Differences Due to Subsampling Algorithm Variants
Prepared by Claude S. Fischer, January 25, 2018

The subsampling algorithm is a procedure to sample up to from names from the list of names provided by the respondent. It is a nota random subsample, but one that goes to specific name-eliciting questions in a specific order to select names across a range of topical areas. See the Main Survey and other documentation for a description.

The purpose of this memo is to provide preliminary information about an unplanned change that occurred during Wave 1 data collection, some observable differences that have been noticed thus far, and implications for how the data are to be handled and analyzed in future. Specifically, the algorithm that is used to select ties within the network for follow-up questions as part of the alter-level subsample, was changed during an update in the system that took place approximately halfway through Wave 1 data collection. The update occurred in order to allow the system to correctly identify participants as Facebook recruits (an unplanned sampling approach that was used to bolster sample size among the young adults). When this update took place, the original algorithm was replaced, unintentionally, with a new algorithm.

The key difference is that the original [and wave 3] algorithm selects, within each specific name-eliciting question, the first name; or the second name if the first is ineligible[1] or already taken; or the third name if the first and second cannot be used, and so on. Once that is complete and either a name is taken or no name is available, the program moves on to the next key name-eliciting question and repeats the process. In the variant, if a name had been listed in a previous question *although not taken there*, then it would be ineligible in subsequent questions. Only novel names in subsequent questions were taken. See footnote for an example:[2]

The first part of this memo simply looks at the descriptive differences between the subsample names produced by the initial, 1st, intended subsampling algorithm (largely applied to the *first half* of our cases and the accidental second version (which applies to the second half of the sample and subsequently, ***all*** the Facebook cases).[3] The later part explores the analytical implications of the difference in some data analyses. The conclusion draws some implications about the consequences of the variation *as of this date*. An appendix provides a coding scheme that links prim_key (the respondent ID number) to whether he or she received the 1st or the 2nd

[1] A relative or partner in the household.
[2] Assume a respondent who gave these names to the first three name-eliciting questions that are part of the subsampling procedure:
b4a (confidant): A, B, C
b7c (emergency): A, B, D
b2a (sociable): E, F, G
The original (and wave 3) algorithm yields: A, B, E.
The variant yields: A, D, E. (It does so because B in answer to question b7c is eliminated by virtue of having already appeared in the b4a question that yielded A for the subsample.)
[3] Specifically: Young sample: version 1 = 94; version 2 = 391 (including all Facebook recruits); Older sample: version 1 = 453; version 2 = 221.

variant in wave 1. Everyone received the 2nd variant in wave 2. And all should receive the 1st in wave 3.

Summary of Descriptive Differences between Subsample Algorithms

In the Table below, I summarize a simple look at how the names provided by the initial subsample algorithm differ from those generated by the second one – within age groups. Although many of the differences are "significant" because of a large N, I essentially focus on (and present) the beta coefficients that are **both "significant" and show meaningful differences**–percentaged by algorithm 1 vs. 2. ***Variables not listed*** *(e.g., b5a, advisors; c1a_20, friends)* ***and blank entries*** *(e.g., b2a, for the young)* ***show no real differences between algorithms***.[4]

One thing you will see is that the names elicited by Alg 1 appear over more name-eliciting questions and thus have greater "exchange" multiplexity. Except for the stark difference in romantic partners appearing in the subsample, there is little difference by role relationships represented in the subsamples selected by Alg 1 and Alg 2. There is little difference in the descriptive check-off variables. Density is lower in the algorithm 2 networks. Looks like respondents were in more frequent contact with the names listed in algo. 1 than in algo 2. That's about it for substantial descriptive differences.

[4] This analysis is based on the w1 data as of about January 1—before the final corrections were made in late January, 2018.

