Supplementary Materials for:

Justifying responses affects the relationship between confidence and accuracy

Daniel B. Wright Sarah M. Wolff
University of Nevada at Las Vegas

Abstract

This document goes through some additional analyses and shows the R code. These are for the response times for math items and for the interest questions, both of these by the three conditions. We had no predictions for these differing by condition. We also show the key for the items. This document is written in **knitr** (Xie, 2015), weaving together LaTeX and R. The code for all analyses in the paper can be found in the rnw file associated with that, which is also on the GitHub page.

The data are loaded. These are from a file that has been cleaned from a Qualtrics data file with identifying information removed. This was run in the same session as another (unpublished) study to ensure no one completed both. The duplicate IP address exclusion was done for the whole session and was done before the data object, below, was created.

```
# This is where file located on computer
setwd( "C:\\Users\\wrighd12\\Documents\\MetaCog\\Replace2")
load("justify.RData")
```

An object is created for the times for answering the mathematics items, and those faster than, on average, 10s are removed.

```
times <- with(justify,cbind(Jtime05_Page.Submit,Jtime07_Page.Submit,
  Jtime08_Page.Submit,Jtime09_Page.Submit,Jtime10_Page.Submit,
  Jtime12_Page.Submit,Jtime13_Page.Submit,Jtime16_Page.Submit,
  Jtime26_Page.Submit,Jtime27_Page.Submit))
totqtimes <- apply(times,1,sum)
tt <- cbind(times,totqtimes)
thresh <- 100
outfortime <- table(totqtimes < thresh)
```

Email: daniel.wright@unlv.edu or dbrookswr@gmail.com.

```
justify <- justify[totqtimes > thresh,]
post <- table(justify$justify)
times <- times[totqtimes > thresh,]
totqtimes <- totqtimes[totqtimes > thresh]
# justify 0 is control, 1 is correct, and 2 is incorrect
```

This is a set of packages downloaded. Not all are used here. Functions for the traditional upper and lower bounds for the 95% confidence intervals are also created.

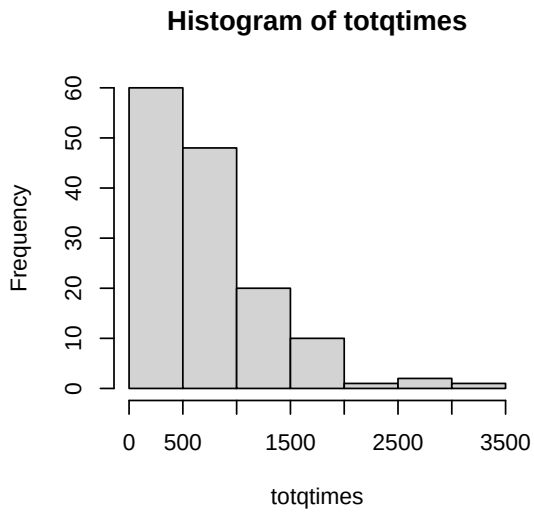
```
library(car)
library(lme4)
library(mirt)
library(xtable)
library(splines)
library(Hmisc)
library(boot)
library(pwr)
library(psych)
library(EnvStats)
lb <- function(x,lev=.95)
  t.test(x,conf.level=lev)$conf.int[1]
ub <- function(x,lev=.95)
  t.test(x,conf.level=lev)$conf.int[2]
```

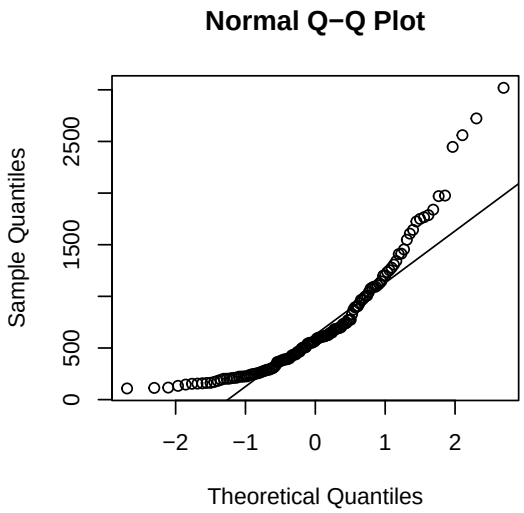
Here is the key.

