

Using SAPA to study the structure of 6600 personality and ability items

Or: Making SAPA techniques work for you

Part of an invited symposium (chaired by Matthias Ziegler):

Measuring personality: What's new and how does it help
personality psychology

European Conference on Personality, Zadar, July, 2018

William Revelle^a & David M. Condon^b

^aDepartment of Psychology, Northwestern University, Evanston, Illinois

^bDepartment 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>



Outline

Introduction

Measuring individual differences: the tradeoff between breadth versus depth

SAPA techniques for everyone: It does not require large samples

SAPA theory

Synthetic Aperture Astronomy as an analogy

Sample items as well as people

Simulating SAPA with real data

The design of the simulation

Simulation Results

Predictive validity

Conclusions

Appendices: R code

The simulation function

Running the code on the data

Overview

1. Using big data techniques has changed the way we study personality. The increase in power due to very large samples allows the detection of small but meaningful effects – structural measures can have finer resolution than previously available and cross-validated predictive accuracy can be substantially enhanced.
2. But these techniques are not limited to big data. We discuss one such approach that can be used for very large or even moderate size samples.
3. We show how SAPA procedures improve scale reliability and validity over using short scales when giving random subsets of items selected from larger scales.

SAPA overview

1. Synthetic Aperture Personality Assessment (SAPA) uses the technique of Massively Missing Completely at Random (MMCAR) data presentation to assess 20K participants per month. Each participant is given a random subset of items chosen from an item pool of more than 6600 items. These items, extended from the [International Personality Item Pool](#) (Goldberg, 1999) and the [International Cognitive Ability Resource](#) (Condon & Revelle, 2014), assess temperament, cognitive ability, interests and attitudes as well as self reported behaviors.
2. Conventional psychometric techniques are used to identify homogeneous scales; empirical item selection procedures are used to develop optimal item composites to predict a wide range of criteria. Data analysis code is done using the *psych* package (Revelle, 2018) in R (R Core Team, 2018).

The basic problem: Fidelity versus bandwidth

- 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.
- This has led to the quandary of should you give many people a few items or a few people, many items?
- Our answer is to do both, but with a *Massively Missing Completely At Random* (MMCAR) data structure.
- We refer to this technique as *Synthetic Aperture Personality Assessment* (SAPA) to recognize the analogy to synthetic aperture radio astronomy (Revelle, Wilt & Rosenthal, 2010; Revelle, Condon, Wilt, French, Brown & Elleman, 2016)
- This is functionally what Frederic Lord (1955, 1977) suggested 63 years ago. It is time to take him seriously.

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 (Costa & McCrae, 1992) with 300, the IPIP (Goldberg, 1999) Big 5 with 100, the BFI (John, Donahue & Kentle, 1991) with 44 items, the BFI2 (Soto & John, 2017) with 60, the 30 item 'Short Five' (Konstabel, Lönnqvist, Leikas, Velázquez, H, Verkasalo, & et al., 2017), the TIPI (Gosling, Rentfrow & Swann, 2003) with 10 and the 10 item BFI (Rammstedt & John, 2007)) to include as part of other surveys.
6. Unfortunately, with this reduction of items, breadth of substantive content is lost. We offer an alternative procedure.

Example studies with subject/item tradeoffs

1. The Potter-Gosling internet project (outofservice.com) has given over 10,000,000 tests since 1997. Originally the 44 items of the Big Five Inventory (BFI) (John et al., 1991) although they are now giving the BFI2 (Soto & John, 2017).
2. 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 (Kosinski, Matz, Gosling, Popov & Stillwell, 2015; Youyou, Kosinski & Stillwell, 2015).
3. Johnson reports [two data sets](#): 300 IPIP-NEO items for 145,388 participants and 120 IPIP-NEO items for 410,376 participants (Johnson, 2014).
4. Smaller scale studies include the [BBC data set](#) of the 44 items of the BFI on 386,375 and the initial report on the BFI-2 (Soto & John, 2017) with several thousand subjects with 60 items.

Exceptions to the shorter and shorter inventory trend

1. Lew Goldberg and his colleagues at the University of Oregon developed the Eugene-Springfield sample (Goldberg & Saucier, 2016) which has given several thousand items to $\approx 1,000$ predominantly white middle class 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).
2. In fact, many of the subsequent attempts at personality scale development have used the Eugene-Springfield sample, e.g., the BFI (John et al., 1991), and the Big Five Aspect Scales (BFAS) of DeYoung, Quilty & Peterson (2007).

The Eugene Springfield sample and the International Personality Item Pool

1. Unfortunately, many of the items that have come out of the E-S sample were prematurely selected to represent the Big 5. That is, even though meant to capture the many dimensions of the lexicon, the adjectival descriptors used had been trimmed to those matching the 5 factors that have been known since the 1950's (Kelly & Fiske, 1950, 1951; Tupes & Christal, 1961; Norman, 1963).
2. Because of the ease of use and the openness of the IPIP, most of the short forms followed the Big Five structure that came out of the E-S sample.

