Embrace Your Missingness

part of a symposium:
Approaching Complex Research Designs From the Perspective of Missing Data
Association for Psychological Science, Chicago

William Revelle[a] & David M. Condon[b]

[a]Department of Psychology, Northwestern University, Evanston, Illinois
[b]Department of Medical Social Sciences, Northwestern University Chicago Illinois

Partially supported by a grant from the National Science Foundation: SMA-1419324

Slides at http://personality-project.org/sapa.html

NORTHWESTERN UNIVERSITY

NORTHWESTERN UNIVERSITY

Outline

The basic problem: Fidelity versus bandwidth

1. Many personality traits, interests and cognitive abilities are multidimensional and have complex structure.
   - To measure these, we need to have the precision that comes with many participants.
   - But we also need the bandwidth that comes with many items.
   - But participants are reluctant to answer very many items.
2. This has led to the quandary of should you give many people a few items or a few people, many items?
3. Our answer is to do both, but with a *Massively Missing Completely At Random* (MMCAR) data structure.
4. We refer to this technique as *Synthetic Aperture Personality Assessment* (SAPA) to recognize the analogy to synthetic aperture radio astronomy.

NORTHWESTERN UNIVERSITY

Breadth vs. depth of measurement

1. Factor structure of domains needs multiple constructs to define structure.
2. Each construct needs multiple items to be measured reliably.
3. This leads to an explosion of potential items .
4. But, people are willing to only answer a limited number of items.
5. This leads to the use of short and shorter forms (the NEO-PI-R with 300, the IPIP big 5 with 100, the BFI with 44 items, the TIPI with 10) to include as part of other surveys.

Example studies with subject/item tradeoffs

1. Eugene-Springfield sample (Goldberg & Saucier, 2016) gave several thousand items to 1,000 participants over 10 years. This sample has been the basis of the development and validation of the International Personality Item Pool (see ipip.ori.org) (Goldberg, 1999).
2. The Potter-Gosling internet project (outofservice.com) has given over 10,000,000 tests since 1997. Tend to be the 44 items of the Big Five Inventory (BFI) (John, Donahue & Kentle, 1991).
3. The Stillwell-Kosinski (mypersonality.org) Facebook application (no longer in service) gave 7,765 people the IPIP version of the NEO-PI-R with facets (300 items), 1,108,472 the IPIP NEO-PI R domains (100 items), and 3,646,237 brief (20 item) surveys. Cross linked to likes and Facebook pages.
4. The Personality project (now at sapa.project.org) has reported item statistics on more than 2,000 items for more than 200,000 participants (but used SAPA procedures).

Trading items for people: Studies, Items, People, Items x People

Table: Data sets vary in their sampling strategy but have similar total information

Study	N	Items	Items/ Person	Items* People
Eugene-Springfield	1,000	3,000?	3,000	$3 * 10^6$
Potter-Gosling	10^7	44	44	$4.4 * 10^8$
Stillwell-Kosinski	$4.5 * 10^6$	20-300	20-300	$1.7 * 10^8$
SAPA	$2 * 10^5$	2,000	100	$2 * 10^7$

But given basic statistical theory, is it worth while to increase the sample size so much. What is the effect of giving more items at the cost of reducing the sample size?

Many items versus many people

1. Not only do want many people, we also want many items.
2. Resolution (fidelity) goes up with sample size, N (standard errors are a function of $\sqrt{N}$)

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N-1}} \qquad \sigma_r = \frac{1-r^2}{\sqrt{N-2}}$$

3. Also increases as number of items, k, measuring each construct (reliability as well as signal/noise ratio varies as number of items and average correlation of the items)

$$\lambda_3 = \alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}} \qquad s/n = \frac{k\bar{r}}{(1-k\bar{r})}$$

4. Thus, we need to increase N as well as k. But how?

Can we increase N and n at the same time?