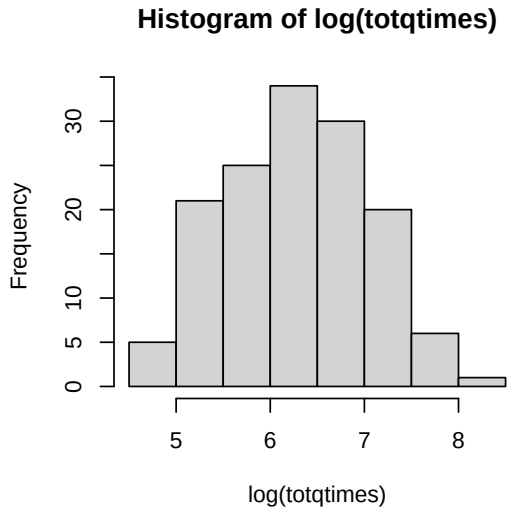
Response times by condition

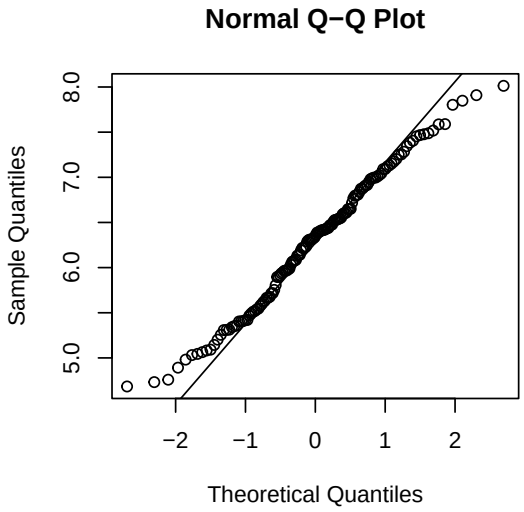
We were asked by a reviewer if the response times varied by condition. The response times were skewed, and this was greatly lessened by using logged response times as shown below. Issues with the traditional standard error of this measure is discussed in Wright and Herrington (2011). We report geometric means along with arithmetic means for descriptive statistics. Hypothesis testing was done with the logged response times and no significant differences emerged.

```
library(e1071)
par(mfrow=c(2,2))
hist(totqtimes)
qqnorm(totqtimes);qqline(totqtimes)
skewness(totqtimes)

## [1] 1.597974

hist(log(totqtimes))
qqnorm(log(totqtimes));qqline(log(totqtimes))
```










```
skewness(log(totqtimes))

## [1] -0.01027005
```

The output for the geometric means betwen conditions. This shows that they are nonsignificant for the two contrasts, and the oneway anova.

```
group <- justify$group <- as.factor(justify$justify) #same name
g12 <- group != 0
g2 <- group == 2
summary(lm(log(totqtimes) ~ g12 + g2))

##
```

```
## Call:
## lm(formula = log(totqtimes) ~ g12 + g2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66893 -0.60603  0.03343  0.52888  1.70527
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.2043     0.1066  58.187   <2e-16 ***
## g12TRUE       0.2235     0.1540   1.452    0.149
## g2TRUE       -0.1621     0.1597  -1.015    0.312
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7615 on 139 degrees of freedom
## Multiple R-squared:  0.0157, Adjusted R-squared:  0.001534
## F-statistic: 1.108 on 2 and 139 DF,  p-value: 0.333
```

```
tapply(totqtimes,group,mean)
```

```
##        0        1        2
## 670.7110 820.0560 663.2856
```

```
geomean <- function(x) exp(mean(log(x)))
tapply(totqtimes,group,geomean)
```

```
##        0        1        2
## 494.8563 618.7818 526.1952
```

```
oneway.test(log(totqtimes)~group)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  log(totqtimes) and group
## F = 1.021, num df = 2.000, denom df = 92.374, p-value = 0.3643
```