SAPA techniques can work for you

1. At the [Personality Project](https://sapa-project.org) (Revelle et al., 2010, 2016) (now at sapa-project.org) we have taken the opposite direction and have given more and more items including measures of temperament, ability, and interests and we are now developing item statistics on more than 6,600 items (Condon & Revelle, 2017) for almost 500,000 participants (by using SAPA procedures).
2. We have reported computer simulations of our procedures but now we want to demonstrate with real data the amazing power of massively missing data.
3. In particular, we want to show that the techniques can work on relatively small samples ($\approx 1000 - 2000$) as well as the larger ones we have been working with.

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

Table: Data sets vary in their sampling strategy and the Potter-Gosling and Stillwell/Kosinski data sets seem to have more data than the others

Study	N	Items (n)	Items/ Person	Items* People
Potter-Gosling	10^7	44	44	$4.4 * 10^8$
Stillwell-Kosinski	$4.5 * 10^6$	20-300	20-300	$1.7 * 10^8$
Johnson	$4.1 * 10^5$	120	120	$4.9 * 10^7$
Johnson	$1.4 * 10^5$	300	300	$4.3 * 10^7$
SAPA* (2010-2017)	$2.5 * 10^5$	2000	100-150	$2.5 * 10^7$
SAPA+ (2017-2018)	$2 * 10^5$	6,600	100-150	$2 * 10^7$
SAPA- (pre 2010)	$6.3 * 10^4$	500	100-150	$9.4 * 10^6$
Eugene-Springfield	10^3	3,000	3,000	$3 * 10^6$

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?

Consider the amount of *information* which varies by number of correlations $\frac{n*(n-1)}{2}$ and $1/(\text{standard error of the } r) \approx \sqrt{N}$.

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

Information varies by the number of correlations ($n * (n - 1) / 2$) weighted by their standard errors which vary by \sqrt{N}

Table: Data sets vary in their sampling strategy and the seemingly smaller sets, by giving many more items actually have more total information

Study	N	Items	Items/ Person	Items* People	Information
SAPA+	$2 * 10^5$	6,600	100-150	$2 * 10^7$	$2.2 * 10^8$
E-S	10^3	3,000?	3,000	$3 * 10^6$	$1.4 * 10^8$
S-Ki	$4.5 * 10^6$	20-300	20-300	$1.7 * 10^8$	$9.5 * 10^7$
SAPA*	$2.5 * 10^5$	1-2,000	100-150	$2.5 * 10^7$	$7.5 * 10^7$
Johnson	$1.4 * 10^5$	300	300	$4.3 * 10^7$	$1.7 * 10^7$
SAPA (pre 2010)	$4.3 * 10^3$	500	100-150	$2.5 * 10^7$	$9.3 * 10^6$
Johnson	$4.1 * 10^5$	120	120	$4.9 * 10^7$	$4.6 * 10^6$
P-G	10^7	44	44	$4.4 * 10^8$	$3.0 * 10^6$

SAPA pairwise:

+SAPA 2017-2018 (100) * SAPA 2013-2017 (1,400) -SAPA 2010-2013 (1100)

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})

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

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)

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

4. Breadth of constructs (band width) measured goes up by number of items (n).
5. Thus, we need to increase N as well as n. But how?

A short diversion: the history of optical telescopes

Resolution varies by aperture diameter (bigger is better)

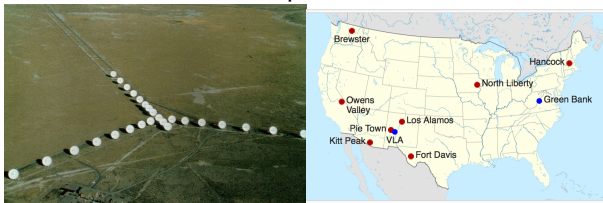


A short diversion: history of radio telescopes

Resolution varies by aperture diameter (bigger is still better)



Aperture can be synthetically increased across multiple telescopes
or even multiple observatories



Can we increase N (subjects) and n (items) 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. Discussed at IMPS (2017) meeting by Rutkowski and Matta in the missing data symposium.
7. Can we use this procedure for the study of individual differences without being a large company?
8. Yes, apply the techniques of radio astronomy to combine measures synthetically and take advantage of the web.

Subjects are expensive, so are items

1. In a survey such as Amazon's Mechanical Turk (MTURK), we would need to pay by the person and by the item.
2. Volunteer subjects are not very willing to answer many items.
3. Why give each person the same items? Sample items, as we sample people.
4. 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) (we sometimes have 99% missing data although our median is only 93% missing!)
5. This is the essence of Synthetic Aperture Personality Assessment (SAPA) (Condon & Revelle, 2014; Condon, 2014; Revelle et al., 2016, 2010).
6. This is a much higher rate of missingness than discussed in the balanced incomplete block design of NAEPS or PISA.

3 Methods of collecting 256 subject * items data

a) 8 x 32 complete

b) 32 x 8 complete

c) 32 x 32 MCAR $p=.25$

46213634521143453443645331212414
21243623166421516154432261516513
51661351155165463622224435623344
11141343362332215612152135614522
25353121264561433433232246526411
61335154566424114612641225353516
24634342151536242425413513435116
11554654453123111162423325516334

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
21461416
33232365

..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.....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.....
.....63.1.....6...5..4..2...5
..2.4..5.....52.4....44...
2.55.....2.....6.....6.....55..
..5.....4...6341.4..2...18/44
...55.....5.....45.....3..32.

