Standard Errors of SAPA Correlations: A Monte Carlo Analysis

Ashley Brown

Department of Psychology
Northwestern University

ISSID: July 31, 2015

Analyzing SAPA Data

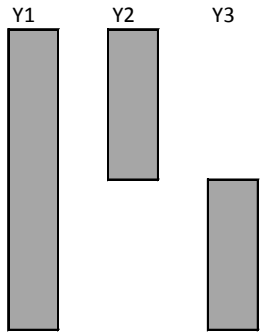
- Available-case analysis (ACA) is used to derive synthetic correlations from 'MMCAR' SAPA data.
- Most large-scale surveys use full-information maximum likelihood (FIML) or multiple imputation (MI) techniques to analyze their incomplete data.
- Why don't we do the same?

Model-based Methods

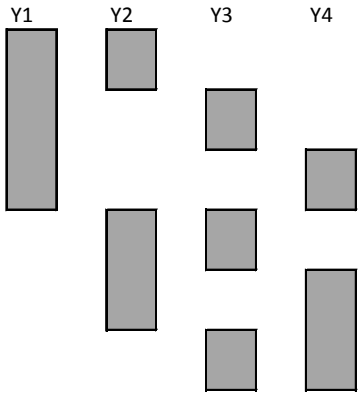
- 'Model-based' methods like maximum likelihood (ML), FIML, and MI require analysts to specify the probability distribution (model) to which they expect their data will conform.
- Pros: Good statistical properties, abundant software (e.g., Expectation Maximization, Dempster et al., 1977)
- Cons: Sensitivity to model misspecification, difficult mathematics/software, data-pattern problems

Patterns in Missing Data

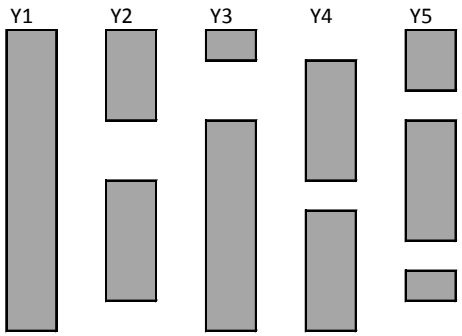
[a] File Matching



[b] Balanced Incomplete Block



[c] General



The Pattern Problem

Randomly Sampled Data Are 'Generally Missing'

- Randomly-sampled SAPA data exhibit a 'general pattern' of missingness, which is hardly a 'pattern' at all.
- Full-information maximum likelihood (FIML) weights *patterns* in sampled data; intuitively, it seems ill-suited to the analysis of SAPA data.
- What about non-model-based methods?

Non-Model-Based Methods

- Non-model-based methods may make assumptions about the model, but they do not require explicit specification thereof.
- Two standard techniques: Complete case analysis and available case analysis (ACA, otherwise known as 'pairwise complete').
- Pros: Often easier to understand and implement; use may be advisable when model is unknown
- Cons: Potential bias, error

Available Case Analysis

What Is It?

- Available-case analysis (ACA) uses all observed scores in parameter estimates, regardless of response-set completeness.
- The interpretation of ACA for univariate statistics (means, variances) is simple.
- The interpretation of ACA for multivariate statistics (covariances, correlations) is not.

Available Case Analysis

ACA Correlations

- The correlation between two random variables X and Y is

$$\rho(X, Y) = \frac{\mathbf{Cov}(X, Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}.$$

- Two common species of ACA correlation: listwise-complete variance and pairwise-complete variance correlations.
- Listwise-complete: Uses all participants with X in **Var**(X), all participants with Y in **Var**(Y).
- Pairwise-complete: Uses only participants with X and Y in **Var**(X) and **Var**(Y).

How to Use ACA Correlations

Proceed with Caution (and Use Pairwise-Complete Variance)

- Many claim that ACA correlations are acceptable estimators when data are MCAR, but no one seems to have *proven* that they are.
- The pairwise-complete variance correlation is generally preferred to the listwise-complete variance correlation (Wilks, 1932; Matthai, 1951; Little and Rubin, 2002).
- SAPA analyses use pairwise-complete variance correlations.
- Specifically, we create an interitem correlation matrix in which all items in all scales appear. Each item in the matrix is a pairwise-complete variance correlation between two items.

Synthetic correlations between SAPA scales

- Correlations between SAPA scales are created 'synthetically' from the interitem correlation matrix.
- Specifically, we divide the sum of the correlations of items between SAPA scales (scales' covariance)...
- ...by the square root of the product of the sums of the correlations of items within SAPA scales (scales' variances).

Justifying the SAPA Method

Simulation and Monte Carlo Analysis

- Goal: Find out whether correlations between SAPA scales obtained from ACA and 'MMCAR' data are good (precise, unbiased) estimators.
- Given the sparse literature, we used simulated data and Monte Carlo analysis to see whether SAPA analyses' results conformed to expectation.
- We wrote our simulation program using R (R Core Team, 2014).
- Five independent variables and two types of statistical analyses (FIML and ACA) yielded 16 dependent variables.

The Parameter Space

Five Independent Variables (IVs)

- N: Number of simulated participants (100, 400, 1600, 6400).
- n: Size of simulated scales (1, 2, 4, 8, or 16 items).
- p: Proportion of items taken (1, .5, .25, .125).
- R_b: Between-scales item intercorrelations (.1, .05, 0).
- R_w: Within-scales item intercorrelations (.2, .3, .4).
- True (latent) value of between-scales correlation is R_b/R_w.
- 720 combinations of IVs.

Simulation Procedure

- 500 replications run on each combination of independent variables (N, n, p, R_w, R_b).
- Each replication produced a randomly-generated, continuous dataset constrained by the active IV combination.
- Each randomly-generated dataset was analyzed using both ACA and FIML to produce a sample correlation (r).
- ACA and FIML sample correlations corrected for reliability were also computed.
- For $n = 16$, a minres factor analysis with oblimin rotation sought a two-factor solution at each replication.
- All statistics and their standard errors (SEs) were calculated by taking the mean and standard deviation, respectively, of the 500 sample statistics for that IV combination and type of analysis (ACA or FIML).