1. Frederic Lord (1955) introduced the concept of sampling people as well as items.
2. Apply basic sampling theory to include not just people (well known) but also to sample items within a domain (less well known).
3. Basic principle of Item Response Theory and tailored tests.
4. Used by Educational Testing Service (ETS) to pilot items.
5. Used by Programme for International Student Assessment (PISA) in incomplete block design (Anderson, Lin, Treagust, Ross & Yore, 2007).
6. Can we use this procedure for the study of individual differences without being a large company?
7. Yes, apply the techniques of radio astronomy to combine measures synthetically and take advantage of the web.

NORTHWESTERN UNIVERSITY

Subjects are expensive, so are items

1. In a survey such as Amazon's Mechanical Turk (MTURK), we need to pay by the person and by the item.
2. Why give each person the same items? Sample items, as we sample people.
3. Synthetically combine data across subjects and across items. This will imply a missing data structure which is
   - Missing Completely At Random (MCAR), or even more descriptively:
   - Massively Missing Completely at Random (MMCAR)
4. This is the essence of Synthetic Aperture Personality Assessment (SAPA) (Condon & Revelle, 2014; Condon, 2014; Revelle, Condon, Wilt, French, Brown & Elleman, 2016; Revelle, Wilt & Rosenthal, 2010).

3 Methods of collecting 256 subject * items data

a) 8 x 32 complete

```
46213634521143453443645331212414
21243623166421516154432261516513
51661351155165463622224435623344
11141343362332215612152135614522
25353121264561433433232246526411
61335154566424114612641225353516
24634342151536242425413513435116
11554654453123111162423325516334
```

b) 32 x 8 complete

```
46323114
25443314
43315423
26314145
41435614
42236153
62421344
35234443
34514166
63415154
44441342
13514321
66365663
12264546
31466135
32645514
66151251
14411441
62443636
33316236
63325425
11531126
61155546
33245361
52241654
63212356
24414663
63661414
45555223
14364433
```

c) 32 x 32 MCAR p=.25

```
..3..2..6.....4.55.......44.....
..........4..6..45..3.4..6....1
6..3.......6.1.....6.2.......5.6
....3522......5.3...3......5....
....3.2.2.......3..2......65..5.
.....51....324.........23......5
....552............25...54.5....
...44.4.5....3..6...6........3..
....61.523.2....2...........3...
5............42.4..6.5......61.
....3....3.6..1.4...1..5......5.
1....54..........2.4.33..6......
4.....52..6.....44.3...........2
..44...1........1..42....5..1...
..1..3.......2..3.521.......6...
.......3.142.........22......12.
.4...2..........3..162...4.....4
..4..6..3.4...1....5.33.........
5..........243..5....41......1..
..5..3..4...4.4..5..1.........4.
.....4.......3..5.2.....64.4..4.
...1.1.2...6....4......55....2..
.....3..2..53.....2..2.3.3......
......1...2..43...3.13........5.
...2.........4..54...2.3..62....
22.......332..1.....5......6....
...5..3.4.....3....5.241.......5
......63.1.......6...5..4..2....
..2.4..5..........52.4.....44...
2.55.....2.....6.....6.....55...
```

Synthetic Aperture Personality Assessment

1. Give each participant a random sample of *pn* items taken from a larger pool of n items.
2. Find covariances based upon "pairwise complete data".
3. Find scales based upon basic covariance algebra.
   - Let the raw data be the matrix $\boldsymbol{X}$ with N observations converted to deviation scores.
   - Then the item variance covariance matrix is $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}'N^{-1}$
   - and scale scores, $\boldsymbol{S}$ are found by $\boldsymbol{S} = \boldsymbol{K}'\boldsymbol{X}$.
   - $\boldsymbol{K}$ is a keying matrix, with $\boldsymbol{K}_{ij} = 1$ if $item_i$ is to be scored in the positive direction for scale j, 0 if it is not to be scored, and -1 if it is to be scored in the negative direction.
   - In this case, the covariance between scales, $\boldsymbol{C}_s$, is

$$\boldsymbol{C}_s = \boldsymbol{K}'\boldsymbol{X}(\boldsymbol{K}'\boldsymbol{X})'N^{-1} = \boldsymbol{K}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{K}N^{-1} = \boldsymbol{K}'\boldsymbol{C}\boldsymbol{K}. \quad (1)$$