Synthetic Aperture Personality Assessment

1. Give each participant a random sample of pn items taken from a larger pool of n items. p_i might be anywhere from .01 to 1.
2. Find covariances based upon “pairwise complete data”. Each pair appears with probability $p_i p_j$ with a median of .01.
3. Find scales based upon basic covariance algebra.
 - Let the raw data be the matrix ${}_N\mathbf{X}_n$ with N observations converted to deviation scores.
 - Then the item variance covariance matrix is ${}_n\mathbf{C}_n = \mathbf{X}'\mathbf{X}N^{-1}$
 - and scale scores, ${}_N\mathbf{S}_s$ are found by $\mathbf{S} = {}_N\mathbf{X}_p\mathbf{K}_s$.
 - ${}_n\mathbf{K}_s$ is a keying matrix, with $k_{ij} = 1$ 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,

$${}_s\mathbf{C}_s = {}_s\mathbf{S}'{}_N{}_N\mathbf{S}_sN^{-1} =$$

$${}_s\mathbf{C}_s = (\mathbf{X}\mathbf{K})'(\mathbf{X}\mathbf{K})N^{-1} = \mathbf{K}'\mathbf{X}'\mathbf{X}\mathbf{K}N^{-1} = \mathbf{K}'{}_n\mathbf{C}_n\mathbf{K}. \quad (1)$$
4. That is, we can find the correlations/covariances between scales from the item covariances, not the raw items.

Total information

1. The information in a correlation varies by its standard error

$$\sigma_r = \sqrt{\frac{1-r^2}{N-2}}$$

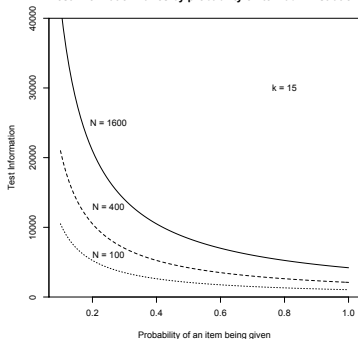
2. In SAPA, k items/person are randomly selected with probability p from a larger number, n ($k = pn$).
3. Thus, the number of subjects per item is pN .
4. The total number of correlations is just $\frac{n*(n-1)}{2}$ and the number of subjects per correlation is p^2N .
5. Total information is number of correlations * $\sqrt{p^2N} = \frac{n*(n-1)}{2} \sqrt{p^2N} = \frac{(k/p)((k/p)-1)}{2} * \sqrt{p^2N} = \frac{k*(k-1)\sqrt{N}}{2*p}$.
6. For the “normal case” where $p = 1$, the information is just what we expect—a quadratic function of k: $I_{kN} = \frac{k*(k-1)\sqrt{N}}{2}$.
7. But the more interesting case (the SAPA case) is for $p < 1$ the information is a hyperbolic function of p:
 $I_{pkN} = \frac{k*(k-1)\sqrt{N}}{2*p}$ but a linear function of the total number of items given ($n = k/p$) $I_{pkN} = \frac{n*(k-1)}{2} * \sqrt{N}$

Total information varies by the number of items (n) and the probability of sampling (p) and total sample size (N)

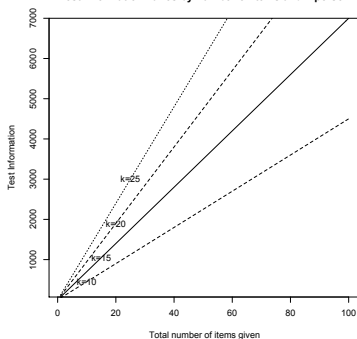
For k items/subject and N subjects, if every item is given with probability p, the information in the test is

$$I_{pkN} = \frac{k*(k-1)\sqrt{N}}{2*p} = \frac{n*(k-1)}{2} * \sqrt{N}$$

Test information varies by probability of item administration



Test information varies by number of items and k/person



Simulations with real data

1. In the past we have shown the power of SAPA using computer simulations of artificial data (Revelle & Condon, 2017, 2016).
2. Now we will show the validity of sampling models based upon samples from a complete data set of 70 Big 5 items.
3. The most recent version of SAPA includes 221K subjects on 6624 items.
4. Of these, there are 38,809 participants with complete data on the SAPA Personality Inventory (SPI) (Condon, 2017), a 135 item inventory with 70 items devoted to the “Big 5” supplemented with 65 additional items to complete an assessment of 27 lower level “facets”.
5. We will examine the 44,341 who have complete data on the 70 Big 5 items in the SPI.

Sampling from complete data

1. We compare three approaches:
 - 1.1 Complete data for 15 items/subject for 20 sets of 2,000 subjects (3 items/factor) (Modeling the naive but common approach of short scales).
 - 1.2 Incomplete data for 15 items/subject, but sampled from 70 items (79% missing or $p = .21$) for 20 sets of 2,000 subjects on 5 factors. This is the approach done in SAPA.
 - 1.3 We also compare 70 items with complete data in sets of 2,000 subjects.
2. For each set, we find 5 factors based upon the covariances.
3. We examine the values and standard errors of the 10 correlations between dimensions.
4. We examine the factor loadings and their standard errors.
5. We compare these to the “population value” based upon the remaining 4,341 subjects with complete data as well as 20 replication samples of 2,000 observations on all 70 items.

The design of the demonstration

1.