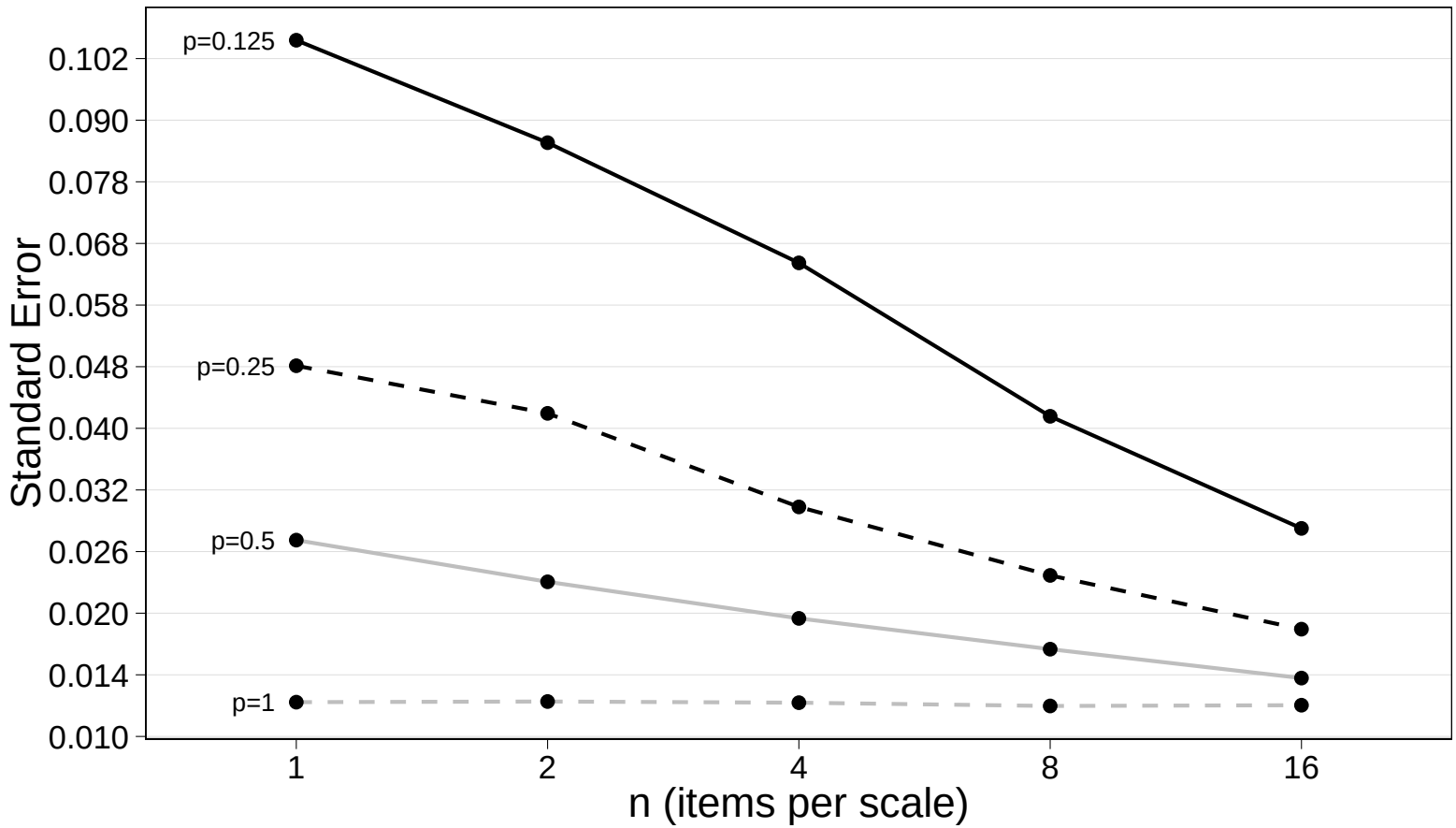
Results: Precision

ACA Uncorrected Correlations' SEs

- Results will focus on both bias and precision (SEs), starting with precision.
- As *N* increases, SE of uncorrected *r* decreases.
- As *p* decreases, SE of uncorrected *r* increases.
- Most importantly: As *n* increases, SE of uncorrected *r* decreases.

SE(r)*: Benefits of aggregation increase with *n*

*SE of ACA Uncorrected *r*, *Rb* = .05, *Rw* = .3, *N* = 6400



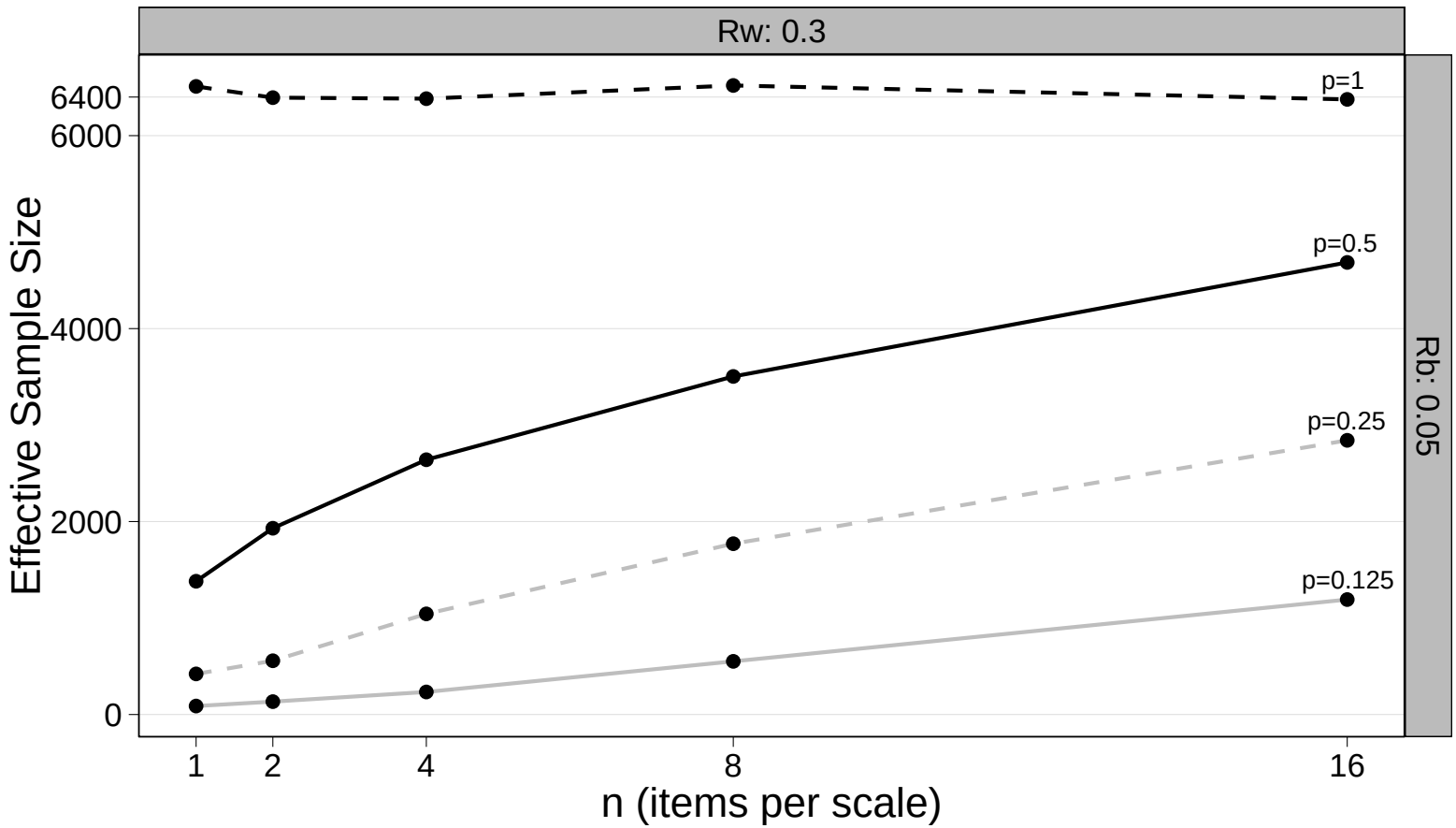
Another Way of Looking at the Effect of n

