Evaluating Different Ability Estimation Methods for Strand Scores in a K-12 Computerized Adaptive Test: Perspective of Test-Retest Reliability

Johnny Denbleyker

Houghton Mifflin Harcourt

Ideas in Testing seminar

Chicago, IL.

November 2015

- Move to adaptive accountability Math assessment
- Based on 3PL model
- Change from reporting percent correct Domain score to estimating theta ability and converting to Scale Scores
- “Stanine-like” 1-9 metric used for reporting Domain scores
- Year 1-2 used MLE estimation of Domain ability based on 3PL model
- Year 3 used EAP *N*(0,1) estimation of Domain ability based on 3PL model
- Year 3 incorporated pre-accountability Fall/Winter “Optional Local Purpose Test” that mimics the Spring NCLB CAT assessment

Background of Study

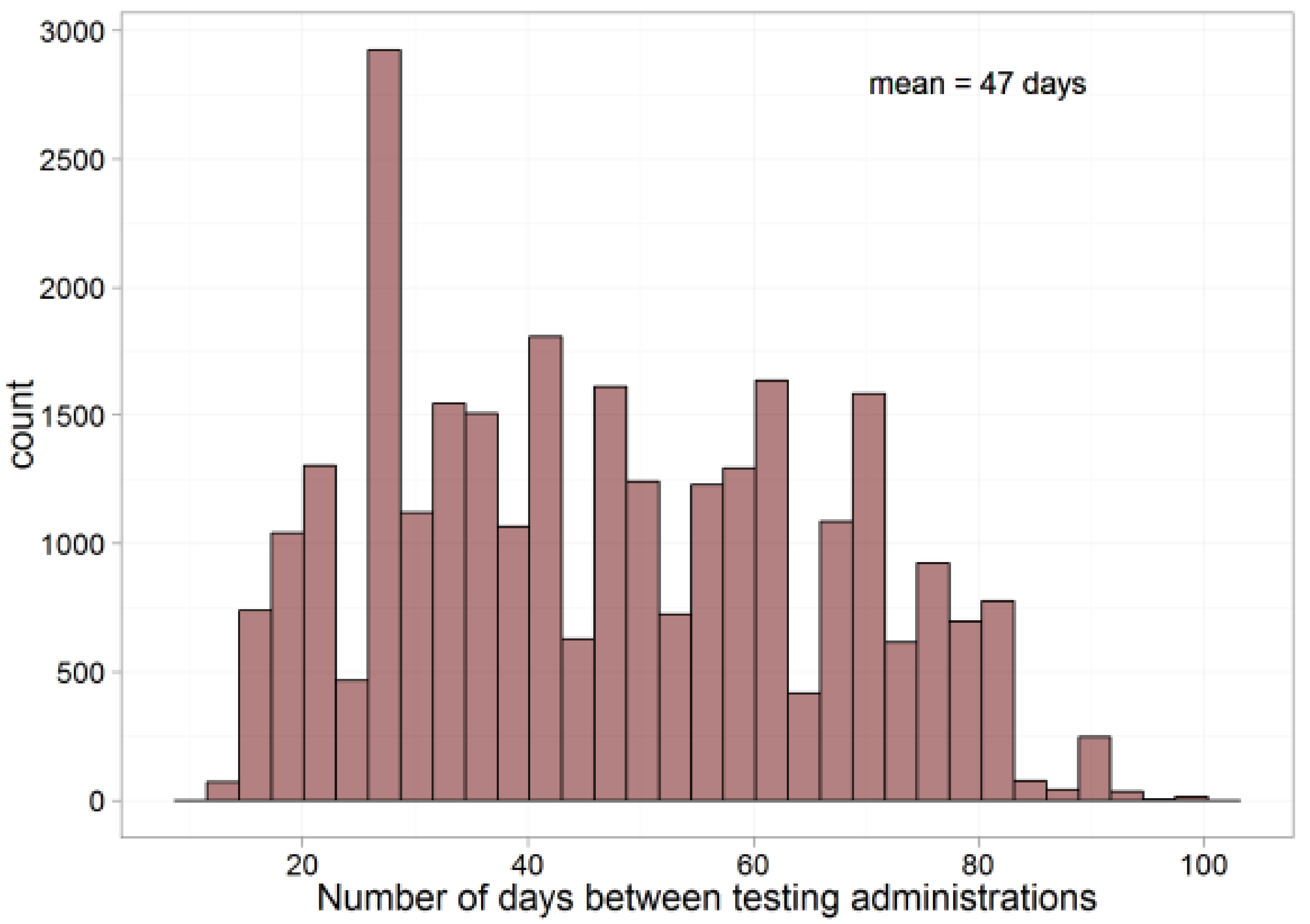
- Previous studies (Bock & Mislevy, 1982; Wang & Vispoel, 1998; Weiss & McBride, 1984; Hanson & Lau, 1999; Wang & Wang, 2001) have noted that the Bayesian methods, such as EAP and MAP, are biased toward the mean of the prior distribution and are thus questionable for use to many standardized testing programs. **BUT…**

- MLE was found to have smaller bias with a direction opposite to that of the Bayesian methods, (i.e., low ability examinees are negatively biased and high ability examinees are positively biased), but have larger standard error (SE) than the Bayesian methods (Warm, 1989)

- In CAT, sub-scores involve more complexity.
- The first set of challenges for sub-scores in CAT involves the number of items per strand typically will decrease (purported benefits of CAT include shorter tests) and examinees get different items. This former will decrease reliability to the extent the sub-scores are not matched properly to the ability level of the examinees.
- The latter makes for raw scores quite problematic for reporting given the raw score involve different sets of items. As a result, IRT pattern scoring will be the logical choice to estimate sub-score ability as similarly done with total score. However, pattern scoring via IRT involve (much) stronger assumptions and increased complexity. Also, the default estimation method for IRT pattern scoring is MLE, which has no way to directly estimate all correct or all incorrect score patterns.
- Finally, the 3PL IRT model is commonly used to estimate ability via MLE for total scores. If the same model is applied to sub-scores, there is a higher risk for problematic likelihood functions, which will generate a level of invalidity into ability estimates for a small percentage (but likely not trivial) of examinees as no sufficient statistic exists under the 3PL model. To add even more issues to consider with sub-score ability estimation under the 3PL model, MLE or MAP estimation methods do not consider the asymmetric likelihood function that exists due to modeling the lower asymptote. Additionally, with the 3PL, expected information and observed information differ)

- Item response data from a grade 6 CAT administered mathematics assessment contained from a statewide accountability assessment is used for the analysis.
- Student’s scores denoted as opportunity 1 and opportunity 2. There were 28,692 matched scores having students testing at least twice within the accountability testing window, which was open from early Feb. to early May.
- Sample consisted of 48.6% of the state-wide population.

Data


mean = 47 days
count
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
Number of days between testing administrations

- The grade 6 math test is comprised of four strands (sub-scores/domains) where the number of items administered varied for each student.
- Strand 1 is “Number & Operation”(14-17 items)
- Strand 2 is “Algebra” (9-12 items)
- Strand 3 is “Geometry & Measurement” (7-9 items)
- Strand 4 is “Data Analysis & Probability” (6-8 items)

Test Blueprint

(a) **To evaluate the stability of domain theta estimates and the corresponding scale scores across ten different ability estimation methods based on a test/retest framework.**

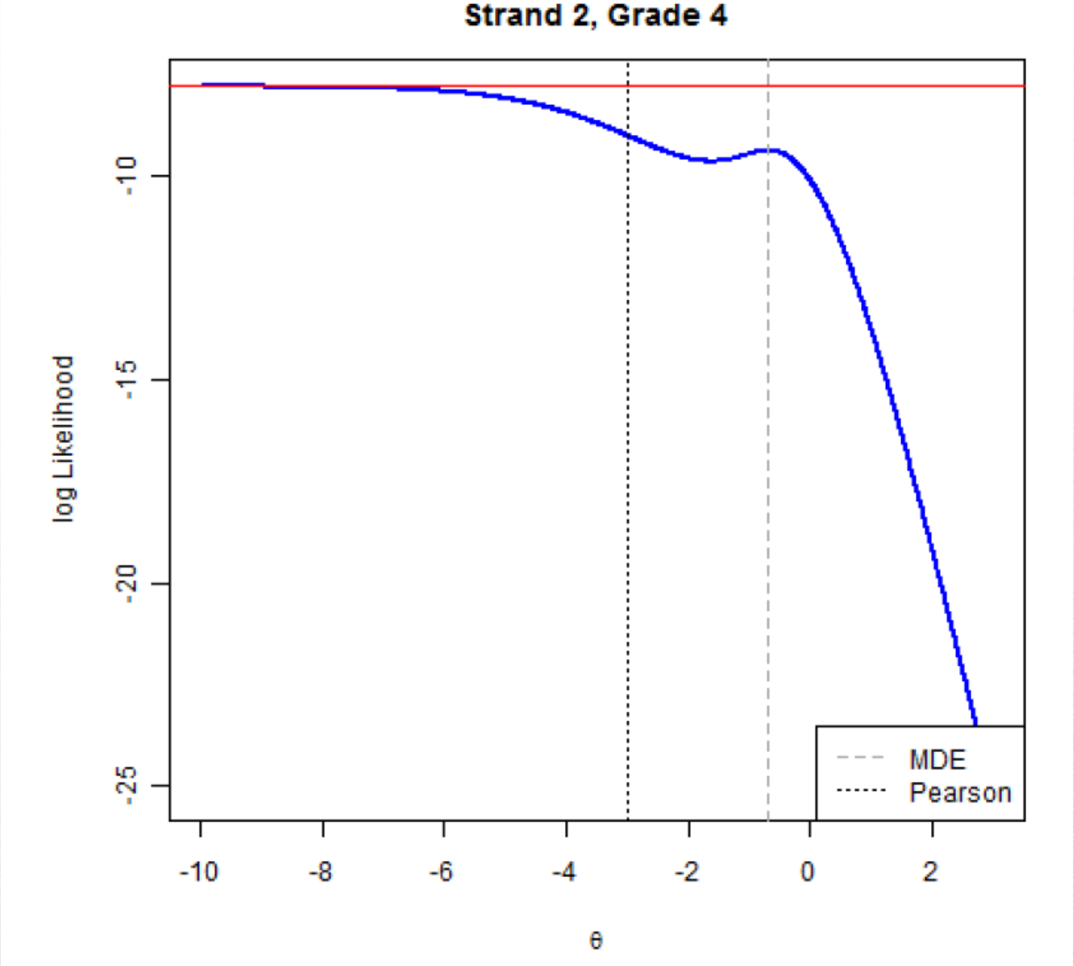
(b) **To evaluate the conditional standard errors of measurement across each method.**

(c) To evaluate how sensitive group mean (aggregate-level) results are to various ability estimators.

