A COMPARATIVE STUDY OF ITEM-LEVEL FIT INDICES IN ITEM RESPONSE THEORY

A DISSERTATION SUBMITTED TO THE FACULTY OF THE
GRADUATE SCHOOL OF THE UNIVERSITY OF MINNESOTA
BY

Jennifer Paige Davis

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY

Professor David J. Weiss, Adviser

June 2009

ACKNOWLEDGEMENTS

I would like to thank my adviser, Dr. David Weiss, for his help and support during my doctoral studies. His feedback on drafts of this dissertation and words of encouragement were invaluable. Thanks also are due to my committee members – Dr. Mark Davison, Dr. Michael Harwell, and Dr. John Campbell – who provided helpful feedback on the preliminary exam and dissertation. Dr. Kathleen Gialluca, from Pearson VUE, also deserves many thanks for her support on the job of my earlier, preliminary exam, doctoral studies.

Additionally, I would like to thank my family for their support during my long journey to the PhD. Finally, special thanks are due to my husband, Keith Davis, who inspired me to persist through difficult times while working on the dissertation; he truly helped keep me going and enabled me to complete this degree.

ABSTRACT

Item-level fit indices (IFI) in item response theory (IRT) are designed to assess the degree to which an estimated item response function approximates an observed item response pattern. There are numerous IFIs whose theoretical sampling distributions are specified; however, in some cases little is known regarding the degree to which these indices follow their theoretical distributions in practice. If an IFI departs substantially from its theoretical distribution, degree of misfit will be misestimated, and test developers will have very little idea of whether their models provide accurate depictions of true item response behavior. Therefore, a Monte Carlo simulation study was conducted to assess the degree to which many available IFIs follow their theoretical distributions. The IFIs examined in this study were (1) Infit (*V*I) and Outfit (*V*O), two IFIs commonly used for the Rasch model; (2) Yen's (1981) χ^2 (*Q*1) and Orlando and Thissen's (2000) χ^2 (*Q*O); (3) three Langrange multiplier statistics [LM(α), LM(β), and LM($\alpha\beta$)] proposed by Glas (1999); and (4) Dragow, Levine, and Williams' (1985) person fit L_z modified by Reise (1990) to assess item fit.

The primary research objective of this study was to determine how a number of factors (listed below) affect Type I error rates and empirical sampling distributions of IFIs. The relationship between IFIs and item parameters was also examined. The crossed between-subjects conditions were: IRT model (1-, 2-, and 3 parameter); data noise, operationalized as strictly unidimensional vs. essentially unidimensional data; item discrimination (high and low); test length (n = 15 and n = 75); and sample size (N = 500 and N = 1,500). There were also two crossed within-subjects factors to capture the impact of item and person parameter estimation error. The dependent variables in this study were IFI Type I error rates and empirical sampling distribution moments across 18,750 replicated items. Data were analyzed and summarized using ANOVA, Pearson correlations, and graphical procedures. The Kolmogorov-Smirnov test was used to directly assess distributional assumptions.

The results of the study indicated that *Q*O was the only statistic to adhere closely to its theoretical sampling distribution across all study conditions. For *V*I, *V*O, L_z, and *Q*1

statistics, sampling distributions were strongly influenced by test length, parameter estimation error, and, to a lesser degree, sample size. In the absence of parameter estimation error, all statistics more closely approximated their theoretical sampling distributions and were affected little by other study conditions. The presence of person parameter estimation error tended to have an inflationary effect on sampling distribution means whereas the presence of item parameter estimation error tended to have a deflationary effect on sampling distribution variances. *V*I, *V*O, and L_z functioned very similarly to one another, with Type I error rates tending to be grossly inflated for $n = 15$ and deflated for $n = 75$ when both person and item parameter error were present. $Q1$ Type I error rates were also grossly inflated for $n = 15$, but were near nominal levels for $n = 75$. Finally, the LM statistics generally exhibited inflated Type I error rates and were moderately influenced by IRT model and discrimination; only for LM(β) did empirical sampling distributions tend to approach theoretical distributions, primarily when discrimination was lower or for the 3-parameter model at both levels of discrimination.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Item response theory (IRT) models have gained widespread use since their introduction some 40 years ago. These models offer many advantages over classical test theory-based approaches to developing tests and scoring examinees (for a discussion of the advantages see Hanbleton & Swaminathan, 1985). However, these advantages are based on certain assumptions about the data to which the models are fit. When using any parametric model, it is crucial to assess model-data fit before using the model for any applied purpose since a model is only as useful as the degree to which it correctly approximates the data. If the model does not fit the data, its use for scoring examinees or equating test forms would be unjustifiable.

Most commonly used IRT models assume that only one latent variable accounts for examinee response to each item on a test; these are unidimensional IRT models. The most general form of a dichotomous unidimensional IRT model, among those most commonly used, is given the by three-parameter logistic (3PL) model item response function (IRF), defined as

$$P_i(u_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)P_i^* \tag{1}$$

where

$$P_i^* = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]}$$

$i = 1, \ldots, n$ items,

$j = 1, \ldots, N$ examinees,

u_{ij} is the item response; $u = 1$ for correct and $u = 0$ for incorrect response,

θ_j is the latent ability for the jth examinee, and

a_i, b_i, and c_i are the discrimination, difficulty, and lower-asymptote parameters (respectively) for the ith item.

When there is no possibility for an examinee to correctly answer an item by guessing, c_i can be fixed to zero and the two-parameter logistic (2PL) model is obtained. Finally, the one-parameter logistic (1PL) model is obtained by setting all a_i in the 2PL equal to each

other. There are also polytomous IRT models available for use with items that have more than two scored response options. These models will not be defined here since they are outside the scope of the present research. Several studies have examined the functioning of item fit statistics for polytomous models. These studies will be discussed in the literature review, but the scope of the research conducted here concerns only the application of item fit statistics to dichotomous IRT models (for a discussion of polytomous IRT models see Thissen & Steinberg, 1986).

Unidimensional IRT models make several assumptions regarding the data to which they are fit. If these assumptions are violated in the data, then the model does not fit the data. These assumptions are (1) adherence to the mathematical form of the IRF and (2) presence of one θ accounting for item response (unidimensionality). Item fit indices have been proposed to address each of these two aspects of fit. Testing the first assumption involves evaluating data on an item-by-item basis whereas testing the second assumption requires data from pairs of items. This study focused solely on methods assessing fit on an item-by-item basis.

Despite the theoretical importance of model-data fit, little research has been done over the last 40 years regarding methods of assessing fit and the practical consequences of model misfit. To date, there are a variety of methods available to assess item-level model-data fit. Group-based chi-square (χ^2) methods (Bock, 1972; Yen, 1981) appear to be most prevalent. However, these methods suffer shortcomings, which will be elucidated below. Methods have been developed within the past decade that attempt to ameliorate some of the problems with the χ^2 statistics (Orlando & Thissen, 2000; Stone, 2000). Increased computing power has opened the frontier for more computationally intensive procedures that utilize Monte Carlo simulation (Sinharay, 2005; Sinharay, Johnson, & Stern, 2006; Stone, 2000). Additionally, several approaches based on the model likelihood function have also been proposed (Glas, 1999; Glas & Falcon, 2003; Orlando & Thissen, 2000; Reise, 1990; Stone, 2000).

Most of the currently available methods for assessing item-level model-data fit were examined in this study. The research focused on item fit because such analysis,

compared with test-level model fit such as dimensionality assessment, potentially can provide more direction to practitioners in terms of how to refine a measurement instrument. For discussions and studies regarding test-level fit assessment, refer to Cai, Maydeu-Olivares, Coffman, and Thissen, 2006; DeChamplain, 1996; Hattie, 1985; Kang and Cohen, 2007; Maydeu-Olivares and Cai, 2006; and McDonald and Mok, 1995. All item fit indices used in this research are presented and briefly described below. Several methods not utilized in this study are also briefly described. After each type of item fit index is defined, a review of past simulation studies involving the indices of interest is presented.

Item Fit Indices: Definition and Literature Review

Item fit indices can be broken down into several types, most of which are based on residuals, defined by

$$e_{ik} = \frac{O(x_{ik}) - E(x_{ik})}{SD(x_{ik})} \tag{2}$$

where

$E(x_{ik}) = n_k P(x_{ik})$,

$SD(x_{ik}) = \{n_k P(x_{ik})[1 - P(x_{ik})]\}^{1/2}$,

$O(x_{ik}) = \sum_{j \in k} x_{ij}$,

i = 1, 2, …, n items,

k = 1, 2, …, K examinee groups created by classifying examinees according to number-correct score or estimated ability ($\hat{\theta}$),

j = 1, 2, …, N examinees,

n_k = number of examinees in the kth group, and

x is an item score.

These residuals can be defined at the individual examinee level or the group level. When defined at the individual level, $K = N$, $n_k = 1$, and $P(x_{ik})$ is defined by a particular model IRF; for 1-, 2-, and 3PL dichotomous models, $P(x_{ik} = 1)$ can be obtained by substituting

model parameter estimates in Equation 1 and $P(x_{ik} = 0)$ can be obtained by $1 - P(x_{ik} = 1)$. Note that, for the individual-level statistics defined below, the subscript *k* will be replaced with *j*. When residuals are defined at the group level, there are several forms of $P(x_{ik})$, which will be presented below.

The residuals in Equation 2 are defined in the most general sense in order to draw attention to the fact that fit assessment is possible in polytomous as well as dichotomous IRT models. However, because this research was primarily concerned with dichotomous IRT models, in subsequent sections *u* (as defined in Equation 1) will be used for notation rather than *x*.

The earliest statistics, proposed for the 1PL model, utilized group-level residuals (Wright & Panchapakesan, 1969; Wright & Mead, 1977). After these were proposed, individual-level indices were also utilized (Wright, Mead, & Bell, 1979). Group-level indices were also independently utilized outside of the 1PL literature (Bock, 1972; Yen, 1981). Early investigations of item model-data fit concerned the 1PL model and its group of fit statistics (Balla & McDonald, 1985; Divigi, 1986; Rogers & Hattie, 1987; Smith, 1991, 1994). Though the investigation of fit has moved on to more complicated models and procedures, investigations of 1PL model fit statistics persist (Wang & Chen, 2005), perhaps because of the more widespread use of the 1PL model in industry and education. Several studies have extended the application of 1PL model fit statistics to other models, including the 2PL and 3PL (Yen, 1981; McKinley & Mills, 1984), a graded unfolding model (DeMars, 2004), and the rating-scale model (Wang & Chen, 2005).

Item fit statistics that utilize the model likelihood function have also been proposed. Some methods use the likelihood function to compute the observed (Stone, 2000) or expected (Orlando & Thissen, 2000) frequencies used to compute the group-level fit indices. One method uses the first and second derivatives of the likelihood function to construct a Lagrange Multiplier test statistic (Glas, 1999; Glas & Falcon, 2003) and another uses the height of the likelihood function at the model parameter estimates to gauge fit (Reise, 1990).

Finally, Monte Carlo simulation has been utilized to assess fit. Because the theoretical distribution of the statistic resulting from Stone's (2000) method is unknown, critical values must be obtained from empirical sampling distributions; these sampling distributions are generated by Monte Carlo data simulation. A method called Posterior Predictive Model Checking (PPMC) has also been proposed, whereby critical values for any particular statistic are obtained from empirical sampling distributions of the statistic in the null case (Sinharay, 2005; Sinharay et al., 2006). PPMC uses the posterior distribution of item parameters when simulating data and thus takes item parameter estimation error into account. Fixed values of item parameters can also be used in generating sampling distributions, but little is known regarding the cost of this simplification.

Individual-Level Indices

Wright et al. (1979) utilized an *unweighted total* item fit statistic as the sum of the individual-level residuals for the *i*th item:

$$VO_{UT} = \sum_{J=1}^{N} \frac{[u_{ij} - P_i(\theta_j)]^2}{P_i(\theta_j)[1 - P_i(\theta_j)]} \tag{3}$$

To reduce the influence of unexpected rare aberrant responses that might occur in large datasets, Wright et al. also divided the sum of squared residuals by their expected variance, yielding the *weighted total* item fit statistic:

$$VI_{WT} = N \frac{\sum_{j=1}^{N} [u_{ij} - P_i(\theta_j)]^2}{\sum_{j=1}^{N} P_i(\theta_j)[1 - P_i(\theta_j)]} \tag{4}$$

The unweighted and weighted total fit statistics are referred to as Outfit and Infit respectively in 1PL model parameter estimation computer programs (Smith, 1994).

Outfit is the sum of squared standardized residuals and thus theoretically would be distributed as a χ^2 with N degrees of freedom (*df*). However, as noted by Rogers and Hattie (1987), the theoretical distribution relies on the binomial approximation to the normal distribution, which is clearly not tenable given that u_{ij} is a Bernoulli variable. The theoretical distribution of Infit is even less clear. Rogers and Hattie noted that "the

statistic appears to be a ratio of variance estimates; if the estimates were independent, this ratio would have an *F* distribution" (p. 48).

In the evaluation of item fit, it has been customary to either divide the statistics in Equations 3 and 4 by sample size (*N*) and evaluate as mean squares (MS) or evaluate as a standard normal (*z*) variable after a cube root transformation (Smith, 1991). The expected value of the mean squares is equal to 1.0 for both statistics and, using $w_{ij} = P_i(\theta_j)[1 - P_i(\theta_j)]$, the standard deviation of the mean squares is given by:

$$SD(\mathrm{VO}_{MS}) = \frac{\left[\sum_{j=1}^{N} w_{ij}^{-1} - 4N\right]^{1/2}}{N} \tag{5}$$

$$SD(\mathrm{VI}_{MS}) = \frac{\left[\sum_{j=1}^{N} w_{ij} - 4\sum_{j=1}^{N} w_{ij}^{2}\right]^{1/2}}{\sum_{j=1}^{N} w_{ij}} \tag{6}$$

Screening items based on the mean squares is difficult since their distribution depends on *N*. Transformations have been proposed in attempt to convert these mean squares to *z* variables. The transformation for both Outfit and Infit is given by (Smith, 1991):

$$V^* = \frac{3[(V^*_{MS})^{1/3} - 1]}{SD(V^*_{MS})} + \frac{SD(V^*_{MS})}{3} \tag{7}$$

where * is replaced by *O* for Outfit and *I* for Infit.

Past research. Smith (1991) found that the sampling distribution mean (μ_{s-dist}) and standard deviation (σ_{s-dist}) of *V*I and *V*O were affected by test length (*n*), *N*, and the difference between *b* and θ ($b - \theta$ offset). The effects of these factors appeared to be byproducts of the effects of parameter estimation error on the sampling distributions of the fit statistics. Sampling distributions were close to a *z* when fit statistics were calculated using true *b* and θ parameters but distorted when using estimated item ($\hat{b}$) and/or person ($\hat{\theta}$) parameters; error in $\hat{\theta}$ led to a reduction in both μ_{s-dist} and σ_{s-dist} whereas $\hat{b}$ estimation error caused a reduction only in σ_{s-dist}.

When fit statistics were computed using both $\hat{b}$ and $\hat{\theta}$ (as would be done in practice), μ_{s-dist} was quite negative and σ_{s-dist} was deflated for small *n,* but both appeared to approach their theoretical values as *n* increased. For *N*, μ_{s-dist} became increasingly more negative as *N* increased; this was less marked for large *n* than small *n*. μ_{s-dist} also became increasingly more negative as the range of *b* increased, but for the short ranges, where μ_{s-dist} were closer to zero, σ_{s-dist} was deflated. For $b-\theta$ offset, when *b* was centered about θ, μ_{s-dist} and σ_{s-dist} were deflated and μ_{s-dist} became increasingly negative as the offset increased for v_{Oz} but not v_{Iz}. There was no clear pattern in which σ_{s-dist} changed across levels of $b-\theta$ offset. Since all of these factors affect parameter estimation error, it is not surprising that they affected the null distributions of 1PL model fit statistics.

Wang and Chen (2005) also found some effect for *n*, *N*, and *b* on the sampling distributions of *V*I and *V*O. However, the effects that they found were much smaller than those found by Smith (1991). μ_{s-dist} was only slightly negative and fluctuated very little with *n* and *N*. σ_{s-dist} was deflated and tended to be somewhat smaller for *V*I than *V*O. Contrary to Smith (1991) there was a slight tendency for σ_{s-dist} to decrease as *n* increased. σ_{s-dist} was near unity for items with *b* near zero and decreased as *b* became extreme; this effect was more pronounced for *V*I than for *V*O.

The discrepancy in results between Smith (1991) and Wang and Chen (2005) is perplexing. For example, *V*O μ_{s-dist} and σ_{s-dist} for a 20-item test with *N* from 100 to 2,000 ranged from −0.05 to −0.49 and 0.86 to 1.07, respectively, in Smith's (1991) study and from −0.02 to −0.07 and 0.81 to 0.91, respectively, in Wang and Chen's study. Smith (1991) used only 10 replications whereas Wang and Chen used 500. This would of course affect the results, but it might be expected that Smith's (1991) results would be more erratic and not exhibit the definite trends summarized above. Smith (1991) also used different sets of item parameters than Wang and Chen, generating then from a U(−1,1)

distribution whereas Wang and Chen generated them from a N(0,1) distribution. This could account for the difference in results between the two studies.

Using critical values suggested by Wright and his colleagues, for a 15-item test with $N = 500$, Rogers and Hattie (1987) found that in the null case *V*O exhibited inflated Type I error and *V*I exhibited somewhat deflated Type I error. Smith (1991) showed that μ_{s-dist} for *V*O tended to be more extreme (and negative) than *V*I for shorter (10-item) tests but not longer tests. Perhaps the difference in Type I error rates between *V*I and *V*O found by Rogers and Hattie was due to a relatively short test.

Several studies have examined the functionality of *V*I and *V*O for polytomous models. DeMars (2004) examined these fit statistics for a graded unfolding model. Across a range of *n* (including short 10-item tests) and *N* conditions, *V*I and *V*O Type I errors were much lower than nominal levels averaging (at $\alpha = 0.05$) 0.0001 for *V*I and 0.002 for *V*O. Correlations between item parameters and fit measures were also examined, and negative correlations between discrimination and both *V*I and *V*O were found.

Wang and Chen (2005) examined the functionality of *V*I and *V*O for a rating-scale model in addition to their examination of the statistics for the 1PL model. The general pattern of results was similar to that for the 1PL model; however, μ_{s-dist} tended to be more negative for the polytomous model than for the 1PL model and σ_{s-dist} appeared to *increase* (rather than decrease as it did for the 1PL) somewhat (toward unity) as *n* increased. But *n* conditions were different between the two models. However, both did have a 20-item condition and in this case σ_{s-dist} was closer to unity for the rating-scale model than for the 1PL model. Rating-scale model *V*I and *V*O μ_{s-dist} was more sensitive to changes in N than for the 1PL model.

Tang (1994) applied VI_{MS} and VO_{MS} to assess item fit at the item category level for a short 10-item partial-credit model test. Using critical values of > -1.2 and ≤ 0.8 (note that these are the mean square versions of the statistics) to identify misfitting items, Type I errors were found to be very high for VO_{MS} relative to VI_{MS}, ranging from

< 0.001 to 0.008 for VI_{MS} and from 0.02 to 0.88 for VO_{MS}. A Similar pattern of results was found by Rogers and Hattie (1987) using the 1PL and the *z*-transformed statistics.

Both Smith (1991) and Wang and Chen (2005) proposed corrections for *V*I and *V*O to restore their sampling distributions to the *z* distribution. Smith (1991) found that, after applying the correction both statistics were much closer to their theoretical sampling distributions and, contrary to the uncorrected versions, had power to detect guessing and startup effects. The corrections to *V*I and *V*O might work well in practice, but the generalizability of these corrections across a range of study conditions and test configurations has not yet been demonstrated.

Group-Level Indices

Most of the various group-level indices that have been proposed have the same form, but differ in the way in which examinees are grouped and expected values (E_{ik}) are calculated. The general form of these statistics is:

$$Q_{BU} = \sum_{k=1}^{K} n_k \frac{[p_{ik} - \pi_{ik}]^2}{\sum_{j \in k} [\pi_{ij}(1-\pi_{ij})]/n_k} = \sum_{k=1}^{K} n_k \frac{[p_{ik} - \pi_{ik}]^2}{\pi_{ik}(1-\pi_{ik}) - \sigma^2_{P_k}} \quad (8)$$

where:

$$\pi_{ik} = E_{ik} / n_k,$$

$$\pi_{ij} = P_i(\hat{\theta}_j),$$

$$p_{ik} = \frac{1}{n_k} \sum_{j \in k} u_{ij}, \text{ and}$$

$$\sigma^2_{P_k} = \frac{1}{n_k} \sum_{j \in k} [\pi_{ij} - \pi_{ik}]^2.$$

For 1-, 2-, and 3PL dichotomous models, $P_i(\hat{\theta}_j)$ can be obtained by substituting model parameter estimates in Equation 1. The statistic in Equation 8 is referred to as an unweighted *between-group* item fit statistic and is distributed as a χ^2 with $K - v$ *df*, where *v* is equal to the number of estimated item parameters.

Wright and Mead's (1977) version of this statistic (*Q*WM) was proposed for use with the 1PL model. To compute *Q*WM, examinees are first grouped according to

number-correct score *t* and the expected values E_{ik} for each group are given by $\sum_{t \in k} n_t P_i(\hat{\theta}_t)$. This statistic includes an adjustment for the expected variance of the predicted probabilities of correct response (McKinley & Mills, 1985; Yen, 1981), which is given by $\sigma^2_{P_k}$ in Equation 8. Though originally proposed for the 1PL model, *Q*WM has been extended for use with the 2PL and 3PL by grouping on $\hat{\theta}$ (Yen, 1981).

For Bock's (1972) and Yen's (1981) version of the statistic (*Q*B and *Q*1 respectively), examinees are classified into groups according to $\hat{\theta}$; Yen (1981) specifies that ten ability groups be used whereas Bock does not make any particular specification. The expected values for Bock's method are given by $n_k P_i(\hat{\theta}_{\text{med}-k})$, where med-*k* refers to the median $\hat{\theta}$ in the *k*th group. The expected values for Yen's method are given by $\sum_{j \in k} P_i(\hat{\theta}_j)$. Additionally, these methods do not include the term $\sigma^2_{P_k}$, though Yen's (1991) extension of *Q*WM to the 2PL and 3PL was equivalent to a version of *Q*1 including $\sigma^2_{P_k}$.

When computing *Q*1 and *Q*B, it is recommended that examinees be divided into quantiles based on $\hat{\theta}$ in order to create groups in which n_k are roughly equal across *k* (Bock, 1972; Yen, 1981). Because expected values of < 5 can disrupt the theoretical sampling distribution of these statistics, it has been recommended that adjacent categories be collapsed when expected values are too low (Orlando & Thissen, 2000), as might be the case with unusually easy or difficult items. Thus, *K* can vary across items. However, for notational simplicity *K* will not be subscripted with *i* here.

Within the context of the 1PL model, a weighted version of the between-group item fit statistic has also been used (Wright et al., 1979). This statistic, expressed as a mean square, is given by:

$$V\text{WB}_{MS} = \frac{K}{K-1} \frac{\sum_{k=1}^{K} \left(\sum_{j \in k} u_{ij} - \sum_{j \in k} \pi_{ij} \right)^2}{\sum_{k=1}^{K} \sum_{j \in k} w_{ij}} \tag{9}$$

QWM can also be divided by its *df* to form an unweighted between-group mean square, $V\text{UB}_{MS}$. As with Infit and Outfit, the magnitude of these mean squares can be used to assess fit, or the cube root transformation in Equation 7 can be applied and the resulting statistic can be tested for significance against a z distribution. The expected values of both $V\text{UB}_{MS}$ and $V\text{WB}_{MS}$ are 1.0 and their standard deviations are:

$$SD(V*_{MS}) = \left[2/(K-1)\right]^{1/2} \tag{10}$$

where * is replaced by *UB* for the unweighted between statistic and *WB* for the weighted between statistic.

$Q1$ and QB can also be formulated as likelihood ratio statistics, which are also distributed as χ^2 with $K - v$ degrees of freedom. This statistic is given by:

$$G^2 = 2\sum_{k=1}^{K}\left[p_{ik} \ln\left(\frac{p_{ik}}{\pi_{ik}}\right) - (1-p_{ik}) \ln\left(\frac{1-p_{ik}}{1-\pi_{ik}}\right)\right] \tag{11}$$

The likelihood ratio versions of $Q1$ and QB will be referred to as $G1$ and GB respectively.

Past research. Whereas fit statistics utilizing person-level model residuals tend to have deflated Type I error, fit statistics based on group-level model residuals tend to exhibit inflated Type I error, except for the case of $V\text{BU}_z$ and $V\text{BW}_z$ which have been shown to have Type I error near or below nominal levels (Rogers & Hattie, 1987; Smith, 1994). Smith (1994) found that μ_{s-dist} and σ_{s-dist} for $V\text{BU}_z$ and $V\text{BW}_z$ were distorted (> 0.0 and < 1.0 respectively) when three ability groups (K) were used but approached expectation (0.0 and 1.0 respectively) as K increased. The values of μ_{s-dist} were somewhat larger for $V\text{BU}_z$ than $V\text{BW}_z$. As N increased, μ_{s-dist} and σ_{s-dist} did not approach their expectations – if anything, they became more aberrant, especially for small n. As n increased, μ_{s-dist} and σ_{s-dist} approached expectation. For both fit statistics, μ_{s-dist} was negative (around –0.20) when there was little dispersion in b parameters and increased to positive values as b dispersion increased. There was no clear pattern for $b - \theta$ offset on the sampling distributions of $V\text{BU}_z$ and $V\text{BW}_z$.

Only two studies could be found that directly examined the distributional properties of the χ^2 group-level fit statistics. Stone and Hansen (2000) examined the sampling distributions of *Q*B and *G*B for a polytomous IRT model across two sample sizes (N = 1,000 and 2,000) and three test lengths (n = 8, 12, and 32). When evaluated with θ, the fit statistics approximated their theoretical sampling distribution fairly closely, though μ_{s-dist} and σ_{s-dist} tended to be somewhat inflated. However, when evaluated with $\hat{\theta}$, the distributions were highly distorted unless n = 32, in which case they were similar to the distributions when θ was used. In general, even when θ was used, the sampling distribution of *G*B was more aberrant than that of *Q*B. Yen (1984) also examined the null distribution of *Q*1 for dichotomous IRT models. Though the investigation showed that *Q*1 was distributed as χ^2 with the appropriate number of *df* for 20- and 40-item tests with 1,000 examinees, the assessment was based on only one replication.

Consistent with the sampling distribution distortions found by Stone and Hansen (2000), the χ^2 group-level fit statistics have been shown to exhibit inflated Type I error rates, especially for small n. Glas and Falcon (2003) found that *Q*1 and *G*1 were sensitive to changes in both N and n. Type I errors (at $\alpha = 0.05$) for *G*1 applied to the 3PL were near 0.5 and 1.0 for 10-item tests with N = 500 and N = 4,000 respectively. For 40-item tests, Type I error was deflated for *G*1 when N = 500 and N = 1,000 but inflated (at 0.3) when N = 4,000. Using the 2PL, Stone and Zhang (2003) also found that *Q*B was sensitive to changes in N and n. Type I error rates increased as N increased and n decreased; Type I error (at $\alpha = 0.05$) was near 1.0 for 10-item tests, and for 40-item tests was 0.06 when N = 500 and 0.32 when N = 2,000.

Using N = 1,000 and 1-, 2-, and 3PL models, Orlando and Thissen (2000) found *Q*1 and *G*1 Type I error rates (at $\alpha = 0.05$) near 1.0 for 10-item tests, between 0.20 and 0.30 for 40-item tests, and between 0.06 and 0.18 for 80-item tests. In general, the Type I error was somewhat more inflated for *G*1 than for *Q*1 and it tended to increase somewhat with model complexity. For a 3PL 50-item test with N = 1,000, Reise (1990) reported

Type I error rates between 0.02 and 0.10 (at $\alpha = 0.01$) and Dodeen (2004) reported Type 1 error rates between 0.09 and 0.28 (at $\alpha = 0.01$) for *Q*B. Dodeen's study conditions were later replicated using *G*B (Sinharay & Lu, 2008), with Type I error rates (at $\alpha = 0.01$) ranging from 0.02 to 0.43. Dodeen, Sinharay and Lu, and Reise varied *a*, *b*, and *c* parameters but only the former two studies found a relationship between item parameters and fit statistics. Perhaps this difference in results is due to Reise's use of uniform distributions for *b* and fixed (within tests) *a* rather than normal distributions for *b* and normal or lognormal distributions for *a* as used in the other two studies. Dodeen found that *Q*B Type I error increased with *a* and *c* whereas Sinharay and Lu found that *G*B Type I error increased only with *a*. Dodeen also found that *Q*B Type I error rates were somewhat higher when *b* was centered about θ (a result also found by McKinley & Mills, 1985), but Sinharay and Lu did not find this result for *G*B. DeMars (2004) (using a polytomous IRT model and normal distribution of θ) also found a positive correlation between *a* and *Q*1 and *G*1 item fit statistics.

McKinley and Mills (1985) compared several versions of the group-level fit statistics for 75-item tests across three IRT models (1-, 2-, and 3PL), three sample sizes (N = 500, 1,000, and 2,000), and three normally distributed θ conditions (low, midrange, and high). Type I errors (at $\alpha = 0.01$) ranged from < 0.01 to 0.08 and, unlike Orlando and Thissen (2000), tended to be lower and somewhat deflated for *G*1 than for the other statistics. Type I error rates were comparable between *Q*1, *Q*B, and *Q*WM. For the 1PL, but not the other models, Type I error tended to be higher when θ was centered at *b*, as opposed to when it was high or low. Type 1 error tended to be somewhat higher for larger *N*; for example, *Q*1 Type I error (at $\alpha = 0.01$) was 0.02 for N = 500 and 0.04 for N = 2,000. This effect was less marked than that found by Glas and Falcon (2003) whose longest test length was 40 items. It is likely that there is an interaction between *N* and *n* such that the effect of *N* on Type I error rates decreases as test length increases.

DeMars (2005) examined Type I error rates (with N = 1,000) for a *G*1 item fit statistic applied to polytomous items. When θ was normally distributed, Type I errors (at $\alpha = 0.05$) for a 10-item test were 0.14 and 0.30 for the two models examined; when θ

was uniform, Type I error rates were higher at 0.30 and 0.46. Type I error was near nominal levels for a 20-item test, though somewhat inflated when θ was uniformly distributed. Very little is known regarding the functionality of these fit statistics for polytomous as opposed to dichotomous models. DeMars' (2005) results suggest that polytomous fit indices are less affected by Type I error inflation for short tests. However, the *df* of the fit statistics used by DeMars (2005) were not adjusted for the estimated item parameters, as is typically done for the χ^2 and G^2 indices; had *df* been used in the same manner as other studies, the Type I error would have been higher.

DeMars (2004) also examined the functionality of a $Q1$ and $G1$ statistic for generalized graded unfolding models. Type I error for the $Q1$ index was much higher than for $G1$; for both indices, Type I error was positively correlated with N, and negatively correlated with K and n. The $Q1$ index was never near nominal levels, but the $G1$ was for $N = 500$ or 1,000, $K = 20$, or $n = 20$ or 30.

The traditional group-level fit indices appear to be sensitive to θ estimation error and thus not useful for assessing fit unless n is large. For large n, θ is estimated reasonably well and the fit statistic theoretical sampling distributions appear to remain relatively unperturbed (at least if most of the items fit the model). It would, therefore, appear that practitioners should, in many instances, abandon use of these group-level indices. However, there is some evidence that this *type* of fit index captures different aspects of fit than other types, such as individual-level indices and those designed specifically to detect violations of local independence (Balla & McDonald, 1985; Smith, 1991, 1994; Yen, 1984). It thus might be advantageous to keep group-level fit indices in the toolbox. But in order to so, their dependence on point estimates of θ must be reduced. Several methods, based on the model likelihood function, have been proposed to reduce this dependence.

