Supplement to 'How to reduce Item Nonresponse in Face-to-Face Surveys? A Review and Evidence from the European Social Survey'

This supplementary material provides the interested reader with more detail on the data, data manipulation, descriptive statistics, and choices made in the analysis. This serves to keep the original text short and still provide full transparency on data, code, and analysis.

Version and Packages

This analysis was written for R version 4.2.2 (2022-10-31) on Linux. The following packages were used with the respective versions as comments.

```
library(tidyverse) # 1.3.1
library(haven) # 2.5.2
library(qqplotr) # 0.0.6
library(fixest) # 0.10.4
library(kableExtra) # 1.3.4
```

Data

Following the literature review and synthesis in the article, I want to test the effectiveness of some potential strategies to reduce item nonresponse. I use the *European Social Survey (ESS)* Round 9 collected between August 2018 and January 2020. The ESS is a biannual face-to-face trend survey on attitudes and beliefs towards social and political topics in Europe established in 2001. In each country and round, the ESS draws a new random sample of the residential population of 15 years and older aiming for a minimum response rate of 70%. Most countries use computer-assisted personal interviews for data collection and the questionnaire is designed to take about one hour. The data release 3.1 includes data from 49,519 respondents from 29 countries (Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Montenegro, Netherlands, Norway, Poland, Portugal, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, and the United Kingdom). For more information on the data, visit europeansocialsurvey.org. I will provide specific information about ESS policies and design in this supplement when needed for the variables discussed.

Data Manipulation

`ESS9e03_1.dta` is the main data file from the ESS 9 version 3.1 and `ESS9INTe03.dta` is the interviewer questionnaire. The former contains predominantly the answers of the respondents but also weights, automatic time stamps, and other metadata. The latter contains information about the interviewer and the answers of the interviewer to a short post-interview questionnaire. The data are available after registration here: https://www.europeansocialsurvey.org/data/download.html?r=9

I combine the two data sets into a single one by matching the ID of the respondent (*idno*) and the country abbreviation (*cntry*).

It is necessary to change the encoding on Linux/Mac to *latin1* to import the data correctly, remove the option if on Windows (see https://haven.tidyverse.org/reference/read_dta.html#character-encoding).

```
ESS9 <- full_join(read_dta("ESS9e03_1.dta", encoding = "latin1"),
                  read_dta("ESS9INTe03.dta", encoding = "latin1"),
                  by = c("idno", "cntry"))
```

Dependent Variables

I use three dependent variables: the number of refusals by a respondent, the number of Don't Know (DK), and their sum. I take the sum of all questions applicable to all respondents. That means, all questions only asked a subset of respondents due to previous answers or survey experiments are left out.

As background information, refusal and DK are not read out to the respondents in the ESS and are not shown on the showcards. But interviewers have them as distinct options within their CAPI instrument. Interviewers are advised to accept them without probing. Whether the respondent refuses or says DK is therefore to some extent based on the interpretation of the interviewer.

The command `select(nwspol:impfun)` selects only the variables that are based on questions presented to the respondent and gets rid of the interviewer questionnaire, weights, etc. for the calculation of item nonresponse.

The ESS distinguishes different types of missings using labeled missings in Stata. Their codes are:

- .a - not applicable (due to filtering, missing by design)
- .b - respondent refused
- .c - Don't know
- .d - not available (should not exist: code for processing errors)

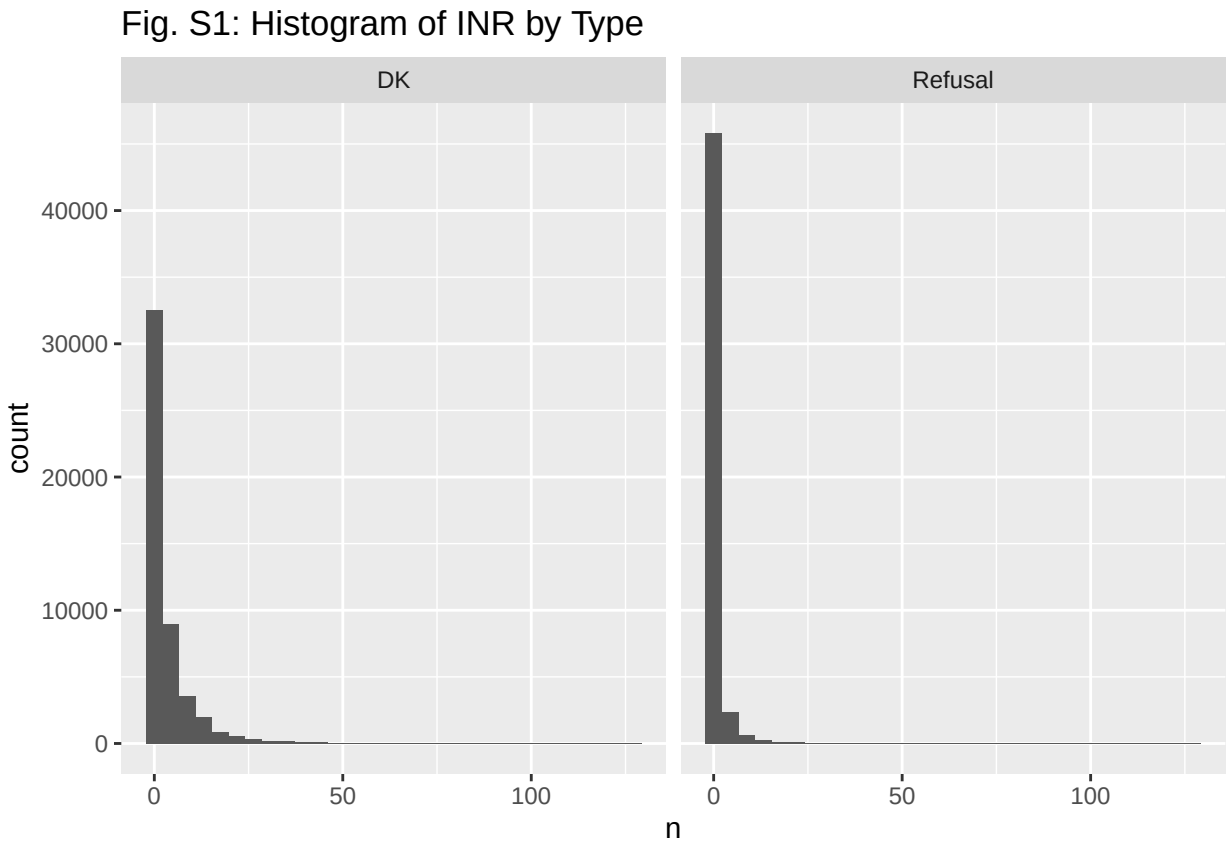
The package *haven* has a function to detect these tagged missings in a Stata dataset in R, see https://haven.tidyverse.org/articles/semantics.html#tagged-missing-values-1