Effective Sample Size

- Expected v. observed effective sample sizes (ESS)
- Expected ESS = Np^2
- Observed ESS = $[(1 - r^2)^2]/ [(1 + r^2)(SE(r)^2)]$

ACA: Expected v. Observed ESS, $N = 6400$

Expected ESS = 100, 400, 1600, 6400 for $p = .125, .25, .5, 1$, respectively



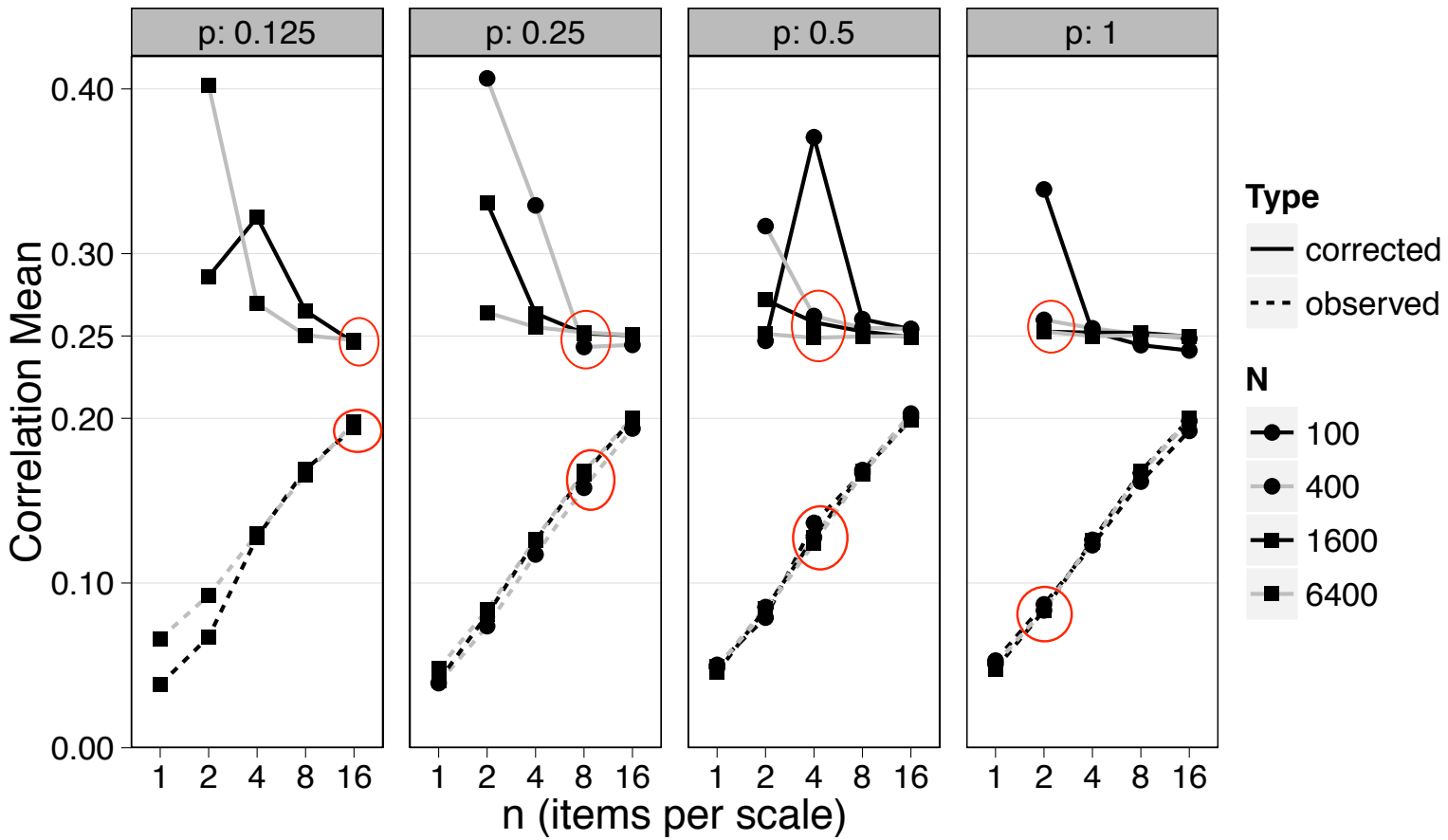
Results: Bias

ACA Uncorrected Correlations

- As n increases, uncorrected r approaches the expected value of corrected r (R_b/R_w).
- Other IVs had little or no effect on uncorrected r.
- Moral: If you're not going to correct for reliability, then you'd better consider aggregating.

Bias of Uncorrected and Corrected Correlations

ACA, *Rb* = .05, *Rw* = .2



Results

Bias and Precision of ACA Alpha-Reliability-Corrected Correlations

- Expected value of corrected r is R_b/R_w.
- Corrected r for incomplete data approached expected (latent) values with more participants and less missingness.
- Like SEs of uncorrected r : SEs of corrected r decrease with more participants, less missingness, and larger scale size.

