# Creating Matched Comparison Groups Using SAS

Colby J. Stoever

TAMUCC

# When should you used matched comparison groups?

- When Random Assignment is
  - Impossible
  - Improbable
  - Unethical
  - Never thought of by Program Director
- When you are asked to evaluate a project after it is completed or after program participants have been assigned to conditions.

# Why should you used match comparison?

- It is good methodology: We need to control for as many nuisance variables as possible.
  - Apples to Apples
  - Program evaluations need to use good methodology too.
- Conclusions will be more sound.
  - You will know if your program is effective.
  - You will know how to change your program.

[redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 48 | 32 | 30 | 28 | 100% | 67% | 63% | 58% |
| 20089 | - | 71 | 56 | 51 | - | 100% | 79% | 72% |
| 20099 | - | - | 78 | 63 | - | - | 100% | 81% |

All first time Freshmen enrolled in [redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 238 | 208 | 188 | 100% | 61% | 54% | 48% |
| 20089 | - | 476 | 318 | 273 | - | 100% | 67% | 57% |
| 20099 | - | - | 487 | 335 | - | - | 100% | 69% |

All first time Freshmen enrolled in BIOL 1[redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 259 | 243 | 226 | 100% | 67% | 63% | 58% |
| 20089 | - | 476 | 375 | 342 | - | 100% | 79% | 72% |
| 20099 | - | - | 487 | 393 | - | - | 100% | 81% |

Projected increase in number of students [redacted]

| | Enrolled term | | | |
|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 |
| 20079 | - | 21 | 35 | 38 |
| 20089 | - | - | 57 | 69 |
| 20099 | - | - | - | 58 |
| Cumulative | | 21 | 92 | 166 |

Tuition and fee impact (assumes retained students enroll in 24 hours per year = $6,019 per student)

| | | | | |
|---|---|---|---|---|
| Cumulative | | 21 | 92 | 166 |
| | | $ 126,399 | $ 553,748 | $ 999,154 |

Program X

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 48 | 32 | 30 | 28 | 100% | 67% | 63% | 58% |
| 20089 | - | 71 | 56 | 51 | - | 100% | 79% | 72% |
| 20099 | - | - | 78 | 63 | - | - | 100% | 81% |

All first time Freshmen enrolled in BIOL 1406

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 238 | 208 | 188 | 100% | 61% | 54% | 48% |
| 20089 | - | 476 | 318 | 273 | - | 100% | 67% | 57% |
| 20099 | - | - | 487 | 335 | - | - | 100% | 69% |

Program X

All first time Freshmen enrolled in BIOL 1406 IF retained @

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 259 | 243 | 226 | 100% | 67% | 63% | 58% |
| 20089 | - | 476 | 375 | 342 | - | 100% | 79% | 72% |
| 20099 | - | - | 487 | 393 | - | - | 100% | 81% |

Program X

Projected increase in number of students retained IF BIOL 1406 students were retained at

| | Enrolled term | | | |
|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 |
| 20079 | - | 21 | 35 | 38 |
| 20089 | - | - | 57 | 69 |
| 20099 | - | - | - | 58 |
| Cumulative | | 21 | 92 | 166 |

Tuition and fee impact (assumes retained students enroll in 24 hours per year = $6,019 per student)

| | | | | |
|---|---|---|---|---|
| Cumulative | | 21 | 92 | 166 |
| | | $ 126,399 | $ 553,748 | $ 999,154 |

# Methodology

- Post-Hoc Evaluation (designed and done after the program started and ended)
- Four Fall Cohorts
  - Separate analyzes to minimalize history and maturation threats to internal validity.
- Matched Groups Design (random selection)
  - Reduces the effects of Selection Bias
  - Gender
  - Ethnicity
  - College
  - Classification (Freshman, Sophomore, etc.)
  - Took a biology course (in Fall)
  - Took a math course (in Fall)

# Methodology (cont.)

- Main variables of interest (dependent variables)
  - Passed Math Course (in Fall)
  - Passed Biology Course (in Fall)
  - Retention rates
    - Statewide- at a 4 year colleges
    - TAMUCC
  - Graduation rate (when possible)
    - Statewide- 4 year degrees
    - TAMUCC

[redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 48 | 32 | 30 | 28 | 100% | 67% | 63% | 58% |
| 20089 | - | 71 | 56 | 51 | - | 100% | 79% | 72% |
| 20099 | - | - | 78 | 63 | - | - | 100% | 81% |

All first time Freshmen enrolled in [redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 238 | 208 | 188 | 100% | 61% | 54% | 48% |
| 20089 | - | 476 | 318 | 273 | - | 100% | 67% | 57% |
| 20099 | - | - | 487 | 335 | - | - | 100% | 69% |

All first time Freshmen enrolled in BIOL 1[redacted]

| | Enrolled term | | | | Retain rate | | | |
|---|---|---|---|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 | 20079 | 20089 | 20099 | 20109 |
| 20079 | 388 | 259 | 243 | 226 | 100% | 67% | 63% | 58% |
| 20089 | - | 476 | 375 | 342 | - | 100% | 79% | 72% |
| 20099 | - | - | 487 | 393 | - | - | 100% | 81% |

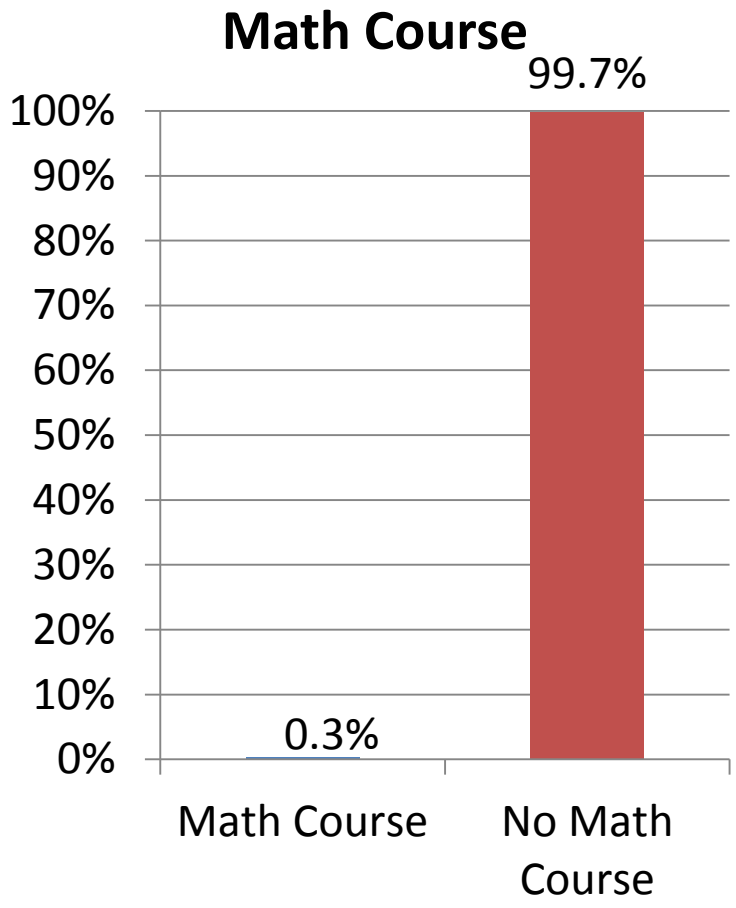
Projected increase in number of students [redacted]