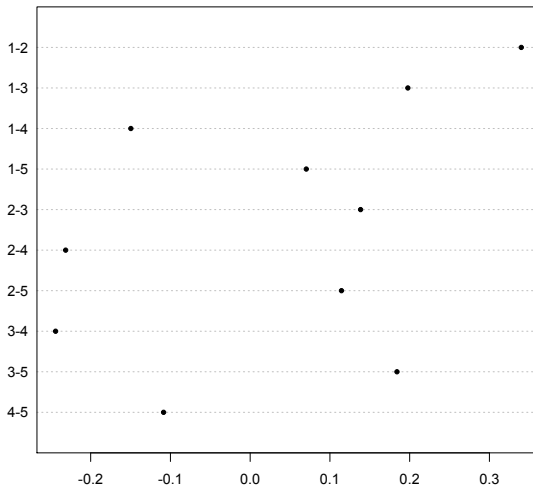
The top 5 items for each SPI factor - total complete data sample

Table: SPI factors (cut = .3)

SPI 5 factors from N=44,341					
Variable	MR2	MR5	MR3	MR1	MR4
Am concerned about others.	0.65				
Sympathize with others feelings.	0.65				
Believe that others have good intentions.	0.63				
Am sensitive to the needs of others.	0.59				
Think of others first.	0.59				
Keep things tidy.		0.66			
Neglect my duties.		-0.61			
Leave a mess in my room.		-0.60			
Start tasks right away.		0.60			
Find it difficult to get down to work.		-0.58			
Dislike being the center of attention.			-0.74		
Like to attract attention.			0.74		
Hate being the center of attention.			-0.73		
Make myself the center of attention.			0.73		
Usually like to spend my free time with people.	0.30		0.58		
Worry about things.				0.74	
Panic easily.				0.72	
Get overwhelmed by emotions.				0.71	
Am a worrier.				0.68	
Would call myself a nervous person.				0.67	
Am full of ideas.					0.70
Am able to come up with new and different ideas.					0.69
Am an original thinker.					0.64
Love to think up new ways of doing things.					0.58
Have a vivid imagination.					0.52

Factor intercorrelations for hold out sample of complete data

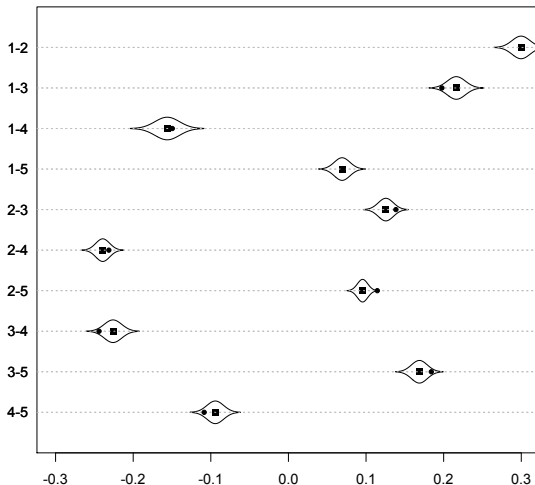
Factor correlations



1. Factor intercorrelations based upon 5 factors with complete data on 70 items.
2. Values based upon hold out sample of 4,341

Factor intercorrelations for 20 replications of complete data

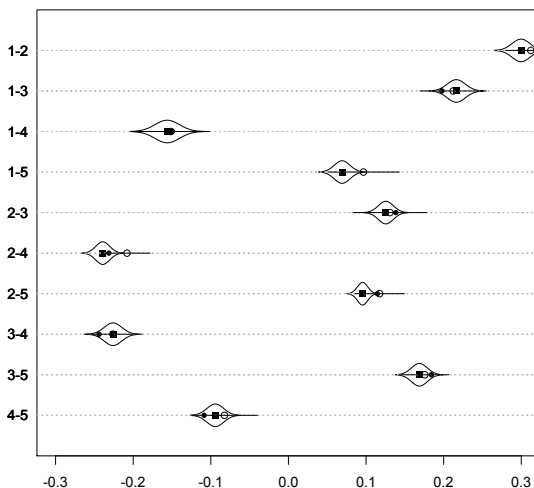
Scale correlations ± 1 sd



1. Factor intercorrelations based upon 5 factors with complete data on 70 items.
2. Mean values for 20 replications of 2,000 subjects each.
3. “Cats Eye” plots show ± 1 standard deviation.
4. Dot shows the hold out values

Comparing full information with SAPA based factor correlations

Scale correlations ± 1 sd



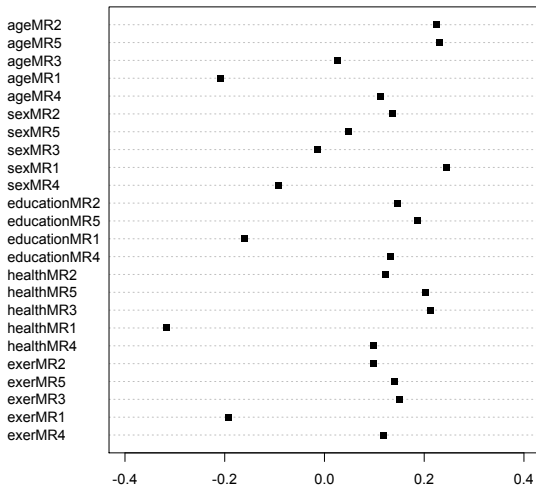
1. Factor intercorrelations based upon 5 factors with sapa data on 15 items/person for 70 total items.
2. Mean values for 20 replications of 2,000 subjects each.
3. Confidence intervals show ± 1 standard deviation.

