Data description

Data sources and access

The analysis uses data from several sources, all containing Norwegian administrative data. Below we detail the sources and contents of the register data sets.

The data on applications and admission cutoffs comes from the Norwegian university and college admission service (NUCAS, http://www.samordnaopptak.no/info/english/). Our request for application data has been handled by Geir Sverre Andersen. More details about use of microdata (including contact information) is provided at http://www.samordnaopptak.no/info/om/personvern-og-sikkerhet/ (only in Norwegian).

Data on completed education, earnings and demographics comes from different national databases maintained by Statistics Norway (SN, http://www.ssb.no/). The procedure for obtaining microdata is described at http://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning.

Application data

The data are administrative data from NUCAS. Applications for higher education are submitted to NUCAS, which handles the appliaction process and stores the application data. The unit of observation is applicant*year*program*institution. Applicants can rank up to 15 program*institution (10 in some years).

Variable	Description
Year	Year of application (1998-2004)
Personal id	Unique national personal id, allowing matching across data sources
Rank	Rank in the application, from 1 (best) to a maximum of 15 (for applicants with 15 ranked programs)
Program id	Unique identifer of program*institution, consisting of an instituion id and program id
NUCAS Institution id	NUCAS' coding of institutions
Application score, main quota	The applicant*program*institution-specific application score for the main quota, calculated by NUCAS
Application score, young quota	The application score for the applicants eligible to compete in the young quota, otherwise as above
Documentation	Whether necessary documentation is provided
Qualified	Whether the applicant satisfies the formal requirements (subjects from upper secondary)
Offer	Whether an offer was sent
Quota	If an offer was sent, from what quota the applicant got an offer
Accepted	Whether the offer was accepted
Started	Whether the applicant actually showed up at the start of the program
Program offered	Whether the program was actually offered or cancelled by the offering institution

Data on admission cutoffs

Calculated and published annually by NUCAS after completion of the admission process[1]. The unit of observation is program*instituion. The files provide cutoffs from two admission rounds, the main admission round (late July, when first offers are sent out the applicants, denoted "HOVED" by NUCAS) and the final admission round (early August, shortly before the start of the semester, denoted "VARA" by NUCAS). We use the cutoffs from the final admission rounds. In some cases where these do not exist, we impute with the admission cutoffs from the main round, if these are available. Cutoffs varies between quotas, and the available quotas varies over time and between institutions in any given year. We retain the cutoffs for the main quota (where all applicants can compete) and for the quota for young applicants (exact definition has changed over time).

Variable	Description
Year	Year of application (1998-2004)
Program id	Unique identifer of program*institution, consisting of an instituion id and program id
NUCAS Institution id	NUCAS' coding of institutions
Cutoff, main quota	Application score of last admitted applicant in the main quota
Cutoff, young quota	Application score of last admitted applicant in the young quota

Data on completed educations

The data comes from the Norwegian national education databased[2]. The source of the information is mostly reports of all enrolled students and completed educations from all Norwegian

---

[1] http://www.samordnaopptak.no/info/opptak/poenggrenser/poenggrenser-tidligere-ar/, only in Norwegian

[2] http://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning/utdanning

schools, colleges and universities. A Norwegian research institution collects data on collected PhDs. Data on education abroad is provided by the Norwegian state educational loan fund and taken from surveys to immigrants. A detailed description of the data base in Norwegian is available at http://www.ssb.no/a/metadata/om_datasamlinger/nudb/nudb.html. The data we use for completed education are the yearly files for the population´s level of education per 1 October[3]. Data on instiution of highest completed education comes from a detailed data set of completed educations[4].

The classification of applications used by NUCAS in the application data is different from SN's classification of completed educations. However, both sources provide detailed classifications. SN classifies educations by level, field and detailed program classification according to the national classification NUS2000, which is comparable to UNESCO's ISCED classification. The 50,083 individuals in our baseline estimation sample are recorded with 1152 different education codes. NUCAS classifies programs by type of education (e.g. teachers college, nursing, engineering) and field (e.g. social science, teaching, health, technical), and also provide program names. Using this information, we aggregate both classifications to our broad fields. We exclude some applicants to programs that we are unable to match to completed fields, see below under sample construction. We use the names of institutions to combine data on applied and completed institutions.