This was repeated using multilevel (cross-classified with randon intercepts for items and people, see Wright & London, 2009) models with the individual trial data, which are more skewed as expected. The same basic finding emerged: no significant effects detected.

```
timeVector <- c(times)
itemno <- rep(1:10,each=nrow(times))
subno <- rep(1:nrow(times),10)
groupno <- rep(group,10)
```

```
group12 <- groupno != 0
group2 <- groupno == 2
skewness(timeVector); skewness(log(timeVector))

## [1] 9.407021
## [1] -0.2789875

tapply(timeVector,groupno,mean)

##        0        1        2
## 67.07110 82.00560 66.32856

tapply(timeVector,groupno,geomean)

##        0        1        2
## 31.71827 43.54277 41.37853

m0 <- lmer(timeVector ~ 1 + (1|itemno) + (1|subno), REML=FALSE)
m1 <- update(m0, .~. + group12)
m2 <- update(m1, .~. + group2)
m2a <- update(m0, .~. + groupno)
anova(m0,m1,m2,m2a)

## Data: NULL
## Models:
## m0: timeVector ~ 1 + (1 | itemno) + (1 | subno)
## m1: timeVector ~ (1 | itemno) + (1 | subno) + group12
## m2: timeVector ~ (1 | itemno) + (1 | subno) + group12 + group2
## m2a: timeVector ~ (1 | itemno) + (1 | subno) + groupno
##     npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0     4 17624 17645 -8807.8    17616
## m1     5 17625 17651 -8807.5    17615 0.5605  1     0.4540
## m2     6 17625 17657 -8806.6    17613 1.7857  1     0.1814
## m2a    6 17625 17657 -8806.6    17613 0.0000  0

anova(m0,m2a)

## Data: NULL
## Models:
## m0: timeVector ~ 1 + (1 | itemno) + (1 | subno)
## m2a: timeVector ~ (1 | itemno) + (1 | subno) + groupno
##     npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m0     4 17624 17645 -8807.8    17616
## m2a    6 17625 17657 -8806.6    17613 2.3463  2     0.3094
```

Reminder: justify 0 is control, 1 is correct, and 2 is incorrect.

```
group12 <- group != 0
group2 <- group == 2
```

Interest Questions

The interest questions are:

1. We are interested in how much people think about how they are answering questions, called metacognition, during quizzes. Examples including thinking why you believe an individual alternative is right or wrong, and thinking about how to answer a question. Using the following scale, how much metacognitive thinking were you doing during the task? (0 to 100, hardly any ... a lot)
2. How interested are you in mathematics? (0 – 100, not interested ... very interested)
3. Some people feel like taking a quiz helps them to learn a topic. Some people do not feel this. Do you feel taking this short quiz helped with your mathematical knowledge? (0–100, Knowledge decreased ... no effect ... Knowledge increased)

```
attqs <- with(justify,cbind(Q3027_1,Q3029_1,Q3031_1,Q3033_1))
attmci <- matrix(ncol=3,nrow=3*4)
for (i in 1:4){
  attmci[3*i - 2,] <- tapply(attqs[,i],group,mean)
  attmci[3*i - 1,] <- tapply(attqs[,i],group,lb)
  attmci[3*i - 0,] <- tapply(attqs[,i],group,ub)
  }
vals <- cor(attqs)[lower.tri(cor(attqs),diag=FALSE)]

ovals <- matrix(ncol=3,nrow=3)
odata <- rowMeans(attqs)
ovals[1,] <- tapply(odata,group,mean)
ovals[2,] <- tapply(odata,group,lb)
ovals[3,] <- tapply(odata,group,ub)
```

Participants were asked four 0–100 attitude questions: how accurately the assessment measures your skills, how much metacognition were you doing, how interested you are in mathematics, and did the quiz help your understanding. These questions were included for exploratory purposes and our analytic approach reflects this. The means and 95% confidence intervals for the three groups are shown in Table 1. The responses on these variables were correlated between $r = .199$ and $r = .575$, which accounts for why these show a similar pattern.

```
vals <- matrix(ncol=3,nrow=5)

for (i in 1:ncol(attqs))
```

Table 1

Means and 95% confidence intervals for the four 0–100 rating scale questions.

		Condition		
		Control	Correct	Incorrect
Assessment is accurate	Mean	65.078	54.489	56.295
	95% CI	(59.739, 70.418)	(47.618, 61.361)	(50.264, 62.327)
Metacognition	Mean	72.745	65.638	61.750
	95% CI	(66.380, 79.111)	(57.314, 73.962)	(52.819, 70.681)
Interest	Mean	75.059	69.277	63.773
	95% CI	(68.565, 81.552)	(60.532, 78.022)	(54.407, 73.139)
Quiz helps understanding	Mean	73.549	68.894	65.386
	95% CI	(67.743, 79.355)	(61.799, 75.989)	(58.557, 72.216)

```
  vals[i,] <-  c(pairwise.t.test(attqs[,i],group)$p.value)[c(1,2,4)]
vals[5,] <- c(pairwise.t.test(rowMeans(attqs),group)$p.value)[c(1,2,4)]
```