To have a comparable set of items for all respondents, I only select the ones applicable to all respondents and exclude items that are affected by filtering questions: `select(where(~ any(is_tagged_na(.x, tag = "a")) == FALSE))`

Attention: the step to calculate the number of missings for each respondent takes some time.

```
count_refusals <- ESS9 %>%
  select(nwspol:impfun) %>%
  select(where(~ any(is_tagged_na(.x, tag = "a")) == FALSE)) %>%
  mutate(across(.fns = ~ is_tagged_na(.x, tag = "b"))) %>%
  rowwise() %>%
  transmute(n_refusals = sum(c_across()))

count_dk <- ESS9 %>%
  select(nwspol:impfun) %>%
  select(where(~ any(is_tagged_na(.x, tag = "a")) == FALSE)) %>%
  mutate(across(.fns = ~ is_tagged_na(.x, tag = "c"))) %>%
  rowwise() %>%
  transmute(n_dk = sum(c_across()))

rbind(tibble(n = count_refusals$n_refusals,
             type = "Refusal"),
      tibble(n = count_dk$n_dk,
             type = "DK")) %>%
  ggplot() +
  aes(n) +
  geom_histogram() +
```

```
  facet_wrap(~type) +
  labs(title = "Fig. S1: Histogram of INR by Type")
```



There are much more DKs than refusals. There are 176490 DKs, on average 3.564 per respondent. The number of refusals amounts to 41821, on average 0.845 per respondent. That totals 218311 cases of item nonresponse. There is a correlation of 0.226 between the number of DKs and refusals.

The standard deviation of DK is 7.194 and the standard deviation of refusals is 3.317. Although ultimately the variance of the residuals matters for the regression, this is a first sign of overdispersion.

Independent Variables

Number of Applicable Items

I hypothesize that longer questionnaires generate proportionally more item nonresponse because respondents' motivation decreases over time and respondents might get tired/exhausted over time reducing their cognitive ability. To test this, I generate a variable that measures how many items apply to the respondent based on the missing code .a. The more items that are not applicable, the fewer questions the respondent is asked.

For better interpretation, I do not use the number of not-applicable items but the number of applicable items. To get this, I subtract the number of not-applicable items from the number of total items. The number of total items is 545.

```
count_applicable <- ESS9 %>%
  select(nwspol:impfun) %>%
  mutate(across(.fns = ~ is_tagged_na(.x, tag = "a"))) %>%
  rowwise() %>%
  transmute(n_applicable = 545 - sum(c_across()))
```

Preparation for Analysis

The variables from the interviewer questionnaire still contain missing values coded as 8 and 9. They are set to NA if their value is above 6. Bulgaria used the wrong format for the interviewer's age and therefore adds many missings to it and is resultantly excluded. The *other* category of ISCED (number code 55) is coded to NA as well. The length of the interviews is top coded to 3 hours because there were errors with the CAPI instrument time stamps on some occasions (see ESS 9 codebook).

Out of an abundance of caution, I merged the interviewer ID with the country abbreviation to make sure that there are no wrong fixed effects if two interviewers have the same ID in different countries although this should not exist.

```
analysis <- ESS9 %>%
  zap_label() %>%
  zap_missing() %>%
  zap_formats() %>% # getting rid of Stata formatting
  transmute(age = agea, # respondent
            age_sq = agea^2,
            educ = ifelse(eisced < 10, eisced, NA),
            duration = ifelse(inwtm < 180, inwtm / 30, 6),
            n_applicable = count_applicable$n_applicable / 10,
            interference = ifelse(preintf < 6, preintf * -1 + 2, NA),
            notprimlang = as.integer(lnghom1 != intlnga),
            gndrmatch = ifelse(intgndr < 6, as.numeric(intgndr == gndr), NA),
            resolder = ifelse(intagea < 200, as.numeric(agea - intagea > 10), NA),
            intolder = ifelse(intagea < 200, as.numeric(intagea - agea > 10), NA),
            showcards = ifelse(resswcd < 6, -1*resswcd + 4, NA),
            clarif = ifelse(resclq < 6, resrelq, NA),
            bestab = ifelse(resbab < 6, resbab, NA),
            underst = ifelse(resundq < 6, resundq, NA),
            intver = paste0(intnum, cntry), # FE
            n_dk = count_dk$n_dk, # DV
            n_ref = count_refusals$n_refusals,
            n_tot = n_dk + n_ref)

# somehow, I had trouble recoding clarif in transmute
analysis$clarif[analysis$clarif > 6] <- NA

labels <- c(age = "Age",
            age_sq = "Age squared",
            educ = "Education (ISCED)",
            underst = "Understood Questions",
            bestab = "Answered to best Ability",
            n_applicable = "Number of applicable Items (10 Items)",
            duration = "Duration of Interview (30 Min)",
            interference = "Interference of Interview",
            notprimlang = "Int. not in primary Language",
            gndrmatch = "Gender Matching",
            resolder = "Respondent 10 years older",
            intolder = "Interviewer 10 years older",
            showcards = "Use of Showcards",
            clarif = "Amount of Clarifications",
            intver = "Interviewer",
            n_dk = "Don't know",
            n_ref = "Refusal",
```

```
            n_tot = "Total")
```

Descriptive Statistics

Variables of Interest

Except of the use of showcards and the number of applicable items, all variables of interest are dummys:

The proportion of interviews where the interviewer is more than 10 years older is 0.366 and the proportion where the respondent is more than 10 years older is 0.295. Both are based of the same original variables, age of the respondent and age of the interviewer and have therefore the same number of NAs: 4045. The large number of NAs is predominantly due to missings in interviewers age. It is missing very frequently. Romania used the wrong format for this item and therefore all interviews from Romania are excluded via missingness in these items.

The proportion of interviews where respondent and interviewer were of the same gender is 0.534 (4045 NAs).

In 8.032 percent of the interviews, someone besides the interviewer and respondent was present in the same room or interfered with the interview (65 NAs).

9.693 percent of the interviews were conducted in a language different from the one the respondent primarily speaks at home. There are no missing data in this item.