ITEM	**YOUNG: ALG 1 v ALG 2**	**OLDER: ALG 1 v ALG 2**
Name-Eliciting (10pp diff.)		
a3b (rom part)	**24 cases v. 1 case**	**34 cases v 0 cases**
b2a (social)		**59% v. 32%**
b4a (confide)		**45 v. 34**
Exch Mltplx: Pct 3+ lists		**51 v. 32**
Role Relationships		
c1a_1 (spouse)	**54% v. 32%**	**11 cases v. 3 cases**
c1a_2 (rom part)	**50 v. 35**	**40 cases v. 0 cases**
c1a_10 (house mates)	**53 v. 34**	**14 cases v. 0 cases**
Alter Descriptions		
(Checked as "close")		(58% v. 53%)
Over 1 hour away	**25 cases v. 9 cases**	
Also caring for home		**39 cases v. 4 cases**
Density Matrix		
Mn. Values of n_x_n_y (i.e., Algo 2 has *lower* density than Algo 1[5])	2.1 v 2.3	2.1 v. 2.3
Subsample Descriptors		
d1a (how met)	(College: 11% v. 20%)	(Family: 29 v 25)
d1c (alter has spouse?)	(No: 34 v 42)	
d1f (see alter 1/wk-plus)	**42 v 30**	
d1g (phone 1/wk-plus)	**42 v 30**	(46 v 37)
d1k (rep feels obliged to)		(63 v 57)
d1l (alter would ask)	(47 v 54)	

[5] Higher value = knows less.

Differences in Modeling Due to Subsampling Variants

Given all the modeling we had already completed on the subsample for the media-communications paper, "Staying in Touch" (Fischer, Child, Lee), we reran the key models of that analysis:

(1) We added in a dummy variable (identifying which algorithm was used) [6] to examine the algorithms' ***independent*** effects on variation in frequency of face-to-face contact, phone contact, e-communications contact, and e-communications contact controlling for in-person and phone contact.

(2) We tested for potential interaction effects in these models using each of the following interaction terms:

algo_sub X alter romantic partner
algo_sub X alter named on b2a
algo_sub X alter named n b4a

and separately, for the e-communication outcome, we tested

algo_sub X freq face-to-face
algo_sub X freq phone contact

Each of these were conducted separately by age cohort.[7]

<u>The results of (1) suggest that</u>

(a) the subsampling variation has limited independent effects on the dependent variables (e.g., a sig neg. effect on in-person freq among the young; a sig pos. effect on e-communications among the "ever-use" young Rs);
(b) does not seen to affect other coefficients noticeably, *except* for the ref-flag and FB_flag cases, which is to be expected since all FB cases were algo #2. Even then, the variable's addition seems not to turn significant effects to insignificant or vice-versa.

<u>The tests of interaction effects (2*)</u> showed two sets of significant results, but one is misleading and the other modest.

(a) *algo_sub X romantic partner* is occasionally significant in the young sample,[8] but there was only one romantic partner among the subsample variant 2 cases for the young (and zero for the latter).

(b) There were significant interaction effects of *algo_sub X frequency phone contact* as a predictor of frequency of e-communications.[9] I took a look at how much was going on by running ANOVA's:

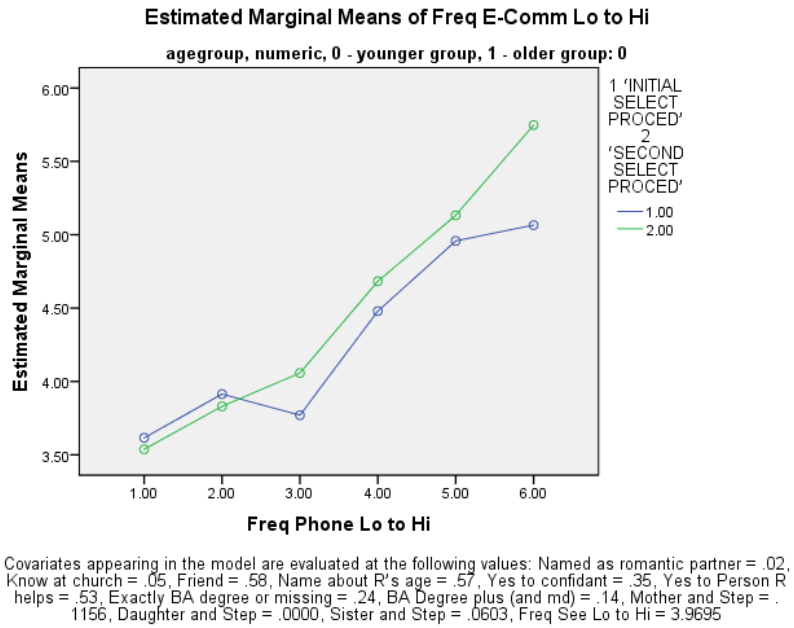
(i) E-communications frequency = *algo_sub X frequency phone contact*

[6] Labeled algo_sub in SC runs; labeled SUBSAMP_ALGO in CF runs.