The unit of observation is individual*year.

Variable	Description
Year	Year of observation (1998-2014)
Personal id	Unique national personal id, allowing matching across data sources
Highest education	SN's detailed classification of the individuals highest recorded completed education
Date of highest educ.	Month of completion
Institution of higest educ.	SN's coding of institution awarding the highest recorded education

Earnings data

The data comes from earnings registers maintained by SN[5]. The source of the information is tax reports. These, in turn, are mostly based on automatic reporting from employers to the tax authorities. The unit of observation is individual*year.

Variable	Description
Year	Year of observation (1998-2014)
Personal id	Unique national personal id, allowing matching across data sources
Labor earnings	Wages and earnings from self-employment, some transfers replacing such earning (e.g. maternity leave, but not unemployme

Demographic data

Residence data and the parent-child link comes from population registers maintained by Norwegian tax authorities and Statistics Norway. Data on parents completed education and earnings (see above) is used to create a file of background characteristics. The unit of observation is the individual (i.e., the applicant).

Variable	Description
Personal id	Unique national personal id, allowing matching across data sources
Mother's highest education	SN's detailed classification, observed at applicant's age 16
Father's highest education	SN's detailed classification, observed at applicant's age 16
Father's earning	Father's pensionable earnings (approx. labor earnings), average of earnings at applicant's age 16 and 19
Municipality	Applicant's municipality of residence at age 16

---

[3] http://www.ssb.no/en/utdanning/statistikker/utniv/aar/2015-06-18?fane=om#content

[4] http://www.ssb.no/a/metadata/om_datasamlinger/nudb/nudb_variabelliste.html#tabell_F_UTD_DEMOGRAFI, only in Norwegian

[5] http://www.ssb.no/en/omssb/tjenester-og-verktoy/data-til-forskning/inntekt

Sample selection and construction

1. Combining the raw application data, pooling all years of applications gives a data set of 366k applicants submitting 604k applications, with 3.0M programs applied altogether
2. Removing applications we cannot use leaves us with 244k applicants, 360k applications, 1.9M programs. Applications removed include:
   (a) "Invalid applications": Denoted as such by NUCAS in personal communications, these are applications that will never be considered because of missing documentation, applicants that do not satisfy formal requirements for the program applied, or because the program is withdrawn (not offered) by the offering institution
   (b) fall-back applications for introductory semester (applications for introductory semester only and not for any further studies, relevant only if an applicant gets no other offer for any of the up to 15 applied programs)
   (c) applications with no data on application score (after imputing with observed application score from same applicant same year)
3. Further restricting the data gives our "population of applicants": 218 824 applicants applying for 1157k programs (83k apply for only one field). Restrictions imposed:
   (a) Keeping only first observed application - this gives us one observed application per individual
   (b) Keep applicants with no higher education when applying - this helps in interpreting the counterfactual higher education of the applicant by excluding the possibility of a pre-existing counterfactual
   (c) Drop applicants admitted in special quotas and, within applications, programs missing data on bounds or ranked lower than a such program - data on bounds is necessary to construct our instruments, we also need to know if the applicant is predicted to be offerend any more-preferred program, finally it is harder to identify the relevant bounds for the few applicants competing in special quotas
4. Within our population of applicants, we construct a data set of 66 796 applicants on the margin between preferred field j and next best field k:
   (a) We drop dominated programs, i.e., programs ranked below one with a lower admission bound, and which thus will never be offered
   (b) We then aggregate programs to fields, construct pairs of preferred and next best fields j,k in the applications and keep pairs where the applicant is predicted to be offered j (k) and would have been offered k (j) if her application score was lower (higher)
   (c) When we observe two margins for an applicant we retain the highest-ranked margin, e.g., if an applicant is predicted to be offered k (rank 2), but could have been offered j (rank 1) if her application score was higher or l (rank 3) if it was lower, we will use the j vs k comparison, and discard the k vs l comparison
5. Finally we exclude some applicants based on their counterfactual field or outcomes to arrive at our baseline estimation sample of 50 083 applicants. Applicants excluded:
   (a) Applicants with ill-defined preferred or next best field: 2889
   (b) Applicants with missing data on completed education or earnings (emigrated or dead): 373
   (c) Applicants with k=medicine: 73
   (d) Applicants with no college or ill-defined completed field: 13 378