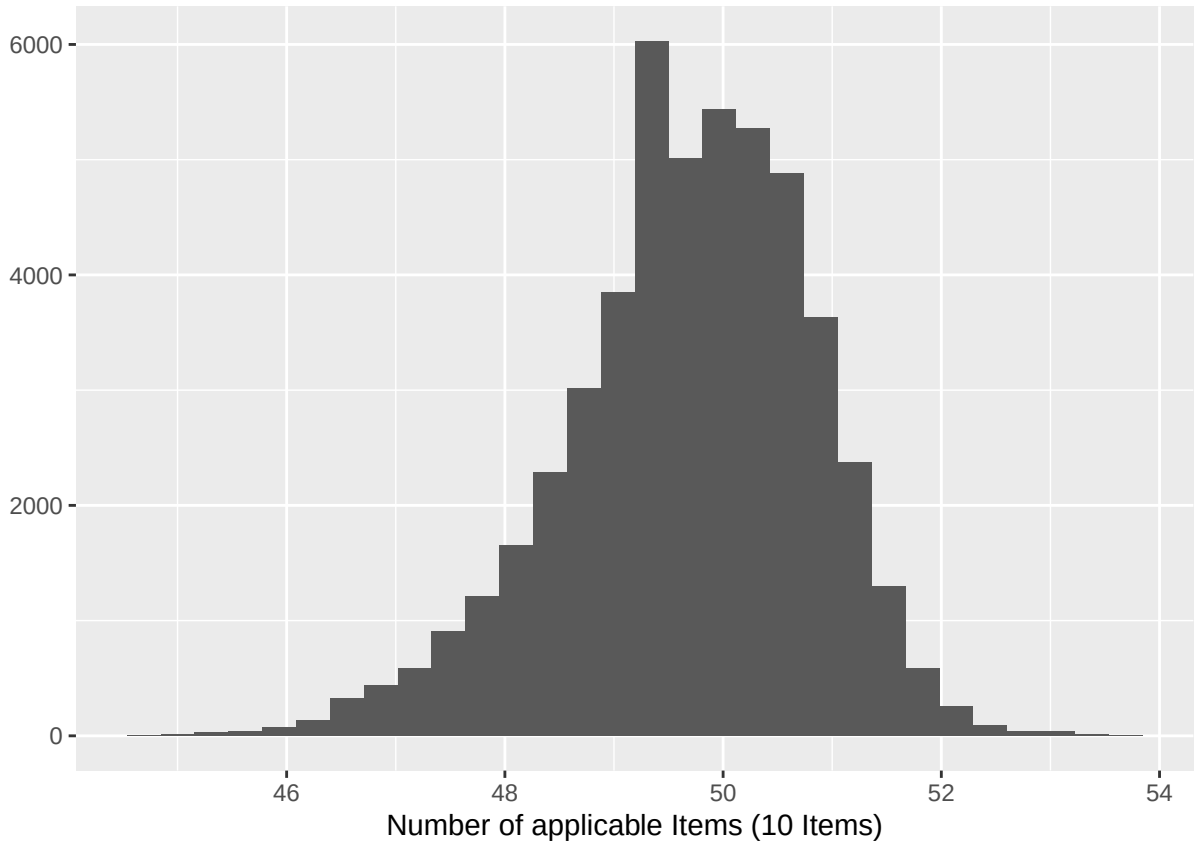
The extent the respondent used showcards as perceived by the interviewer is measured on a three point scale: respondent used all the applicable showcards, respondent used only some applicable showcards, respondent refused/ was unable to use the showcards at all. This question is asked to the interviewer after the interview, see the ESS9 source questionnaire. I flipped the order to ease interpretation: higher number represents more frequent use. Most respondents used all the applicable showcards (0.792), 0.163 respondents used them only sometimes and 0.045 never used them. There are 129 NAs.

The number of applicable items is described below. A potential bias in the check of the mechanism that longer questionnaires increase INR because respondents get tired would be if the respondents that are asked more or less questions differ systematically in their ability as this could confound the expected decrease in motivation. The scatter plot of the number of applicable items and education as a proxy for ability serves to check that. It does not seem to be the case that a different number of applicable questions is related to ability and differences in other respondent characteristics should not affect the outcome according to the theoretical model outlined in the article.

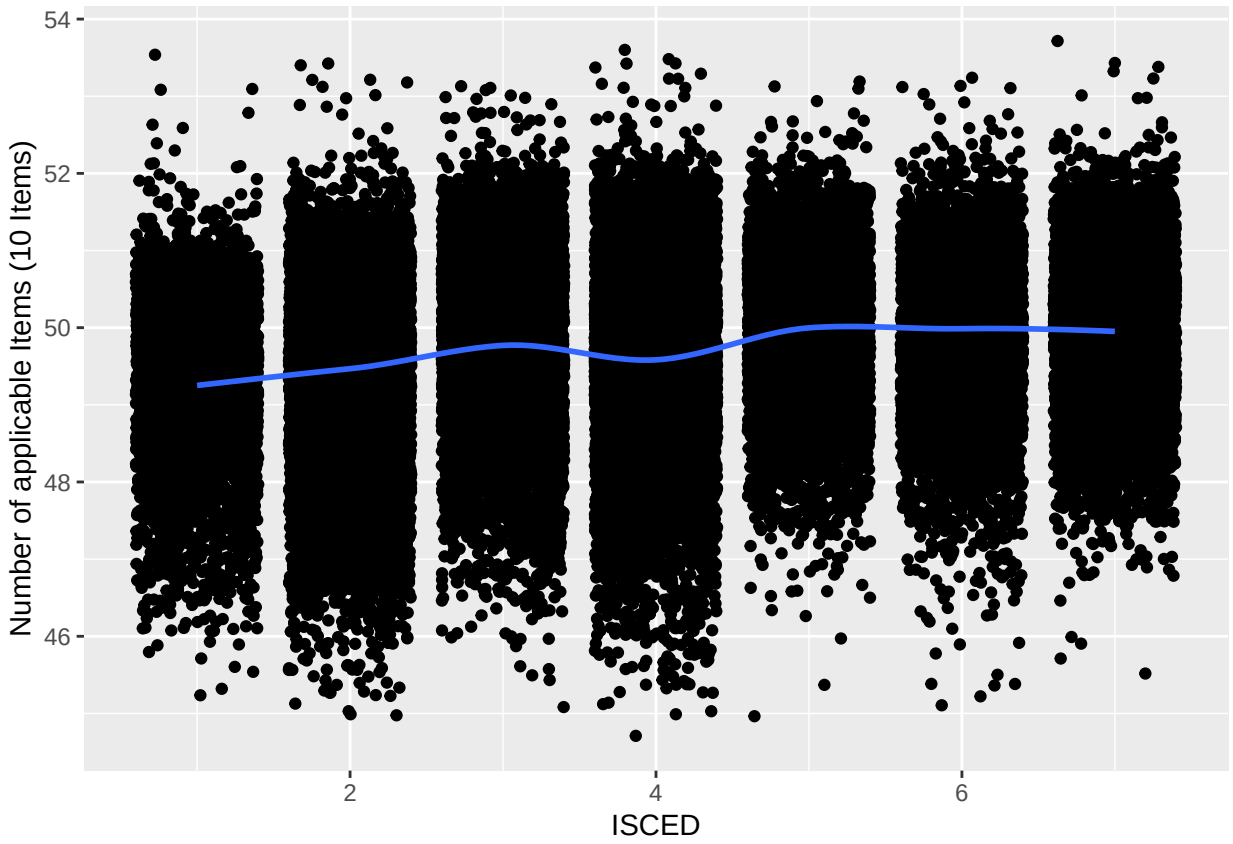
```
summary(analysis$n_applicable)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    44.7    49.0    49.8    49.7    50.5    53.7
```

```
ggplot(analysis) +
        aes(x = n_applicable) +
        geom_histogram() +
        labs(x = "Number of applicable Items (10 Items)", y = "")
```



