Self-regulation ontology project: Sampling simulations

Russell Poldrack

November 26, 2015

1 Introduction

At the kickoff meeting for this project on Nov 23, 2015, there was discussion regarding the most optimal sampling schemes for the proposed study. The goal of this document is to lay out the work done to assess the questions raised in this discussion.

The original sampling plan was to attempt to enroll a number of MTurk workers to complete the entire battery of tasks (currently estimated at about 70 tasks). This has the benefit of providing a relatively large number of subjects with substantial coverage of the full battery across 10 sessions, but it also raises practical concerns. In particular, the rate of attrition to be expected from this procedure is currently unknown, and a high rate of non-completion of the full battery would raise difficult modeling issues, because the data would be missing not at random. Dave Mackinnon suggested that it might be more optimal to instead sample a larger number of indviduals on a smaller portion of the battery using random assignment of tests to subjects, so that missing data would be missing completely at random, which allows the use of a much wider and more convenient set of modeling tools. At the meeting it was decided that simulations were in order to assess this question.

2 Simulation methods

Simulations were based on real data obtained from the UCLA Consortium for Neuropsychiatric Phenomics (CNP). This dataset consists of data from 1254 subjects across a large number of measures. We selected a subset of these measures for further examination, including both performance and questionnaire measures relevant to the proposed work. Initial exploration led to identification of a set of 12 variables (listed in Table 1, which were relatively well clustered into four clusters (see Figure 1). The total sample size for these measures (including only subjects with complete data for all measures) was 1126.

Based on these clusters, a confirmatory factor analysis (CFA) model was created, with four latent factors, each of which was associated with three behavioral measures. In addition, a second model was created that includes only two factors and combines across measures that should clearly be separated; this is meant to serve as a comparison model,

Table 1: Description of variables used in CFA model.

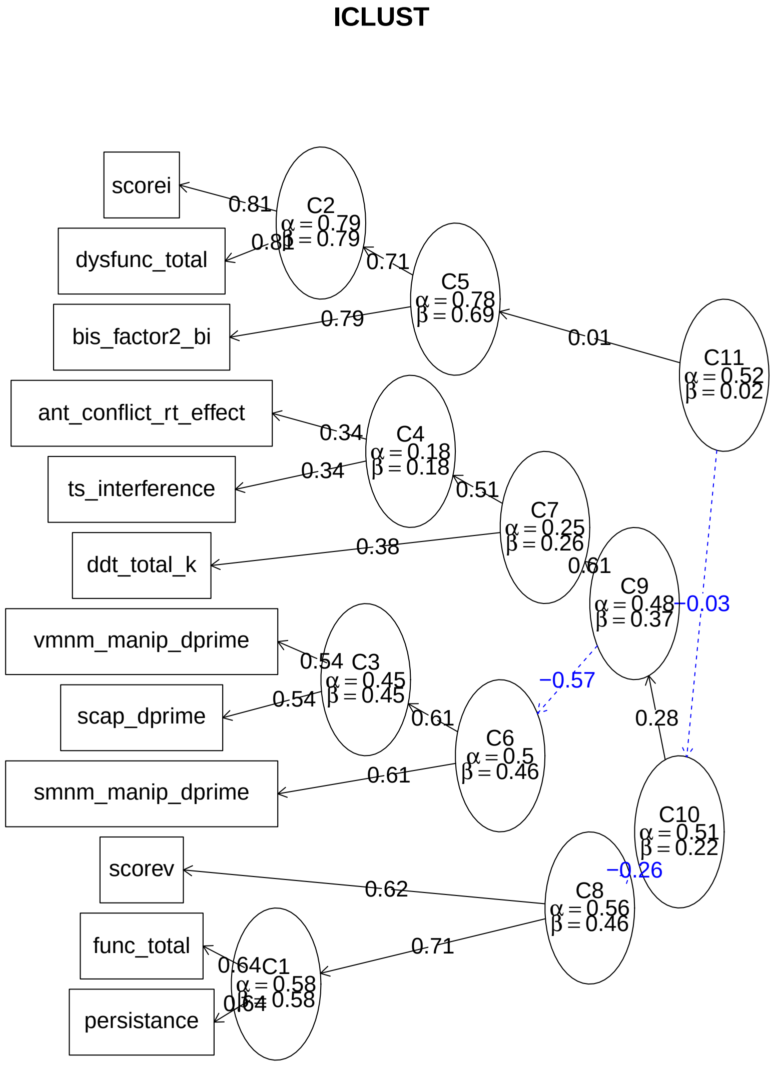
Factor 1	
scorei	Eyesenk Impulsiveness Inventory, Impulsivity Subscale
dysfunc_total	Dickman, Dysfunctional Impusivity Subscale
bis_factor2_bi	Barratt Impulsiveness Scale, Behavioral Impulsivity factor
Factor 2	
ant_conflict_rt_effect	Attention Networks Task, RT conflict effect
ts_interference	Task switching interference effect
ddt_total_k	Delay Discounting task, mean discount rate
Factor 3	
vmnm_manip_dprime	Verbal WM manipulation, d-prime
scap_dprime	Spatial working memory capacity
smnm_manip_dprime	Spatial WM manipulation, d-prime
Factor 4	
scorev	Eyesenk Impulsiveness Inventory, Venturesomeness Subscale
dysfunc_total	Dickman, Functional Impusivity Subscale
persistence	TCI Persistence subscale

