Pre-analysis Plans: An Early Stocktaking Replication of Descriptive Analysis

George K. Ofosu and Daniel N. Posner

January 25, 2021

1.Hypotheses

```
hdata <- select(pap_dat, clearhyp:pointoutnew,
    primary_secondary_hyp_maintained)
# names(hdata)

stargazer::stargazer(as.data.frame(hdata),
    title = "Hypotheses: clarity, quantity, and reporting",
    covariate.labels = c("PAP had clear hypotheses",
        "PAP specified more than one hypothesis",
        "Unclear whether PAP specified more than one hypothesis",
        "# hypotheses specified", "# hypotheses was unclear",
        "PAP distinguished primary verus secondary hypotheses",
        "# primary hypotheses", "# primary hypotheses unclear",
        "# primary hypotheses presented in main body published paper",
        "# primary hypotheses presented in main body of paper unclear",
        "# primary hypotheses presented in the appendix of paper",
        "# primary hypotheses presented in the appendix of paper (revised)",
        "# primary hypotheses presented in appendix unclear",
        "# primary hypotheses in article's main body supported",
        "Unclear: # hypotheses presented in main body of paper supported",
        "# primary hypotheses presented in the appendix supported",
        "# primary hypotheses presented in the appendix supported (revised)",
        "Unclear: # of primary hypotheses presented in appendix supported",
        "# of secondary hypotheses in main body of the paper",
        "# of secondary hypotheses in the appendix of paper",
        "New hypotheses reported in published article",
        "Author(s) acknowledge mew hypotheses",
        "Primary vs. secondary hypotheses distinction maintained in paper"),
    type = "text", summary.stat = c("n",
        "mean", "sd"))
```

```
##
## Hypotheses: clarity, quantity, and reporting
## ==========================================================================================
## Statistic                                                          N   Mean  St. Dev.
## ------------------------------------------------------------------------------------------
## PAP had clear hypotheses                                          195 0.900   0.300
## PAP specified more than one hypothesis                            191 0.960   0.200
## Unclear whether PAP specified more than one hypothesis            195 0.021   0.140
## # hypotheses specified                                            169 26.000  76.000
## # hypotheses was unclear                                          182 0.130   0.330
## PAP distinguished primary verus secondary hypotheses              191 0.430   0.500
## # primary hypotheses                                              84  12.000  20.000
```

```
## # primary hypotheses unclear                                          75  0.040   0.200
## # primary hypotheses presented in main body published paper           87  6.900   8.900
## # primary hypotheses presented in main body of paper unclear          86  0.023   0.150
## # primary hypotheses presented in the appendix of paper               54  0.370   1.200
## # primary hypotheses presented in the appendix of paper (revised)     52  0.390   1.200
## # primary hypotheses presented in appendix unclear                    57  0.088   0.280
## # primary hypotheses in article's main body supported                 75  3.300   4.600
## Unclear: # hypotheses presented in main body of paper supported       86  0.160   0.370
## # primary hypotheses presented in the appendix supported              12  4.200   9.100
## # primary hypotheses presented in the appendix supported (revised)    11  1.700   2.700
## Unclear: # of primary hypotheses presented in appendix supported      12  0.083   0.290
## # of secondary hypotheses in main body of the paper                   25  7.000   15.000
## # of secondary hypotheses in the appendix of paper                    16  1.500   2.900
## New hypotheses reported in published article                          93  0.180   0.390
## Author(s) acknowledge mew hypotheses                                  17  0.180   0.390
## Primary vs. secondary hypotheses distinction maintained in paper      37  0.920   0.280
## ------------------------------------------------------------------------------------
```

2. Clarity of variable definition

```
vdata <- pap_dat %>% select(cleardepvar,
    changevar, describechangedepvar,
    seconddepvarsclear, clearindvar,
    changeindvar, describechangeindvar,
    ncontrols, ncontrols_unclear)

stargazer::stargazer(as.data.frame(vdata),
    type = "text", summary.stat = c("n",
        "mean", "sd"), title = "Clarity of definition, changes, and reporting of variables",
    covariate.labels = c("Specified clear dependent variable (DV)",
        "Changed primary DV in main paper",
        "Acknowledged change in DV",
        "Specified clear dependent variable (DV)",
        "Specified clear treatment variable (IV)",
        "Changed independent variable in paper",
        "Pointed out change in IV",
        "# clear control variables",
        "# of control variables unclear"))
```

```
##
## Clarity of definition, changes, and reporting of variables
## ==========================================================
## Statistic                                  N  Mean  St. Dev.
## ----------------------------------------------------------
```

```
## Specified clear dependent variable (DV) 195 0.770  0.420
## Changed primary DV in main paper        73  0.068  0.250
## Acknowledged change in DV                7  0.290  0.490
## Specified clear dependent variable (DV) 34  0.650  0.480
## Specified clear treatment variable (IV) 195 0.930  0.250
## Changed independent variable in paper   91  0.044  0.210
## Pointed out change in IV                 4  0.000  0.000
## # clear control variables               109 4.100  5.200
## # of control variables unclear          180 0.440  0.500
## ------------------------------------------------------------
```

