Some exploration of ProQuest data quality

Flavio & Christoph

03 October, 2022

Contents

Counts 3
- Number of authors 3
- Number of fields per author 4

Advisors 4

Information on field 6
- Fraction of authors with field in MAG (overall) 6
- Fraction of authors with field in MAG (as of those with at least one field) 6
- Number of fields MAG level 0 per author 7
- Number of field 0 per author by first reported field0 (position 0) 8

ProQuest vs Survey of Earned Doctorates 9
- Distribution of how the links to the field were made 12

N graduates by detailed fields 13
- Comments 16
- What is happening to psychology? 16
- Conclusion 18

```
q <- "select * from pq_authors
        left join (
                select university_id, location
                from pq_unis
        ) using (university_id)
        "

authors <- tbl(con, sql(q)) %>%
    collect() %>% # focus only on U.S. institutions
    filter(grepl("United States", location))

advisors <- tbl(con, "pq_advisors") %>%
    collect()

fields <- tbl(con, "pq_fields") %>%
    collect()

q <- "select * from pq_fields_mag a
        inner join (select FieldOfStudyId, NormalizedName as fieldname_mag
                    from FieldsOfStudy) b on (a.mag_field0 = b.FieldOfStudyId)"
fields_mag <- tbl(con, sql(q)) %>%
    collect()
```

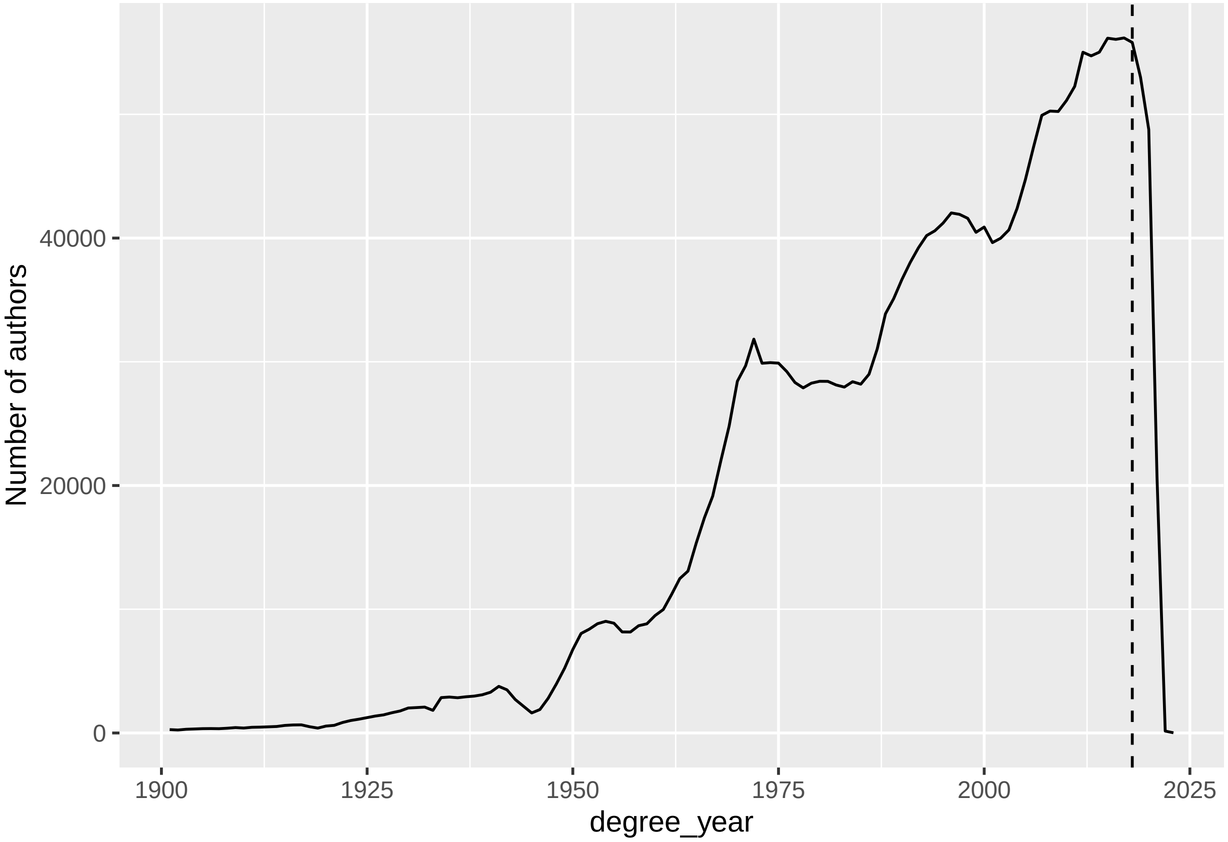
```
path_nces <- paste0(datapath, "NCES_NSF/processed/")
```

```
data_graduates <- fread(paste0(path_nces, "graduate_counts.csv"), data.table = FALSE)

data_graduates_fld0 <- fread(paste0(path_nces, "counts_uni_field0.csv"),
                             data.table = FALSE)

fields_nces <- fread(paste0(path_nces, "fields.csv"), data.table = FALSE)

data_graduates <- data_graduates %>%
  left_join(fields_nces %>%
              select(field_id, fld = shortname),
            by = "field_id") %>%
  select(-field_id)
```

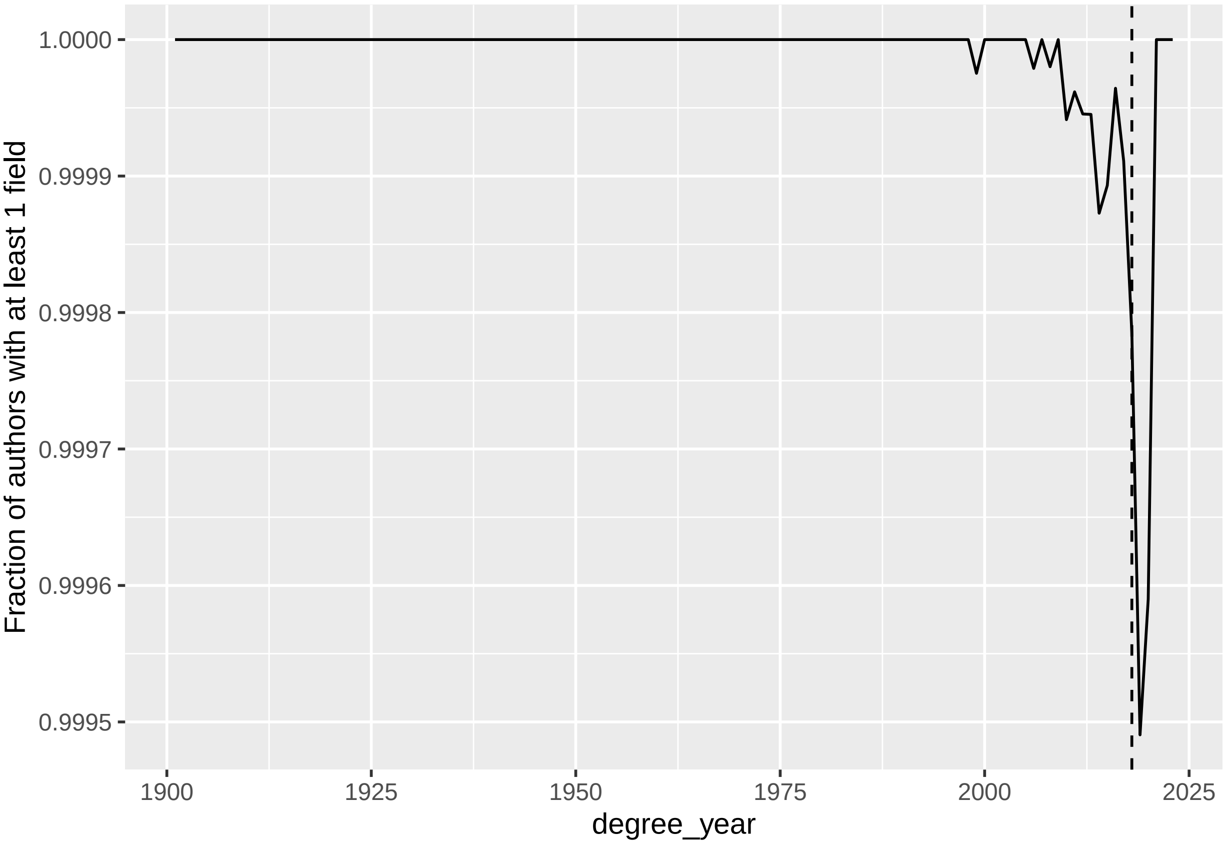
```
d_main <- authors %>%
    mutate(has_advisor = ifelse(goid %in% advisors$goid, 1, 0),
           has_field = ifelse(goid %in% fields$goid, 1, 0),
           has_field_mag = ifelse(goid %in% fields_mag$goid, 1, 0))

fields_perauthor <- fields_mag %>%
    group_by(goid) %>%
    summarise(n_field0 = n(), .groups = "drop")

d_main <- d_main %>%
    left_join(fields_perauthor, by = "goid")
```