in order to asses the ability to identify the correct model. These models are presented in Figure 2.

Simulated datasets were generated by sampling subjects with replacement from the original 1126. The base sample size for the simulations was 400 subjects, which was used for simulations with complete cases (no missing data). Datasets with missing data were generated by sampling larger datasets and then setting a particular proportion of variables to NA for each subject; the sample size was increased in order to keep the total number of complete cells constant across simulations.

The two CFA models were fit to each simulated dataset using the lavaan package in R. Two approaches to missing data were explored. In the first, the missing data were imputed using two methods: predictive mean imputation with the mice package in R as well, and imputation using Amelia II. Note that in each case only a single imputed dataset was processed. The models were then fit to each imputed dataset with no missing values. In the second, the model was fit using full information maximum likelihood (FIML) with no imputation (which is appropriate given that the data are missing completely at random). The primary measures of interest were root mean square approximation error (RMSEA) which is a measure of goodness of fit of the correct model, and the p-value for the model comparison between the correct and incorrect models, which was used to estimate statistical power to distinguish correct vs. incorrect models. For each model, 1000 simulations were run for four levels of sampling (100%, 50%, 25%, and 20% of the 12 tasks present for each

Figure 1: **Clustering of the 12 variables of interest.**

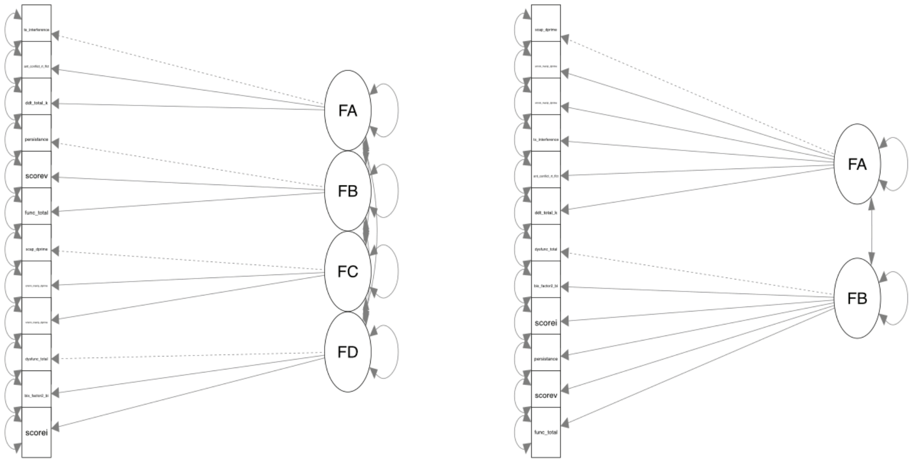


subject).

3 Results

Results of the simulations are presented in Figure 3. An important finding was that model fitting failed regularly, with increasing rates of failure as the amount of missing data increased. Model fitting was substantially more successful when using FIML with missing