3. Sampling and power analysis

```
sdata <- pap_dat %>% select(popofinterest:poweranalysis)

stargazer::stargazer(as.data.frame(sdata),
    type = "text", summary.stat = c("n",
        "mean", "sd"), title = "Clarity of definition, changes, and reporting of variables",
    covariate.labels = c("Clearly specified the population of interest",
        "Clearly specified the sampling frame",
        "Clearly specified the sampling strategy",
        "Specified the conditions under which to exclude units",
        "Use power analysis to justify sample size"))
```

```
##
## Clarity of definition, changes, and reporting of variables
## ============================================================================
## Statistic                                                   N  Mean  St. Dev.
## ----------------------------------------------------------------------------
## Clearly specified the population of interest               195 0.770  0.420
## Clearly specified the sampling frame                       195 0.590  0.490
## Clearly specified the sampling strategy                    195 0.610  0.490
## Specified the conditions under which to exclude units      194 0.190  0.390
## Use power analysis to justify sample size                  195 0.280  0.450
## ----------------------------------------------------------------------------
```

4. Data collection

```
ddata <- pap_dat %>% select(treatmentundercontrol:imbalance)

stargazer::stargazer(as.data.frame(ddata),
    summary.stat = c("n", "mean", "sd"),
```

```
    type = "text", covariate.labels = c("Treatment assignment controlled by author(s)",
        "Unclear whether author(s) controlled  treatment assignment",
        "Specified randomization procedure",
        "Specified manipulation checks (balance table)",
        "Specified how to deal with covariate imbalance"))
```

```
##
## ===============================================================================
## Statistic                                                      N  Mean  St. Dev.
## -------------------------------------------------------------------------------
## Treatment assignment controlled by author(s)                  191 0.900  0.300
## Unclear whether author(s) controlled treatment assignment     183 0.011  0.100
## Specified randomization procedure                             189 0.600  0.490
## Specified manipulation checks (balance table)                 191 0.240  0.420
## Specified how to deal with covariate imbalance                179 0.200  0.400
## -------------------------------------------------------------------------------
```

5. Inclusion and exclusion rules

```
inexdata <- pap_dat %>% select(rulesmissing:followноncomp)


stargazer::stargazer(as.data.frame(inexdata),
    summary.stat = c("n", "mean", "sd"),
    type = "text", covariate.labels = c("Specified the rules dealing with missingness/attrition",
        "Specified the rules for dealing with outliers",
        "Specified the rules for dealing with noncompliance",
        "Followed rules for dealing with missingness and attrition",
        "Followed rules for dealing with outliers",
        "Followed rules for dealing with noncompliance"))
```

```
##
## ===============================================================================
## Statistic                                                      N  Mean  St. Dev.
## -------------------------------------------------------------------------------
## Specified the rules dealing with missingness/attrition        195 0.250  0.430
## Specified the rules for dealing with outliers                 195 0.077  0.270
## Specified the rules for dealing with noncompliance            175 0.130  0.330
## Followed rules for dealing with missingness and attrition      26 1.000  0.000
## Followed rules for dealing with outliers                        7 1.000  0.000
## Followed rules for dealing with noncompliance                  10 1.000  0.000
## -------------------------------------------------------------------------------
```

6. Statistical model specification

```
statdata <- pap_dat %>% select(statmodel:covars) %>%
    select(-statdevdescr)


stargazer::stargazer(as.data.frame(statdata),
    summary.stat = c("n", "mean", "sd"),
    type = "text", covariate.labels = c("Specified a clear statistical model",
        "Deviated from specified statistical model",
        "Pointed out deviation from statistical model",
        "Specified how to estimate standard errors",
        "Specified how to adjust multiple hypotheses testing",
        "Researcher commit to simple difference in means test",
        "Specified direction of hypotheses testing",
        "Specified whether and how control variables  will be included in analysis"))
```

```
##
## ================================================================================================
## Statistic                                                                       N  Mean  St. Dev.
## ------------------------------------------------------------------------------------------------
## Specified a clear statistical model                                           194 0.680  0.470
## Deviated from specified statistical model                                     73  0.190  0.400
## Pointed out deviation from statistical model                                  14  0.071  0.270
## Specified how to estimate standard errors                                     184 0.370  0.480
## Specified how to adjust multiple hypotheses testing                           184 0.230  0.420
## Researcher commit to simple difference in means test                          195 0.180  0.390
## Specified direction of hypotheses testing                                     36  0.360  0.490
## Specified whether and how control variables will be included in analysis 44  0.640  0.490
## ------------------------------------------------------------------------------------------------
```