Indices Utilizing the Model Likelihood Function

The likelihood of a persons × items data matrix $\mathbf{u_{ij}}$ is given by:

$$L(\mathbf{u_{ij}} \mid \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{N} P_i(\theta_j)^{u_{ij}} \left[1 - P_i(\theta_j)\right]^{(1-u_{ij})} \tag{12}$$

where $\boldsymbol{\xi} = (\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_n})$ is the set of item parameters (i.e., for the 2PL $\boldsymbol{\xi_i} = \{a_i, b_i\}$) and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ contains the person parameters. This persons × items data matrix is comprised of person response vectors ($\mathbf{u_{.j}}$) and item response vectors ($\mathbf{u_{i.}}$). Several fit indices utilize the likelihood functions of these response vectors, which are given by:

$$L(\mathbf{u_{.j}} \mid \boldsymbol{\xi}, \theta_j) = \prod_{i=1}^{n} P_i(\theta_j)^{u_{ij}} \left[1 - P_i(\theta_j)\right]^{(1-u_{ij})} \tag{13}$$

$$L(\mathbf{u_{i.}} \mid \boldsymbol{\xi_i}, \boldsymbol{\theta}) = \prod_{j=1}^{N} P_i(\theta_j)^{u_{ij}} \left[1 - P_i(\theta_j)\right]^{(1-u_{ij})} \tag{14}$$

In the marginal maximum likelihood (MML) item parameter estimation procedure, the person parameters are treated as unknown parameters sampled from some particular distribution $g(\theta \mid \boldsymbol{\tau})$, where $\boldsymbol{\tau}$ is the set of parameters that defines the distribution (e.g. $\boldsymbol{\tau} = \{\mu, \sigma\}$ for the normal distribution). In this case $L(\mathbf{u_{.j}})$ is given by:

$$L(\mathbf{u_{.j}} \mid \boldsymbol{\xi}, \boldsymbol{\tau}) = \int_{\theta=-\infty}^{\infty} L(\mathbf{u_{.j}} \mid \boldsymbol{\xi}, \theta) g(\theta \mid \boldsymbol{\tau}) \, \partial\theta \tag{15}$$

The *j* subscript has been removed from θ indicating that θ is considered an unknown value randomly sampled from some distribution. Orlando and Thissen (2000) used this function to compute π_{ik}. In their approach, examinees are divided into score groups based upon their number-correct scores and the likelihood of observing each number-correct score $k = 0, 1, \ldots, n$ is then found as:

$$S(k \mid \hat{\boldsymbol{\xi}}, \theta) = S_k = \sum_{p=1}^{w_k} L(\mathbf{u_{.p}} \mid \hat{\boldsymbol{\xi}}, \theta) = \sum_{p=1}^{w_k} \prod_{i=1}^{n} P_i(\theta)^{u_i} [1 - P_i(\theta)]^{(1-u_i)} \tag{16}$$

where

$w_k = \dfrac{n!}{k!(n-k)!}$ is the number of possible distinct response patterns p for number-correct score k,

$\mathbf{u}_{.\mathbf{p}}$ is the pth response vector (across the n items) in set w_k, and

$\hat{\boldsymbol{\xi}}$ is the set of item parameter estimates.

The π_{ik} are then found by the following ratio:

$$\pi_{ik} = \frac{\int_{\theta=-\infty}^{\infty} P_i(\theta) S_{k-1}^{*i} g(\theta \mid \boldsymbol{\tau}) \partial\theta}{\int_{\theta=-\infty}^{\infty} S_k \, g(\theta \mid \boldsymbol{\tau}) \partial\theta} \tag{17}$$

where S_{k-1}^{*i} is the likelihood of number-correct score $k - 1$ excluding the ith item (the item for which fit is being assessed). Numerical quadrature is then used to approximate Equation 17 by:

$$\pi_{ik} \equiv \frac{\sum_{s=1}^{S} P_i(\theta_s) S_{(k-1)s}^{*i} A(\theta_s)}{\sum_{s=1}^{S} S_{ks} A(\theta_s)} \tag{18}$$

This method can be used in to obtain the π_{ik} for use in both the χ^2 and G^2 tests described above and defined in Equations 8 and 11; these will be referred to as *Q*O and *G*O respectively. Orlando and Thissen (2000) computed *Q*O and *G*O by summing across the $K - 2$ number-correct score groups, $k = 1, \ldots, n - 1$. They also employed a procedure to collapse across cells if expected frequencies were too low.

An alternative method for grouping examinees and obtaining the p_{ik} used in group-level fit statistics has been proposed by Stone, Ankenmann, Lane, and Liu (1993). This method utilizes the Bayesian posterior distribution of θ conditioned on the observed data and item parameter estimates:

$$P(\theta \mid \mathbf{u}_{.\mathbf{j}}, \hat{\boldsymbol{\xi}}, \boldsymbol{\tau}) = \frac{L(\mathbf{u}_{.\mathbf{j}} \mid \hat{\boldsymbol{\xi}}, \theta) g(\theta \mid \boldsymbol{\tau})}{L(\mathbf{u}_{.\mathbf{j}} \mid \hat{\boldsymbol{\xi}}, \boldsymbol{\tau})} \tag{19}$$

Other methods use point estimates from this distribution, but Stone's method essentially uses the posterior distribution of θ to distribute expected probabilities across K groups

defined by the discrete quadrature points of θ used to approximate Equation 19. This quadrature approximation is given by:

$$P(\theta_k \mid \mathbf{u}_{.\mathbf{j}}, \hat{\boldsymbol{\xi}}, \boldsymbol{\tau}) = \frac{L(\mathbf{u}_{.\mathbf{j}} \mid \hat{\boldsymbol{\xi}}, \theta_k) A(\theta_k)}{\sum_{k=1}^{K} L(\mathbf{u}_{.\mathbf{j}} \mid \hat{\boldsymbol{\xi}}, \theta_k) A(\theta_k)} \tag{20}$$

The observed number-correct for the *k*th group, O_{ik}, is then found by:

$$O_{ik} = \sum_{j=1}^{N} u_{ij} P(\theta_k \mid \mathbf{u}_{.\mathbf{j}}, \hat{\boldsymbol{\xi}}, \boldsymbol{\tau}) \tag{21}$$

These are actually the pseudocounts that are computed in the E step of the EM algorithm used in MML item parameter estimation (Baker & Kim, 2004). The observed number of examinees at each of the *K* points is found by:

$$n_{ik} = \sum_{j=1}^{N} P(\theta_k \mid \mathbf{u}_{.\mathbf{j}}, \hat{\boldsymbol{\xi}}, \boldsymbol{\tau}) \tag{22}$$

Dividing Equation 21 by Equation 22 yields p_{ik}. The value for π_{ik} is obtained by evaluating the IRF at each of the *K* quadrature points. These π_{ik} and p_{ik} can then be used to compute the χ^2 and G^2 statistics in Equations 8 and 11. Stone's versions of these statistics will be referred to as *Q*S and *G*S.

A drawback of Stone's procedure is that the resulting *Q*S and *G*S cannot be assumed to follow a χ^2 distribution since the cell data are no longer independent. In order to conduct significance tests for Stone's fit indices, Monte Carlo sampling is used to determine the empirical sampling distribution of the test statistics. Stone's method involves the following steps:

1. Simulate data using item parameter estimates from the calibration on the real data,
2. Calibrate the simulated data,
3. Calculate fit statistics using parameter estimates from the simulated data calibration,
4. Repeat steps 1– 3 *R* times; *R* = 1,000 in Stone et al. (1993).

Stone (2000) reported that the statistics appear to be distributed as a scaled χ^2. The scaling factor γ and *df* for this distribution can be estimated directly from the mean and variance of the empirical sampling distribution using the method of moments. As described by Stone (2000), first consider a scaled χ^2 statistic C_s (e.g. either *Q*S or *G*S): $C_s \sim \gamma C$ where C is distributed as a χ^2 with *v df*.

Since $E(C) = v$ and $Var(C) = 2v$ then

$E(C_s) = E(\gamma C) = \gamma E(C) = \gamma v$ and

$Var(C_s) = Var(\gamma C) = \gamma^2 Var(C) = 2v\gamma^2$.

The mean and variance from the empirical sampling distribution of C_s can be used for $E(C_s)$ and $Var(C_s)$, respectively, and the two equations can then be solved for γ and *v*. A significance test can then be conducted by evaluating C_s / γ as a χ^2 with *v df*.

Stone (2000) proposed using a resampling method to obtain estimates of γ and *v*. The resampling method simply involves skipping step 2 above and conducting step 3 using the item parameter estimates from the real data; thus, this method does not take into account uncertainty in item parameter estimation. Stone (2000) showed that this method functions very similarly to the original (full sequence of steps 1 – 4 above) method but has slightly less power. The χ^2 and G^2 fit statistics computed using the resampling method will be referred to as *Q*S* and *G*S*, respectively.

Glas (1999) proposed a Lagrange Multiplier (LM) test statistic to assess item fit. This statistic gauges the degree to which the item likelihood function is restricted by imposing constraints on model parameters. In this procedure, a general form of the IRF is posited in which item parameters are allowed to vary across score levels. For example, a 2PL model in which item parameters vary across *K* examinee groups, could be proposed as

$$P_{ik}(u_{ij} = 1 \mid j \in k, \theta, \boldsymbol{\xi}_i) = \frac{1}{1 + \exp(-z_{ik})} \tag{23}$$

where

$k = 1, 2, \ldots, K$ groups,

$z_{ik} = (a_i + \alpha_{ik})[\theta - (b_i + \beta_{ik})]$, and

$\alpha_{iK} = \beta_{iK} = 0$ in order to identify the model.

The parameters of such a model are partitioned into a set of free parameters ($\boldsymbol{\xi}_1$) and a set of fixed parameters ($\boldsymbol{\xi}_2$). In the 2PL example given above, $\boldsymbol{\xi}_1 = (a_i, b_i)$ and $\boldsymbol{\xi}_2 = (\alpha_{i1}, \ldots, \alpha_{ik}, \beta_{i1}, \ldots, \beta_{ik})$. The composite null hypothesis tested by the procedure is that $\alpha_{ik} = 0$ and $\beta_{ik} = 0$ for $k = 1, \ldots, K-1$. The LM statistic used to test this hypothesis is given by:

$$\mathrm{LM}(\alpha\beta) = \mathbf{h}(\boldsymbol{\xi}_2)' \mathbf{W}^{-1} \mathbf{h}(\boldsymbol{\xi}_2) \tag{24}$$

where

$\mathbf{W} = \mathbf{H}_{22} - \mathbf{H}_{21}\mathbf{H}_{11}^{-1}\mathbf{H}_{12}$,

$$\mathbf{h}(\boldsymbol{\xi}_2) = \frac{\partial \ln L}{\partial \boldsymbol{\xi}_2} = \sum_{j=1}^{N} E\left[\mathbf{b}_j(\boldsymbol{\xi}_2) \mid \mathbf{u}_j, \boldsymbol{\xi}\right],$$

$$\mathbf{b}_j(\boldsymbol{\xi}_2) = \frac{\partial \ln L(\mathbf{u}_j, \theta_j \mid \boldsymbol{\xi})}{\partial \boldsymbol{\xi}_2},$$

$\mathbf{H}_{pq} = \dfrac{\partial^2 \ln L}{\partial \boldsymbol{\xi}_p \boldsymbol{\xi}_q'}$ for p = 1,2 and q = 1,2, and

$\ln L$ is the log of the item likelihood function for the general model.

Any cross-group elements in the computation of these matrices are equal to zero. For example, letting k and l index the K score groups, $E\left[\mathbf{b}_{jk}(\boldsymbol{\xi}_{2l}) \mid \mathbf{u}_{jk}, \boldsymbol{\xi}\right] = 0$. The vector $\mathbf{h}(\boldsymbol{\xi}_2)$ and matrices $\mathbf{H}_{pq}(\boldsymbol{\xi})$ are of order v, where v equals the number of elements in $\boldsymbol{\xi}_2$. For the 2PL example above, $v = 2K$. Models can also be tested in which only *a* or *b* parameters are allowed to vary. For example, to test the simple null hypotheses that $\alpha_{ik} = 0$ for $k = 1, \ldots, K-1$, an LM(α) statistic would be defined by Equation 24 using $\boldsymbol{\xi}_2 = (\alpha_{i1}, \ldots, \alpha_{ik})$; to test the null hypothesis that $\beta_{ik} = 0$ an LM(β) statistic would be constructed in the same manner, but with $\boldsymbol{\xi}_2 = (\beta_{i1}, \ldots, \beta_{ik})$. According to Glas, The LM

statistics defined by Equation 24 theoretically follow a χ^2 distribution with $df = v - f$, where f is the number of parameters that are fixed in order to identify the model.

Reise (1990) used the height of the likelihood function evaluated at the maximum likelihood item and person parameter estimates ($\hat{\boldsymbol{\xi}}_{\mathbf{i}}$ and $\hat{\boldsymbol{\theta}}$ respectively) to assess item fit. This index is given by

$$L_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}} \tag{25}$$

where

$$l_0 = \ln L(\mathbf{u}_{\mathbf{i.}} \mid \hat{\boldsymbol{\xi}}_{\mathbf{i}}, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{N} u_{ij} \ln[P_i(\hat{\theta}_j)] + (1 - u_{ij}) \ln[1 - P_i(\hat{\theta}_j)] \quad \text{and}$$

$$Var(l_0) = \sum_{j=1}^{N} P_i(\hat{\theta}_j)[1 - P_i(\hat{\theta}_j)] \left\{ \ln \left[\frac{P_i(\hat{\theta}_j)}{[1 - P_i(\hat{\theta}_j)]} \right] \right\}^2 .$$

L_z was originally proposed by Dragow, Levine, and Williams (1985) to assess person fit. The person fit index is of the same form as Equation 25 but, for each person, is applied to the person response vector ($\mathbf{u}_{\mathbf{j}}$) rather than the item response vector ($\mathbf{u}_{\mathbf{i.}}$). Reise (1990) expanded L_z to assess item fit by applying it to item response vectors.

The statistic L_z, whether used to assess item or person fit, is asymptotically distributed as a standard normal variable. Negative values indicate that the observed response vector ($\mathbf{u}_{\mathbf{i.}}$ or $\mathbf{u}_{\mathbf{j}}$) is less consistent than the model predicts and positive values indicate that the response vector is more consistent than predicted by the model.

While the promise of L_z for assessing item fit remains relatively unknown, research indicates that person fit L_z departs from a standard normal distribution when computed with estimated model parameters (Molenaar & Hoijtink, 1990; Nering, 1995; Reise, 1995; Seo & Weiss, 2009; Snijders, 2001; Van Krimpen-Stoop & Meijer, 2001). A review of the L_z person fit literature is outside the scope of this paper; for such a review see Ro (2001) or Seo and Weiss.

Past research. A handful of studies could be found systematically examining the functioning of the fit statistics that utilize the model likelihood function. There have been

two studies by Orlando and Thissen (2000, 2003) in which *Q*O/*G*O were compared with *Q*1/*G*1; three studies involving Stone's method (Roberts, 2003; Stone, 2003; Stone & Zhang, 2003); one study by Glas and Falcon (2003) in which LM, *Q*O, *G*O, *Q*1, and *G*1 were compared; and one study by Reise (1990) in which L_z was compared with *Q*B. The indices utilizing the model likelihood function generally outperformed the traditional indices in the studies that compared the two types.

Orlando and Thissen (2000) compared *Q*O/*G*O and *Q*1/*G*1 for the 1-, 2- and 3PL across three test lengths (n = 10, 40, and 80) with N = 1,000. Type I error remained closest to nominal levels for *Q*O (between 0.04 and 0.07 at $\alpha = 0.05$), tended to increase somewhat as n increased for *G*O (up to 0.13 at $\alpha = 0.05$), was grossly inflated for *Q*1/*G*1 on short tests (in the 0.90s at $\alpha = 0.05$), and still somewhat above nominal levels for the long test (between 0.06 and 0.18). Type I error also appeared to increase somewhat as the number of model item parameters (or degree of model complexity) increased, but the effect was more marked for *Q*1/*G*1 than *Q*O/*G*O. The authors also examined the moments from the null conditions to assess how well *Q*O and *G*O followed their theoretical distributions. As might be expected, based on the patterns of Type I error rates mentioned above, *Q*O was closer to a χ^2 than *G*O and better approximated a χ^2 as n increased.

The power of *Q*O to detect several types of misfit has also been examined (Orlando & Thissen, 2000, 2003). First, in 2000, *Q*O was assessed in instances where the calibrating model had fewer parameters than the generating model. *Q*O had difficulty identifying misfit due to calibrating 3PL data with the 2PL model (power ranged from 0.11 to 0.13) but had reasonable power to detect misfit when 2PL/3PL items were calibrated with the 1PL model (power ranged from 0.46 to 0.58). Test length did not appear to markedly affect *Q*O power. Finally, when 2PL or 3PL items were calibrated with the 1PL, power varied as a function of generating model parameters. As might be expected, power was lowest (between 0.20 and 0.30) for items with midrange *a*s and highest for items with more extreme *a*s (in the 0.70s and 0.80s for the most extreme *a*s).

For items with lower *a*s, power was higher for difficult than easy items; this pattern was reversed for items with higher *a*s.

In 2003, Orlando and Thissen examined *Q*O in instances where the generating model was one of three nonlogistic forms: nonmonotonic, upper asymptote < 1, and plateau at middle θ but logistic at lower and higher θ. Data were simulated such that one "bad" item of each type was on an exam at a time; thus, false positive rates could be assessed in addition to power. The same test lengths as the 2000 study were used but, unlike the previous study, only the 3PL was used to calibrate the items and three levels of sample size were used (N = 500, 1,000, and 2,000).

The results were presented with receiver operating curves, which illustrate the ratio of true to false positive rates. In the plots, false positive rates between 0.01 and 0.10 were displayed and it was unclear whether rates outside of this range were observed. Across the three bad item types, power increased with *N* and *n*. Performance was better for *Q*O than *Q*1, except when *n* = 80 and *N* = 500 or 1,000. Thus, for long tests and smaller sample sizes it might be better to use *Q*1. Strictly in terms of *Q*O performance, when *N* = 2,000, power was consistently near 1.0 except for the second and third type of bad items when *n* = 10. For the smaller sample sizes, power and false positive rates were positively related.

Instead of conducting significance tests, misfitting items can also be identified from the magnitude of their χ^2/df ratio. Orlando and Thissen (2003) also presented results in terms of the proportion of simulations in which the largest χ^2/df corresponded with the misfitting item. When *N* = 500 and *n* = 10, the misfitting items were identified 56% of the time using *Q*O and only 8% of the time using *Q*1. However, when *n* = 80 (for the same *N*), the detection rates were 19% for *Q*O and 29% for *Q*1. When *N* = 2,000, detection rate increased with test length and *Q*O performed more favorably at the long test length; when *n* = 10, detection rates for *Q*O and *Q*1 were 66% and 12% and when *n* = 80 they were 94% and 61% respectively.

Stone and Zhang (2003) compared Type I error rates and power to detect model misspecifications for *Q*S*, *Q*O, and *Q*B across three sample sizes (*N* = 500, 1,000, and

2,000) and test lengths (n = 10, 20, and 40). Type I error rates were near nominal levels for QS* and QO across all conditions and highly inflated for QB in all conditions other than $N = 500/n = 40$. In general, QS* exhibited more power than QO to detect misfit due to the use of improper item parameters; this was more marked for smaller N. However, it is unknown whether the greater power for QS* comes at a cost of a higher false positive rate. The authors did not present any results regarding false positive rates, though they were clearly available since one of the misfit conditions involved alterations to item parameters for only two items on each test.

To date, the other available studies examining Stone's method have used only polytomous models. Using a graded unfolding model, Roberts (2003) examined Type I error and power to detect misfit when the generating model was more complicated than the calibrating model. Both QS and QS* were used. Across a range of n and N conditions, Type I error was near nominal levels for QS and somewhat deflated for QS*. Both statistics had adequate power (in the 0.80s to 0.90s) to detect misfit due to the use of incorrect item parameters.

Stone (2003) examined Type I error rates and power for QS* applied to graded-response items across two test length conditions (n = 6 and 12), three sample size conditions (N = 500, 1,000 and 2,000) and two θ distribution conditions (normal and positively skewed). Misfit was created by altering the parameters used to assess fit: items were altered such that there was only one misfitting item, with either incorrect discrimination or difficulty parameters, on the test at a time. When all items on the test fit the model, Type I errors were at or near nominal levels when θ was normally distributed, though they tended to be a bit higher for n = 6 than n = 12. When θ was skewed, Type I errors were inflated; for example, with n = 12 and N = 1,000 Type I error (at $\alpha = 0.05$) was 0.16. Power was not examined when θ was skewed due to the inflated Type I error rates in the null case.

QS* had more power to detect aberrations in difficulty than in discrimination. In general, power increased with N and was quite a bit lower for n = 6 than n = 12. False positive rates were fairly close to their nominal levels; for n = 6 and n = 12, average false positive/power rates (at $\alpha = 0.05$, in percentages) were 7.4/74.0 and 6.2/84.9 respectively.

These rates are more favorable than those reported by Glas and Falcon (2003) for the LM and *G*O, but the study conditions upon which Glas and Falcon's and Stone's (2003) results are based are too disparate to allow firm conclusions regarding relative superiority of methods.

Glas and Falcon (2003) compared LM(β), *Q*O/*G*O, and *Q*1/*G*1 for the 3PL across three test lengths (n = 10, 20, and 40) and three sample sizes (N = 500, 1,000, and 4,000) and found that Type I error rates were much closer to nominal levels for the LM(β) and *Q*O/*G*O statistics than for *Q*1/*G*1. Type I error rates (at $\alpha = 0.05$) ranged from 0.04 to 0.09 for LM(β) with K = 5, 0.04 to 0.11 for LM(β) with K = 9, and 0.05 to 0.08 for *G*O (exact results were reported only for the G^2 statistics). Type I error for LM(β) tended to increase somewhat as n increased; unlike Orlando and Thissen's (2000) results, the reverse was true for *G*O. Type I error rates for *G*1 were comparable to LM(β) and *G*O only for the 20-item test with N = 500. In all other conditions, Type I error for *G*1 was highly inflated except for 40-item tests with N = 1,000 or 5,000, in which case it was deflated.

The power of the fit statistics to detect violations of the IRF was also examined by Glas and Falcon (2003). Misfitting data were generated by adding parameters to the model that caused b to vary within each of five ability groups. Both the percentage of misfitting items and the magnitude of misfit were manipulated. For both LM(β) and *G*O, power increased with N. Though main effects for n were reported, such that power increased somewhat with n, this pattern was not consistent across all conditions. It also was less evident for *G*O and LM(β) with K = 9 than for LM(β) with K = 5. Glas and Falcon note of the results for n that "they are…hard to explain, and may be a complicated interaction effect of various simulation settings" (p. 96). In the least severe misfit condition (10% misfitting items and 0.25 difference in b-values across ability groups), power ranged from 0.11 to 0.62 for LM(β) with K = 5, 0.08 to 0.99 for LM(β) with K = 9, and from 0.10 to 0.71 for *G*O. In the most severe misfit condition (20% misfitting items and 0.50 difference in b-values across ability groups), power ranged from 0.23 to 1.00 for LM(β) with K = 5 and 9, and from 0.09 to 0.61 for *G*O. *G*O was clearly less powerful

than the LM(β) when violations were severe, perhaps because the violations were more specifically tested by the LM(β) statistic. The superior power of the LM(β) statistic might therefore not generalize across other cases of IRF misspecification.

Finally, across most of the study conditions the false positive rate (Type I error, at $\alpha = 0.05$, for model-fitting items) was inflated for all of the statistics; for LM(β) with K = 5, LM(β) with K = 9, and GO the average false positive/power rates (in percentages) were 17.8/42.8, 26.4/52.7, and 11.0/30.4 respectively. False positive rates increased as the magnitude of individual-item misfit became more severe, but appeared to be relatively unaffected by the frequency of misfitting items on the test. False positive rates also were positively correlated with N and somewhat negatively correlated with n. The *LM*'s increased power does appear to come at a cost of higher Type I error rates for correctly specified items. In the most extreme misfit condition, false positive rates for LM(β) (with K = 5) ranged from 0.12 to 0.78 (the rate was higher with K = 9) whereas for GO it ranged from 0.07 to 0.23.

Reise (1990) found that L_z Type I error rates were very near nominal levels for a 50-item 3PL test with N = 1,000. The statistic was examined across eight different datasets created by crossing four levels of a with two levels of b. There was some evidence that L_z and b were mildly correlated. Correlations between L_z and b were > 0.13 in five of the datasets; for comparison, correlations between QB and b were near zero for all eight datasets. Despite the correlations between L_z and b, Type I error rates for L_z did not appear to fluctuate systematically or depart from nominal levels across any of the conditions studied. The power of L_z and QB to detect misfit when incorrect item parameters were used was also examined. The detection rates for L_z were near 100% when as were 0.50 to 0.75 greater than their true value. At lower levels of misfit, the b distribution appeared to affect power; detection rates were 58% when as were 0.25 greater than their true value and b was U(−1.5,1.5), but they dropped to 38% when b was U(−2.5,2.5). The power for QB in these latter two scenarios was greater, but this increase in power came at the cost of the higher Type I error rates, which were near 0.20 for a =

0.05. Clearly L_z needs to be examined across a wider range of data conditions to determine the generalizability of the promising results mentioned above.

Conclusions

Studies have been conducted examining the functionality of the currently available fit statistics across a number of data conditions, but to date no study has been conducted in which all available fit statistics are methodically compared across a full range of data conditions. To date, evidence is fairly clear that the traditional group-level indices (e.g. $Q1$ and QB) do not function properly unless the test is very long. Though indices based on the model-likelihood function appear to be a promising alternative, it has not been adequately demonstrated that these statistics follow their theoretical sampling distributions across a range of data conditions. It remains unknown whether item parameter values and estimation error affect these sampling distributions. Furthermore, most studies have used only normally distributed θ, so little is known about the effect of other θ distributions on the sampling distributions of the various fit statistics. There is also evidence that the sampling distributions for these statistics are distorted for model-fitting items when there is misfit elsewhere in the data matrix. False positive rates (i.e., Type I error rates for model-fitting items when there is misfit elsewhere in the data matrix) have been examined for most of the likelihood-based indices, but the conditions under which they have been examined across studies are not comparable enough to allow conclusions to be drawn regarding which method is best at minimizing false positive rates. If false positive rates become inflated, model-fitting items might be unduly discarded in the process of removing identified misfitting items from a test. This is potentially a problem and its severity in and consequences for existing item-fit assessment methods warrant further investigation.

Rationale for the Present Study

This study is based on a "whittling down" philosophy regarding item fit statistic research. First, the adherence of (a representative group of) fit statistics to their theoretical sampling distributions, rather than simple adherence to Type I error rates,

should be fully examined under a limited number of conditions at the extremes of what might be used in practice (i.e., small and large *N* and *n*). Additionally, understanding the impact of parameter estimation error on the different types of fit indices would also enable (1) generalization of results to other fit indices of similar type and (2) projections of fit statistic functionality in other conditions across which parameter estimation error would be expected to vary. Finally, because there is always some degree of noise in real item response data, it would be useful to demonstrate the degree to which fit statistics might function differently when item response data are unrealistically well-behaved (the conditions under which fit statistics have always been studied) as opposed to more realistically somewhat noisy.

Those statistics that appear to follow their theoretical sampling distributions across the conditions mentioned above should receive further study. A decision was also made that it would be more efficient to initially suspend examination of the Monte Carlo fit assessment procedures (i.e., Sinharay, 2005; Sinharay, Johnson, & Stern, 2006; Stone et al. 1993; Stone 2000) until it has been demonstrated that none of the other less labor intensive methods (i.e., those that have specific theoretical sampling distributions) is adequate. If one of the less labor intensive methods is shown to function well across a range of data conditions, it then might be worthwhile to compare it to fit statistics yielded by methods such as Stone's to determine which of the two are better.

CHAPTER 2: METHOD

The primary research questions addressed by this study were:

1. How do Type I error rates and empirical sampling distributions for item fit indices vary between 1-, 2-, and 3PL IRT models and across different conditions or levels of (1) parameter estimation error, (2) item discrimination, (3) sample size, and (4) test length?
2. How robust are item fit indices to minor violations of unidimensionality?
3. What is the relationship between fit indices and item parameters?

The item fit indices examined in this study were: $Q1$, $Q\text{O}$, LM($\alpha\beta$), LM(α), LM(β), L_z, $V\text{I}$, and $V\text{O}$.

The between-subjects conditions examined were: (1) IRT model (1-, 2-, and 3PL), (2) data noise, which was operationalized as strictly unidimensional (SU) vs. essentially unidimensional (EU) data, (3) item discrimination (high and low), (4) test length (n = 15 and n = 75), and (5) sample size (N = 500 and N = 1,500). All between-subjects factors were crossed, resulting in a total of 48 ($3 \times 2 \times 2 \times 2 \times 2$) cells. There were also two crossed within-subjects factors to capture the impact of model parameter estimation error on fit statistic functionality. These were: (1) item parameter estimation error, defined by the use of true vs. estimated item parameters ($\boldsymbol{\xi}$ and $\hat{\boldsymbol{\xi}}$ respectively) in computing fit statistics; and (2) person parameter estimation error, defined by the use of true vs. estimated θ parameters ($\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ respectively) in computing fit statistics. Type of item fit statistic can also be considered a within-subjects factor in this study since, like the parameter estimation error conditions, the levels of the factor were applied within each between-subjects cell. The dependent variables in this study were fit statistic Type 1 error rates at three α levels (0.01, 0.05, and 0.10) and empirical sampling distribution moments.

Replication can be considered at both the test and item level. Because the primary interest in this study was to examine properties of individual test items, the choice was made to maintain an equal number of item replications across conditions. There were

18,750 item replications within each of the 48 between-subjects study conditions. This was achieved by creating 1,250 test replications for the n = 15 conditions and 250 for n = 75 conditions.

Item Response Data

A persons × items data matrix of dichotomous item responses was generated for each replication within each of the 48 study cells created by crossing the between-subjects factors. Data were generated for each cell by first computing the model-expected probability of correctly answering each item using the model parameter sets for that cell obtained by the methods discussed below. The 1-, 2-, and 3PL models defined by Equation 1 were used to obtain the SU expected probabilities, and MIRT models defined by Equation 26 below were used to find EU expected probabilities. These expected probabilities were then compared to a U(0, 1) random variable. The item was scored 1 if the value of the random variable was greater than or equal to the model probability of answering the item correctly, otherwise the item was scored 0. The R package for statistical computing (R Development Core Team, 2007) was used to generate all data used in this study. The R code used to generate item response datasets can be found in Appendix A.

Introduction of Data Noise

In the SU condition, all item responses were generated from 1-, 2-, and 3PL models defined by Equation 1. The data in this condition can be considered SU since only one θ accounted for item responses. However, it is generally acknowledged that this strong form of the unidimensionality assumption is unrealistic in real data (Hambleton & Swaminathan, 1985; Nandakumar, 1991).

A weaker form of the unidimensionality assumption requires that a dominant θ account for performance on a set of test items. This has been referred to as *essential* unidimensionality (Nandakumar, 1991; Stout, 1990) and in a factor analytic context implies that the proportion of variance in the data matrix recovered by the primary factor is large with respect to that recovered by additional factors. There are seemingly an

infinite number of ways in which EU data can be generated by varying the number of minor factors, orthogonality of the dimensions, and degree of simple structure. Some researchers have used factor analytic results from real data to obtain parameters for use in simulating EU data with a large number of minor dimensions (Davey, Nering, & Thompson, 1997). Another approach is to simulate data from a multidimensional IRT (MIRT) model in which one primary θ and a minor θ account for responses to each item (Ackerman, 1989; Nandakumar, 1991; Way, Ansley, & Forsyth, 1988).

Though the former approach of Davey et al. (1997) might be more realistic, since it is conceivable that myriad minor factors could account for item responses (such as anxiety, motivation, alertness), its use in the present research posed problems. First, no guidance could be found in the literature regarding how to generate such data in the absence of parameters obtained from real data examples. Such data could be generated from a factor analytic model by sampling minor factor loadings such that they would account for a significantly smaller proportion of the item response variance than the dominant factor. However, the best choice of distribution from which the minor factor loadings should be sampled is not straightforward. Secondly, Davey et al.'s approach uses the nonlinear factor-analytic model implemented in the NOHARM computer program rather than logistic IRT models. Though the data generated by the NOHARM and logistic IRT models should be very similar, it is unknown whether the slight differences between the generating models could affect item fit statistic functionality. Because of this potential confound, the scope of the present study was limited to data generated from a two-dimensional logistic MIRT model.