As these are exploratory analyses all three pairwise t-tests were were conducted for each question and then the p-values were adjusted using Holm's procedure (Holm, 1979). The only difference that was statistically significant was participants in the justify correct condition gave lower ratings for the perception that the assessment was accurate than participants in the control condition.

```
cd <- matrix(ncol=3,nrow=5*3)
cd[1,] <- cohen.d(attqs[group == 0 | group == 1,1],
                  group[group == 0 | group == 1])$cohen.d
cd[2,] <- cohen.d(attqs[group == 0 | group == 2,1],
                  group[group == 0 | group == 2])$cohen.d
cd[3,] <- cohen.d(attqs[group == 1 | group == 2,1],
                  group[group == 1 | group == 2])$cohen.d
cd[4,] <- cohen.d(attqs[group == 0 | group == 1,2],
                  group[group == 0 | group == 1])$cohen.d
cd[5,] <- cohen.d(attqs[group == 0 | group == 2,2],
                  group[group == 0 | group == 2])$cohen.d
cd[6,] <- cohen.d(attqs[group == 1 | group == 2,2],
                  group[group == 1 | group == 2])$cohen.d
cd[7,] <- cohen.d(attqs[group == 0 | group == 1,3],
                  group[group == 0 | group == 1])$cohen.d
cd[8,] <- cohen.d(attqs[group == 0 | group == 2,3],
                  group[group == 0 | group == 2])$cohen.d
cd[9,] <- cohen.d(attqs[group == 1 | group == 2,3],
                  group[group == 1 | group == 2])$cohen.d
cd[10,] <- cohen.d(attqs[group == 0 | group == 1,4],
                  group[group == 0 | group == 1])$cohen.d
```

```
cd[11,] <- cohen.d(attqs[group == 0 | group == 2,4],
                   group[group == 0 | group == 2])$cohen.d
cd[12,] <- cohen.d(attqs[group == 1 | group == 2,4],
                   group[group == 1 | group == 2])$cohen.d
cd[13,] <- cohen.d(rowMeans(attqs[group == 0 | group == 1,]),
                   group[group == 0 | group == 1])$cohen.d
cd[14,] <- cohen.d(rowMeans(attqs[group == 0 | group == 2,]),
                   group[group == 0 | group == 2])$cohen.d
cd[15,] <- cohen.d(rowMeans(attqs[group == 1 | group == 2,]),
                   group[group == 1 | group == 2])$cohen.d
colnames(cd) <- c("lb","d","ub")
rownames(cd) <-
  matrix(t(outer(paste0("q",c(1:4,"tot")),c("0v1","0v2","1v2"),paste)))
```

```
print.xtable(xtable(cd,caption="The 95\\% confidence intervals for Cohen's $d$
       for the pairwise comparisons for the interest questions.
       The first 12 rows are for the four questions, the final
       3 for the sum of these responses. These are not making any adjustment
       for there being multiple comparisons. "),hline.after=c(-1,0,12,15))
```

	lb	d	ub
q1 0v1	-0.91	-0.50	-0.10
q1 0v2	-0.87	-0.46	-0.05
q1 1v2	-0.33	0.08	0.50
q2 0v1	-0.68	-0.28	0.12
q2 0v2	-0.83	-0.43	-0.02
q2 1v2	-0.55	-0.14	0.28
q3 0v1	-0.62	-0.22	0.18
q3 0v2	-0.83	-0.42	-0.01
q3 1v2	-0.60	-0.18	0.23
q4 0v1	-0.61	-0.21	0.19
q4 0v2	-0.79	-0.38	0.02
q4 1v2	-0.56	-0.15	0.26
qtot 0v1	-0.79	-0.39	0.01
qtot 0v2	-1.03	-0.62	-0.20
qtot 1v2	-0.56	-0.15	0.27

Table 2

The 95% confidence intervals for Cohen's d for the pairwise comparisons for the interest questions. The first 12 rows are for the four questions, the final 3 for the sum of these responses. These are not making any adjustment for there being multiple comparisons.

References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70. doi: 10.2307/4615733

Wright, D. B., & Herrington, J. A. (2011). Problematic standard errors and confidence intervals for skewness and kurtosis. *Behavior Research Methods*, *43*, 8-17. doi: 10 .3758/s13428-010-0044-x

Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*, 439-456. doi: 10.1348/ 000711008X327632

Xie, Y. (2015). *Dynamic documents with R and **knitr*** (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.