5-2010

A comparison of limited-information and full-information methods in Mplus for estimating IRT parameters for non-normal populations

Christine E. DeMars
James Madison University, demarsce@jmu.edu

RUNNING HEAD: Limited- and Full-Information

A Comparison of Limited-Information and Full-Information Methods in M*plus* for Estimating IRT Parameters for Non-normal Populations

Christine E. DeMars

James Madison University

Paper to be presented at the annual meeting of the National Council on Measurement in Education, Denver (May 2010).

Author Note

Christine E. DeMars, Center for Assessment and Research Studies, James Madison University.

Correspondence concerning this manuscript should be addressed to Christine DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg VA 22807.

Abstract

In structural equation modeling software, either limited-information (bivariate proportions) or full-information item parameter estimation routines could be used for the 2PL IRT model. Limited-information methods assume the continuous variable underlying an item response is normally distributed. For skewed and platykurtic latent variable distributions, three methods were compared in M*plus*: limited-information, full-information integrating over a normal distribution, and full-information integrating over the known underlying distribution. For the most discriminating easy or difficult items, limited-information estimates of both parameters were considerably biased. Full-information estimates obtained by integrating over a normal distribution were somewhat biased. Full-information estimates obtained by integrating over the true latent distribution were essentially unbiased. For the *a*-parameters, standard errors were larger for the limited-information estimates when the bias was positive but smaller when the bias was negative. For the *b*-parameters, standard errors were generally similar for the limited- and full-information estimates. Sample size did not substantially impact the differences between the estimation methods; limited-information did not gain an advantage for smaller samples.

keywords: item parameter estimation, limited-information, full-information

A Comparison of Limited-Information and Full-Information Methods in M*plus* for Estimating Item Parameters for Non-normal Populations

Most item response theory (IRT) software uses full-information methods which operate on the entire response string. However, in the structural equation modeling (SEM) field, there is a long history of estimating 2PL models (or the equivalent in the common factor metric) for dichotomous responses from limited (pairwise) information, based on the univariate and bivariate proportions. Assuming the continuous response underlying the dichotomous response is normally distributed, these proportions can be used to calculate the item thresholds and the tetrachoric correlation matrix, which can then be used, along with a weight matrix consisting of the asymptotic variances (and possibly covariances) of the thresholds/correlations, to estimate the item slopes (loadings) and other parameters in the structural model (Jöreskog, 1990; Muthén, 1984). M*plus* 5.21 (Muthén & Muthén, 2009) allows the user to choose between this limited-information approach and the full-information approach more often used in IRT. LISREL has historically only included limited-information but recently introduced full-information estimation in LISREL 8.8 (Jöreskog & Sörbom, 2007). In LISREL, full-information is available only for exploratory factor analysis and can not be used with a full structural model. Thus, limited-information methods are still in use. Flora and Curran (2004) found limited-information analysis of ordinal items to be reasonably robust to violations of normality in the underlying variable, but they did not compare the results to full-information approaches nor did they vary the item discrimination or report the accuracy of the threshold estimates. Boulet (1996) found limited-information item parameter estimates were more biased than full-information estimates when the underlying distribution was skewed, but using software more familiar to IRT modelers than to SEM modelers. Finch (2010) found both limited-information and full-information item parameter estimates were more biased and had larger standard errors when the underlying

distribution was skewed. The objective of this study is to assess how violations of the normality assumption impact the item parameter estimates derived from limited-information methods and full-information methods in one popular SEM package, M*plus*.

2-Parameter Logistic Model

The 2-parameter logistic (2PL) model is often used to describe dichotomous responses to test items or surveys.

$$P(x = 1 \mid \theta, a, b) = \frac{e^{1.7a(\theta-b)}}{1 + e^{1.7a(\theta-b)}} \quad (1)$$

where P(x = 1) is the probability that the observed response, x, is correct (or in agreement), θ is the ability or trait measured by the test or survey, *a* is the item discrimination, and *b* is the item difficulty. The 1.7 puts the *a* parameter on the same metric as a normal ogive metric; if omitted, the *b* and θ parameters would be unchanged and the *a* parameters would be multiplied by a factor of 1.7. The parameters of the 2PL model can be estimated by either limited or full-information procedures.

Limited-information Estimation

Dichotomous responses to test items or surveys may be conceptualized in terms of a continuous latent variable, y*. When y* is above a threshold, the response is 1 (correct or agree), and otherwise the response is 0 (incorrect or disagree). y* is a function of another latent variable, symbolized θ in IRT, the ability or trait measured by the test or survey.

$$y^* \propto 1.7a\theta + r, \quad (2)$$

where *a* is the item discrimination from Equation 1 and r is an error or residual term. Because y* is a latent variable, the residuals are not measurement errors but instead are the part of y* not explained by θ. In other words, the residual for item *i* is the variance unique to item *i*. The higher the item discrimination, the greater the proportion of y* attributable to θ. Typically, r is assumed

to be either normally or logistically distributed. If data follow the 2PL model, r is logistically distributed. Generally, y* is standardized to have a variance of 1. The standardized coefficient λ then replaces the 1.7*a*, where $\lambda = \frac{a}{\sqrt{1 - a^2}}$.

Weighted least squares (WLS) is one of the most common limited-information (LI) methods for estimating the parameters of the 2PL model. A clear, simple explanation of WLS is provided in Flora and Curran (2004) and it should be consulted by readers interested in more details and or in more advanced references. Briefly, WLS typically involves several stages. First, the standardized threshold, τ, is estimated for each item by finding the inverse of the standard normal distribution corresponding to the percent correct for the item (if there are covariates in the model, this stage is more complicated). Then the tetrachoric correlations between every pair of items are estimated. The standardized loadings, λ, are then estimated from the tetrachoric correlation matrix, in a generalized least squares procedure with a weight matrix composed of the asymptotic error variances and covariances of the tetrachoric correlations. Sometimes the error variances and covariances of the thresholds are included in the weight matrix, as in M*plus*. A simpler alternative is diagonally weighted least squares (DWLS), where only the diagonal elements, the error variances, are used in the weight matrix. In M*plus*, this is labeled weighted least squares mean and variance adjusted (WLSMV) because of the adjustments made to the fit indices (Muthén, 1993). In the M*plus* WLSMV, the off-diagonals of the weight matrix are used for computing the standard errors of the parameter estimates, but not for estimating the parameters themselves (Finney & DiStefano, 2006; Flora & Curran, 2004). τ and λ are standardized coefficients. They can be transformed to the unstandardized coefficients in Equations 1 and 2 through the following relationships: $a = \frac{\lambda}{\sqrt{1 + \lambda^2}}$ and $b = \frac{-\tau}{\sqrt{1 + \lambda^2}}$.

Other LI methods also exist. Knol and Berger (1991) describe several additional factor analytic procedures that utilize the tetrachoric correlation matrix: unweighted least squares (ULS), adjusted minimum residuals (MINRES), maximum likelihood (ML), and iterative principal factor analysis. Another LI approach is McDonald's (1997) harmonic analysis based on the bivariate item proportions, which, like the various methods that utilize tetrachoric correlations, assumes an underlying latent distribution. The details of these other LI procedures can be found in the cited references.

Clearly, the LI approach was intended for use with normally-distributed y*. This implies normally distributed θ as well as normally distributed residuals. Muthén (1984; 1993) expressed concerns that the assumption of underlying normality may not always be appropriate and presented a method for testing trivariate normality (Muthén, 1993; Muthén and Hofacker, 1988). Jöreskog (1990) also discussed the assumption of normality of the underlying variables. But in practical applications with substantive research questions, the analysts may simply proceed without testing for normality in the latent trait, perhaps believing that most psychological and cognitive traits are normally distributed or that it does not make much difference in parameter estimation.

Full-information

Currently, maximum marginal likelihood (MML; known alternatively as marginal maximum likelihood) is one of the most popular full-information (FI) methods. In MML, the θ distribution is approximated on a series of quadrature points. Each time the item parameters are updated, the likelihood of each observed response pattern is calculated at each quadrature point, and then averaged (marginalized) over the θ distribution. The goal of the estimation is to find the item parameters which maximize the marginal likelihood of the observed response patterns. See Bock and Atkin (1981) or Baker (1992, Chapters 6 and 7) for further details.

In MML estimation, either a normal distribution can be assumed for θ or the distribution can be estimated empirically. In BILOG (Zimowski, Muraki, Mislevy, & Bock, 2003), for example, the distribution may be estimated as a histogram, with the rectangles centered on a set of discrete quadrature points. In M*plus*, the distribution might be estimated as a mixture of normal distributions, though this is considerably more complex then the histogram approach. Woods and Lin (2009) detailed other methods of estimating the distribution, and showed that erroneously assuming a normal distribution of θ yielded less accurate item parameter estimates. Other work has also shown that in MML estimation, integrating over a normal distribution when the true distribution is skewed produces at least slightly greater RMSE (Seong, 1990; Stone, 1992; Woods, 2006; Zwinderman & van den Wollenberg, 1990).

Comparing Full-information and Limited-information

Boulet (1996) compared FI MML with McDonald's LI method with normal and skewed θ distributions. The FI method involved integrating over a normal θ distribution, regardless of the true θ distribution. When data were generated for a normal θ distribution, the LI estimates had somewhat less bias and smaller RMSE. The difference between estimation methods was more pronounced for the discrimination estimates than for the difficulty estimates. When θ was skewed, the LI parameter estimates for the most discriminating, easy or difficult items were far more likely to fall into an extreme range (a estimate > 4.5 when the true value was 1.5, or $|b|$ > 4.5 when the true value was 2), particularly for small samples. For the discriminating, easy items, LI estimates were also more biased and had larger RMSE.

Finch (2010) compared FI MML in BILOG with LI in NOHARM and M*plus*. He examined both normal and nonnormal θ distributions. The models were two-dimensional with simple structure, so the LI models took the other factor into account in the estimation, but the FI model estimated each the item parameters for each factor separately, which may affect the

comparison. Additionally, in some conditions the data followed a 3PL model, which was correctly specified in BILOG and NOHARM but necessarily was mis-specified as a 2PL model in M*plus*, blurring estimation differences with model specification differences whenever results were averaged over the 2PL and 3PL data. The 2PL data with no correlation between the factors is most relevant to the current study. In this condition, the *a*-parameters had a small negative bias, which was larger for NOHARM than the other two procedures. SEs for the *a*-parameters were smallest for FI and largest for M*plus* LI; when the θ distribution was skewed, the SEs were larger and the differences among the methods were greater. For the *b*-parameters, results were separated only by main effects so the 2PL no-corrrelation condition could not be described separately. For the skewed θ distribution overall, SEs were similar across methods, the NOHARM *b*-parameters were negatively biased and the M*plus* LI and BILOG *b*-parameters were positively biased. Finch averaged over all items, so there were no separate results for the items with the most extreme *a* or *b*-parameters, and on average there would not have been many extreme true parameters because the parameters were drawn from normal distributions.

A number of other studies have compared limited and FI estimation, but only with normally distributed θ. Reiser and VandenBerg (1994) found that WLS LI parameter estimates were less accurate than FI parameter estimates with smaller samples (N =500), though the difference decreased with sample size. LI discrimination parameters were overestimated and difficulty parameters were biased outward somewhat. Forero and Maydeu-Olivares (2009) compared ULS LI to FI under a number of conditions, including variation in the skew of the observed variables, but the underlying trait was apparently normally distributed in every condition. In most conditions, there was little difference in the bias and RMSE of the limited and FI item parameter estimates. LI yielded slightly more accurate parameter estimates, but FI yielded more accurate standard errors. Finger (2001) compared ULS LI and McDonald's method

to FI. Generally, the RMSE of the *a*-parameters was similar for the three methods, but the RMSE of the *b*-parameters was higher for FI when the a-parameters were higher and more disperse. FI tended to underestimate the *a*-parameters for the most discriminating items. McDonald's method had the highest correlations between estimated and true *a*-parameters, though the three methods had more similar correlations for the *b*-parameters. For the overall item response function, there was little difference in RMSE for lower-discriminating items, but for more discriminating items McDonald's method, followed closely by ULS, had the lowest RMSE. Muraki and Engelhard (1985) compared FI estimation to a MINRES factor analysis of the tetrachoric correlations. For items of moderate difficulty, there was little difference between the estimates, but for the most extreme items, the FI estimates had less bias and smaller standard errors. Knol and Berger (1991) compared FI estimation to several different LI procedures. Compared to FI, some of the LI methods were as accurate or nearly as accurate with one and two dimensions, and more accurate with three dimensions. Again, all of these studies utilized normally distributed θ.

In summary, nearly all of the previous simulation studies, with the exception of Boulet (1996) and Finch (2010), have used normally distributed θ. These studies have found LI methods provide estimates that are similar to, or sometime more accurate than, FI. The two studies that included non-normal θ found FI estimates were more accurate, particularly for the *a*-parameters. Because LI methods are more commonly used in an SEM framework, this study will further explore LI and FI estimates from one SEM software package, M*plus*. By focusing on a single software package, other differences between the methods should be minimized. Other software is explored in Appendices A and B.