Personality is about prediction

1. The fact that we can find the personality structure of our 70 items by using SAPA techniques is nice, but so what?
2. Can we use the factors to predict anything?
3. As part of the SAPA data collection, we ask about a number of demographic variables as well as various behaviors.
4. Can we predict those criteria using our factor scores?
5. Once again, we compare the hold out sample of 4,300 people to 20 random samples of 2,000 people with full information on all 70 items, to SAPA procedures with 15 items/person to “short form” procedures with 15 items given to everyone.
6. We note that these are demonstrations of short form versus SAPA procedures on sample sizes that are attainable by many of us.

Predictive validity

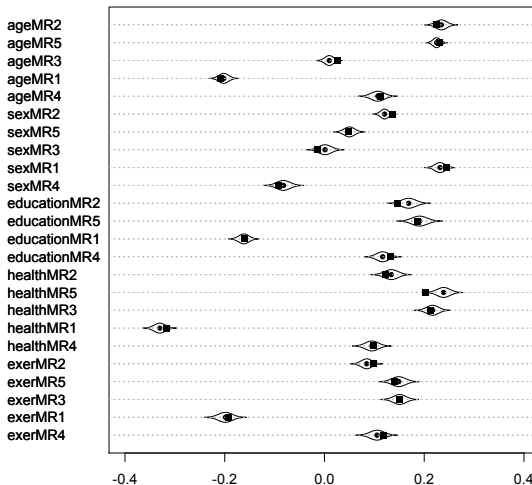
Predictive validity of factor scales



1. Factor validities based upon 5 factors with complete data on 70 items.
2. Hold out sample
3. Criteria predicted are age, sex, education, health, and exercise.

Comparing full information factor criterion validity with SAPA validity

Hold out sample and Full scale Validity coefficients



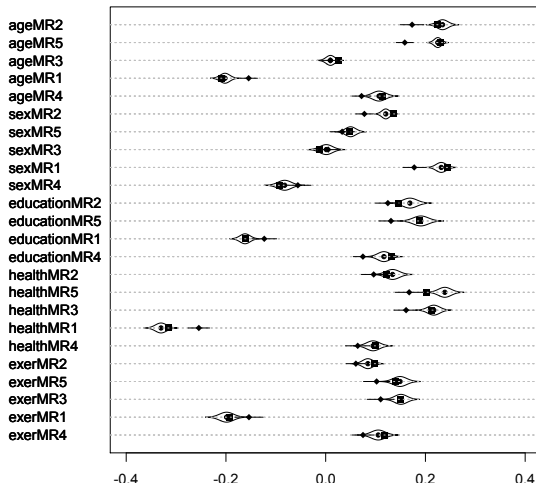
Factor validities based upon 5 factors with complete data on 70 items.

Criteria predicted are age, sex, education, health, and exercise.

“Cats Eye” plots show ± 1 standard deviation.

Comparing full information factor criterion validity with SAPA validity

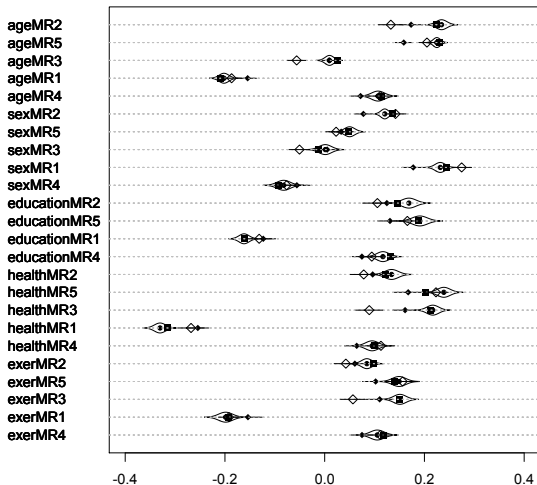
Hold out sample and Full scale Validity coefficients



Factor validities based upon 5 factors with sapa data on 15 items/person for 70 total items. Criteria predicted are age, sex, education, health, and exercise. Confidence bars are ± 1 sd.

Comparing full sample with SAPA and short form validity

Hold out, Full scale, SAPA, short scale validity coefficients



Factor validities based upon 5 factors with sapa data on 15 items/person for 70 total items. Criteria predicted are age, sex, education, health, and exercise. Confidence bars are ± 1 sd.

Some comments

1. These simulations were done with 2,000 simulated (but real) subjects
2. With $p=.21$, the number of pairwise correlations was $p^2N = .04 * 2000 \approx 80/pair$ and yet the results are very stable!
3. We did it again with just 1,000 simulated subjects and results are similar.
4. With such missingness, the correlation matrices are improper, but the `fa` function will give a minres solution anyway.