(d) To evaluate the distributional properties of scale scores for each of the different estimation methods.

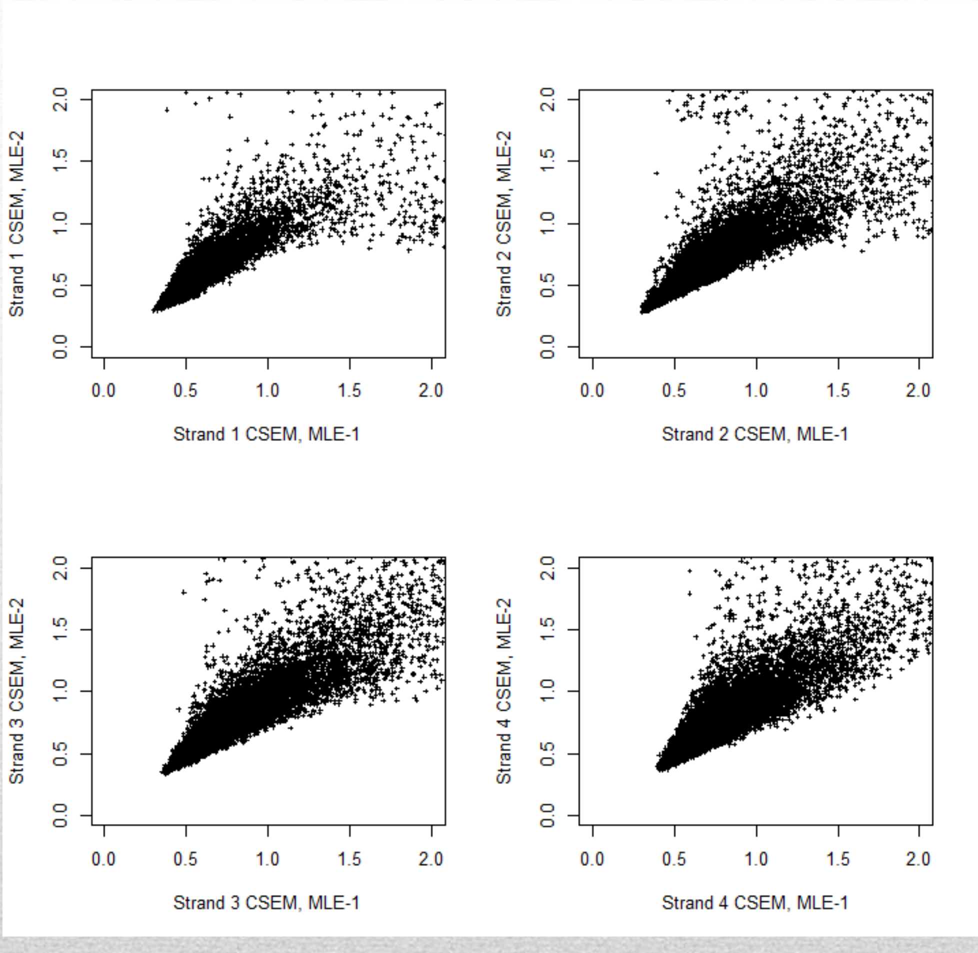
Purposes of Study

3PL log-likelihood function



Will the Real MLE Please Stand Up?

3PL MLE comparison



Will the Real CSEM Please Stand Up?

Methodology

- Ten different estimation methods are compared using matched students across the opportunity 1 and opportunity 2 testing administrations.
- Each estimation method is implemented within R (The R Project for Statistical Computing) using the `irtoys` R package, which allows for flexibility to estimate ability via MLE, EAP, and MAP.
- There are three different variants of MLE, six variants of EAP, and two with MAP. These are described in some detail in the next section.

The `irtoys` functions used in this study include the following:

- **mlebme**—to conduct Maximum likelihood and Bayes Modal estimation of ability
- **eap**—to conduct EAP estimation of ability
- **normal.qu**—to weight likelihood function with EAP and MAP (e.g., specifying prior information)

- ***MLE-1***—Maximum likelihood in the R package `irtoys` using the default direct optimizer routine

- ***MLE-2***—Maximum likelihood in `irtoys` using a reconfigured function to specify “Brent’s” method” using a -10 to 10 search range, and 2) to estimate standard errors from the Hessian evaluated at the MLE.

Ability Estimation Methods

Ability Estimation Methods

- ***EAP-1***—EAP estimation with *N*(0,1) prior

Here, the standard normal prior is used. The criterion applied to the EAP-1 method was to have 61 quadrature points and a boundary range from -3 to +3. By default, the irtoys R package uses a boundary range from-4 to +4 and 15 quadrature points where quadrature points and weights are based on the Normal distribution.

- ***EAP-2***—EAP estimation with *N*(0,10) prior

Provides a very uniform-like weighting along the quadrature points over a boundary range from-0.3 to +0.3 (+/-3 divided by s.d.=10).

EAP-3**—**EAP estimation with *N*(MLE,1) prior – ‘**Empirical Bayes**’

Here, the prior mean for each examinee is their respective total MLE ability estimate (MLE-1 based) that is provided a strong weight (variance=1).

EAP-4**—**EAP estimation with *N*(MLE,3) prior, where MLE is student’s total MLE ability estimate and a diffuse prior variance of 9—which provides a moderate uniform-like weighting along the quadrature points over a boundary range from-1 to +1 (+/-3 divided by s.d.=3).

Ability Estimation Methods

EAP-5**—**EAP estimation with *N*(EAP,1) prior, where EAP is student’s total EAP ability estimate

Here, the prior mean for each examinee is their respective total EAP ability estimate that is provided a strong weight (variance=1). The prior is supplied as a data frame of EAP estimates that is looped over examinees.

EAP-6**—**EAP estimation with *N*(EAP,3) prior, where EAP is student’s total EAP ability estimate.

The same criteria used for *EAP-5*was applied with *EAP-6*except diffuse prior variance of 9—which is coded as a standard deviation within the function, was used that provides a moderate uniform-like weighting along the quadrature points over a boundary range from-1 to +1 (+/-3 divided by s.d.=3).