(ii) E-communications frequency = *algo_sub X frequency phone contact,* controlling for 12 covariates that had been significant at p<.01 in the regression equations.

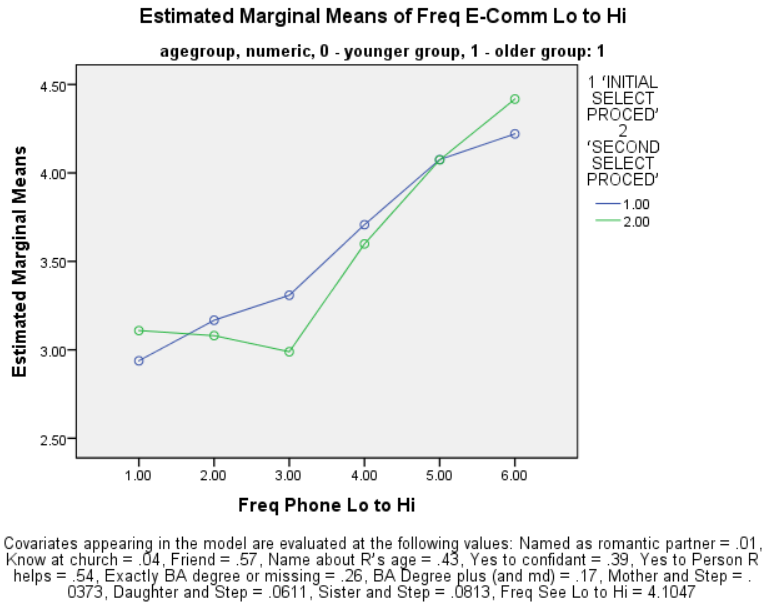
The figures below show the results: [cf: *Not much substantive.*]
These two figures illustrate the interaction of algorithm with a predictor of e-communications. Although statistically significant, the effects seem minor substantively. The blue lines are alter-communication frequencies for respondents with a variant 1 subsample, the green for a variant 2.

EST. EFFECT of phone frequency on e-communications frequency X SUBSAMP-ALGO, controlling for about a dozen covariates: YOUNG sample



EST. EFFECT of phone frequency on e-communications frequency X SUBSAMP-ALGO, controlling for about a dozen covariates: OLD sample

[7] (See Communication_algo_redo_SC_1-8-18.docx)
[8] (pp. 21-23 of \ Communication_algo_redo_SC_1-8-18.docx)
[9] (pp. 52, 55).



Conclusions and Recommendations:

1. The subsample variation is associated with age group (81% of the young versus 33% of the older respondents got variant #2). All Facebook recruits got variant #2.
2. Subsample variant #2 reduced the number of subsample ties we can observe compared to variant #1.
3. Variant #2 shifts the composition of those ties, of the subsample we have, moderately in the direction of alters who are:
    a. (a) singleplex—i.e., who appear in few or only one of the key eliciting questions and are thus probably less central to Rs' networks;
    b. (b) less often contacted by the Rs;
    c. (c) know fewer of the other alters; and
    d. (d) are less often co-residents.
4. Much of those differences are accounted for by the traits of relationships. (E.g., if fewer romantic partners or intimate friends are in the list, the list will be more peripheral.) So, the independent effect of the subsampling beyond the constitution of the subsample is modest.
5. The subsampling algorithm seems to have negligible or limited effects on substantive associations --- *at least as far as we can see to date.*
6. But*: A general good practice would be to introduce subsample variant as another methods-control variable (along with recruitment, mode, and interviewer) in analyses of the subsample data and to test for interaction effects* (especially considering the connection of subsample variant with age group and recruitment method).

CODING FOR WHICH RESPONDENTS ON Wave 1 GOT ALGORITHIM #1 (INTENDED) OR ALGORITHIM #2 (ACCIDENTAL).