SAPA or MMCAR procedures are very powerful

1. At the SAPA-project we estimate difficulty parameters and covariance structures for 1,000s of items even though only 100-150 items are answered per subject.
2. More importantly, the same procedures can be used for people with smaller sample sizes with fewer items (e.g. MTurk research).
3. Structure of ability measures using the open source ability test from the International Cognitive Ability Resource (ICAR)
<http://icar-project.com>
4. Data sharing: <https://dataverse.harvard.edu/dataverse/SAPA-ProjectCode/manuscript/>
5. SPI development (Condon, 2017): <https://sapa-project.org/research/SPI/SPIdevelopment.pdf>
6. SPI scales, norms, IRT parameters:
<https://sapa-project.org/research/SPI>
7. Today's slides at <http://personality-project.org/sapa>
8. Join the ICAR and SAPA projects.

R code

The next few slides show the R code used for these analyses First we show the function that forms the various scales We then show the commands to the graphics.

R code

```
###
###The function to do the simulations

sim.complete <- function(x,nvar=NULL,n.trials=20,keys = NULL, n.items=15,n.obs=2000,
nf=5,criteria=NULL, short=NULL,seed=NULL,sapa=TRUE) {
  set.seed(seed) #allows us to either randomly choose seed or not

  total.n.obs <-nrow(x)
  if(is.null(nvar)) nvar <- ncol(x) - length(criteria)
  if(is.null(short)) short <- n.items/nf
  ids <- 1:total.n.obs
  Phi <- list()
  Phi.f <- list()
  orig.Phi <- list()
  fa.list <- list()
  s.keys <- list()
  facong <- list()
  long.loadings <- sapa.loadings <- list()
  short.scales <- list()
  overlap <- sapa.long <- short.long <- list()
  long.cors <- list()
  sapa.validity <- short.validity <- long.validity <- list()
  validity <- NA

  #continues
```

R code

The next few slides show the R code used for these analyses First we show the function that forms the various scales We then show the commands to the graphics.

R code

```
###
###The function to do the simulations

sim.complete <- function(x,nvar=NULL,n.trials=20,keys = NULL, n.items=15,n.obs=2000,
nf=5,criteria=NULL, short=NULL,seed=NULL,sapa=TRUE) {
  set.seed(seed) #allows us to either randomly choose seed or not

  total.n.obs <-nrow(x)
  if(is.null(nvar)) nvar <- ncol(x) - length(criteria)
  if(is.null(short)) short <- n.items/nf
  ids <- 1:total.n.obs
  Phi <- list()
  Phi.f <- list()
  orig.Phi <- list()
  fa.list <- list()
  s.keys <- list()
  facong <- list()
  long.loadings <- sapa.loadings <- list()
  short.scales <- list()
  overlap <- sapa.long <- short.long <- list()
  long.cors <- list()
  sapa.validity <- short.validity <- long.validity <- list()
  validity <- NA

  #continues
```

○○○○
○○○○○○○
○○○○○○○○○
○○○○○●○○○
○○○

R code

R code

```

# we can either figure out the keys, or use them from input
if(is.null(keys)) {
  #First, define the target based upon all the cases
  #do it from the hold out sample

  if(is.null(n.obs)) {n.obs <- trunc(total.n.obs/n.trials)}
  ss <- sample(ids,replace=FALSE) #a random permutation of the entire data set
  x <- x[ss,] #this is a random organization of the data
  validity <- x[c((n.trials*n.obs +1):total.n.obs),1:nvar] #this is the hold out sample

  if(is.na(validity)) {validity <- x} #this is the complete data set
  f.all <- fa(validity[1:nvar],nfactors=nf)
  f.all <- fa.organize(f.all) #organize the factors in quasi-echelon form
  f.all.sorted <- fa.sort(f.all)
  keys <- factor2cluster(f.all.sorted)
  #keys <- fa.organize(keys) #organize the keys into quasi-echolon form
  key.list <- keys2list(keys)
  select <- NULL
  for (k in 1:nf) {select <- c(select,selectFromKeys(key.list[[k]][1:short])}
    s.keys[[k]] <- key.list[[k]][1:short]}
    names(s.keys) <- names(key.list)
  }
}

#short keys and select are the keys for the top 3 items per scale
combined.keys <- c(key.list,s.keys)

```

R code

R code

```

validity.scores <- scoreItems(keys,validity) #this is normally the hold out sample
if(!is.null(criteria)) { predictive.validity = cor(validity.scores$scores,x[c((n.trials*n.obs
validity.cors <- validity.scores$cor[lower.tri(validity.scores$cor)]
validity.f <- fa(validity,nf=nf)
validity.f.cors <- validity.f$Phi[lower.tri(validity.f$Phi)]

#now, do the replications of various scoring algorithms
k <- 1
for(trials in 1:n.trials) {
  data <- x[ss[k:(n.obs*trials)],] #this is a derivation sample
  #find the scores with all the data using all the items in the sample
  long.scores <- scoreFast(key.list,data)
  l.cor <-cor(long.scores,use="pairwise")
  long.cors[[trials]] <- l.cor[lower.tri(l.cor)]
  #Now, find the short scales
  short.scores <- scoreFast(s.keys,data[select])
  short.cor <- cor(short.scores,use="pairwise")
  if(!is.null(criteria)) {short.val <- cor(short.scores,data[criteria],use="pairwise")} else {
  #short.cor <- scoreItems(s.keys,data[select])$cor
  short.scales[[trials]] <- short.cor[lower.tri(short.cor)]
  sl <- cor(short.scores,long.scores,use="pairwise") #but these have an overlap problem

  short.long[[trials]] <- diag(sl)
  short.validity[[trials]] <- short.val

R <- cor(data[1:nvar],use="pairwise")
overlap[[trials]] <- diag( scoreOverlap(combined.keys,R)$cor[1:nf,(nf+1):(2*nf)]) #these are
f.long <- fa(R,nf)
long.l <- apply(f.long$loadings,1,function(x) which(abs(x) ==max(abs(x))))
long.loadings[[trials]] <- apply(f.long$loadings,1,function(x) x[which(abs(x) ==max(abs(x)))])