Ability Estimation Methods

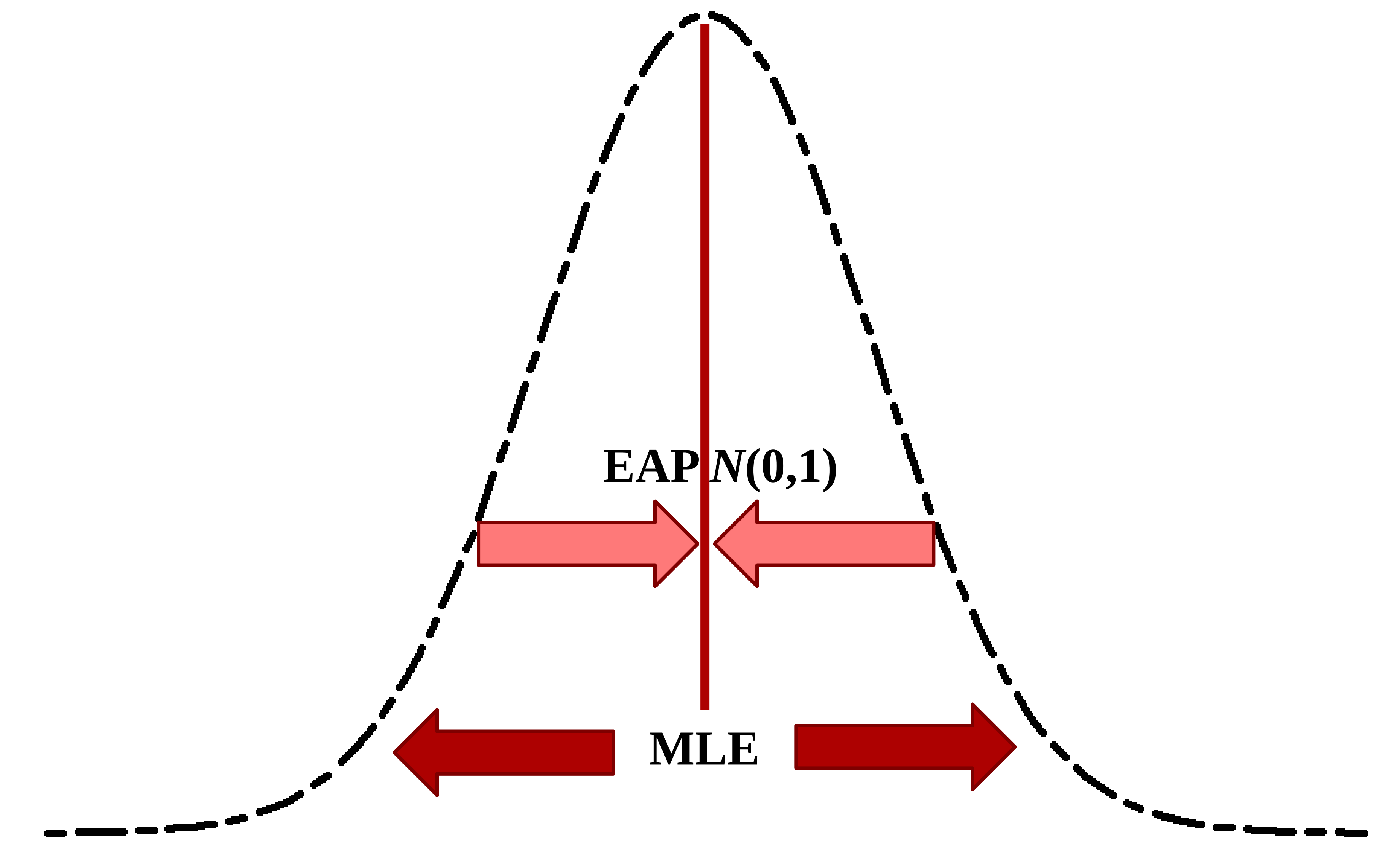
Ability Estimation Methods

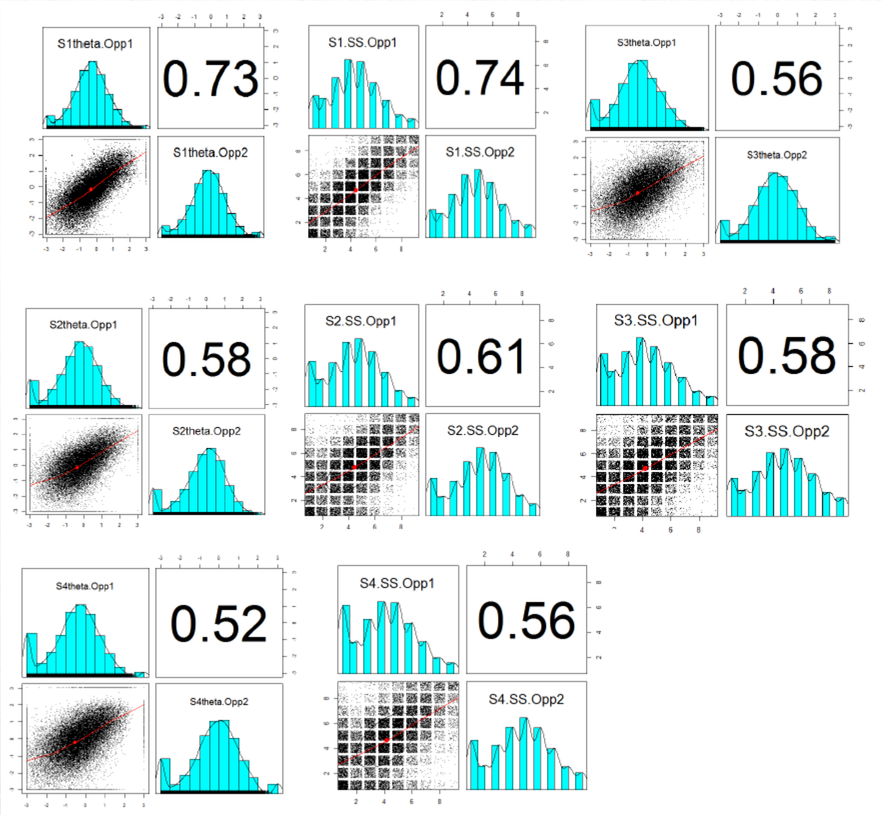
MAP-1—MAP estimation with *N*(0,1) prior

- Here, the standard normal prior is used. The criterion applied to the MAP-1 method was to have 61 quadrature points and a boundary range from -3 to +3.

MAP-2—MAP estimation with *N*(0,3) prior

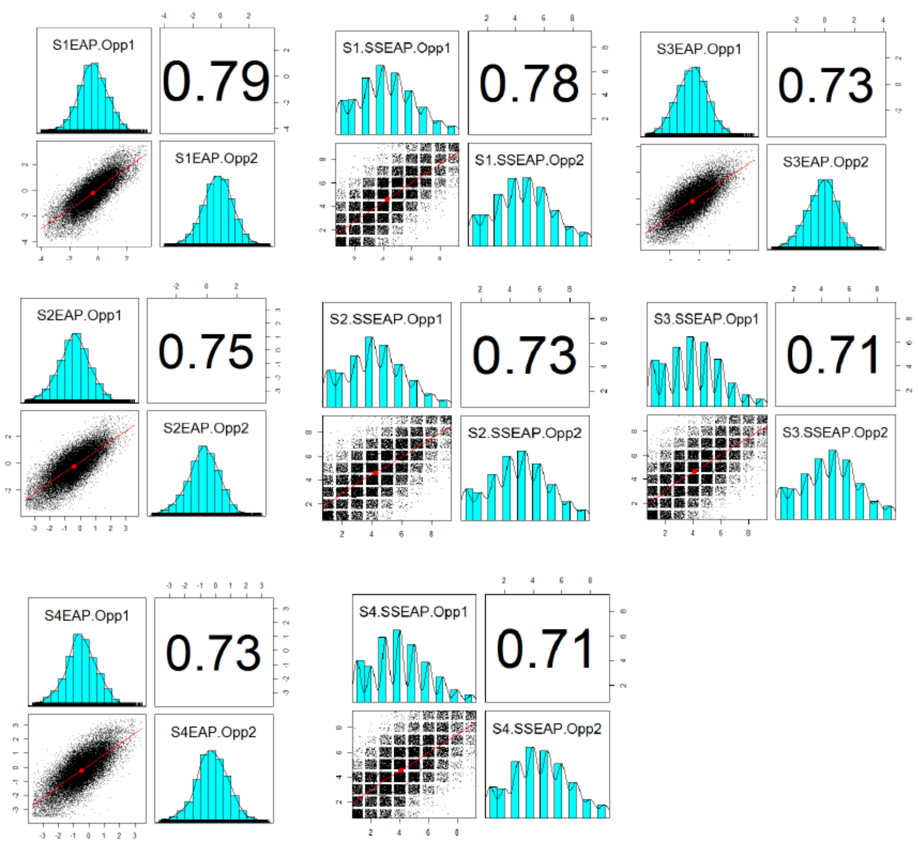
- The same criteria used for *MAP-1* was applied with *MAP-2* except diffuse prior variance of was used that provides a moderate uniform-like weighting along the quadrature points over a boundary range from-1 to +1 (+/-3 divided by s.d.=3).

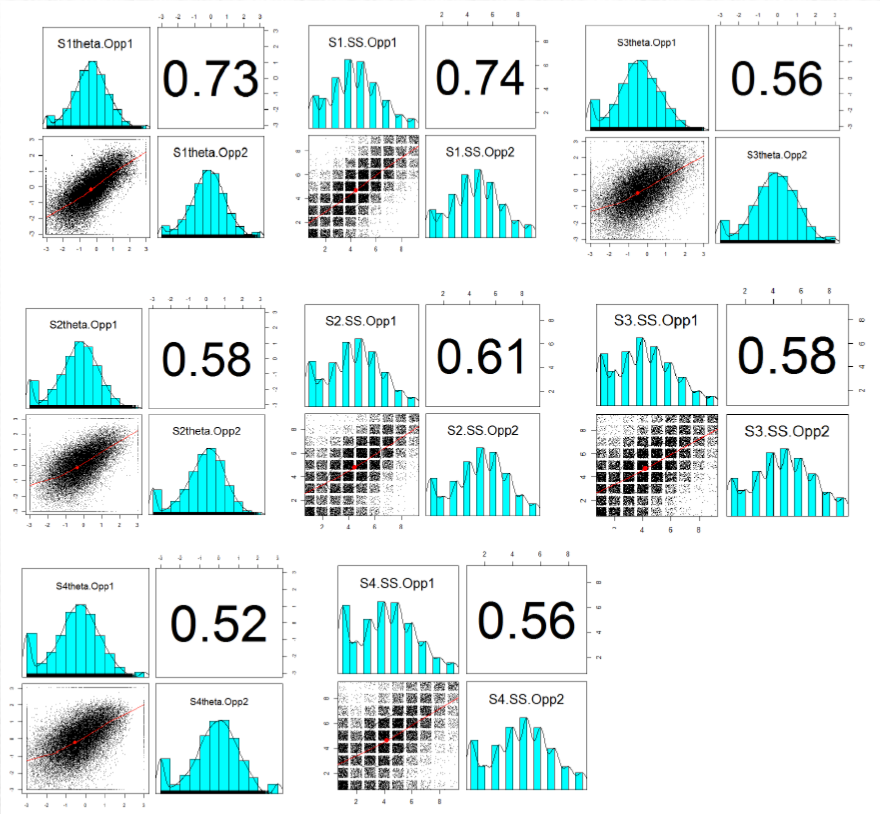

EAP $N(0,1)$
MLE


S1theta.Opp1
0.73
S1theta.Opp2
S1.SS.Opp1
0.74
S1.SS.Opp2
S3theta.Opp1
0.56
S3theta.Opp2
S2theta.Opp1
0.58
S2theta.Opp2
S2.SS.Opp1
0.61
S2.SS.Opp2
S3.SS.Opp1
0.58
S3.SS.Opp2
S4theta.Opp1
0.52
S4theta.Opp2
S4.SS.Opp1
0.56
S4.SS.Opp2