Counts

Number of authors

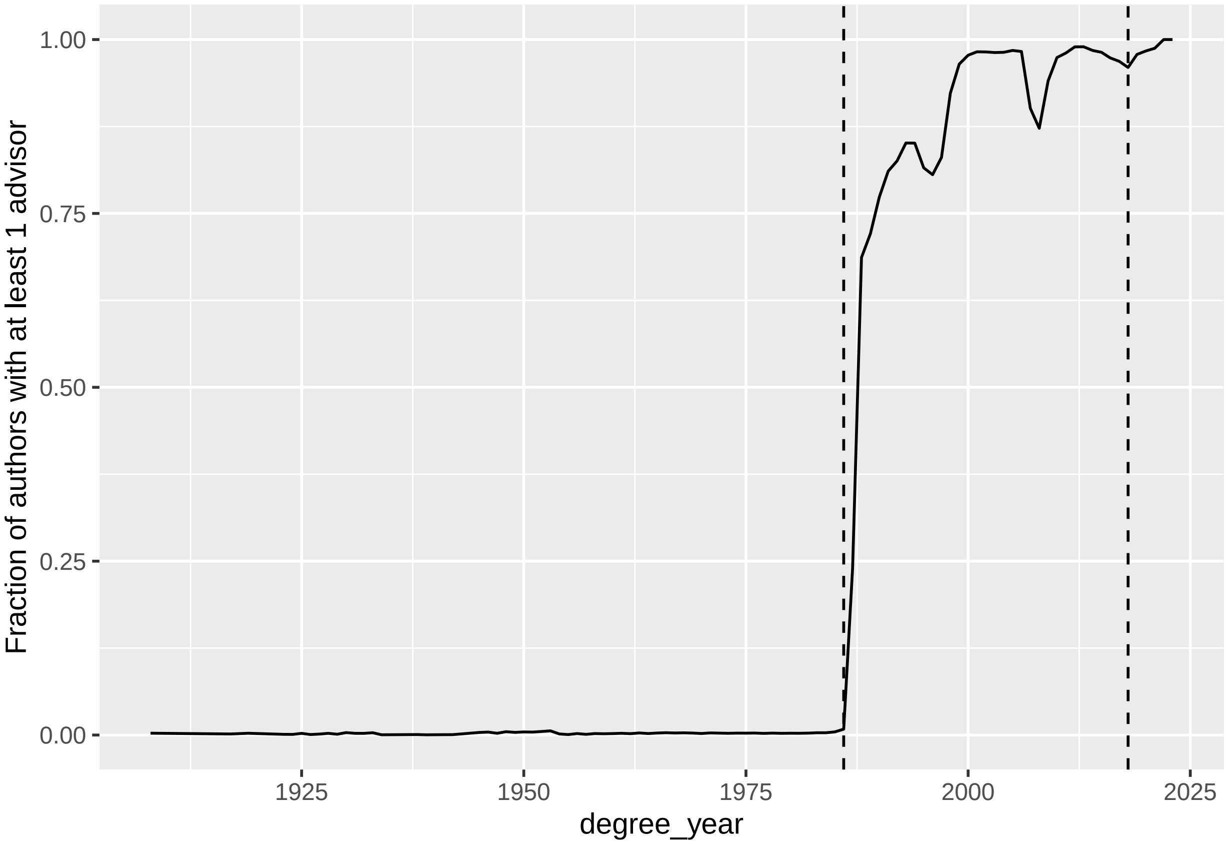


Number of fields per author



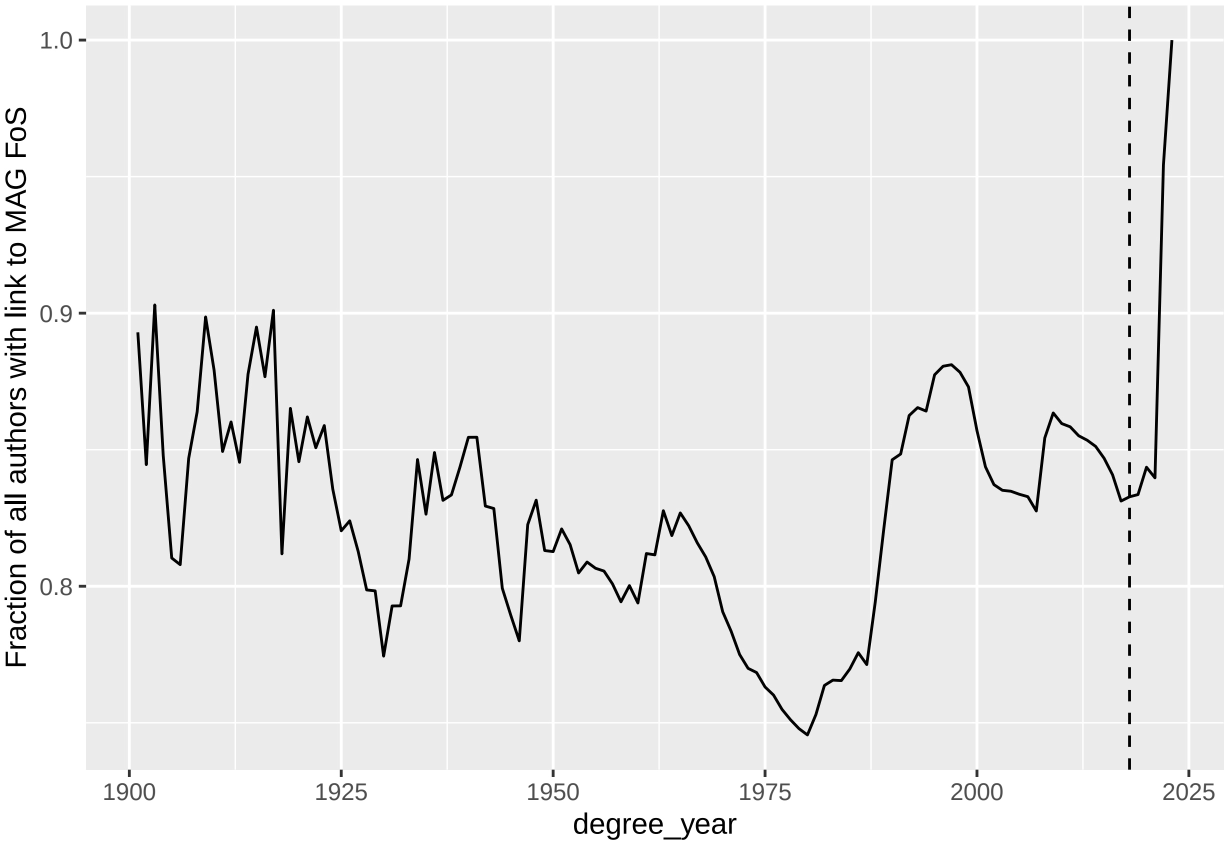
Advisors

- Advisors present per author
- advisors are only present since 1986


Fraction of authors with at least 1 advisor
1.00
0.75
0.50
0.25
0.00
1925
1950
1975
2000
2025
degree_year