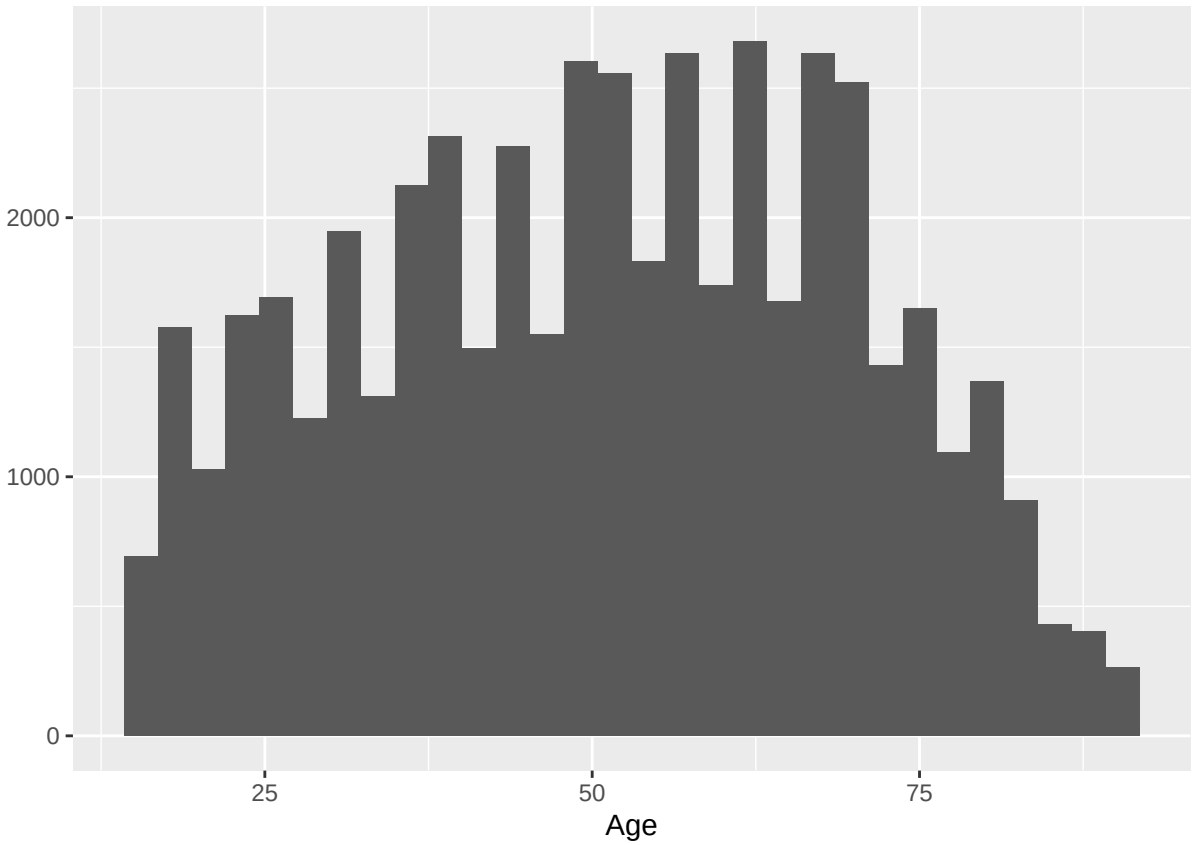

```
ggplot(analysis) +
        aes(x = educ,
            y = n_applicable) +
        geom_jitter() +
        geom_smooth(method = "loess",
                    se = FALSE) +
        labs(x = "ISCED",
             y = "Number of applicable Items (10 Items)")
```



Control Variables

The ESS surveys European populations over the age of 15. This is therefore the minimum age in the sample. The maximum age is 90. The mean age is 51.066 with a standard deviation of 18.647. There are 222 NAs.

```
ggplot(analysis) +
        aes(x = as.numeric(age)) +
        geom_histogram() +
        labs(x = "Age", y = "")
```



Education is operationalised using the established ISCED scale that orders the latest educational degrees by level. The number codes and levels are:

- 1: ES-ISCED I , less than lower secondary
- 2: ES-ISCED II, lower secondary
- 3: ES-ISCED IIIb, lower tier upper secondary
- 4: ES-ISCED IIIa, upper tier upper secondary
- 5: ES-ISCED IV, advanced vocational, sub-degree
- 6: ES-ISCED V1, lower tertiary education, BA level
- 7: ES-ISCED V2, higher tertiary education, >= MA level

The following table shows absolute frequencies of the educational levels:

Var1	Freq
1	3800
2	8329
3	8027
4	11212
5	6079
6	5517
7	6281
NA	274

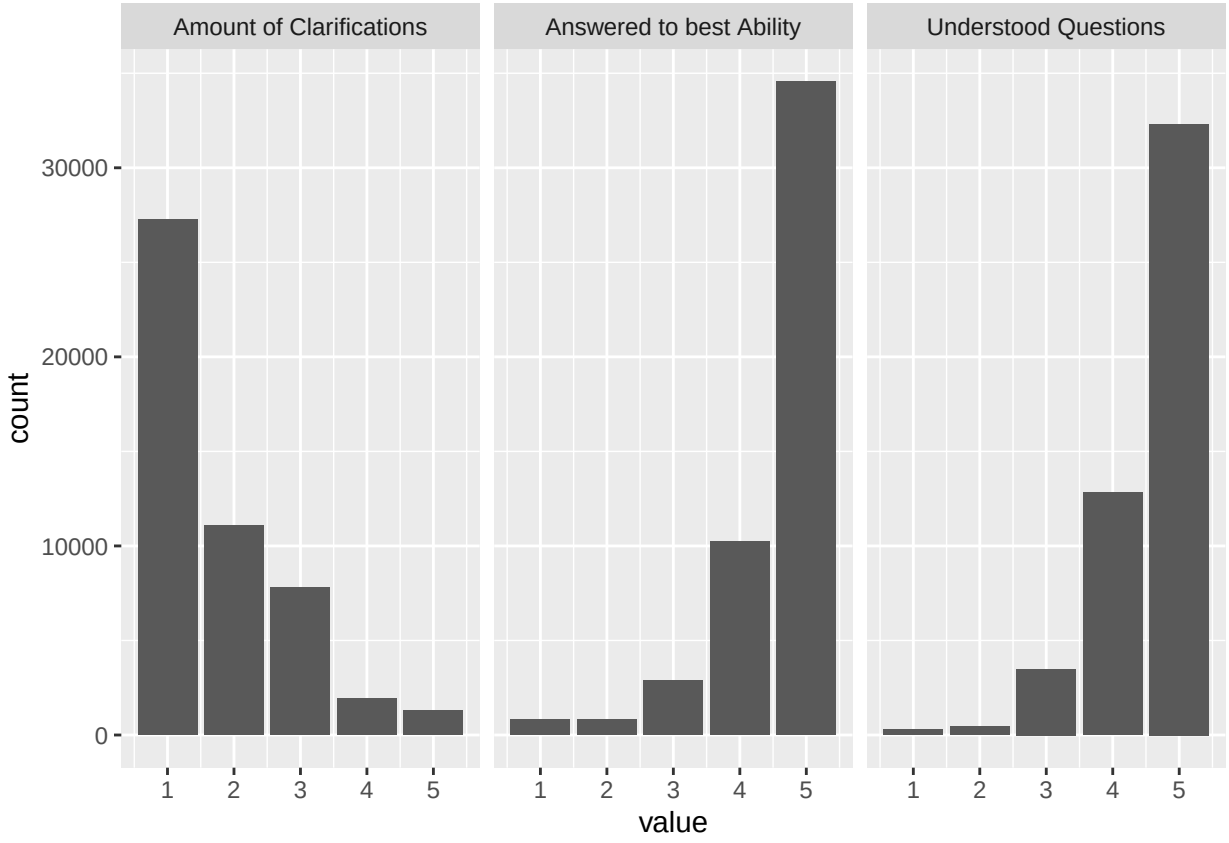
All control variables from the interviewer questionnaire are measured on a five point scale: Never, Almost never, Now and then, Often, Very often. The respective number codes range from 1 to 5.

I selected the assessment of the interviewer how often the respondent... - asked for clarifications, - answered to the best of their ability and - understood the question.

These items have two functions as controls. They are proxies of ability (and motivation) as well as the other

controls. And they control for the interviewers' assessment of the interview which might influence the answer on showcard use. In multiple regression, only the effect of a variable *net of all others* is estimated, therefore dampening the endogeneity problem with the showcard use item.

```
rbind(tibble(value = analysis$clarif,
             var = "Amount of Clarifications"),
      tibble(value = analysis$bestab,
             var = "Answered to best Ability"),
      tibble(value = analysis$underst,
             var = "Understood Questions")) %>%
  ggplot() +
    aes(x = value) +
    geom_bar() +
    facet_wrap(~var)
```




```
rbind(tibble(value = analysis$clarif,
             var = "Amount of Clarifications"),
      tibble(value = analysis$bestab,
             var = "Answered to best Ability"),
      tibble(value = analysis$underst,
             var = "Understood Questions")) %>%
  group_by(var) %>%
  summarise(Mean = mean(value, na.rm = TRUE),
            SD = sd(value, na.rm = TRUE))
```

```
## # A tibble: 3 x 3
##   var                       Mean    SD
##   <chr>                    <dbl> <dbl>
```

```
## 1 Amount of Clarifications  1.76 1.02
## 2 Answered to best Ability  4.56 0.815
## 3 Understood Questions      4.55 0.715
```

Multivariate Analyses

I analyze the data using a negative binomial regression with interviewer fixed effects. I choose this model because my dependent variable is a count variable that has the typical skewed shape of count variables making it unfit for linear regression. I chose a negative binomial link function instead of a Poisson regression because there are signs of overdispersion and because it is generally recommended. A more thorough discussion on model choice can be found below.

I use interviewer fixed effects to control for any variation on the interviewer level. Standard errors are clustered by the interviewer as well. This is common practice with fixed effects because the fixed effect only gets rid of the differences in means between groups but there might be differences in variance between groups as well. Not acknowledging potential differences in variance between groups in fixed-effects regression would lead to heteroscedasticity and therefore wrong statistical inference. Normal parametric standard errors tend to be too large (Cameron and Trivedi 2013). Clustered standard errors are therefore the default in the fixest package (see https://cran.r-project.org/web/packages/fixest/vignettes/standard_errors.html). I replicated the regressions in Stata using bootstrap standard errors (not shown) and only get marginally different results, likely from differences in the optimization algorithm.

The ESS does not draw simple random samples in all countries but the country samples are often stratified. In usual analysis, this requires the use of design weights acknowledging different probabilities of being in the sample. It is additionally recommended to use weights to correct for different sampling probabilities due to unit nonresponse. But since this analysis is not concerned with making inferences about the population of countries but inferences about realized interviews, no weighting is applied.

The missings are deleted listwise. There are two major sources of missing data in my analysis. The first one is the age of the interviewer is often missing and Romania used the wrong format for the interviewer's age altogether. I resultingly had to completely exclude Romania. But this would only be consequential if the nonresponse in the age of the interviewer is related to the interaction between respondent and interviewer since I control for interviewer effects using the fixed effects. Although such a large loss of observations is indeed unfortunate, they can be assumed to be random and therefore not bias the results. The levels of item nonresponse in the other variables used for analysis is always reported with the descriptive statistics above and are not particularly high. Since imputation in multivariate analysis makes as many assumptions, I am not sure whether it would be beneficial compared to listwise deletion and its assumption of being missing at random.