MLE-1

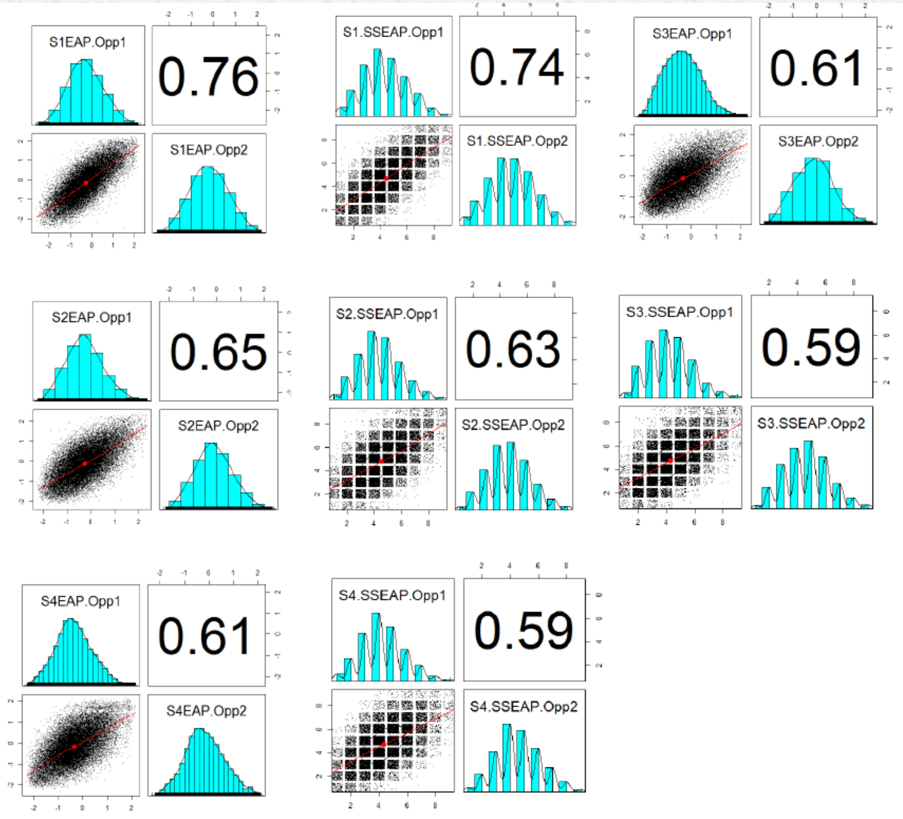
EAP-3

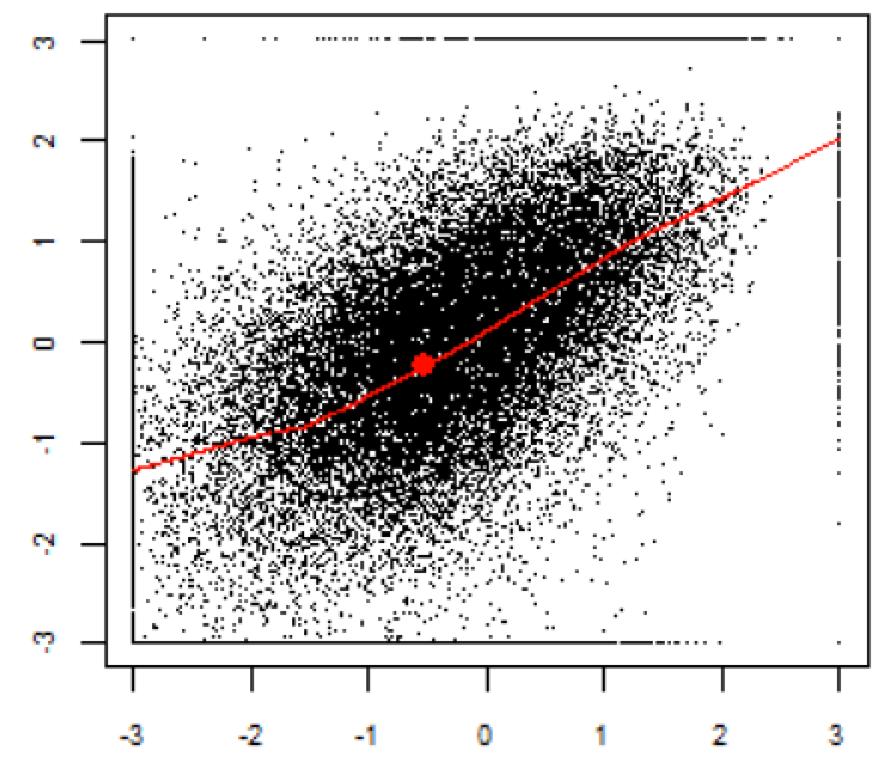

S1EAP.Opp1
0.79
S1EAP.Opp2
S1.SSEAP.Opp1
0.78
S1.SSEAP.Opp2
S3EAP.Opp1
0.73
S3EAP.Opp2
S2EAP.Opp1
0.75
S2EAP.Opp2
S2.SSEAP.Opp1
0.73
S2.SSEAP.Opp2
S3.SSEAP.Opp1
0.71
S3.SSEAP.Opp2
S4EAP.Opp1
0.73
S4EAP.Opp2
S4.SSEAP.Opp1
0.71
S4.SSEAP.Opp2


S1theta.Opp1
0.73
S1theta.Opp2
S1.SS.Opp1
0.74
S1.SS.Opp2
S3theta.Opp1
0.56
S3theta.Opp2
S2theta.Opp1
0.58
S2theta.Opp2
S2.SS.Opp1
0.61
S2.SS.Opp2
S3.SS.Opp1
0.58
S3.SS.Opp2
S4theta.Opp1
0.52
S4theta.Opp2
S4.SS.Opp1
0.56
S4.SS.Opp2

MLE-1

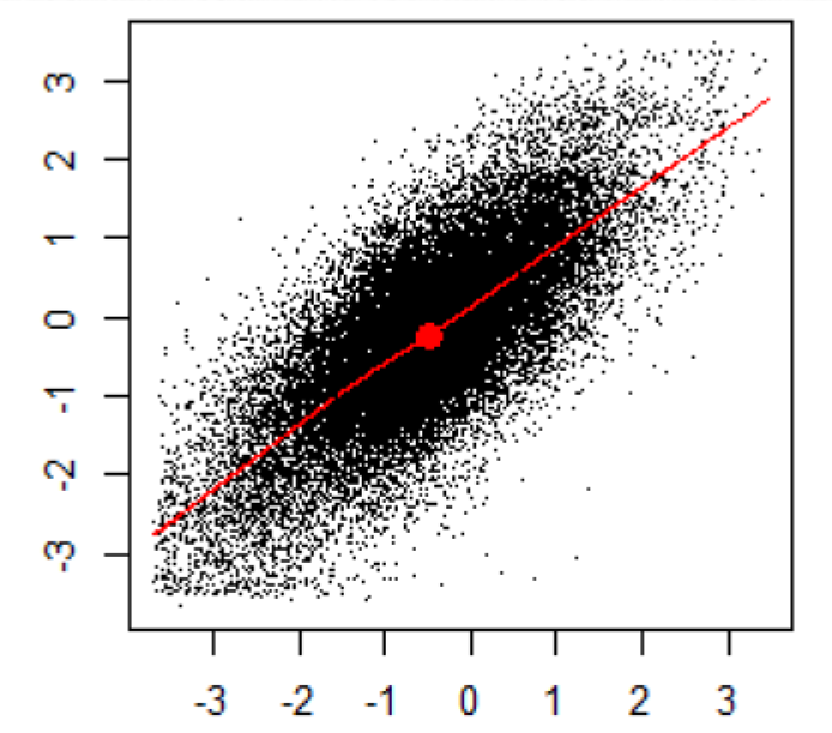
EAP-1


S1EAP.Opp1
0.76
S1EAP.Opp2
S1.SSEAP.Opp1
0.74
S1.SSEAP.Opp2
S3EAP.Opp1
0.61
S3EAP.Opp2
S2EAP.Opp1
0.65
S2EAP.Opp2
S2.SSEAP.Opp1
0.63
S2.SSEAP.Opp2
S3.SSEAP.Opp1
0.59
S3.SSEAP.Opp2
S4EAP.Opp1
0.61
S4EAP.Opp2
S4.SSEAP.Opp1
0.59
S4.SSEAP.Opp2