7. Other features

```
otherdata <- pap_dat %>% select(updated:pages)


names(otherdata)
```

```
## [1] "updated"        "howmanyupdates" "clearchanges"   "irb"
## [5] "sop"            "pages"
```

```
stargazer::stargazer(as.data.frame(otherdata),
    summary.stat = c("n", "mean", "sd"),
    type = "text", covariate.labels = c("Plan was updated",
        "If so, number of times updated",
        "Changes were clearly marked",
        "Study had IRB approval", "Study pre-specify how to deal with unexpected events (e.g., SOP)",
        "Pages (in single-spaced pages)"))
```

```
##
## ==================================================================================
## Statistic                                                          N   Mean  St. Dev.
## ----------------------------------------------------------------------------------
## Plan was updated                                                 195 0.120   0.340
## If so, number of times updated                                   53  0.980   1.700
## Changes were clearly marked                                      22  0.730   0.460
## Study had IRB approval                                           193 0.680   0.470
## Study pre-specify how to deal with unexpected events (e.g., SOP) 194 0.031   0.170
## Pages (in single-spaced pages)                                   194 15.000  16.000
## ----------------------------------------------------------------------------------
```

8. Computations

a. new hypotheses restricted to whether hypotheses were clearly specified

```
mean(pap_dat$newhyp[pap_dat$clearhyp ==
    1], na.rm = T)
```

```
## [1] 0.18
```

```
mean(pap_dat$pointoutnew[pap_dat$clearhyp ==
    1 & pap_dat$newhyp == 1], na.rm = T)
```

```
## [1] 0.2
```

```
table(pap_dat$pointoutnew[pap_dat$clearhyp ==
    1 & pap_dat$newhyp == 1])
```

```
##
##  0  1
## 12  3
```

b. categorizing the number of hypotheses specified

```
pap_dat <- mutate(pap_dat, cat_nhyp = ifelse(howmanyhyp <=
    5, 1, ifelse(howmanyhyp > 5 & howmanyhyp <=
    10, 2, ifelse(howmanyhyp > 10 &
    howmanyhyp <= 20, 3, ifelse(howmanyhyp >
    20 & howmanyhyp <= 50, 4, ifelse(howmanyhyp >
    50, 5, NA))))), cat_nhyp_dummy = ifelse(cat_nhyp ==
    1, 1, ifelse(cat_nhyp > 1, 2, NA)),
    cat_nprimaryhyp = ifelse(howmanyprimary <=
        5, 1, ifelse(howmanyprimary >
        5 & howmanyprimary <= 10, 2,
        ifelse(howmanyprimary > 10 &
            howmanyprimary <= 20, 3,
            ifelse(howmanyprimary >
                20 & howmanyprimary <=
                50, 4, ifelse(howmanyprimary >
                50, 5, NA))))), pages_cat = ifelse(pages <=
        5, 1, ifelse(pages > 5 & pages <=
        10, 2, ifelse(pages > 10 & pages <=
        15, 3, ifelse(pages > 15 & pages <=
        20, 4, ifelse(pages > 20 & pages <=
        25, 5, ifelse(pages > 25 & pages <=
        30, 6, ifelse(pages > 30 & pages <=
        40, 7, ifelse(pages > 40, 8,
        NA)))))))))

# prop.table(table(pap_dat$cat_nhyp))

# tab1 <- table(pap_dat$cat_nhyp)

# number of hypotheses



p1 <- ggplot(filter(pap_dat, !is.na(cat_nhyp)),
    aes(x = factor(cat_nhyp), group = 1)) +
    geom_bar(aes(y = (..count..)/sum(..count..))) +
    scale_y_continuous(labels = scales::percent,
        limits = c(0, 0.45)) + scale_x_discrete(labels = c("1-5",
    "6-10", "11-20", "21-50", "50+")) +
    geom_text(aes(label = str_c("(",
        ..count.., ")", sep = ""), y = ..prop..),
```

```
        stat = "count", vjust = -0.5,
        nudge_x = -0.1, size = 6) +
    geom_text(aes(label = scales::percent(..prop..),
        y = ..prop..), stat = "count",
        vjust = -0.5, nudge_x = 0.3,
        size = 6) + labs(x = "Number of hypotheses specified",
    y = "Percentage of studies", title = "Panel A") +
    theme_tufte() + theme(axis.title = element_text(face = "bold",
    size = 24, colour = "black"), legend.title = element_text(face = "bold"),
    legend.key = element_rect(fill = "white",
        color = "white"), axis.ticks = element_line(size = 2),
    panel.background = element_blank(),
    axis.line = element_line(size = 1,
        linetype = "solid"), title = element_text(family = "serif",
        size = 13, colour = "black",
        face = "bold"), axis.text.y = element_text(family = "serif",
        size = 18, colour = "black",
        face = "bold"), axis.text.x = element_text(family = "serif",
        size = 18, colour = "black",
        face = "bold"), strip.text = element_text(size = 12,
        face = "bold"), legend.text = element_text(family = "serif",
        size = 8, colour = "black"),
    panel.grid.minor = element_line(colour = "grey",
        size = 0.5))



p2 <- ggplot(filter(pap_dat, !is.na(cat_nprimaryhyp) &
    cat_nhyp_dummy == 2 & primary ==
    1), aes(x = factor(cat_nprimaryhyp),
    group = 1)) + geom_bar(aes(y = (..count..)/sum(..count..))) +
    scale_y_continuous(labels = scales::percent,
        limits = c(0, 0.45)) + scale_x_discrete(labels = c("1-5",
    "6-10", "11-20", "21-50", "50+")) +
    geom_text(aes(label = str_c("(",
        ..count.., ")", sep = ""), y = ..prop..),
        stat = "count", vjust = -0.5,
        nudge_x = -0.1, size = 6) +
    geom_text(aes(label = scales::percent(..prop..),
        y = ..prop..), stat = "count",
        vjust = -0.5, nudge_x = 0.3,
        size = 6) + labs(x = "Number of primary hypotheses specified",
    y = "Studies with more than 5 hypotheses that distinguished \n primary from secondary hypotheses",
```