4. That is, we can find the correlations/covariances between scales from the item covariances, not the raw items.

The basic tradeoff: standard errors and effective sample size

1. Standard error of correlations between any *single pair of items* is just
$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 2}}$$
2. However, simulation (and some theory) shows that the standard error of correlations of *synthetic* correlations of *scales of length k* decreases as a joint function of the number of items in the scale and the *inverse of the probability* of any two items being administered.
3. Effectively, this is because what ever causes error in any correlation does not aggregate across k independent pairs of items.

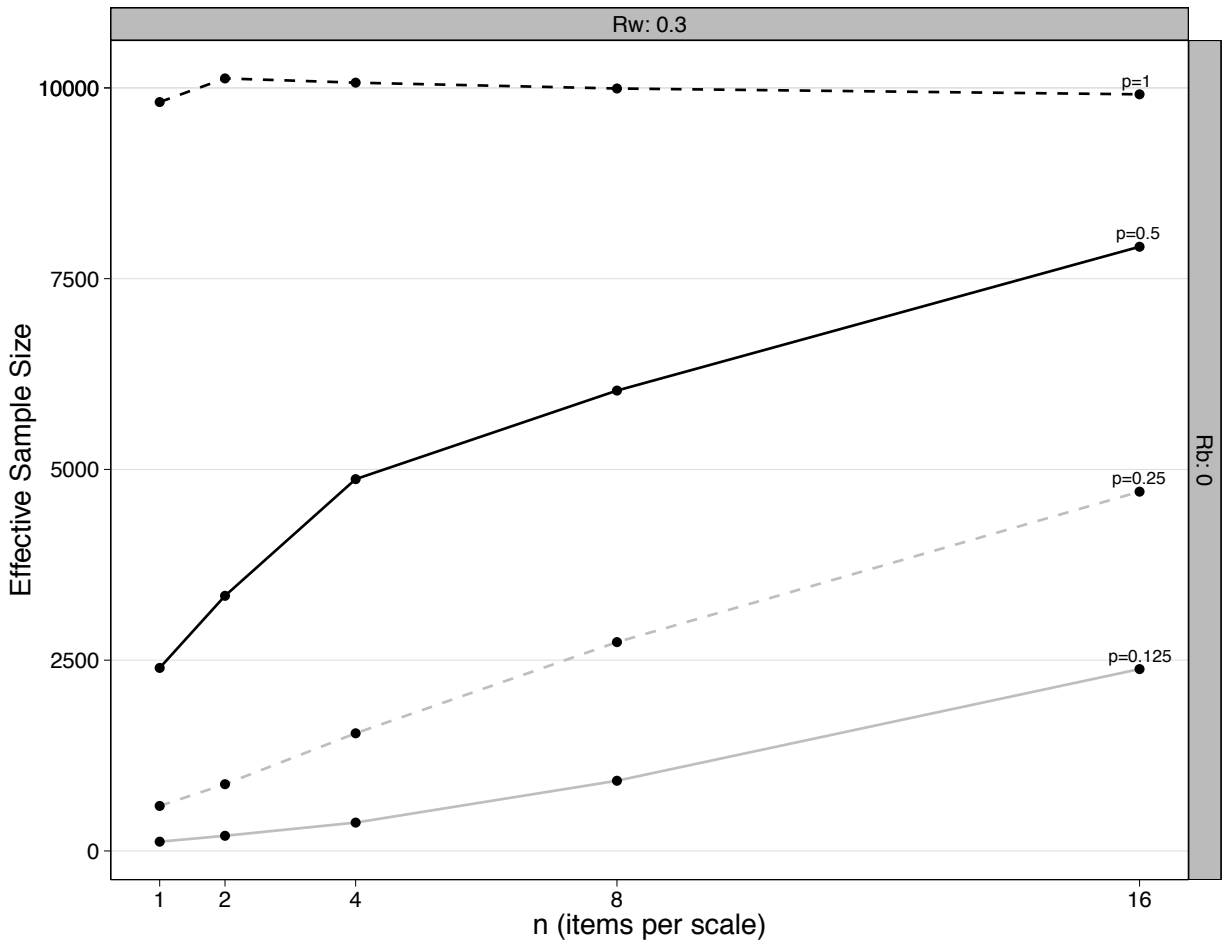
SAPA standard errors and effective sample size