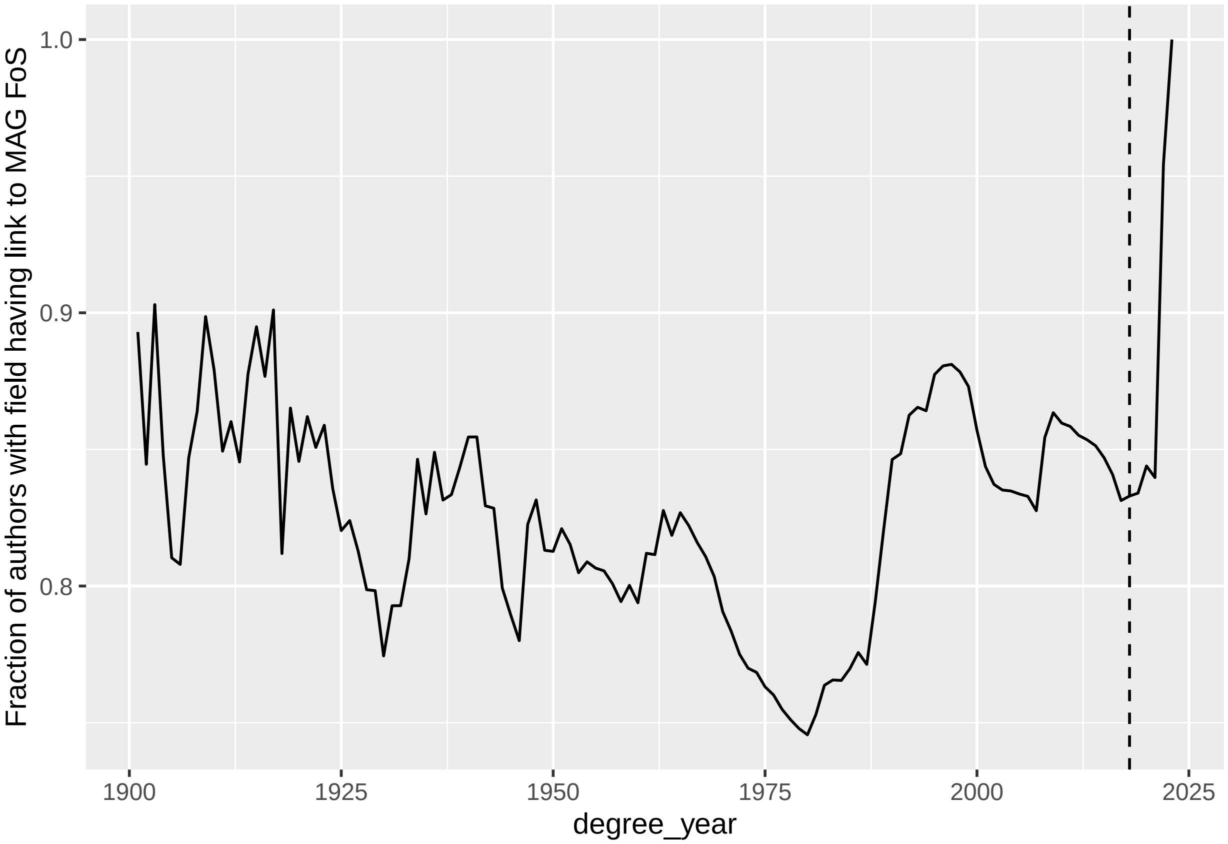
Information on field

Fraction of authors with field in MAG (overall)



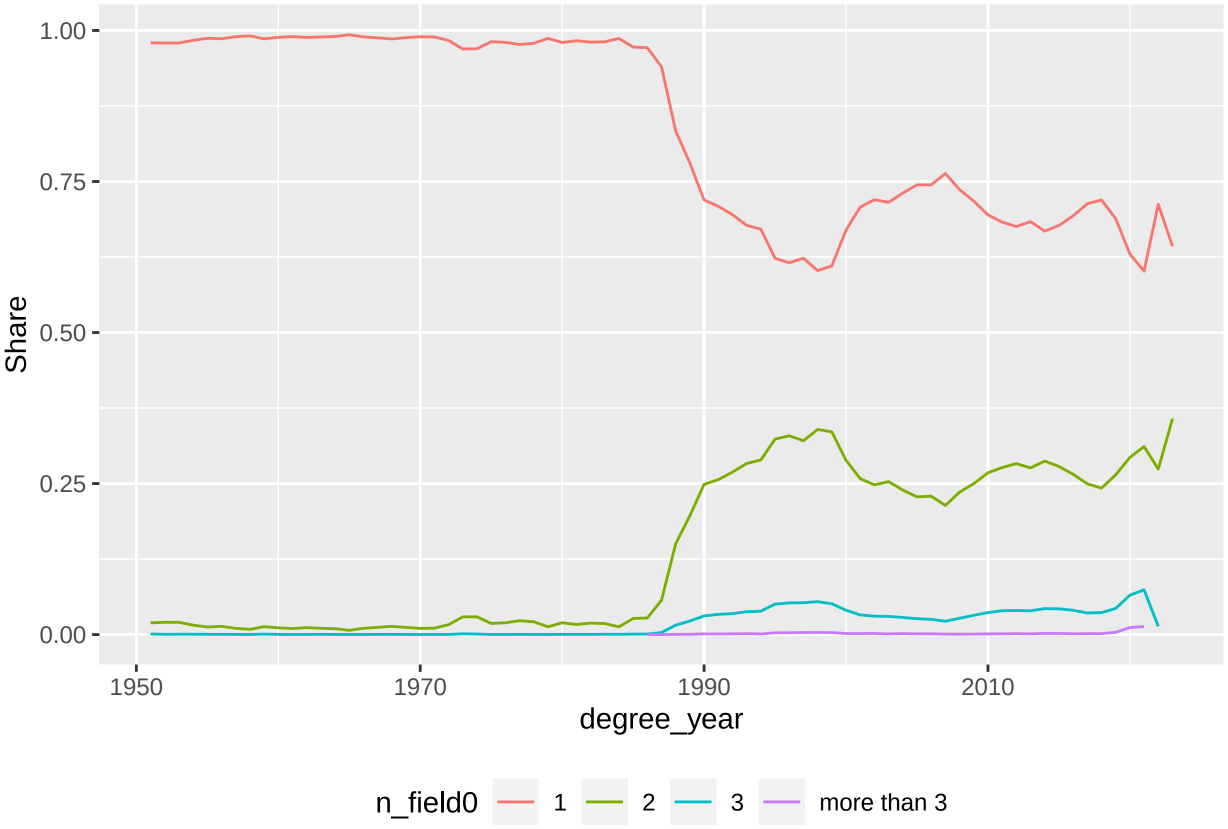
Fraction of authors with field in MAG (as of those with at least one field)

```
d_main %>%
    filter(has_field == 1) %>%
    group_by(degree_year, has_field_mag) %>%
    summarise(nb = n(), .groups = "drop") %>%
    group_by(degree_year) %>%
    mutate(s = nb / sum(nb)) %>%
    ungroup() %>%
    filter(has_field_mag == 1) %>%
    ggplot(aes(x = degree_year, y = s)) +
    geom_line() +
    labs(y = "Fraction of authors with field having link to MAG FoS") +
    geom_vline(xintercept = 2018, linetype = "dashed")
```