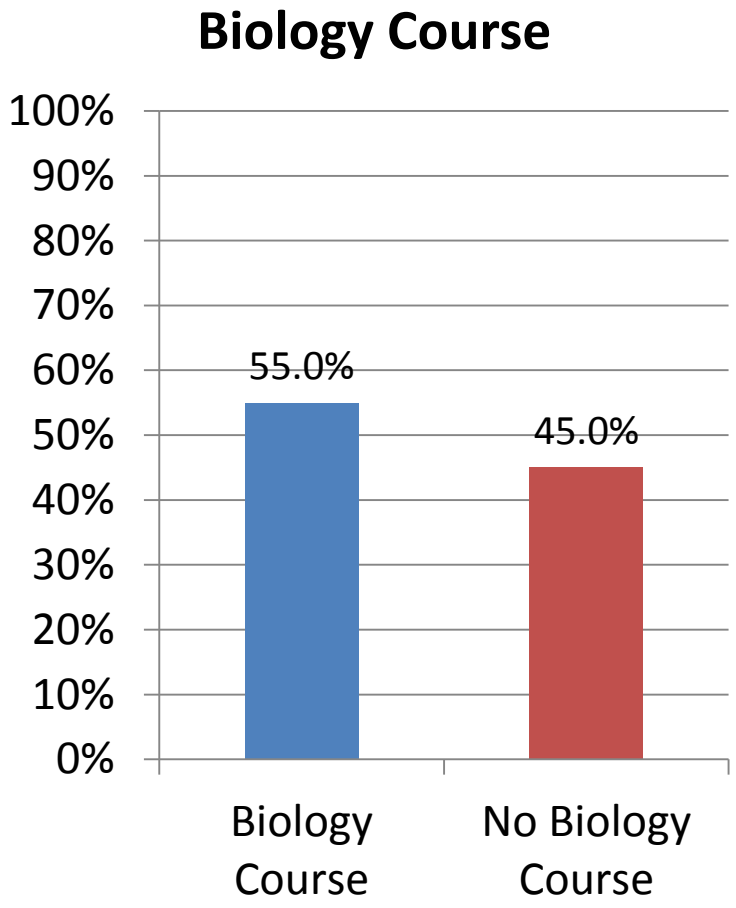
| | Enrolled term | | | |
|---|---|---|---|---|
| entry term | 20079 | 20089 | 20099 | 20109 |
| 20079 | - | 21 | 35 | 38 |
| 20089 | - | - | 57 | 69 |
| 20099 | - | - | - | 58 |
| Cumulative | | 21 | 92 | 166 |

Tuition and fee impact (assumes retained students enroll in 24 hours per year = $6,019 per student)

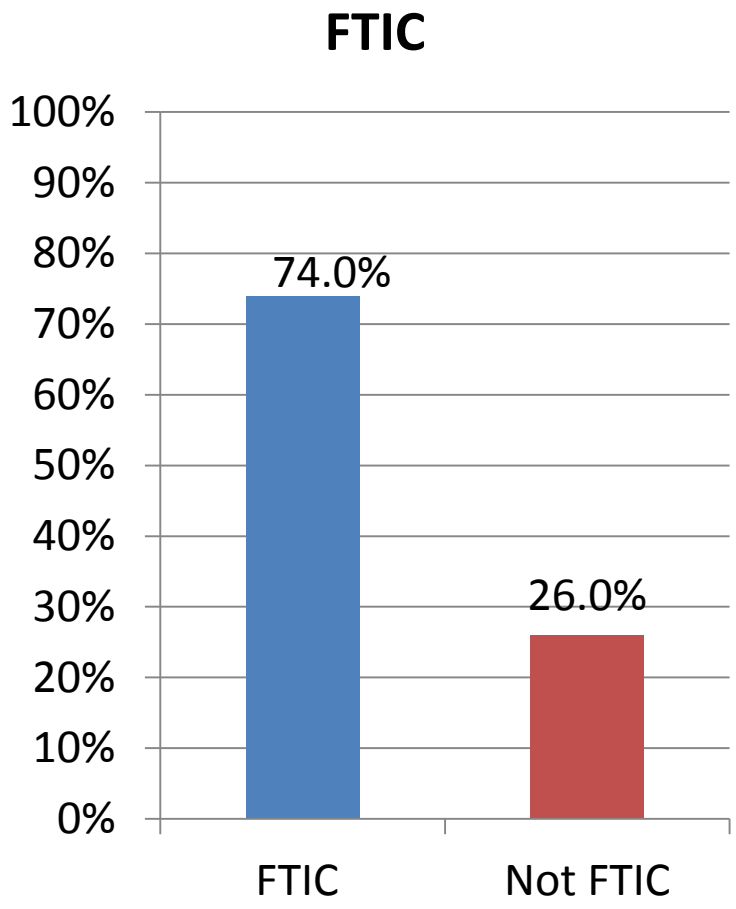
| | | | | |
|---|---|---|---|---|
| Cumulative | | 21 | 92 | 166 |
| | | $ 126,399 | $ 553,748 | $ 999,154 |