1. When forming synthetic scales from MMCAR based items, the standard error of correlations decreases as a function of the Total number of subjects (N), the *the inverse of the percentage* of items sampled (p), and the *number of items forming the scale* (k).
2. Ashley Brown has shown this quite clearly by simulation (Brown, 2014).
3. A good way to visualize this is to examine the standard error of correlations as a function of N, p, and k.
4. An even more dramatic way is to plot the *Effective Sample Size* (N_{eff}) which because

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 2}} \text{ is merely } N_{eff} = \frac{(1 - r^2)^2}{\sigma_r^2} + 2$$

Effective sample size varies by the size of the composite scale.

Simulating N= 10,000 with probability of any item (Brown, 2014) ($p = .125, .25, .5$, or 1) and items in the composite 1 , 2, 4, 8, 16.



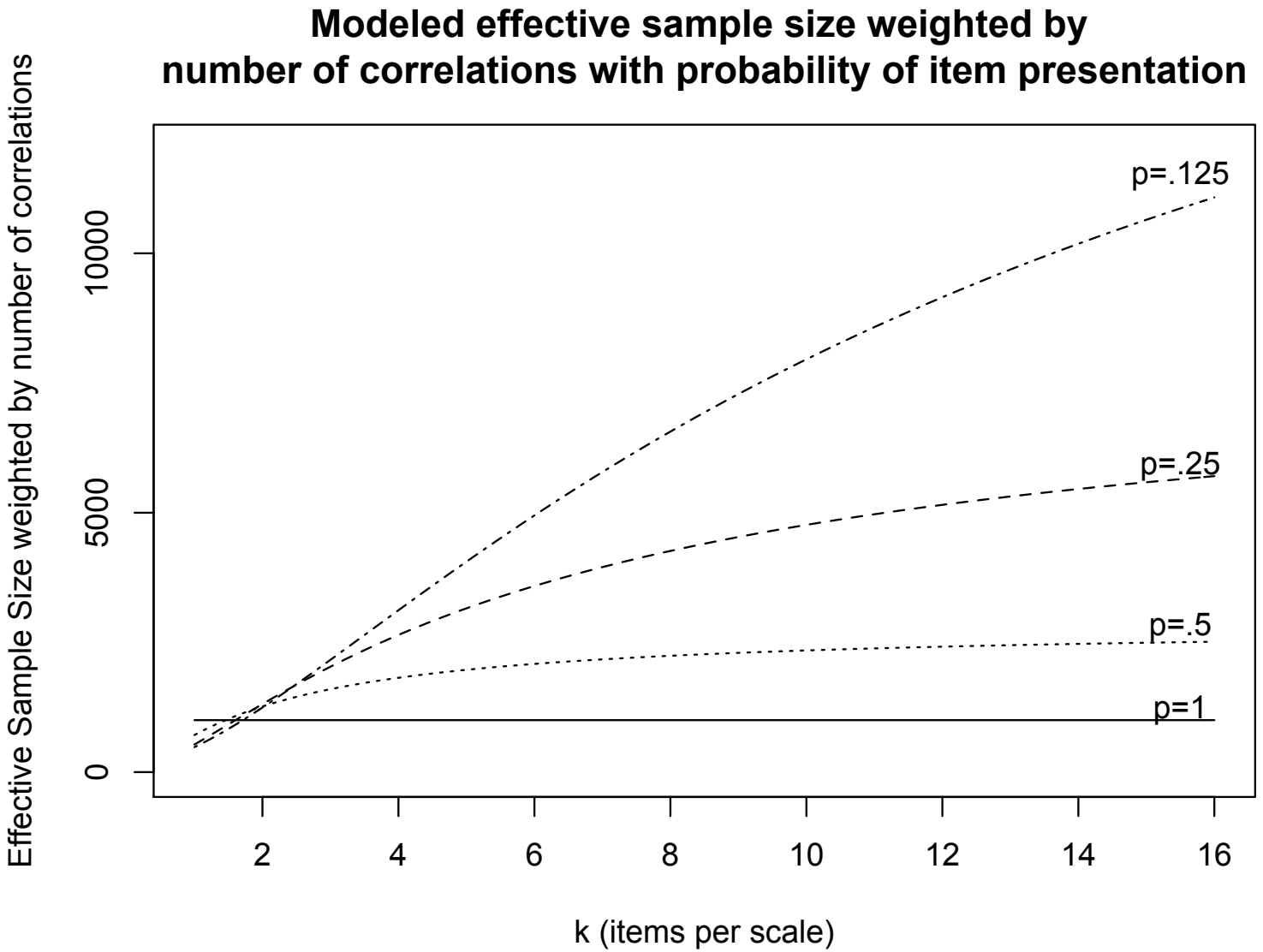
Comments on simulation values

1. These simulations are based upon N = 10,000
2. Although for k = 1, the effective sample size is, of course, just Np^2 and thus for $p = .25 = 10,000 * .25^2 = 625$ this provides a relatively small standard error ($\sigma_r = .04$).
3. Had we not sampled, we would have a standard error of .01 but for 1/4 the number of items and thus 1/16 the number of correlations.
4. Is this extra precision worth the reduction in bandwidth?
5. More importantly, the standard error of 4 items scales with an even more dramatic sampling (p = .125) would also be roughly .04 but with 8 times as many items and thus 64 times as many correlations.

Correlation size weighted sample size

1. If we weight the effective sample size by the number of correlations found we come across an interesting observation.
2. Giving more items and then random sampling, and then forming composite scales leads to a weighted effective sample size that exceeds the actual sample size!
3. This observation is supported by bootstrapped resampling of our SAPA data sets.
4. In this next figure we consider what would happen if we applied SAPA procedures to 1000 participants.