Figure 2: **Correct (left) and incorrect (right) CFA models.**

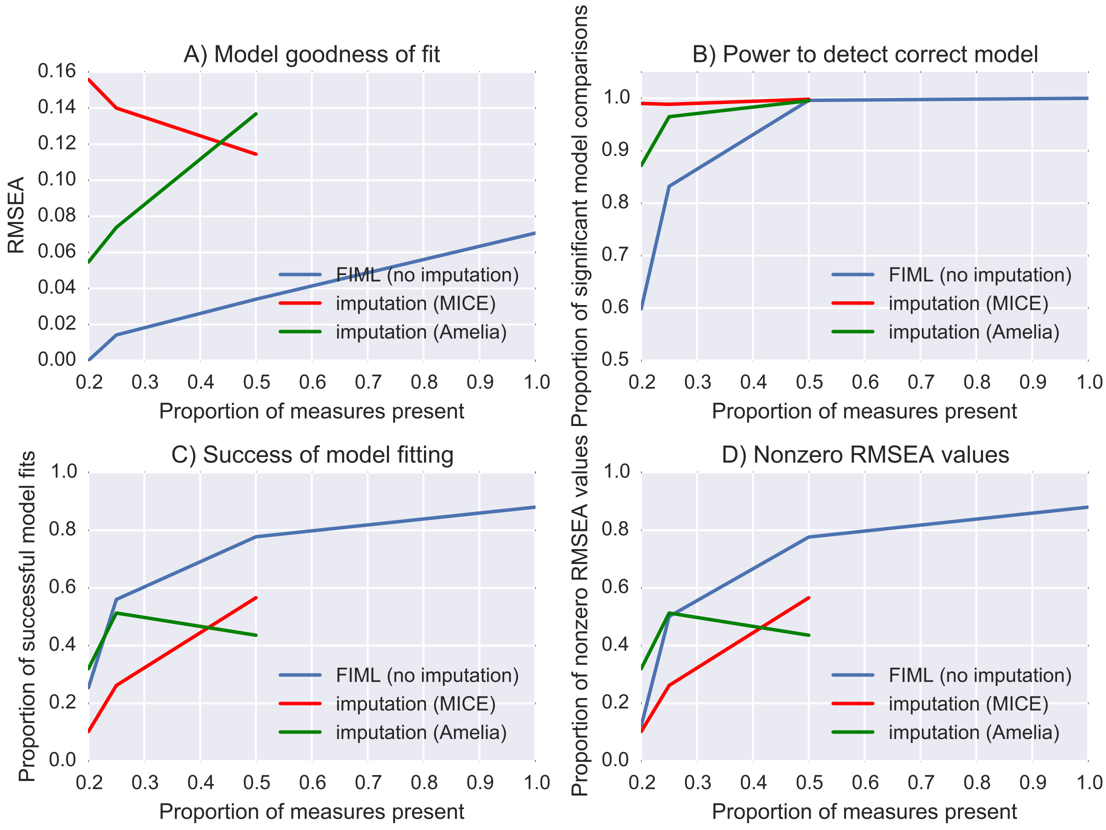


data compared to analyses using imputation. All subsequent analyses are conditionalized upon successful model fitting.

Divergent results were observed between analyses using imputation and those using FIML with missing data. Model goodness of fit increased as the degree of subsampling increased for FIML, whereas it decreased (and was much lower overall) for models using imputation. One concern is that a substantial number of models had an RMSEA value of zero, which may suggest model degeneracy even for models that did not fail outright (Figure 3D).

For model comparison, results also differed between missing data approaches. For models using imputation, there was very high power to identify the correct model even at the lowest levels of sampling; for models using FIML without imputation, power decreased as the proportion of sampled measures fell below 0.5, and fell below the nominal 80% level at a sampling proportion of 20%.

Figure 3: **Results from simulations**



4 Summary

The foregoing results suggest that we can reasonably move to sampling a very small proportion of tasks per subject, though this decision depends on a complex set of tradeoffs between different figures of merit across the simulations. In particular, it appears that the power to detect differences between correct and incorrect models remains high when using imputation, but is reduced when using FIML will missing data. On the other hand, model goodness of fit was very good for FIML and increased with proportion of missing data, due to its dependence on sample size. It appears that overall, imputation was more effective using Amelia versus MICE.

The one major remaining concern is the high degree of model fitting failure; even in the complete case simulations there was more than 10% failure of model fitting, and this increased to almost 80% modeling failure rate in the case with 80% of values missing, both for FIML and for imputation using Amelia. It would be useful to gain more insight into these failures so that we can avoid a case where our proposed models cannot be fit.

Another open question is whether these results will extend to the case of a larger dataset like the one being proposed here, with more than 60 tasks and likely more than 150 dependent variables. We need to devise a more specfiic plan for the analyses that will be performed on the data, so that we can better assess the impact of planned missingness.