EU data were generated from a compensatory MIRT model (McKinley & Reckase, 1983) with one common primary ability (θ_1) and one minor ability (θ_2). The dichotomous 3PL compensatory model is given by:

$$P_i(u_{ij} = 1 \mid \boldsymbol{\theta}_\mathbf{j}, \mathbf{a_i}, \mathbf{b_i}, c_i) = c_i + (1 - c_i)P_i^* \tag{26}$$

where

$$P_i^* = \frac{\exp\left[\sum_{r=1}^{R}[a_{ir}(\theta_{jr} - b_{ir})]\right]}{1 + \exp\left[\sum_{r=1}^{R}[a_{ir}(\theta_{jr} - b_{ir})]\right]},$$

$r = 1, \ldots, R$ dimensions,

$\mathbf{\theta_j}$ is the vector of R ability parameters for the jth examinee,

$\mathrm{a_i}$ is the vector of R discrimination parameters for the ith item,

$\mathbf{b_i}$ is the vector of R difficulty parameters for the ith item, and

u_{ij} and c_i are as defined in Equation 1.

The logit for this model can alternatively be expressed in a slope-intercept form as $\sum_{r=1}^{R} a_{ir}\theta_{jr} + d_i$ where $d_i = -\sum_{r=1}^{R} a_{ir}b_{ir}$; note that this is the form in which it was originally proposed by McKinley and Reckase. As with unidimensional dichotomous models, the 2PL version can be obtained by fixing c_i in Equation 26 to zero, and a 1PL version can be obtained by constraining the a parameter in each dimension to be equal across all items.

Data Generation

The datasets created across the data noise and IRT model conditions were designed to be as similar as possible in probability structure. For each test length and sample size condition in the study, and each replication within each of these conditions, MIRT item and person parameters were first sampled to create the 3PL EU datasets in the high discrimination condition. The parameters used to generate datasets in the other conditions were obtained from the 3PL EU model parameters by several methods, which are described below.

3PL Data Parameters

For the EU conditions θ_1 and θ_2 were sampled as uncorrelated random variables from a standard normal distribution. In the high discrimination conditions a_1 was drawn

from a N(1.03, 0.30) distribution and a_2 was drawn from a N(0.49, 0.10) distribution. a_1 and a_2 were also mildly negatively correlated ($r = -0.29$). Both b_1 and b_2 were drawn from normal distributions with $\sigma = 0.82$, but μ for b_1 was 0.30 and μ for b_2 was −0.03. b_1 and b_2 were also weakly negatively correlated ($r = -0.10$). The moments of the distributions from which **a** and **b** were sampled were similar to those used by Way et al. (1988). The *c* parameters were sampled from a lognormal distribution with $\mu = \ln(0.2)$ and $\sigma = 0.15$. For the low discrimination conditions, the sampled **a** parameters were multiplied by 0.6. These, along with the same **b** and *c* parameters as in the high discrimination condition and newly sampled $\boldsymbol{\theta}$ parameters, were used to generate data in the low discrimination conditions.

For each sampled 3PL EU item and person parameter set a corresponding SU item and person parameter set was obtained as follows:

1. To obtain the SU item parameters, a very large ($N = 10{,}000$) item response data matrix was simulated (in the manner described above under *Item Response Data*) with the sampled 3PL MIRT parameters and MIRT model in Equation 26.
2. This data matrix was then calibrated with the unidimensional 3PL IRT model to obtain item parameters for use in generating the SU data. At this step all *c* were fixed at their MIRT generating values to ensure that the EU and SU datasets had similar levels of simulated random guessing.
3. To obtain the SU θ parameters, the EU item and person parameters from Step 1 and the corresponding SU item parameter values obtained from Step 2 were used to find the SU θ that satisfied the equality:

$$\sum_{i=1}^{n} \ln \frac{P^*_{EU_i}}{Q^*_{EU_i}} - \sum_{i=1}^{n} \ln \frac{P^*_{SU_i}}{Q^*_{SU_i}} = 0 \qquad (27)$$

Thus, the SU θ for each simulee was that which yielded an expected score equal to that in the EU data condition. Applying some algebra, Equation 28 can be solved for θ as:

$$\theta_j = \frac{\sum_{i=1}^{n} \ln \frac{P^*_{EU_i}}{Q^*_{EU_i}} + \sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} a_i} \tag{28}$$

Data were generated in this manner in order to minimize potential confounds due to unidimensional model parameter distribution differences between SU and EU conditions. Descriptive statistics for SU generating parameters can be found in Appendix B (Tables B-1 – B-4).

1PL and 2PL Data Parameters

Data generation for the 1PL and 2PL proceeded in a manner similar to that for the 3PL. However, the same distribution of item parameters could not be used for these models since this would result in the datasets being more difficult and more discriminating than the 3PL datasets. Thus, the first step was to rescale the 3PL parameters such that they would yield 1PL and 2PL datasets of comparable difficulty and discriminations as the 3PL dataset. A method similar to that used by Ackerman (1989) in order to create matched item response data generated from two different types of MIRT models was used to match response data generated by 2PL and 3PL MIRT models. Namely, a least squares approach was used to minimize the probability of correct response between the 2PL and 3PL MIRT model.

Specifically, a 2PL MIRT model was defined as:

$$P_{2PLi}(u_{ij} = 1 \mid \boldsymbol{\theta}_\mathbf{j}, \mathbf{a}_\mathbf{i}, \boldsymbol{\alpha}_\mathbf{i}, \mathbf{b}_\mathbf{i}, \delta_i) = \frac{\exp\left(\sum_{r=1}^{2} (a_{ir} + \alpha_{ir})[\theta_{jr} - (b_{ir} + \delta_i)]\right)}{1 + \exp\left(\sum_{r=1}^{2} (a_{ir} + \alpha_{ir})[\theta_{jr} - (b_{ir} + \delta_i)]\right)} \tag{29}$$

where

$\mathbf{a}_\mathbf{i}$ is the vector of sampled 3PL discrimination parameters for the *i*th item,

$\boldsymbol{\alpha}_\mathbf{i}$ is the vector of 2PL discrimination scaling factors for the *i*th item,

$\mathbf{b}_\mathbf{i}$ is the vector of sampled 3PL difficulty parameters for the *i*th item, and

δ_i is the 2PL difficulty scaling factor for the *i*th item.

In order to maintain the relative strength between the major and minor dimension, this model was subject to the constraint that $a_1 / a_2 = \alpha_1 / \alpha_2$. Using a sample of 5,000 simulees with θ_1 and θ_2 uncorrelated and drawn from N(0,1), the 2PL MIRT parameters were then sought by finding the $\boldsymbol{\alpha}_\mathbf{i}$ and $\delta_\mathbf{i}$ minimizing:

$$\sum_{j=1}^{5000} (P_{2PLi} - P_{3PLi})^2 \tag{30}$$

Where P_{3PLi} is defined by Equation 26.

An attempt was made to use this approach to rescale each of the generated 3PL **a** and **b** parameters for use in the 2PL conditions; however, in many cases there did not exist a unique or reasonable solution to the least squares equation. A reasonable solution was obtained for item parameters at the mean of the generating distributions, so 1PL and 2PL parameters were generated from distributions in which the μ was rescaled. The scaling factor (α) for $\mu(a_1)$ was –0.05 and that for $\mu(a_2)$ was $(0.49/1.03)\alpha$. The scaling factor (δ) for $\mu(b_1)$ and $\mu(b_2)$ was –0.38. The σ of the distributions from which item parameters were sampled remained the same as those used to generate 3PL data, except for the 1PL where it was set to zero. Use of these values to generate model parameters resulted in 1PL and 2PL datasets that were fairly close to 3PL datasets in terms of probability structure (see *Checking Generated Data* section below). The 2PL SU item and person parameters were found in the same manner as described above for the 3PL data; descriptive statistics for these parameters can be found in Appendix B (Tables B-1, B-2, and B-4).

Checking Generated Data

The adequacy of the data generation algorithm was inspected. First, an adequate routine should result in generated data in which empirical proportions of correct response (p) converged to model-predicted probabilities (π). Second, since the data generation procedures described above were designed to maintain equivalence in probability

structure across study conditions, p should be similar across all study conditions. Therefore, within each study condition several quantities were examined. First, the mean and SD of item p values across the 18,750 item replications were computed. Next, the difference between p and $\pi(\Delta)$ was obtained for each item. The mean and SD of Δ across the 18,750 item replications was used to evaluate the fidelity of the simulation algorithm.

Table 1 presents the results of this analysis. The fidelity of the algorithm appeared to be very good, with Δ ranging from −0.0002 to 0.0002. As might be expected, SD(Δ) was smaller when N = 1,500 as opposed to N = 500. The average empirical item difficulty also remained fairly consistent across study conditions, ranging from 0.548 to 0.564. Item p values tended to be somewhat more variable in the high discrimination conditions than the low discrimination conditions. In high discrimination conditions SD(p) ranged from 0.159 to 0.200 and in the low discrimination condition SD(p) ranged from 0.131 to 0.160

Table 1. Mean and SD of Empirical Item Proportion-Correct (p) and the Difference (Δ) Between p and Model-Predicted Proportions (π) Within Each Study Cell

N	n	Data Noise	a	3PL				2PL				1PL			
				p		Δ		p		Δ		p		Δ	
				Mean	SD	Mean	SD	M	SD	Mean	SD	M	SD	Mean	SD
500	15	SU	high	0.551	0.161	0.00011	0.019	0.564	0.196	−0.00009	0.016	0.562	0.196	0.00000	0.017
			low	0.561	0.133	−0.00012	0.020	0.550	0.159	0.00020	0.019	0.548	0.159	0.00012	0.019
		EU	high	0.552	0.163	−0.00018	0.019	0.564	0.197	0.00004	0.017	0.561	0.199	−0.00003	0.017
			low	0.561	0.133	0.00019	0.020	0.550	0.159	−0.00010	0.019	0.549	0.160	0.00018	0.019
	75	SU	high	0.549	0.161	−0.00011	0.019	0.562	0.197	0.00000	0.017	0.564	0.198	0.00012	0.017
			low	0.559	0.133	0.00005	0.020	0.549	0.159	0.00000	0.019	0.551	0.159	0.00009	0.019
		EU	high	0.551	0.162	0.00009	0.019	0.562	0.198	0.00014	0.017	0.563	0.199	−0.00011	0.016
			low	0.560	0.133	−0.00002	0.020	0.549	0.160	−0.00011	0.019	0.551	0.160	0.00016	0.019
1,500	15	SU	high	0.548	0.160	0.00008	0.011	0.562	0.198	0.00017	0.010	0.562	0.195	−0.00001	0.009
			low	0.559	0.131	0.00006	0.012	0.550	0.160	−0.00007	0.011	0.550	0.158	0.00013	0.011
		EU	high	0.550	0.161	0.00007	0.011	0.562	0.199	0.00005	0.010	0.562	0.197	−0.00009	0.010
			low	0.559	0.132	−0.00002	0.012	0.549	0.160	0.00009	0.011	0.549	0.158	0.00000	0.011
	75	SU	high	0.550	0.159	0.00000	0.011	0.563	0.195	0.00015	0.010	0.563	0.200	0.00000	0.010
			low	0.561	0.131	−0.00010	0.012	0.549	0.158	−0.00007	0.011	0.550	0.160	−0.00016	0.011
		EU	high	0.552	0.161	0.00002	0.011	0.562	0.196	0.00003	0.010	0.563	0.200	0.00008	0.009
			low	0.561	0.132	0.00009	0.012	0.549	0.158	−0.00007	0.011	0.551	0.160	−0.00016	0.011

Model Parameter Estimation

The item and person parameters for each simulated item response dataset were estimated using MML as implemented in the MULTILOG computer program. Priors were specified for the item parameters in order to stabilize the estimates. In order to determine appropriate priors, the impact of various priors was tested in a small number of replicated tests. The goal of this endeavor was to find priors that would be tight enough to reduce the presence of aberrant estimated parameters but loose enough to allow the estimated parameter distributions to be fairly close to their true distributions. The priors used can be found in Table 2. The priors for *a* include the normal metric scaling factor 1.7, since specification in MULTILOG is on this metric. Additionally, in MULTILOG the *c* parameters are obtained by estimating *d* parameters such that:

$$c = \frac{\exp(d)}{1 + \exp(d)} \tag{31}$$

The priors for the lower asymptote are specified on the *d* metric. Descriptions of estimated model parameters for all study conditions can be found in Appendix B where Tables B-1 through B-12 present descriptive statistics for estimated model parameters, Tables B-13 through B-24 contain statistics summarizing the relationship between estimated EU and SU model parameters, and Tables B-25 through B-36 contain statistics summarizing the relationship between generating and estimated model parameters.

Table 2. Item Parameter Specifications and Priors Used in Dataset Calibrations

	a			
Model	High Discrimination	Low Discrimination	*b*	*d*
1PL	fixed at 1.7	fixed at 1.11	N(−0.21, 1.5)	
2PL	N(1.70, 0.60)	N(1.08, 0.30)	N(−0.23, 1.5)	
3PL	N(1.95, 0.60)	N(1.20, 0.40)	N(0.15, 1.5)	N(−1.35,0.40)

Fit Statistic Computation

Code was written in R (see Appendix A) to compute each of the item fit indices, $Q1$, QO, LM($\alpha\beta$), LM(α), LM(β), L_z, VI, and VO. These statistics were computed for each item in each of the generated item response datasets using estimated model

parameters. In addition, several versions of each statistic were computed (described below), each designed to isolate a particular source of parameter estimation error.

When computing *Q*O and the LM statistics, θ was assumed to come from a N(0,1) distribution. Guassian quadrature with 21 points (the same number used by MULTILOG in the item calibrations) evenly-spaced across θ from −4.0 to 4.0 was used to evaluate the integrals in the *Q*O and LM equations. Since Glas (2003) reported using 40 quadrature points, LM and *Q*O in some of the study conditions were also computed using this larger number of points. However, the values of both statistics were virtually unchanged when 40 points were used as opposed to 21. Additionally, as recommended by Glas (personal communication, July 6, 2008) an approximation to the second derivatives used to compute **W** (in Equation 24) was used to compute the LM statistics. This approximation is given by (Glas, 2003) as:

$$\mathbf{H}_{pq} \cong \sum_{j=1}^{N} E\left[\mathbf{b}_j(\xi_p) \mid \mathbf{u}_j, \xi\right] E\left[\mathbf{b}_j(\xi_p) \mid \mathbf{u}_j, \xi\right]^{\mathrm{T}} \tag{32}$$

where

p = 1, 2,

q = 1, 2, and

$$\mathbf{b}_j(\xi_q) = \frac{\partial \ln L(\mathbf{u}_j, \theta_j \mid \xi)}{\partial \xi_q}$$

Any cross-group elements in the computation of the matrices are equal to zero. For example, letting k and l index the K score groups, $E\left[\mathbf{b}_{jk}(\xi_{2l}) \mid \mathbf{u}_{jk}, \xi\right] = 0$.

The *Q* and LM statistics are group-level fit statistics, requiring that examinees be somehow grouped according to ability level. When computing *Q*1, simulees were initially divided into deciles based on θ. For *Q*O, number-correct (NC) scores were used to classify simulees, with one NC score per group, initially. When computing a χ^2 from contingency tables, the presence of low expected frequencies of correct or incorrect response can disrupt the statistic. Therefore, as recommended by Orlando and Thissen (2000), score groups were collapsed when necessary to maintain expected frequencies of

at least 1.0. Finally, for the LM statistics simulees were divided into five groups based on NC score as was done by Glas (2003).

Estimation Error Conditions

Different versions of each fit statistic were computed in each dataset by crossing the conditions of (1) item parameter estimation error, defined by the use of $\boldsymbol{\xi}$ or $\hat{\boldsymbol{\xi}}$ in computing fit statistics and (2) person parameter estimation error, defined by the use of $\boldsymbol{\theta}$ or $\hat{\boldsymbol{\theta}}$ in computing fit statistics. The item fit statistics for the full estimation error condition ($\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$) were computed using MML $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\theta}}$ as noted above. The statistics for the $\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}$ condition were computed using the true item and estimated person parameters, those for the $\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$ condition were computed using estimated item and true person parameters, and those for the $\boldsymbol{\xi}, \boldsymbol{\theta}$ condition were computed using both true item and person parameters.

The estimation error conditions were crossed with the data noise (SU vs. EU) conditions. For the SU data, the true parameters were defined as those used to generate the item response data. However, in the EU condition, MIRT model parameters were the generating parameters and, since the fit statistics were evaluated with unidimensional model parameter estimates, these MIRT parameters could not be used as the parameters in the estimation error conditions involving $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. In this case, large sample unidimensional model estimates of $\boldsymbol{\xi}$ were used as the true parameters, and the θs resulting from Equation 28 were used as the true person parameters. Note that these are the same methods by which SU model parameter sets were generated. Thus, fit statistics computed in the SU and EU conditions were evaluated in the estimation error conditions with the same true parameter sets.

All four estimation error conditions were examined for $Q1$, L_z, VI, and VO. However, the θ estimation error conditions were not relevant for QO because this statistic does not rely on $\hat{\theta}$. Thus, QO was examined only under the ξ estimation error conditions. Finally, because the LM statistics, by definition, account for both person and

item parameter estimation error it did not make sense to examine their functionality when computed with true item parameters, so the impact of parameter estimation error was not examined for LM statistics.

Data Analysis

The rate at which these statistics rejected the null hypothesis of model-data fit (Type I error) and the distribution of these statistics across the 18,750 item replications were the primary dependent variables.

First, sums of squares (SS) and effect sizes for study factors on Type I error rates in the $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ condition were obtained by estimating an ANOVA model in which all between-subjects study factors were specified. Type of fit statistic was also specified in the model as a within-subjects factor (though this distinction did not really matter since significance tests were not conducted). This analysis summarized the effects of the various study factors on fit statistic functionality when model parameter estimates were used to compute fit statistics. The effect size index (η^2) used in this study is given by:

$$\eta^2 = \frac{SS_{effect}}{\sum SS_{effects}} \tag{33}$$

In order to assess the impact of parameter estimation error (PE), ANOVAs were computed separately for each fit statistic (except the LM statistics, where PE impact was not assessed) in each of the PE conditions. The resulting η^2 for study effects were compared across PE conditions to determine the degree to which influence of study effects varied across PE conditions. Additionally, in order to assess the gross impact of PE on Type I error, SS for the mean (model intercept) and all study effects combined were also obtained from each of these ANOVAs. The SS and η^2 (note that for these analyses $SS_{\text{intercept}}$ was also included in the denominator of Equation 33) for the model intercept and effects were compared across PE conditions to determine the relative impact of PE on fit statistic functionality. If the Type I error rate for a particular fit statistic is near nominal levels and unaffected by any study factors, the SS for the intercept at $\alpha = 0.01$, 0.05, and 0.10 should be near $\alpha^2/N_{\text{cells}}$ (where $N_{\text{cells}} = 48$), or

0.0048, 0.12, and 0.48 respectively. Additionally, the SS would be near zero for all model factors combined.

The degree to which the fit statistic distributions approximated their theoretical sampling distributions was assessed by conducting Kolmogorov-Smirnov (KS) tests and examining the moments of the fit statistic empirical sampling distributions. The KS procedure tests for whether an observed distribution matches a theoretical distribution (DeGroot, 1986). In this test the observed cumulative density function (c.d.f) $F^*(x)$ of a variable (x) is compared with a theoretical one $F(x)$ and the following hypotheses are tested:

$$\begin{aligned} H_0: &\quad F(x) = F^*(x) \quad \text{for } -\infty < x < \infty, \\ H_1: &\quad H_0 \text{ is not true.} \end{aligned} \tag{34}$$

The test statistic is:

$$D = \max_{-\infty<x<\infty} | F(x) - F^*(x) | \tag{35}$$

and can be tested for significance using the c.d.f. of the Kolmogorov distribution:

$$P(K \le t) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2t^2) \tag{36}$$

and the asymptotic result that if H_0 is true:

$$\sqrt{N_{D^*}}\, D \xrightarrow{N_{D^*}\to\infty} K \tag{37}$$

where N_{D^*} are the number of cases upon which $F^*(x)$ is based. An α level of 0.05 was used to determine KS test significance in this study.

For the LM and z statistics, the first four distributional moments (mean, SD, skewness, and kurtosis) were computed over the 18,750 item replications and inspected across study conditions. The KS test results were summarized by computing the percentage of replicated tests (π_R) for which the KS test null hypothesis was rejected. If fit statistics follow their theoretical distribution, π_R should be near 0.05 (the α level used for the KS tests).

The manner in which distributional analyses were carried out was somewhat different for $Q1$ and QO. Because categories (K) were collapsed for the Q statistics in

order to maintain expected counts ≥ 1, K varied across items, and hence theoretically items came from χ^2 distributions with different degrees of freedom. Thus, distributional analyses were conducted conditional on K for the Q statistics. Items were grouped according to K, and within each group the first two distributional moments were examined and the KS test was conducted. The relationship between Q statistics and K was also examined in $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ conditions by constructing 95% confidence intervals (CI) around the mean and variance estimates in each K group. For the mean, these intervals were obtained using the expected SD of the sampling distribution for the mean, or $\sqrt{2df / N_k}$. Because the distribution for the sample variance is unknown, the CIs were constructed from the quantiles of empirical distributions of the sample variance obtained from a Monte Carlo sampling procedure. To carry out this procedure, for each K group, 100,000 random samples of size N_K were drawn from a χ^2 distribution with the appropriate df (i.e., K – model). The variance of each sample was obtained and the 2.5th and 97.5th quantile points were used as the lower and upper bounds for the CIs. Only moments from groups in which $N_K \geq 30$ were examined.

To summarize the departure of Q statistic moments from expectation in the PE conditions, the bias and mean error (ME) of each was computed by:

$$\text{Bias(Mean)} = \sum_{k=1}^{K} \left(\overline{Q}_k - df_k \right) / K \tag{38}$$

$$\text{ME(Mean)} = \sum_{k=1}^{K} \left| \overline{Q}_k - df_k \right| / K \tag{39}$$

$$\text{Bias(SD)} = \sum_{k=1}^{K} \left(SD\left[\overline{Q}_k \right] - \sqrt{2df_k} \right) / K \tag{40}$$

$$\text{ME(SD)} = \sum_{k=1}^{K} \left| SD\left[\overline{Q}_k \right] - \sqrt{2df_k} \right| / K \tag{41}$$

where $k = 1, \ldots, K$ indexes the number of groups (in which $N_K \geq 30$) used to compute $Q1$.

Bias was inspected across study conditions to determine patterns of negative or positive bias; therefore, the pattern of signs was more important in interpreting Bias than its magnitude. The magnitude of ME was inspected across study conditions to determine

conditions in which moments were more likely to be distorted. Because the statistics in each study condition had different constitutions of K and N_K, both of which will theoretically affect the magnitude of ME, a reference point was needed to interpret the magnitude of these values. Therefore a Monte Carlo procedure was again used to obtain 95% CIs about each ME. This was accomplished by several steps. Within each study condition:

1. The observed K and N_K were obtained and used as data simulation inputs,
2. For each K, N_K observations were randomly drawn from a χ^2 distribution with the appropriate *df,*
3. The mean and SD were obtained from each of the K groups and used to compute ME in Equations 39 and 41,
4. Steps 1–3 were repeated 10,000 times and the 2.5th and 97.5th quantiles were obtained for ME(Mean) and ME(SD) as the lower and upper bounds of the 95% CIs.

KS tests were conducted for each K with 15 or more items. To summarize the results of the KS tests, the frequency (F_K) and proportion (π_K) of tests across K (where $N_K \geq 15$) in which the null hypothesis was rejected were obtained. If fit statistics followed their theoretical distribution, π_K should be near 0.05 (the α level used for the KS tests).

CHAPTER 3: RESULTS

Type I Error Rates

Relative Effect Sizes for Study Factors

The sums of squares (SS) and effect sizes (η^2) from the ANOVA conducted on Type 1 error (T1) rates can be found in Table 3 for all main effects and all interactions with small ($\eta^2 \geq 0.005$) effects, and in Table C-1 of Appendix C for all effects. There was a large main effect for type of fit statistic (FS), where η^2 ranged from approximately 0.20 to 0.25 across the three α levels, and *n*, where η^2 ranged from approximately 0.22 to 0.30. Small effects were evident for IRT model (M), where η^2 ranged from about 0.02 to 0.03 across the three α levels, and *N,* where η^2 ranged from about 0.01 to 0.03. Overall, level of discrimination (D) and data noise (DN) had a negligible effect on T1 rates. Effects were observed for all two-way interactions involving FS, except that with DN. Small effects (ranging from 0.01 to 0.02) were evident for FS × D, moderate effects (ranging from 0.03 to 0.08) were evident for FS × M and FS × *N*, and large effects (ranging from 0.24 to 0.31) were evident for FS × *n*. There were also small to moderate effects (ranging from 0.01 to 0.05) for the interaction between *N* and *n* and the three-way interaction between FS, *N*, and *n*. The η^2 for all effects involving DN were uniformly < 0.00005, indicating no effect of data noise on T1 rates.

Table 3. Sums of Squares and Effect Sizes for Study Factors on Type I Error Rates

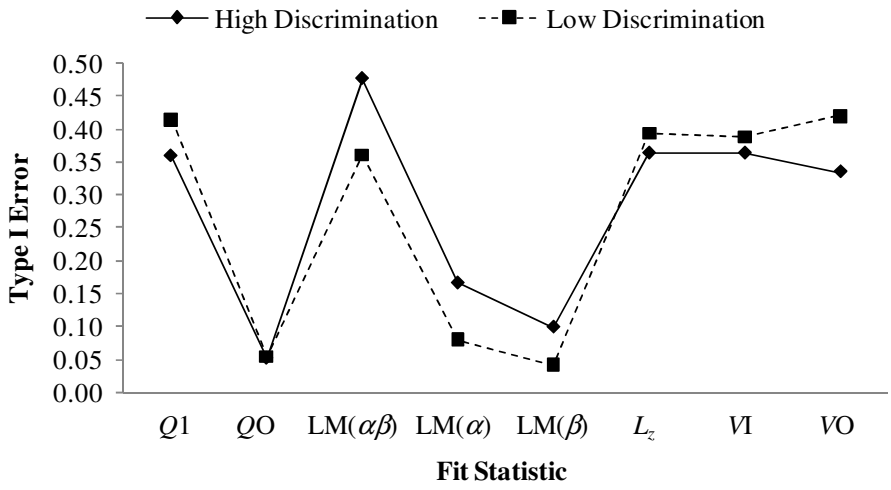
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
	SS	η^2	SS	η^2	SS	η^2
Main Effects						
Fit Statistic (FS)	5.7779	0.1970	8.6115	0.2311	9.6266	0.2450
Discrimination (D)	0.0004	--	0.0051	--	0.0038	--
Data Noise (DN)	< 0.0001	--	< 0.0001	--	< 0.0001	--
Model (M)	0.6126	0.0209	0.8380	0.0225	0.9909	0.0252
Sample Size (*N*)	1.0235	0.0349	0.4761	0.0128	0.2904	0.0074
Test Length (*n*)	6.4998	0.2217	10.4955	0.2816	11.6854	0.2974
Interactions						
FS × D	0.5372	0.0183	0.4272	0.0115	0.3849	0.0098
FS × M	2.2884	0.0780	2.2767	0.0611	2.1412	0.0545
FS × *N*	2.1665	0.0739	1.4552	0.0390	1.1174	0.0284
FS × *n*	7.0748	0.2413	11.0728	0.2971	12.1892	0.3102
N × *n*	1.4165	0.0483	0.5824	0.0156	0.2149	0.0055
FS × *N* × *n*	1.4155	0.0483	0.6387	0.0171	0.2749	0.0070
TOTAL WS	19.5776	0.6676	24.7350	0.6637	25.9756	0.6611
TOTAL BS	9.7467	0.3324	12.5357	0.3363	13.3146	0.3389
TOTAL	29.3243		37.2707		39.2902	

Overall, T1 rates were inflated for all fit statistics other than *Q*O, and to a lesser extent, LM(β); across all study conditions combined. at $\alpha = 0.01$, T1 rates were 0.01 for *Q*O, 0.04 for LM(β), 0.08 for LM(α), between 0.27 and 0.28 for *Q*1 and the *z* (L_z, *V*I, and *V*O) statistics, and 0.34 for LM($\alpha\beta$). At $\alpha = 0.05$ T1 rates were 0.05 for *Q*O, 0.07 for LM(β), 0.12 for LM(α), between 0.38 and 0.39 for *Q*1 and the *z* statistics, and 0.42 for LM($\alpha\beta$). Finally, at $\alpha = 0.10$ T1 rates were 0.10 for *Q*O, 0.11 for LM(β), 0.17 for LM(α), between 0.43 and 0.44 for the *z* statistics, 0.46 for *Q*1, and 0.47 for LM($\alpha\beta$).

The interaction between FS and D (illustrated in Figure 1 for $\alpha = 0.05$ and in Table C-2 of Appendix C for all α levels) was such that T1 rates were somewhat higher in the high than the low discrimination condition for the LM statistics, with the reverse being true for *Q*1 and the *z* statistics. *Q*O appeared to be unaffected by D. The effects of D for the other fit statistics were mild; T1 was far from nominal levels for *Q*1, LM($\alpha\beta$), and the *z* statistics in the D condition that yielded lower T1 rates. LM(α) and LM(β)

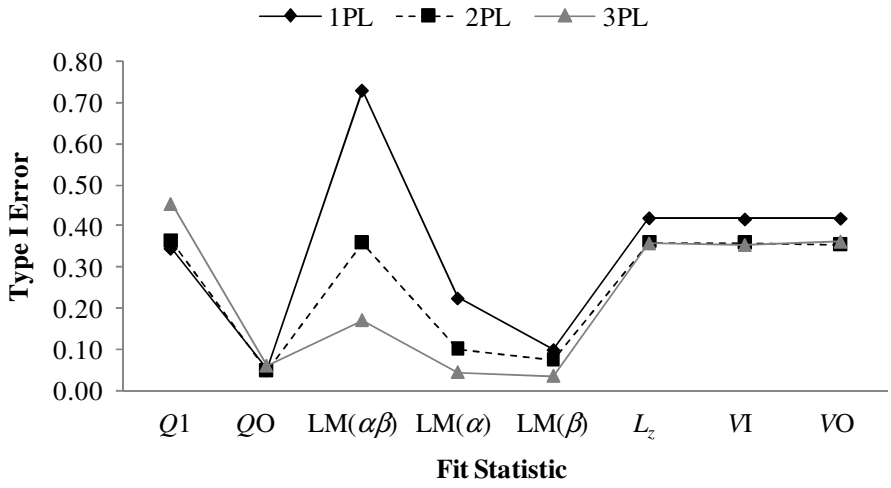
exhibited T1 rates close to nominal levels for the low discrimination condition, with LM(α) being somewhat higher than nominal levels and LM(β) being somewhat lower, except at $\alpha = 0.01$ where LM(β) rates were not deflated. In the high discrimination condition, T1 for LM(α) and LM(β) approximately doubled, except for $\alpha = 0.01$ where the rate of inflation was somewhat higher.