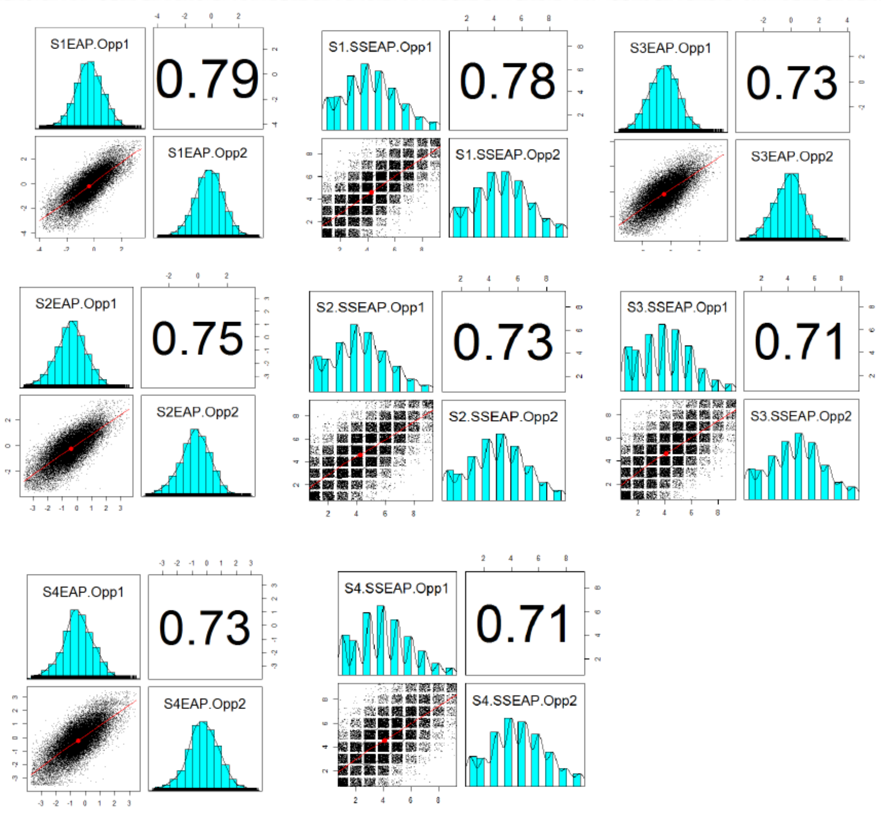


MLE-1

EAP-3

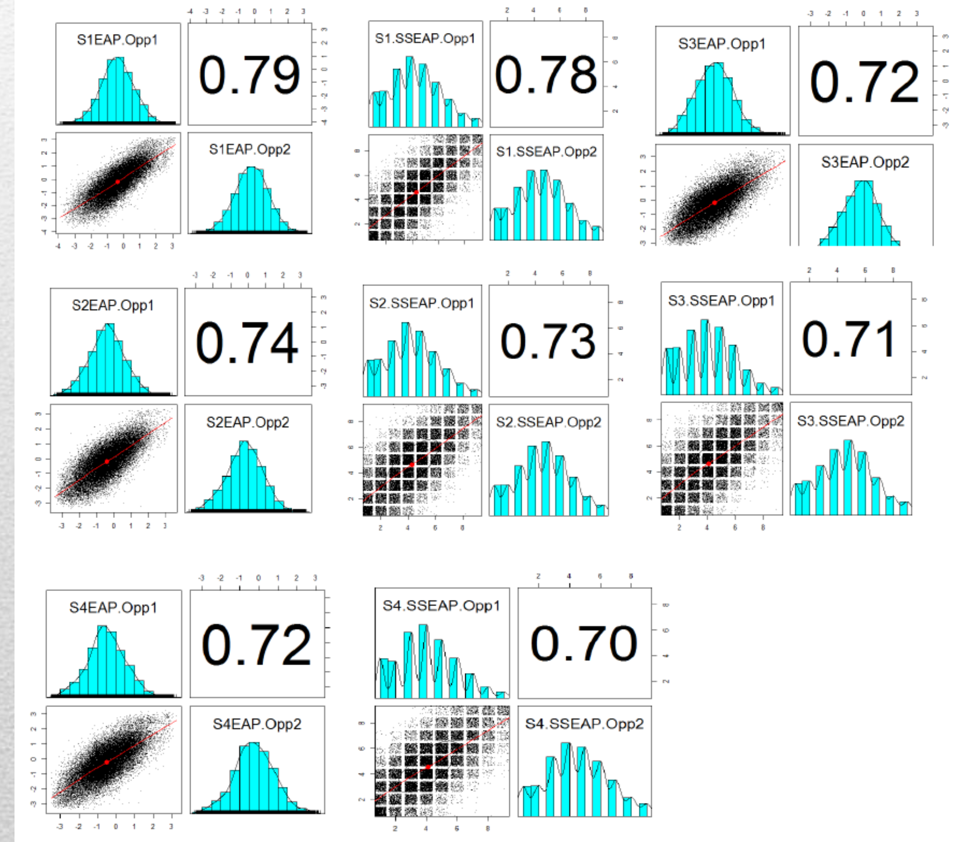


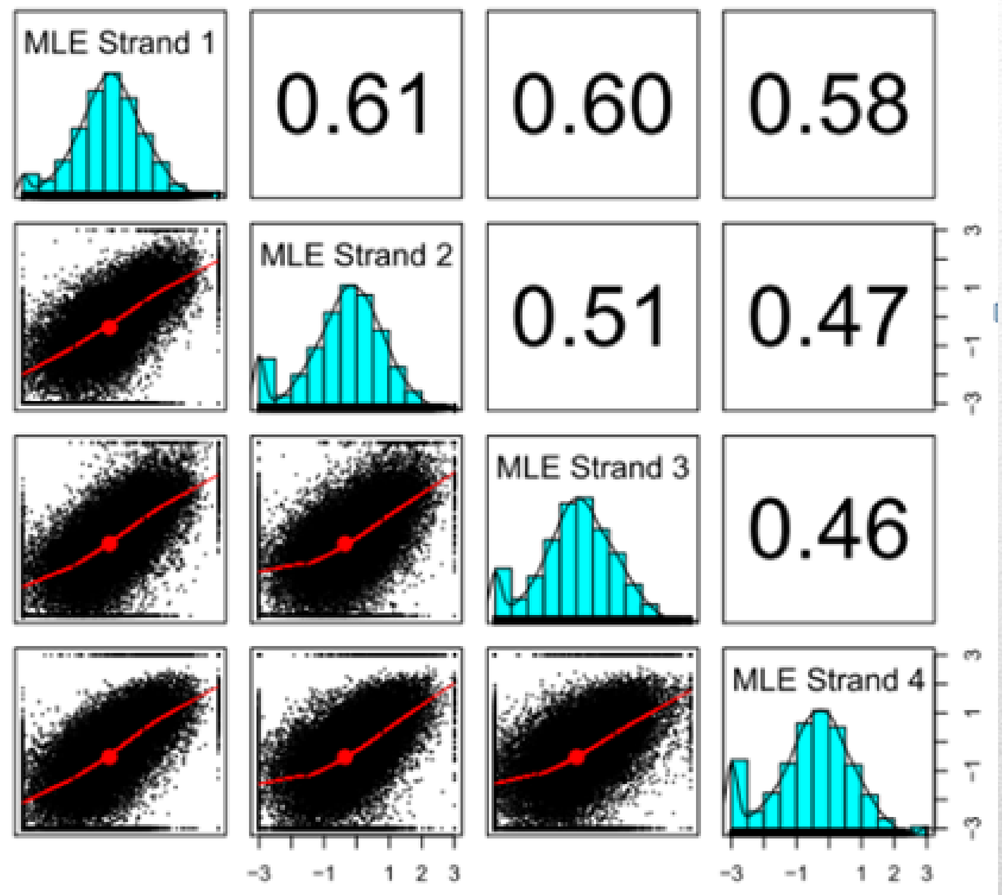
Opp1/Opp2 Scatter Plots


S1EAP.Opp1
0.79
S1EAP.Opp2
S1.SSEAP.Opp1
0.78
S1.SSEAP.Opp2
S3EAP.Opp1
0.73
S3EAP.Opp2
S2EAP.Opp1
0.75
S2EAP.Opp2
S2.SSEAP.Opp1
0.73
S2.SSEAP.Opp2
S3.SSEAP.Opp1
0.71
S3.SSEAP.Opp2
S4EAP.Opp1
0.73
S4EAP.Opp2
S4.SSEAP.Opp1
0.71
S4.SSEAP.Opp2

