**On-line Supplementary Material**

For “*P*-curving as a safeguard against *p*-hacking in SLA research: A case study”

By Seth Lindstromberg

**Contents**

p. 2 **Table A**: The ten accepted primary studies.

Note. This table does not take account of interactions. For that information see tables 2 and 3 in the main text.

p. 9 **Table B**: The excluded primary studies

p. 11 **Reference list for Table B**

p. 14 **R script:** Script for the simulations, for the corresponding figures, and for generating input to the p-curve app

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table A. *The disclosure table regarding the primary studies figuring in the case study* | | | | | | |
| *Key*   * CEF = The Common European Framework for labelling proficiency levels– e.g., A2 ≈ elementary * DV = Dependent (or outcome) variable * F = Forewarning of the (near) immediate posttest * Format = Format of the (near) immediate posttest * MC = multiple choice gloss * *N*items = Number of targeted vocabulary items * Pilot = Pilot study to assess familiarity of candidate target items * Pretest = Pretest to filter out known candidate target items or overly proficient learners * With respect to approximate levels of learner proficiency, quote marks indicate authors’ own assessments. * SG = Single word gloss * RA = Random Assignment of learners to learning conditions * RA intact = RA within intact groups | | | | | | |
| *Primary Study* | *Features of Procedure & Learner Proficiency* | *Experimental design (usually omitting information about delayed posttests, tests of form recognition, and tests of reading comprehension)* | *Quoted text from original paper with statistical results unless results were given only in a table* | | | *Further comments plus the recalculated p-value for the comparison between L1 & L2 glossing in the immediate test. Usually omitting information about interactions.* |
| 1  Arpaci (2016) | Pilot, RA, No F*, N*items = 16, format MC, level: “A2” (CEF) | 1-way Anova “with planned comparisons”; 3 gloss conditions (L1 L2, None) | “Post-hoc comparisons using Bonferroni test demonstrated that the mean scores of each group were significantly different from each other (L1 gloss>L2 gloss, p = .043." | | | This recalculated *p-*value is much lower than Arpaci’s at least part because Arpaci applied a Bonferroni adjustment:  p =.000517 |
| 2  Ertürk (2016) | No RA; pilot, F, *N*items = 16, level: “elementary”, format: MC | 1-way ANOVA; 3 gloss conditions (L1, L2, None) | “Post hoc analysis results indicated that there was a statistically significant difference between no-gloss (M = 21.19, SD = 5.04) and L2 gloss (M = 18.21, SD = 6.18) conditions (no gloss > L2 gloss, p = .023) and L1 (M = 22.90, SD = 4.16) and L2 gloss (M = 18.21, SD = 6.18) conditions (L1 > L2, p = .000). However, there was not any significant difference between no-gloss and L1gloss groups.” | | | p = .0000883 |
| 3  Farvardin & Biria (2012) | RA, pretest, F unclear; *N*items = 25 (an item could be known by up to 25% of the learners), level: apparently mid- upper intermediate; format: give meaning in L1 or L2 | 1-way ANOVA:  L1 SG, L2 SG, L2 MC | “As Table 3 depicts, the effect of gloss remained significant throughout the two posttests; F (2, 96) = 23.07, p<.05 on the immediate posttest…” | | | I left the scores for the L2 MC gloss condition out of account because there was no L1 MC gloss condition.  p = .000001455  . |
| 4  Ko (1995) | No RA, pretest,  +/-F a factor;  *N*items = 15,  level unclear; likely mixed but above low intermediate, format: MC | An implied 2-way ANOVA within MANOVA.  Incidental learning (+/-) x Glossing (L1, L2, None). | "There was a significant difference between glossed conditions (Ll gloss and L2 gloss) and the no gloss condition, as well as between Ll gloss and L2 gloss conditions." | | | The following p-value stems from L1 & L2 glossing scores collapsed across +/- F:  p = .00003089 |
| 5  Ko (2017) | RA, multiple pilots, no F, *N*items = 14, format: MC, likely level of the low proficiency group: intermediate | 4 x 2 between subjects ANOVA. Gloss (L1, L2, L1+L2, None ) x Proficiency (low vs high). | "The L1 gloss and the L1+L2 gloss were effective for lower level  learners, and the L2 gloss and the L1+L2 gloss were effective for higher-level learners." | | | p = .01447 |
| 6  Mitarai & Aizawa (1999) | No RA, no F, pretest, *N*items = 14, Likely level: low to mid-intermediate, format: MC | 2 x 2 ANOVA: Language (L1 vs L2) x Gloss type (SG vs MC). Oddly, the researchers eliminated learners from the MC condition who made 4 or more wrong gloss choices during the treatment. | "The language effect was significant both in the SG and MCG [conditions], F(1,181) = 4.90,P < .05, F(1,181) = 24.71, p < .001, respectively." | | | p = .000007215 |
| 7  Öztürk &Yorgancı (2017) | No RA, pretest,  no F;  *N*items = 21, level: “intermediate” | L1 SG vs L2 SG.  IS t-test on the scores from the immediate posttest of meaning recall | “L1 gloss was found to be superior to L2 gloss” | | | p = .01972 |
| 8  Pishghadam & Ghahari (2011) | RA intact, pretest, *N*items = 10, probably F, “participants were at pre-intermediate, intermediate, and upper-intermediate, Format: MC | 1-way ANOVA:  L1 SG,  L1 MC,  L2 SG,  L2 MC | "Single L1 gloss > Single L2 gloss / MCL1 gloss/ MC L2 gloss"  For the case study the result was:  L1 (both gloss types combined) > L2 (both gloss types combined) | | | p =.000004537 |
| 9  Shiki (2008) | RA, Pseudowords, No F, *N*items = 16, “intermediate” (max TOEFL score = 450), format: give meaning in L1 or L2 | 1-way ANOVA  L1 SG,  L1 MC,  L2 SG,  L2 MC | "L1 glosses appeared more effective for the participants in this study  than L2 glosses in order to retain words incidentally during reading." | | | p = .00002881 |
| 10  Yoshii (2006) | RA, no F, a pretest to exclude high scorers. Target: 14 verbs. Tests: Supply meaning in L1. It was “assumed that the participants were low intermediate or intermediate” | ANOVA: “2 (L1, L2) X 2 (picture, no picture) X 2 (immediate test, delayed test)”. Glossing: “(1) L1 text only; (2) L2 text only; (3) L1 text plus picture; and (4) L2 text plus picture". | | "a significant interaction effect between languages and tests (*F* = 9.45, p < .05, η2 = .05), indicating that patterns of vocabulary retention differed for L1 and L2 gloss groups". | When L1 & L2 scores are collapsed across +/- picture,  *p* = .4474. When the + picture scores are omitted, *p* = .01828.  L1 text only scores always > L2 text only scores, but significantly more so on the immediate test than on the delayed test. | |