Figure 1. FS × D Interaction for Type I Error Rates at $\alpha = 0.05$



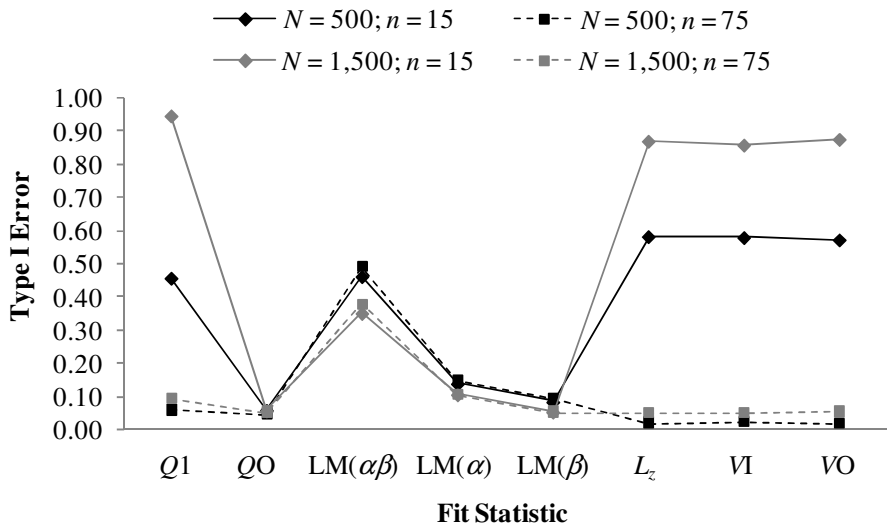
Overall, T1 rates were mildly negatively related to model complexity (see Table C-3 of Appendix C); for example, at $\alpha = 0.05$, overall T1 rates for the 1PL, 2PL, and 3PL were 0.34, 0.25, and 0.23. The interaction between FS and M (illustrated in Figure 2 for $\alpha = 0.05$ and in Table C-3 of Appendix C for all α levels) appeared to be attributable to (1) a reversal of this pattern for $Q1$ such that T1 was somewhat higher for the 3PL than for the 1PL and 2PL, (2) a stronger mode1 effect for LM($\alpha\beta$) and LM(α), and (3) the lack of a discernible model effect for QO. The effect of model on LM($\alpha\beta$) was most dramatic; for example, at $\alpha = 0.05$ T1 rates for the 1PL, 2PL, and 3PL were 0.73, 0.36, and 0.17 respectively.

Figure 2. FS × M Interaction for Type I Error Rates at $\alpha = 0.05$



Finally, the three-way interaction between FS, *N*, and *n* (displayed in Figure 3 for $\alpha = 0.05$ and Table C-4 of Appendix C for all α levels) revealed a strong impact of test length on T1 rates for *Q*1 and the *z* statistics, but not for the other fit statistics. Sample size had a small impact on T1 for the LM statistics, with rates somewhat more inflated when *N* = 500 as opposed to when *N* = 1,500. Sample size had a much stronger impact on T1 rates for *Q*1 and the *z* statistics when *n* = 15 than when *n* = 75. In fact, when *n* = 75 T1 rates were near nominal levels for both levels of *N*, though somewhat deflated for the *z* statistics when *N* = 500 and somewhat inflated for Q1 when *N* = 1,500. Compared with the other fit statistics, *Q*O T1 rates were unperturbed by either *N* or *n*.

Figure 3. FS × N × n Interaction for Type I Error Rates at $\alpha = 0.05$

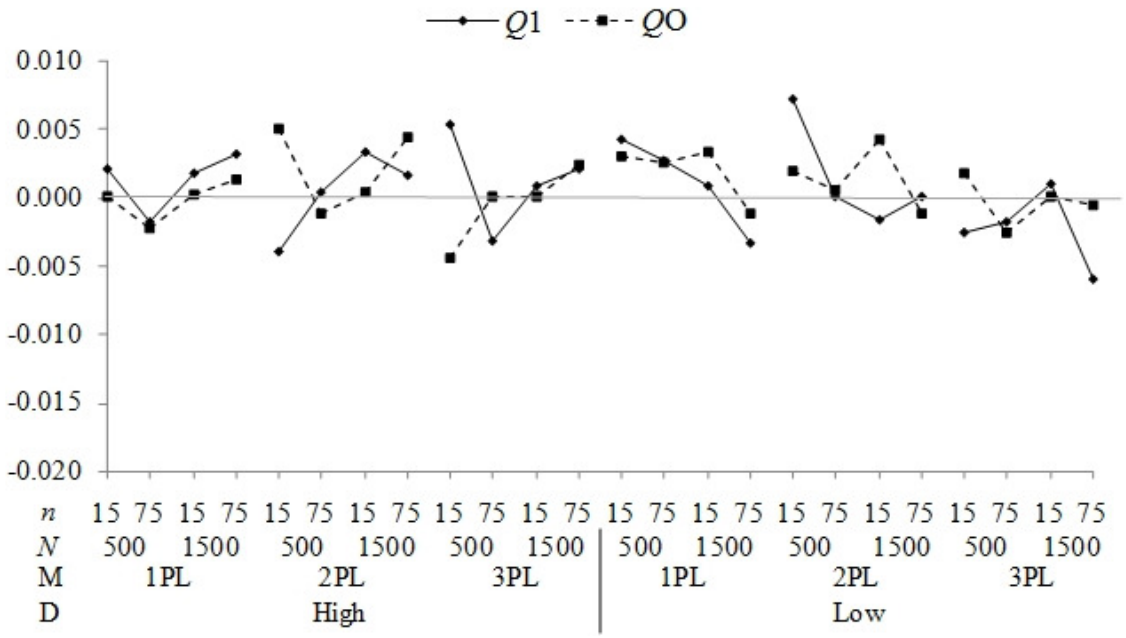


Examination of Type I Error Rate Bias in Essentially Unidimensional Data

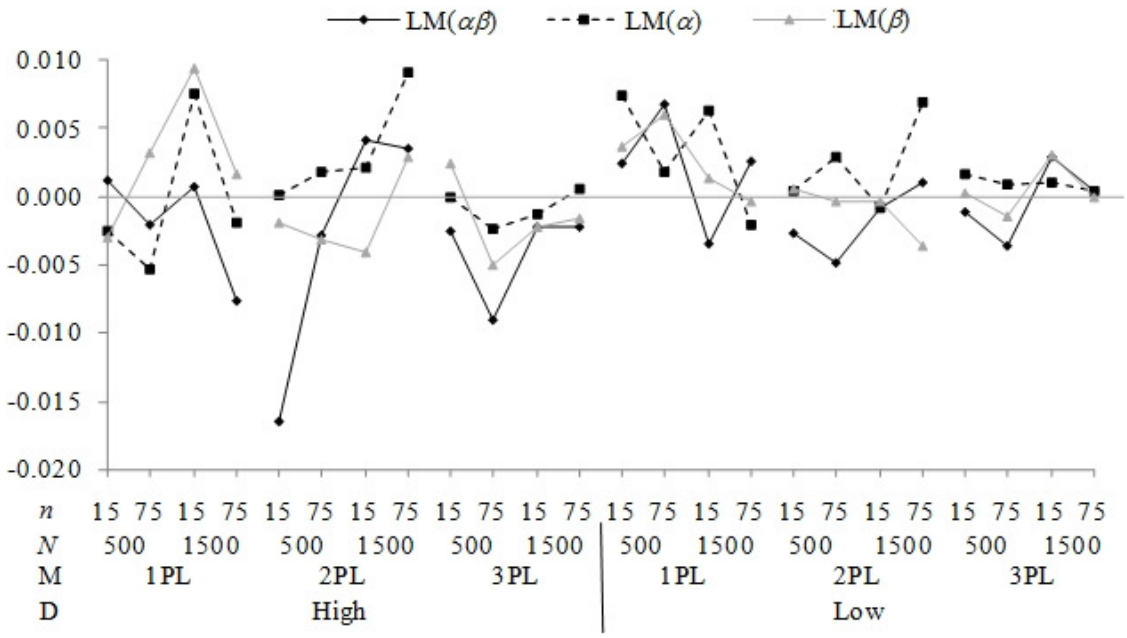
The effect size statistics did not provide any evidence that DN affected the functionality of the fit statistics. In order to further examine the effect for DN and ensure that it had no impact on T1 rates, the difference in T1 rates between SU and EU conditions was computed and plotted. This difference, which was created by subtracting SU T1 rates from EU T1 rates, will be referred to as DN bias. By its definition, positive values indicate that T1 rates were inflated in conditions in which more data noise was present, as might be expected if the fit statistics had enough power to detect the minor data perturbance introduced in the EU conditions. However, the plots of DN Bias (which can be found in Figure 4 for $\alpha = 0.05$ and in Figures C-1 to C-6 of Appendix C for other α levels) lacked any clear systematic pattern indicative of such perturbance for any of the fit indices.

Figure 4. DN Bias by Study Condition at $\alpha = 0.05$

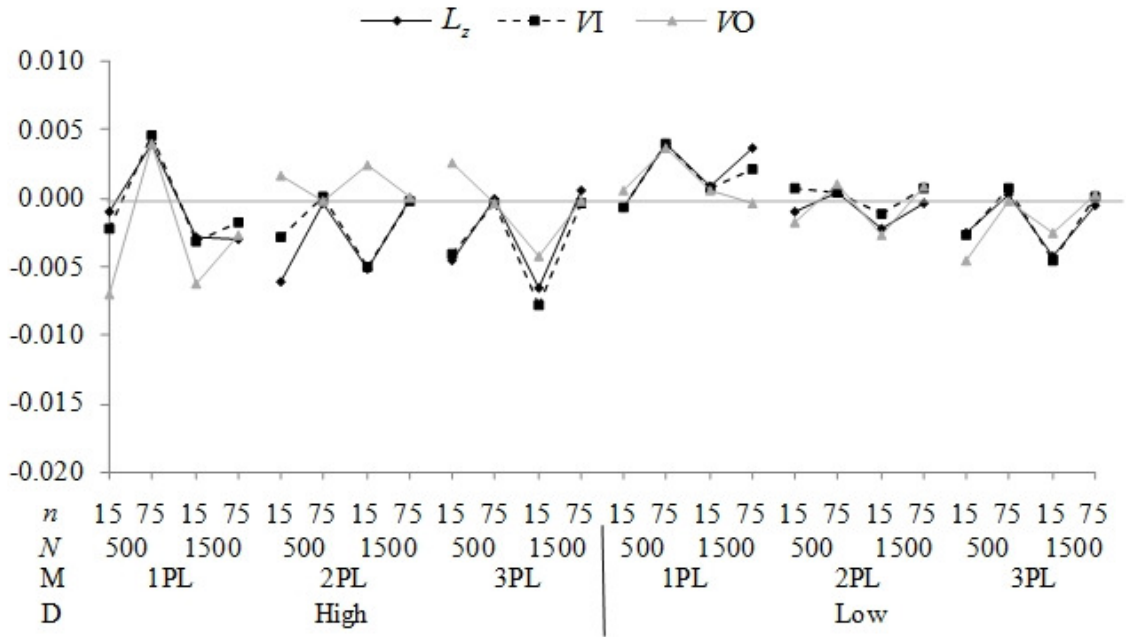
a. Q Sttistics



b. LM Statistics



c. z Statistics



If the presence of minor dimensionality were to have an impact, it might also be more pronounced for conditions in which parameters were estimated more precisely. For example, there would be the most DN Bias for the 1PL and the least for the 3PL, since more complex models tend to have less accurate parameter estimates, all other things being equal. Alternatively, since parameters tend to be estimated with more precision in larger sample sizes, the longer test length might be expected to exhibit more DN Bias. However, the first pattern was only subtly evident (by the slight difference in y-axis level between the three models in the DN Bias figures) for the LM statistics and the second pattern was not evident in the figures. On average, the 1PL did exhibit the most positive DN Bias values; at $\alpha = 0.01$, DN Bias values for the 1PL, 2PL, and 3PL were 0.0013, −0.0008, and −0.0002 respectively. At $\alpha = 0.05$ the respective values were 0.0008, −0.0001, and −0.0011 and at $\alpha = 0.10$ they were 0.0012, 0.0003, and −0.0014. No overall trends were evident for sample size. Because the ratio between sample size and the number of estimated parameters (SSR) has been shown to impact item parameter estimation error more than sample size alone (De Ayala, 1999), DN Bias was also

examined by SSR. As evident in Table 4, DN Bias did not systematically increase with SSR for any of the three models.

Table 4. DN Bias (Averaged Across D and FS) by Model and SSR

Model	SSR	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1PL	7	0.0017	0.0022	−0.0001
	20	0.0005	−0.0006	0.0025
	33	0.0021	0.0005	0.0023
	100	0.0011	0.0012	−0.0001
2PL	3	−0.0005	−0.0004	−0.0031
	10	0.0006	0.0006	0.0003
	17	−0.0014	−0.0012	0.0008
	50	−0.0019	0.0016	0.0032
3PL	2	−0.0005	−0.0017	−0.0014
	7	−0.0006	−0.0003	−0.0019
	11	0.0012	−0.0009	−0.0017
	33	−0.0007	−0.0016	−0.0007

There was, thus, virtually no evidence that DN had any impact on fit statistic functionality. Though DN Bias, on average, was consistently positive for the 1PL across α levels, but not for the 2PL and 3PL, the magnitude of the values in all cases was so small that these variations, though logically interpretable as a DN effect, were likely merely attributable to random fluctuations. Due to the lack of any perceptible effects for DN, T1 results for this condition will not be further discussed; results will be presented for the SU conditions only, though results for some EU conditions are presented in the appendices to demonstrate equivalence between SU and EU conditions.

Adherence to Type I Error Rates Across Study Conditions When Estimated Model Parameters are Used

T1 rates at $\alpha = 0.01$, 0.05, and 0.10 for all SU conditions can be found in Tables 5–7 and rates at all α levels for the EU conditions can be found in Appendix C (Tables C-5 to C-7). T1 rates for each cell in the study design were classified into one of three

categories: below nominal levels, near nominal levels, and above nominal levels. T1 rates were considered near nominal levels if they were within 0.03 above or below the α level.

*Q*O was the only fit statistic that yielded T1 rates consistently near nominal levels across all study conditions and α levels, except $\alpha = 0.10$ for the 3PL with $N = 500$ and $n = 15$. However, in this condition T1 rates were just slightly above the cutoff value for both levels of discrimination, reaching 0.132 in the high discrimination and 0.140 in the low discrimination condition.

T1 rates for LM(β) were near nominal levels in many conditions, but the proximity to nominal levels varied across α level. For example, at $\alpha = 0.01$, T1 in the high discrimination conditions was considered inflated for all 1PL and 2PL conditions, and the 3PL when $N = 500$ and $n = 75$. However, at $\alpha = 0.05$ T1 was considered inflated only for all 1PL conditions, and the 2PL when $N = 500$. The T1 pattern when $\alpha = 0.10$ was the same as that when $\alpha = 0.05$, except for the T1 rate deflation for the 3PL when $N = 1{,}500$. In the low discrimination conditions, T1 remained close to nominal levels across most study conditions and α levels, except for (1) the 3PL when $N = 1{,}500$ and $n = 75$ at $\alpha = 0.05$, (2) the 2PL when $N = 1{,}500$ at $\alpha = 0.05$, and (3) all 3PL conditions at $\alpha = 0.10$.

LM(α) exhibited a similar T1 pattern to LM(β) but with elevated T1 rates in general, which resulted in fewer cells in which T1 rates were near nominal levels. LM($\alpha\beta$) performed very poorly in all study conditions, and began to approach nominal T1 levels only for the 3PL low discrimination conditions.

*Q*1 generally had T1 rates near nominal levels when $n = 75$, except when used with the 3PL where it tended to be somewhat inflated. For the longer test length, T1 rates also were closer to nominal levels when $N = 500$ than when $N = 1{,}500$; this pattern of results for *Q*1 is similar to that found in past research (McKinley & Mills, 1985; Glas & Falcon, 2003).

Finally, the *z* statistics had patterns of T1 rates similar to one another. In both discrimination conditions, these statistics had T1 rates near nominal levels across all α levels only for the 1PL when $N = 500$ and $n = 75$. For the other two models, T1 rates were quite deflated for the larger *n* condition, with the deflation being more extreme for

N = 500 than N = 1,500. This difference was more marked in the low discrimination conditions.

Table 5. Type I Error Rates at $\alpha = 0.01$ for all SU Conditions

Model	N	n	Fit Statistic							
			Q1	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO
High Discrimination										
1PL	500	15	0.089	**0.010**	0.758	0.236	0.111	0.318	0.344	0.283
		75	**0.008**	**0.011**	0.778	0.253	0.129	**0.009**	**0.013**	**0.004**
	1,500	15	0.746	**0.011**	0.719	0.193	0.049	0.799	0.801	0.743
		75	**0.011**	**0.012**	0.707	0.180	0.046	0.032	0.033	**0.019**
2PL	500	15	0.112	**0.007**	0.404	0.136	0.111	0.189	0.226	0.117
		75	**0.008**	**0.006**	0.471	0.161	0.136	**0.000**	**0.000**	**0.000**
	1,500	15	0.729	**0.013**	0.198	0.055	0.042	0.737	0.743	0.630
		75	**0.013**	**0.009**	0.285	0.063	0.054	**0.000**	**0.000**	**0.000**
3PL	500	15	0.282	**0.014**	0.171	0.036	**0.025**	0.227	0.256	0.150
		75	**0.017**	**0.009**	0.235	0.058	0.049	**0.000**	**0.000**	**0.000**
	1,500	15	0.853	**0.014**	0.063	**0.013**	**0.009**	0.727	0.727	0.689
		75	0.035	**0.010**	0.120	**0.024**	**0.020**	**0.000**	**0.000**	**0.000**
Low Discrimination										
1PL	500	15	0.150	**0.010**	0.574	0.075	**0.018**	0.391	0.400	0.436
		75	**0.009**	**0.009**	0.608	0.081	**0.018**	**0.009**	**0.012**	**0.009**
	1,500	15	0.870	**0.012**	0.548	0.067	**0.012**	0.817	0.804	0.865
		75	**0.017**	**0.012**	0.578	0.073	**0.012**	0.038	0.039	0.044
2PL	500	15	0.215	**0.009**	0.269	**0.030**	**0.015**	0.395	0.407	0.426
		75	**0.009**	**0.006**	0.279	**0.030**	**0.020**	**0.000**	**0.000**	**0.000**
	1,500	15	0.878	**0.011**	0.113	**0.013**	**0.007**	0.789	0.778	0.834
		75	**0.015**	**0.008**	0.115	**0.010**	**0.007**	**0.000**	**0.001**	**0.001**
3PL	500	15	0.443	**0.015**	0.088	**0.011**	**0.006**	0.404	0.416	0.437
		75	**0.018**	**0.009**	0.100	**0.011**	**0.008**	**0.000**	**0.000**	**0.000**
	1,500	15	0.939	**0.014**	**0.024**	**0.004**	**0.003**	0.776	0.764	0.824
		75	0.037	**0.011**	0.032	**0.005**	**0.004**	**0.000**	**0.000**	**0.000**

NOTE. T1 rates near nominal levels, defined by $0.00 \leq T1 \leq 0.03$, are in bold.

Table 6. Type I Error Rates at $\alpha = 0.05$ for all SU Conditions

Model	N	n	Fit Statistic							
			$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO
High Discrimination										
1PL	500	15	0.276	**0.047**	0.813	0.315	0.169	0.572	0.578	0.530
		75	**0.041**	**0.048**	0.829	0.326	0.176	**0.052**	**0.057**	**0.032**
	1,500	15	0.908	**0.048**	0.775	0.285	0.110	0.890	0.886	0.860
		75	**0.058**	**0.052**	0.770	0.261	0.092	0.122	0.119	0.096
2PL	500	15	0.316	**0.045**	0.491	0.177	0.139	0.518	0.517	0.422
		75	**0.044**	**0.042**	0.547	0.199	0.166	*0.000*	*0.000*	*0.002*
	1,500	15	0.904	**0.049**	0.277	0.089	**0.068**	0.851	0.847	0.778
		75	**0.061**	**0.045**	0.368	0.099	**0.079**	*0.001*	*0.003*	*0.004*
3PL	500	15	0.535	**0.069**	0.250	**0.064**	**0.047**	0.526	0.524	0.469
		75	0.081	**0.048**	0.309	0.086	**0.073**	*0.000*	*0.001*	*0.001*
	1,500	15	0.944	**0.063**	0.121	**0.035**	**0.026**	0.830	0.825	0.821
		75	0.126	**0.051**	0.177	**0.043**	**0.036**	*0.001*	*0.004*	*0.001*
Low Discrimination										
1PL	500	15	0.386	**0.048**	0.656	0.148	**0.063**	0.623	0.615	0.673
		75	**0.044**	**0.044**	0.690	0.161	**0.061**	**0.053**	**0.055**	**0.057**
	1,500	15	0.961	**0.051**	0.634	0.138	**0.054**	0.897	0.880	0.942
		75	**0.079**	**0.053**	0.661	0.150	**0.053**	0.136	0.133	0.154
2PL	500	15	0.492	**0.050**	0.376	**0.072**	**0.040**	0.620	0.614	0.660
		75	**0.048**	**0.042**	0.388	**0.070**	**0.046**	*0.002*	*0.003*	*0.002*
	1,500	15	0.956	**0.050**	0.207	**0.043**	**0.029**	0.879	0.860	0.929
		75	**0.075**	**0.048**	0.210	**0.039**	**0.031**	*0.018*	*0.019*	**0.035**
3PL	500	15	0.710	**0.072**	0.174	**0.037**	**0.028**	0.620	0.613	0.664
		75	0.087	**0.051**	0.179	**0.036**	**0.030**	*0.001*	*0.001*	*0.001*
	1,500	15	0.978	**0.068**	**0.077**	**0.024**	**0.020**	0.860	0.841	0.907
		75	0.153	**0.055**	0.082	**0.024**	*0.019*	*0.017*	*0.019*	**0.023**

NOTE. T1 rates near nominal levels, defined by $0.02 \leq T1 \leq 0.08$, are in bold, those below nominal levels are in italics, and those above nominal levels are in regular font.

Table 7. Type I Error Rates at $\alpha = 0.10$ for all SU Conditions

Model	N	n	Fit Statistic							
			$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO
High Discrimination										
1PL	500	15	0.428	**0.092**	0.842	0.375	0.221	0.698	0.691	0.655
		75	**0.086**	**0.092**	0.858	0.382	0.224	**0.106**	**0.112**	**0.076**
	1,500	15	0.954	**0.094**	0.807	0.343	0.171	0.923	0.917	0.907
		75	**0.119**	**0.100**	0.805	0.323	0.144	0.206	0.199	0.182
2PL	500	15	0.477	**0.096**	0.550	0.217	0.171	0.674	0.669	0.572
		75	**0.092**	**0.091**	0.600	0.236	0.197	*0.001*	*0.004*	*0.004*
	1,500	15	0.955	**0.094**	0.341	**0.125**	**0.098**	0.895	0.889	0.843
		75	**0.122**	**0.093**	0.432	0.135	**0.110**	*0.012*	*0.019*	*0.034*
3PL	500	15	0.682	0.132	0.316	**0.101**	**0.077**	0.666	0.659	0.626
		75	0.157	**0.104**	0.367	**0.117**	**0.103**	*0.003*	*0.006*	*0.002*
	1,500	15	0.971	**0.124**	0.179	*0.064*	*0.051*	0.873	0.865	0.875
		75	0.216	**0.103**	0.230	*0.068*	*0.060*	*0.013*	*0.021*	*0.008*
Low Discrimination										
1PL	500	15	0.538	**0.096**	0.706	0.207	**0.112**	0.730	0.712	0.780
		75	**0.094**	**0.094**	0.734	0.229	**0.115**	**0.104**	**0.107**	**0.118**
	1,500	15	0.981	**0.101**	0.682	0.201	**0.107**	0.927	0.910	0.967
		75	0.146	**0.104**	0.710	0.212	**0.106**	0.223	0.218	0.248
2PL	500	15	0.650	**0.099**	0.453	**0.115**	**0.072**	0.730	0.716	0.772
		75	**0.103**	**0.089**	0.460	**0.112**	**0.079**	*0.012*	*0.015*	*0.015*
	1,500	15	0.975	**0.096**	0.281	**0.081**	*0.059*	0.919	0.897	0.959
		75	0.144	**0.099**	0.284	**0.077**	*0.062*	**0.077**	**0.072**	**0.118**
3PL	500	15	0.818	0.140	0.245	**0.071**	*0.058*	0.717	0.704	0.764
		75	0.170	**0.106**	0.247	*0.068*	*0.058*	*0.007*	*0.008*	*0.007*
	1,500	15	0.989	**0.126**	0.136	*0.052*	*0.045*	0.896	0.875	0.937
		75	0.266	**0.110**	0.133	*0.049*	*0.041*	**0.087**	**0.086**	**0.106**

NOTE. T1 rates near nominal levels, defined by $0.07 \leq T1 \leq 0.13$, are in bold, those below nominal levels are in italics, and those above nominal levels are in regular font.

Impact of Parameter Estimation Error on Type I Error Rates

For each fit statistic, sums of squares (SS) for the mean (model intercept) and all study factors combined were obtained from a univariate ANOVA of T1 rates in each of the PE conditions; complete ANOVA tables with disaggregated study effects for T1 rates at $\alpha = 0.05$ can be found in Appendix D (Tables D-1 to D-5). The SS and effect sizes for the model intercept and factors were compared across PE conditions to determine the

relative impact of PE on fit statistic functionality. If T1 rates for a fit statistic were near nominal levels and unaffected by any study factors, the SS for the intercept at $\alpha = 0.01$, 0.05, and 0.10 should be near α^2/N_{cells} (where $N_{cells} = 48$), or 0.0048, 0.12, and 0.48 respectively. Additionally, the SS would be near zero for all model factors combined.

The results of this analysis for *Q*O can be found in Table 8. At $\alpha = 0.01$ and 0.05, intercept SS were closer to target values in the $\hat{\xi}$ condition than the ξ condition, with the reverse being true at $\alpha = 0.10$. At all α levels, factor SS were closer to zero and accounted for less of the total variation in T1 rates in the ξ than the $\hat{\xi}$ condition. These results indicate that the study factors tended to have somewhat less impact on T1 rates when ξ were used to compute *Q*O, but that the levels of T1 tended to be slightly more inflated compared to when $\hat{\xi}$ were used, except at $\alpha = 0.10$ where the reverse was true. However, the degree of difference in T1 levels between the two conditions was very small, with $\hat{\xi}$ and ξ T1 rates averaging, respectively, 0.0108 and 0.0123 at $\alpha = 0.01$, 0.0521 and 0.0536 at $\alpha = 0.05$, and 0.1040 and 0.1031 at $\alpha = 0.10$ across the 48 study cells.

Figures in which with *Q*O T1 rates are compared between PE conditions for each study cell can be found in Figures D-1 to D-6 of Appendix D. From inspection of these figures, it is apparent that model affected T1 rates more in the $\hat{\xi}$ than ξ conditions, particularly for the 3PL when $n = 15$. To quantify these effects, M and the M × n interaction accounted for 45.8% and 19.5% of T1 variation for $\hat{\xi}$ and only 8.2% and 2.5% for ξ. Additionally, DN had a greater effect on T1 error in ξ than $\hat{\xi}$ conditions, accounting for less than 1% of T1 variation (at $\alpha = 0.05$) for $\hat{\xi}$ and nearly 19.5% for ξ.

Table 8. *Q*O Type I Error Rate Sums of Squares and Effect Sizes for Model Intercept and Study Factors by Parameter Estimation Error Condition

		Parameter Estimation Error Condition			
		$\hat{\boldsymbol{\xi}}$		$\boldsymbol{\xi}$	
		SS	η^2	SS	η^2
$\alpha = 0.01$					
	Intercept	0.0056	0.9544	0.0072	0.9815
	Factors	0.0003	0.0456	0.0001	0.0185
	TOTAL	0.0058		0.0074	
$\alpha = 0.05$					
	Intercept	0.1301	0.9779	0.1380	0.9937
	Factors	0.0029	0.0221	0.0009	0.0063
	TOTAL	0.1331		0.1389	
$\alpha = 0.10$					
	Intercept	0.5190	0.9837	0.5101	0.9965
	Factors	0.0086	0.0163	0.0018	0.0035
	TOTAL	0.5276		0.5119	

The impact of PE on T1 rates for the other fit statistics (which can be seen in Figures D-7 to D-12 of Appendix D) was more marked than that for *Q*O, whose T1 rates remained relatively unperturbed even in the $\hat{\boldsymbol{\xi}}$ condition. For *Q*1 and the *z* statistics, T1 rates remained closest to nominal levels, and were least affected by study factors in the $\boldsymbol{\xi}, \boldsymbol{\theta}$ condition. Person-parameter estimation error also had a greater impact on T1 for both types of fit statistics. However, the pattern of results across the four PE conditions was somewhat different between *Q*1 and the *z* statistics.

*Q*1 T1 rate SS and η^2 in each PE condition can be found in Table 9. The minor impact of ξ estimation error is evident by the slight reduction in factor η^2 and the slight decrease in intercept SS within each θ estimation error condition when *Q*1 was computed with $\boldsymbol{\xi}$ as opposed to $\hat{\boldsymbol{\xi}}$. For example, at $\alpha = 0.05$, when $\hat{\boldsymbol{\theta}}$ was used, study factors accounted for 47.2% of T1 variation in the $\hat{\boldsymbol{\xi}}$ condition and 44.5% of the T1 variation in the $\boldsymbol{\xi}$ condition. Furthermore, intercept SS decreased from 7.16 to 6.47 – though still far from the target value of 0.12. Dramatic changes were evident when $\boldsymbol{\theta}$ was used to

compute $Q1$, in which case study factors accounted for 8.0% of T1 variation in the $\hat{\xi}$ condition, 1.3% of the T1 variation in the ξ condition, and intercept SS decreased from 0.25 in the $\hat{\xi}$ condition to 0.14 in the ξ condition.

Figures comparing $Q1$ T1 rates at $\alpha = 0.10$ between PE conditions for each study cell can be found in Appendix D (Figures D-7 to D-8). Figures are presented only for $\alpha = 0.10$ because the $Q1$ T1 error scale (across all study conditions) was too large to informatively display T1 rates at lower α levels. Similar to QO, DN had a greater impact on T1 rates in the condition in which true model parameters were used to compute fit statistics; for example, at $\alpha = 0.05$, DN accounted for 42.1% of T1 error variation in the $\xi, \boldsymbol{\theta}$ condition and less than 1% of variation in the other PE conditions. This effect is not as evident in the appendix figures as it was for QO due to the expanded y-axis scale for $Q1$. The differential effect for model across PE conditions is more evident in the figures. This effect accounted for 71.7% of T1 variation in the $\hat{\xi}, \boldsymbol{\theta}$ condition, 10.6% in the $\xi, \boldsymbol{\theta}$ condition, and only near 1% in the two PE conditions involving $\hat{\boldsymbol{\theta}}$. Finally, test length had a very large impact on T1 rates when $\hat{\boldsymbol{\theta}}$ was used, where it accounted for near 70% of T1 variation, but not when $\boldsymbol{\theta}$ was used, where it accounted for 11% and < 1% of T1 variation in the $\hat{\xi}, \boldsymbol{\theta}$ and $\xi, \boldsymbol{\theta}$ conditions respectively.

Table 9. *Q*1 Type I Error Rate Sums of Squares and Effect Sizes for Model Intercept and Study Factors by Parameter Estimation Error Condition

	Parameter Estimation Error Condition							
	$\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$		$\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$		$\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$		$\boldsymbol{\xi},\boldsymbol{\theta}$	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
$\alpha = 0.01$								
Intercept	3.5322	0.3836	0.0123	0.8876	3.1537	0.4097	0.0070	0.9661
Factors	5.6759	0.6164	0.0016	0.1124	4.5430	0.5903	0.0002	0.0339
TOTAL	9.2080		0.0138		7.6966		0.0073	
$\alpha = 0.05$								
Intercept	7.1640	0.5285	0.2486	0.9195	6.4710	0.5547	0.1363	0.9874
Factors	6.3901	0.4715	0.0218	0.0805	5.1941	0.4453	0.0017	0.0126
TOTAL	13.5541		0.2703		11.6651		0.1380	
$\alpha = 0.10$								
Intercept	10.3386	0.6309	0.9107	0.9372	9.3904	0.6547	0.5134	0.9923
Factors	6.0484	0.3691	0.0610	0.0628	4.9531	0.3453	0.0040	0.0077
TOTAL	16.3869		0.9717		14.3436		0.5174	

The impact of PE on T1 rates was similar across the *z* statistics (see Tables 10–12). Unlike *Q*1, the use of $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\xi}}$ resulted in a dramatic deflation of intercept SS indicating that T1 rates tended to fall below nominal levels. However, relative to the $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ condition, this condition did not show reduction in the proportion of T1 variation due to study factors for L_z and *V*I; As with *Q*O, but less marked, some reduction was evident for *V*O. The use of $\boldsymbol{\xi}$ with $\hat{\boldsymbol{\theta}}$ did result in slightly less influence of study factors, as with *Q*1, but intercept SS actually increased somewhat (rather than decreasing as it did for *Q*1) in this condition relative to the $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ condition. Only in the $\boldsymbol{\xi},\boldsymbol{\theta}$ condition did SS values approach the target, suggesting that the *z* statistics approximate their theoretical distributions when there is no error in the model parameters.