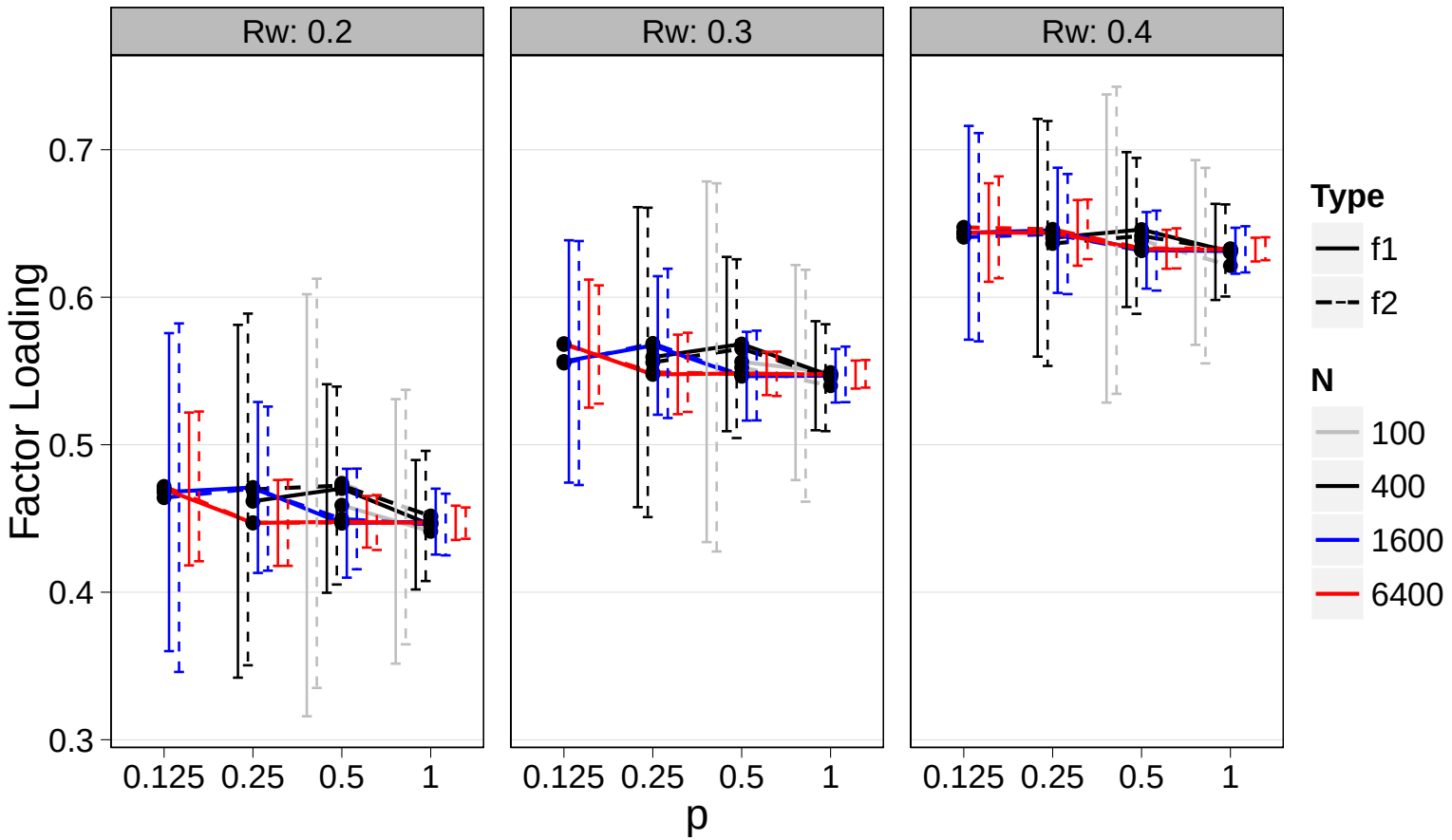
Results: ACA Factor Analysis

Precision and Bias of Loadings and Intercorrelations

- Factor analyses were performed for $n = 16$ only.
- Intercorrelations' (*phi*) latent value = R_b/R_w
- As N and p increase, *phi* approaches its latent value.
- Intercorrelations' SEs decreased as N and p increased.
- Loadings' latent value = square root of R_w.
- Only R_w strongly affected loadings' value.
- Loadings' SEs decreased as N, p, and R_w increased.

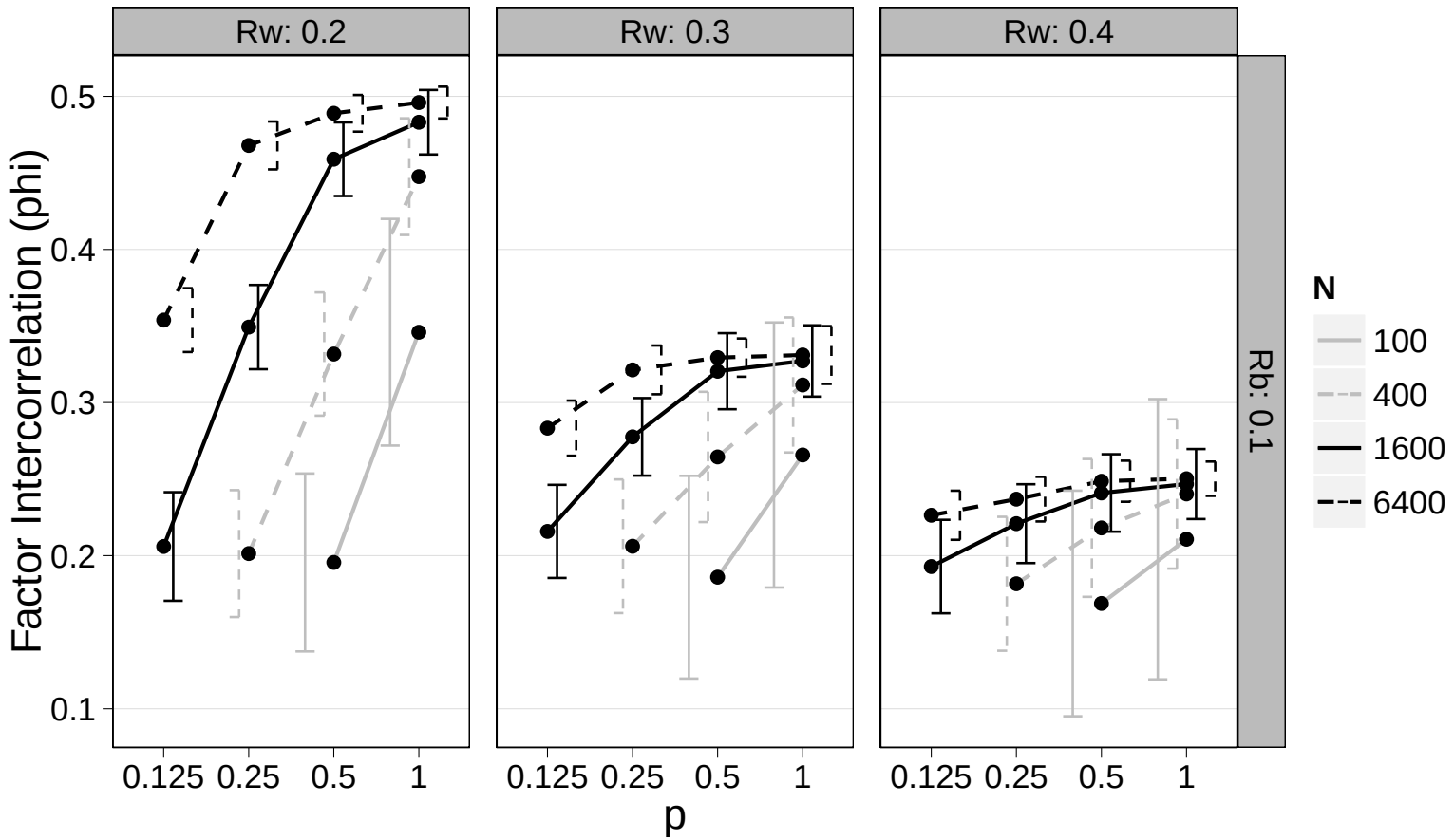
ACA Factor Loadings

Means and Standard Errors, *Rb* = 0



ACA Factor Intercorrelations

Means and Standard Errors, *Rb* = .1



Results: FIML

Precision and Bias Relative to ACA

- The patterns in the FIML data were the same as in the ACA data.
- However (and unsurprisingly), FIML tended to fail more often than ACA did.
- (Precision of ACA relative to FIML) = (ACA SE/FIML SE); in general, FIML was slightly more precise.
- (Bias of ACA relative to FIML) = (ACA statistic - FIML statistic); in general, FIML and ACA did not differ in bias.

Results in Context

- Summary: The combination of ACA and MCAR data produce estimates of correlations that are only slightly less precise and no more biased than those obtained using FIML.
- Important: Effect of *n* on correlations' SEs.
- Comforting: Most results conformed to expectation; no problem with finding correlation from ACA factor analyses.
- Limitations: Simulated rather than real data, clean rather than messy simulated data, no rigorous theoretical model.

Conclusion

- SAPA's correlations have good statistical properties.
- SAPA techniques can be applied to many survey designs where one wants to increase breadth of coverage but is limited in the number of items that can be presented.
- With as few as 500-1000 subjects, it is clear that it is better to present random samples taken from longer scales than it is to present short forms of equivalent length to all participants.