|  |  |  |
| --- | --- | --- |
| Table B. *The excluded studies* | | |
| *Study* | *T, df, p* (if re-calculated) | *Reason(s)* |
| Barabadi, Asma, & Panahi (2018) | *t* = 1.6588, df = 123.99,  *p* = .09968 | *p* > .05; Within subs, odd contrasts (e.g., 1 L2 gloss vs 2 English glosses (with vs without Farsi glosses) |
| Choi (2016) | *t* = 0.48412, df = 116.19,  *p* = .6292 | *p* > .05 |
| Ha (2016) |  | Reading Comprehension only |
| Jacobs, Dufon, & Hong. (1994) |  | *p* > .05 |
| Kang, Kweon, & Choi (2020) | *t* = 1.1657, df = 46.698,  *p* = .2496 | *p* > .05 |
| Kim & Choi (2017) |  | Apparently *p* > .05 for low level learners and all learners combined, and *p* < .05 only for upper-intermediates |
| Ko (2005) |  | *p* > .05; reading comprehension only |
| Ko (2012) | *t* = -1.0734, df = 53.064,  *p* = .288 | *p* > .05 |
| Kongtawee & Sappapan (2018) | *t* = 16.201, df = 163.72,  *p* < .00001 | Within subjects. Learner proficiency may have been extremely low. |
| Miyasako (2002) | *t* = -2.1093, df = 93.853,  *p* = .03758 | Negative *t*-statistic (*p* < .05) |
| Palmer (2003)\* |  | Unpublished dissertation |
| Rouhi & Mohebbi (2012) | *t* = 1.2773, df = 24.671,  *p* = .2134 | *p* > .05 |
| Salehi & Naserieh (2013) | *t* = 0.94279, df = 33.303,  *p* = .3526 | *p* > .05 |
| Salimi & Mirian (2019) | *t* = 5.7026, df = 57.428,  *p* = .0000004305 | Reading Comprehension only |
| Taheri & Zade (2014) | *t* = -3.0011, df = 119.66,  *p* = .003275 | Negative *t*-statistic (p < .05).  Within subjects. Learners apparently mainly upper intermediate. |
| Zarei & Hasani (2011) | *t* = 1.035, df = 155.55,  *p* = .3023 | *p* > .05; no immediate posttest. L1 & L2 gloss totals aggregated across 4 conditions of gloss placement. |
| *Note.*  \* Several additional unpublished studies covered by Yanagisawa et al. (2020) have not been included in this table. | | |