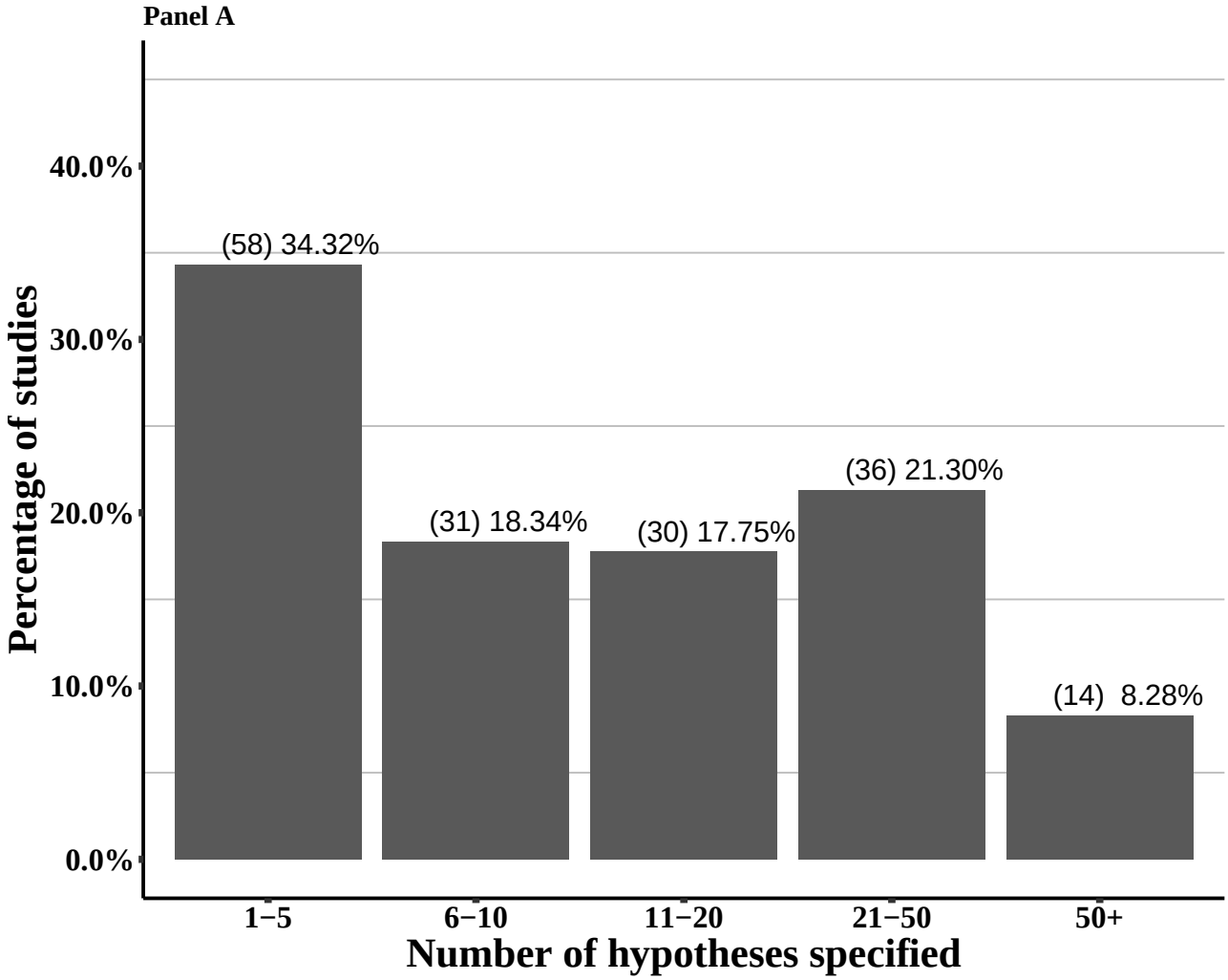
```
    title = "Panel B") + theme_tufte() +
    theme(axis.title = element_text(face = "bold",
        size = 24, colour = "black"),
        legend.title = element_text(face = "bold"),
        legend.key = element_rect(fill = "white",
            color = "white"), axis.ticks = element_line(size = 2),
        panel.background = element_blank(),
        axis.line = element_line(size = 1,
            linetype = "solid"), title = element_text(family = "serif",
            size = 13, colour = "black",
            face = "bold"), axis.text.y = element_text(family = "serif",
            size = 18, colour = "black",
            face = "bold"), axis.text.x = element_text(family = "serif",
            size = 18, colour = "black",
            face = "bold"), strip.text = element_text(size = 12,
            face = "bold"), legend.text = element_text(family = "serif",
            size = 8, colour = "black"),
        panel.grid.minor = element_line(colour = "grey",
            size = 0.5))




# ggsave(filename =
# 'num_hyp_twoparts.pdf',arrangeGrob(p1,
# p2,nrow = 1,ncol = 2),
# width=20,height = 12)
```