```

R code

```

fa.cong <- matrix(unlist(facong),ncol=nf,byrow=TRUE)
Phi.f.vec <- matrix(unlist(Phi.f),ncol = nf*(nf-1)/2,byrow=TRUE)
Phi.vec <- matrix(unlist(Phi),ncol = nf*(nf-1)/2,byrow=TRUE)
short.scales <- matrix(unlist(short.scales),ncol = nf*(nf-1)/2,byrow=TRUE)
short.long <- matrix(unlist(short.long),ncol=nf,byrow=TRUE)
sapa.long <- matrix(unlist(sapa.long),ncol=nf,byrow=TRUE)
long.cors <- matrix(unlist(long.cors),ncol = nf*(nf-1)/2,byrow=TRUE)
sapa.loadings <- matrix(unlist(sapa.loadings),ncol=nvar,byrow=TRUE)
long.loadings <- matrix(unlist(long.loadings),ncol=nvar,byrow=TRUE)
overlap <- matrix(unlist(overlap),ncol=nf,byrow=TRUE)
if(!is.null(criteria)) {sapa.validity <- matrix(unlist(sapa.validity),ncol=nf*length(criteria),byrow=TRUE)
                        long.validity <- matrix(unlist(long.validity),ncol=nf*length(criteria),byrow=TRUE)
                        short.validity <- matrix(unlist(short.validity),ncol=nf*length(criteria),byrow=TRUE)
                        }

k <- 1
cnN <- 1:nf
vecnames <- 1:(nf*(nf-1)/2)
  for(i in 1:(nf-1)) {for (j in (i+1):nf) {
    vecnames[k] <- paste(cnN[i],cnN[j],sep="-")
    k<- k + 1 }}

colnames(Phi.f.vec) <- vecnames
colnames(Phi.f.vec) <- vecnames
colnames(short.scales ) <- vecnames
colnames(long.cors) <- vecnames
colnames(Phi.vec) <- vecnames
colnames(short.long) <- colnames(sapa.long) <- colnames(overlap)<- paste0("S",1:nf)

colnames(sapa.validity) <- colnames(long.validity) <- colnames(short.validity) <- paste0(rep(1:nf),1:nf)

result <- list(validity.f.cors =validity.f.cors,validity.scaled.cors=validity.cors ,long.cors=long.cors,
sapa.long=sapa.long, predictive.validity=predictive.validity,sapa.validity=sapa.validity,long.validity=long.validity)
return(result)
}

```


R code

R code

```
#match factors based upon congruences

factor.match <- function(f1,f2) {
  cong <- fa.congruence(f1,f2)
  match <- apply(cong,1,function(x) which.max(abs(x)))
  f2 <- fa.organize(f2,o=match)
  cong <- fa.congruence(f1,f2)
  flip <- diag(sign(diag(cong)))
  f2$loadings <- f2$loadings %**% flip
  phi <- flip %**% f2$Phi %**% flip
  return(list(f=f2,phi=phi))
}
```

More R

R code

```
#Code for ECP.18
#first get the data ready

read.file() #get the latest data set
sapa <- SAPAdata07feb2017thru05jul2018
spi.5.items <- selectFromKeys(keys.list[391:395])
spi.demo.items <- c(spi.5.items, cs(age,sex,health,education,height,weight,BMI,exer,smoke))
spi.demo <-sapa[spi.demo.items]

cc <- apply(spi.demo[1:70],1,function(x) sum(!is.na(x)))
spi5.com <- spi.demo[cc==70,]

spi5.com <- char2numeric(spi5.com) #the criterion variables are categorical
spi.com <-scrub(spi5.com,"sex",max=2) #make sex binary

criteria <- cs(age,sex, education, health, exer )
sim5 <- sim.complete(spi5.com,nvar=70,criteria=criteria)

error.dots(sim5$validity.scaled.cors,add=FALSE,pch=20,sort=FALSE,main="Factor correlations",)
error.dots(sim5$long.cor,sort=FALSE,pch=15,main="Scale correlations +/- 1 sd",eyes=TRUE,sd=TRUE)

names(sim5$validity.scaled.cors) <- colnames(sim5$sapa.cor)
error.dots(sim5$validity.scaled.cors,add=TRUE,pch=20,sort=FALSE,main="",xlim=c(-.3,.3))
error.dots(sim5$sapa.cor,sort=FALSE,add=TRUE,main="",sd=TRUE,xlim=c(-.3,.3))

error.dots(abs(sim5$long.cor),sort=FALSE,pch=15,main="Scale correlations +/- 1 sd",eyes=TRUE)
error.dots(abs(sim5$sapa.cor),sort=FALSE,add=TRUE,main="")
```