**References for Table B**

Barabadi, E., Asma, A., & Panahi, A. (2018). The relative impact of L1 and L2 glosses along with computer-generated phonological guidance on EFL learners’ vocabulary learning. *Cogent Education, 5*, 1–13. <https://www.tandfonline.com/toc/oaed20/5/1?nav=tocList>

Choi, S. (2016). Effects of L1 and L2 glosses on incidental vocabulary acquisition and lexical representations. *Learning and Individual Differences, 45,* 137–143. <https://doi.org/10.1016/j.lindif.2015.11.018>

Ha, M.-J. (2016). L1 and L2 glosses: Their effects on Korean EFL learners’ reading comprehension. *Studies in Linguistics, 40*, 403–421. <https://doi.org/10.14257/astl.2015.92.04>

Jacobs, G., Dufon, P., & Hong, F.C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading, 17*, 19–28. <https://doi.org/10.1111/j.1467-9817.1994.tb00049.x>

Kang, H., Kweon, S.-O., & Choi, S. (2020). Using eye-tracking to examine the role of first and second language glosses. *Language Teaching Research*. Epub ahead of print 3 June 2020. https://doi.org/10.1177/1362168820928567

Kim, H.-S., & Choi, U.-Y. (2017). L1 vs. L2 glosses, learner perceptions, and L2 vocabulary learning. *Modern English Education, 18*, 67–89. <https://doi.org/10.1177/1362168820981394>

Ko, M. (2005). Glosses, comprehension, and strategy use. *Reading in a Foreign Language, 17*, 125–143. https://www.researchgate.net/publication/291826015\_Glosses\_Comprehension\_and\_Strategy\_Use

Ko, M. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly, 46*, 56–79. <https://doi.org/10.1002/tesq.3>

Kongtawee, P., & Sappapan, P. (2018). The effects of L1 and L2 hypertext glosses on reading comprehension and vocabulary retention among Thai secondary school students. *Arab World English Journal, 9*, 367–380. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3258839>

Palmer, R. (2003). *A comparison of the effect of glossed self-instruction reading materials and traditional teacher fronted instruction.* Unpublished doctoral dissertation, Indiana University of Pennsylvania, Indiana, PA, USA. https://search.proquest.com/openview/c14435046f898594f8a39db2504ddc4c/1?pq-origsite=gscholar&cbl=18750&diss=y