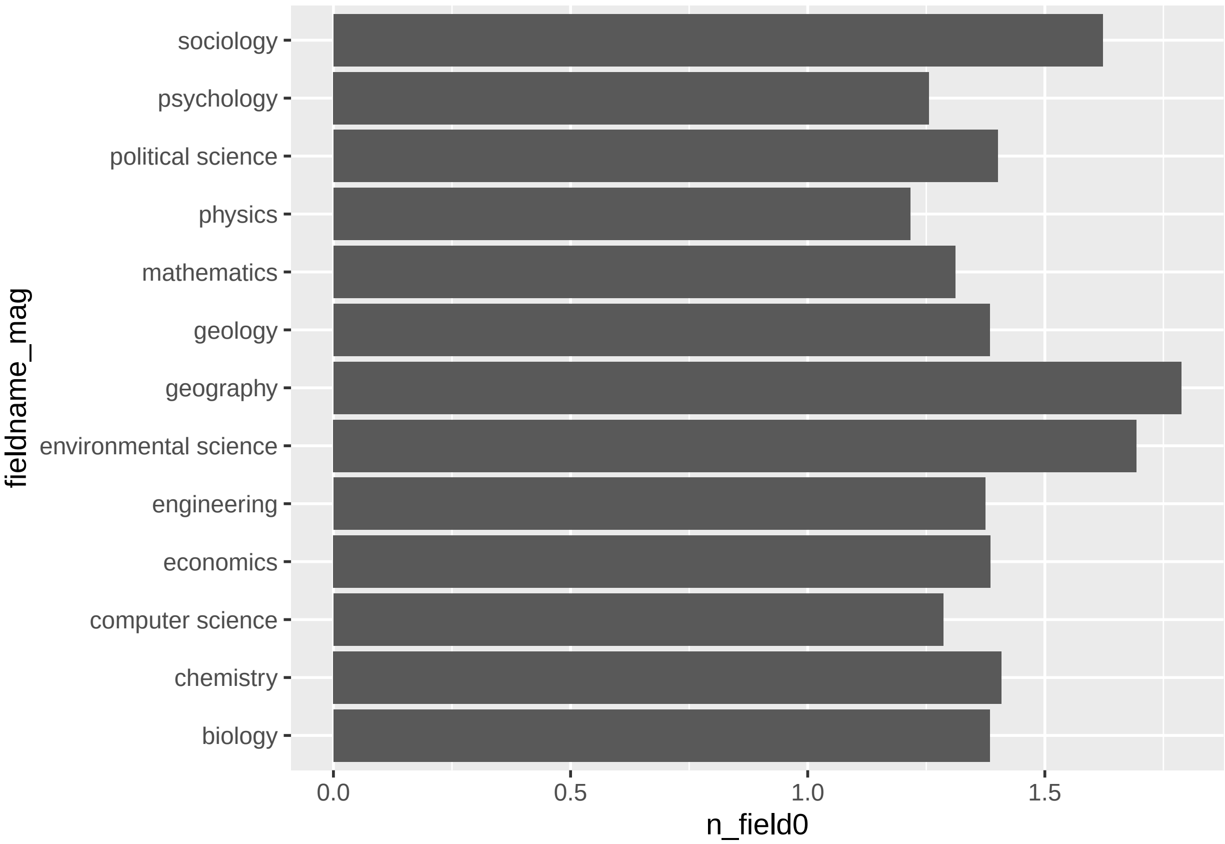
Number of fields MAG level 0 per author

```
d_main %>%
    filter(has_field_mag == 1 & degree_year > 1950) %>%
    mutate(n_field0 = case_when(
        n_field0 <= 3 ~ as.character(n_field0),
        n_field0 > 3 ~ "more than 3"
    )) %>%
    mutate(n_field0 = factor(n_field0)) %>%
    group_by(degree_year, n_field0) %>%
    summarise(nb = n(), .groups = "drop") %>%
    group_by(degree_year) %>%
    mutate(s = nb / sum(nb)) %>%
    ungroup() %>%
    ggplot(aes(x = degree_year, y = s, color = n_field0)) +
    geom_line() +
    labs(y = "Share") +
    theme(legend.position = "bottom")
```



Number of field 0 per author by first reported field0 (position 0)

```
main_fields <- c("chemistry",
        "sociology",
        "mathematics",
        "biology",
        "computer science",
        "political science",
        "engineering",
        "psychology",
        "environmental science",
        "physics",
        "geology",
        "geography",
        "economics")

d_main %>%
  select(goid, n_field0, degree_year) %>%
  filter(degree_year > 1990 & degree_year <= 2020) %>%
  left_join(fields_mag %>%
              filter(position == 0 ) %>%
              select(goid, fieldname_mag),
            by = "goid") %>%
  filter(!is.na(n_field0)) %>%
  filter(fieldname_mag %in% main_fields) %>%
  group_by(fieldname_mag) %>%
  summarise(n_field0 = mean(n_field0),
            .groups = "drop") %>%
  ggplot(aes(x = fieldname_mag, y = n_field0)) +
  geom_bar(stat = "identity") +
  coord_flip()
```



ProQuest vs Survey of Earned Doctorates

```
# Prepare the proquest data
nb_pq <- authors %>%
    filter(grepl("United States", location)) %>%
    inner_join(fields_mag %>% # NOTE: drops graduates w/o field in mag, most of them "education"
                filter(position == 0) %>% # there can be multiple fields per goid, need to take one
                select(goid, fieldname_mag),
            by = "goid") %>%
    rename(year = degree_year) %>%
    group_by(year, fieldname_mag) %>%
    summarise(nb = n(),
             .groups = "drop")

# assign to broad field
humanities <- c("art", "history", "philosophy")
math_compsc <- c("mathematics", "computer science")
socsci <- c("economics", "political science", "psychology",
            "sociology")
phys_earth <- c("chemistry",
                "geography", "geology", "physics")
```

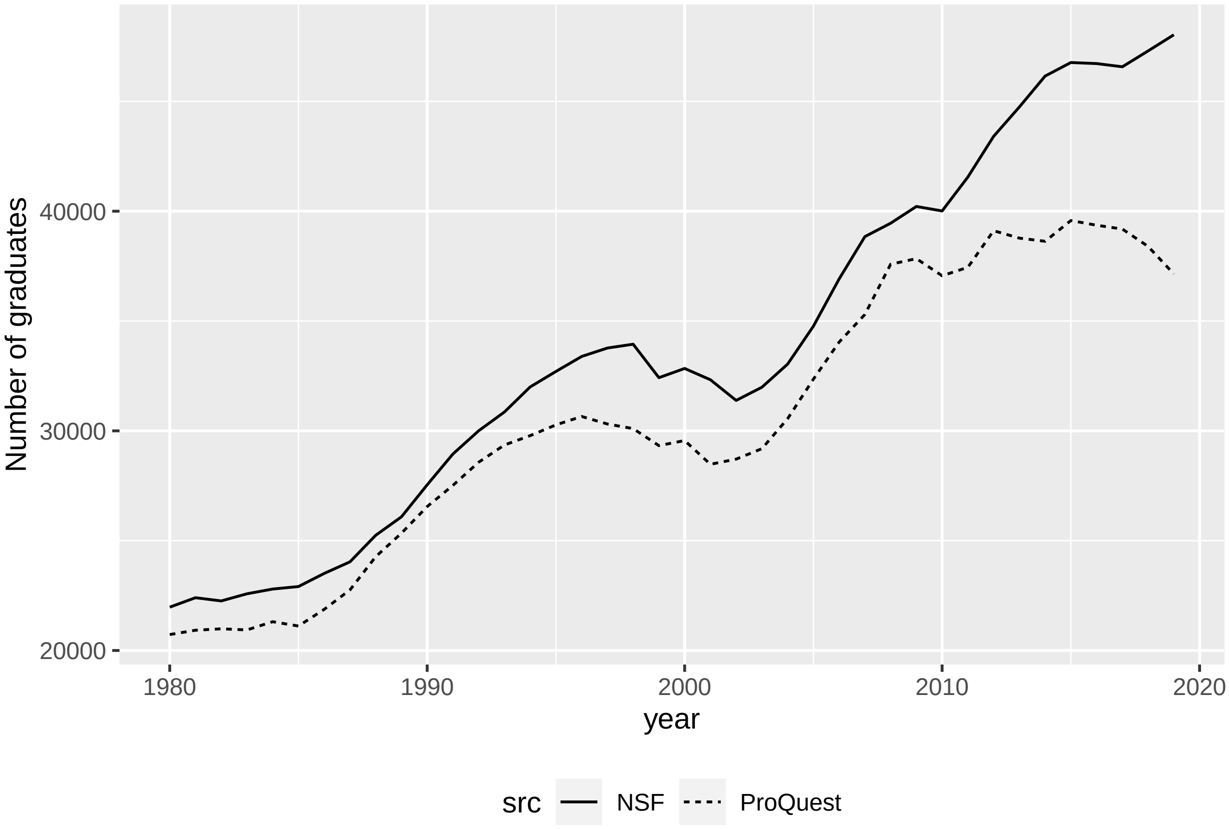
```
life_sc <- c("biology", "medicine")
engn <- c("engineering", "materials science")
unclassified <- c("environmental science", "business") # business is classified as "other social scienc
nb_pq <- nb_pq %>%
    mutate(fld = case_when(
        fieldname_mag %in% humanities ~ "humanities",
        fieldname_mag %in% math_compsc ~ "math_compsc",
        fieldname_mag %in% socsci ~ "socsci",
        fieldname_mag %in% life_sc ~ "life_sc",
        fieldname_mag %in% phys_earth ~ "phys_earth",
        fieldname_mag %in% engn ~ "engn",
        fieldname_mag %in% unclassified ~ "unclassified"
    ))

nb_pq <- nb_pq %>%
    group_by(year, fld) %>%
    summarise(nb = sum(nb),
              .groups = "drop")
```

```
nb_sed <- data_graduates %>%
    group_by(year, fld) %>%
    summarise(nb = sum(nb),
              .groups = "drop")

s_theses <- nb_pq %>%
    rename(nb_pq = nb) %>%
    left_join(nb_sed %>%
                rename(nb_sed = nb),
              by = c("fld", "year")) %>%
    filter(year >= 1980) %>%
    mutate(s_pq = nb_pq / nb_sed)

n_theses <- s_theses %>%
    select(-s_pq) %>%
    gather(key = src, value = nb, nb_sed:nb_pq) %>%
    mutate(src = ifelse(src == "nb_pq", "ProQuest", "NSF")) %>%
    group_by(year, src) %>%
    summarise(nb = sum(nb, na.rm = TRUE), # nb_sed missing for unclassified
              .groups = "drop")
```