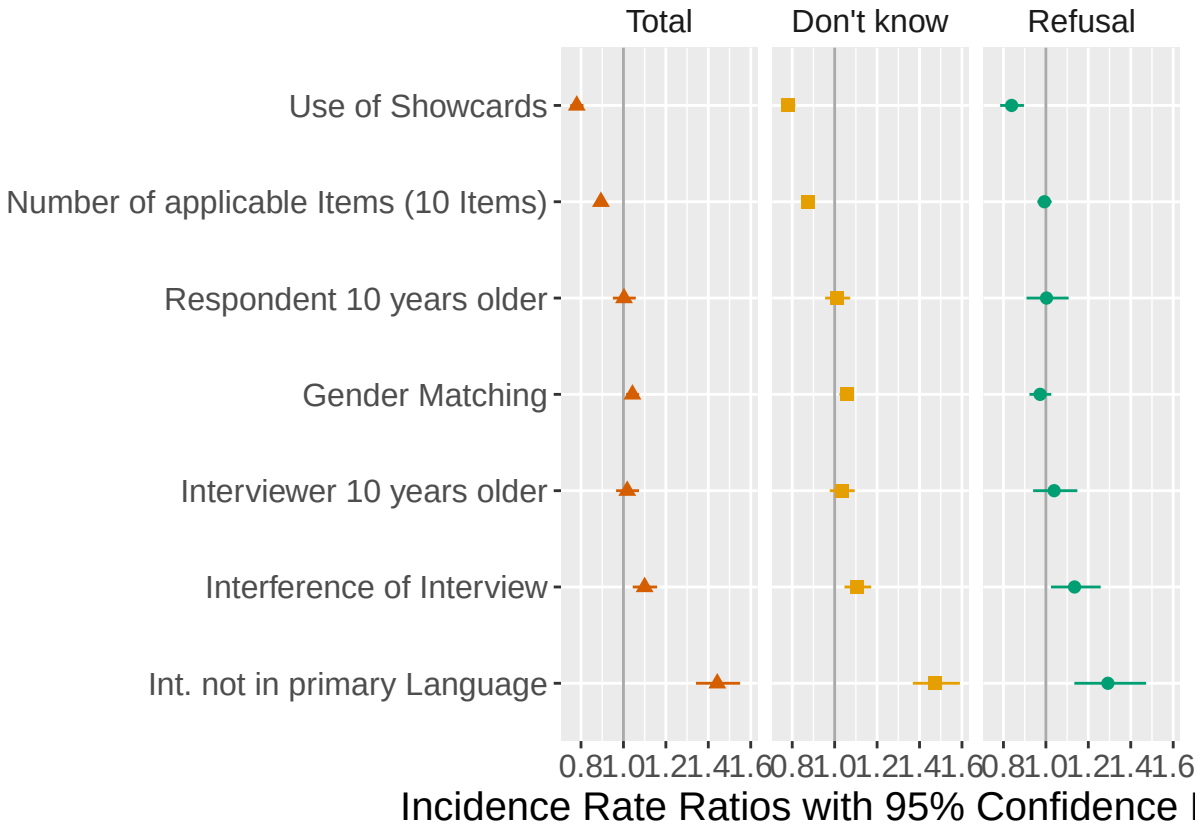
```
model <- fenegbin(data = analysis,
                  fml =  c(n_dk, n_ref, n_tot) ~
                    n_applicable + interference + showcards + notprimlang +
                      gndrmatch + resolder + intolder +
                      educ + age + age_sq + underst + bestab +  clarif |
                      intver,
                  se = "cluster")

setFixest_dict(labels)
etable(model,
       file = "results_inr_ess9.tex", replace = TRUE, tex = TRUE,
       title = "Regression Results",
       digits = 3, se.below = TRUE,
       signifCode = c("***"=0.001, "**"=0.01, "*"=0.05),
       fitstat = c("n", "pr2", "wpr2", "bic", "theta", "f"))
etable(model,
```

```
       tex = FALSE,
       title = "Regression Results",
       digits = 3, se.below = TRUE,
       signifCode = c("***"=0.001, "**"=0.01, "*"=0.05),
       fitstat = c("n", "pr2", "wpr2", "bic", "theta", "f"))
```

```
##                                        model 1   model 2    model 3
## Dependent Var.:                     Don't know   Refusal      Total
##
## Number of applicable Items (10 Items) -0.133***  -0.007    -0.112***
##                                     (0.010)    (0.017)   (0.009)
## Interference of Interview            0.102***   0.127*    0.095***
##                                     (0.029)    (0.053)   (0.027)
## Use of Showcards                    -0.247***  -0.176*** -0.249***
##                                     (0.022)    (0.034)   (0.020)
## Int. not in primary Language         0.389***   0.256***  0.366***
##                                     (0.038)    (0.067)   (0.037)
## Gender Matching                      0.055**   -0.028     0.042**
##                                     (0.017)    (0.027)   (0.015)
## Respondent 10 years older            0.012      0.003     0.003
##                                     (0.030)    (0.051)   (0.028)
## Interviewer 10 years older           0.033      0.038     0.018
##                                     (0.029)    (0.051)   (0.027)
## Education (ISCED)                   -0.115***   0.068*** -0.085***
##                                     (0.005)    (0.009)   (0.005)
## Age                                 -0.034***   0.009    -0.029***
##                                     (0.003)    (0.005)   (0.003)
## Age squared                          0.0004*** -4.07e-5   0.0003***
##                                     (2.54e-5)  (4.79e-5) (2.41e-5)
## Understood Questions                -0.348***  -0.052    -0.298***
##                                     (0.017)    (0.027)   (0.016)
## Answered to best Ability            -0.029     -0.180*** -0.057***
##                                     (0.015)    (0.023)   (0.014)
## Amount of Clarifications             0.254***   0.598***  0.330***
##                                     (0.011)    (0.021)   (0.011)
## Fixed-Effects:                      ---------- --------- ----------
## Interviewer                                Yes       Yes        Yes
## ----------------------------------- ---------- --------- ----------
## S.E.: Clustered                     by: intver by: int.. by: intver
## Observations                            43,745    36,745     44,000
## Pseudo R2                              0.12036   0.18128    0.12289
## Within Pseudo R2                       0.05886   0.08776    0.05999
## BIC                                  199,301.2  92,395.0  214,873.6
## Over-dispersion                        0.89873   0.75842     1.0379
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
results <- bind_rows(bind_cols(var = rownames(model$n_dk$coeftable),
                               model$n_dk$coeftable,
                               y = "Don't know"),
                     bind_cols(var = rownames(model$n_ref$coeftable),
                               model$n_ref$coeftable,
                               y = "Refusal"),
                     bind_cols(var = rownames(model$n_tot$coeftable),
```

```
                                model$n_tot$coeftable,
                                y = "Total")) %>%
  rename(estimate = `Estimate`,
         se = `Std. Error`) %>%
  filter(var != ".theta") %>%
  filter(var %in% c("duration", "n_applicable", "interference", "showcards",
                    "notprimlang", "gndrmatch", "resolder", "intolder")) %>%
  mutate(irr = exp(estimate),
         upper = exp(estimate + se * (-qnorm((1-0.95)/2))),
         lower = exp(estimate - se * (-qnorm((1-0.95)/2))))

ggplot(results) +
  aes(y = reorder(var, -irr),
      x = irr,
      colour = y,
      shape = y) +
  geom_vline(xintercept = 1,
             colour = "darkgrey") +
  geom_point(size = 2) +
  geom_linerange(aes(y = var,
                     xmin = lower,
                     xmax = upper),
                 lwd = .5) +
  facet_wrap(~factor(y, levels=c("Total", "Don't know", "Refusal"))) +
  scale_color_manual(values = c("#E69F00", "#009E73", "#D55E00")) +
  scale_shape_manual(values = 15:17) +
  scale_y_discrete(labels = labels) +
  theme(legend.position = "none",
        strip.background = element_blank(),
        text = element_text(size = 15)) +
  labs(y = "", x = "Incidence Rate Ratios with 95% Confidence Intervals")
```




```
ggsave(filename = "results_inr_ess9.pdf",
       width = 22, height = 10, units = "cm")
```