Contact Information

- SAPA's correlations have good statistical properties.
- SAPA techniques can be applied to many survey designs where one wants to increase breadth of coverage but is limited in the number of items that can be presented.
- With as few as 500-1000 subjects, it is clear that it is better to present random samples taken from longer scales than it is to present short forms of equivalent length to all participants.
- If you have questions or comments, please email me at AshleyBrown2011@u.northwestern.edu
- Thank you!

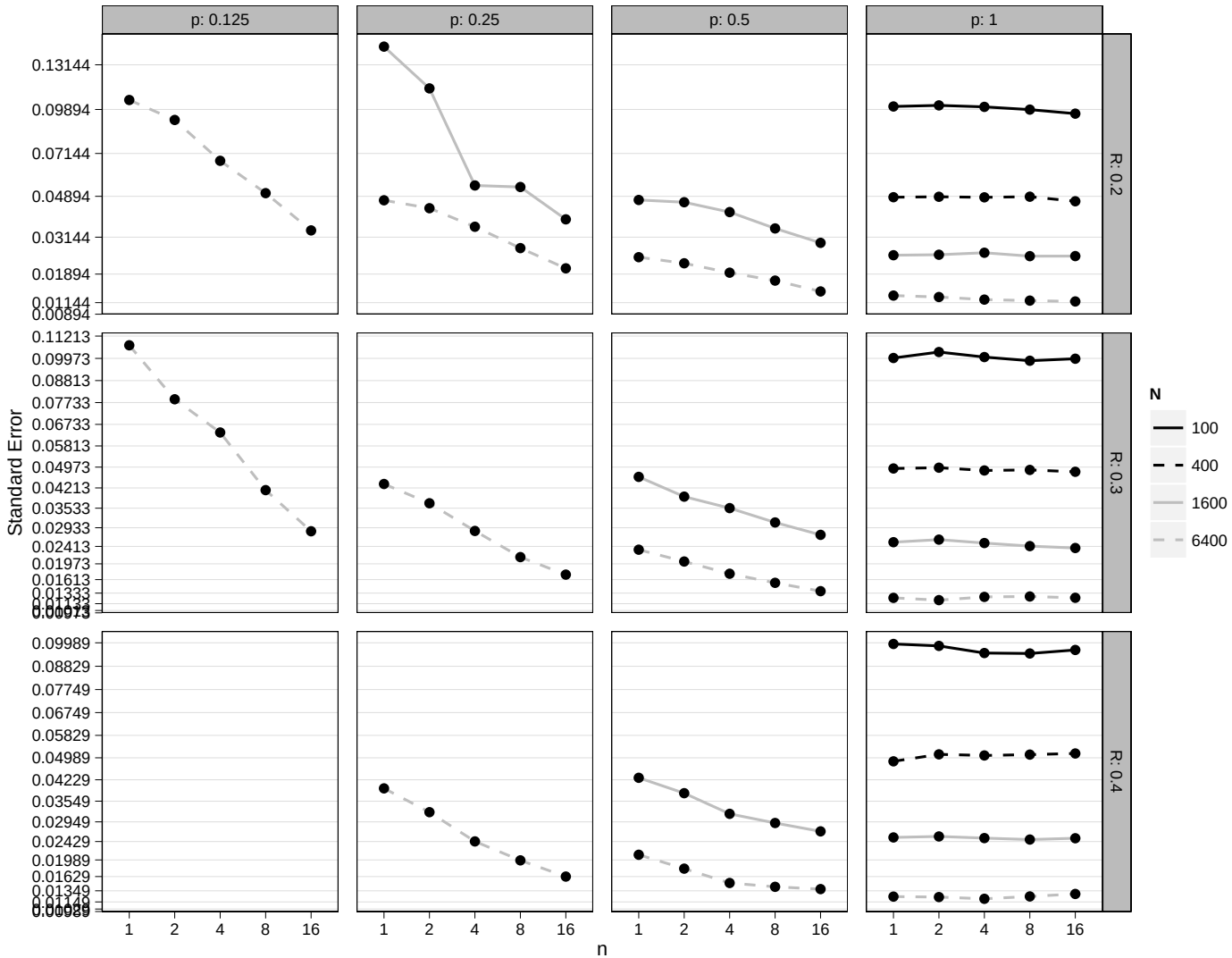
Precision and Bias of ACA Relative to FIML

All DVs

- Precision (ACA/FIML) range:
- Uncorrected correlations' SEs: 0.60 - 1.75
- Corrected correlations' SEs: 0.42 - 2.15
- Factor intercorrelations' SEs: 0.48 - 1.32
- Factor loadings' SEs: 0.83 - 2.54
- Bias (ACA - FIML) range:
- Uncorrected correlations: -0.05 - 0.05
- Corrected correlations: -0.21 - 0.21
- Factor intercorrelations: -0.06 - 0.05
- Factor loadings: -0.01 - 0.01

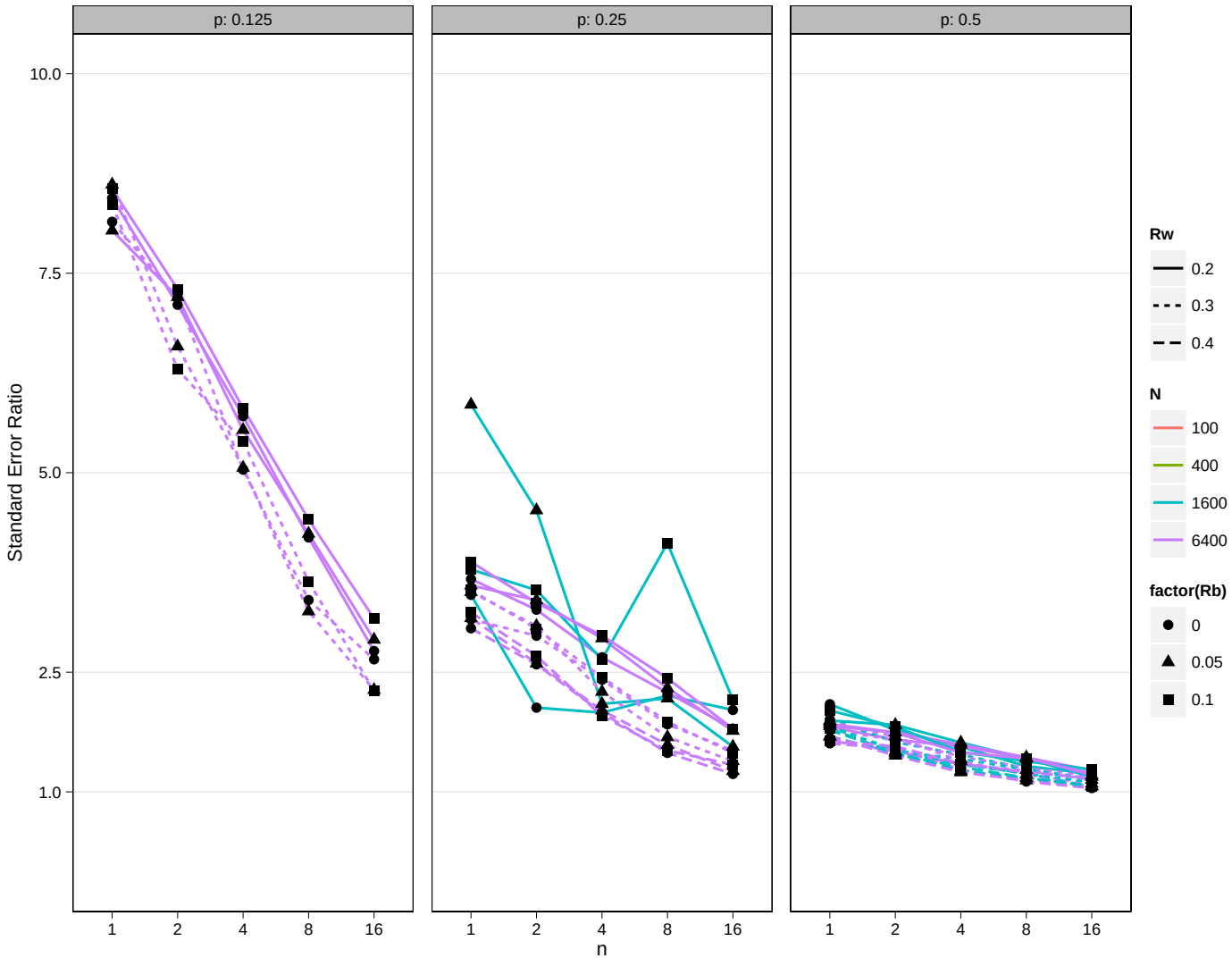
SE(r)*: Benefits of aggregation increase with *n*