Rouhi, A., & Mohebbi, H. (2012). The effect of computer assisted L1 and L2 glosses on L2 vocabulary learning. *The Journal of Asia TEFL, 9*, 1–19. http://journal.asiatefl.org/

Salehi, V., & Naserieh, F. (2013). The effects of verbal glosses on vocabulary learning and reading comprehension. *Asian EFL Journal, 15*, 24–64. https://www.asian-efl-journal.com/main-editions-new/the-effects-of-verbal-glosses-on-vocabulary-learning-and-reading-comprehension/

Salimi, E., & Mirian, E. (2019). The effect of L1 glosses for abstract words on English reading comprehension. *The Reading Matrix: An International Online Journal, 19*, 181–196. <http://www.readingmatrix.com/files/20-th6ehz28.pdf>

Taheri, P., & Zade, M. (2014). On the incidental learning of abstract and concrete vocabulary items in L1 gloss and L2 gloss. *English Language Teaching, 1*, 77–91. <https://www.semanticscholar.org/paper/On-the-Incidental-Learning-of-Abstract-and-Concrete-Taheri/c4715cb557b382e7bb3366fdc03a425089333b4b>

Zarei, A., & Hasani, S. (2011). The effects of glossing conventions on L2 vocabulary recognition and production. *The Journal of Teaching Language Skills, 3*, 209–233. <https://jtls.shirazu.ac.ir/article_390.html>

**R Script**

Notes.

1. The number of simulations was always one million. Consequently, on some computers an overall simulation can take more than ten minutes to complete. Running time can be reduced by taking fewer samples: But this will only work if the y-axis is re-specified. Also, the resulting *p-*curves will look somewhat different than they do in the corresponding figure in the main text.
2. From here on, a line must not begin with a capital letter, lines beginning with # excepted.. If the following script is copied into a Word document, some line-initial letters may be automatically capitalized. Any such capital letters should be changed back to small case.

#FIGURES 1 – 6

#

############Figure 1

##Simulation for figure 1

set.seed=1 # Set a random number generator for replicability

pvals.fig1 = c() # To make an empty container for the results

# generated by the next line of code

for (i in 1:10^6) pvals.fig1[i] = y<-t.test(rnorm(25,10,1),

rnorm(25,10,1),var.equal=T)$p.val

options(scipen=999) # So as not to get numbers in scientific notation

length(pvals.fig1[pvals.fig1<=.05])/length(pvals.fig1) # To get the % of p values under .05

## Make Figure 1

par( tck = -0.01, family = "serif", cex =1.0,

pin=c(4,2),bty="n")

hist(pvals.fig1,breaks=39, ylim =c(0, 60000),cex.lab= 1.2,

col="lightgreen", border = "white", main="",

xlab = "P-values",ylab="Count")

arrows(x0 =.05, y0=20000, x1 = .05, y1= -1, # For the vertical line.

length = 0, col = "black", lwd = 3, lty = 3)

legend("top",

c("Here H0 is true and Cohen's d = 0.\n\nSo, 5% of these p-values are < .05."),

bty="n", cex = 1.15)

## end fig 1

#

#######Figure 2

## The simulation for figure 2

set.seed=1

pfig2 = c()

for (i in 1:10^6) pfig2[i] = y<-t.test(rnorm(25,10,1), rnorm(20,11,1),var.equal=T)$p.val

length(pfig2[pfig2>.05])/length(pfig2)

length(pfig2[pfig2<=.05])/length(pfig2)