Method

Data Simulation

Dichotomous item responses were simulated to follow a 2PL IRT model. The underlying continuous response, y*, thus was a function of θ and error (item-specific variance). The errors were logistically distributed, which is a minor departure from the normal distribution. However, the distribution of θ was skewed negative, skewed positive, or was platykurtic. The negatively skewed and platykurtic distributions are shown in Figure 1; the positively skewed distribution was a mirror-image of the negatively skewed distribution. The mean of each θ distribution was zero and the standard deviation was one. For each θ distribution, 1000 samples of 5000 or 300 examinees were drawn. The sample size of 300 was used to show results with a sample smaller than generally recommended for the 2PL model but not unrealistically small. LI estimation might work better than FI for the smaller sample.

Responses to 45 items were simulated. Three levels of discrimination (a = 0.3, 0.8, and 1.3) were crossed with 15 levels of item difficulty evenly spaced from -2 to 2. These values were chosen to systematically illustrate the effects of discrimination and difficulty. Because the contribution of θ to y* is proportional to the item discrimination, the distribution of y* was less normal for higher values of discrimination.

Parameter Recovery

Item parameters were estimated in M*plus* 5.2 (Muthén & Muthén, 2009) using both limited and FI methods. The WLSMV (mean and variance adjusted weighed least-squares) estimator was used for the LI estimates. This is a robust WLS method, also known as diagonally-weighted least squares (DWLS), in which only the variances, not the covariances, of the tetrachoric correlations and intercepts are used in the weight matrix for the parameter estimation. The covariances of the correlations are used only in estimating the standard errors, not in

parameter estimation. The solution was estimated in the default standardized parameter metric and then transformed to the IRT metric. Forero and Maydeu-Olivares (2009) explained that it should not matter whether the minimization was with respect to standardized or unstandardized parameters, but their results showed that, for a few replications, the estimates in the unstandardized metric were far less accurate, leading to higher standard errors. Although only WLS is presented in the main study, an alternative LI method is presented in Appendix A.

The M*plus* MLR estimator with the EM algorithm and numerical integration was used for the FI estimates. This procedure is often called marginal maximum likelihood or maximum marginal likelihood (MML) in the IRT literature because it estimates the item parameters that maximize the marginal likelihood of the observed data. In M*plus*, the likelihood of the data is integrated over a normal θ distribution before solving for the item parameters (the spacing of the quadrature points can be rectangular or Gauss-Hermite or randomly selected Monte Carlo points, but the weights are the densities for a normal distribution). The inappropriate use of a normal distribution was not expected to influence the M*plus* FI estimates nearly as much as the M*plus* LI estimates, but it could have some impact. Thus, in a third condition, the θ distribution was also specified as a known mixture of normal distributions which yielded the skewed/platykurtic distribution used to generate the data—this was not intended to be realistic but instead to be a best case scenario. These estimates will be referred to as the FI-Nonnormal estimates. Estimating the θ distribution would be more realistic than using the known θ distribution; to keep the results simple, estimating the θ distribution was relegated to Appendix B.

Accuracy was assessed by bias and standard error (standard deviation of estimates around the mean estimate). Empirical standard errors were also compared to theoretical standard errors. The metric was set for a group with mean $\theta = 0$ and standard deviation = 1. The replication samples used for the calibration differed randomly from these values, so before computing the

bias and standard deviation, the replication sample mean θ was subtracted from each *b*-parameter and the difference was divided by the replication sample standard deviation. Similarly, each *a*-parameter was multiplied by the standard deviation of θ in that replication sample. This had no effect on the bias as the scaling constants averaged 0 and 1 over the 1000 replications, but it slightly reduced the standard deviation of the item parameters across replications.

Results

The results are presented in a series of figures. Within each figure, from left to right, there are three sets of points. The leftmost is for the least discriminating items and the rightmost is for the most discriminating items. Within each set, item difficulty increases from left to right.

Figures 2 and 3 show the bias in the *b*-parameters for the negatively skewed data for the larger sample and small sample, respectively. The LI estimates were positively biased for either easy or difficult items, and negatively biased for middle-difficulty items. An exception to this pattern occurred for the small sample, least discriminating, easy items, where the bias was negative instead of positive. The severity of the bias increased in absolute value as the *a*-parameter increased. The FI estimates assuming a normal θ distribution followed the same pattern, but to a lesser degree. For example, in the large sample with $a = 1.3$, the LI estimate of the difficulty parameter for the easiest item would average approximately -1.65 (bias near 0.35), but the FI assuming normal θ would average approximately -1.85 (bias near 0.15). The FI estimates based on the correct skewed distribution were virtually unbiased in the large sample and had only a small bias in the small sample.

Figures 4 and 5 show the bias in the *a*-parameters for the negatively skewed data. There was very little bias for the least discriminating items, regardless of difficulty or estimation method. For the two higher discrimination levels, the LI estimates were positively biased for easy items and negatively biased for difficult items. The FI estimates assuming a normal θ

distribution followed the same pattern, but to a lesser degree. For both the LI and FI-normal estimates, the bias was sometimes greater with the smaller sample size. The FI estimates based on the correct skewed distribution were again virtually unbiased for the sample size of 5000, but slightly positively biased for the sample size of 300. Another difference between the sample sizes occurred for the easiest items with $a = 1.3$; for N = 5000, the easiest items were the most positively biased, but for N = 300, LI estimates for the two easiest items were less biased than the next two items.

The bias for the positively skewed θ were a mirror image (reflected both right/left and up/down) of the results for the negatively skewed data and are not shown.

For the platykurtic θ distribution (Figures 6-9), bias was minimal for the FI methods, with the exception of the *b*-parameter bias for the small sample with low *a*-parameters. For the LI method, bias was much lower in absolute value than it was for the skewed distributions. For the easiest and most difficult items, the LI *b*-parameters were biased away from the mean (with the exception noted for the *b*-parameter bias for the small sample with low *a*-parameters) and the high and moderate *a*-parameters were negatively biased. These effects again increased with increasing true discrimination.

Standard errors were another criterion for accuracy. Although the LI estimates were biased for the easiest and most difficult items, small standard errors would indicate the bias was consistent across replications. Conversely, unbiased estimates are not very useful for any one sample if they have large standard errors. The empirical standard errors are shown in Figures 10-17 for the negatively skewed and platykurtic distributions; values for the positively skewed distribution are not shown because they were a right/left mirror image of the values for the negatively skewed distribution. Standard errors were of course much smaller for the larger sample size and for the middle difficulty items. The standard errors of the *b*-parameters were

quite similar for the LI and FI estimation methods; they were slightly larger for the LI method only when the *b*-parameter estimates were most biased. For the *a*-parameters, there were greater differences between the LI and FI methods, particularly for the small sample and moderate or large discriminations. For these conditions, for both skewed and platykurtotic distributions, when the bias in the *a*-parameter was positive, the standard error was larger for the LI estimates than for the FI estimates. But when the bias in the *a*-parameter was negative, the standard error was smaller for the LI estimates than for the FI estimates. When LI underestimated the *a*-parameters, the underestimates were consistent. Similarly, again for the small samples and moderate or large *a*-parameters, the standard errors for the FI estimates integrating over a normal distribution were somewhat larger than the standard errors for the FI estimates integrating over the true θ distribution when the bias in the *a*-parameter was positive, and conversely when the bias in the *a*-parameter was negative.

The theoretical SEs supplied in the output for the item parameters were also compared to the empirical SEs summarized above. When analyzing real data, these analytically-derived standard errors are generally used because there is only one sample, unless standard errors are estimated by bootstrap methods. Tables 1 and 2 show the average ratio of the theoretical SE to the empirical SE. The ratio was calculated for each item, then averaged over the 15 items with the same discrimination because item difficulty had little effect on the ratio.

For the *b*-parameters (Table 1), when the *a*-parameter was low and the sample size was large, the theoretical SEs from M*plus* were accurate for all estimation methods. For the large sample size, as *a* increased, the theoretical SEs were underestimated for all methods. For the smaller sample size, the SEs of *b* were somewhat overestimated when the a-parameter was low, but were more accurate for moderate or high *a*-parameters. Again, ratios were similar for all estimation methods.

For the *a*-parameter (Table 2), when the *a*-parameter was low, the theoretical SEs from M*plus* were accurate for all estimation methods and both sample sizes. For the large sample size, as *a* increased, the theoretical SEs were somewhat underestimated, especially using the LI estimates. For the smaller sample size, only the LI standard errors were increasingly underestimated with increasing *a*-parameter. Item difficulty had an effect only in the large sample condition and only for the LI method; the SEs were most underestimated for the easiest and most difficult items.

Limitations

Findings are limited to the 2PL model. The data were generated to follow a 2PL model because a 3PL model can not be specified in M*plus*. With unidimensional data, when a 2PL model is estimated for 3PL data, item difficulties and discriminations are underestimated (Yen, 1981). The underestimation of the discrimination parameters is particularly evident for more difficult items, because the estimated slope flattens out to better capture the responses in the range impacted by the non-zero *c*-parameter. This negative bias in the *a*-parameter extends to multidimensional models (Finch, 2010) but Finch found a positive bias in the *b*-parameter for multidimensional models when a 2PL model was mis-specified for 3PL data. To avoid confounding mis-specification of the θ distribution with mis-specification of the functional form of the model, only 2PL data were included in the present study.

For simplicity, only a unidimensional context was explored. Finch (2010) showed that the correlation between the factors impacts parameter recovery, so these unidimensional results should be extended to multidimensional data with caution. The general principle that violations of normality can bias the LI item parameter estimates for the most extreme items likely generalizes because it follows from the theory behind LI estimation. What is more difficult to know is whether FI would empirically continue to perform better than LI for multidimensional

data. The theoretical advantages of FI might break down empirically as the model becomes more complex.

Discussion and Conclusion

For well-discriminating easy or difficult items, LI analysis yielded biased estimates of both difficulty and discrimination when the θ distribution was not normal. The bias in discrimination for the easiest and most difficult items was far greater than the average bias Flora and Curran (2004) found in loadings for skewed distributions, but they did not report the estimates by item difficulty. For positively skewed θ, Boulet (1996) also found that the *b*-parameters for easy items were negatively biased, the *a*-parameters for easy items were negatively biased, and the *a*-parameters for difficult items were positively biased. This pattern was similar to the pattern in the current study for positively skewed θ, except that difficult items also had bias in the *b*-parameters in the current study but little bias in Boulet's study.

Bias was lower for the platykurtic distributions than for the skewed distributions. This might be because the selected playkurtic distribution departed less from the normal distribution, or because the differences from the normal were balanced on either side of the mean.

Because the bias was greatest for the most extreme items, LI estimates would be particularly problematic if test developers planned a test to be easy (or difficult) relative to the calibration examinees. With a negatively-skewed sample, the developers would think the easiest items were more difficult than they actually were, and thus construct a test that was even easier than planned. The LI discrimination parameters for these easy items, based on the negatively-skewed sample, would also lead the developers to think the test discriminated better than it really did.

Because the bias of the *b*-parameters was higher for more discriminating items, two items with the same difficulty could have different degrees of *b*-bias within the same examinee

population. Similarly, because the bias of the *a*-parameters varied depending on the *b*-parameter, two items with the same discrimination could have different degrees of *a*-bias within the same examinee population. Two examinee groups with the same skew and variance but different means would have different bias for the same item depending on how far the item's difficulty was from the group mean. This could lead to poor estimates of DIF. Thus, the bias in the LI estimates could have practical consequences.

FI analysis integrating, incorrectly, over a normal distribution of θ produced noticeably less biased estimates, and FI analysis integrating over the non-normal distribution of θ used to generate the data reduced the bias even more. The empirical standard errors were generally no larger for the FI estimates than for the LI estimates, except when the LI *a*-parameter estimates were biased low. Thus, full-information techniques should be preferred to limited-information techniques for models with dichotomous items if there is any suspicion of non-normality in the latent distribution. Estimation of the θ distribution as a mixture of normals could potentially yield more accurate estimates than assuming a normal distribution during the FI analysis, but only if the θ distribution is estimated accurately. A known θ distribution was used in this study for the non-normal FI estimation to simulate a best-case scenario; estimation of the θ distribution should be examined in further research (a step in this direction is discussed in Appendix B). In summary, incorrectly assuming a normal distribution is more problematic for limited-information analysis than for full-information analysis.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Bock, R. D., & Atkin, M. (1981). Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika, 46*, 443-459.

Boulet, J. R. (1996). The effect of nonnormal ability distributions on IRT parameter estimation using full-information and limited-information methods. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Publication No. AAT NN15701, ProQuest Document ID 739847191)

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor-analysis based models. *Applied Psychological Measurement, 34*, 10-26.

Finger, M. S. (2001). A comparison of full-information and unweighted least-squares limited-information methods used with the 2-parameter normal Ogive model. (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3026475)

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.). *A second course in structural equation modeling* (pp. 269 - 314). Greenwich, CT: Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Forero, C. G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14*, 275-299.