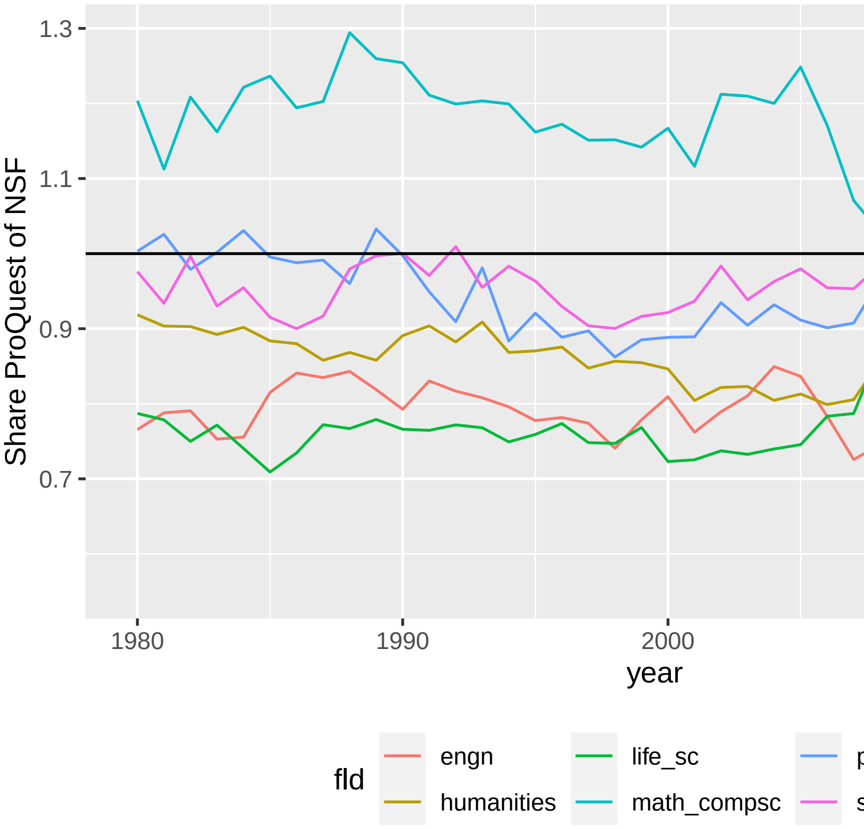
Totals over time

- Note: proquest has some theses classified as post-2020, they are missing from here
- In both counts, the field "education" is missing

```
n_theses %>%
    filter(year <= 2019) %>%
    ggplot(aes(x = year, y = nb, linetype = src)) +
    geom_line() +
    theme(legend.position = "bottom") +
    labs(y = "Number of graduates")
```



```
s_theses %>%
    filter(!(fld %in% c("unclassified")) & year < 2020) %>%
    ggplot(aes(x = year, y = s_pq, color = fld)) +
    geom_line() +
    theme(legend.position = "bottom") +
    labs(y = "Share ProQuest of NSF") +
    geom_hline(yintercept = 1)
```



Fraction ProQuest of NSF by field and year

Notes:

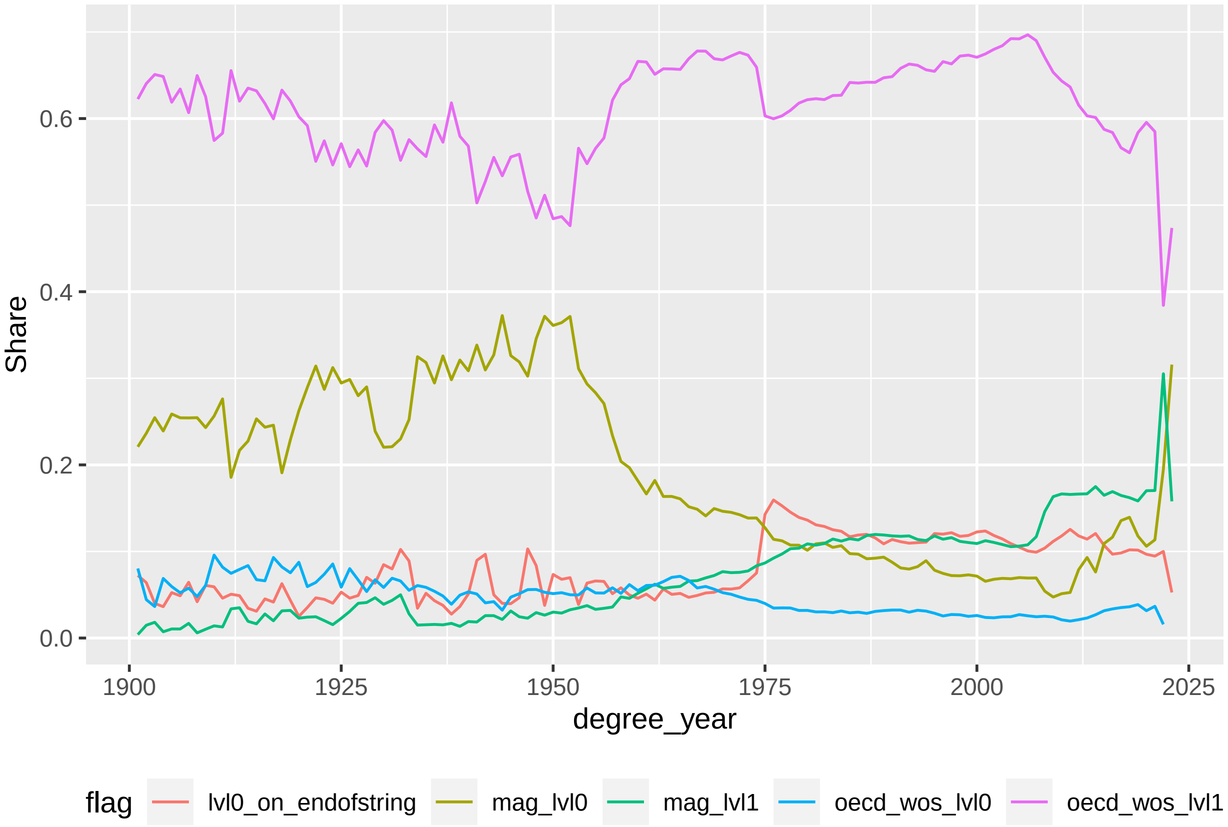
- if an author has multiple fields, the first one is used.
- life_sc includes medicine.
- missing is the interdisciplinary field "environmental science" from ProQuest.
- Todo: repeat this with more aggregate fields: e.g. GEEMP and LPS (to the extent possible). does this get rid of the over-representation of the math/computer science fields? (for instance because of engineering?).