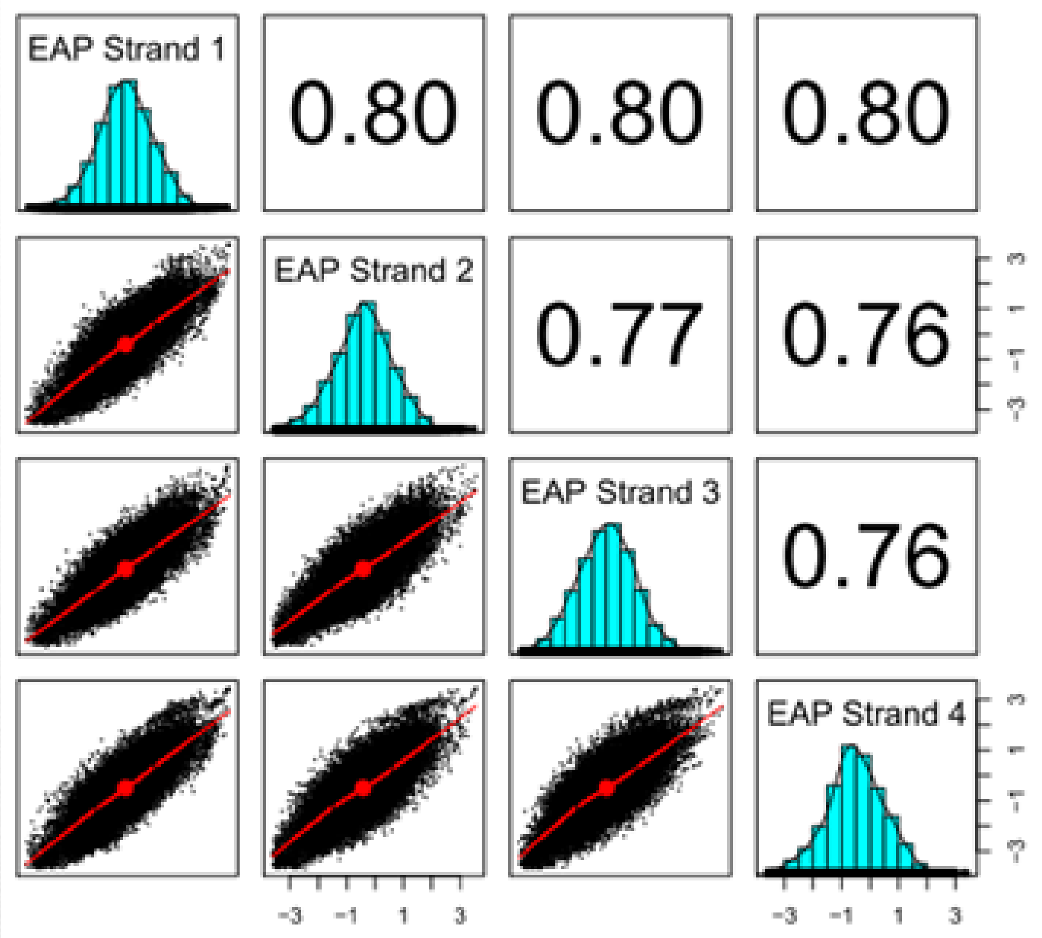
EAP-3

EAP-5

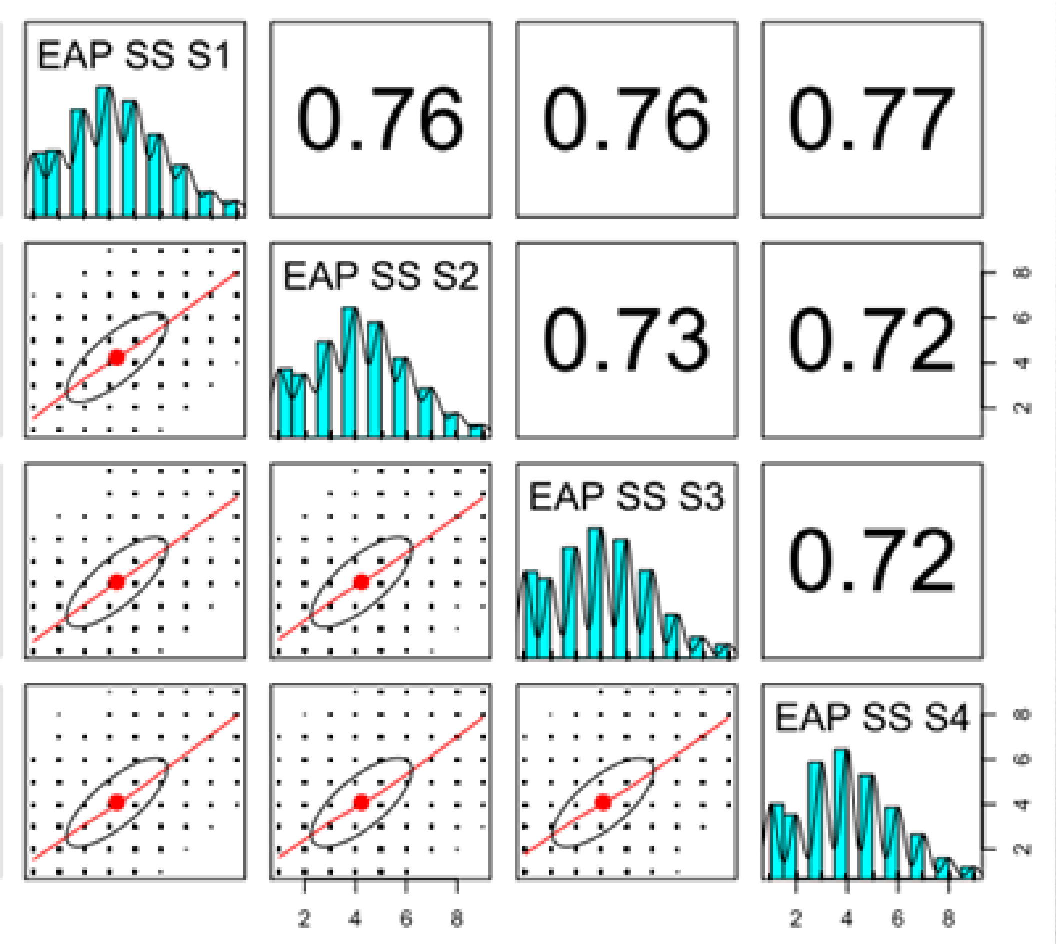

S1EAP.Opp1
0.79
S1EAP.Opp2
S1.SSEAP.Opp1
0.78
S1.SSEAP.Opp2
S3EAP.Opp1
0.72
S3EAP.Opp2
S2EAP.Opp1
0.74
S2EAP.Opp2
S2.SSEAP.Opp1
0.73
S2.SSEAP.Opp2
S3.SSEAP.Opp1
0.71
S3.SSEAP.Opp2
S4EAP.Opp1
0.72
S4EAP.Opp2
S4.SSEAP.Opp1
0.70
S4.SSEAP.Opp2