*SE of FIML Uncorrected *r*, *Rb* = .05



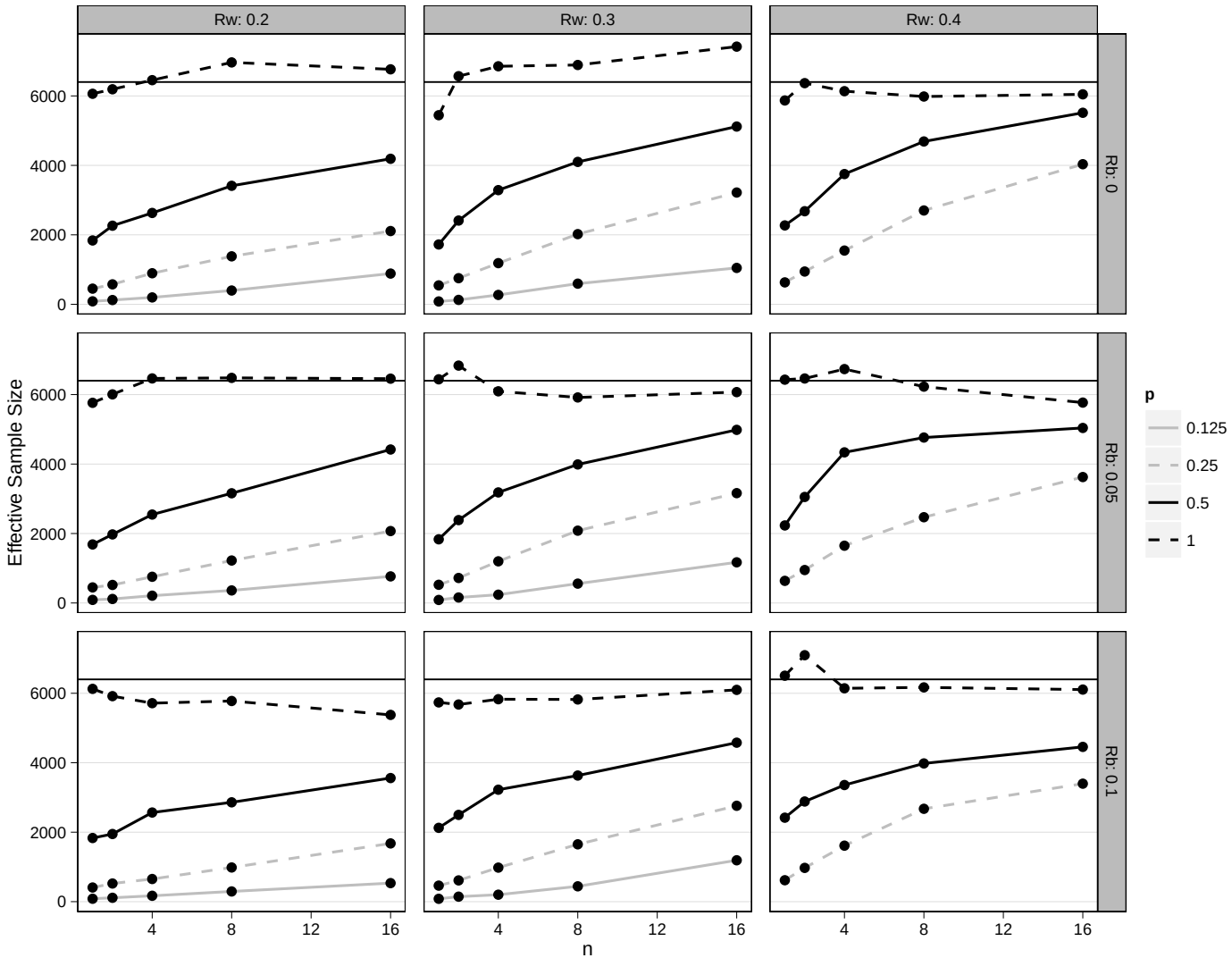
SE(r) ratio: Benefits of aggregation increase with *n*