More R

R code

```
#now for the validities
```

```
error.dots(sim5$long.validity, sort=FALSE, eyes=TRUE, pch=20, sd=TRUE, main="SAPA and Full scale
error.dots(sim5$sapa.validity, sort=FALSE, add=TRUE, pch=18, sd=TRUE, main="", xlim=c(-.4, .4))
error.dots(sim5$short.validity, sort=FALSE, add=TRUE, pch=5, sd=TRUE, main="", xlim=c(-.4, .4))
predictive <- as.vector(sim5$predictive.validity)
names(predictive) <- colnames(sim5$short.validity) #need these to keep the scaling correct
error.dots(predictive, add=FALSE, sort=FALSE, sd=TRUE, pch=15, xlim=c(-.4, .4), main="Predictive va

error.dots(sim5$long.validity, sort=FALSE, eyes=TRUE, pch=20, sd=TRUE, main="Hold out sample and
error.dots(predictive, add=TRUE, sort=FALSE, sd=TRUE, pch=15, xlim=c(-.4, .4), main="")
error.dots(sim5$sapa.validity, sort=FALSE, add=TRUE, pch=18, sd=TRUE, main="", xlim=c(-.4, .4))

error.dots(sim5$long.validity, sort=FALSE, eyes=TRUE, pch=20, sd=TRUE, main="Hold out sample,
error.dots(predictive, add=TRUE, sort=FALSE, sd=TRUE, pch=15, xlim=c(-.4, .4), main="")
error.dots(sim5$sapa.validity, sort=FALSE, add=TRUE, pch=18, sd=TRUE, main="", xlim=c(-.4, .4))

error.dots(sim5$long.validity, sort=FALSE, eyes=TRUE, pch=20, sd=TRUE, main="Hold out, Full
error.dots(predictive, add=TRUE, sort=FALSE, sd=TRUE, pch=15, xlim=c(-.4, .4), main="")
error.dots(sim5$sapa.validity, sort=FALSE, add=TRUE, pch=18, sd=TRUE, main="", xlim=c(-.4, .4))
error.dots(sim5$short.validity, sort=FALSE, add=TRUE, pch=5, sd=TRUE, main="", xlim=c(-.4, .4))

error.dots(sim5$long.cor, sort=FALSE, eyes=TRUE, sd=TRUE, pch=20, main="Full scale Interfactor c
error.dots(sim5$long.cor, sort=FALSE, eyes=TRUE, sd=TRUE, pch=20, main="SAPA and Full scale Inter
error.dots(sim5$sapa.cor, sort=FALSE, add=TRUE, pch=18, sd=TRUE, main="", xlim=c(-.4, .4))
```

R code

```
test.info <- function(n,k=15, N= 100) {  
  p <- n/k  
  info <- n * (k-1) * sqrt(N)/(2)  
  return(info)  
}  
  
test.info.p <- function(p,k=15,N=100) {n <- k/p  
  info <- n * (k-1) * sqrt(N)/(2)  
  return(info) irwise = 1600",xlim=c(0,1))
```

- 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.
- 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. (2017). The sapa personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. Technical report.
- Condon, D. M. & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence*, 43, 52–64.
- Condon, D. M. & Revelle, W. (2017). The times they are

a-changin' (in personality assessment). In *European Conference on Psychological Assessment*.

Costa, P. T. & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology*, 93(5), 880–896.

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.

- Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6), 504 – 528.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of Research in Personality*, 51, 78 – 89.
- Kelly, E. L. & Fiske, D. W. (1950). The prediction of success in the VA training program in clinical psychology. *American Psychologist*, 5(8), 395 – 406.
- Kelly, E. L. & Fiske, D. W. (1951). *The prediction of performance in clinical psychology*. Ann Arbor, Michigan: University of Michigan Press.
- Konstabel, K., Lönnqvist, J.-E., Leikas, S., Velàzquez, R. G., H, H. Q., Verkasalo, M., , & et al. (2017). Measuring single

constructs by single items: Constructing an even shorter version of the “short five” personality inventory. *PLoS ONE*, 12(8), e0182714.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6), 543.

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

Lord, F. M. (1977). Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction. *Applied Psychological Measurement*, 1(3), 447–455.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factors structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 66(6), 574–583.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of Research in Personality*, 41(1), 203 – 212.

Revelle, W. (2018). *psych: Procedures for Personality and Psychological Research*.
<https://cran.r-project.org/web/packages=psych>: Northwestern University, Evanston. R package version 1.8.4.

Revelle, W. & Condon, D. M. (2016). Embrace your missingness. In *Part of symposium: approaching complex research designs from the perspective of missing data.*, Chicago, IL. Association for Psychological Science.

Revelle, W. & Condon, D. M. (2017). Using MMCAR to explore

the structure of personality and ability. In *Annual Meeting of the International Psychometric Society*, Zurich.

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* (2nd ed.). chapter 37, (pp. 578–595). Sage Publications, Inc.

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.

Soto, C. J. & John, O. P. (2017). The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1), 117–143.

Tupes, E. C. & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. Technical Report 61-97, USAF ASD Technical Report, Lackland Air Force Base (reprinted in *Journal of Personality* (1992) 60: 225–251).

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences*, 112(4), 1036–1040.