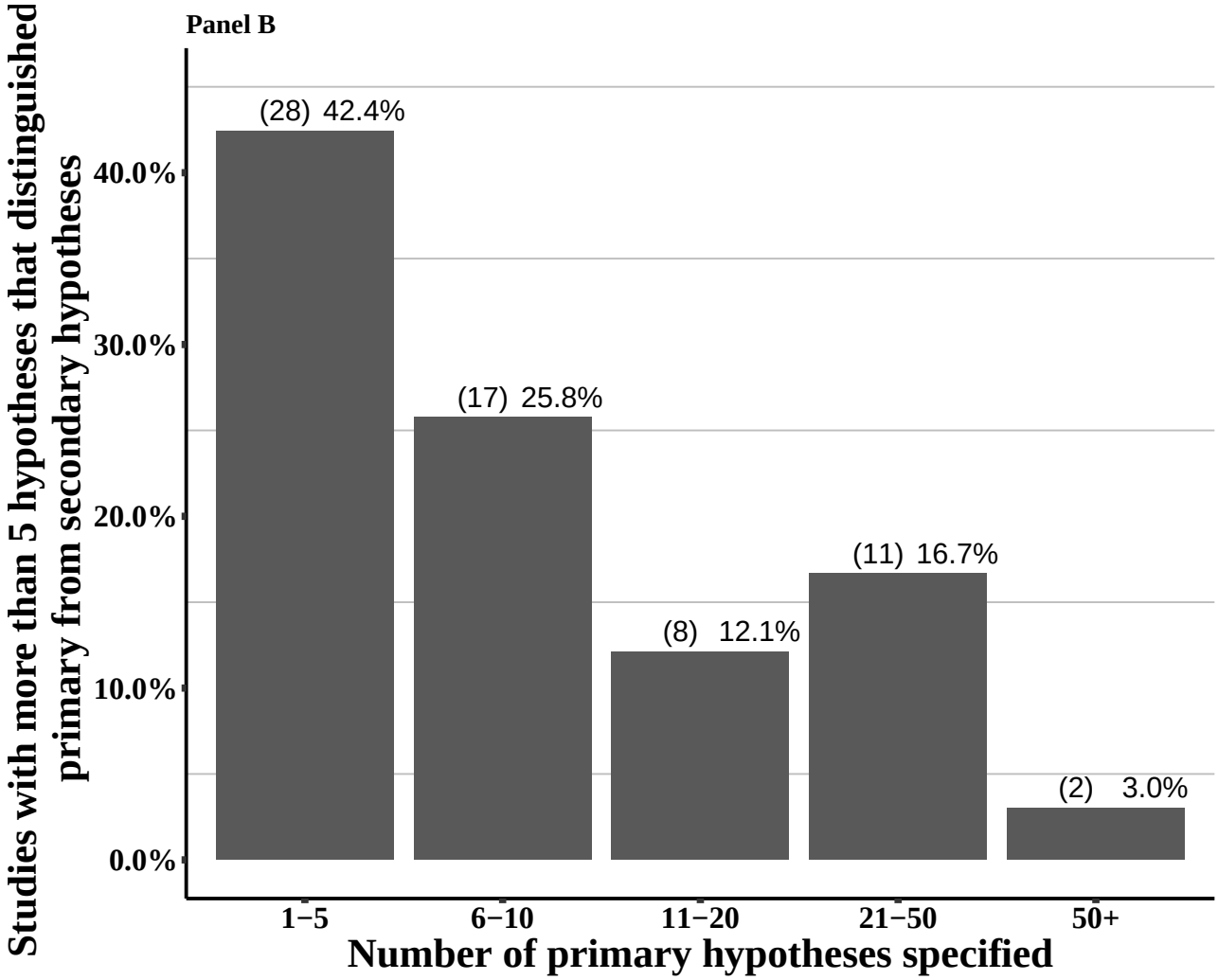
```
p1
```

```
p2
```

c. Proportion that distinguish between primary and secondary hypotheses by number of hypotheses

```
## number of specified hypotheses by
## whether author distinguished
## primary vs. secondary hypotheses

prop.table(table(pap_dat$cat_nhyp, pap_dat$primary),
    1)
```


Panel A
40.0%
30.0%
20.0%
10.0%
0.0%
Percentage of studies
(58) 34.32%
(31) 18.34%
(30) 17.75%
(36) 21.30%
(14) 8.28%
1−5
6−10
11−20
21−50
50+
Number of hypotheses specified


Panel B
Studies with more than 5 hypotheses that distinguished primary from secondary hypotheses
40.0%
30.0%
20.0%
10.0%
0.0%
(28) 42.4%
(17) 25.8%
(8) 12.1%
(11) 16.7%
(2) 3.0%
1−5
6−10
11−20
21−50
50+
Number of primary hypotheses specified

```
##
##         0     1
##   1 0.793 0.207
##   2 0.452 0.548
##   3 0.567 0.433
##   4 0.333 0.667
##   5 0.071 0.929
```

```
prop.table(table(pap_dat$cat_nhyp_dummy,
    pap_dat$primary), 1)
```

```
##
##        0    1
##   1 0.79 0.21
##   2 0.40 0.60
```

```
##
summary(pap_dat$howmanyprimary[pap_dat$cat_nhyp_dummy ==
    2 & pap_dat$primary == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       1       4       6      13      15     122       9
```

```
tab2 <- table(pap_dat$cat_nprimaryhyp[pap_dat$cat_nhyp_dummy ==
    2 & pap_dat$primary == 1])

prop.table(table(pap_dat$cat_nprimaryhyp[pap_dat$cat_nhyp_dummy ==
    2 & pap_dat$primary == 1]))
```

```
##
##    1    2    3    4    5
## 0.42 0.26 0.12 0.17 0.03
```

```
# stargazer::stargazer(as.matrix(tab2))
```

d. Proportion of researchers who distinquished primary and secondary hypotheses who maintained such distinction in the resulting paper

```
## restricted to PAPs specifying more
## than 5 primary hypotheses
prop.table(table(pap_dat$primary_secondary_hyp_maintained[pap_dat$cat_nhyp_dummy ==
    2 & pap_dat$primary == 1]))
```

```
##
##     0     1
## 0.097 0.903
```

```
### proportion who maintained
### distinction in main paper
prop.table(table(pap_dat$primary_secondary_hyp_maintained))
```

```
##
##     0     1
## 0.081 0.919
```

f. multiple hypotheses testing among PAPs that specify more that 5 hypotheses

```
prop.table(table(pap_dat$cat_nhyp_dummy,
    pap_dat$multipletest), 1)
```

```
##
##        0    1
##   1 0.89 0.11
##   2 0.71 0.29
```