Correlation Matrix: MLE-1

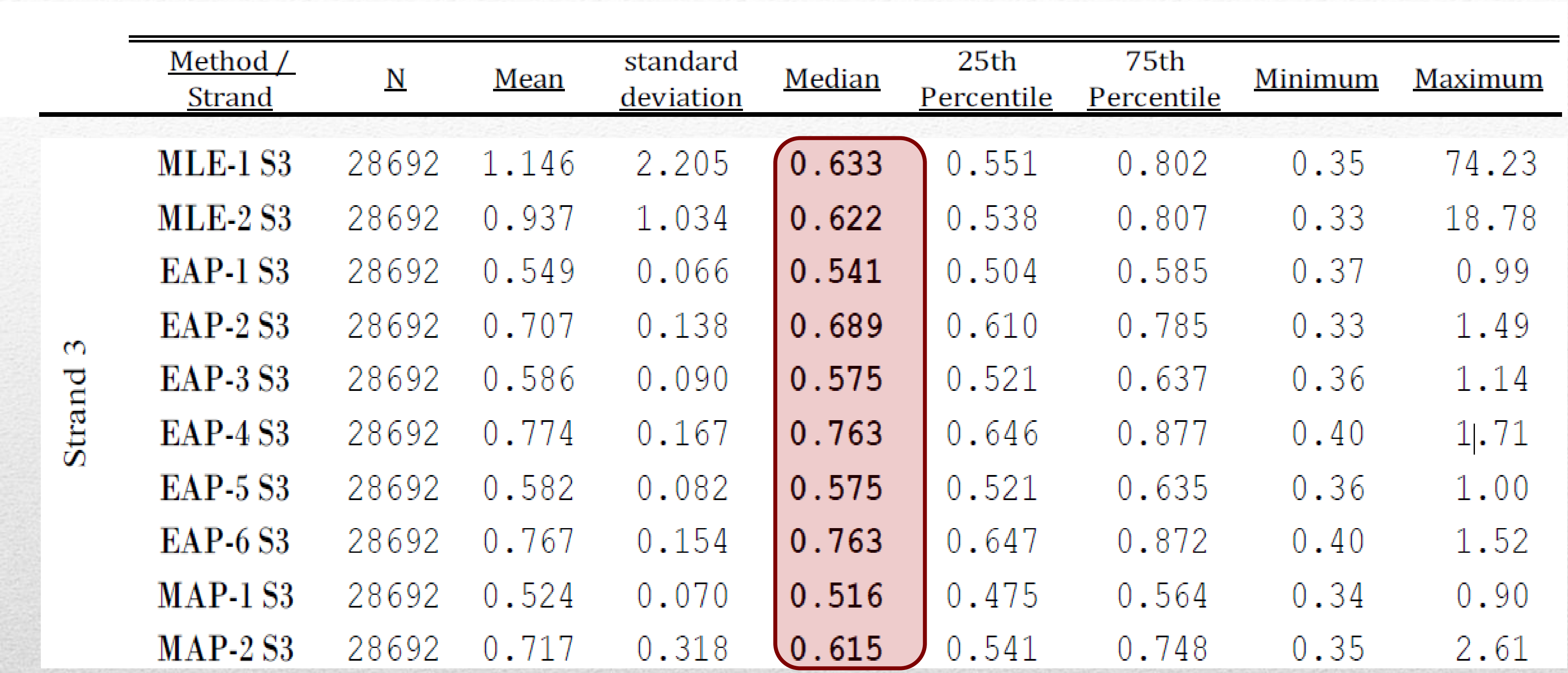


Correlation Matrix: EAP-3

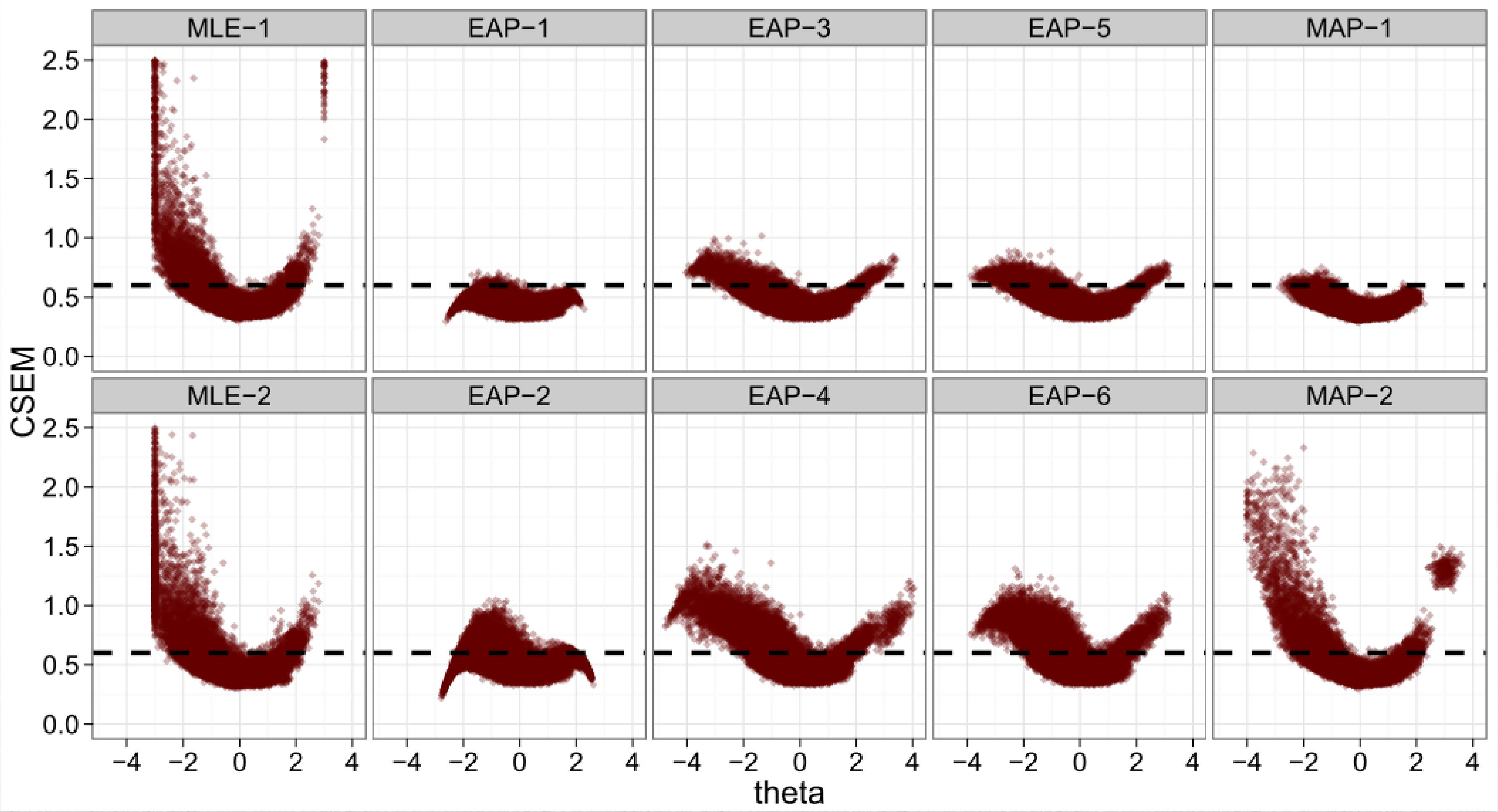


Correlation Matrix: EAP-3

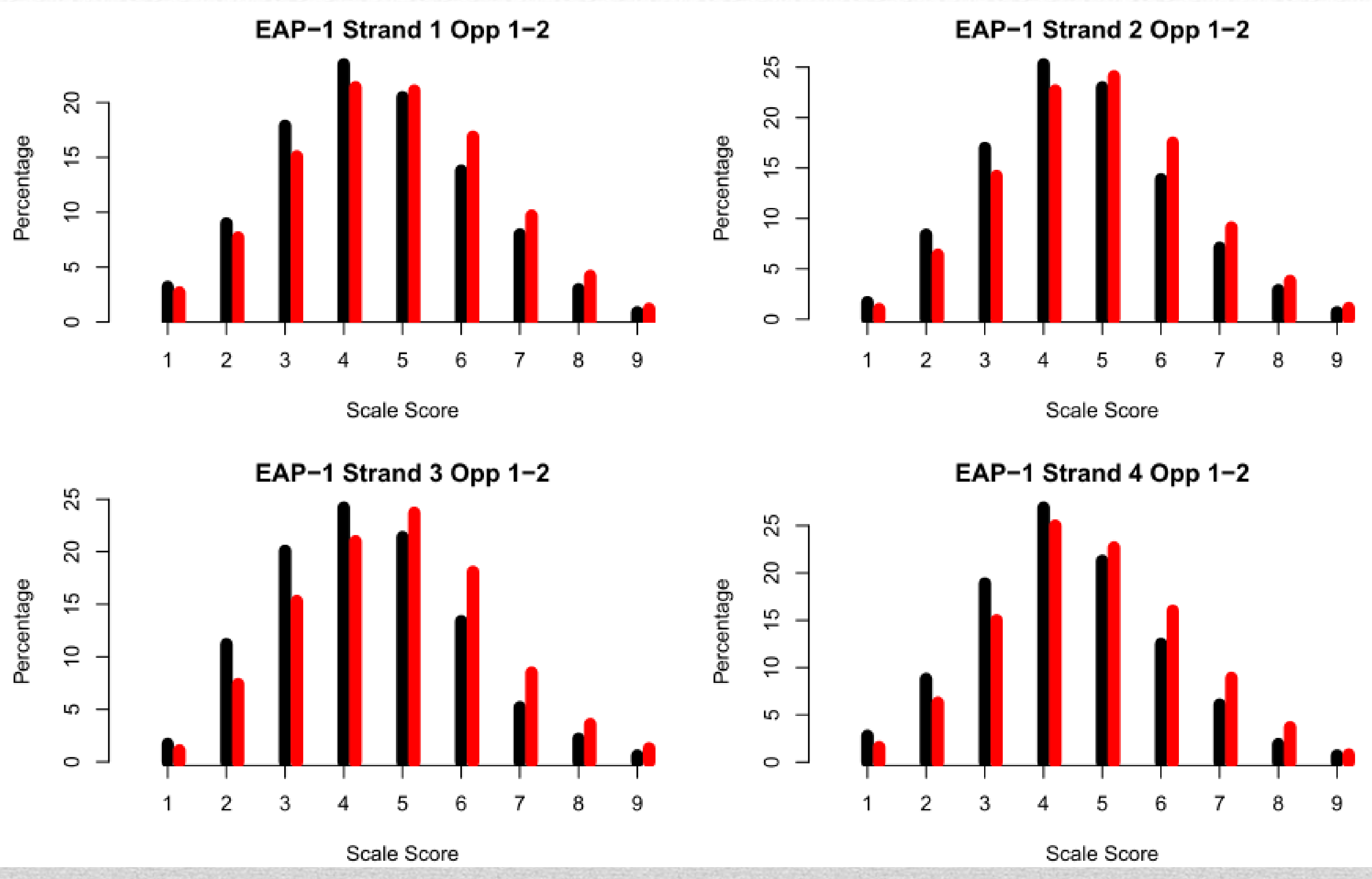
Tabular Results of CSEMs by Method



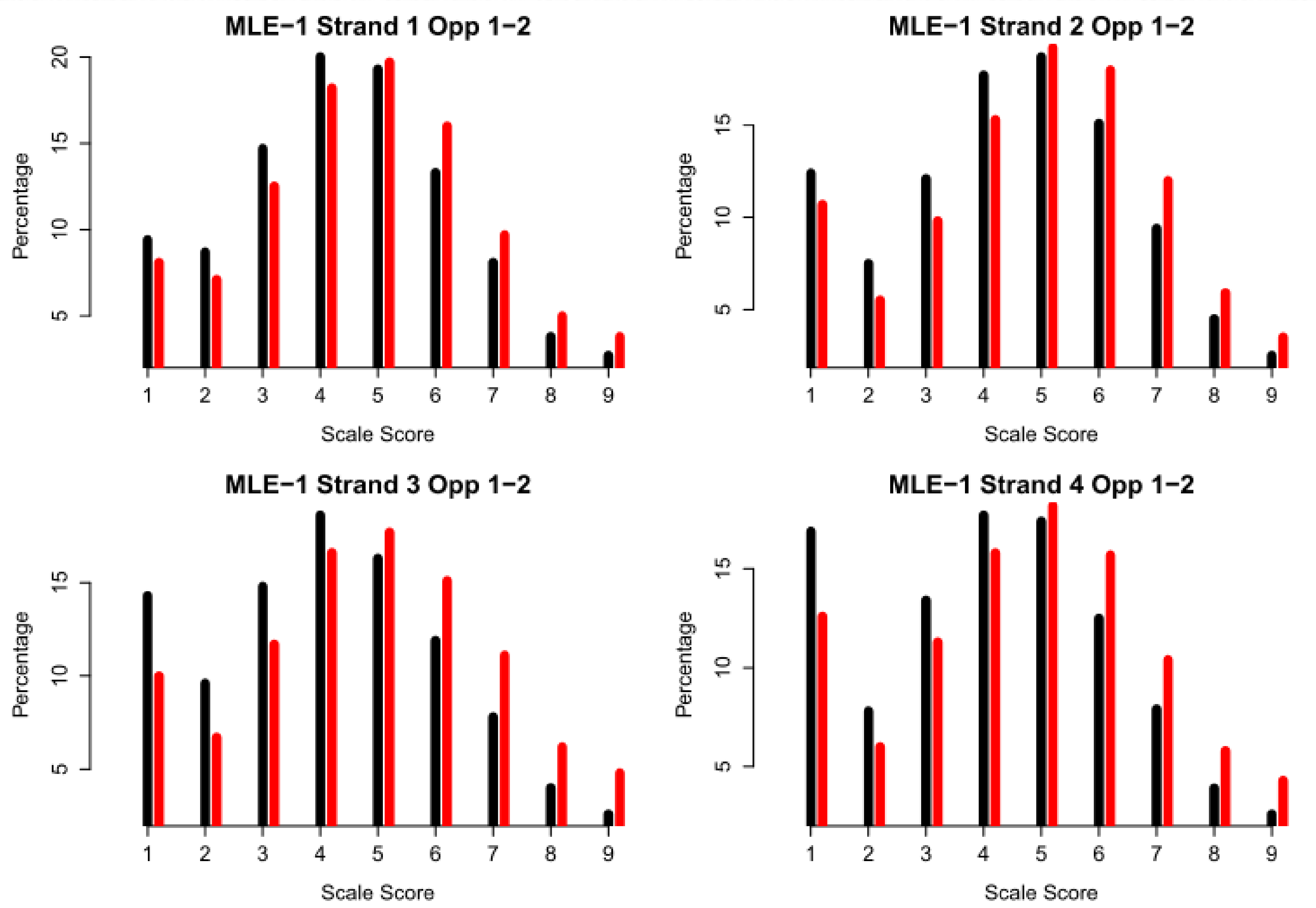
	Method / Strand	N	Mean	standard deviation	Median	25th Percentile	75th Percentile	Minimum	Maximum
Strand 3	MLE-1 S3	28692	1.146	2.205	0.633	0.551	0.802	0.35	74.23
	MLE-2 S3	28692	0.937	1.034	0.622	0.538	0.807	0.33	18.78
	EAP-1 S3	28692	0.549	0.066	0.541	0.504	0.585	0.37	0.99
	EAP-2 S3	28692	0.707	0.138	0.689	0.610	0.785	0.33	1.49
	EAP-3 S3	28692	0.586	0.090	0.575	0.521	0.637	0.36	1.14
	EAP-4 S3	28692	0.774	0.167	0.763	0.646	0.877	0.40	1.71
	EAP-5 S3	28692	0.582	0.082	0.575	0.521	0.635	0.36	1.00
	EAP-6 S3	28692	0.767	0.154	0.763	0.647	0.872	0.40	1.52
	MAP-1 S3	28692	0.524	0.070	0.516	0.475	0.564	0.34	0.90
	MAP-2 S3	28692	0.717	0.318	0.615	0.541	0.748	0.35	2.61



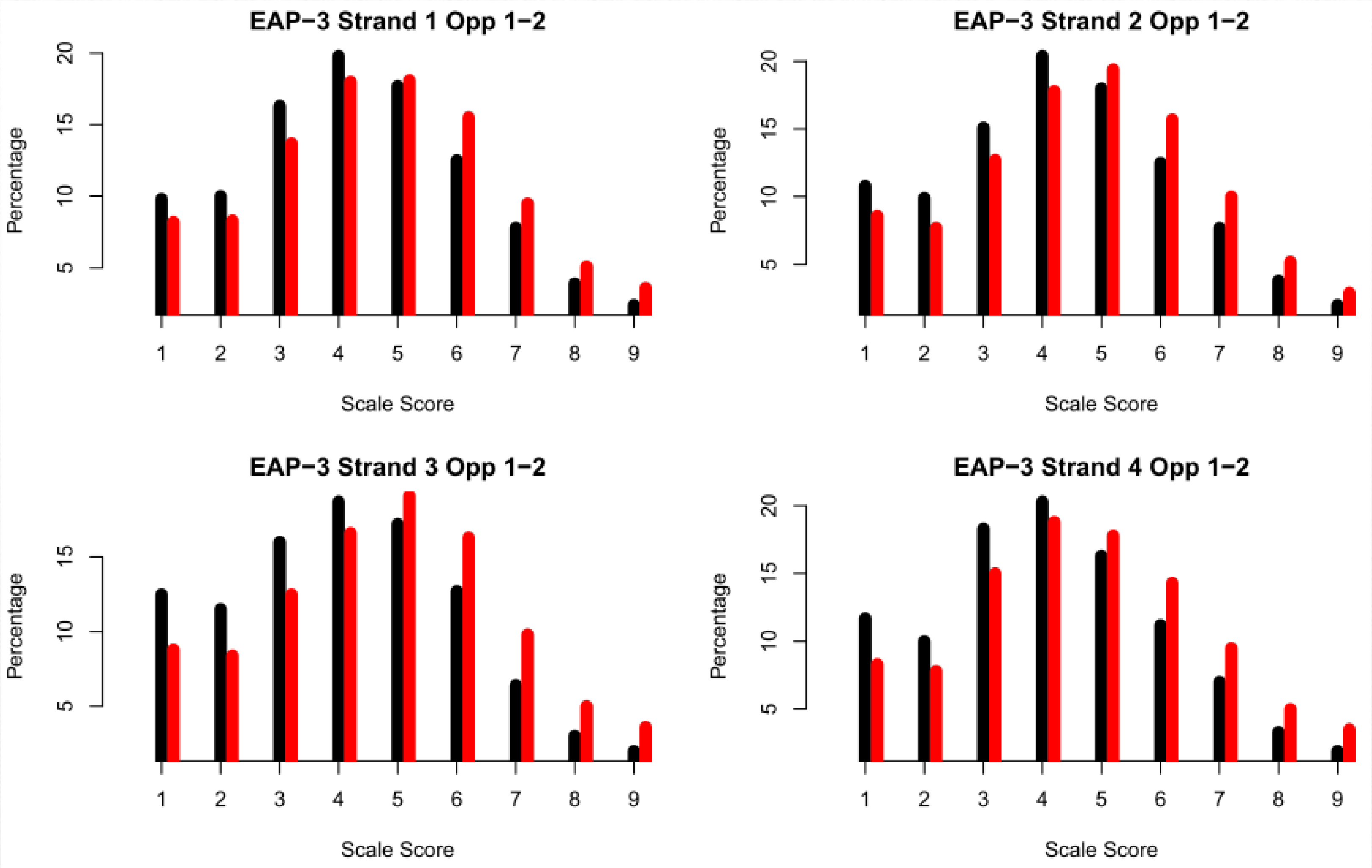
Plot of Strand 1 CSEMs as a Function of Ability across each Method