Frasier, C., & McDonald, R. P. (2003). NOHARM 3. [Computer software]. Author: Niagara College, Ontario, Canada. Downloaded 1/10/2008 from http://people.niagaracollege.ca/cfraser/download/

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity, 24*, 387-404.

Jöreskog, K. G., & Sörbom, D. (2007). LISREL 8.8 [computer software]. Lincolnwood, IL: Scientific Software International.

Knol, D. L., & Berger, M. P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26*, 457-477.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 257-269). New York: Springer.

Muraki, E. & Engelhard, G. Jr. (1985). Full-information item factor analysis: Applications of EAP scores. *Applied Psychological Measurement, 9*, 417-430.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49*, 115-132.

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 205-234). Newbury Park, CA: SAGE.

Muthén, B., & Hofacker, C. (1988). Testing the assumptions underlying tetrachoric correlations. *Psychometrika, 53*, 563-578.

Muthén, L. K., & Muthén, B. O. (2009). *Mplus 5.21: Statistical analysis with latent variables* [computer software]. Author: Los Angeles, CA.

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology, 47*, 85-107.

Seong, T.-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distribution. *Applied Psychological Measurement, 14*, 299-311.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Woods, C. M. (2006). Ramsey-Curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253-270.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement, 33*, 102-117.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, D. (2003). BILOG-MG [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement, 14*, 73-81.

Table 1

Mean Ratio of Theoretical SE to Empirical SE for the b-Parameter.

		a = 0.3			*a* = 0.8			*a* = 1.3		
N	θ	LI	FI	FI-NN	LI	FI	FI-NN	LI	FI	FI-NN
5000	NS	0.96	0.97	0.97	0.85	0.86	0.85	0.77	0.80	0.78
5000	PL	0.97	0.97	0.96	0.88	0.88	0.85	0.80	0.83	0.79
5000	PS	0.97	0.97	0.97	0.84	0.85	0.85	0.77	0.79	0.78
300	NS	1.13	1.16	1.14	0.99	1.04	1.02	1.03	1.04	0.99
300	PL	1.08	1.11	1.11	1.00	1.04	1.01	1.02	1.05	0.99
300	PS	1.10	1.13	1.12	0.98	1.03	1.00	1.02	1.03	0.98

Note. LI = limited-information, FI = full-information, using a normal distribution, FI-NN = full-information, using the true non-normal distribution used to generate the data, NS = negative skew, PL = platykurtic, PS = positive skew.

Table 2

Mean Ratio of Theoretical SE to Empirical SE for the a-Parameter.

		a = 0.3			*a* = 0.8			*a* = 1.3		
N	θ	LI	FI	FI-NN	LI	FI	FI-NN	LI	FI	FI-NN
5000	NS	0.97	0.97	0.98	0.89	0.92	0.93	0.83	0.92	0.92
5000	PL	0.97	0.97	0.97	0.92	0.94	0.94	0.82	0.94	0.94
5000	PS	0.97	0.97	0.98	0.89	0.92	0.92	0.84	0.92	0.92
300	NS	0.94	1.02	1.01	0.83	1.01	1.01	0.82	0.98	0.97
300	PL	0.95	1.01	1.00	0.83	1.00	0.99	0.82	0.99	0.98
300	PS	0.94	1.01	1.01	0.82	1.00	0.99	0.82	0.98	0.97

Appendix A

Limited Information in NOHARM

NOHARM 3 (Frasier & McDonald, 2003) uses McDonald's harmonic analysis, which approximates the normal ogive function with a polynomial series. In the estimation, the difference between the observed bivariate proportions and model-based proportions is minimized by unweighted least squares (ULS). To assess this limited-information approach, the item parameters were also calibrated in NOHARM3. The model was specified as a unidimensional model, using default starting values.

When the sample size was 5000, the bias and standard errors were nearly identical to those from M*plus*' WLSMV procedure. The same was true when the sample size was 300, with the following exceptions: For the negatively-skewed θ distribution, when $a = 1.3$, the absolute value of the bias in the *a*-parameters was greater for NOHARM for the easiest items but less for NOHARM for the most difficult items (Figure A1). The standard errors of the *a*-parameters for the discriminating easy or difficult items was greater for NOHARM (Figure A2). These relationships were mirrored for the positively-skewed θ distribution. With the platykurtic θ distribution, again only for the sample size of 300, when $a = 1.3$, the absolute value of the bias in both the *a*- and *b*-parameters for the easiest and most difficult items was somewhat less for NOHARM (Figures A3 and A4), but the standard errors for these items was greater for NOHARM (Figure A5).

Boulet (1996) documented a problem with extreme parameter estimates in NOHARM. In the current study, *b*-parameters greater than 4 in absolute value were considered extreme, given that the true values ranged from -2 to 2. Similarly, *a*-parameters greater than 3 in the normal metric (5.1 in the logit metric) were considered extreme, given that the highest true *a*-parameter

was 1.7. Before summarizing the data, as noted in the main text, extreme estimates were set equal to -4 or 4 for the *b*-parameters and 3 for the *a*-parameters. Extreme estimates were virtually never a problem when the sample size was 5000, but there were some occurrences when the sample size was 300. For the small sample size, Table A1 shows the number of extreme estimates, per 1000 estimates, for NOHARM and each of the M*plus* methods. For the *b*-parameters, extreme estimates occurred most frequently for the M*plus* LI method, followed by the NOHARM LI method. Nearly all extreme *b*-estimates were due to the less-discriminating, easy or hard items, which thus had much higher rates than the means displayed in the table. Boulet also found more extreme *b*-estimates for easy or difficult items, but in contrast to the present study these extreme *b*'s were associated with highly discriminating items. For the *a*-parameters, extreme estimates occurred much more frequently for NOHARM. Nearly all extreme *a*-estimates were due to the most-discriminating easy items for the negatively-skewed θ and to the most-discriminating difficult items for the positively-skewed θ, consistent with Boulet's findings of more frequent extreme *a*-estimates for NOHARM for highly discriminating items.

In summary, although the limited-information procedure applied in NOHARM is very different from the limited-information procedure applied in M*plus*, the results were similar. The parameters for extreme items have large bias when the θ is not normal. The NOHARM method does not seem to be more accurate than the M*plus* LI method, and NOHARM cannot be used for the full structural model, only for the measurement part of the model.

Table A1

Extreme Values per 1000 Estimates for N = 300

	Negatively Skewed θ	Platykurtic θ	Positively Skewed θ
b-parameters			
NOHARM	6.6	3.7	6.4
Mplus LI	9.9	3.8	10.2
Mplus FI - normal	4.0	3.0	4.0
Mplus FI - nonnormal	3.4	3.0	3.5
a-parameters			
NOHARM	12.9	0.0	13.9
Mplus LI	3.6	0.0	3.4
Mplus FI - normal	0.4	0.1	0.4
Mplus FI - nonnormal	0.4	0.2	0.3

Appendix B

Estimating the Latent Distribution

With real data, the underlying distribution would not be known. One method of estimating the distribution of θ is the histogram method available in BILOG. An alternative method, available in M*plus*, is to estimate it as a mixture of normal distributions. Both methods were briefly examined for this study.

Estimating the Distribution of θ with Quadrature Points in Bilog

To make the BILOG estimation as comparable as possible to M*plus*, no prior distributions were used on the item parameters and 15 quadrature points were specified. Prior distributions are often recommended for the *a*-parameters, but that would add another difference between the BILOG and M*plus* estimates as priors can not be used in M*plus*. The EMPIRICAL option was specified so that the θ distribution would be estimated instead of assumed normal. With this option, the proportions of examinees at each quadrature point are estimated after each cycle of the item parameter estimation. The metric is determined by adjusting the resulting θ distribution to mean = 0 and standard deviation = 1. Default starting values were used.

In Figures B1 and B2, the M*plus* LI estimates have been removed and the BILOG FI estimates are compared only to the M*plus* FI estimates to simplify the graphs. The scale has been kept the same as that in the main text.

For the *b*-parameters, the bias was very similar to the bias using the known θ distribution in M*plus*. The BILOG estimates were slightly biased away from the mean. This parallels Boulet's (1996) TESTFACT results. Figure B1 illustrates this outward bias for the negative skew distribution, large sample size, but this was true for all distributions and was not in the opposite direction for the positively skewed distribution. BILOG users may more typically see bias

towards the mean because BILOG is often implemented with Bayesian priors on each item parameter, which were omitted here. Notably, this outward bias was much smaller than the bias of the M*plus* FI-normal estimates or the LI estimates. For the *a*-parameters, the bias was virtually identical to the bias using the known θ distribution in M*plus*.

For the large sample size, empirical standard errors of both *a* and *b* were almost identical to standard errors using the known θ distribution in M*plus*. For the small sample, the standard errors of the BILOG *a*-parameter estimates (but not the *b*-parameter estimates) were slightly higher than the standard errors using the known θ distribution when the *a*-parameter was high (Figure B2). This difference was minimal, but it was observed for all three θ distributions.

Sometimes the accuracy of the estimated θ distribution itself may be of interest. One could test the difference between the true and estimated distributions with a statistical significance test such as the Kolmogorov-Smirnov, but for substantive interpretation it may be more meaningful to judge whether the estimated shape seems reasonably close. The estimated distribution for nine replications for the large sample size are shown in Figures B3 (negatively skewed) and B4 (platykurtic). There is some variance, but in general the distributions appear to have the correct substantive form. Variation was greater for the smaller sample (not shown). Additionally, the densities for the smaller samples appeared less smooth; often, one quadrature point would have a higher (or lower) density than the points on either side instead of a smooth progression over adjacent points. Although this did not have a large impact on the standard errors of the item parameter estimates, compared to the standard errors from integrating over the true θ distribution, it could lead to incorrect inferences about the substantive form of the θ distribution for small samples.

Overall, even with small samples, very little accuracy in item parameter estimation was lost by estimating the θ distribution in BILOG instead of integrating over the true θ distribution. Additionally, BILOG has the flexibility of estimating the 3PL model, not just the 2PL model available in M*plus* and LISREL. One can also use prior distributions on the item parameters in BILOG, decreasing the standard errors for small samples. However, BILOG will only estimate unidimensional models, and it only includes the measurement part of the model, so it is of little use to those running multidimensional models or full structural models. This demonstration does show the feasibility of estimating the θ distribution using quadrature points.

Estimating the Distribution of θ with Mixture Models in M*plus*

The θ distributions were generated by mixing normal distributions, which is consistent with the mixture model implemented in M*plus*. Because of the computer time needed for estimating IRT mixture models, only the first 100 replications were run and the positive-skewed distribution was not used because with the other estimation methods the results had been comparable to those with the negative-skewed distribution. The true number of classes, 2 for the skewed θ distribution and 3 for the polykurtic distribution, was specified. No effort was made to compare models with different numbers of classes. One thousand random start values were used, with optimization completed on four. Up to 50 iterations were carried out in the initial stage, with the default of up to 500 in the final optimization. The default 15 quadrature points were used. All *a*-parameters were assigned starting values of 1.7 in the unperturbed starting-value set. The distribution for one class was fixed to a mean of 0 and variance of 1. For the skewed θ, the distribution of the other class was assigned starting values of (-1, 1). For the polykurtic θ, the distribution of class 2 was assigned starting values of (-1, 1) and the distribution of class 3 was assigned starting values of (1, 1). For one replication that would not terminate normally, these

starting values were adjusted. Before summarizing the results, the metric of the item parameters and class means and variances were adjusted based on the constants needed to standardize the total population to a mean of 0 and standard deviation of 1.

For the sample size of 5000, both the *a*- and *b*-parameters were virtually unbiased. The standard errors were either the same as or only *slightly* larger than when integration was over the known θ distribution. This was true for both the skewed and platykurtic θ distribution. For the sample size of 300, the bias in both the *a*- and *b*-parameters was close to the bias from the true θ distribution. The standard errors of the *b*-parameters were only slightly larger than they were for the true θ distribution, but the standard errors of the *a*-parameters were noticeably larger, especially for the platykurtic θ distribution (Figure B5).

As discussed for BILOG, sometimes users may be interested in the shape of the estimated θ distribution. Additionally, M*plus* users may be interested in the distributions of the classes. Figures B6 and B7 show the first nine replications for the large sample size. The class densities are shown with dotted lines and the total density is shown with a solid line. The total distribution was relatively consistent but there was somewhat more variance across replications in the class distributions, especially for the platykurtic θ distribution (Figure B7). The estimated shape of the small sample distribution (not shown) varied considerably more, and there were sometimes spikes due to one class with a very small variance, especially for the platykurtic distribution, which had three classes. These differences in the estimated θ distribution across replications likely contributed to the increased standard error of the *a*-parameter. Additionally, poor inferences about the shape of the θ distribution would be frequent for the small sample.

Sample sizes of 300 appear to be too small to consistently estimate the *a*-parameters while simultaneously estimating the θ distribution as a mixture of normal distributions. With

small samples, there may be less error from incorrectly integrating over a normal distribution than from estimating the θ distribution. However, with larger samples, estimating the distribution tentatively appears to be a feasible approach.