Although I removed all NAs before, there are different numbers of observations for each dependent variable. This is due to some interviewers having no within-variance (that have only respondents with no item nonresponse). Since a within-estimator needs within-variance, interviewers without within-variance are excluded from the analysis. This is particularly often the case for the number of refusals since the number of refusals is relatively low. Although such a large loss of observations is indeed unfortunate, they again can be assumed to be random and therefore not bias the results.

Regression Diagnostics and Model Justification

Fixed effects count data models are quite contested. Wooldridge (1999) claims that fixed effects negative binomial regression (FENB) is not a true fixed effects estimator and that overdispersion is not of much concern in these models. He advises always using FE Poisson regression (FEPois) over FENB.

Allison and Waterman (2002) agree that FENB in its previous form was not a true FE estimator but propose a modified version that they advise to always use over FEPois. This improved version is the one implemented in the fixest package.

I compare estimates and diagnostic plots for both regression models using the total sum of item nonresponse as the dependent variable to assess the reliability of my results.

```
fepois <- fepois(data = analysis,
                 fml =  n_tot ~
                   n_applicable + interference + showcards + notprimlang +
                     gndrmatch + resolder + intolder +
                     educ + age + age_sq + underst + bestab +  clarif |
                     intver,
                 se = "cluster")
```

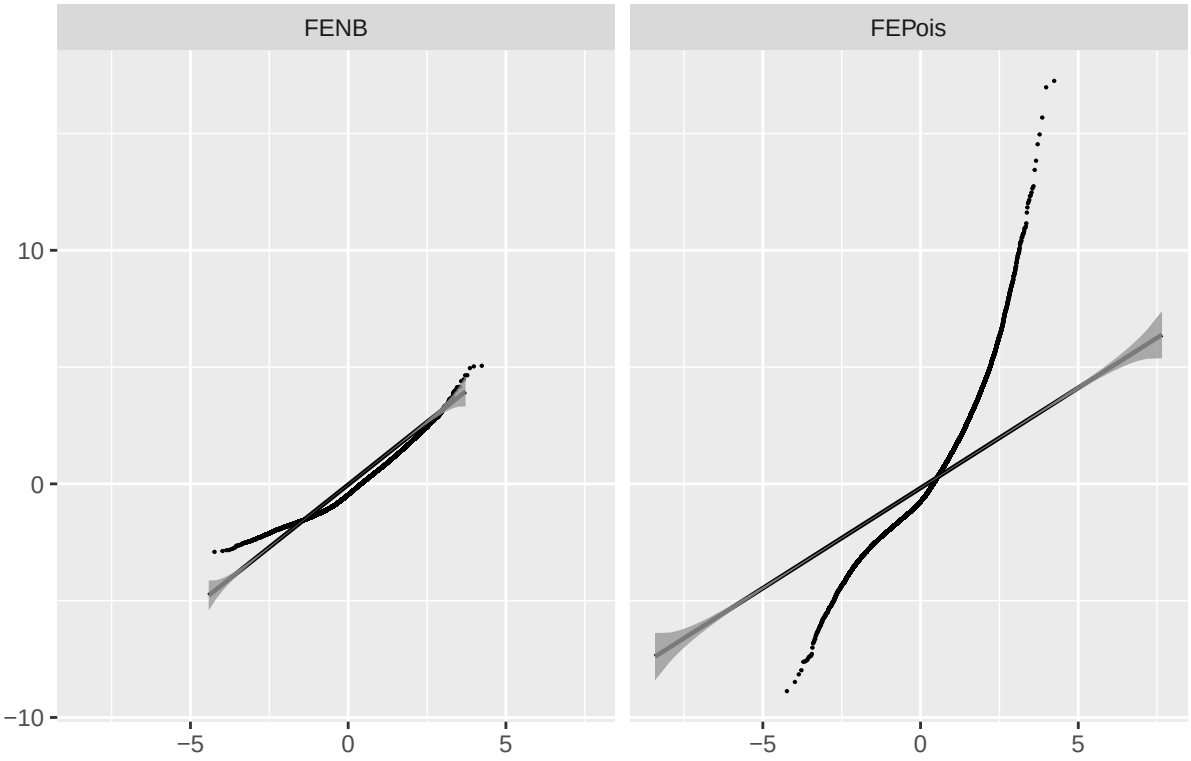
```
setFixest_dict(labels)
etable(model$n_tot, fepois,
       tex = FALSE,
       title = "Tab. S2: FENB vs. FEPois",
       digits = 3, se.below = TRUE,
       signifCode = c("***"=0.001, "**"=0.01, "*"=0.05),
       fitstat = c("n", "pr2", "wpr2", "bic", "theta", "f"))
```

```
##                                      model$n_..     fepois
## Dependent Var.:                           Total      Total
##
## Number of applicable Items (10 Items) -0.112***  -0.095***
##                                       (0.009)    (0.009)
## Interference of Interview              0.095***   0.102***
##                                       (0.027)    (0.028)
## Use of Showcards                      -0.249***  -0.241***
##                                       (0.020)    (0.021)
## Int. not in primary Language           0.366***   0.396***
##                                       (0.037)    (0.036)
## Gender Matching                        0.042**    0.041**
##                                       (0.015)    (0.016)
## Respondent 10 years older              0.003     -0.017
##                                       (0.028)    (0.028)
## Interviewer 10 years older             0.018      0.038
##                                       (0.027)    (0.027)
## Education (ISCED)                     -0.085***  -0.074***
##                                       (0.005)    (0.005)
## Age                                   -0.029***  -0.027***
##                                       (0.003)    (0.003)
## Age squared                            0.0003***  0.0003***
##                                       (2.41e-5)  (2.32e-5)
## Understood Questions                  -0.298***  -0.243***
##                                       (0.016)    (0.015)
## Answered to best Ability              -0.057***  -0.071***
##                                       (0.014)    (0.014)
## Amount of Clarifications               0.330***   0.268***
##                                       (0.011)    (0.011)
## Fixed-Effects:                        ---------- ----------
## Interviewer                                  Yes        Yes
## ------------------------------------ ---------- ----------
## Family                                Neg. Bin.    Poisson
## S.E.: Clustered                      by: intver by: intver
## Observations                             44,000     44,000
## Pseudo R2                                0.12289    0.42287
## Within Pseudo R2                         0.05999    0.16196
## BIC                                    214,873.6  286,571.5
## Over-dispersion                          1.0379         --
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
diagnostics <- rbind(tibble(resid = resid(fepois,
                                          type = "deviance"),
                            fitval = fepois$fitted.values,
                            model = "FEPois"),
```