Table 10. L_z Type I Error Rate Sums of Squares and Effect Sizes for Model Intercept and Study Factors by Parameter Estimation Error Condition

	Parameter Estimation Error Condition							
	$\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$		$\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$		$\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$		$\boldsymbol{\xi},\boldsymbol{\theta}$	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
$\alpha = 0.01$								
Intercept	3.6917	0.4335	0.0003	0.4194	4.0914	0.5072	0.0075	0.9154
Factors	4.8235	0.5665	0.0004	0.5806	3.9755	0.4928	0.0007	0.0846
TOTAL	8.5151		0.0006		8.0668		0.0082	
$\alpha = 0.05$								
Intercept	6.8638	0.5226	0.0104	0.4861	7.9436	0.6469	0.1486	0.9717
Factors	6.2698	0.4774	0.0110	0.5139	4.3362	0.3531	0.0043	0.0283
TOTAL	13.1335		0.0214		12.2798		0.1530	
$\alpha = 0.10$								
Intercept	9.1431	0.5734	0.0562	0.5460	11.0235	0.7315	0.5622	0.9854
Factors	6.8016	0.4266	0.0468	0.4540	4.0455	0.2685	0.0084	0.0146
TOTAL	15.9447		0.1030		15.0691		0.5706	

Table 11. *V*I Type I Error Rate Sums of Squares and Effect Sizes for Model Intercept and Study Factors by Parameter Estimation Error Condition

	Parameter Estimation Error Condition							
	$\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$		$\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$		$\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$		$\boldsymbol{\xi},\boldsymbol{\theta}$	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
$\alpha = 0.01$								
Intercept	3.8049	0.4447	0.0003	0.4226	4.2083	0.5212	0.0071	0.9334
Factors	4.7520	0.5553	0.0004	0.5774	3.8656	0.4788	0.0005	0.0666
TOTAL	8.5568		0.0007		8.0738		0.0076	
$\alpha = 0.05$								
Intercept	6.7570	0.5256	0.0108	0.4992	7.8988	0.6543	0.1480	0.9760
Factors	6.0987	0.4744	0.0108	0.5008	4.1739	0.3457	0.0036	0.0240
TOTAL	12.8558		0.0216		12.0728		0.1517	
$\alpha = 0.10$								
Intercept	8.9177	0.5761	0.0575	0.5580	10.8265	0.7359	0.5643	0.9864
Factors	6.5623	0.4239	0.0455	0.4420	3.8861	0.2641	0.0078	0.0136
TOTAL	15.4800		0.1030		14.7126		0.5721	

Table 12. *V*O Type I Error Rate Sums of Squares and Effect Sizes for Model Intercept and Study Factors by Parameter Estimation Error Condition

	Parameter Estimation Error Condition							
	$\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$		$\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$		$\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$		$\boldsymbol{\xi},\boldsymbol{\theta}$	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
$\alpha = 0.01$								
Intercept	3.5520	0.4195	0.0017	0.7269	3.9035	0.4810	0.0078	0.9633
Factors	4.9156	0.5805	0.0007	0.2731	4.2123	0.5190	0.0003	0.0367
TOTAL	8.4676		0.0024		8.1158		0.0081	
$\alpha = 0.05$								
Intercept	6.8316	0.5158	0.0180	0.7272	7.5230	0.6100	0.0989	0.9517
Factors	6.4123	0.4842	0.0068	0.2728	4.8097	0.3900	0.0050	0.0483
TOTAL	13.2439		0.0248		12.3327		0.1040	
$\alpha = 0.10$								
Intercept	9.2993	0.5749	0.0700	0.7126	10.4559	0.6935	0.3582	0.9509
Factors	6.8752	0.4251	0.0283	0.2874	4.6215	0.3065	0.0185	0.0491
TOTAL	16.1746		0.0983		15.0773		0.3767	

Figures comparing L_z and *V*O T1 rates at $\alpha = 0.10$ between PE conditions for each study cell can be found in Appendix D (Figures D-9 to D-12). The figures clearly depict the deflationary effect of coupling accurate person parameters (either $\hat{\boldsymbol{\theta}}$ when *N* is large, or $\boldsymbol{\theta}$) with less accurate item parameter estimates, especially for the 2PL and 3PL. Figures are not presented for *V*I since it functioned essentially the same as L_z (see empirical sampling distribution results for the *z* statistics). Unlike for the *Q* statistics, ξ estimation error seemed to cause T1 rate deflation. This result, coupled with the T1 rate inflation caused by θ estimation error accounts for the interesting T1 patterns in Tables 6 and 7 where T1 rates were quite deflated for the larger *n* condition. Furthermore, the result that this deflation was more extreme for *N* = 500 than *N* = 1,500 is not surprising, since, in smaller samples, less θ error would be accumulated in the computation of the *z* statistics.

For all fit statistics, there was little aberrance in T1 rates when statistics were computed with true model parameters, especially in SU conditions. For example, at $\alpha = 0.05$, T1 rates in SU and EU conditions ranged, respectively, from (1) 0.046 to 0.052 and 0.047 to 0.072 for *Q*1, (2) 0.046 to 0.056 and 0.048 to 0.065 for *Q*O, (3) 0.047 to 0.052

and 0.048 to 0.082 for L_z, (4) 0.046 to 0.053 and 0.051 to 0.078 for *V*I, and (5) 0.029 to 0.050 and 0.032 to 0.073 for *V*O. Tables of T1 rates at all three α levels for all $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions can be found in Appendix D (Tables D-6 to D-11).

Type I Error Rate Summary

There appeared to be little perturbance in T1 rates for all fit statistics (except for LM, which were not examined in that condition) when no parameter estimation error was present. However, in realistic conditions, when both person and item parameters were used to compute fit statistics, *Q*O was the only statistic that adhered closely to nominal T1 rates across the conditions of this study. T1 rates for the other fit statistics were affected by all study factors other than DN and were, in some conditions, too inflated or deflated for practical use of the statistics. Though T1 rates were near nominal levels for some of these statistics in some conditions (i.e., when $n = 75$ for *Q*1 with the 2PL and 3PL; when $n = 75/N = 500$ for the z statistics with the 1PL; and for LM(α) and LM(β) when discrimination was lower), the influence of test length, sample size, IRT model, and item discrimination indicate that there is no guarantee that T1 rates will be maintained in somewhat different conditions. The results presented above indicate that *Q*O should be a useful statistic for fit screening across a range of test conditions, though it might have somewhat higher than nominal T1 rates when used with the 3PL and somewhat lower than nominal rates when item parameters are poorly estimated (i.e., when tests are very long and samples are small).

Fit Statistic Empirical Sampling Distributions

Most of the fit statistics clearly did not follow their theoretical distributions, based upon the aberrances in their T1 rates. However, an examination of their sampling distributions is still informative because it can lead to better understanding of how these statistics are malfunctioning. Furthermore, the observation that T1 rates were near nominal levels for *Q*O and some of the other fit statistics in some conditions provides evidence, but not proof, that theoretical sampling distributions are being followed; strong

evidence can only be obtained through a direct examination of empirical sampling distributions.

*Q*O

Because categories (*K*) were collapsed for *Q*O in order to maintain expected counts ≥ 1, *K* varied across items, and hence theoretically items came from χ^2 distributions with different degrees of freedom (*df*). For each study condition, frequencies of items within each *K* (N_K) can be found in Appendix E (Tables E-1 to E-8). In a few rare cases, *Q*O could not be computed because *K* became too small, resulting in zero or negative *df*. Across all study conditions, this occurred for only 14 of the *Q*O statistics computed and was attributable to items with the most extreme difficulty and discrimination parameters.

Observed *K* was related to item parameter values. The 1PL items at the higher and lower difficulty tended to have fewer *K* and items in the middle difficulty range tended to have more *K*. For the 2PL, this was tempered by discrimination: items at the difficulty extremes with moderate to high discrimination tended to have fewer *K*. For the 3PL, only lower difficulty items tended to have fewer *K*, though again this was tempered by discrimination: high difficulty items did not require as much category collapsing because of the presence of non-zero lower asymptotes that prevent expected counts from being near zero. For both the 2PL and 3PL, items with low discrimination also tended to have more *K*. These relationships are depicted in Figures E-1 to E-12 of Appendix E.

Relationship between QO moments and K. Graphical depictions of *Q*O empirical sampling distribution first and second moments were created in which de-trended moments (i.e., Mean *Q*O − *df* and Var *Q*O − 2*df*) were plotted across *K*. The theoretical moments were subtracted from the empirical moments in order to allow for a more space-economical presentation of the results. No crucial information is lost in this representation since the *level* of Mean *Q*O or Var QO (which theoretically will increase across *K*) is not of central interest; what is of most interest is the *degree* to which the empirical moments depart from expectation.

Figures were inspected across all study conditions for discernible patterns regarding any relationship between *Q*O moments and *K*. Only patterns that were apparent

across both EU and SU conditions are discussed since, due to lack of evidence for DN impact, EU and SU conditions were considered to essentially be replications. In general, patterns were very similar across levels of discrimination. Only those cases in which patterns notably differed between the two levels are mentioned below.

Figures 5–8 present mean plots for SU high discrimination conditions, and plots for the other study conditions can be found in Tables E-13 to E-24 of Appendix E. When $N = 500$ and $n = 15$ (Figures 5, E-13, E-17, E-18), for both the 1PL and 2PL the means remained largely within or very near the limits of the confidence intervals (CIs) across K, except for the 2PL where means were above the upper limit in the highest few K. The means for the 3PL were consistently higher than expectation and above the upper limit of the CIs. When N was large (Figures 6, E-14, E-19, E-20) the tendency for 2PL QO to exceed the upper CI limit was not apparent. In both sample size conditions, the 1PL showed the fewest aberrations and the 3PL showed the most. For $N = 500$, across all $n = 15$ study conditions, the percentage of times that the mean fell outside of the CIs for the 1PL, 2PL, and 3PL were 25.0%, 37.5%, and 87.5% respectively; when $N = 1{,}500$, the respective percentages were 11.1%, 12.5%, and 100%.

When $N = 500$ and $n = 75$ (Figures 7, E-15, E-21, E-22), like the previously discussed condition for $n = 15$, the mean for the 2PL tended to exceed expectation at higher K. The positive bias for the 3PL was less apparent then when $n = 15$, with the means being largely within the CIs except for larger K. When N was large (Figures 8, E-16, E-23, E-24), these patterns became less marked. In fact, in the high discrimination condition, the pattern was not evident for the 2PL and barely evident for the 3PL. Like the smaller n conditions, in both sample size conditions, the 1PL showed the fewest aberrations and the 3PL showed the most. For $N = 500$, across all $n = 75$ study conditions, the percentage of times that the mean fell outside of the CIs for the 1PL, 2PL, and 3PL were 5.5%, 16.8%, and 43.1% respectively; when $N = 1{,}500$ the respective percentages were 3.4%, 8.1%, and 26.0%.

Figure 5. De-trended *QO* Means by *K* for SU High Discrimination *N* = 500 *n* = 15 Condition

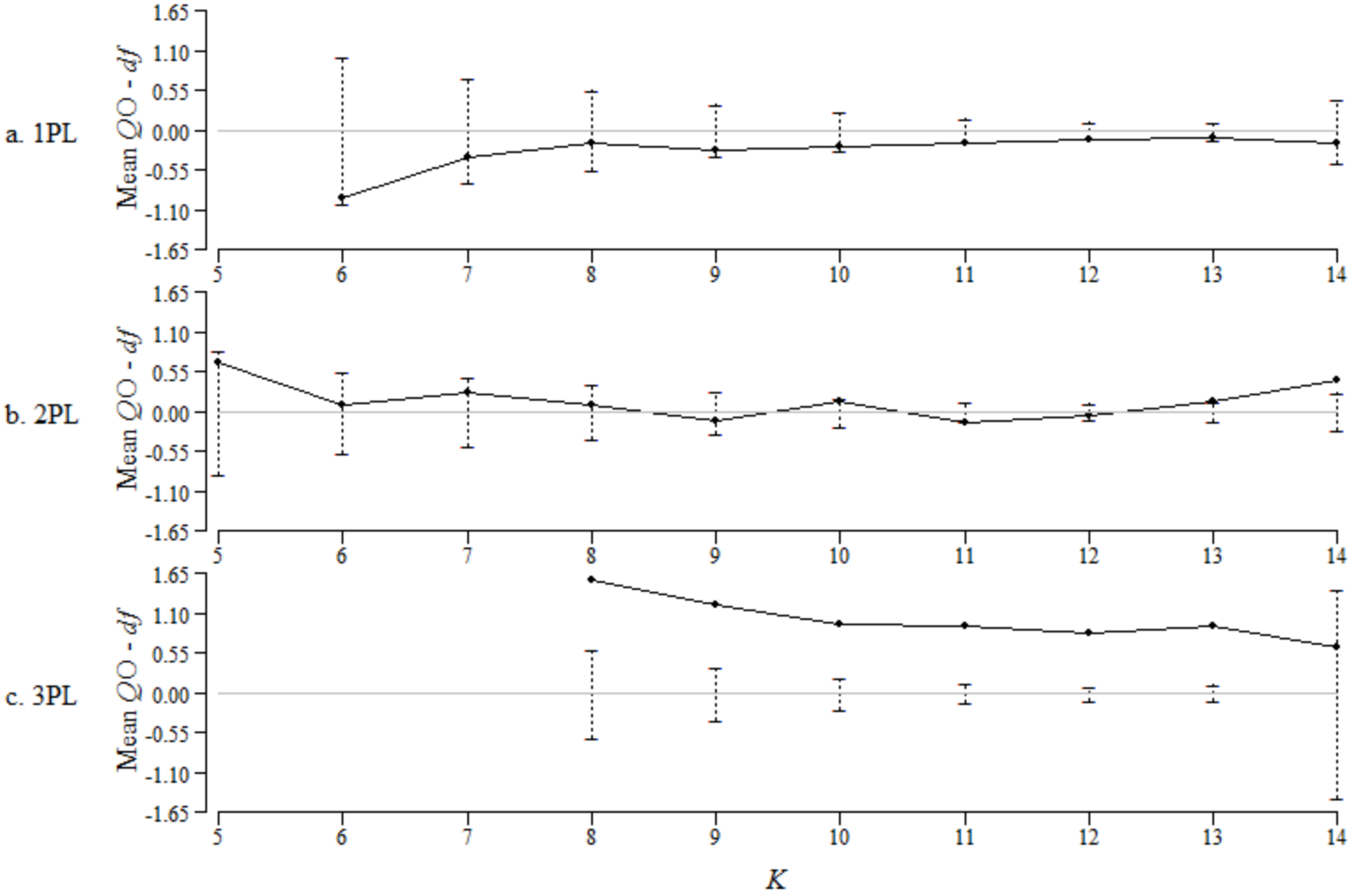


Figure 6. De-trended *Q*O Means by *K* for SU High Discrimination *N* = 1,500 *n* = 15 Condition

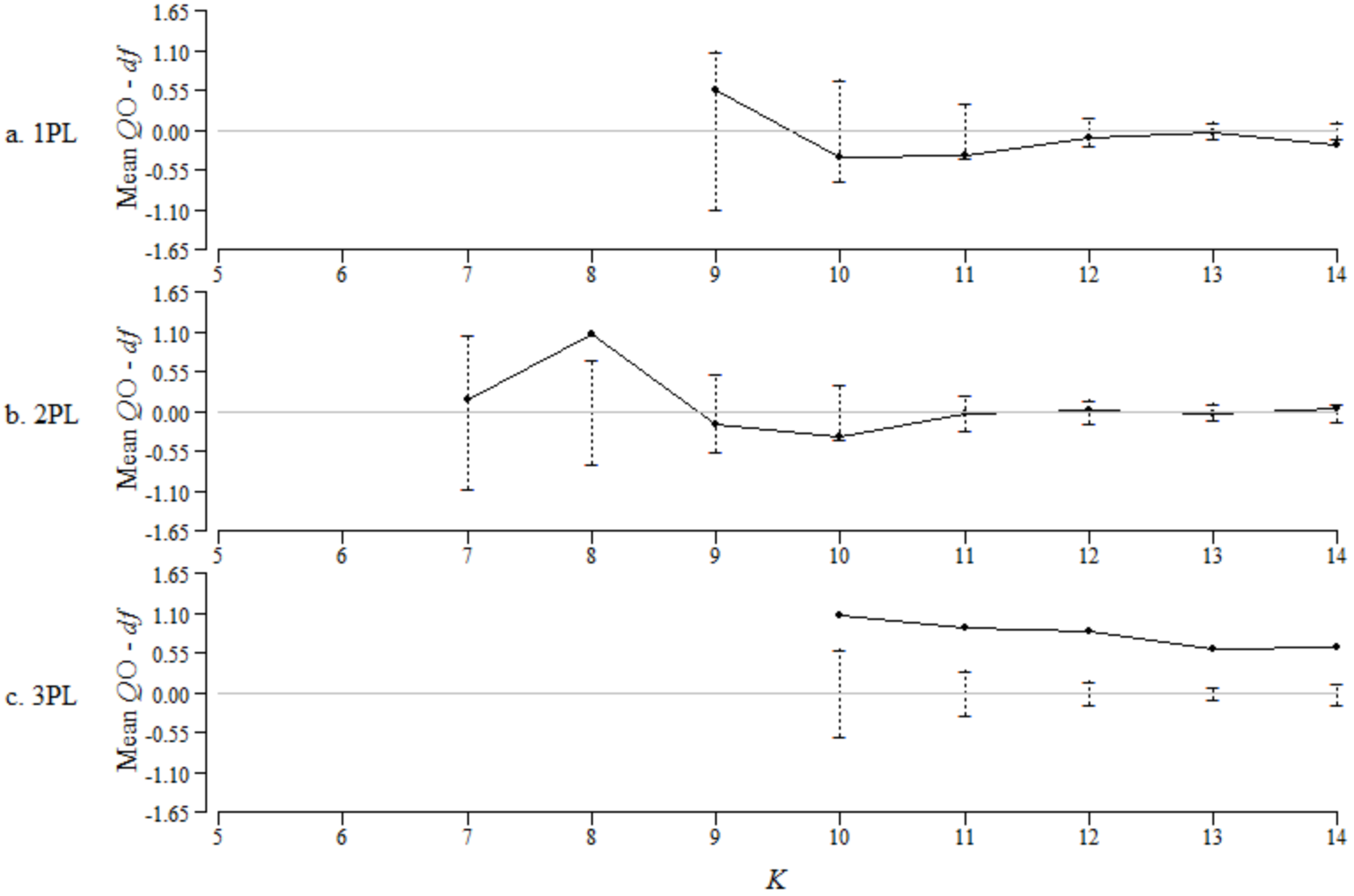


Figure 7. De-trended *QO* Means by *K* for SU High Discrimination *N* = 500 *n* = 75 Condition

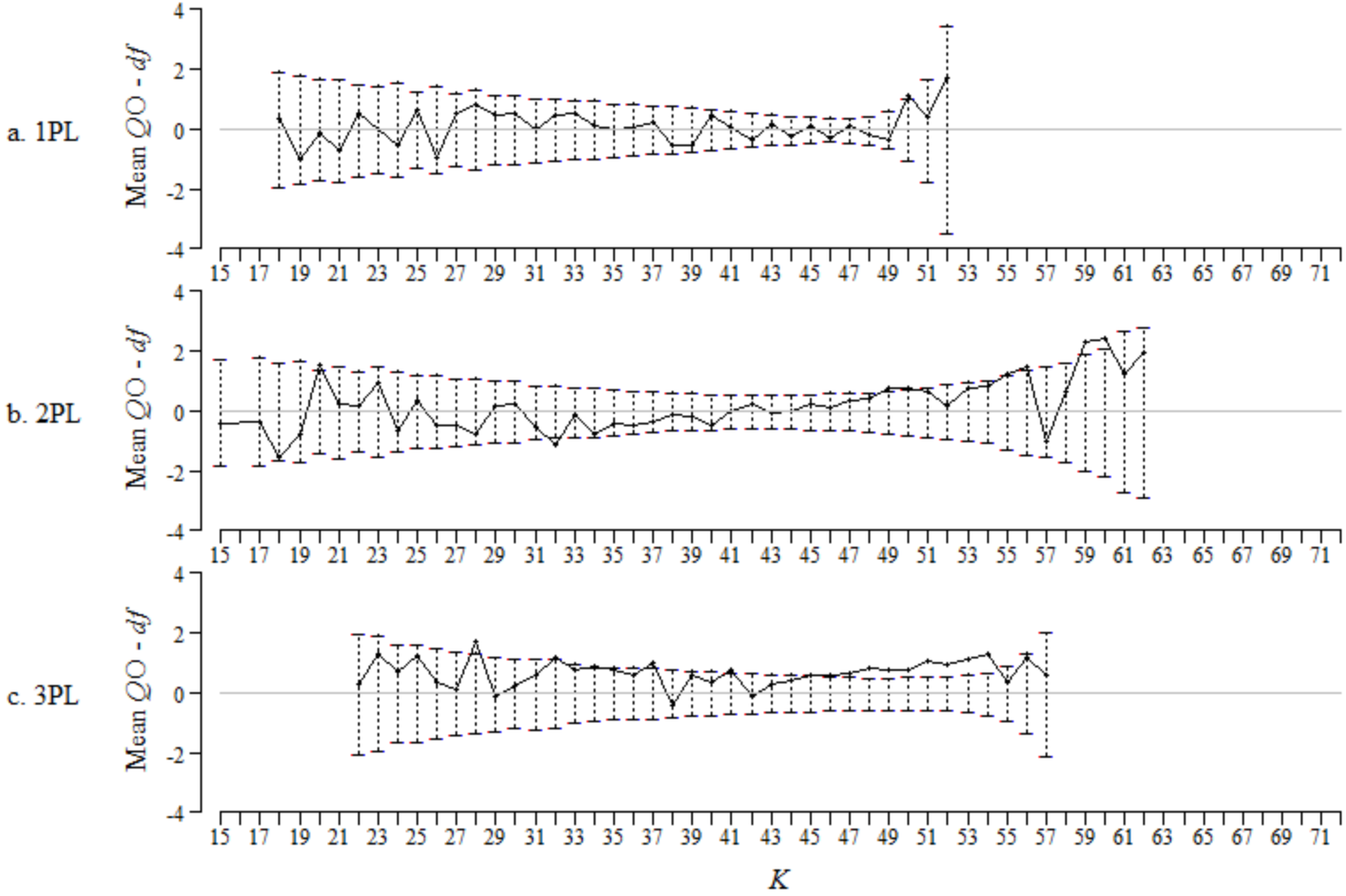
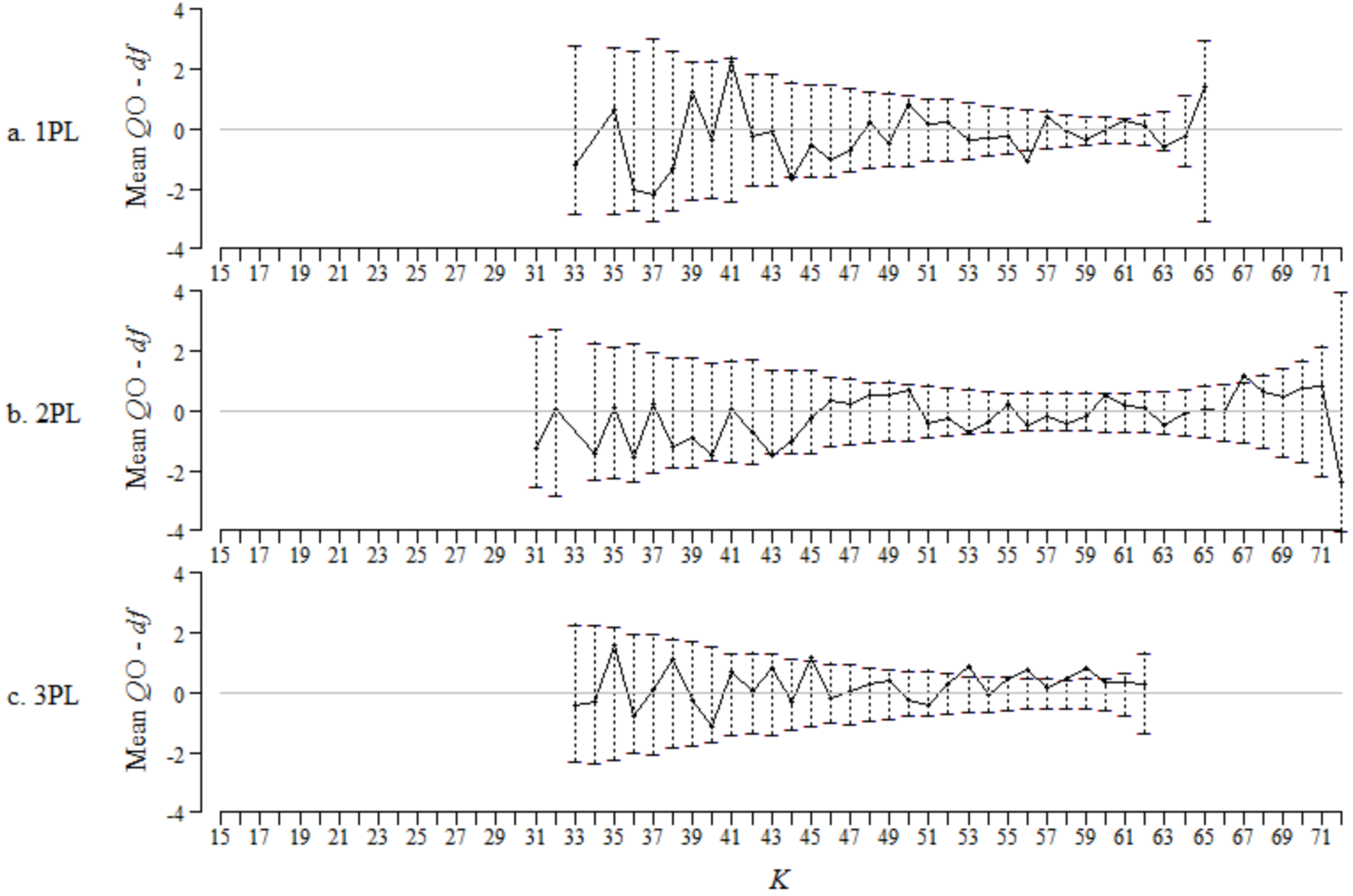


Figure 8. De-trended *Q*O Means by *K* for SU High Discrimination *N* = 1,500 *n* = 75 Condition



Figures 9–12 present variance plots for SU high discrimination conditions, and plots for the other study conditions can be found in Tables E-25 to E-36 of Appendix E. When n = 15 and N = 500, QO variance tended to be consistently somewhat higher than expectation for the 3PL in the low discrimination condition (Figures E-25 and E-30). When N was large, variances for the 3PL were fairly consistently somewhat above expectation at both levels of discrimination (Figures 10, E-26, E-31, E-32). In the large N condition (when n = 15), the percentage of variances that were outside their CIs were higher for the 3PL (at 26.3%), and to a lesser extent the 1PL (at 22.2%), than the 2PL (at 12.5%). However, when N was smaller, this pattern reversed, with the 2PL having a higher percentage (at 37.5%) than the 1PL and 3PL, which were both at 25.0%.

When N = 500 and n = 75, variances tended to be somewhat consistently below expectation across K for the 2PL (Figures 11, E-27, E-33, E-34), falling slightly below the lower CI limit for many K. When N = 1,500 (Figures 12, E-28, E-35, E-36), the consistent bias for 2PL variances to fall below expectation was less apparent. The bias for variance to be consistently somewhat above expectation that was evident for the 3PL when n = 15 was not present when n = 75 (Figures 11, 12, E-27, E-28, E-33 – E-36). In fact, in the smaller sample size, 3PL QO variance tended to be more likely to fall below expectation (Figures 11, E-27, E-33, E-34). For N = 500 (when n = 75), across all study conditions, the percentage of times that variances fell outside of the CIs for the 1PL, 2PL, and 3PL was 11.9%, 36.2%, and 32.8% respectively; when N = 1,500 the respective percentages were 2.2%, 12.1%, and 8.3%. Clearly, variances were less aberrant when N was larger and less aberrant for the 1PL than the other two models, though this difference became less marked when N was large.