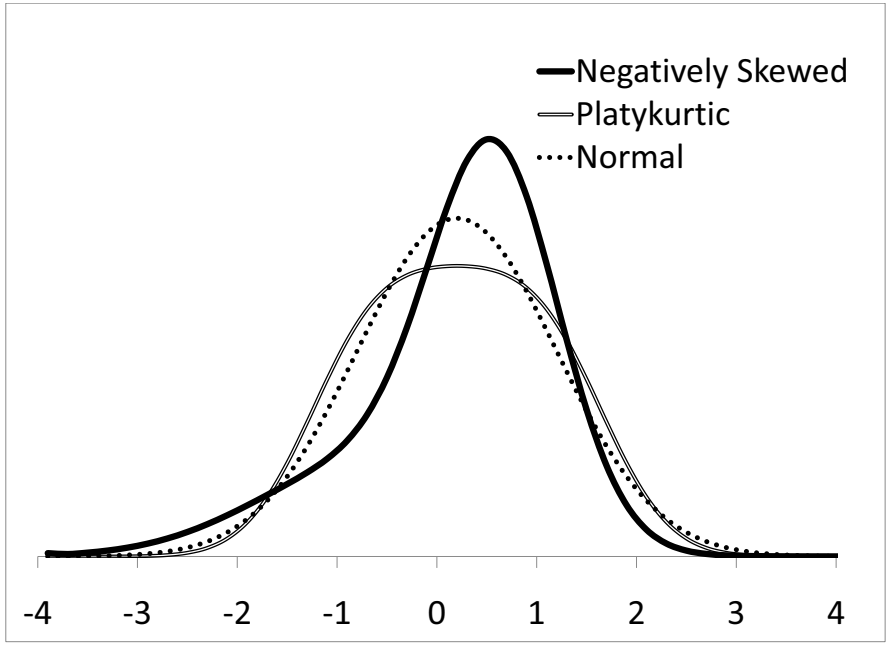


Figure 1. Distributions used to generate the data. A normal distribution is also shown for comparison.

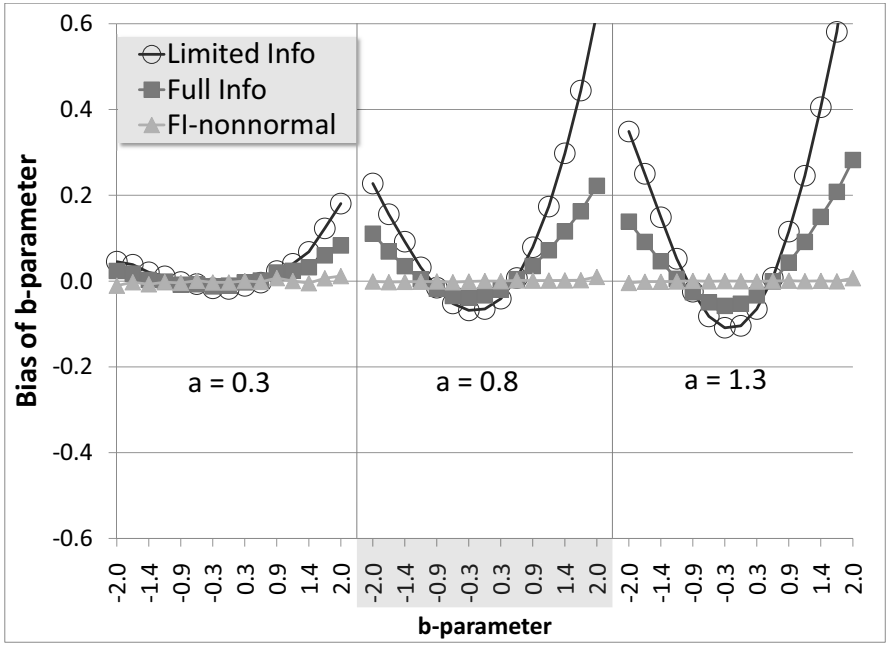


Figure 2. Bias of estimated b-parameter, negatively skewed distribution, N = 5000.

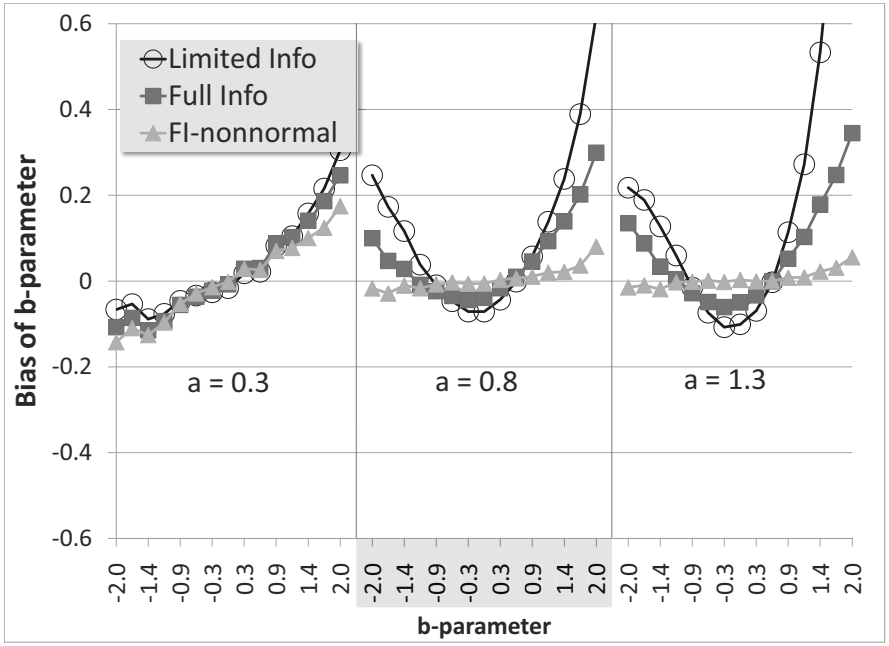


Figure 3. Bias of estimated b-parameter, negatively skewed distribution, N = 300.

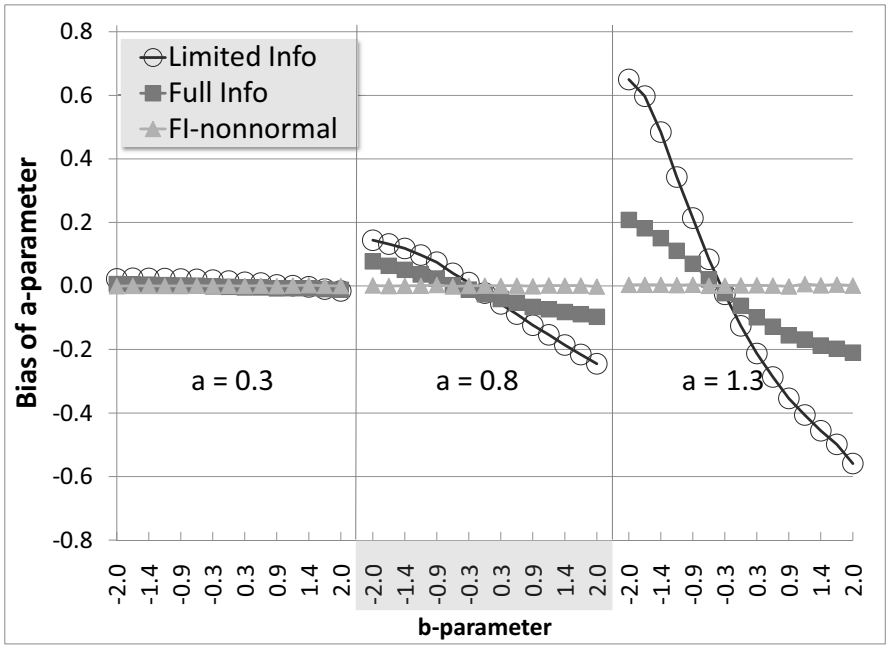


Figure 4. Bias of estimated a-parameter, negatively skewed distribution, N = 5000.

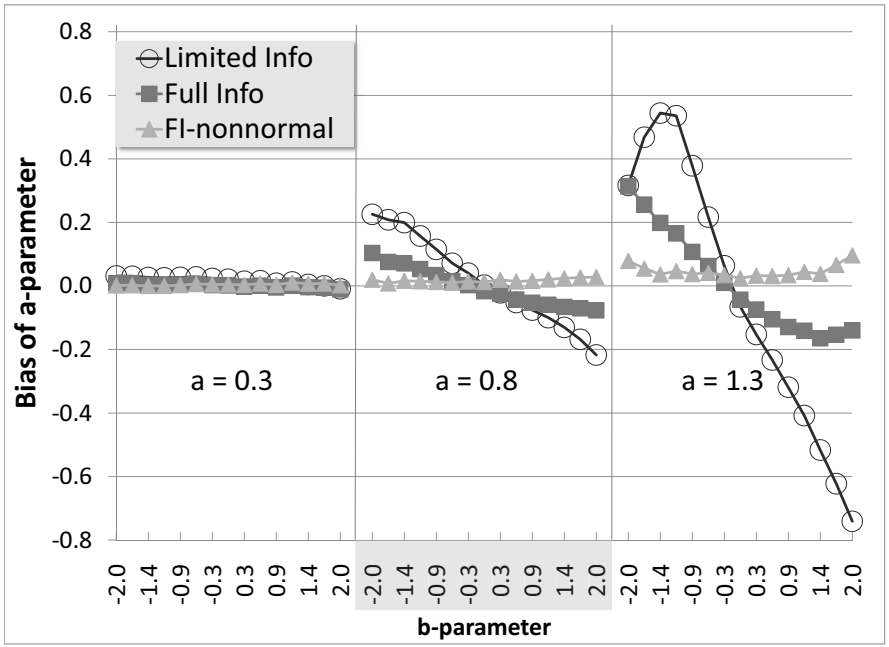


Figure 5. Bias of estimated a-parameter, negatively skewed distribution, N = 300.

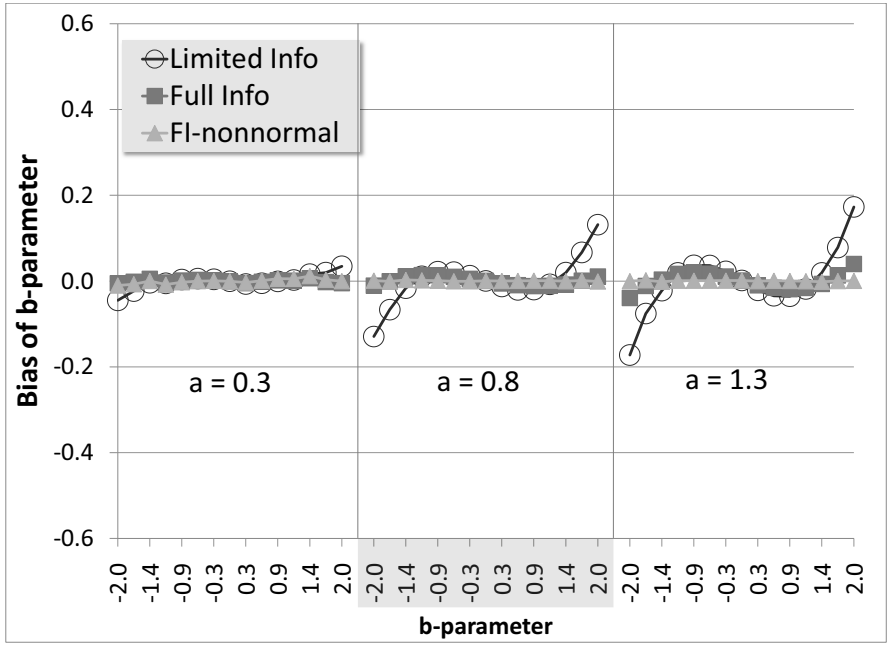


Figure 6. Bias of estimated b-parameter, platykurtic distribution, N = 5000.

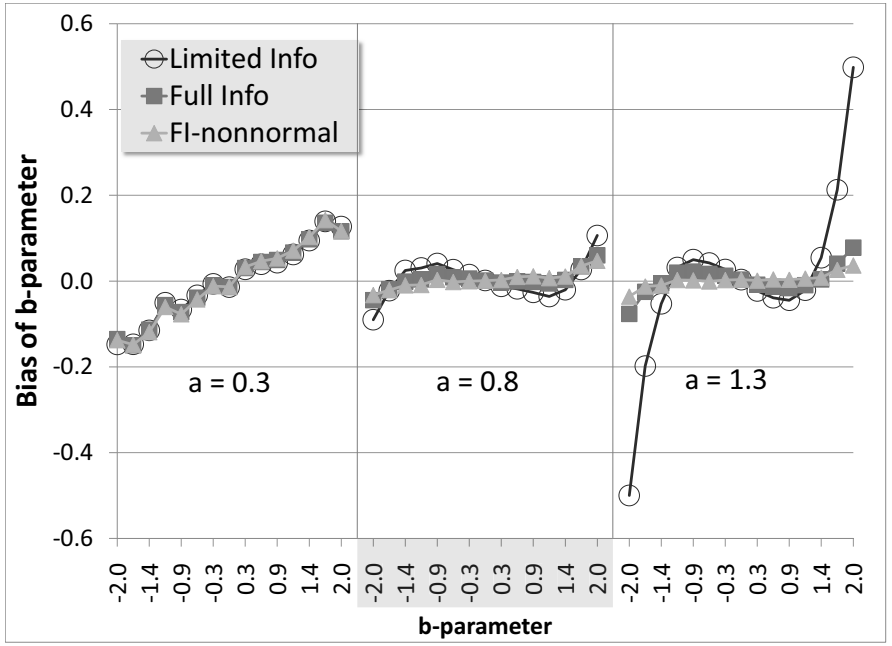


Figure 7. Bias of estimated b-parameter, playtkurtic distribution, N = 300.