Coding for subsampling variants follows:

This is an SPSS Code, but it is relatively simple:

(1) **IF** the prim_key ID variable is **GT** 7200000010400 **OR**
(2) the prim_key number is explicitly listed below, **THEN** are subsample = variation #2
(3) **ELSE**, the case is subsample = variation #1

```
IF (prim_key GT  7200000010400
or prim_key eq   5000000000127
or prim_key eq   6000000002795
or prim_key eq   7000000001373
or prim_key eq   7000000006034
or prim_key eq   7100000000382
or prim_key eq   7100000000656
or prim_key eq   7100000001915
or prim_key eq   7100000002688
or prim_key eq   7100000003372
or prim_key eq   7100000003841
or prim_key eq   7100000003852
or prim_key eq   7100000004333
or prim_key eq   7100000004437
or prim_key eq   7100000004675
or prim_key eq   7100000004855
or prim_key eq   7100000010174
or prim_key eq   7100000010428
or prim_key eq   7100000010529
or prim_key eq   7100000010726
or prim_key eq   7100000010995
or prim_key eq   7100000012055
or prim_key eq   7100000012342
or prim_key eq   7100000012635
or prim_key eq   7100000012656
or prim_key eq   7100000012932
or prim_key eq   7100000013009
or prim_key eq   7100000013224
or prim_key eq   7100000013401
or prim_key eq   7200000000424
or prim_key eq   7200000000515
or prim_key eq   7200000000552
or prim_key eq   7200000000642
or prim_key eq   7200000000848
or prim_key eq   7200000001266
```

```
or prim_key eq   7200000001393
or prim_key eq   7200000001853
or prim_key eq   7200000002209
or prim_key eq   7200000003021
or prim_key eq   7200000003101
or prim_key eq   7200000003232
or prim_key eq   7200000003257
or prim_key eq   7200000003361
or prim_key eq   7200000003477
or prim_key eq   7200000003557
or prim_key eq   7200000003734
or prim_key eq   7200000003799
or prim_key eq   7200000003966
or prim_key eq   7200000003967
or prim_key eq   7200000004132
or prim_key eq   7200000004432
or prim_key eq   7200000004477
or prim_key eq   7200000004636
or prim_key eq   7200000004701
or prim_key eq   7200000004805
or prim_key eq   7200000004964
or prim_key eq   7200000004998
or prim_key eq   7200000005224
or prim_key eq   7200000005401
or prim_key eq   7200000005431
or prim_key eq   7200000005518
or prim_key eq   7200000005681
or prim_key eq   7200000005948
or prim_key eq   7200000006095
or prim_key eq   7200000006206
or prim_key eq   7200000006239
or prim_key eq   7200000006533
or prim_key eq   7200000006544
or prim_key eq   7200000006582
or prim_key eq   7200000006708
or prim_key eq   7200000006750
or prim_key eq   7200000006944
or prim_key eq   7200000007004
or prim_key eq   7200000007058
or prim_key eq   7200000007153
or prim_key eq   7200000007368
or prim_key eq   7200000007407
or prim_key eq   7200000007512
or prim_key eq   7200000007522
or prim_key eq   7200000007555
or prim_key eq   7200000007592
or prim_key eq   7200000007621
or prim_key eq   7200000007868
or prim_key eq   7200000007924
or prim_key eq   7200000008038
or prim_key eq   7200000008043
or prim_key eq   7200000008079
```

```
or prim_key eq   7200000008132
or prim_key eq   7200000008252
or prim_key eq   7200000008261
or prim_key eq   7200000008307
or prim_key eq   7200000008431
or prim_key eq   7200000008479
or prim_key eq   7200000008598
or prim_key eq   7200000008661
or prim_key eq   7200000008737
or prim_key eq   7200000008774
or prim_key eq   7200000008799
or prim_key eq   7200000008842
or prim_key eq   7200000008912
or prim_key eq   7200000008919
or prim_key eq   7200000008999
or prim_key eq   7200000009081
or prim_key eq   7200000009121
or prim_key eq   7200000009193
or prim_key eq   7200000009246
or prim_key eq   7200000009339
or prim_key eq   7200000009490
or prim_key eq   7200000009540
or prim_key eq   7200000009730
or prim_key eq   7200000009762
or prim_key eq   7200000009791
or prim_key eq   7200000009904
or prim_key eq   7200000009958
or prim_key eq   7200000010010
or prim_key eq   7200000010068
or prim_key eq   7200000010131
or prim_key eq   7200000010430
or prim_key eq   7200000010482)
SUBSAMP_ALGO = 2.
IF (SUBSAMP_ALGO NE 2) SUBSAMP_ALGO = 1.
VARIABLE LABELS SUBSAMP_ALGO 1 'INITIAL SELECT PROCED' 2 'SECOND
SELECT PROCED'.
```