Effective Sample Size weighted by the number of correlations

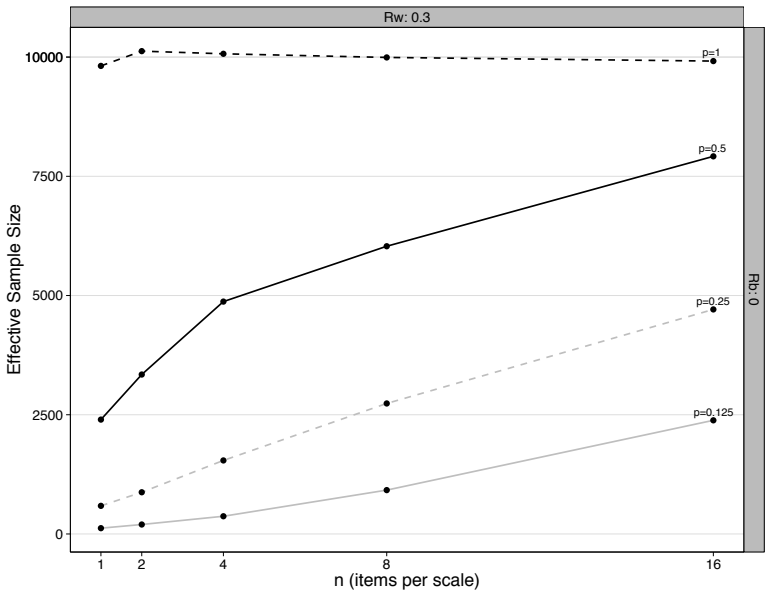


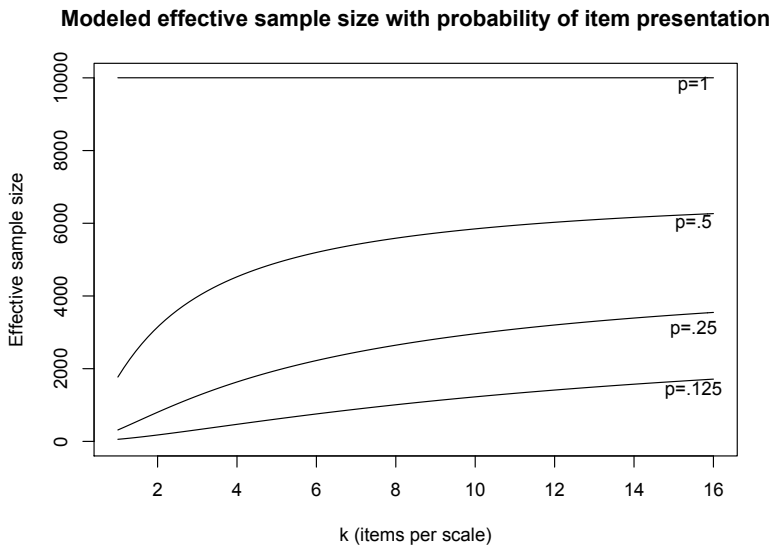
Comments upon the modeled values

1. We are modeling the simulated values (Brown, 2014) with an estimate based upon sample size and number of items given.
2. Without sampling, the effective sample size of a composite of any length is the same as that given by the traditional formula for standard errors.
3. But, with sampling the errors of any pair of correlations are *mostly independent* of the errors of any other pair.