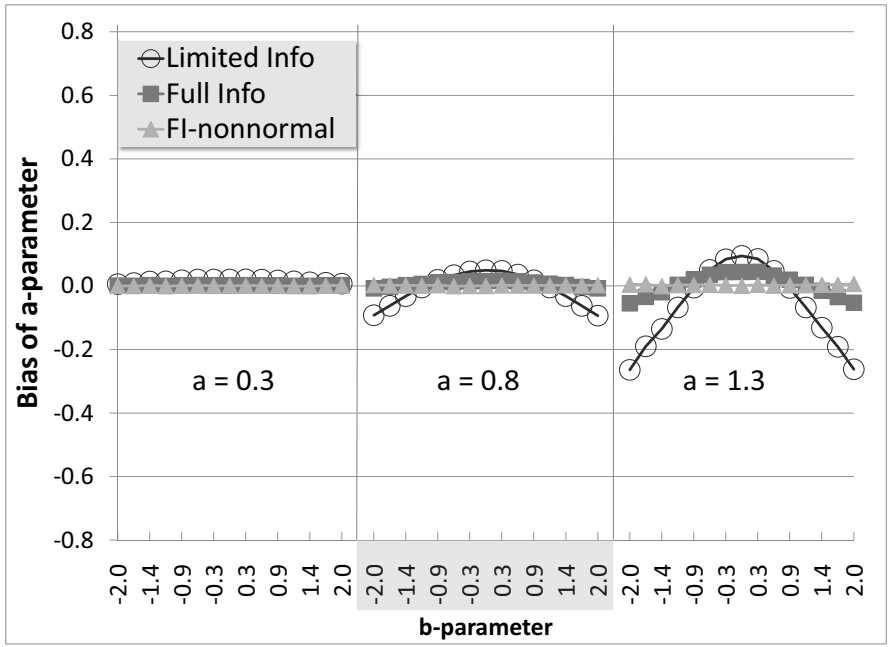


Figure 8. Bias of estimated a-parameter, playtkurtic distribution, N = 5000.

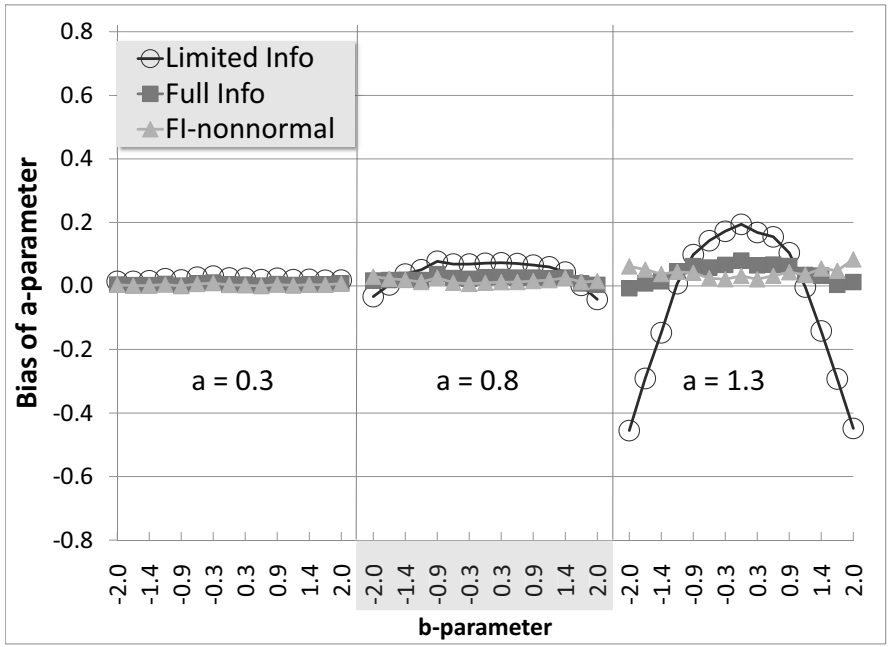


Figure 9. Bias of estimated a-parameter, playkurtic distribution, N = 300.

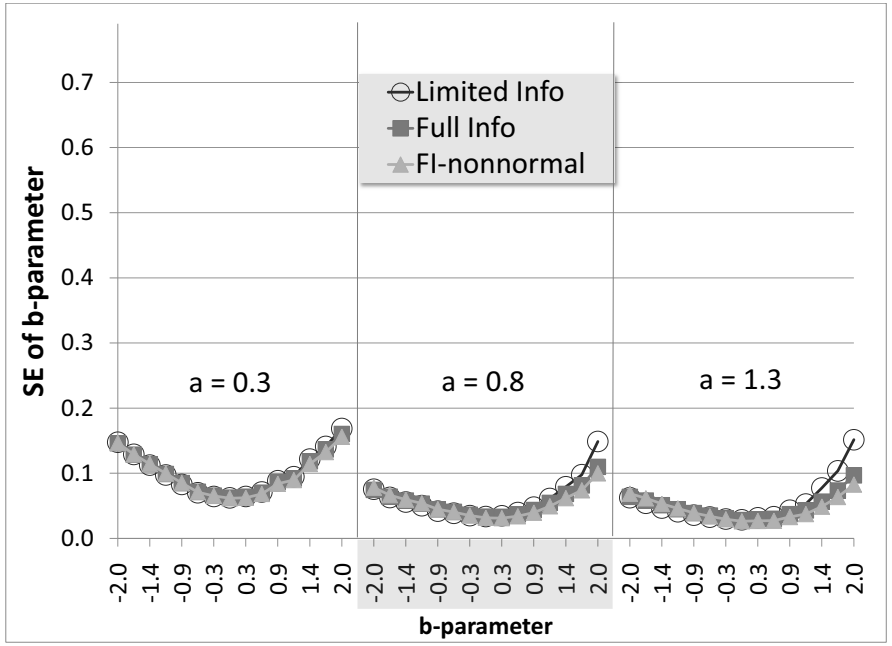


Figure 10. SE of estimated b-parameter, negatively skewed distribution, N = 5000.

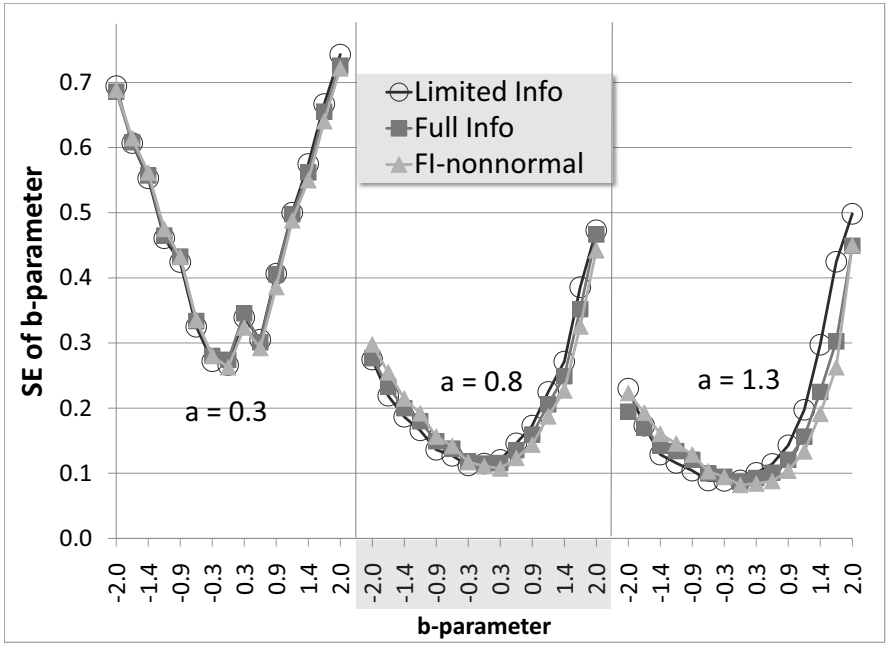


Figure 11. SE of estimated b-parameter, negatively skewed distribution, N = 300.

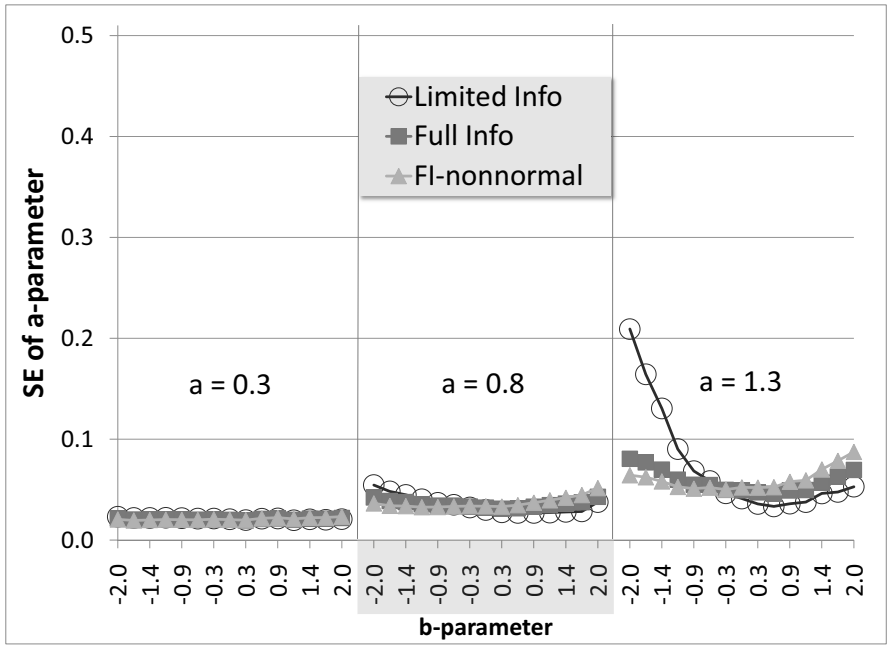


Figure 12. SE of estimated a-parameter, negatively skewed distribution, N = 5000.

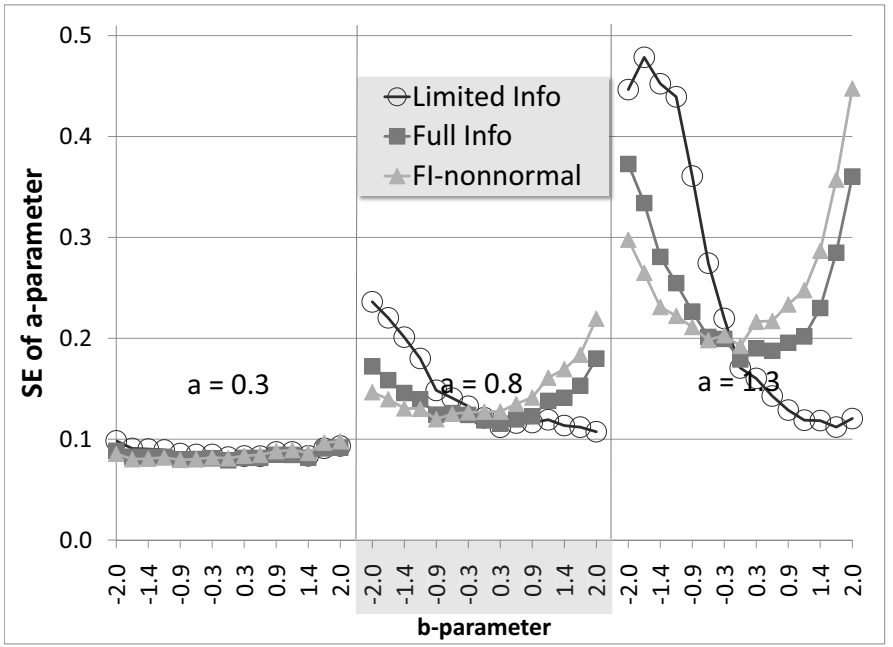


Figure 13. SE of estimated a-parameter, negatively skewed distribution, N = 300.