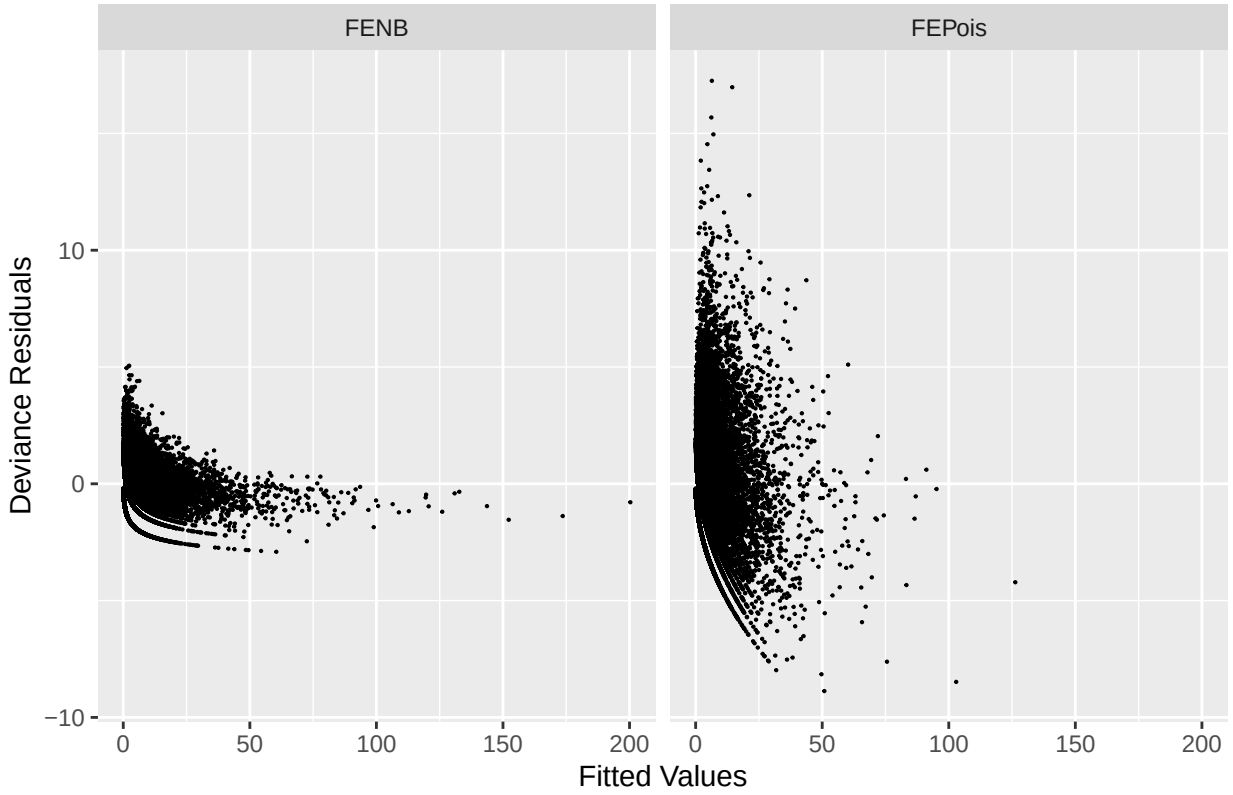
```
                    tibble(resid = resid(model$n_tot,
                                         type = "deviance"),
                           fitval = model$n_tot$fitted.values,
                           model = "FENB"))

ggplot(diagnostics) +
  aes(sample = resid) +
  stat_qq(size = 0.1) +
  stat_qq_line() +
  stat_qq_band() +
  facet_wrap(~model) +
  labs(title = "Fig. S2: QQ-Plot of Deviance Residuals",
       x = "", y = "")
```

Fig. S2: QQ−Plot of Deviance Residuals




```
ggplot(diagnostics) +
  aes(x = fitval,
      y = resid) +
  geom_point(size = 0.1) +
  facet_wrap(~model) +
  labs(title = "Fig. S3: Residuals vs Fitted: FENB",
       x = "Fitted Values",
       y = "Deviance Residuals")
```

Fig. S3: Residuals vs Fitted: FENB



As can be seen in the diagnostic plots, the Poisson Regression with fixed effects exhibits a much larger residual variance and the deviance residuals are much less normal than the negative binomial regression with fixed effects. The average deviance residual of the FEPois is -0.381 with a standard deviation of 1.897 compared to -0.35 with a standard deviation of 0.961 for FENB.

Although normality of the residuals is not such a strong criterion with count data models, Cameron and Trivedi (2013, chapter 9.4) expect normality in deviance residuals in correctly specified models. While not perfect, the deviance residuals of the FENB model are much closer to normally distributed than FEPois. And the unconditional FENB by Allison and Waterman (2002) implemented in the fixest package is a true fixed effects model, unlike previous versions. It has an incidental parameter problem though, but it is usually very small. Wooldridge (1999) argued for Poisson regression when using fixed effects but this seems to be the worse option in this case. The overdispersion seems to be still consequential even with fixed effects.

On the other hand, the substantial results are very similar. Coefficients differ slightly between models but signs and significance are largely identical. I, therefore, settle on the FENB model in the paper.

Literature

Allison, Paul D. and Richard P. Waterman (2002). “Fixed-Effects Negative Binomial Regression Models”. In: Sociological Methodology 32.1, pp. 247–265. https://doi.org/10.1111/14679531.00117.

Cameron, A. Colin, and Pravin Trivedi. Regression Analysis of Count Data. 2nd ed. Cambridge: Cambridge University Press, 2013. https://doi.org/10.1017/CBO9781139013567.

Wooldridge, Jeffrey M. (1999). “Distribution-Free Estimation of Some Nonlinear Panel Data Models”. In: Journal of Econometrics 90.1, pp. 77–97. https://doi.org/10.1016/S0304-4076(98)00033-5.