4. But the higher the probability of any item being given, the less the independence on pairs of pairs of items.
5. The lower the probability of any item being given, the more the pairs are independent, but the lower the likelihood of the pair being given.
6. Combining these into a relatively simple formula leads to the following figure which matches pretty well (but not perfectly) the estimates from Brown (2014)

Modeled and bootstrapped effective sample size.

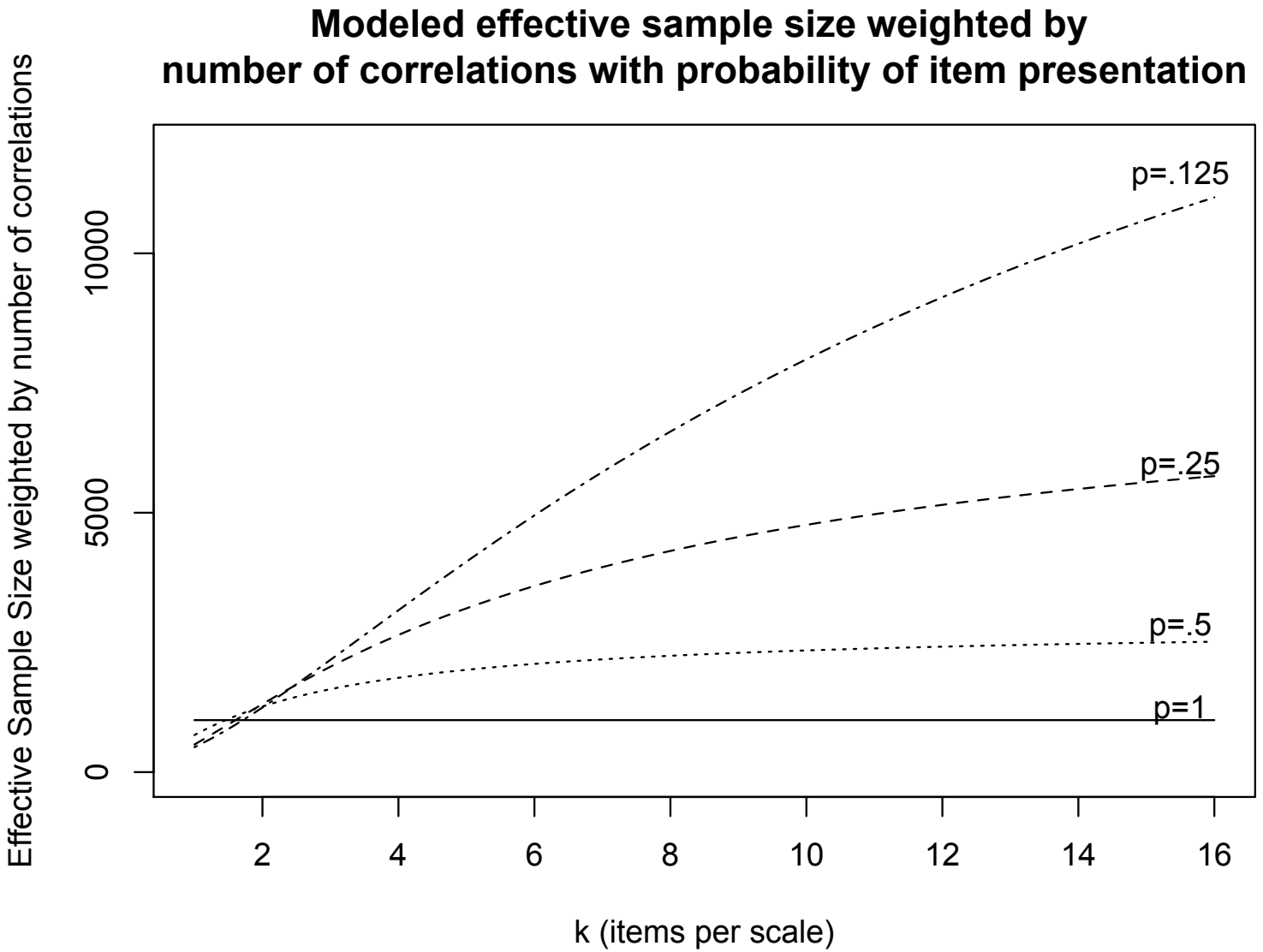
Comparing the bootstrapped values to modeled values.