Distribution of how the links to the field were made

- Note: this is the fraction of total links per degree_year; if an author has two links, she enters twice here

```
fields_mag %>%
    left_join(d_main %>%
        select(goid, degree_year),
        by = "goid") %>%
    group_by(degree_year, flag) %>%
    summarise(nb = n(), .groups = "drop") %>%
    group_by(degree_year) %>%
    mutate(s = nb / sum(nb)) %>%
    ungroup() %>%
    ggplot(aes(x = degree_year, y = s, color = flag)) +
    geom_line() +
    labs(y = "Share") +
    theme(legend.position = "bottom")
```

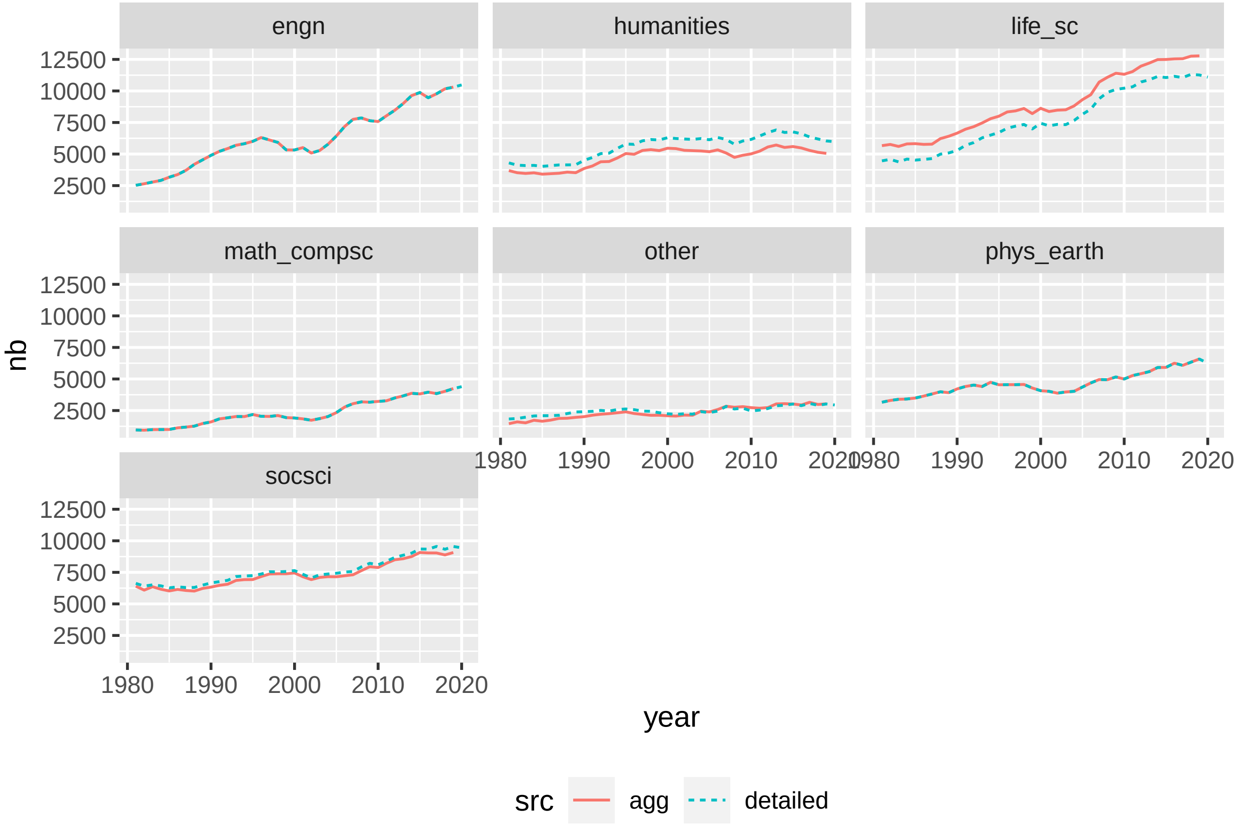
```
## Warning: Removed 5 row(s) containing missing values (geom_path).
```



N graduates by detailed fields

First we validate the two data sets we have from the SED: the aggregate data used above should be roughly equal to the data by detailed field and university used below

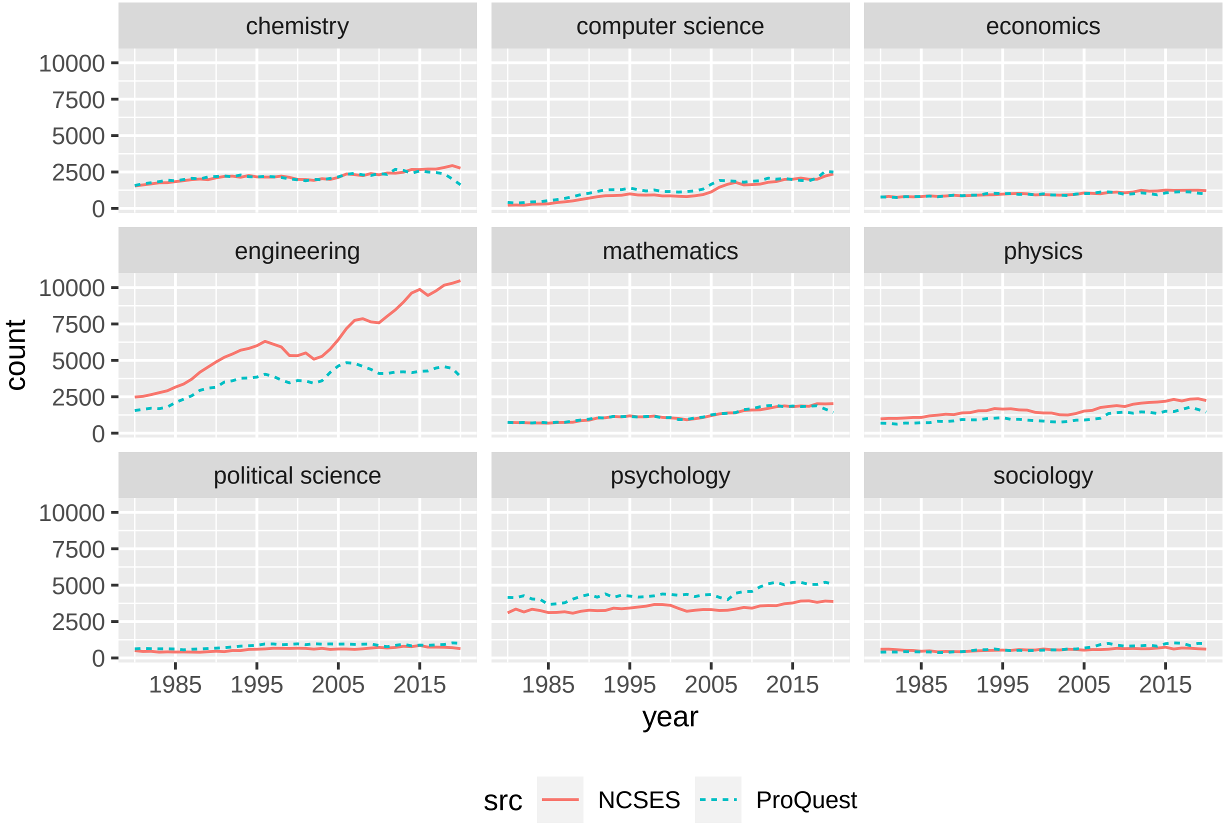
```
data_from_detailed <- data_graduates_fld0 %>%
    mutate(fld = case_when(
        fieldname0_mag %in% c("humanities", humanities) ~ "humanities",
        fieldname0_mag %in% math_compsc ~ "math_compsc",
        fieldname0_mag %in% c(socsci, "other socsci") ~ "socsci",
        fieldname0_mag %in% c(life_sc, "health sciences") ~ "life_sc",
        fieldname0_mag %in% phys_earth ~ "phys_earth",
        fieldname0_mag %in% engn ~ "engn",
        fieldname0_mag %in% unclassified ~ "other"
    )) %>%
  group_by(year, fld) %>%
  summarise(nb = sum(nb), .groups = "drop")


# df_agg <- nb_sed %>%
#   filter(fld == "socsci")

dk <- bind_rows(
  data_from_detailed %>%
    select(year, fld, nb) %>%
    mutate(src = "detailed"),
  nb_sed %>% mutate(src = "agg")
) %>%
  filter(year > 1980)

dk %>%
  ggplot(aes(x = year, y = nb))  +
  geom_line(aes(linetype = src, color = src)) +
  facet_wrap(~fld) +
  theme(legend.position = "bottom")
```