Description of final analysis file

```
Contains data from ../wk48/data.dta
  obs:       115,734
 vars:            29                          5 Feb 2016 14:02
 size:     7,059,774
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
appyear         int     %8.0g                 year of application
fnr             double  %011.0f               person id
female          byte    %8.0g                 1 if female
age             byte    %8.0g                 categories: <18, one-year
                                                categories 18-29, 30-39, 40+
immpar          byte    %8.0g                 both parents are immigrants
mor_highed      byte    %9.0g                 mother has higher eductation
far_highed      byte    %9.0g                 father has higher eductation
far_earnings    float   %9.0g                 average of father's earnings at
                                                applicant age 16 and 19
gpa             float   %9.0g                 application score
gpa2            float   %9.0g                 application score squard
gpa3            float   %9.0g                 application score cubic
f1              byte    %20.0g     field_agg
                                              preferred field
field_det_f1    byte    %27.0g     field_det
                                              preferred detailed field
inst_f1         byte    %8.0g      inst_f1    institution of preferred field
d_f1            float   %9.0g                 distance from admission cut-off of
                                                preferred field
z1              byte    %9.0g                 above cut-off
offer_f1        byte    %9.0g                 recorded offered preferred field
f0              byte    %20.0g     field_agg
                                              next-best field
field_det_f0    byte    %27.0g     field_det
                                              next-best detailed field
inst_f0         byte    %8.0g      inst_f1    institution of next-best detailed
                                                field
peer_pred       float   %9.0g                 predicted average peer gpa
pred_exp        byte    %9.0g                 predicted potential experience 8
                                                yrs after applying
inst_pred       byte    %9.0g      inst_f1    predicted institution
f               byte    %14.0g     field_agg
                                              field of highest completed
                                                education 8 yrs after applying
f8field_comp_~t byte    %8.0g      field_det
                                              detailed field of highest
                                                education after 8 yrs
inst_compl      byte    %8.0g      inst_f1    institution that awarded highest
                                                education
f8yrc           float   %9.0g                 earnings 8 yrs after applying
lnf8yrc         float   %9.0g                 log f8yrc
sample          byte    %9.0g                 indicate baseline sample: f8yrc<.
                                                & f<99 & f0<10
--------------------------------------------------------------------------------
Sorted by:
```

Program files

1. **replication**.do: Master file, runs all other files in correct order. Calls the following files, in that order.
2. **data_outcomes**.do: Manages outcome data, combines data sources into one file. Calls:
   (a) **classify_completed**.do: Classifies data on completed education into fields
3. **data_applications**.do: Manages application data, defines population, merges with outcome data, creates complete file with applicant*course applied and all data. Calls:
   (a) **replace_navn**.do: Makes text edits to facilitate merging on course names with data on admission cut-offs
   (b) **classify_applied**.do: Classifies applied courses into fields, also codes nominal duration
   (c) **vlabels**.do: Labels fields
   (d) **ilabel**.do: Labels institutions
4. **data**.do: Creates final analysis file, with applicant as unit of observation and variables for preferred and next-best fields (as well as background and outcomes)
   (a) **varlabels**.do: Add labels to variables in analysis data set
5. **descriptives**.do: Makes Figures I and II**, replication**.do then makes Table III
6. **rdgraph**.do: Makes Figures III, IV, IV, VI and Xa
7. **estimation**.do: Estimates pay-offs for several specifciations, saves results for use by later files
8. **esttables**.do: Uses saved estimation results, makes Tables IV, B.I, B.IV and B.VI
9. **estfigures**.do: Uses saved estimation results, makes Figures VII, VIII, IX, Xb, XI, XII, B.I, B.IV, B.VI and B.VII
10. **pooled**.do: Estimates OLS and 2SLS pay-offs not fixing counterfactual field, makes Figures B.II and B.III
11. **nonsep**.do: Estimates (non-separable returns) to field*next-best*predicted institutions, **replication**.do then makes Table B.V
12. **estimation_det**.do: Estimates pay-offs to subfields, makes Figure B.V
13. **testmodels**.do: Makes Tables B.II and B.III