Although we capture the effect for $k \leq 8$ we underestimate the effect for larger k.

Effective Sample Size weighted by the number of correlations



Suggestions for the future

1. Applying the SAPA or MMCAR technique is practical for any data set with more than about 500 participants.
2. Clearly, the fewer participants involved, the less aggressively one should sample.
3. But with $N > 1,000$, p can be as low as .25 and still can relatively stable pairwise estimates and very stable estimates for scales of 4 or more items.
4. These estimates of effective sample size are consistent with various bootstrapped resampling estimates from our real data.
5. Giving more items/construct provides a less biased estimate of between construct correlations.
6. The further power of sampling is that estimates of means are found for Np participants and also benefit from sampling more items.

Trading items for people: Studies, Items, People, Items x People

Table: Data sets vary in their sampling strategy but have similar total information

Study	N	Items	Items/ Person	Items* People	r's * $\sqrt{N}$
ES	1,000	3,000?	3,000	$3 * 10^6$	$2.8 * 10^8$
PG	10^7	44	44	$4.4 * 10^8$	$6 * 10^6$
SK	$4.5 * 10^6$	20-300	20-300	$1.7 * 10^8$	$1.9 * 10^7$
SAPA	$2 * 10^5$	2,000	100	$2 * 10^7$	$1.6 * 10^8$

Anderson, J., Lin, H., Treagust, D., Ross, S., & Yore, L. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591–614.

Brown, A. D. (2014). Simulating the MMCAR method: An examination of precision and bias in synthetic correlations when data are 'massively missing completely at random'. Master's thesis, Northwestern University, Evanston, Illinois.

Condon, D. M. (2014). *An organizational framework for the psychological individual differences: Integrating the affective, cognitive, and conative domains*. PhD thesis, Northwestern University.

Condon, D. M. & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.

NORTHWESTERN UNIVERSITY

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe*, volume 7 (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R. & Saucier, G. (2016). The Eugene-Springfield Community Sample: Information Available from the Research Participants. Technical Report 56-1, Oregon Research Institute, Eugene, Oregon.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institue of Personality and Social Research.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection

using planned missing designs. In N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *SAGE Handbook of Online Research Methods* chapter 37: Mobile Methods. Sage Publications, Inc. (in press).

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control* chapter 2, (pp. 27–49). New York, N.Y.: Springer.