Figure 9. De-trended *Q*O Variance by *K* for SU High Discrimination *N* = 500 *n* = 15 Condition

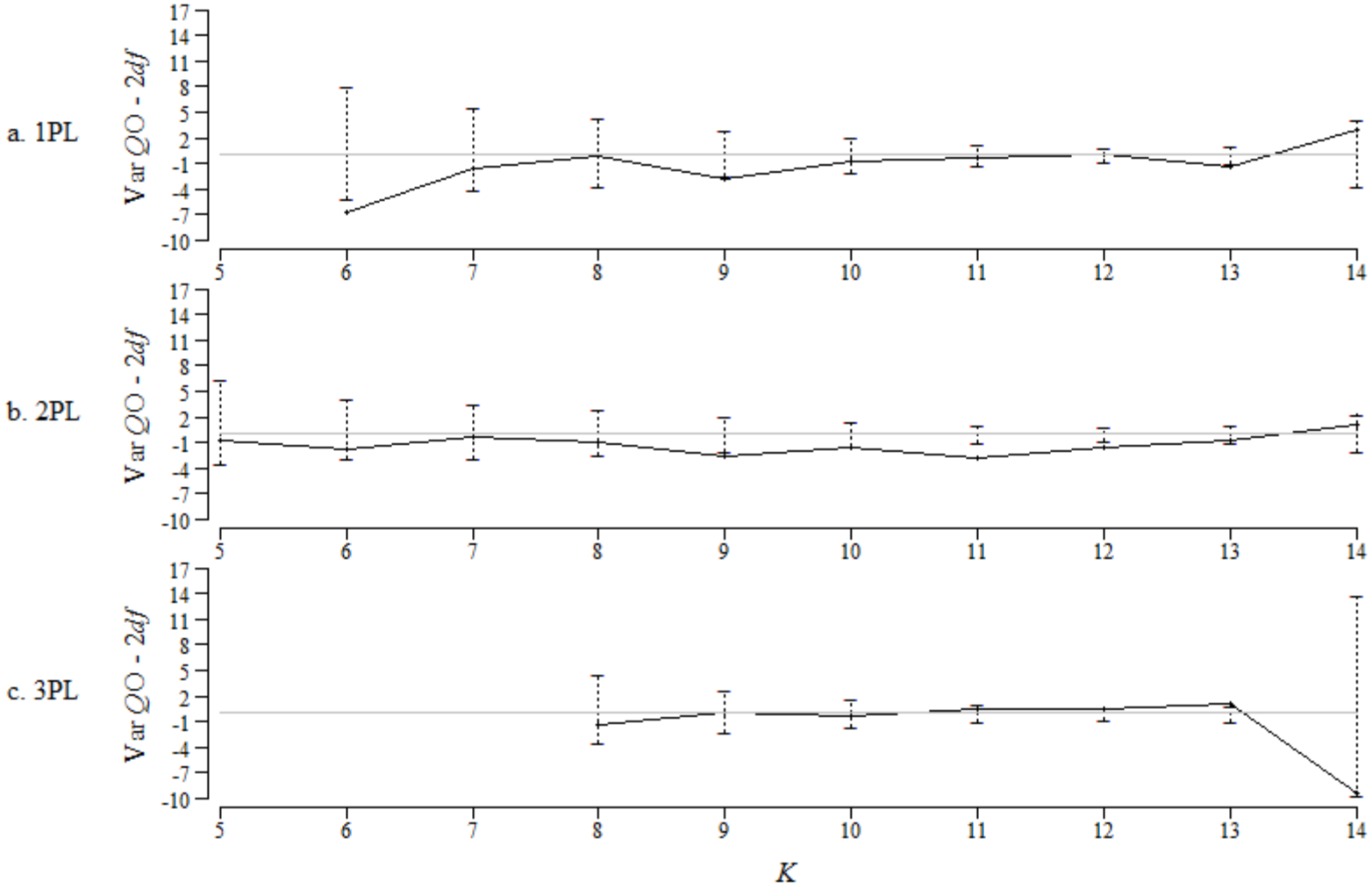


Figure 10. De-trended *Q*O Variance by *K* for SU High Discrimination *N* = 1,500 *n* = 15 Condition

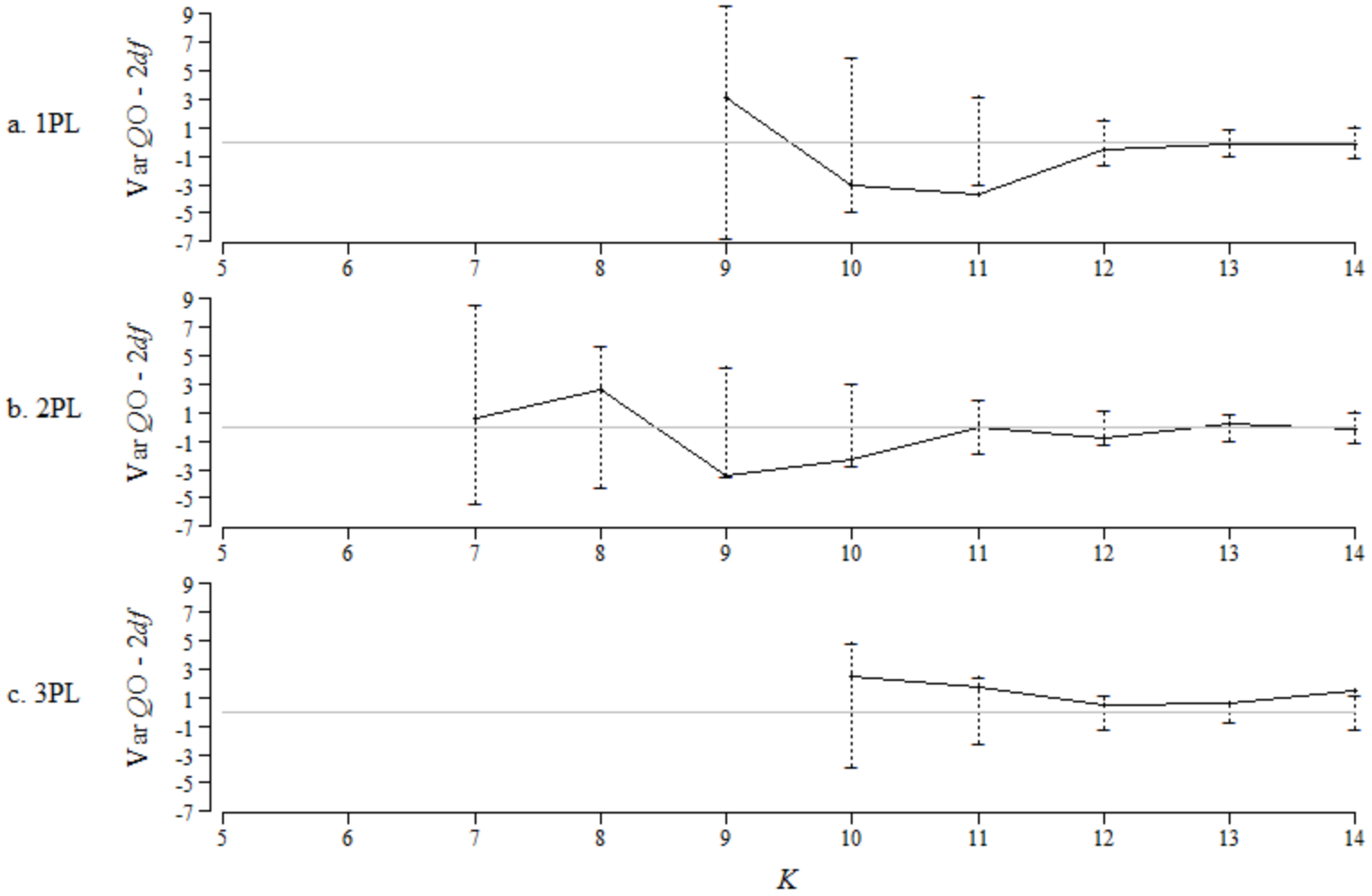


Figure 11. De-trended *Q*O Variance by *K* for SU High Discrimination *N* = 500 *n* = 75 Condition

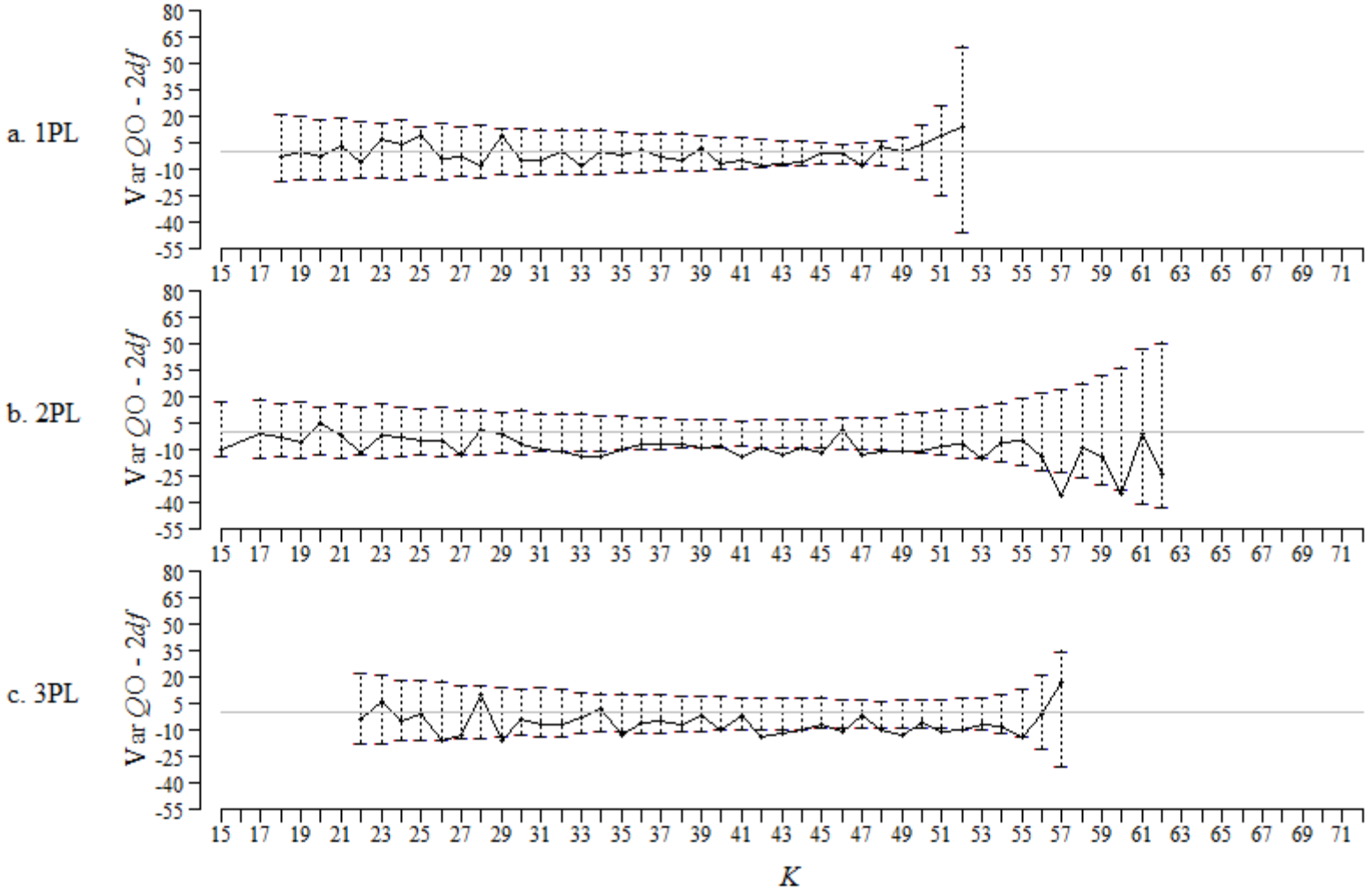
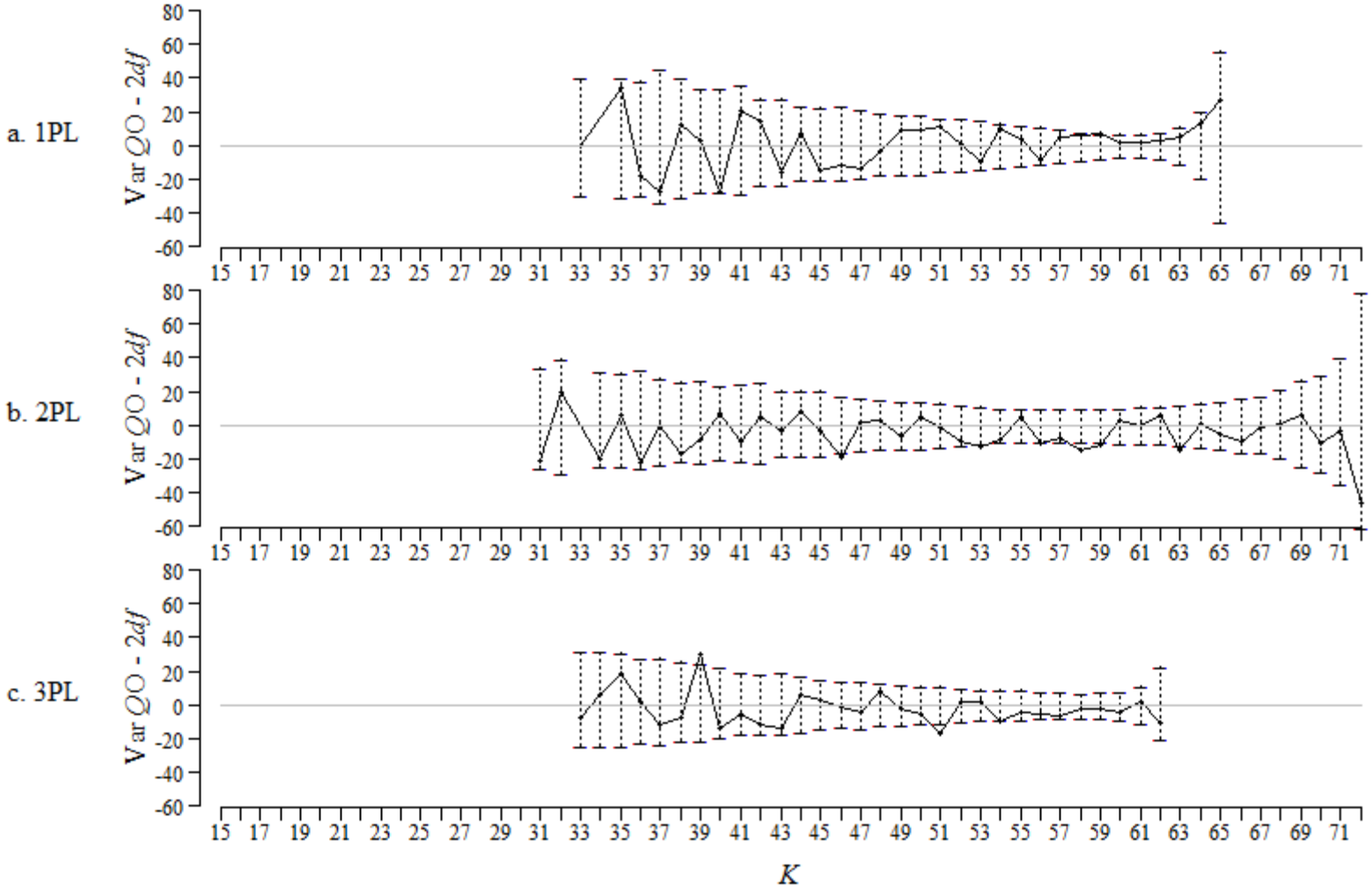


Figure 12. De-trended *Q*O Variance by *K* for SU High Discrimination *N* = 1,500 *n* = 75 Condition



Kolmogorov-Smirnov tests. The frequency (F_K) and proportion (π_K) of tests across K (where $N_K \geq 15$) in which the null hypothesis was rejected can be found in Tables 13 and 14. In general, π_K for each model (Model Total rows in Tables 13 and 14) were quite a bit higher than 0.05, exceeding 0.10 in all cases, except for the 1PL when n = 75 and the 2PL when n = 75 and N = 1,500 (Table 14). Generally, π_K increased with model complexity, which is consistent with the previous analysis of sampling distribution means that showed the most aberrant results for the 3PL, fewer for the 2PL, and fewer still for the 1PL. KS rejections were also more likely to occur when n = 15, however, this could be due, at least partially, to an artifact of N_K since KS test power increases with sample size and the n = 15 conditions tended to have much larger N_K than the n = 75 conditions. However, there was a rather large effect size for n (η^2 = 0.21) in the ANOVA of QO T1 rates, so it does appear justified to conclude that QO is more likely to depart from its theoretical distribution when n is small as opposed to larger.

Table 13. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that QO Followed Its Theoretical Distribution (n = 15)

			N = 500				N = 1,500			
M	DN	D	F_K	π_K	No. $N_K \geq 15$	Avg N_K	F_K	π_K	No. $N_K \geq 15$	Avg N_K
1PL	EU	High	4	0.444	9	2081	2	0.286	7	2676
		Low	0	0.000	5	3748	0	0.000	4	4688
	SU	High	4	0.444	9	2081	2	0.286	7	2677
		Low	1	0.200	5	3748	0	0.000	4	4688
	Model Total		**9**	**0.321**	**28**	**2676**	**4**	**0.182**	**22**	**3408**
2PL	EU	High	3	0.273	11	1703	1	0.111	9	2082
		Low	4	0.667	6	3123	0	0.000	5	3750
	SU	High	5	0.455	11	1704	1	0.111	9	2082
		Low	4	0.667	6	3124	1	0.200	5	3749
	Model Total		**16**	**0.471**	**34**	**2205**	**3**	**0.107**	**28**	**2678**
3PL	EU	High	6	0.857	7	2676	4	0.667	6	3124
		Low	4	0.800	5	3749	4	1.000	4	4686
	SU	High	6	0.750	8	2343	6	1.000	6	3124
		Low	5	1.000	5	3749	6	1.000	6	3124
	Model Total		**21**	**0.840**	**25**	**2999**	**20**	**0.909**	**22**	**3408**
Grand Total			**46**	**0.529**	**87**	**2585**	**27**	**0.375**	**72**	**3124**

Table 14. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that *Q*O Followed Its Theoretical Distribution (n = 75)

			N = 500				N = 1,500			
M	DN	D	F_K	π_K	No. $N_K \geq 15$	Avg N_K	F_K	π_K	No. $N_K \geq 15$	Avg N_K
1PL	EU	High	4	0.103	39	479	2	0.059	34	549
		Low	2	0.083	24	779	0	0.000	15	1248
	SU	High	1	0.027	37	505	1	0.029	34	549
		Low	1	0.042	24	779	3	0.200	15	1248
	Model Total		**8**	**0.065**	**124**	**603**	**6**	**0.061**	**98**	**763**
2PL	EU	High	6	0.111	54	346	2	0.043	46	405
		Low	6	0.188	32	583	2	0.080	25	748
	SU	High	10	0.189	53	352	2	0.044	45	414
		Low	10	0.333	30	621	6	0.250	24	779
	Model Total		**32**	**0.189**	**169**	**442**	**12**	**0.086**	**140**	**534**
3PL	EU	High	13	0.333	39	478	3	0.086	35	534
		Low	14	0.560	25	748	7	0.350	20	936
	SU	High	16	0.400	40	467	3	0.088	34	550
		Low	12	0.480	25	748	7	0.368	19	984
	Model Total		**55**	**0.426**	**129**	**579**	**20**	**0.185**	**108**	**693**
Grand Total			**95**	**0.225**	**422**	**531**	**38**	**0.110**	**346**	**648**

***Impact of parameter estimation error*.** When *Q*O was computed with true item parameters it appeared to better approximate its theoretical distribution according to the KS test results. In the ξ conditions, π_K tended to be closer to 0.05, especially in SU conditions. Aggregated across all n = 15 conditions, π_K in SU and EU ξ conditions were 0.06 and 0.27 respectively; this is considerably lower than π_K in n = 15 SU and EU $\hat{\xi}$ conditions, which were 0.51 and 0.41 respectively. Across all n = 75 conditions, π_K in SU and EU ξ conditions were 0.05 and 0.08 respectively; this is also somewhat lower than π_K in n = 75 SU and EU $\hat{\xi}$ conditions, which were 0.19 and 0.16 respectively. Disaggregated (across all study conditions except D) F_K and π_K can be found in Table E-9 of Appendix E.

Despite the higher than expected π_K in $\hat{\xi}$ conditions, *Q*O appeared to follow its theoretical sampling distribution to a reasonable degree. Across all $\hat{\xi}$ study conditions (1)

Bias(Mean) ranged from −0.40 to 0.11 for the 1PL, from −0.24 to 0.30 for the 2PL, and from 0.17 to 1.23 for the 3PL; (2) ME(Mean) ranged from 0.02 to 0.77 for the 1PL, from 0.11 to 0.75 for the 2PL, and from 0.49 to 1.23 for the 3PL; (3) Bias(SD) ranged from −0.41 to 0.18 for the 1PL, from −0.60 to 0.06 for the 2PL, and from −0.44 to 0.35 for the 3PL; and (4) ME(SD) ranged from 0.06 to 0.59 for the 1PL, from 0.12 to 0.69 for the 2PL, and from 0.09 to 0.52 for the 3PL. These ranges are not markedly farther from zero than those in the SU ξ conditions (in which π_K were most on target) where, across all M, N, n and D study conditions, Bias(Mean) ranged from −0.50 to 0.27, ME(Mean) ranged from 0.03 to 0.69, Bias(SD) ranged from −0.31 to 0.44, and ME(SD) ranged from 0.09 to 0.64. Tables of Bias and ME values in all study conditions can be found in Appendix E (Tables E-10 to E-17).

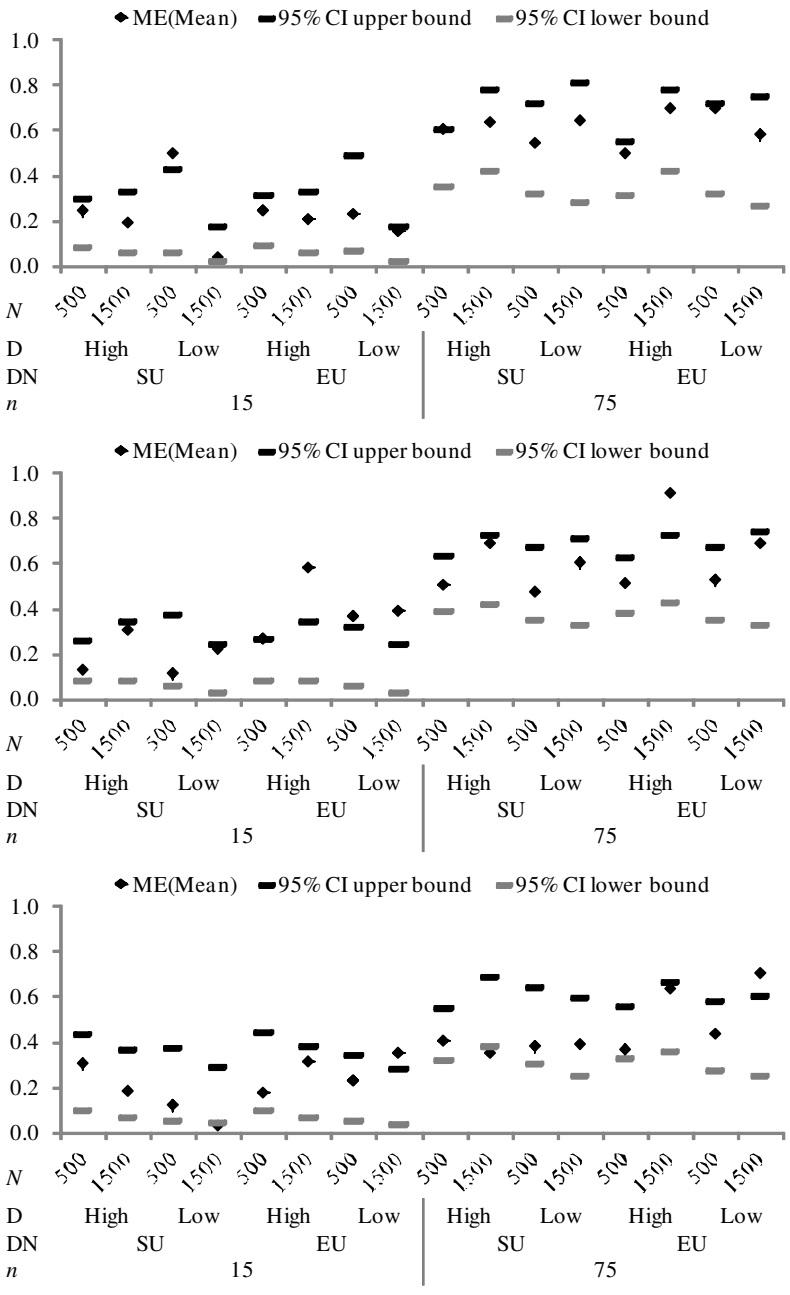
Though the magnitude of ME appeared relatively small, some marked differences were evident between $\hat{\xi}$ and ξ conditions when ME were inspected in relation to their CIs. Plots of ME(Mean) for QO in ξ conditions can be found in Figure 13. The effect for DN seen in the KS test results was again evident as ME(Mean) was less likely to fall above the upper bound of CIs in SU conditions (where this occurred in two of the 24 conditions) than in EU conditions (where this occurred in six of the 24 conditions). If QO truly followed its theoretical distribution, only one or two ME(Mean) estimates across the 48 study conditions should have exceeded the upper bound. However, the margins by which estimates exceeded the upper bounds were quite small, ranging from 0.001 to 0.072 (averaging 0.036) in SU conditions and 0.041 to 0.239 (averaging 0.130) in EU conditions.

Figure 13. Estimates of ME(Mean) and 95% CIs About the Estimates for *Q*O and All ξ Study Conditions

a. 1PL

b. 2PL

c. 3PL

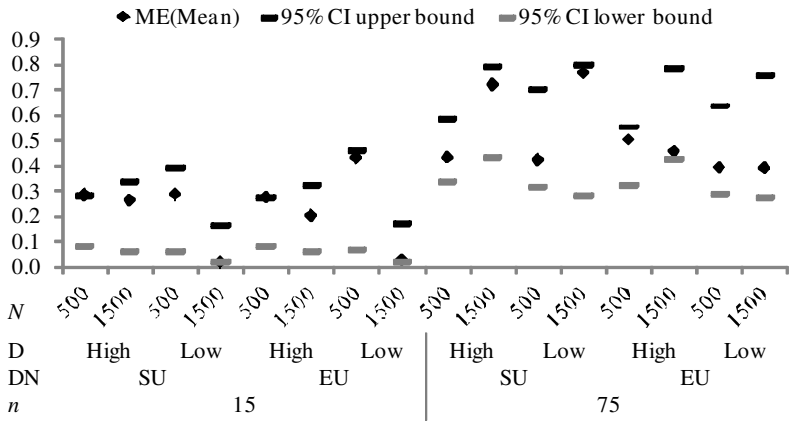


In $\hat{\xi}$ conditions, ME(Mean) was more likely to exceed CI upper limits, doing so for 12 of the SU conditions and nine of the EU conditions. The margins by which estimates exceeded the upper bounds were larger than in the ξ conditions, ranging from

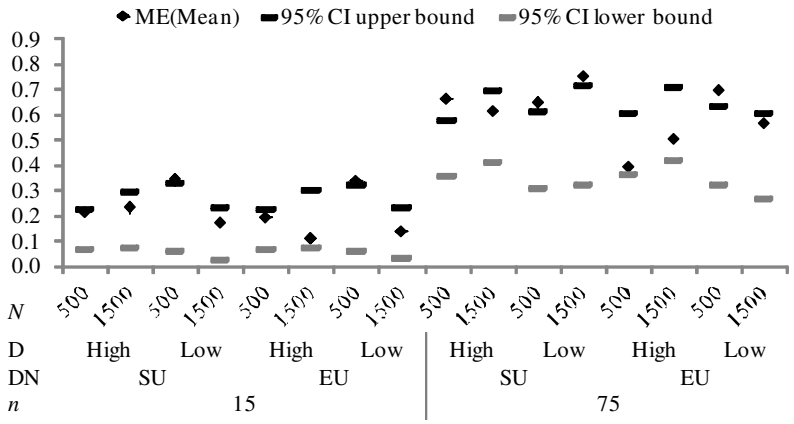
0.002 to 0.902 (averaging 0.286) in SU conditions and 0.004 to 0.727 (averaging 0.344) in EU conditions. However, as is evident in Figure 14, which contains plots of ME(Mean) in $\hat{\xi}$ conditions, the inflated ME were primarily attributable to the 3PL; ME(Mean) exceeded the CI upper limits in 50% (13/16) of the 3PL conditions, 37.5% (6/16) of the 2PL conditions, and 12.5% (2/16) of the 1PL conditions.

Figure 14. Estimates of ME(Mean) and 95% CIs About the Estimates for *Q*O and All $\hat{\xi}$ Study Conditions

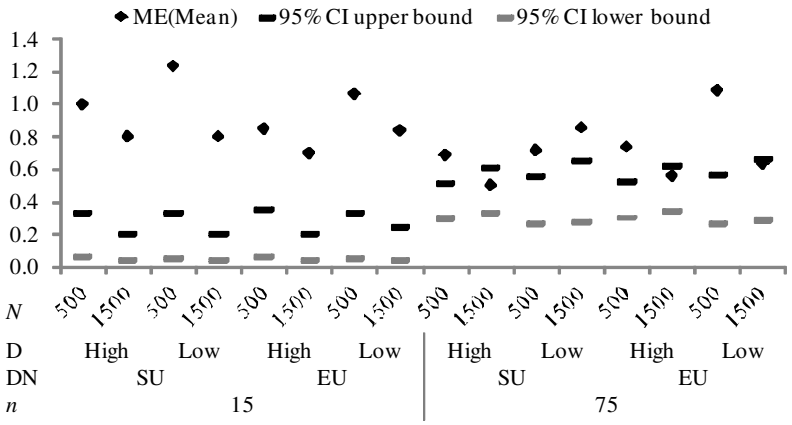
a. 1PL



b. 2PL



c. 3PL



ME(SD) did not markedly differ between $\hat{\xi}$ and ξ conditions and tended to be more erratic than expectation in both. Plots of ME(SD) in $\hat{\xi}$ and ξ conditions can be found in Appendix E (Figures E-37 and E-38). In ξ conditions ME(SD) fell above the CI upper limit in six of the SU conditions and 14 of the EU conditions; the margins by which these estimates exceeded the upper bounds were again quite small, ranging from 0.006 to 0.096 (averaging 0.045) in SU conditions and 0.002 to 0.225 (averaging 0.067) in EU conditions. In $\hat{\xi}$ conditions ME(SD) fell above the CI upper limit in eight of the SU conditions and five of the EU conditions; the margins by which these estimates exceeded the upper bounds was somewhat larger than in ξ conditions, ranging from 0.015 to 0.228 (averaging 0.071) in SU conditions and < 0.001 to 0.184 (averaging 0.076).

Finally, there was a noteworthy, though minor, deflationary impact of ξ estimation error on the variance of *Q*O's sampling distribution. In nearly every study condition, Bias(SD) in $\hat{\xi}$ conditions was less than Bias(SD) in ξ conditions (see Tables E-10 to E-13). To summarize this result, Table 15 presents Bias(SD) averaged across M and D conditions.

Table 15. Bias(SD) Averaged Across M and D Conditions

DN	PE	n = 15		n = 75	
		N = 500	N = 1,500	N = 500	N = 1,500
SU	$\hat{\xi}$	−0.17	−0.01	−0.36	−0.15
	ξ	0.01	0.14	0.02	0.26
EU	$\hat{\xi}$	−0.03	0.02	−0.42	−0.10
	ξ	0.16	0.29	0.03	0.43

Relationship between QO and item parameters. From inspection of Figures 5–12 and E-13 to E-36, there appeared to be little discernible relationship between the mean and variance of the *Q*O sampling distributions and *K*, except for the tendency for the 2PL means to be somewhat inflated at larger *K* in some conditions (when $N = 500$ or when $N = 1{,}500$ and discrimination was low). As a final step in determining the invariance of *Q*O's distribution, the relationship between item *Q*O values and item parameters was also examined. However, due to the fairly close relationship between *K* and item parameter values, and the lack of relationship between *K* and *Q*O moments, relationships between *Q*O and item parameters were not expected. Nonetheless, Pearson correlations (r) between *Q*O and item parameters within each *K* were obtained for all *K* with 30 or more items ($N_K \geq 30$). Correlations tended to be more extreme for smaller N_K, as might be expected since correlations are more influenced by outlying observations in smaller sample sizes. Figures 15–17 contain scatterplots between these correlations and N_K, which is plotted on a log scale in order to better represent the data at smaller N_K. In these figures. the correlations are scattered fairly symmetrically about zero, with the range decreasing as N_K increases. This pattern indicates lack of any systematic relationship between *Q*O and item parameters. Scatterplots between item parameters and *Q*O within each N_K were also inspected for non-linear relationships, but none were apparent.