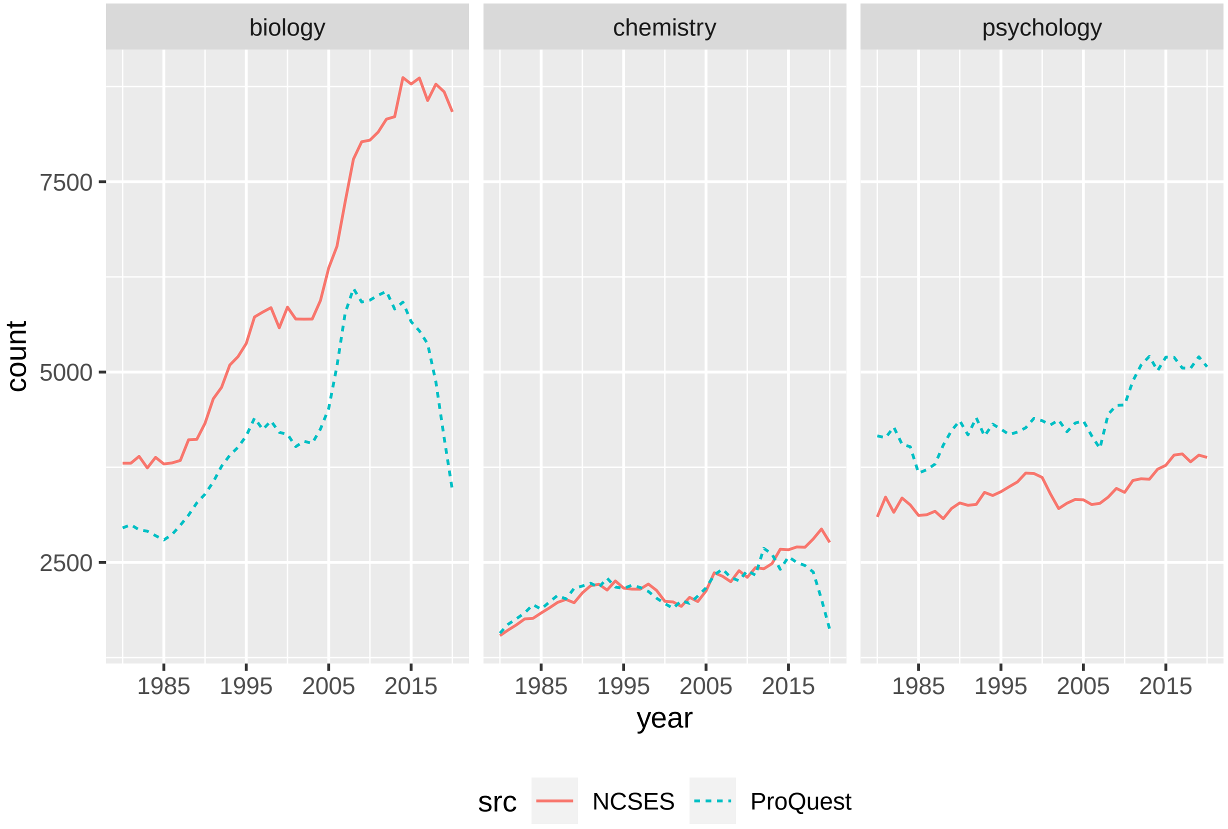
Now we plot the detailed counts from SED with the counts from ProQuest

```
nb_pq_fld0 <- authors %>%
    filter(grepl("United States", location)) %>%
    inner_join(fields_mag %>% # NOTE: drops graduates w/o field in mag, most of them "education"
                 filter(position == 0) %>% # there can be multiple fields per goid, need to take one
                 select(goid, fieldname_mag),
           by = "goid") %>%
    rename(year = degree_year) %>%
    group_by(year, fieldname_mag) %>%
    summarise(nb = n(),
              .groups = "drop")

nb_ncses_fld0 <- data_graduates_fld0 %>%
  group_by(year, fieldname0_mag) %>%
  summarise(nb = sum(nb), .groups = "drop")

d_fld0 <- nb_pq_fld0 %>%
  filter(year >= 1980 & year <= 2020) %>%
  rename(nb_pq = nb) %>%
  left_join(nb_ncses_fld0 %>%
              rename(nb_ncses = nb),
            by = c("year", "fieldname_mag" = "fieldname0_mag")) %>%
  pivot_longer(cols = starts_with("nb_"),
               values_to = "count",
               names_to = "src") %>%
  mutate(src = ifelse(src == "nb_pq", "ProQuest", "NCSES"))
```

```
keep_fields <- c("chemistry", "computer science",
                 "psychology", "economics", "sociology", "mathematics",
                 "political science", "physics", "engineering")

d_fld0 %>%
  filter(fieldname_mag %in% keep_fields) %>%
  ggplot(aes(x = year, y = count)) +
  geom_line(aes(linetype = src, color = src)) +
  facet_wrap(~fieldname_mag) +
  theme(legend.position = "bottom") +
  scale_x_continuous(breaks = seq(1985, 2015, 10))
```



```
d_fld0 %>%
  filter(fieldname_mag %in% c("biology", "psychology", "chemistry")) %>%
  ggplot(aes(x = year, y = count)) +
  geom_line(aes(linetype = src, color = src)) +
  facet_wrap(~fieldname_mag) +
  theme(legend.position = "bottom") +
scale_x_continuous(breaks = seq(1985, 2015, 10))
```



Comments

- Looks good overall, worrisome: Biology post-2010, psychology
- Biology: the gap could also come from the fact that we do not cover medicine in proquest
- Psychology: the aggregates by social sciences seem to match up between pq and ncses; have we misclassified some other soc sci psychology?
  - we cover the social sciences well *overall* (see figures above)
  - but the match in the aggregate could also just come from the other fields having slight undercoverage, which dominates the excess numbers in proquest. but this is implausible since we cover the other social sciences well
  - the social sciences from SED also includes "other social sciences"; perhaps they are mostly in pschology in ProQuest?

What is happening to psychology?

- Hypothesis 1: We have classified at least some PQ graduates in psychology which in SED are in "other social sciences"
  - How to test it: compare counts of psychology including and excluding other social sciences to ProQuest
  - In the aggregate cross-section over time

    - Across universities: the correlation between the number of psychology graduates from ProQuest and from the SED should be higher when we include the other social sciences in the counts.

