632) = -4.69, *p* < .001 |
| *t*(1632) = -5.56, *p* < .001 |

-not sure if we will need random vs. not random but I suck and ran the wrong ones so here they are ☺