Figure 15. Scatterplots Between $r(QO,b)$ and N_K (on Log10 Scale)

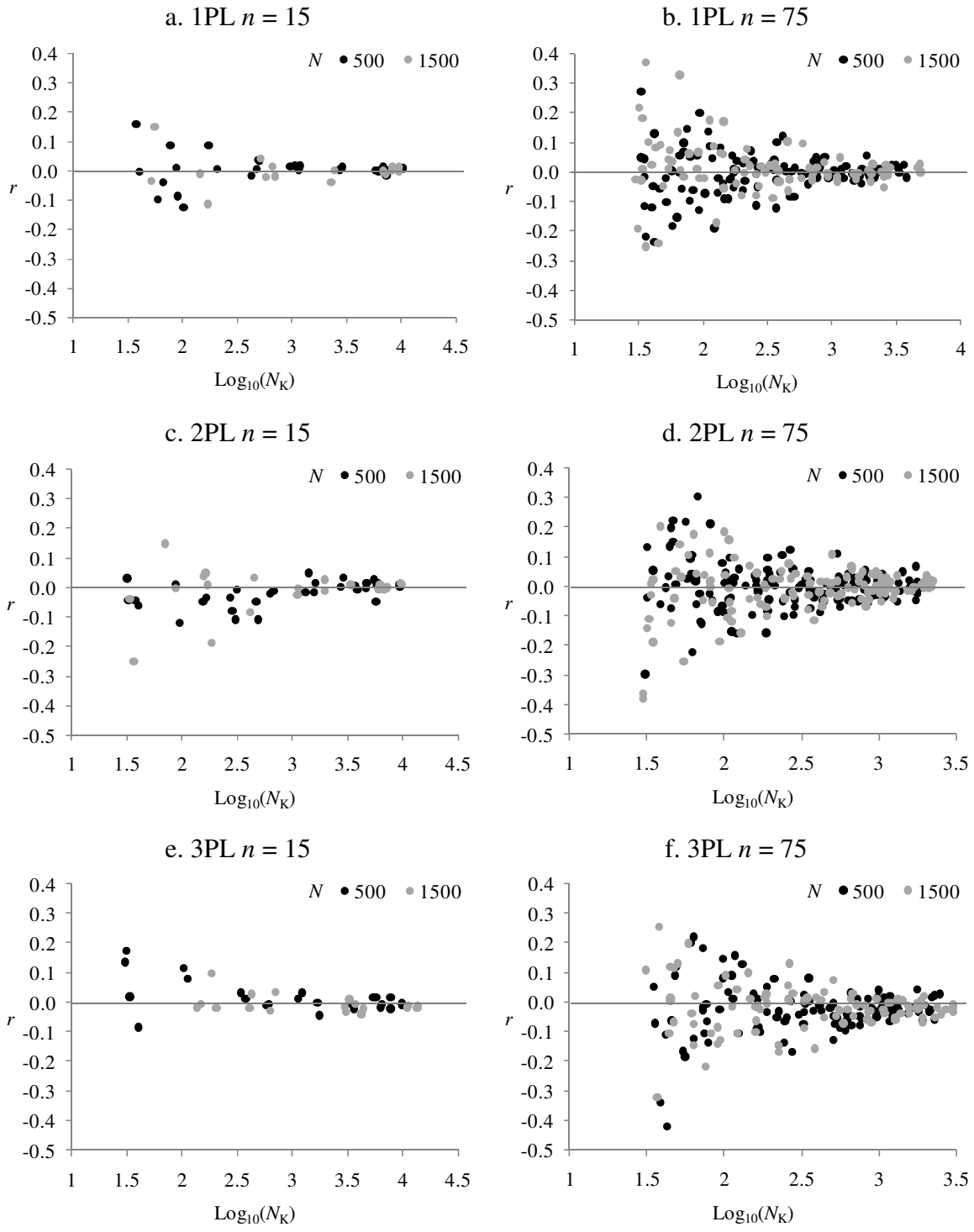


Figure 16. Scatterplots Between $r(QO,a)$ and N_K (on Log10 Scale)

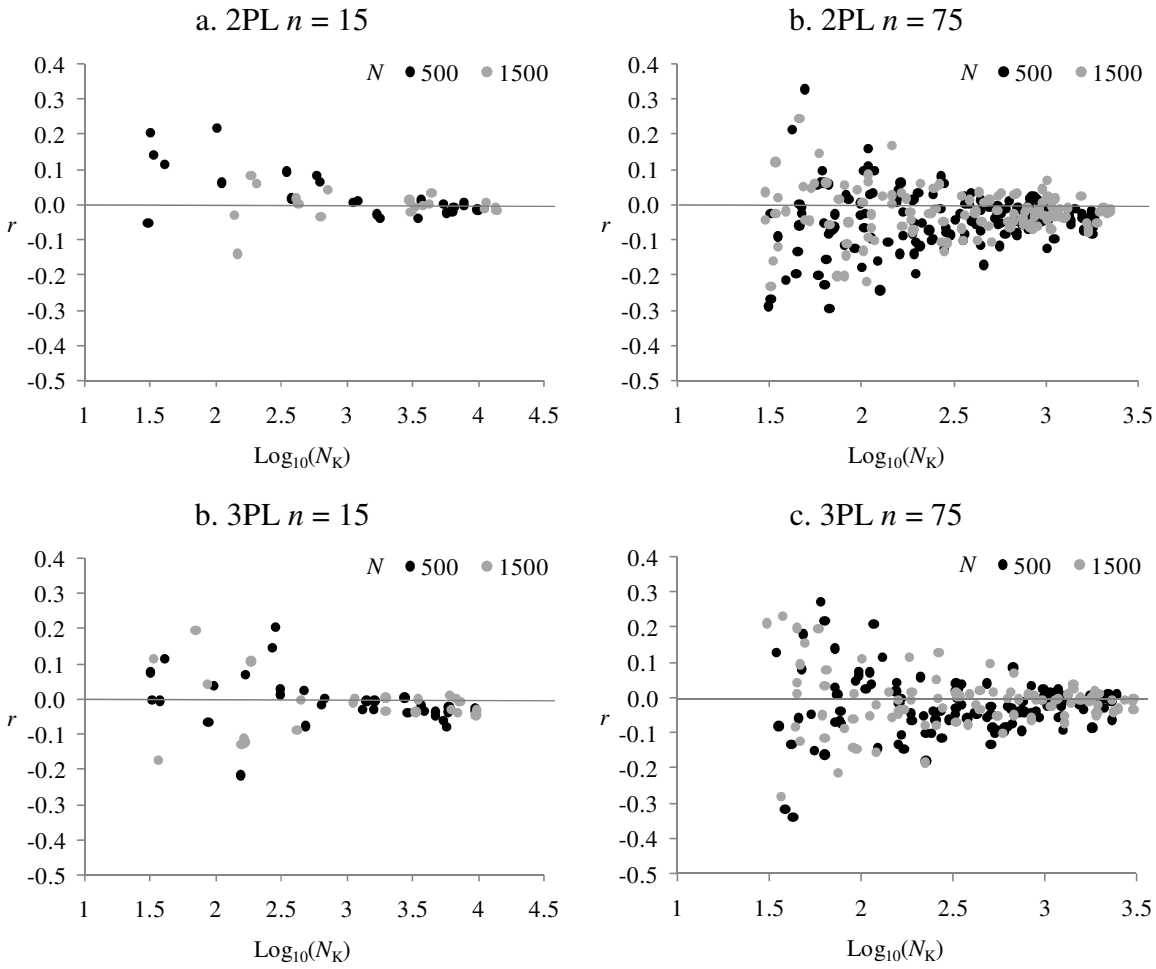
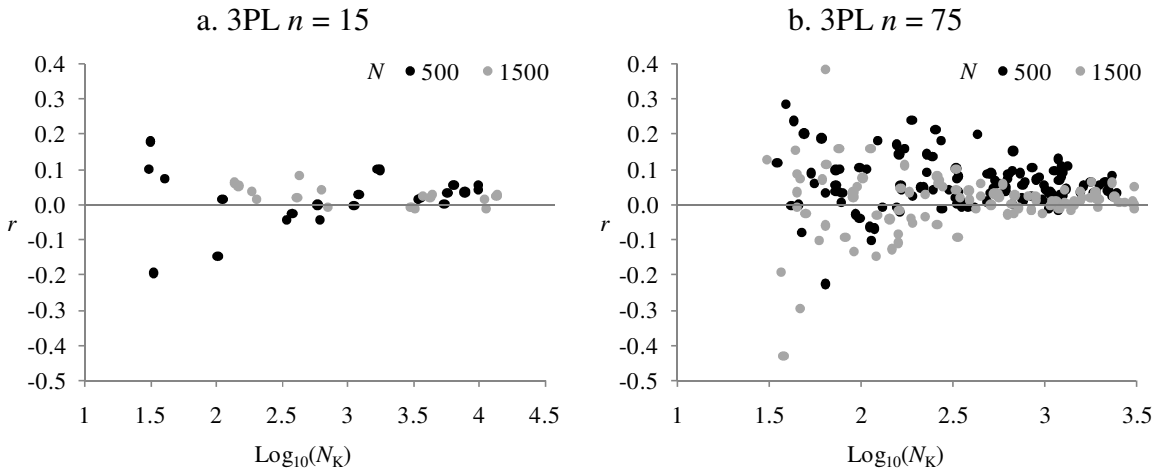


Figure 17. Scatterplots Between $r(QO,c)$ and N_K (on Log10 Scale)



QO summary. In the absence of item parameter estimation error and data noise, *Q*O appeared to generally approximate its theoretical distributions. If *Q*O strictly followed its theoretical distributions, ME(Mean) and ME(SD) should be expected to fall outside the limits of expectation in 0 to 1 (of the 24) SU conditions. In SU conditions ME(Mean) exceeded the limits of expectation in only 2 (of 24) study conditions (Figure 13) but ME(SD) (Figure E-37) showed more (6 of 24), but minor, departures from expectation. Nevertheless, the proportion (over all study conditions) of KS test rejections was near expectation, at 0.06.

The introduction of data noise in the absence of parameter estimation error led to a small but noticeable disruption in *Q*O's sampling distribution. ME(Mean) (Figure 13) and ME(SD) (Figure E-37) were somewhat more likely to be aberrant in EU than in SU conditions, exceeding the upper CI limit in 6 (of 24) conditions for ME(Mean) and 14 (of 24) conditions for ME(SD). An effect for DN was also present in KS test results, where, relative to SU conditions, π_K was elevated at 0.11.

When *Q*O was computed with estimated item parameters, its sampling distribution was further, though very mildly, disrupted and there was little noticeable impact of data noise. In SU and EU conditions ME(Mean) exceeded the limits of expectation in 12 (of 24) and 9 (of 24) study conditions respectively. In SU and EU

conditions ME(SD) exceeded the limits of expectation in 8 (of 24) and 5 (of 24) study conditions respectively. In both SU and EU conditions π_K (at 0.19 and 0.16 respectively) were somewhat higher than when no item parameter estimation error was present. Much of the sampling distribution disturbance was due to the 3PL for which, compared to the other two models, *Q*O means tended to be more inflated, especially when $n = 15$. Additionally, when there was item parameter estimation error present, *Q*O variances tended to become mildly deflated for all models. For example, in SU conditions, Bias(SD) averaged –0.16 when $\hat{\xi}$ were used as opposed to 0.09 when ξ were used. This deflation in variances was most marked when $n = 75$ and $N = 500$ and somewhat more prevalent for the 2PL than the other two models. Finally, there did not appear to be any marked relationship between *Q*O sampling distribution moments and *K* or between *Q*O and item parameter values.

*Q*1

Like *Q*O, items varied in the number of categories used to compute *Q*1 due to cell collapsing required to maintain expected values ≥ 1. Item parameters bore the same relationship to *K* for *Q*1 as they did for *Q*O. There was a somewhat large number of items for which *Q*1 *p* values were unavailable due to very small *K*, resulting in zero or negative *df*. This occurred for 52 of the *Q*1 statistics computed. For each study condition, frequencies of items within each *K* (N_K) can be found in Appendix F (Tables F-1 to F-8).

Relationship between Q1 moments and K. As for *Q*O, graphical depictions of *Q*1 empirical sampling distribution first and second moments were created in which de-trended moments were plotted across *K*. However, the plots for the sample variance were slightly modified for *Q*1. Because variances became very inflated for *Q*1 in some cases, the second moments were plotted in standard deviation units in order to decrease the y-axis scale to more informatively display the data. Charts for the sample means and SDs in the SU high discrimination conditions can be found in Figures 19–21. Appendix F contains charts for the sample means (Figures F-1 to F-8) and SDs (Figures F-9 to F-16) in all conditions.

From inspection of these plots it was apparent that $Q1$ distributions consistently tended to become less aberrant as K increased for the 3PL, but not the other two models. Not surprisingly (from the inflated T1 rates observed for $Q1$) sample means were consistently well above the upper CI limit in all instances, except for the 1PL when N = 500 and n = 75 and the 2PL in this condition at higher K. Variances were consistently well above expectation across K at n = 15 and fairly consistently somewhat below expectation at n = 75 for the 1PL and 2PL when N = 500. For the 3PL, sample variances, like the means, were inflated but decreased as K increased, and in fact, were at or very near expectation when n = 75.

Figure 18. De-trended $Q1$ Means and SDs by K for the SU High Discrimination $N = 500$ $n = 15$ Condition

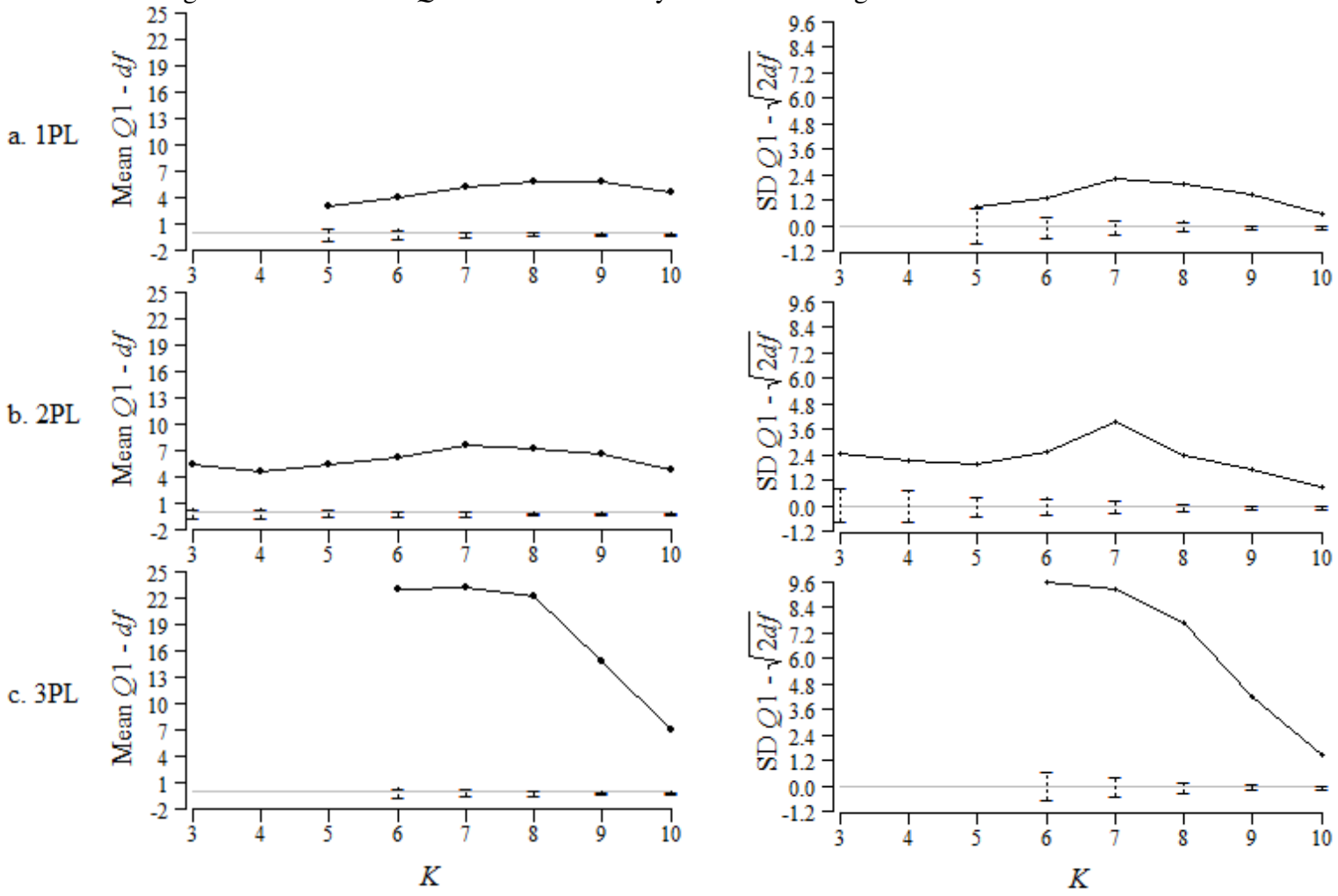


Figure 19. De-trended $Q1$ Means and SDs by K for the SU High Discrimination N = 1,500 n = 15 Condition

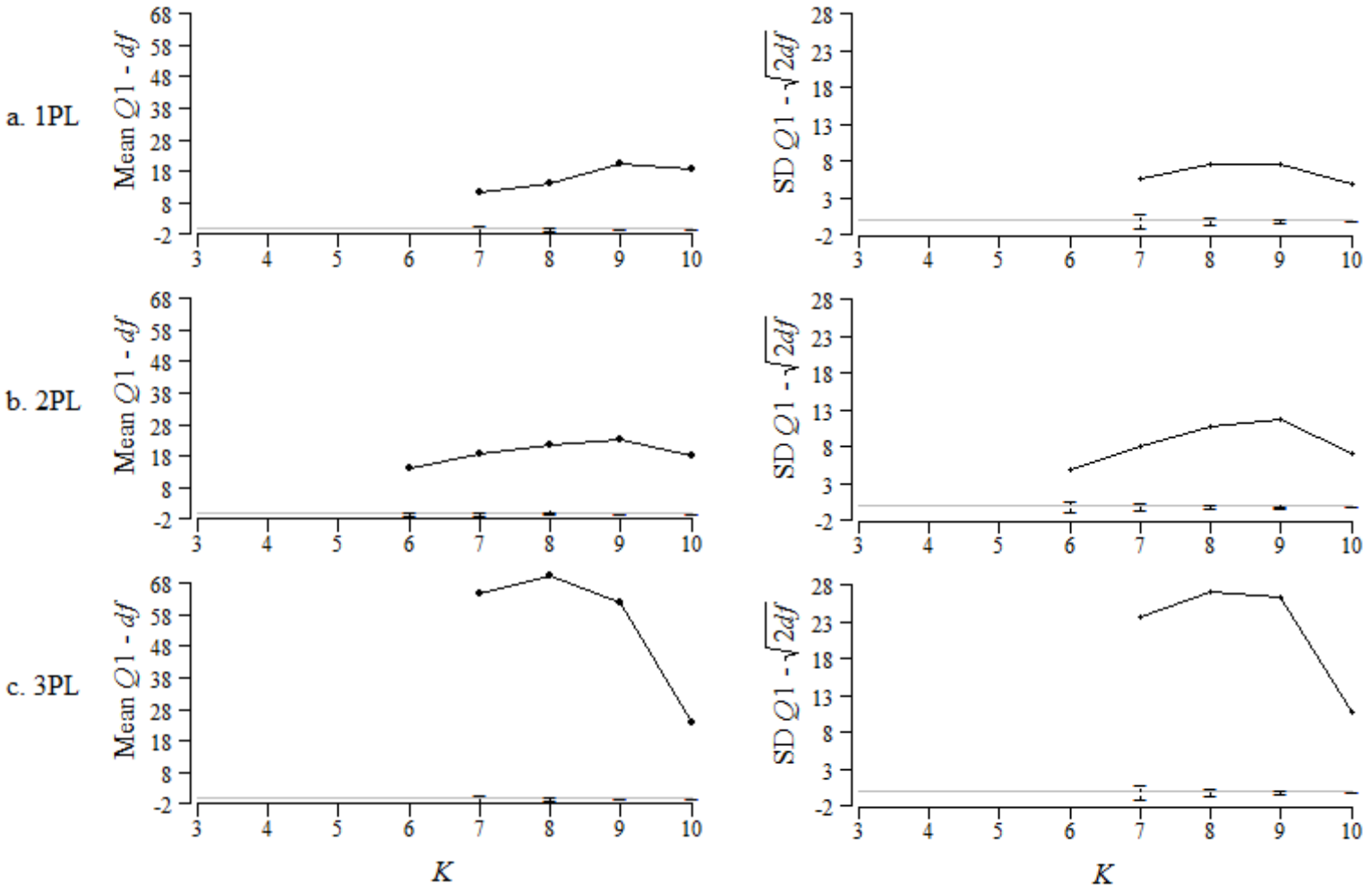


Figure 20. De-trended $Q1$ Means and SDs by K for the SU High Discrimination $N = 500$ $n = 75$ Condition

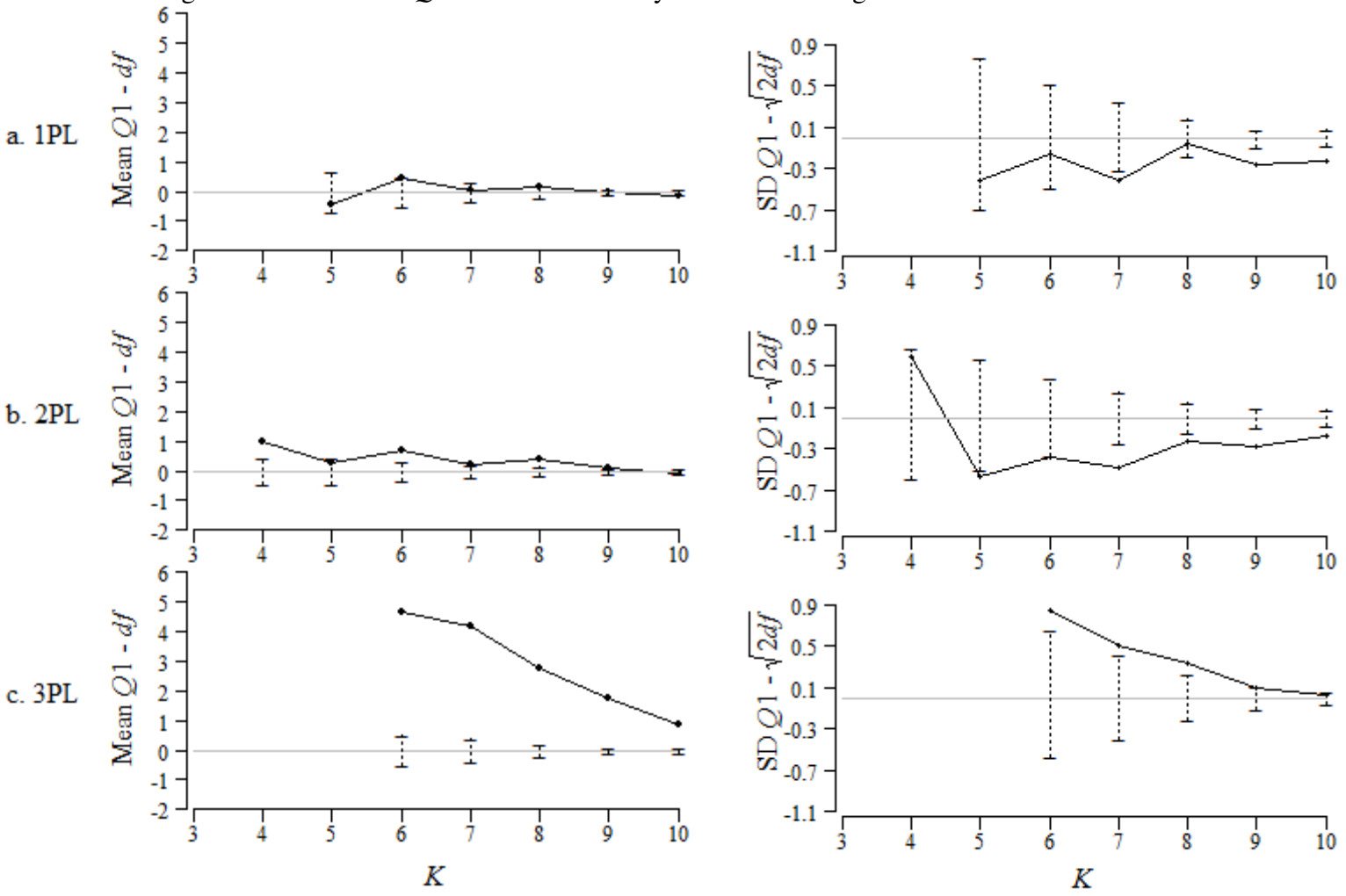
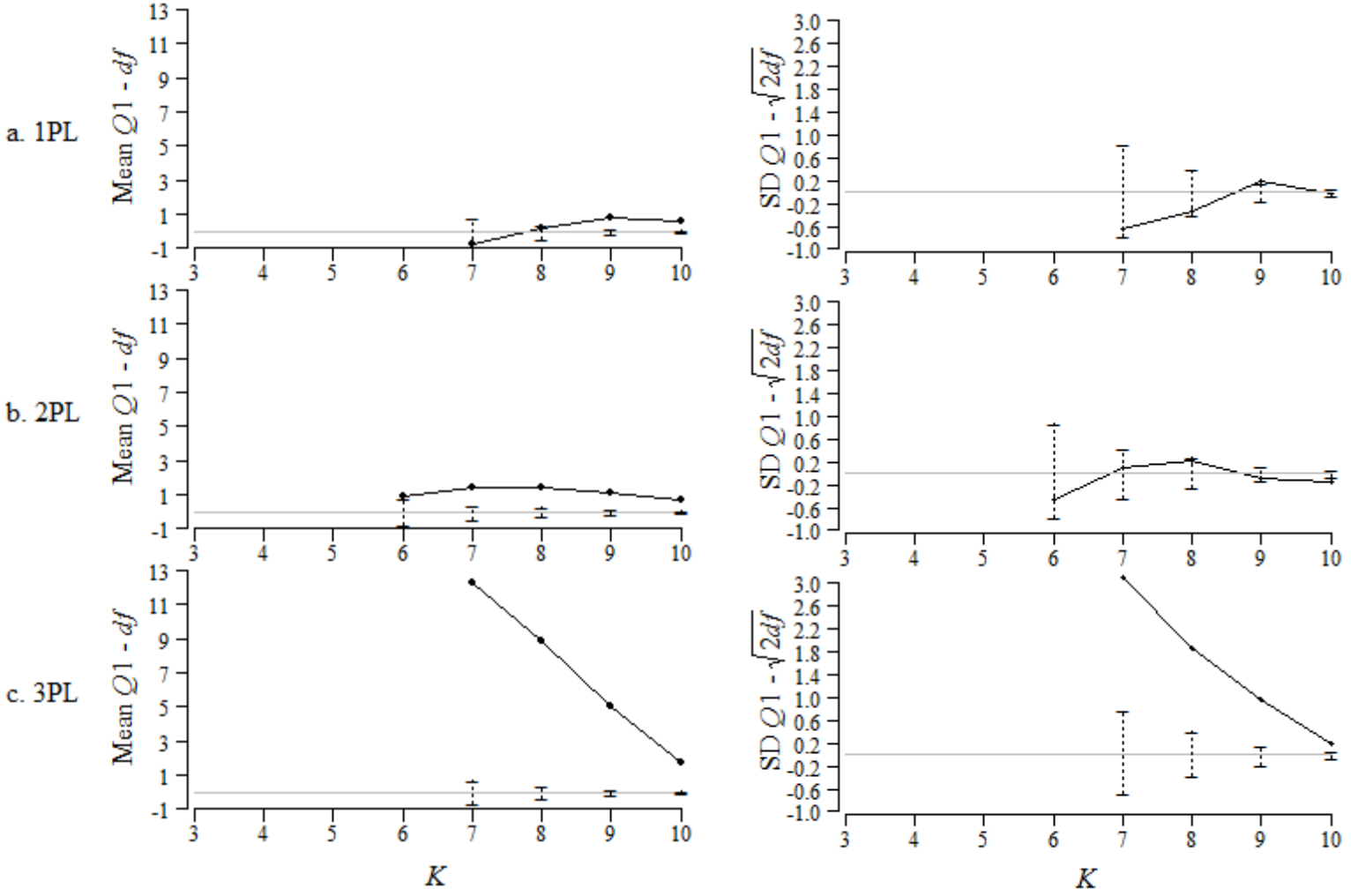


Figure 21. De-trended $Q1$ Means and SDs by K for the SU High Discrimination N = 1,500 n = 75 Condition



Kolmogorov-Smirnov tests. From the analysis of moments, $Q1$ clearly did not follow its theoretical distribution in most study conditions, and not surprisingly, all KS tests were statistically significant (at $\alpha = 0.05$) for all models when $n = 15$ and for the 3PL when $n = 75$. For the 1PL and 2PL when $n = 75$, where $Q1$ appeared to function near expectation, the proportion of KS tests across K (π_K) that were significant was less than 1.0 but still far from the 0.05 that would be expected if the statistics followed their theoretical distributions. When $N = 500$ π_K were 0.50 for the 1PL and 0.88 for the 2PL; when $N = 1{,}500$ they were 0.67 for the 1PL and 0.94 for the 2PL.

The significance of the KS tests in the cases where $Q1$ means and SDs appeared to function near expectation was likely due to the fact that there was no condition in which *both* the moments were on target. In those conditions where variances were closest to expectation (e.g. the 1PL and 2PL at the highest K when $n = 75$ and $N = 1{,}500$), the means, which were closest to expectation when $N = 500$, departed more from expectation.

Impact of parameter estimation error. Because there tended to be a great deal of bias in $Q1$ moments, the results presentation will focus on patterns of results with respect to Bias rather than ME. The values of Bias and ME in SU conditions can be found in Appendix F (Tables F-9 to F-12 for Bias and Tables F-13 to F-16 for ME). The first and second moments of the empirical sampling distributions tended to move closer to expectation when $\boldsymbol{\theta}$, as opposed to $\hat{\boldsymbol{\theta}}$, was used to compute $Q1$. The presence of θ estimation error caused inflations in both $Q1$ mean and variance. This effect was most evident when $n = 15$, where Bias(Mean) and Bias(SD) in $\hat{\boldsymbol{\theta}}$ conditions were consistently greater than in $\boldsymbol{\theta}$ conditions. When $n = 15$, Bias(Mean) ranged from 4.86 to 57.06 in $\hat{\boldsymbol{\theta}}$ conditions and from −0.30 to 2.07 in $\boldsymbol{\theta}$ conditions, and Bias(SD) ranged from 1.40 to 25.65 in $\hat{\boldsymbol{\theta}}$ conditions and from −0.24 to 0.52 in $\boldsymbol{\theta}$ conditions.

A smaller amount of inflation in $\hat{\boldsymbol{\theta}}$ conditions was evident in $n = 75$ $\boldsymbol{\xi}$ conditions where Bias(Mean) in $\hat{\boldsymbol{\theta}}$ conditions was consistently greater (but to a lesser degree than in $n = 15$ conditions) than in $\boldsymbol{\theta}$ conditions, ranging from 0.02 to 10.29 in $\hat{\boldsymbol{\theta}}$ conditions and from −0.36 to 0.15 in $\boldsymbol{\theta}$ conditions. Compared to $\boldsymbol{\theta}$ conditions, Bias(SD) was also

greater in all $\hat{\boldsymbol{\theta}}$ conditions (except one), ranging from 0.08 to 5.03 in $\hat{\boldsymbol{\theta}}$ conditions and from −0.21 to 0.21 in $\boldsymbol{\theta}$ conditions. Finally, the inflationary pattern was not as clear in the $n = 75$ $\hat{\xi}$ conditions, where only Bias(Mean) was consistently greater in $\hat{\boldsymbol{\theta}}$ than in $\boldsymbol{\theta}$ conditions when $N = 1{,}500$.

When $\hat{\boldsymbol{\theta}}$ was used to compute $Q1$, the presence of ξ estimation error had a consistent and more marked deflationary impact on $Q1$ SDs than means when $n = 15$. This was more evident in the high discrimination condition where, for example: (1) Bias(Mean) values for the 1PL in $\hat{\xi}$ and ξ conditions, respectively, were 4.86 and 6.15 whereas Bias(SD) doubled from 1.40 in $\hat{\xi}$ to 2.82 in ξ conditions; (2) Bias(Mean) values for the 2PL in $\hat{\xi}$ and ξ conditions respectively were 6.02 and 5.81, whereas Bias(SD) increased by a factor of 1.8, from 2.26 in $\hat{\xi}$ to 4.01 in ξ conditions; and (3) Bias(Mean) values for the 3PL in $\hat{\xi}$ and ξ conditions respectively were 18.07 and 17.73, whereas Bias(SD) increased by a factor of 1.6, from 6.43 in $\hat{\xi}$ to 10.09 in ξ conditions. When $n = 75$, the deflationary effect of ξ estimation error was still evident as Bias(SD) in $\hat{\xi}$ conditions was consistently less than Bias(SD) in ξ conditions, though many of the Bias(SD) values became negative in the $\hat{\xi}$ conditions (reaching −0.27).

When $\boldsymbol{\theta}$ was used to compute $Q1$, the consistent deflationary effect of ξ estimation error on SDs was not evident. In the $\boldsymbol{\theta}$ conditions, Bias(SD) was much closer to zero, ranging from −0.29 to 0.52 when $\hat{\xi}$ was used and from −0.22 to 0.28 when ξ was used; these ranges are much tighter than those in the $\hat{\boldsymbol{\theta}}$ conditions, which were: 1.40 to 21.94 across $n = 15$ $\hat{\xi}$ conditions, 1.83 to 25.65 across $n = 15$ ξ conditions, −0.27 to 1.53 across $n = 75$ $\hat{\xi}$ conditions, and 0.08 to 5.03 across $n = 75$ ξ conditions.

Finally, when $\boldsymbol{\theta}$ was used to compute $Q1$, the use of ξ, as opposed to $\hat{\xi}$, eliminated the trend for Bias(Mean) to increase with model complexity. For example, when $n = 15$ and $N = 500$, in the high discrimination $\hat{\xi}$ condition, Bias(Mean) values for the 1PL, 2PL, and 3PL were −0.01, 0.86, and 1.88, respectively, whereas in the ξ

condition the respective values were −0.26, −0.06, and −0.13. This pattern was evident across all other N, n, and discrimination conditions.

The π_K within each PE condition (in Table 16 aggregated across DN and D conditions) help summarize the impact of parameter estimation error on the $Q1$ sampling distributions. As a digression, it must be noted that DN did not appear to impact π_K in $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions as it did with QO; aggregated over all $\boldsymbol{\xi}, \boldsymbol{\theta}$ study conditions, QO π_K for SU and EU conditions were 0.06 and 0.11 respectively, whereas, $Q1$ π_K was 0.30 in both DN conditions.

The absence of θ and ξ estimation error had the strongest impact on the 3PL when N = 500, in which case only about 10% of KS tests were significant; for the 1PL and 2PL these percentages ranged from 21% to 33%. When N = 1,500, the percentage of KS test rejections ranged from 31% (for the 2PL when n = 15) to 57% (for the 3PL when n = 75). Tables with disaggregated counts of KS test rejections can be found in Appendix F (Tables F-17 to F-20).