- Hypothesis 2: Proquest covers some other people with special degrees
    - Test: drop these people from the counts in ProQuest

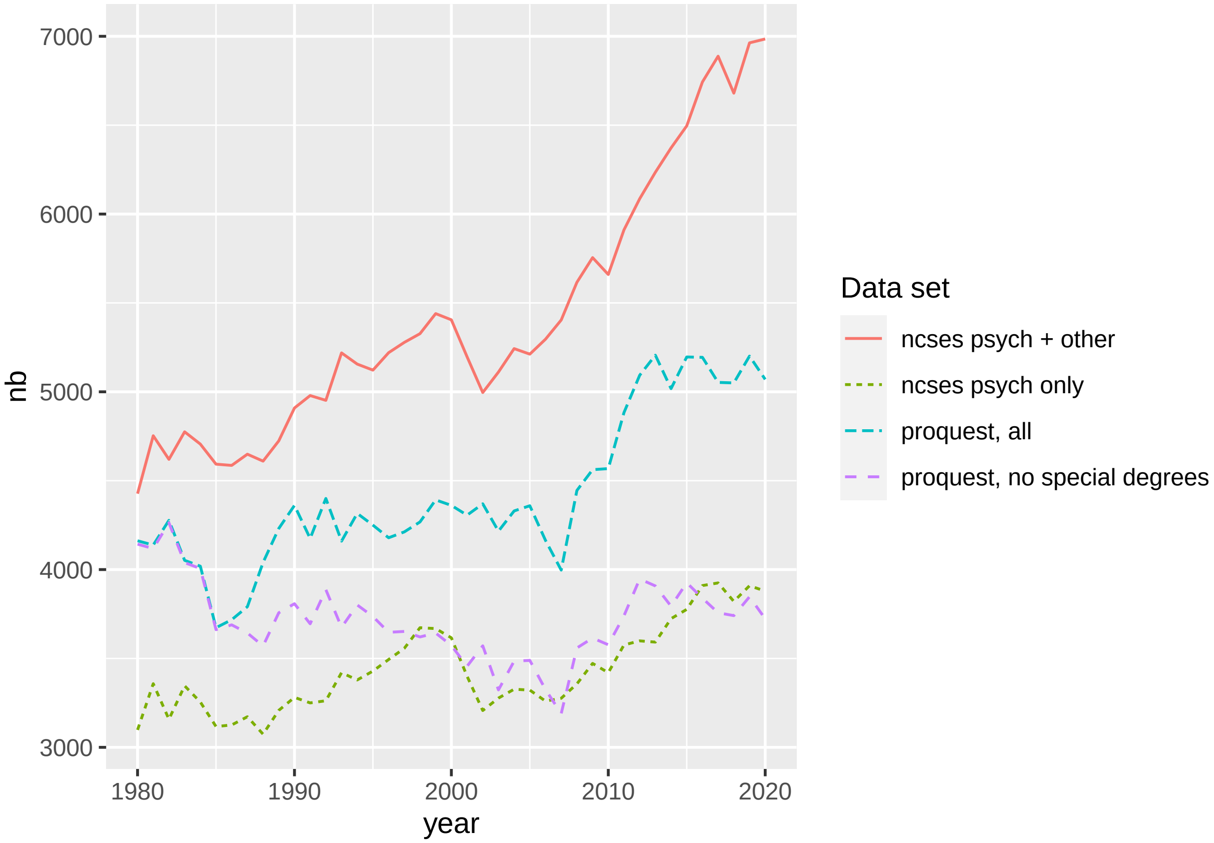
Compare the aggregates

```
d_psych_ncses <- bind_rows(
  data_graduates_fld0 %>%
    filter(fieldname0_mag == "psychology") %>%
    mutate(type = "ncses psych only"),
  data_graduates_fld0 %>%
    filter(fieldname0_mag %in% c("psychology", "other socsci")) %>%
    mutate(fieldname0_mag = "psychology") %>%
    group_by(unitid, year, fieldname0_mag) %>%
    summarise(nb = sum(nb), .groups = "drop") %>%
    mutate(type = "ncses psych + other")
) %>%
  filter(year >= 1980)
```

```
special_degrees <- c("Psy.D.", "Ed.D.", "D.Ed.")

ls_pq_psych <- list(
  base = authors,
  no_special_degrees = authors %>%
    filter(!(degree_level %in% special_degrees))
)

ls_pq_psych <- map(
  .x = ls_pq_psych,
  .f = ~.x %>%
    inner_join(fields_mag %>%
                 filter(position == 0) %>%
                 select(goid, fieldname_mag),
               by = "goid") %>%
    filter(fieldname_mag == "psychology") %>%
    rename(year = degree_year) %>%
    group_by(year, fieldname_mag) %>%
    summarise(nb = n(),
              .groups = "drop") %>%
    rename(fieldname0_mag = fieldname_mag)
)
```

```
d_aggregate <- bind_rows(
  d_psych_ncses %>%
    group_by(year, fieldname0_mag, type) %>%
    summarise(nb = sum(nb), .groups = "drop"),
  ls_pq_psych$base %>%
    mutate(type = "proquest, all"),
  ls_pq_psych$no_special_degrees %>%
    mutate(type = "proquest, no special degrees")
)
```

```
d_aggregate %>%
  filter(year >= 1980 & year <= 2020) %>%
  ggplot(aes(x = year, y = nb)) +
  geom_line(aes(linetype = type, color = type)) +
  guides(color=guide_legend(title="Data set"),
         linetype=guide_legend(title="Data set"))
```



Conclusion

- Dropping graduates with the special degrees brings the aggregate lines much closer together
  - On average, the counts are still slightly higher in ProQuest than in NCSES
- See below; cross-section may still be problematic but we need to compare with other fields
- Implication: drop these people from linking / from the analysis

In what follows, I tried out some more things. It still looks as if in the cross-section, proquest over-predicts the actual number of graduates.

Compare across universities

```
crosswalk_ipeds <- tbl(con, "links_to_cng") %>%
  filter(from_dataset == "pq") %>%
  select(unitid, proquest_id = from_id, link_score) %>%
  collect()

cng_institutions <- tbl(con, "cng_institutions") %>% collect()

psych_pq <- authors %>%
```

```r
  filter(degree_level != "Ed.D." & degree_level != "Psy.D.") %>%
  filter(grepl("United States", location)) %>%
  inner_join(fields_mag %>% # NOTE: drops graduates w/o field in mag, most of them "education"
               filter(position == 0) %>% # there can be multiple fields per goid, need to take one
               select(goid, fieldname_mag),
             by = "goid") %>%
  rename(year = degree_year) %>%
  filter(fieldname_mag == "psychology") %>%
  group_by(year, university_id) %>%
  summarise(nb = n(),
            .groups = "drop") %>%
  left_join(crosswalk_ipeds %>%
              select(-link_score),
            by = c("university_id" = "proquest_id")) %>%
  select(unitid, year, nb_pq = nb) %>%
  filter(!is.na(unitid))

dk <- d_psych_ncses %>%
  select(-fieldname0_mag) %>%
  mutate(type = ifelse(type == "ncses psych only", "nb_ncses_psych_only", "nb_ncses_psych_add_other")) %
  pivot_wider(names_from = type, values_from = nb, values_fill = 0)

d_micro <- psych_pq %>%
  left_join(dk, by = c("unitid", "year")) %>%
  filter(!is.na(nb_ncses_psych_only))
```

Regression

- If they lined up perfeclty, intercept = 0 and coefficient = 1
- Use log to account for different sizes of the programs

```r
r3_2_1_unis <- cng_institutions %>%
  filter(basic2021 %in% 15:17) %>%
  pull(unitid)

d_est <- d_micro %>% #%>% filter(unitid %in% r3_2_1_unis)
  mutate(research_intensive = ifelse(unitid %in% r3_2_1_unis, 1, 0)) %>%
  filter(year %in% 1990:2015)

m_psych <- feols(log(nb_pq) ~ log(nb_ncses_psych_only) ,
                 weights = d_est$nb_ncses_psych_only,
                 data = d_est)
```

```
## NOTES: 467 observations removed because of 0-weight.
##        467 observations removed because of infinite values (RHS: 467).
```

```r
m_add <- feols(log(nb_pq) ~ log(nb_ncses_psych_add_other),
               weights = d_est$nb_ncses_psych_only,
               data = d_est)
```

```
## NOTES: 467 observations removed because of 0-weight.
##        280 observations removed because of infinite values (RHS: 280).
```

```r
estlist <- list(psych = m_psych, add_other = m_add)

etable(estlist)
```

```
##                                              psych          add_other
## Dependent Var.:                        log(nb_pq)         log(nb_pq)
##
## (Intercept)                   0.3120*** (0.0380) 0.2052*** (0.0403)
## log(nb_ncses_psych_only)      0.8656*** (0.0127)
## log(nb_ncses_psych_add_other)                    0.7905*** (0.0118)
## _____________________________ __________________ __________________
## S.E. type                                    IID                IID
## Observations                               5,116              5,116
## R2                                       0.47774            0.46740
## Adj. R2                                  0.47764            0.46730
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```