## Make Figure 2

par(mgp = c(2, .5, 0), tck = -0.01, family = "serif",

cex =1.0, bty="n")

hist(pfig2,breaks=99,ylim=c(0,87000), cex.lab= 1.2,

col="lightgreen", border = "white", main="",xlab = "P-values",ylab="Count")

arrows(x0 =.05, y0=30000, x1 = .05, y1= -1, length = 0,

col = "black", lwd = 3, lty = 3)

legend("top", c("Now d = 1.00. So with n.x = n.y = 20,\n\na priori statistical power = 87%.\n\nThus, 87% of these p-values are < .05."),bty="n", cex = 1.15)

##end fig 2

#

######### Figure 3

## Make Figure 3 (no new simulation needed)

plot.new()

par(mgp = c(2, .5, 0), tck = -0.01, family = "serif",

cex =1.1, bty="n")

hist(pfig2[pfig2 < .05],breaks=39,ylim=c(0,120000), cex.lab= 1.1,

col="lightgreen", border = "white", main="",

xlab = "Significant p-values",ylab="Count")

legend("center",

c("These are the p-values from figure 2\n\n(87% of the total) that lie below .05."),

bty="n", cex = 1.15)

arrows(x0 =.025, y0=10000, x1 = .025, y1= -1, length = 0, col = "black", lwd = 3, lty = 3)

## end fig 3

#

####### Figure 4

## Simulation for figure 4

set.seed=123

pfig4 = c() # This line makes a receptacle for the p-values

for (i in 1:10^6) {

x<-rnorm(25,20,4.50)

y<-rnorm(25,20,4.00)

t<-t.test(x,y, var.equal=T)$p.val

w<-wilcox.test(x,y)$p.val

xout<-x[-c(which.min(x),which.max(x))]

yout<-y[-c(which.min(y),which.max(y))]

tout<-t.test(xout,yout, var.equal=T)$p.val

wout<-wilcox.test(xout,yout)$p.val

pfig4[i] <- ifelse(test = t < .05, yes = t,

ifelse(test = w < .05, yes = w,

ifelse(test = tout<.05, yes = tout, no = wout))) }

## Make Figure 4

plot.new()

options(scipen = 999)

# Top histogram

par(mfrow=c(2,1)) # To get 2 plots in 1 display

par(mgp = c(2, 0.5, 0), tck = -0.01, family = "serif",

mai= c(0.2, 0.75, 0.5, 0.5),

cex =1.0, pin=c(5,2.5),bty="n")

hist(pfig4,breaks = 19, ylim = c(0, 110000),

cex.lab= 1.0, col="lightgreen", border = "white",

main="",xlab = "P values",ylab="Number of p values")

arrows(x0 = 0.05, y0=100000, x1 = .05, y1= 0, length = 0,

col = "black", lwd = 3, lty = 3)

legend(c(.02), c(90000), c("The dashed line marks p = .05."),bty="n", cex = 1.15)

# Bottom histogram

par(mgp = c(2, 0.5, 0), tck = -0.01, family = "serif",

mai = c(1.0,0.75, 0.0, 0.5),

cex =1.0, pin=c(5,1.5),bty="n")

hist(pfig4[pfig4 <.05],breaks = 19, cex.lab= 1.1, col="lightgreen",

border = "white", main="",xlab = "P-values",ylab="")

## end

#

############### Figure 5, H0 = F, considerable selection bias

## Simulation for fig 5

set.seed=123

pfig5 = c() # To make an empty container for the results

# generated by the following lines.

for (i in 1:10^6) {

x<-rnorm(25,20,4.50)

y<-rnorm(25,22,4.00)

t<-t.test(x,y, var.equal=T)$p.val

w<-wilcox.test(x,y)$p.val

xout<-x[-c(which.min(x),which.max(x))]

yout<-y[-c(which.min(y),which.max(y))]

tout<-t.test(xout,yout, var.equal=T)$p.val

wout<-wilcox.test(xout,yout)$p.val

pfig5[i] <- ifelse(test = t < .05, yes = t,

ifelse(test = w < .05, yes = w,

ifelse(test = tout<.05, yes = tout, no = wout))) }

## Put both 2 plots in 1 display

par(mfrow=c(2,1))

par( tck = -0.01, family = "serif", cex =1.0, mgp=c(1.5,0.5,0),

pin=c(5,2.2),bty="n",mai= c(0.5, 0.75, 0.5, 0.5))

hist(pfig5,breaks = 19, ylim = c(0, 120000), cex.lab= 1.1, col="lightgreen",

border = "white", main="",xlab = "",ylab="")

abline(v = .05, col = "black", lwd = 3, lty = 3)