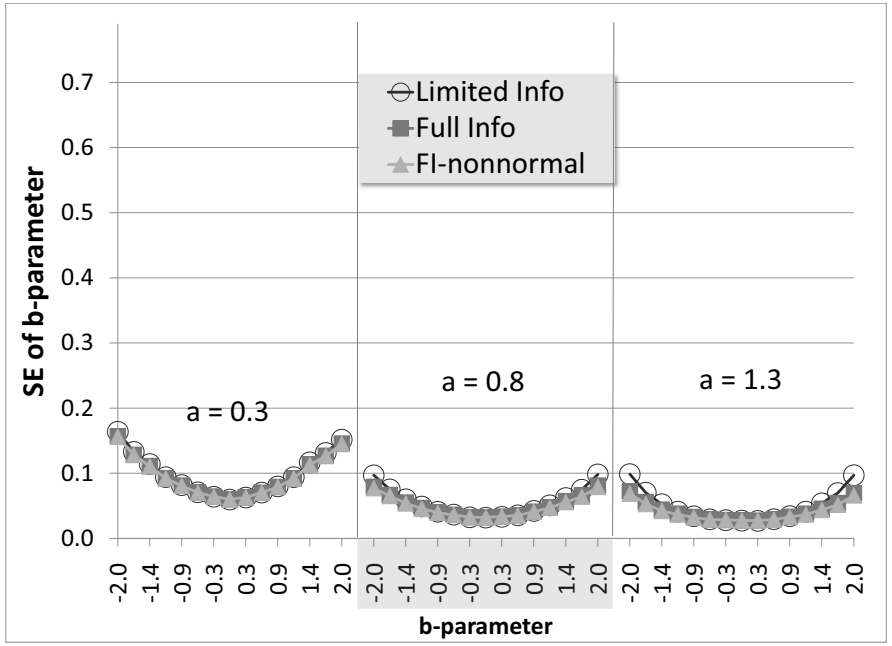


Figure 14. SE of estimated b-parameter, platkurtic distribution, N = 5000.

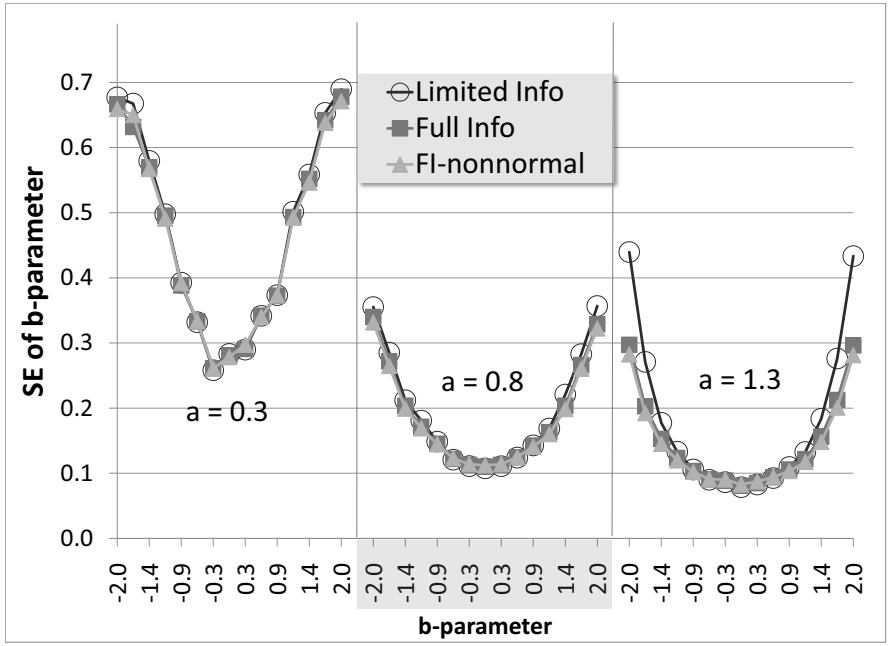


Figure 15. SE of estimated b-parameter, platykurtic distribution, N = 300.

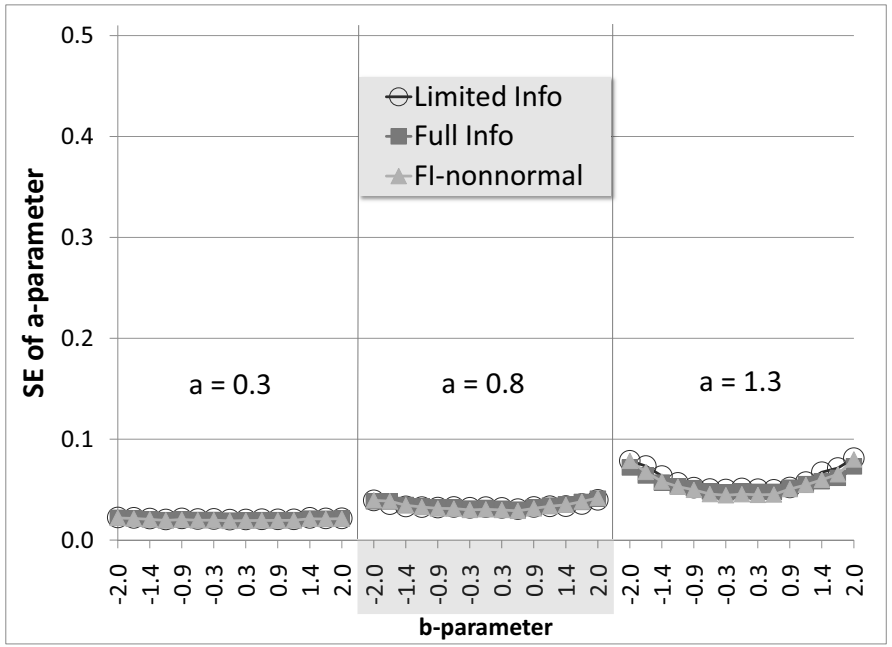


Figure 16. SE of estimated a-parameter, platykurtic distribution, N = 5000.

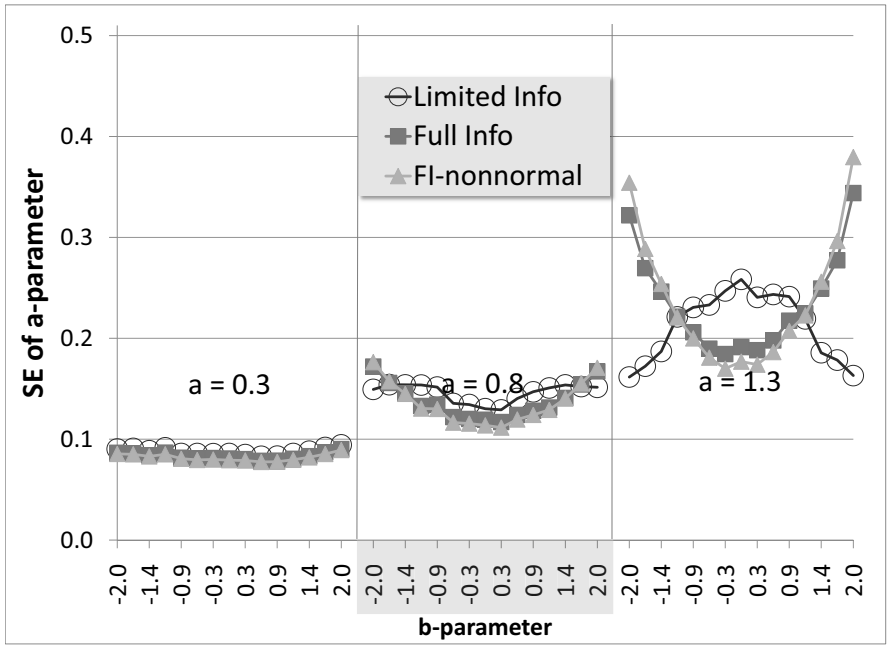


Figure 17. SE of estimated a-parameter, playkurtic distribution, N = 300.

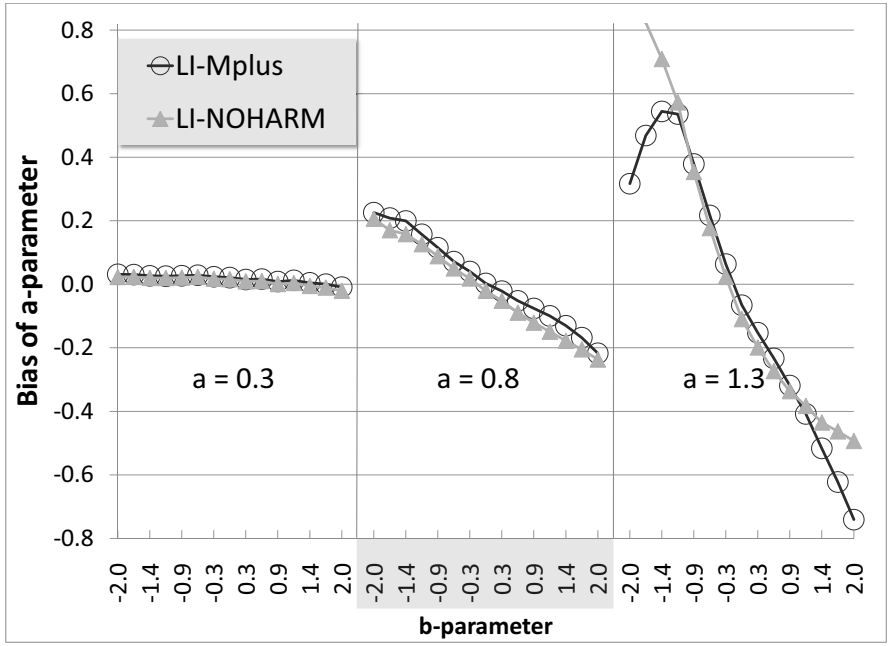


Figure A1. Bias of estimated *a* -parameter, negatively skewed distribution, N = 300. Limited-information NOHARM is compared to limited-information M*plus* .

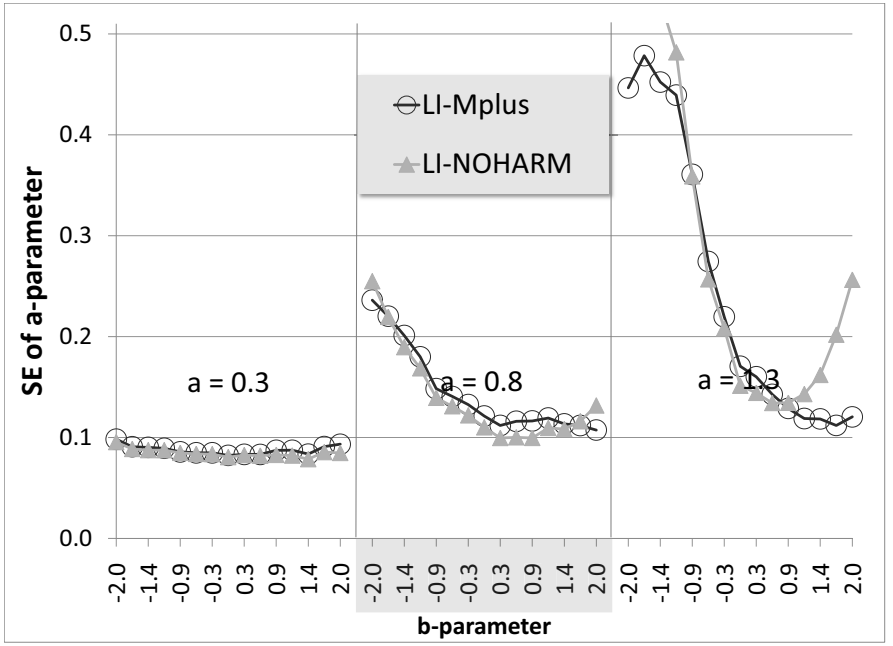


Figure A2. SE of estimated *a* -parameter, negatively skewed distribution, N = 300. Limited-information NOHARM is compared to limited-information M*plus* .

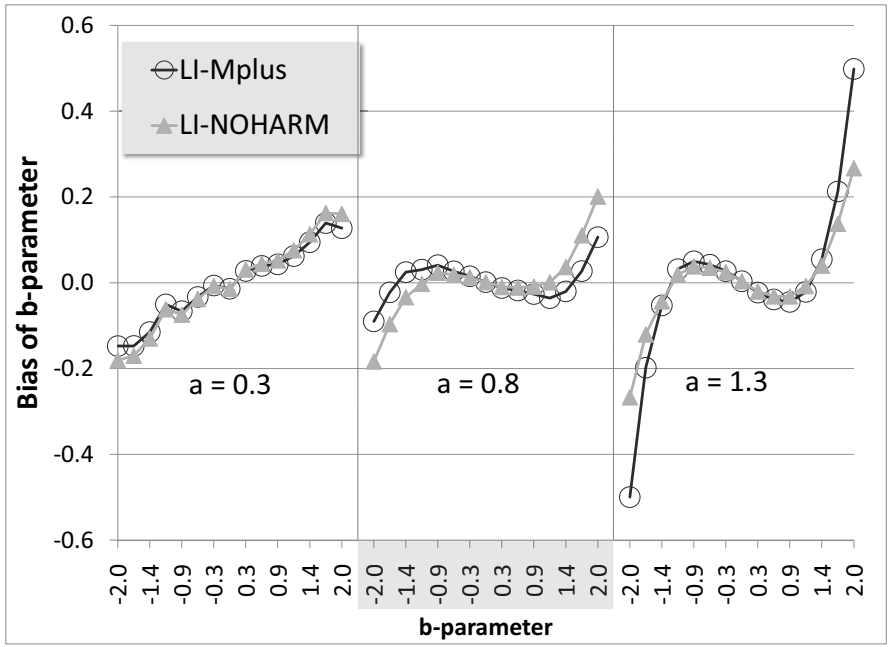


Figure A3. Bias of estimated *b* -parameter, playtkurtic distribution, N = 300. Limited-information NOHARM is compared to limited-information M*plus* .