# Demographic Overall Cohorts

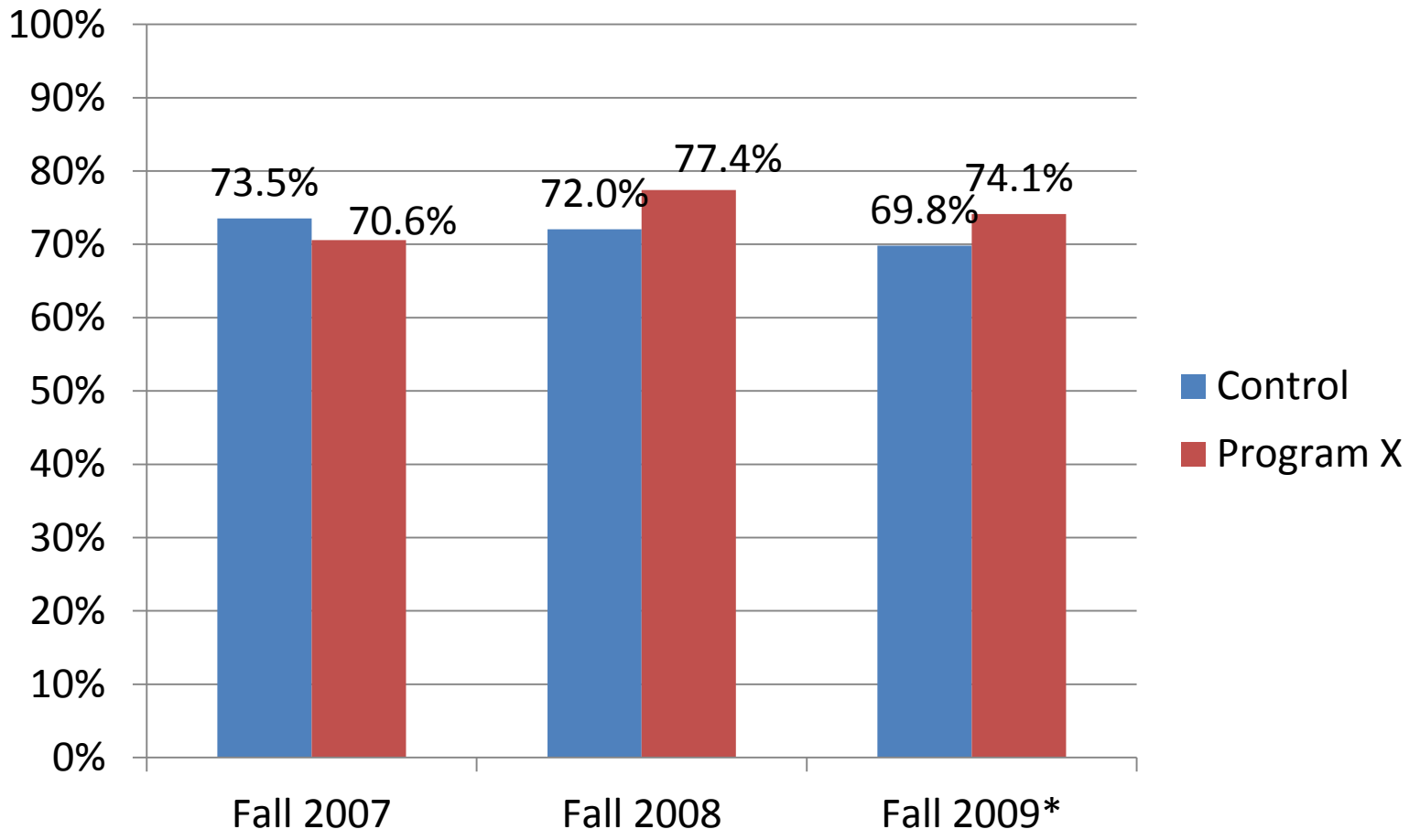


# Demographic Overall Cohorts



# One Year Persistence Rates



# When should I used the proposed Methodology

- When random assignment can be used.
- When you have data on individuals who are not in the program or project you are evaluating.
- When you do not have the sample to use propensity scoring (very different topic for very large DB's)

```
DATA CTRL&&Cohort&i (KEEP=STUSSN INDEX);
SET  Controlset&&Cohort&i (KEEP=STUSSN stugen stueth collegenum stuclass
Anybiol anyMath );
INDEX = stugen||collegenum||stuclass||Anybiol||anyMath||stueth;
*INDEX  = COMPRESS(INDEX ,'');
Run;


DATA case&&Cohort&i (KEEP=STUSSN INDEX);
SET expset&&Cohort&i (KEEP=STUSSN stugen stueth collegenum stuclass
Anybiol anyMath);
INDEX = stugen||collegenum||stuclass||Anybiol||anyMath||stueth;
*INDEX = COMPRESS(INDEX ,'');
Run;
```

```
PROC FREQ DATA= CTRL&&Cohort&i NOPRINT;
TABLES INDEX/LIST MISSING OUT=CTRLCNT (KEEP=INDEX COUNT
RENAME=(COUNT=CTRLCNT));
run;
```

```
PROC FREQ DATA= case&&Cohort&i NOPRINT;
TABLES INDEX/LIST MISSING OUT=caseCNT (KEEP=INDEX COUNT
RENAME=(COUNT=CaseCNT));
run;
```

```
DATA ALLCOUNT;
MERGE  casecnt(IN=A) CTRLCNT (IN=B);
BY INDEX;
IF CASECNT > 0;
IF A AND NOT B THEN CTRLCNT = 0;
_NSIZE_ = MIN(CASECNT,CTRLCNT);
IF _NSIZE_ GT 0;
Run;
```

```
PROC SQL;
CREATE TABLE WORK.ELIGIBLE_CONTROLS AS
SELECT *
FROM CTRL&&Cohort&i
WHERE INDEX IN (SELECT INDEX FROM ALLCOUNT);
PROC SORT DATA = WORK.ELIGIBLE_CONTROLS;