g. Reporting of specified hypotheses

```
pap_dat <- mutate(pap_dat, presentedprimaryappendix_rev0 = ifelse(is.na(presentedprimaryappendix),
    0, presentedprimaryappendix), primaryreported_perc = presentedpimarymainbody/howmanyprimary,
    primaryreported_perc_dum = ifelse(primaryreported_perc <
        1, 0, ifelse(primaryreported_perc ==
        1, 1, NA)), primarymain_appendix = presentedpimarymainbody +
        presentedprimaryappendix_rev0,
    total_primary_reported = ifelse(primarymain_appendix ==
        howmanyprimary, 1, ifelse(primarymain_appendix <
        howmanyprimary, 0, NA)), howmanyprimary_rev = ifelse(primary ==
        0 & is.na(howmanyprimary), howmanyhyp,
        howmanyprimary), primaryreported_perc2 = presentedpimarymainbody/howmanyprimary_rev,
    primaryreported_perc_dum2 = ifelse(primaryreported_perc2 <
        1, 0, ifelse(primaryreported_perc2 ==
        1, 1, NA)), primarymain_appendix2 = presentedpimarymainbody +
        presentedprimaryappendix_rev0,
    total_primary_reported2 = ifelse(primarymain_appendix ==
        howmanyprimary_rev, 1, ifelse(primarymain_appendix <
        howmanyprimary_rev, 0, NA)),
```

```
    presentedpimarymainbody_rev = ifelse(is.na(presentedpimarymainbody),
        0, presentedpimarymainbody),
    presentedsecondarymainbody_rev = ifelse(is.na(presentedsecondarymainbody),
        0, presentedsecondarymainbody),
    presentedsecondaryappendix_rev = ifelse(is.na(presentedsecondaryappendix),
        0, presentedsecondaryappendix),
    total_hyp_reported = presentedpimarymainbody_rev +
        presentedprimaryappendix_rev0 +
        presentedsecondarymainbody_rev +
        presentedsecondaryappendix_rev,
    total_hyp_reported_dum = ifelse(total_hyp_reported ==
        howmanyhyp, 1, ifelse(total_hyp_reported <
        howmanyhyp, 0, NA)), prop_reported = ifelse(publication ==
        "PAP with publication", total_hyp_reported/howmanyhyp,
        NA))

## proportion of papers reporting
## primary hypotheses reported in
## main paper (excluding appendix)
prop.table(table(pap_dat$primaryreported_perc_dum[pap_dat$publication ==
    "PAP with publication"]))
```

```
##
##    0    1
## 0.41 0.59
```

```
## proportion of papers reporting all
## pre-specified primary hypotheses
## in main and appendix of paper
prop.table(table(pap_dat$total_primary_reported2[pap_dat$publication ==
    "PAP with publication"]))
```

```
##
##    0    1
## 0.39 0.61
```

```
## Proportion of paper reporting less
## than pre-specified total
## hypotheses
prop.table(table(pap_dat$total_hyp_reported_dum[pap_dat$publication ==
    "PAP with publication"]))
```