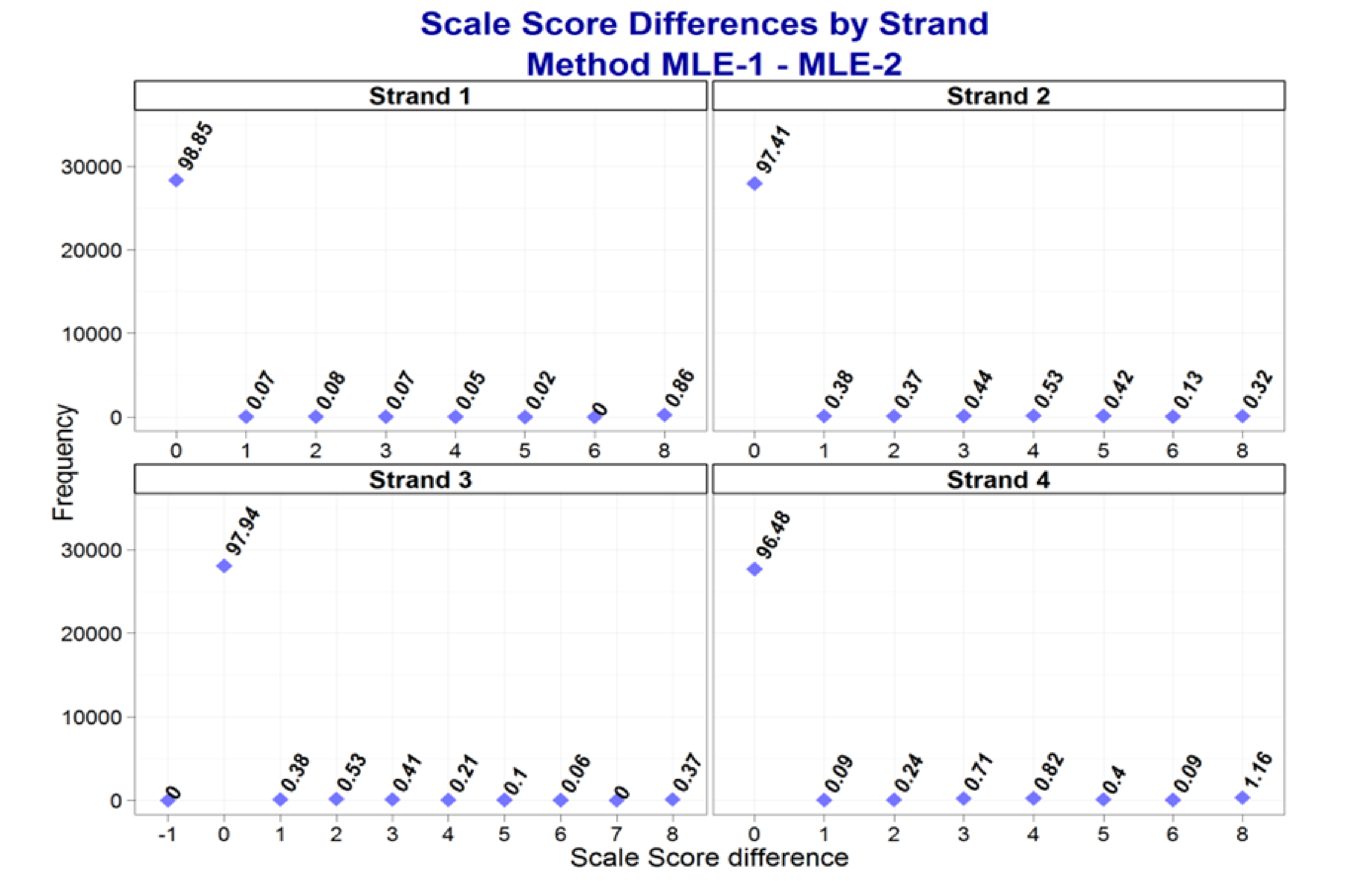
Histogram of EAP-1 Opp1/Opp2 Scale Scores



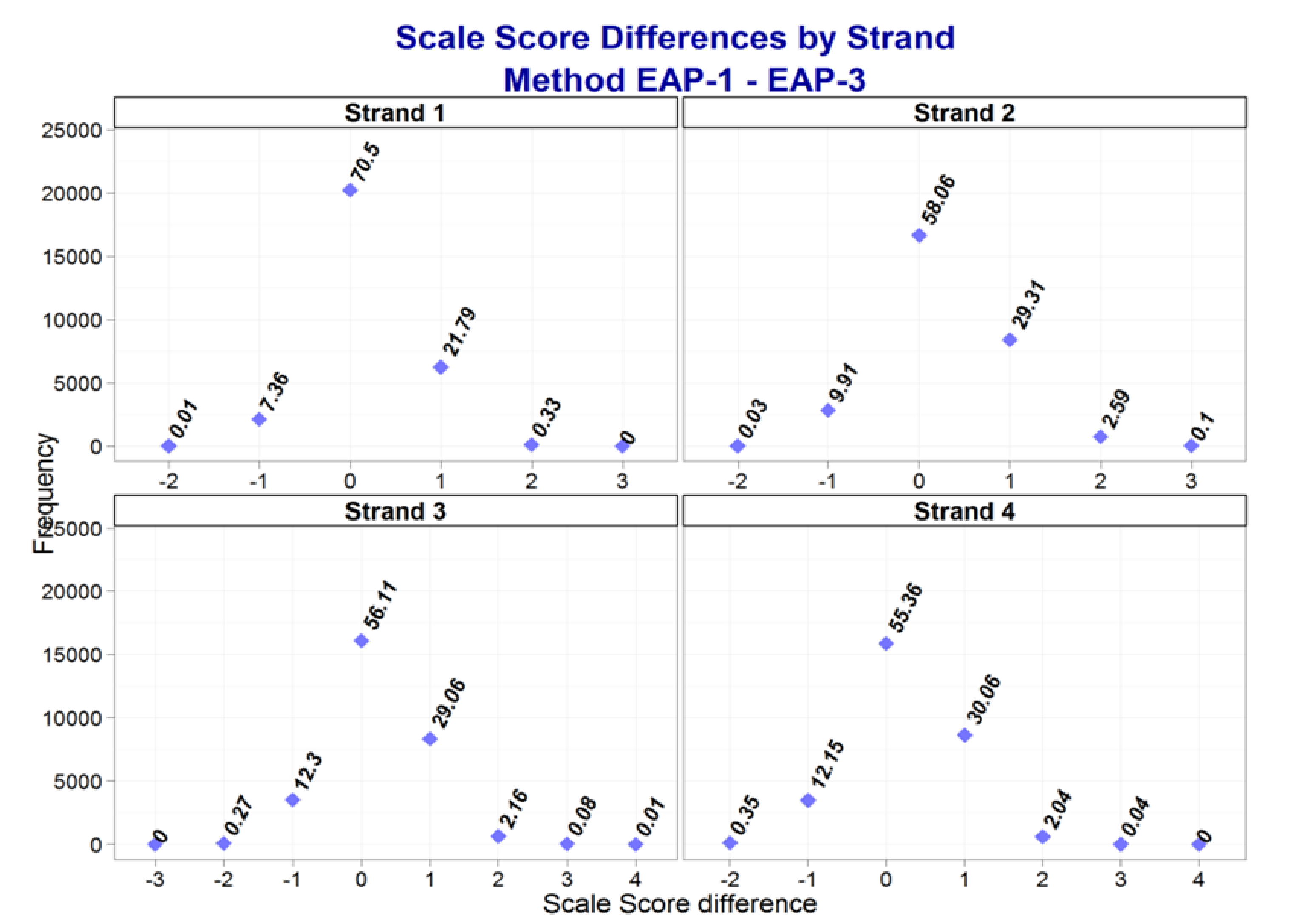
Histogram of MLE-1 Opp1/Opp2 Scale Scores



Histogram of EAP-3 Opp1/Opp2 Scale Scores



Distributional difference of scale scores for MLE-1 and MLE-2



Distributional difference of scale scores for EAP-1 and EAP-3

Summary: 3PL Domain Score Estimation

- EAP with MLE total score used as prior an informed prior was easily the most reliable under the replication perspective
- MLE strongly biased outward
- standard EAP N(0,1) biased inward
- EAP / MAP underestimates error (CSEMs)
- MAP with vague prior nearly same as MLE
- EAP with vague prior bad choice