BY INDEX;
run;

PROC SQL;
CREATE TABLE WORK.ELIGIBLE_cases AS
SELECT *
FROM case&&Cohort&i
WHERE INDEX IN (SELECT INDEX FROM ALLCOUNT);
PROC SORT DATA = WORK.ELIGIBLE_cases;
BY INDEX;
run;
```

•

```
PROC SURVEYSELECT DATA = WORK.ELIGIBLE_CONTROLS
SAMPSIZE = ALLCOUNT
METHOD = SRS
SEED=542178
OUT=WORK.SELECTED_CONTROLS;
STRATA INDEX;
run;

PROC SURVEYSELECT DATA = WORK.ELIGIBLE_cases
SAMPSIZE = ALLCOUNT
METHOD = SRS
SEED=607329
OUT=WORK.SELECTED_cases;
STRATA INDEX;
run;
```

•

```
DATA CC&&Cohort&i (KEEP=stussn INDEX CCID);
SET WORK.SELECTED_CONTROLS (IN=A KEEP=stussn INDEX)
WORK.SELECTED_cases (IN=B KEEP=stussn INDEX);
IF A THEN CCID = 1; *CONTROLS;
ELSE IF B THEN CCID = 0; *CASES;
run;

PROC SORT DATA= CC&&Cohort&i;
BY INDEX CCID;
run;
```

```
DATA CC&&Cohort&i (KEEP=StuSSN INDEX CCID MATCHID);
SET CC&&Cohort&i;
BY INDEX CCID;
LENGTH CTKTR CAKTR IDXID 8 IDA $6 MATCHX $50 MATCHID 8;
ATTRIB MATCHID FORMAT =20.;
RETAIN CTKTR CAKTR IDXID;
IF CCID = 1 THEN CTKTR +1; * COUNTER FOR CONTROLS;
ELSE IF CCID = 0 THEN CAKTR +1; * COUNTER FOR CASES;
IF FIRST.INDEX THEN IDXID +1; * INCREASE INDEX COUNT;
IDA = COMPRESS(SUBSTR(INDEX,4,6),'*'); * RETAIN PART OF INDEX;
IDX= PUT(IDXID,4.); * COUNTER (CHARACTER);
IF CCID = 1 THEN MATCHX = IDX||IDA||CTKTR; * MATCHID FOR CONTROLS;
ELSE IF CCID = 0 THEN MATCHX = IDX||IDA||CAKTR;* MATCHID FOR CASES;
MATCHX = COMPRESS(MATCHX,'');
MATCHID = INPUT(MATCHX, 20.); * NUMERIC MATCHID;
run;
```

# Contact Information


- Colby Stoever
- colbystoever@tamucc.edu