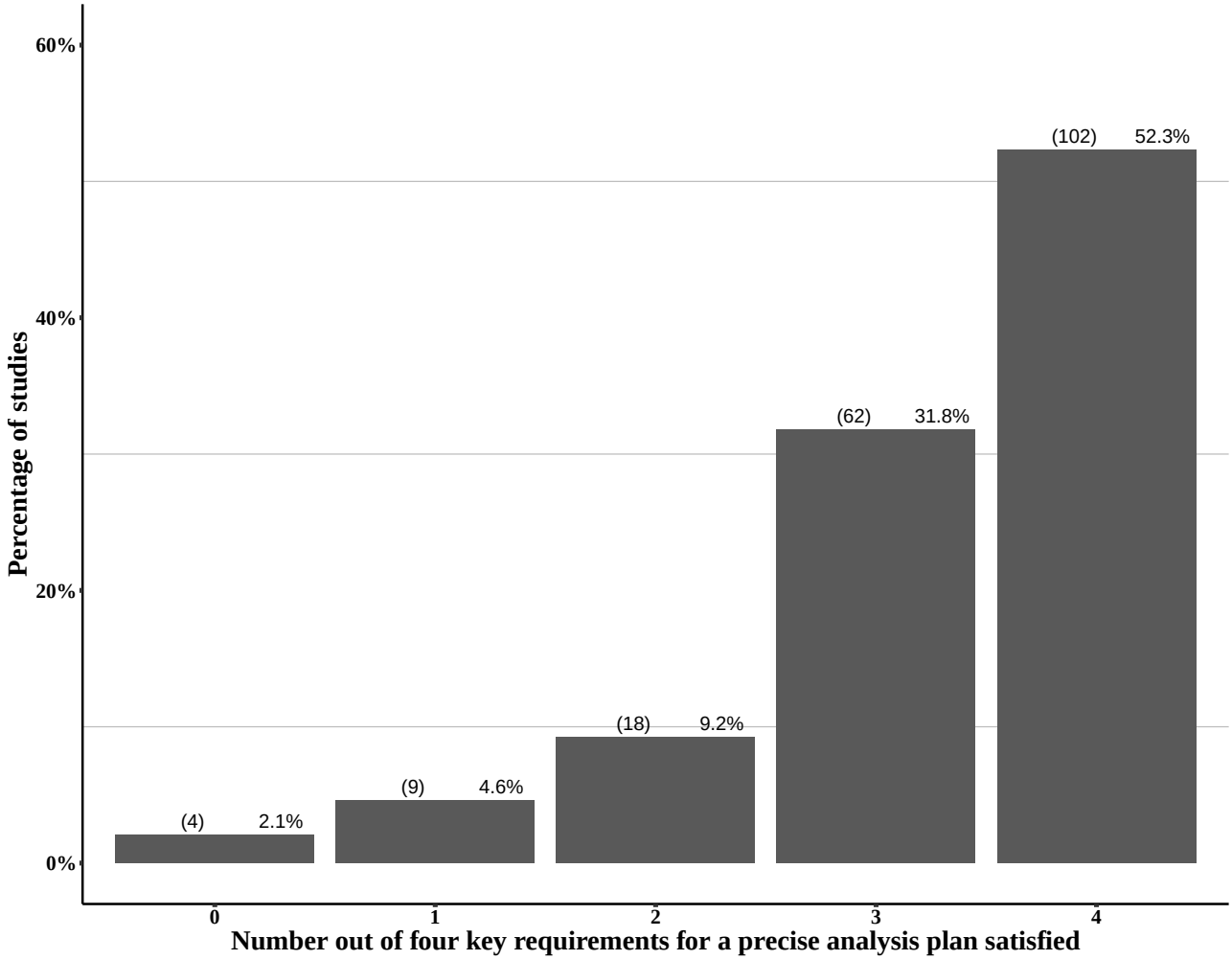
```
##
##    0    1
## 0.62 0.38
```

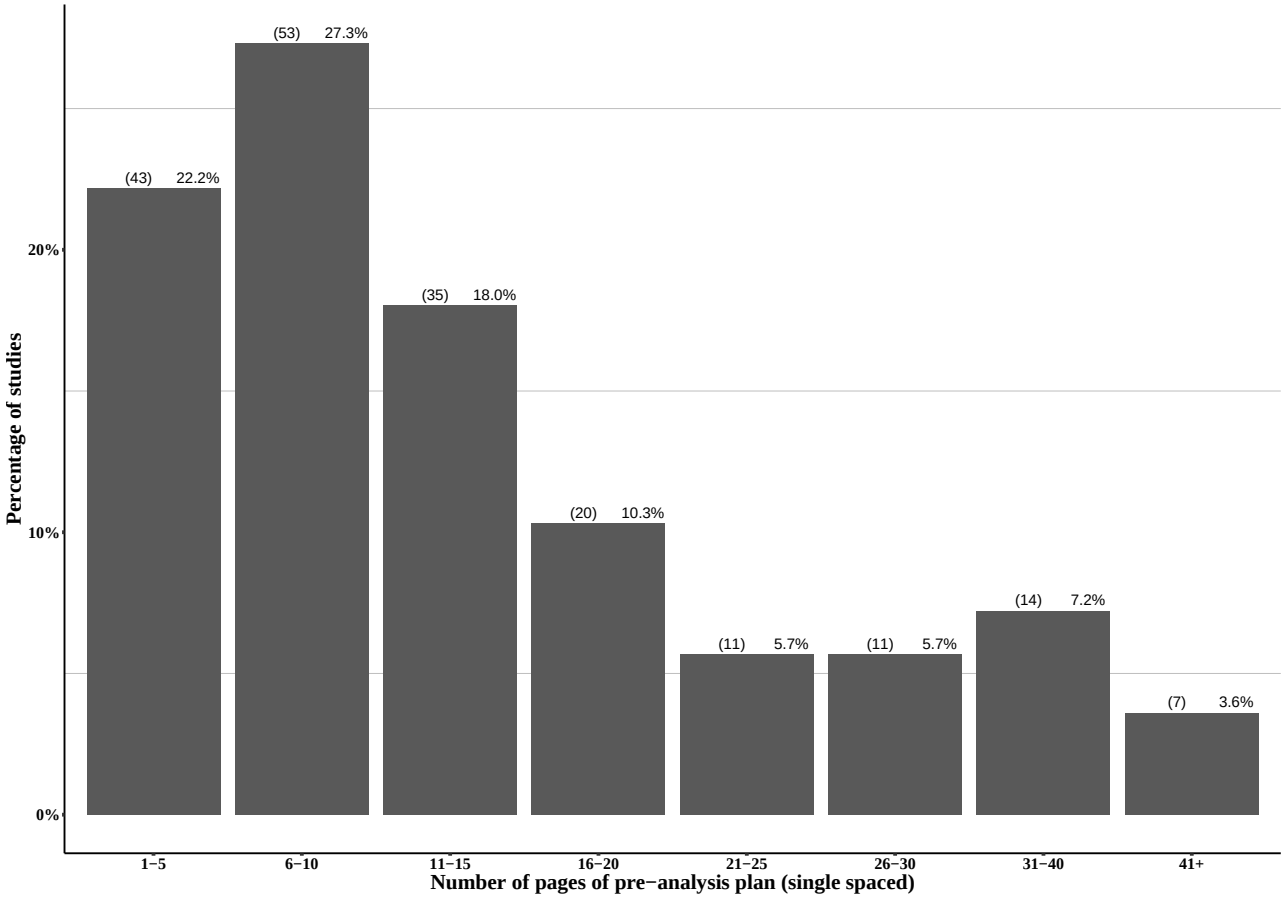
```
## Median proportion of total
## pre-specified hypotheses reported

median(pap_dat$prop_reported, na.rm = T)
```

```
## [1] 0.75
```

Number and share of PAPs satisfying the four key requirements of a complete

```
##
## 2011-2015      2016
##       107        88
```



i.number of pages

```
##
##  1  2  3  4  5  6  7  8
## 43 53 35 20 11 11 14  7
```