FIML



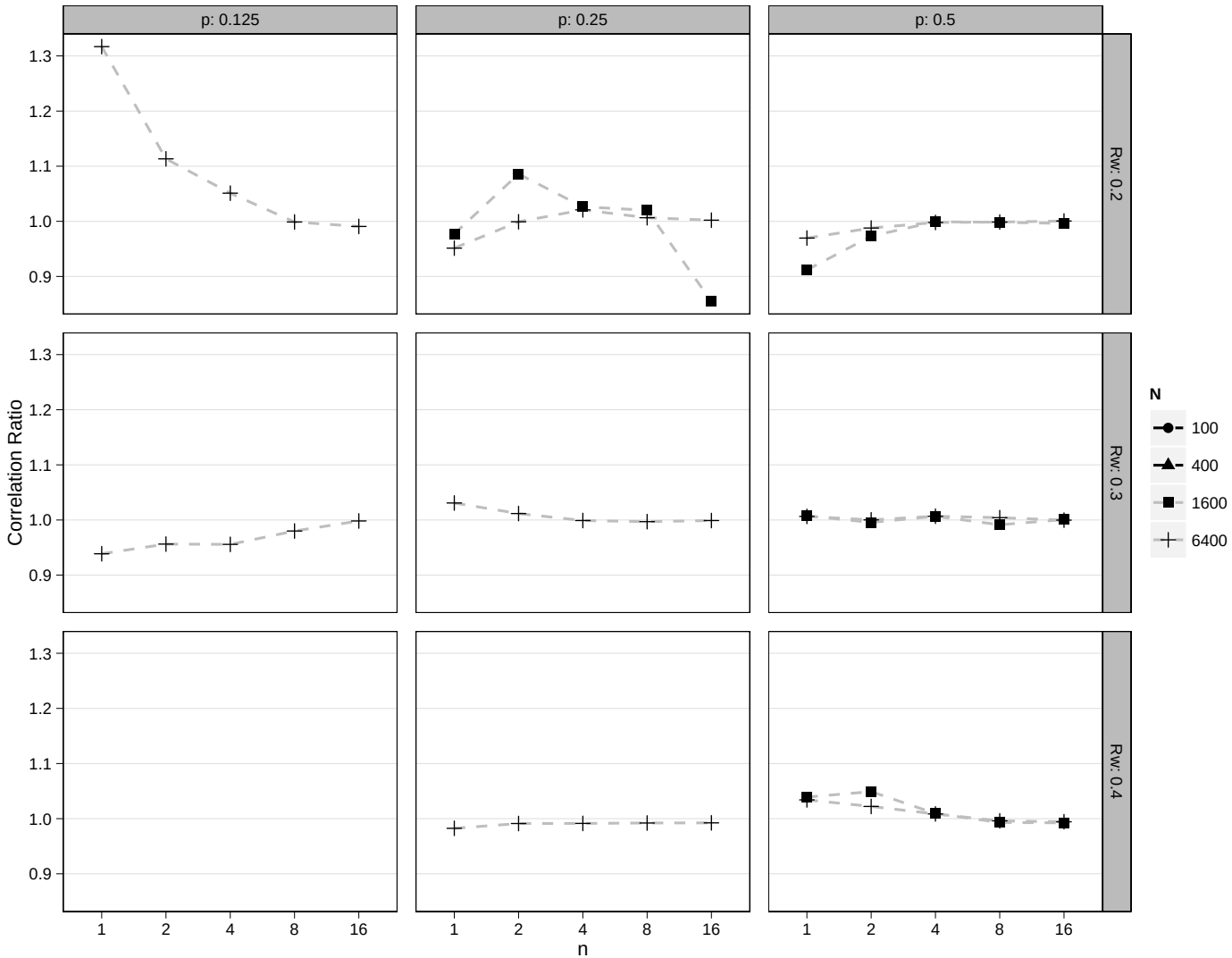
FIML: Expected v. Observed ESS, $N = 6400$