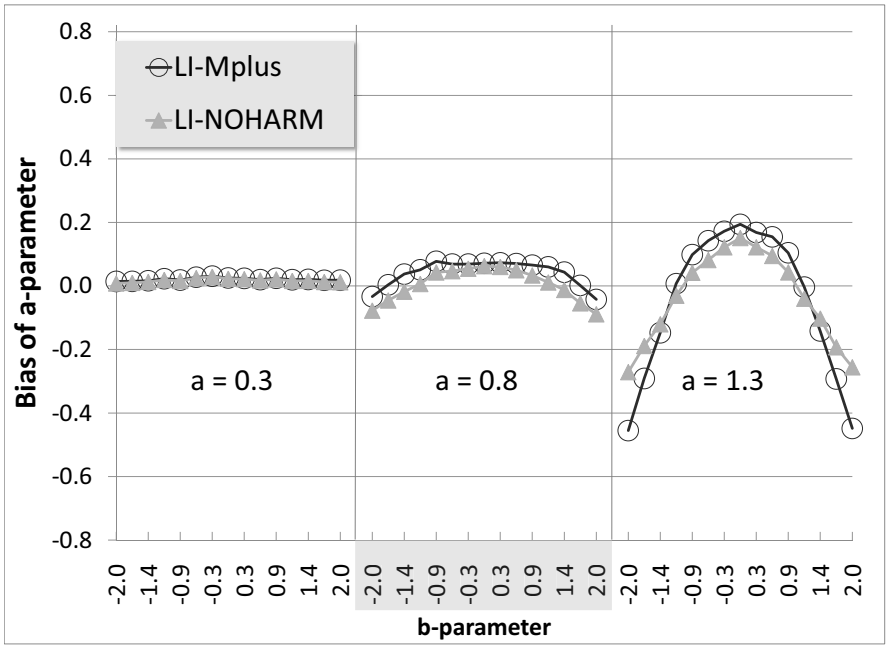


Figure A4. Bias of estimated *a* -parameter, playkurtic distribution, N = 300. Limited-information NOHARM is compared to limited-information M*plus* .

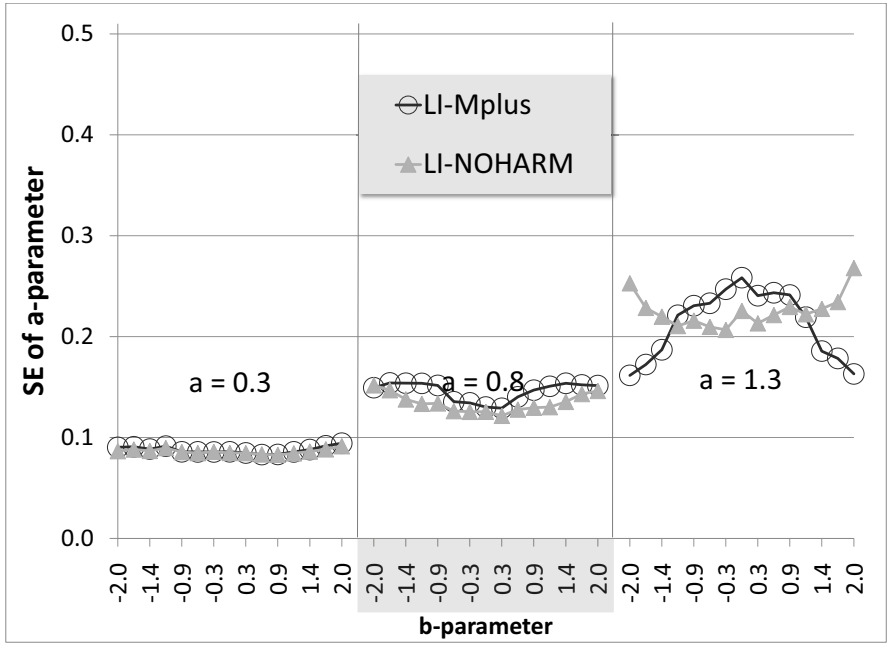


Figure A5. SE of estimated *a* -parameter, playkurtic distribution, N = 300. Limited-information NOHARM is compared to limited-information M*plus* .

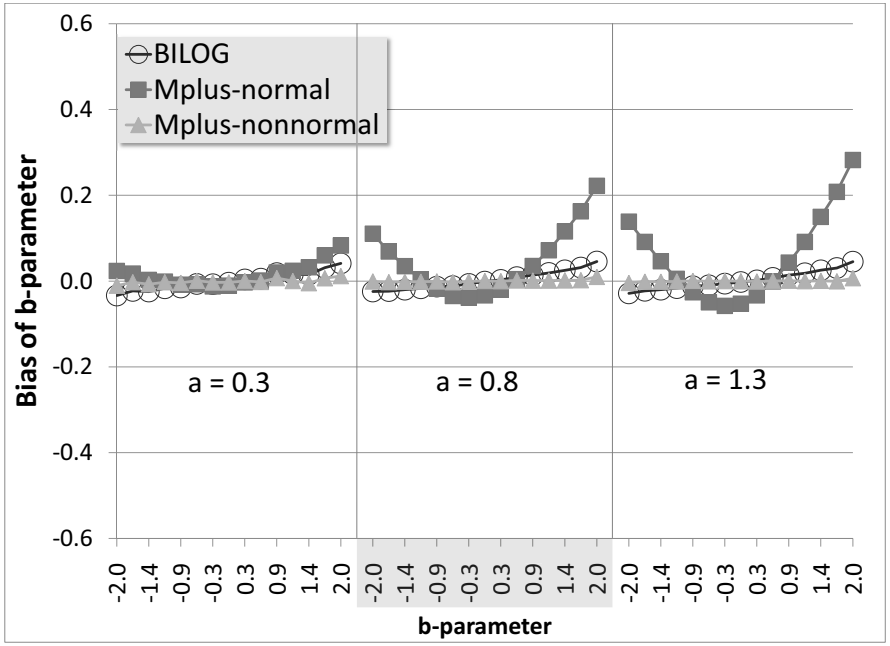


Figure B1. Bias of estimated *b* -parameter, negatively skewed distribution, N = 5000.

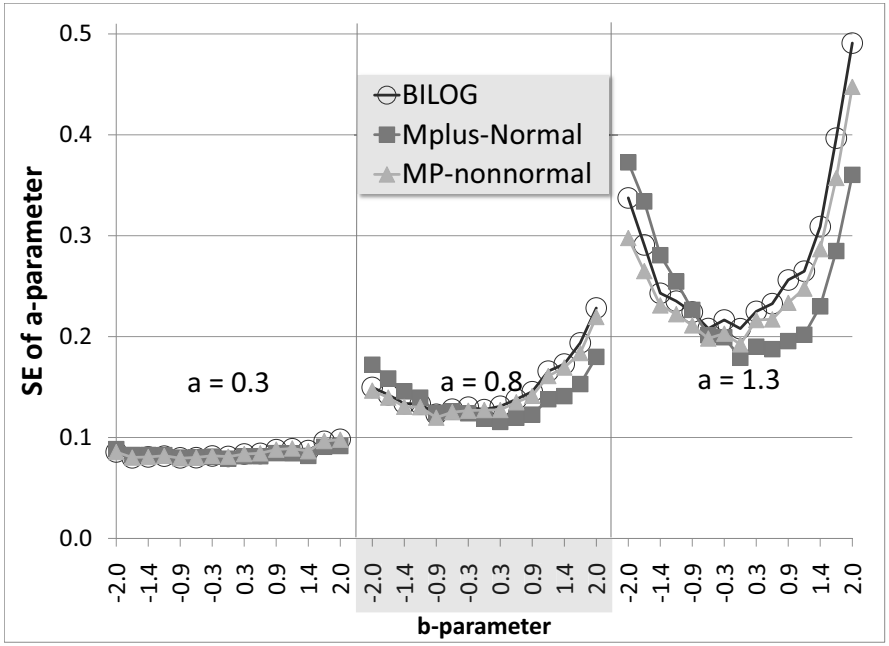
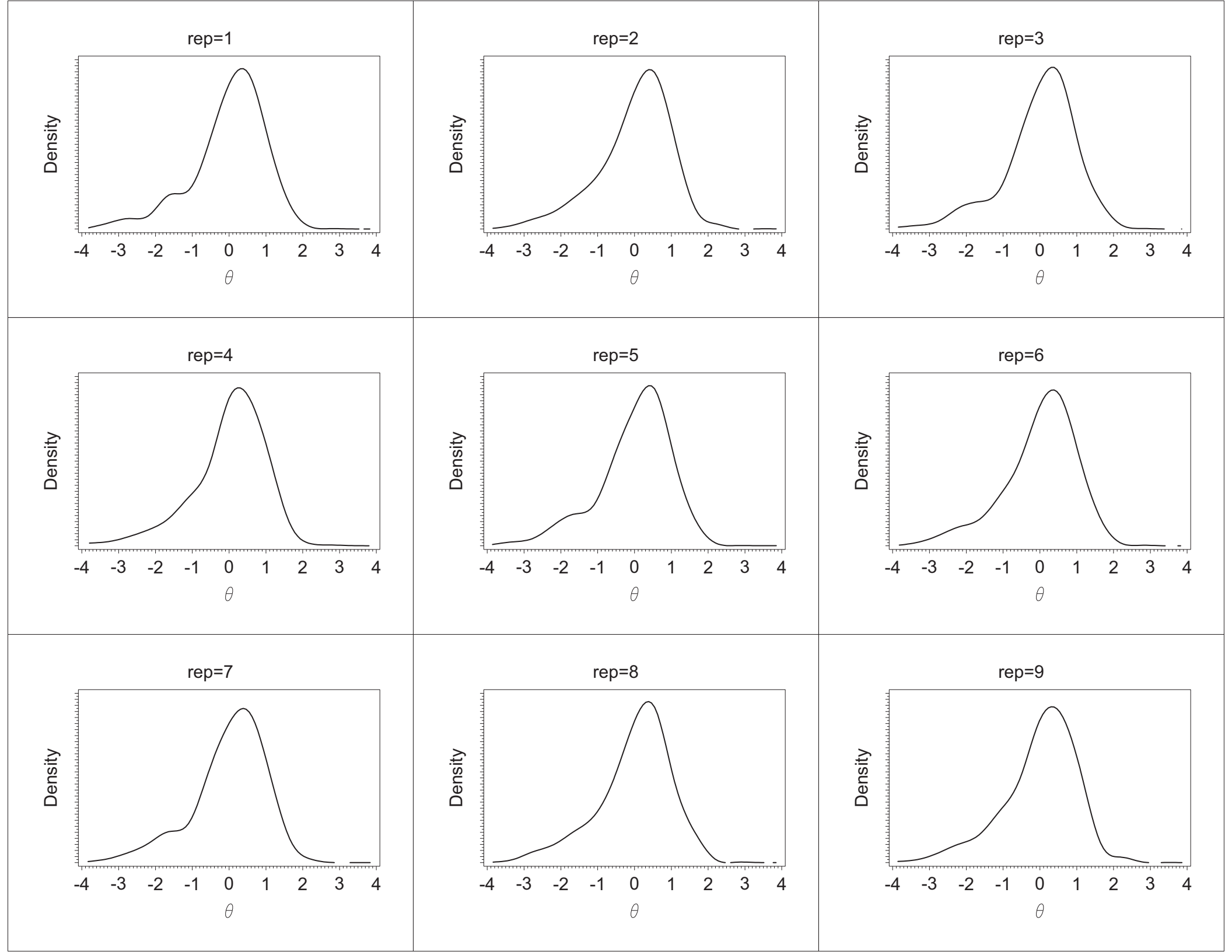
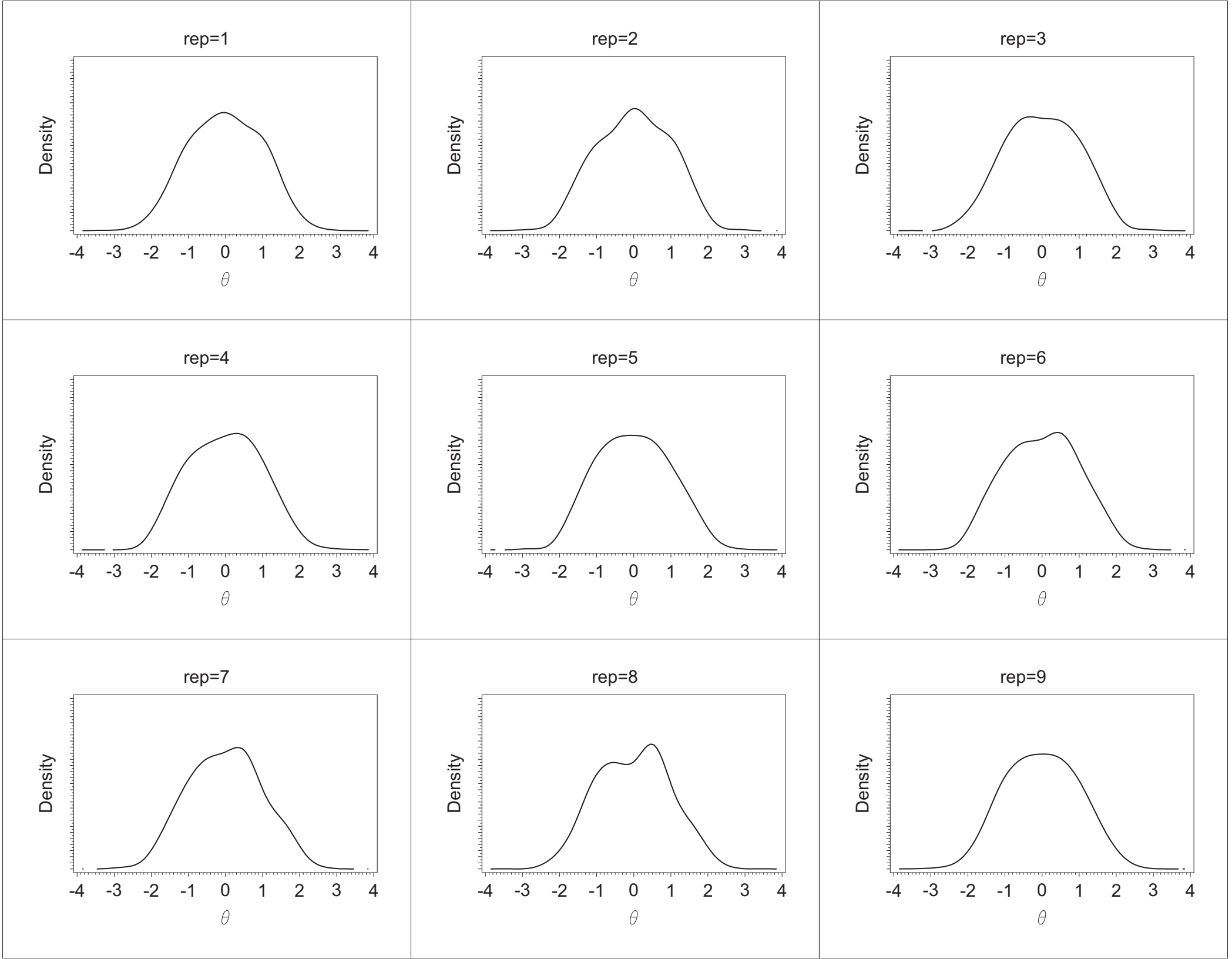