#legend(-0.08, 96000,c("p = .05"),bty="n")

#

par(mgp = c(2, .5, 0), tck = -0.01, family = "serif",

cex =1.0, mai = c(1.0,0.75, 0.0, 0.5), bty="n")

hist(pfig5[pfig5 <.05],breaks = 19, cex.lab= 1.1, col="lightgreen",

border = "white", main="",xlab = "P-values",ylab="")

abline(h = 12900, col = "red", lwd = 3.0, lty = 3)

arrows(x0 = 0.025, y0=18100, x1 = .025, y1= 0, length = .0, col = "black", lwd = 3.0, lty = 3)

legend(0.018, 32000,c("p = .025"),bty="n")

## end fig 5

#

####### To judge the effect size in simulation 5:

install.packages(“pwr”)

library(pwr)

set.seed=123

pt = c() # To make an empty container for the results

# generated by the following lines.

for (i in 1:10^5) {

x<-rnorm(25,20,4.50)

y<-rnorm(25,22,4.00)

pt[i]<-t.test(x,y, var.equal=T)$p.val

}

length(p[p<.05])/100000 # This gives the apriori statistical power

pwr.t.test(n = 25, d = NULL, sig.level = 0.05, power = .37055, # Power calculator

type = "two.sample", alternative = "two.sided")

#

############## Figure 6

plot.new()

percent = c(78, 2, 4, 4, 11)

pval = c(.01, .02, .03, .04, .05)

dashed.percent = c(45, 19.5, 12.5, 11, 10.8)

dashed.pval = c(.01, .018, .03, .04, .05)

#

par(cex.lab = 1.2, type = "n")

plot(pval, percent, xlab = "p-value", ylab = "Percentage of Test Results", xlim= c(.01, .05), ylim = c(0,100), frame.plot = F)

lines(pval, percent, pval, type = "l", lty = 1, lwd = 2.5, col = "#3399FF")

lines(dashed.pval, dashed.percent, type = "l", lty = 2, lwd = 2.5, col = "forestgreen")

text(.01,78, "78%", pos = 3)

text(.02,2, "2%", pos = 2)

text(.03,4, "4%", pos = 3)

text(.04,4, "4%", pos = 3)

text(.05, 11, "11%", pos = 3)

abline(h = 20, col = "red", lty = 3, lwd =2)

#

############### Figure 6

plot.new()

percent = c(78, 2, 4, 4, 11)

pval = c(.01, .02, .03, .04, .05)

dashed.percent = c(45, 19.5, 12.5, 11, 10.8)

dashed.pval = c(.01, .018, .03, .04, .05)

#par(cex.lab = 1.2, type = "n")

plot(pval, percent, xlab = "p-value", ylab = "Percentage of Test Results", xlim= c(.01, .05), ylim = c(0,100), frame.plot = F)

lines(pval, percent, pval, type = "l", lty = 1, lwd = 2.5, col = "#3399FF")

lines(dashed.pval, dashed.percent, type = "l", lty = 2, lwd = 2.5, col = "forestgreen")

text(.01,78, "78%", pos = 3)

text(.02,2, "2%", pos = 2)

text(.03,4, "4%", pos = 3)

text(.04,4, "4%", pos = 3)

text(.05, 11, "11%", pos = 3)

abline(v = .02, col = "red", lty = 3, lwd =2)

abline(v = .03, col = "red", lty = 3, lwd =2)

abline(v = .04, col = "red", lty = 3, lwd =2)

abline(v = .05, col = "red", lty = 3, lwd =2)

abline(h = 18.5, col = "red", lty = 3, lwd =2)

abline(h = 13, col = "red", lty = 3, lwd =2)

abline(h = 11.5, col = "red", lty = 3, lwd =2)

abline(h = 11, col = "red", lty = 3, lwd =2)