Expected ESS = 100, 400, 1600, 6400 for $p = .125, .25, .5, 1$, respectively



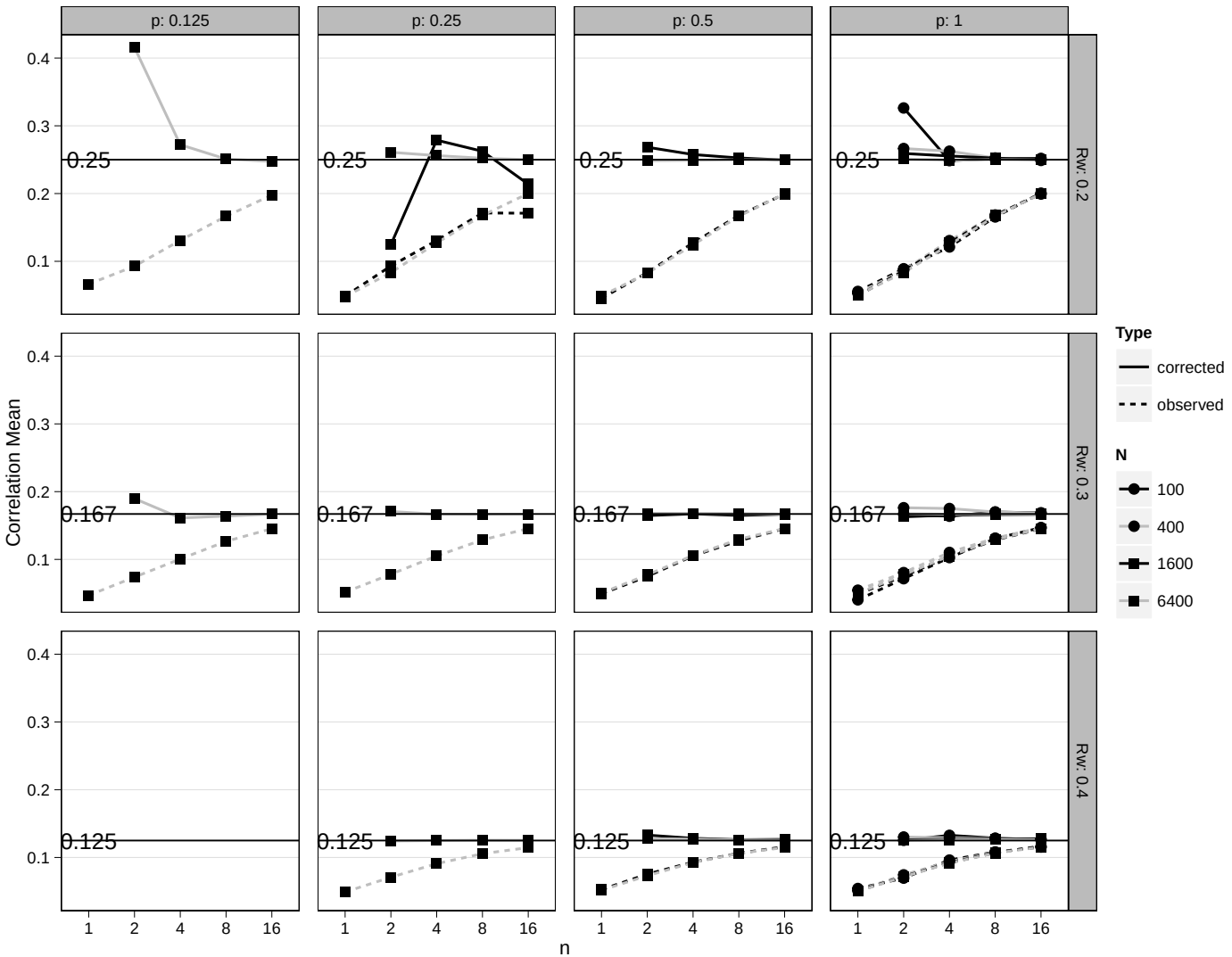
Results, Part 2: FIML

Uncorrected *r* (*p* <1)/ Uncorrected r (*p* = 1), *Rb* = .05



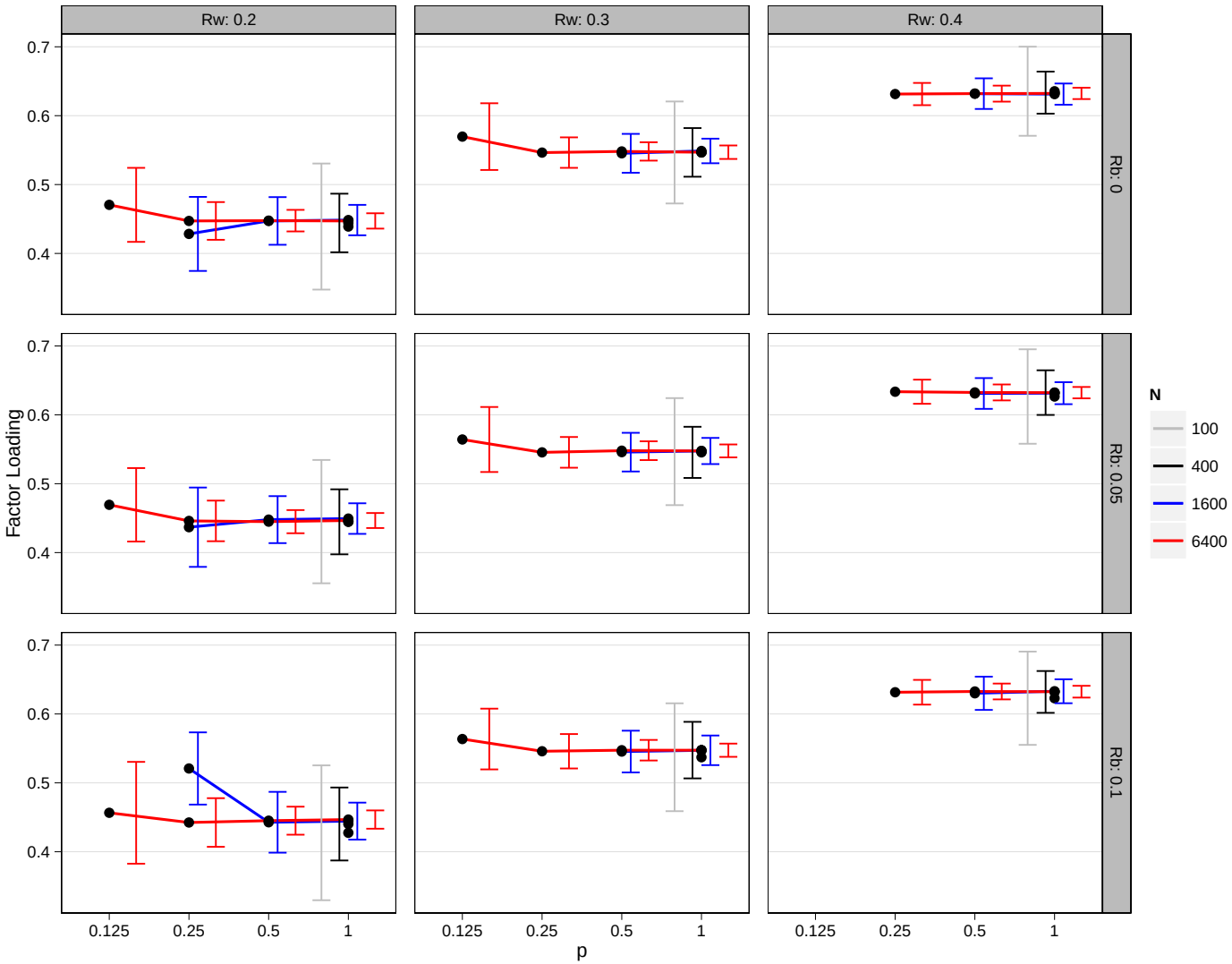
Uncorrected (dot) v. Corrected (solid) Correlations

FIML, *Rb* = .05



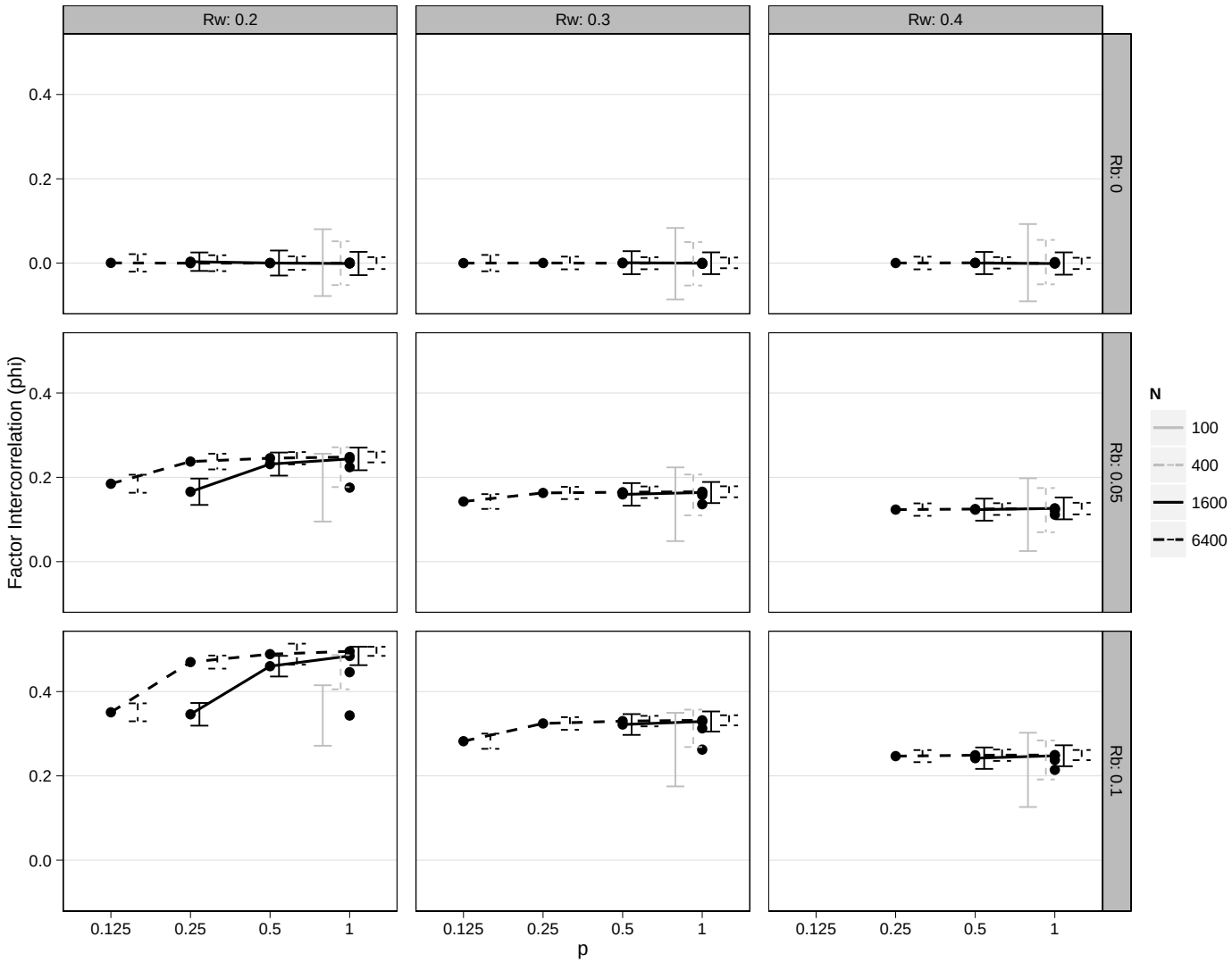
Results, Part 3: FIML

Factor Loadings: Means and Standard Errors



FIML Factor Intercorrelations

Means and Standard Errors



References I

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data.

Matthai, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 11(2):145–152.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Wilks, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):163–195.