Figure B2. SE of estimated *a* -parameter, negatively skewed distribution, N = 300.


rep=1
rep=2
rep=3
rep=4
rep=5
rep=6
rep=7
rep=8
rep=9
Density
θ
-4 -3 -2 -1 0 1 2 3 4


rep=1
rep=2
rep=3
rep=4
rep=5
rep=6
rep=7
rep=8
rep=9
Density
-4 -3 -2 -1 0 1 2 3 4
θ

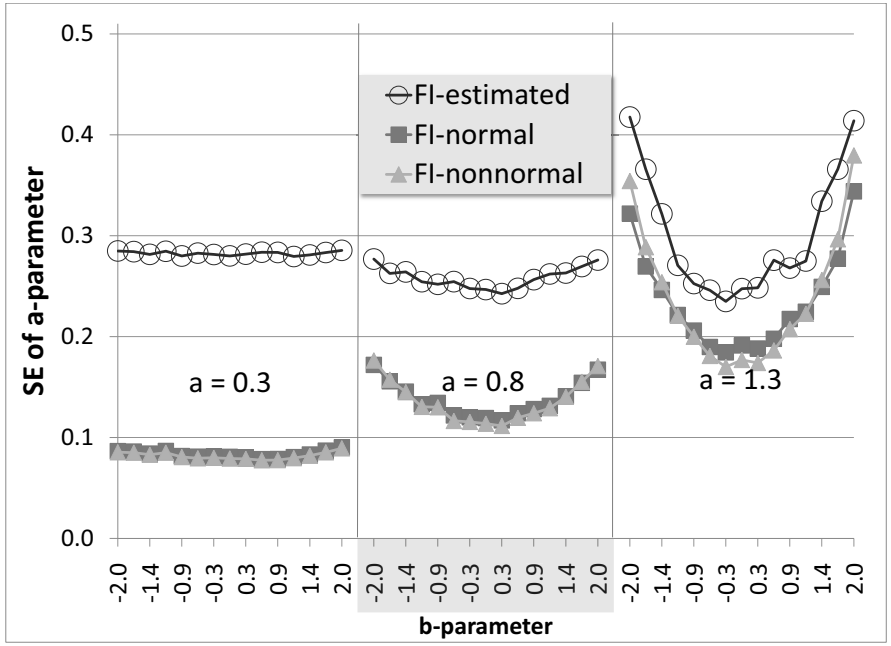
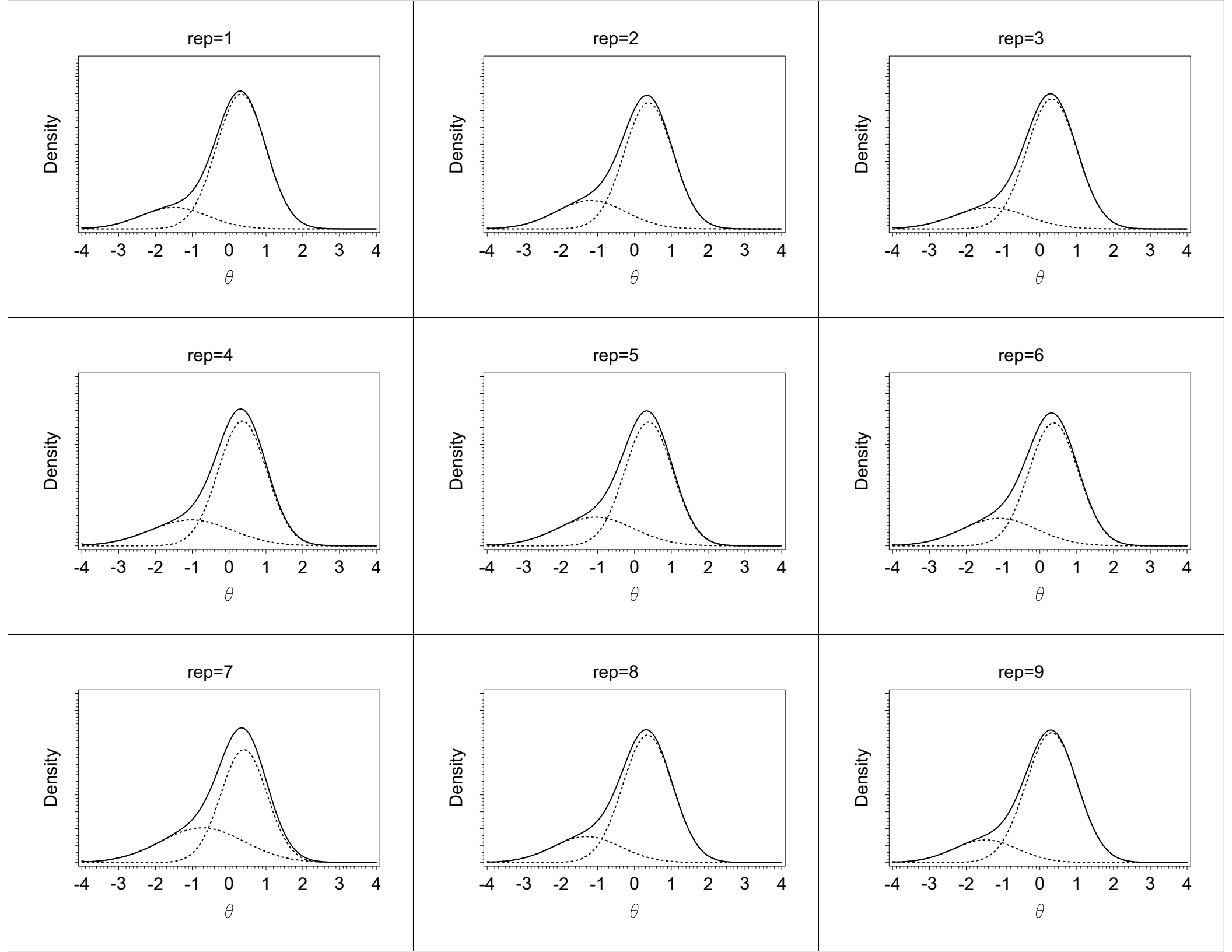
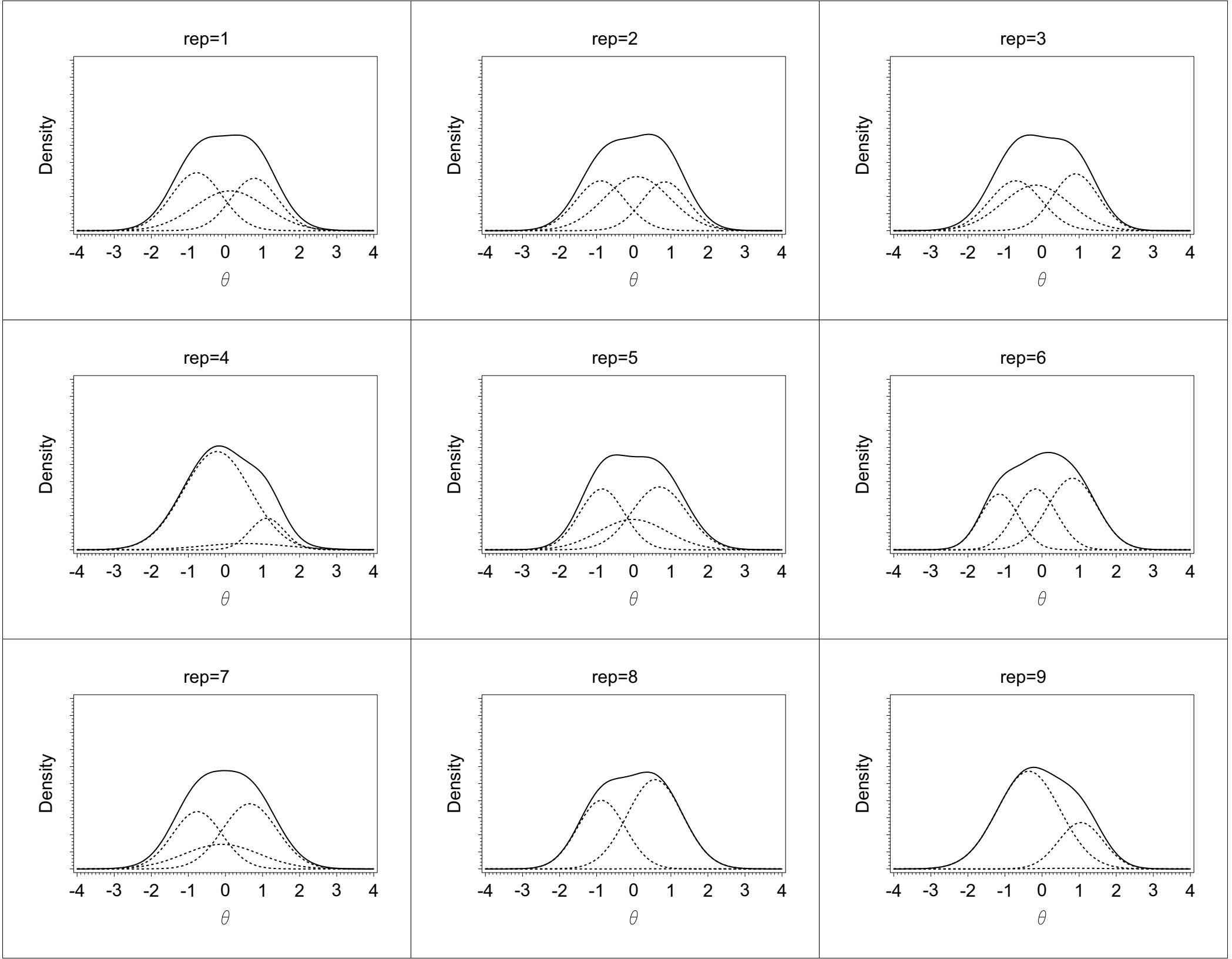


Figure B5. SE of estimated *a* -parameter, playkurtic distribution, N = 300. Three M*plus* full-information methods are compared.


rep=1
rep=2
rep=3
rep=4
rep=5
rep=6
rep=7
rep=8
rep=9
Density
-4 -3 -2 -1 0 1 2 3 4
θ


rep=1
rep=2
rep=3
rep=4
rep=5
rep=6
rep=7
rep=8
rep=9
Density
θ
-4 -3 -2 -1 0 1 2 3 4