####### R script for calculating statistics to submit to the p-curve app

install.packages("BSDA") # This line installs the BSDA package.

library(BSDA) # This line activates that package.

options (scipen = 999) # This line prevents use of scientific notation in the output

# The statistics below relate to Arpaci (2016 )

# In the line below, “x” & “y” refer to L1 & L2, respectively; "s" = “SD”.

#The numbers are descriptive statistics to be provided by the user

tsum.test(mean.x= 12.79, s.x=1.663, n.x=28, mean.y= 10.07, s.y=3.42, n.y=28)

# t = 3.7847, df = 39.092, p = .000517 # This is the output from the previous line

###### Figure 6

### Make the plot

plot.new()

percent = c(78, 2, 4, 4, 11)

pval = c(.01, .02, .03, .04, .05)

dashed.percent = c(45, 19.5, 12.5, 11, 10.8)

dashed.pval = c(.01, .018, .03, .04, .05)

# The plot

par(cex.lab = 1.2, bty = "n")

plot(pval, percent, xlab = "p-value", ylab = "Percentage of Test Results", xlim= c(.01, .05), ylim = c(0,100), frame.plot = F)

lines(pval, percent, pval, type = "l", lty = 1, lwd = 2.5, col = "#3399FF")

lines(dashed.pval, dashed.percent, type = "l", lty = 2, lwd = 2.5, col = "forestgreen")

text(.01,78, "78%", pos = 3)

text(.02,2, "2%", pos = 2)

text(.03,4, "4%", pos = 3)

text(.04,4, "4%", pos = 3)

text(.05, 11, "11%", pos = 3)

abline(h = 20, col = "red", lty = 3, lwd =2)

### end fig 6

#### Extra figure, not in the paper

#### Example of slight p hacking when H0 = True

############### Make an extra figure, not in the paper

## This figure shows an effect of one type of mild *p*-hacking, namely,

## the *p*-curve has a hump just below (to the left of) .05 and

## a dip just above (to the right of) .05.

## effect size = 0

set.seed=123

pfig3 = c()

for (i in 1:10^6) {

x<-rnorm(25,20,4.50); y<-rnorm(25,20,4.00)

t<-t.test(x,y, var.equal=T)$p.val

w<-wilcox.test(x,y)$p.val

pfig3[i] <- ifelse(test = t>= .05 & t<= .06, yes = ifelse(test = w < t , yes = w, n = t), no = t)

}

## Make the figure

windows(7, 4)

par(mgp = c(2, 0.5, 0), tck = -0.01, family = "serif", cex =1.1 , pin=c(5,2),bty="n")

hist(pfig3,breaks = 19, ylim = c(0, 120000), cex.lab= 1.1, col="lightgreen",

border = "white", main="",xlab = "P-values",ylab="Number of p-values")

abline(h = 50000, col = "red", lwd = 2, lty = 3)

arrows(x0 = 0.05, y0=76000, x1 = .05, y1= 53200, length = .1, col = "black", lwd = 2, lty = 1)

legend(-0.08, 100000,c("p = .05"),bty="n")

## end extra figure

# Zimmerman & Zumbo (2009) found that running Student's *t*-test and the WMW test on the

# same data and choosing the one giving the lowest *p* value raises the type I

# raises error rate to about .055 or perhaps.060.

# Zimmerman, D., & Zumbo, B. (2009). Hazards in choosing between pooled and

# separate-variances t tests. *Psicológica, 30*, 371–390.

# End