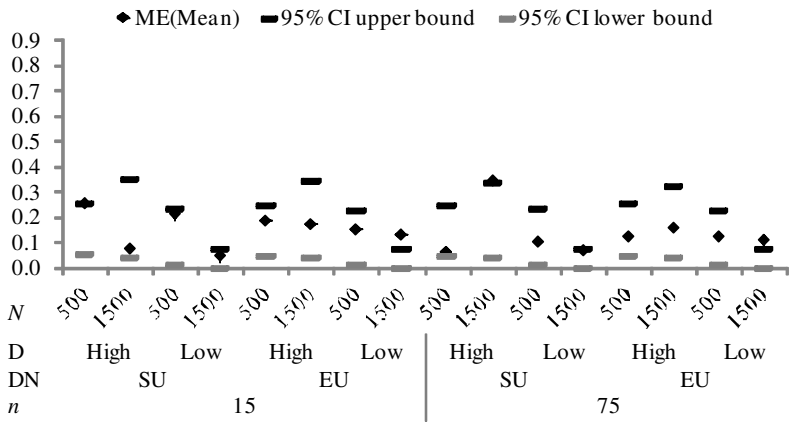
Table 16. π_K Across DN and D Conditions Within n, N, PE, and Model Conditions

n	N	PE		1PL		2PL		3PL	
		θ	ξ	π_K	No. $N_K \geq 15$	π_K	No. $N_K \geq 15$	π_K	No. $N_K \geq 15$
15	500	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	1.000	20	1.000	22	1.000	16
			$\boldsymbol{\xi}$	1.000	20	1.000	24	1.000	20
		$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.444	18	0.913	23	0.944	18
			$\boldsymbol{\xi}$	0.333	18	0.292	24	0.105	19
	1,500	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	1.000	13	1.000	16	1.000	12
			$\boldsymbol{\xi}$	1.000	14	1.000	16	1.000	12
		$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.700	10	0.875	16	1.000	12
			$\boldsymbol{\xi}$	0.417	12	0.313	16	0.385	13
75	500	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.500	18	0.875	24	1.000	18
			$\boldsymbol{\xi}$	0.556	18	0.417	24	0.900	20
		$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.444	18	0.917	24	1.000	18
			$\boldsymbol{\xi}$	0.222	18	0.208	24	0.100	20
	1,500	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.667	12	0.944	18	1.000	12
			$\boldsymbol{\xi}$	0.750	12	0.722	18	1.000	14
		$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.538	13	0.882	17	0.917	12
			$\boldsymbol{\xi}$	0.417	12	0.444	18	0.571	14

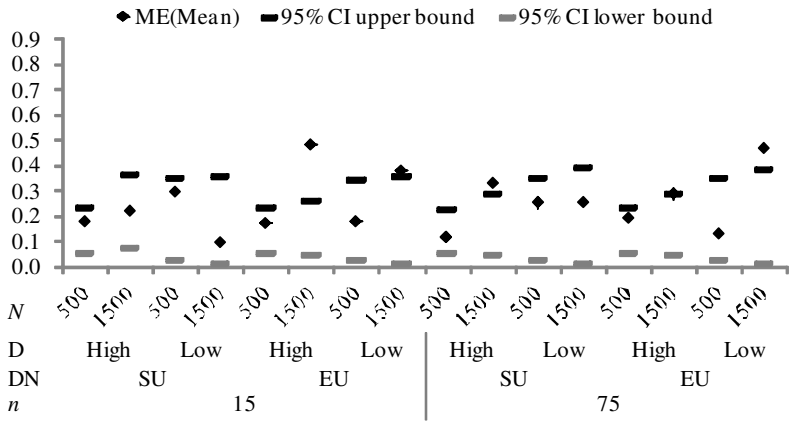
Though π_K still exceeded expectation even in $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions, the magnitude by which the means and SDs of $Q1$ departed from their expected values was relatively small. Plots of ME(Mean) and ME(SD) for $Q1$ in $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions can be found in Figures 22 and 23. Unlike the KS test results, there did appear to be some impact of DN on $Q1$'s sampling distribution. ME(Mean) fell above the upper bound of CIs in three (of 24) SU conditions and eight (of 24) EU conditions, and ME(SD) fell above the CI upper limits in one SU condition and eight EU conditions. The margins by which ME(Mean) estimates exceeded the upper bounds were quite small, ranging from 0.006 to 0.040 (averaging 0.020) in SU conditions and 0.023 to 0.491 (averaging 0.174) in EU conditions. ME(SD) also exceeded upper CI bounds by small margins, at 0.105 for the single SU case and ranging from 0.050 to 0.330 (averaging 0.149) in SU conditions.

Figure 22. Estimates of ME(Mean) and 95% CIs About the Estimates for $Q1$ in All $\boldsymbol{\xi}$,$\boldsymbol{\theta}$ Study Conditions

a. 1PL



b. 2PL



c. 3PL

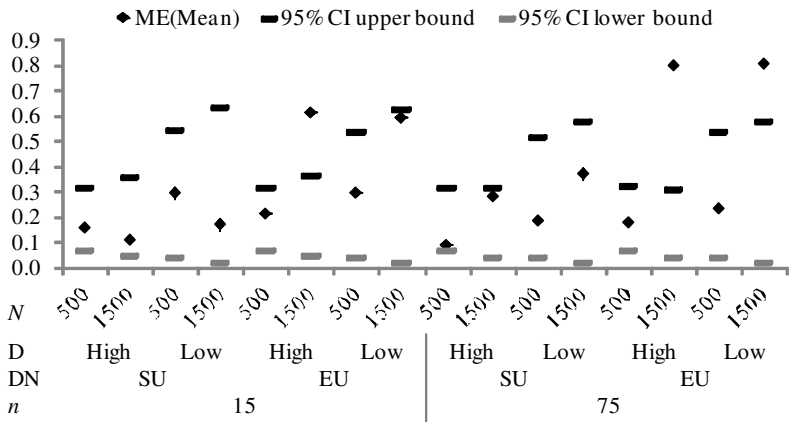
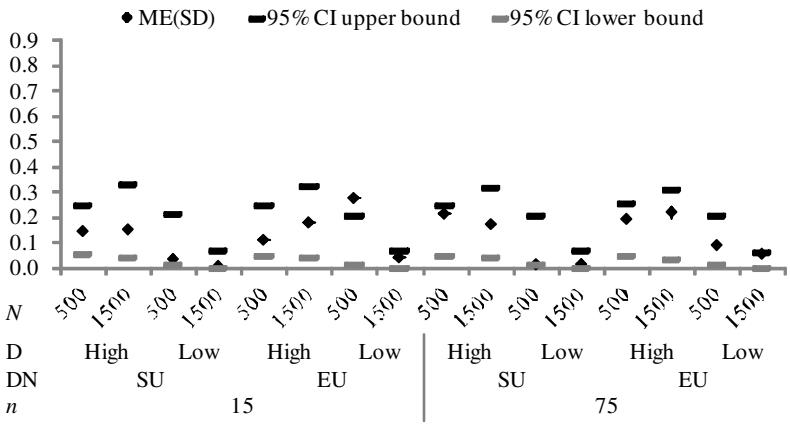
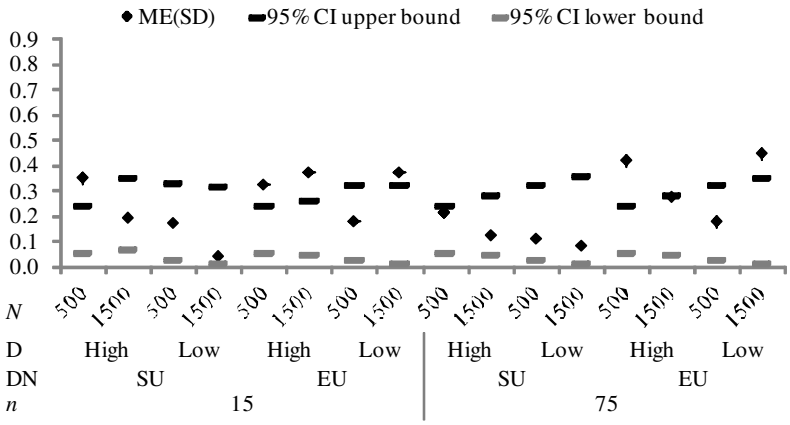


Figure 23. Estimates of ME(SD) and 95% CIs About the Estimates for $Q1$ in All $\boldsymbol{\xi},\boldsymbol{\theta}$ Study Conditions

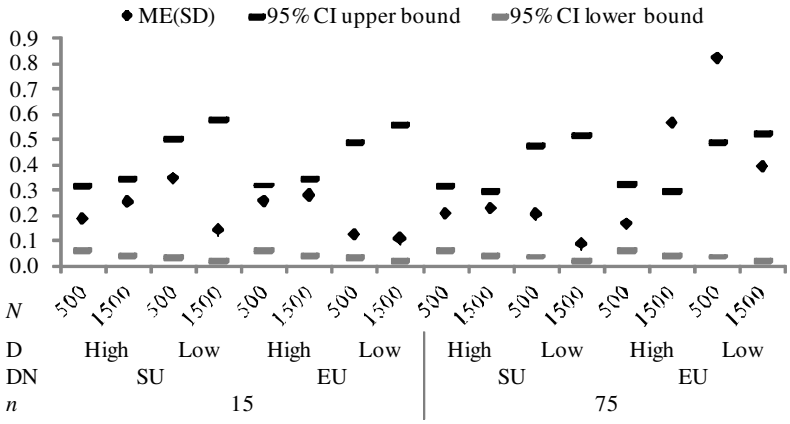
a. 1PL



b. 2PL



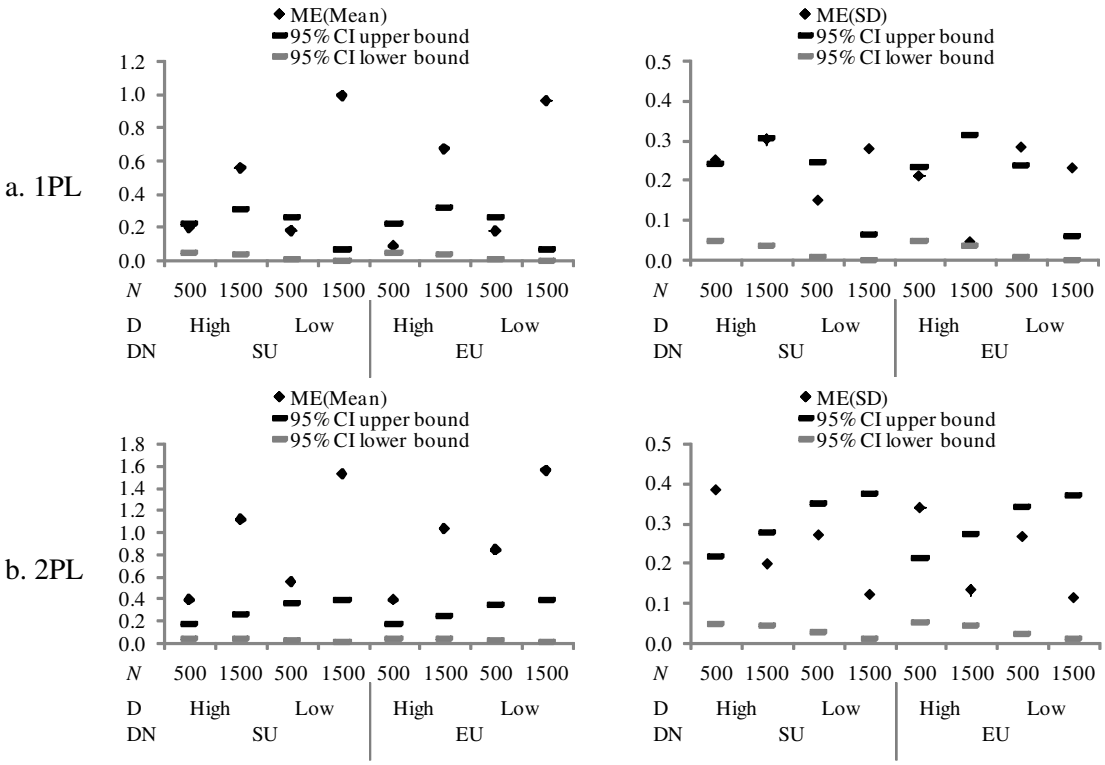
c. 3PL



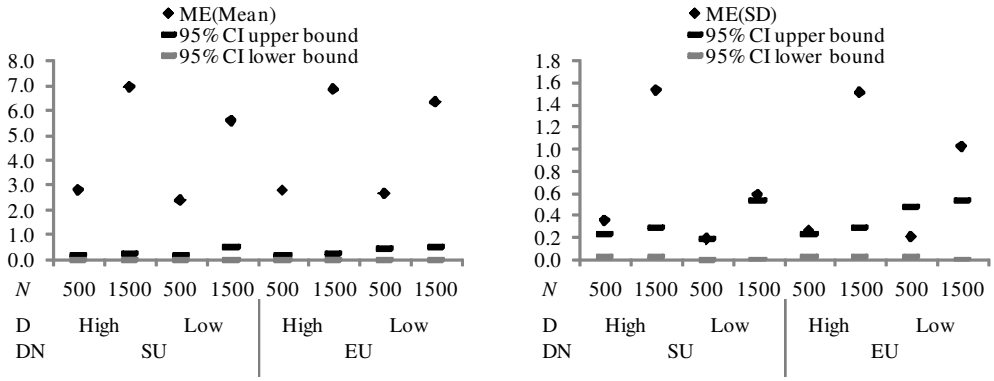
As a point of comparison, ME(Mean) and ME(SD) plots for the $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ conditions where $Q1$ moments appeared to least aberrant will be mentioned. These plots, for $n = 75$ conditions, are shown in Figure 24 where it is evident that, on average, means were near

expectation only for the 1PL when $N = 500$. SDs generally showed a lesser degree of disturbance. Though they exceeded the CI upper bounds in four (of the eight) 1PL conditions, two 2PL conditions and seven 3PL conditions, the margins by which they exceeded the CI upper limits were fairly small for the 1PL and 2PL, ranging from 0.011 to 0.216. For the 3PL, margins were quite small for some conditions but exceeded 1.2 in high discrimination conditions when $N = 1{,}500$.

Figure 24. Estimates of ME(Mean) and ME(SD) and 95% CIs About the Estimates for $Q1$ in All $n = 75$ $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Study Conditions



c. 3PL



Relationship between Q1 and item parameters. Unlike *Q*O, *Q*1 was correlated with item parameters, though these correlations were much more evident when $n = 15$ than when $n = 75$. Table 17 contains Pearson correlations (r), within each K group, between item b parameters and *Q*1 for each SU study condition. For both the 1PL and 2PL, when $n = 15$ *Q*1 was positively correlated with b across K, ranging (across K where $N_K > 1{,}000$) from 0.10 to 0.35 for the 1PL and 0.17 to 0.42 for the 2PL. These correlations tended to be stronger when $N = 1{,}500$ than when $N = 500$. For the 3PL in the shorter test length condition, correlations were positive for smaller K (in which N_K were also smaller) but substantially negative for $K = 10$ (which contained the majority of items); when $N_K = 10$, correlations ranged from −0.26 to −0.52. At the longer test length, for all models the patterns mentioned above were not as clear and correlations tended to approach zero across K (as N_K increased), except for the 3PL when $N = 1{,}500$ where correlations were mildly to moderately negative at the largest K.

Table 17. Correlations Between $Q1$ and b Within K

Model	K	n = 15				n = 75			
		N = 500		N = 1,500		N = 500		N = 1,500	
		N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination									
1PL	5	53	0.28			69	−0.04		
	6	168	0.18			164	−0.06		
	7	399	0.30	44	0.45	416	0.06	69	−0.25
	8	1110	0.31	161	0.36	1549	0.05	298	0.07
	9	5431	0.22	1564	0.35	7213	0.04	2525	0.15
	10	11560	0.14	16959	0.32	9308	0.01	15831	0.07
2PL	3	33	0.27						
	4	54	0.30			71	0.11		
	5	149	0.41			116	0.31		
	6	257	0.22	68	0.40	281	0.15	53	0.13
	7	494	0.38	180	0.38	704	0.10	219	0.28
	8	1596	0.36	533	0.50	2067	0.08	667	0.20
	9	5247	0.39	2356	0.42	6550	0.04	3192	0.17
	10	10913	0.20	15569	0.34	8929	0.02	14565	0.07
3PL	6	75	0.56			87	0.38		
	7	203	0.46	34	0.62	225	0.21	69	0.11
	8	638	0.12	186	0.19	870	−0.01	281	0.14
	9	2707	0.01	1034	−0.07	3504	−0.06	1542	−0.32
	10	15093	−0.26	17479	−0.53	14028	−0.03	16837	−0.17
Low Discrimination									
1PL	9	735	0.08	258	0.31	323	0.10		
	10	17982	0.10	18492	0.20	18420	0.03	18744	0.05
2PL	8	42	0.42			67	0.16		
	9	504	0.31	59	0.30	965	0.12	118	0.26
	10	18184	0.17	18685	0.26	17696	0.02	18625	0.08
3PL	9	264	0.55			484	0.17	52	0.29
	10	18469	−0.36	18727	−0.52	18230	−0.07	18697	−0.23

Correlations between a parameters and $Q1$ can be found in Table 18 for each SU study condition. For both models, when n = 15 $Q1$ was positively correlated with a across K. For the 2PL these correlations were somewhat stronger when N = 1,500 than when N = 500. Correlations were much stronger in the low discrimination than the high discrimination conditions, reaching as high as 0.81 for the 2PL when n = 15 and N = 500.

When n = 75, these patterns were not as clear and correlations tended to approach zero across K (as N_K increased) except in the low discrimination conditions where it remained mildly positive at the highest K in most instances.

Table 18. Correlations Between $Q1$ and a Within K

Model	K	n = 15				n = 75			
		N = 500		N = 1,500		N = 500		N = 1,500	
		N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination									
2PL	3	33	0.45						
	4	54	0.56			71	0.08		
	5	149	0.57			116	0.16		
	6	257	0.40	68	0.61	281	0.26	53	0.27
	7	494	0.40	180	0.53	704	0.07	219	0.15
	8	1596	0.36	533	0.52	2067	0.11	667	0.18
	9	5247	0.29	2356	0.33	6550	0.04	3192	0.12
	10	10913	0.15	15569	0.30	8929	0.00	14565	0.06
3PL	6	75	0.60			87	0.46		
	7	203	0.51	34	0.61	225	0.29	69	0.27
	8	638	0.19	186	0.15	870	0.06	281	0.20
	9	2707	0.13	1034	–0.03	3504	0.00	1542	–0.16
	10	15093	0.33	17479	0.31	14028	0.02	16837	0.02
Low Discrimination									
2PL	8	42	0.81			67	0.33		
	9	504	0.65	59	0.54	965	0.24	118	0.27
	10	18184	0.63	18685	0.76	17696	0.08	18625	0.18
3PL	9	264	0.60			484	0.25	52	0.34
	10	18469	0.61	18727	0.63	18230	0.11	18697	0.17

Finally, correlations between c parameters and $Q1$ can be found in Table 19 for each SU study condition. Correlations were generally mildly to moderately negative across most K and conditions, and tended to be somewhat more extreme in the low discrimination conditions when n = 15.

Table 19. Correlations Between $Q1$ and c Within K

K	$n = 15$				$n = 75$			
	$N = 500$		$N = 1{,}500$		$N = 500$		$N = 1{,}500$	
	N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination								
6	75	−0.42			87	−0.37		
7	203	−0.41	34	−0.48	225	−0.34	69	0.11
8	638	−0.27	186	−0.25	870	−0.34	281	0.14
9	2707	−0.32	1034	−0.19	3504	−0.19	1542	−0.32
10	15093	−0.18	17479	−0.13	14028	−0.06	16837	−0.17
Low Discrimination								
9	264	−0.49			484	−0.31	52	−0.32
10	18469	−0.32	18727	−0.27	18230	−0.10	18697	−0.17

When θ and ξ estimation error were absent from $Q1$ it was uncorrelated with item parameters. Correlations between item parameters and $Q1$ in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions can be found in Appendix F (Tables F-21 to F-23). Across all K (where $N_K > 1{,}000$) and study conditions, correlations with b ranged from −0.03 to 0.05, correlations with a ranged from −0.03 to 0.02, and correlations with c ranged from −0.03 to 0.02.

Q1 summary. In the absence of item parameter estimation error and data noise, $Q1$ appeared to generally approximate its theoretical distributions. If $Q1$ strictly followed its theoretical distributions, ME(Mean) and ME(SD) should be expected to fall outside the limits of expectation in 0 to 1 (of the 24) SU conditions. In SU conditions ME(Mean) exceeded the limits of expectation in 3 (of 24) study conditions (Figure 22) and ME(SD) exceeded the limits of expectation in 1 (of 24) study conditions (Figure 23). Though there appeared to be few aberrations in the first and second moments, the proportion (over all study conditions) of KS test rejections was higher than expectation, at 0.30, and, unlike for QO, decreased as model complexity increased, from 0.40 for the 1PL to 0.15 for the 3PL. Overall, $Q1$ π_K was higher than QO π_K, which was near expectation; this result could be due, at least partially, to greater KS test power for $Q1$ since it had fewer K than QO resulting in larger sample sizes for many KS tests.

The introduction of data noise in the absence of parameter estimation error led to a small but noticeable disruption in $Q1$'s sampling distribution. ME(Mean) (Figure 22)

and ME(SD) (Figure 23) were more likely to be aberrant in EU than in SU conditions, each exceeding the CI upper limits in 8 (of 24) conditions. However, this main effect for DN was not present in KS test results, where π_K in EU conditions was the same as that in SU conditions.

When $Q1$ was computed with estimated model parameters, its sampling distribution was disrupted and there was little noticeable impact of data noise. The presence of θ estimation error caused inflations in both $Q1$'s mean and variance and, as with QO, the presence of ξ estimation error decreased sampling distribution variances. When n = 15, both $Q1$'s mean and SD were grossly inflated for all models. Across SU conditions, Bias(Mean) ranged from 4.86 to 55.14 and Bias(SD) ranged from 1.40 to 21.94 (Tables F-9 and F-10); Bias for both moments increased with model complexity and N. When n = 75, there tended to be less positive bias in means and SDs, where, across SU conditions, Bias(Mean) ranged from −0.14 to 6.95 and Bias(SD) ranged from −0.27 to 1.53 (Tables F-11 and F-12). Like n = 15 conditions, Bias for both moments tended to increase with model complexity and N. Means were closest to expectation for the 1PL and 2PL when n = 75 and N = 500; however, in this condition, where means appeared to be most on target, SDs were fairly consistently somewhat below expectation (see Figures F-13 and F-14).

Finally, there were relationships between $Q1$ sampling distribution moments and K, and between $Q1$ and item parameter values. For the 3PL, $Q1$ means and SDs generally moved closer to expectation as K increased (Figures F-1 – F-16), though one or both moments were still above expectation for large K. Correlations between $Q1$ and item parameters were much more evident when n = 15 than when n = 75, where correlations tended to approach zero as sample sizes increased. $Q1$ tended to be: positively correlated with b for the 1PL and 2PL, and negatively correlated with b for the 3PL (Table 17); positively correlated with a for the 2PL and 3PL (Table 18); and negatively correlated with c for the 3PL (Table 19). Correlations between item parameters and $Q1$ were near zero when there was no parameter estimation error in the statistics (Tables F-21 – F-23).

LM Tests

In a handful of cases LM statistics were negative, which is in clear violation of distributional assumptions. Negative values occurred only for 1PL and 2PL LM($\alpha\beta$) in the n = 15 high discrimination conditions. Across all study conditions, only 26 of the LM($\alpha\beta$) were negative. These were excluded from all analyses.

Tables 20–25 present descriptive statistics for the LM statistics across SU study conditions. The first row of data in these tables contains the expected values for the descriptive statistics if LM statistics followed their theoretical distributions. The last two columns in the tables contain a summary of the KS tests that were conducted for each replicated test. The average of the KS D values ($\overline{D}$) is reported, as well as the proportion of replicated tests (π_R) in which D was statistically significant (at the 0.05 α level). Histograms of observed LM statistics overlaid with the theoretical distributions can be found in Appendix G (Figures G-1 to G-9).

The most striking feature of the LM distributions was the inordinate percentage of statistics with extremely high values. Additionally, for the 2PL and 3PL lower values were somewhat overrepresented for LM(α) and LM(β); this is evident from the medians in Tables 22–25 which tended to be lower than expectation for the 2PL and 3PL (also see Figures G-5, G-6, G-8, and G-9). Distributional aberrances were much more marked for LM($\alpha\beta$) (where π_R averaged 0.76 across SU conditions) than for LM(α) or LM(β), the π_R for which averaged 0.43 and 0.37 respectively. All other study factors (other than DN) had an impact on LM distributions (see Appendix G, Figures G-10 to G-12). KS test π_R tended to be larger in high discrimination conditions, when N was small, and n was large. The effect for n was likely attributable, at least partially, to differences in KS test power between the two levels of n. T1 rates varied little between n = 15 and n = 75 conditions but did tend to be slightly lower in the former than the latter conditions, indicating very minor effects of n on LM sampling distributions; for example, across all SU conditions T1 rates (at α = 0.05) when n = 15 and n = 75 were, respectively, 0.404 and 0.434 for LM($\alpha\beta$), 0.119 and 0.124 for LM(α), and 0.066 and 0.072 for LM(β).

IRT model also impacted the KS test results, but the nature of the effect differed between the three LM statistics; π_R tended to (1) decrease with model complexity for LM($\alpha\beta$), (2) increase with model complexity for LM(β), and (3) be lower for the 2PL than 1PL and 3PL for LM(α). Finally, in some low discrimination conditions LM(α) and LM(β) did tend to approach their theoretical distributions. However, the observed distributions were only very close to the theoretical target for LM(β) with the 1PL in the low discrimination conditions when N = 1,500.

Table 20. Descriptive Statistics for LM($\alpha\beta$) in SU High Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test	
									$\overline{D}$	π_R
Expected		8.00	7.33	4.00	0.26[a]	37.33[b]	1.00	1.50	--	0.05
1PL										
500	15	7.78E+12	66.09	5.89E+14	0.64	6.65E+16	93.47	9.54E+03	0.81	1.00
	75	1.83E+04	81.57	2.67E+05	0.36	2.50E+07	61.01	4.94E+03	0.80	1.00
1,500	15	4.69E+03	53.56	8.43E+04	0.74	6.23E+06	46.31	2.71E+03	0.77	1.00
	75	3.62E+04	50.27	3.41E+06	0.96	4.67E+08	136.46	1.87E+04	0.74	1.00
2PL										
500	15	4.98E+10	15.14	5.22E+12	0.61	6.60E+14	115.57	1.40E+04	0.53	0.93
	75	3.11E+05	18.08	3.33E+07	0.30	4.55E+09	135.79	1.85E+04	0.53	1.00
1,500	15	5.54E+10	10.04	7.58E+12	0.45	1.04E+15	136.89	1.87E+04	0.36	0.52
	75	4.56E+04	11.65	2.51E+06	0.36	3.08E+08	104.39	1.23E+04	0.37	1.00
3PL										
500	15	264.26	9.62	9.89E+03	0.39	1.29E+06	119.35	1.54E+04	0.33	0.45
	75	1.96E+03	10.17	4.04E+04	0.35	3.02E+06	53.70	3.59E+03	0.30	1.00
1,500	15	115.06	7.83	3.63E+03	0.36	3.05E+05	67.30	5.24E+03	0.26	0.20
	75	1.57E+03	8.25	4.24E+04	0.33	4.33E+06	68.20	6.16E+03	0.17	0.63

[a]0.00001[th] quantile point of a $\chi^2(8)$ distribution; [b](1 – 0.00001)[th] quantile point of a $\chi^2(8)$ distribution

Table 21. Descriptive Statistics for LM($\alpha\beta$) in SU Low Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		8.00	7.33	4.00	0.26[a]	37.33[b]	1.00	1.50	--	0.05
1PL										
500	15	83.14	25.50	627.57	0.56	5.87E+04	55.82	4.43E+03	0.67	0.99
	75	134.58	28.99	916.58	0.72	4.32E+04	26.71	907.73	0.66	1.00
1,500	15	51.34	23.41	136.69	0.53	1.34E+04	53.50	4.89E+03	0.65	0.99
	75	88.67	26.01	2.33E+03	0.42	3.12E+05	129.44	1.73E+04	0.64	1.00
2PL										
500	15	53.07	12.22	876.00	0.48	8.52E+04	63.50	5.27E+03	0.45	0.78
	75	167.79	12.37	3.90E+03	0.61	3.54E+05	58.45	4.34E+03	0.40	1.00
1,500	15	22.86	9.44	580.08	0.54	5.58E+04	77.93	6.64E+03	0.33	0.40
	75	29.04	9.56	619.37	0.49	4.63E+04	59.45	3.92E+03	0.25	0.94
3PL										
500	15	15.74	9.03	129.36	0.46	9.52E+03	44.29	2.42E+03	0.31	0.34
	75	95.63	8.89	7.06E+03	0.48	9.16E+05	120.84	1.53E+04	0.20	0.82
1,500	15	9.20	7.62	53.29	0.44	6.86E+03	115.94	1.46E+04	0.27	0.21
	75	14.72	7.50	522.42	0.33	6.53E+04	112.93	1.35E+04	0.11	0.16

[a]0.00001^{th} quantile point of a $\chi^2(8)$ distribution; [b]$(1 - 0.00001)^{th}$ quantile point of a $\chi^2(8)$ distribution

Table 22. Descriptive Statistics for LM(α) in SU High Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01[a]	28.47[b]	1.41	3.00	--	0.05
1PL										
500	15	245.61	5.43	5.84E+03	0.02	5.33E+05	73.85	6.28E+03	0.37	0.57
	75	387.73	5.46	5.53E+03	0.00	4.78E+05	51.49	3.68E+03	0.32	1.00
1,500	15	135.90	5.08	3.50E+03	0.01	3.06E+05	55.54	3.92E+03	0.34	0.48
	75	691.11	4.81	6.36E+04	0.03	8.69E+06	136.12	1.86E+04	0.26	0.98
2PL										
500	15	717.34	3.43	1.84E+04	0.02	1.24E+06	45.29	2.35E+03	0.26	0.16
	75	7.21E+03	3.60	7.22E+05	0.00	9.83E+07	134.57	1.83E+04	0.18	0.70
1,500	15	472.15	2.99	2.57E+04	0.03	2.68E+06	86.04	8.02E+03	0.24	0.13
	75	1.13E+03	3.03	4.32E+04	0.03	3.05E+06	52.94	3.13E+03	0.12	0.20
3PL										
500	15	25.55	2.83	450.82	0.01	4.10E+04	53.05	4.02E+03	0.26	0.20
	75	67.10	2.80	1.20E+03	0.02	6.91E+04	36.26	1.56E+03	0.15	0.41
1,500	15	18.85	2.52	584.86	0.01	4.85E+04	59.15	4.06E+03	0.29	0.30
	75	120.34	2.50	3.45E+03	0.01	2.51E+05	49.74	2.90E+03	0.19	0.74

[a]0.00001^{th} quantile point of a $\chi^2(4)$ distribution; [b]$(1 - 0.00001)^{th}$ quantile point of a $\chi^2(4)$ distribution

Table 23. Descriptive Statistics for LM(α) in SU Low Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01[a]	28.47[b]	1.41	3.00	--	0.05
1PL										
500	15	5.95	3.96	11.63	0.03	693.12	29.27	1.40E+03	0.26	0.19
	75	6.67	4.14	28.17	0.01	2.02E+03	47.71	2.75E+03	0.18	0.68
1,500	15	5.42	3.92	5.50	0.02	124.40	4.02	36.56	0.26	0.19
	75	5.64	4.05	6.04	0.03	127.14	5.28	59.24	0.17	0.60
2PL										
500	15	5.95	3.11	83.75	0.01	9.62E+03	89.79	9.59E+03	0.25	0.14
	75	8.79	3.07	155.63	0.02	1.03E+04	52.28	3.00E+03	0.12	0.18
1,500	15	3.85	2.82	19.63	0.01	2.57E+03	119.70	1.56E+04	0.27	0.22
	75	4.15	2.86	29.88	0.01	2.51E+03	63.98	4.54E+03	0.14	0.33
3PL										
500	15	4.08	2.76	30.24	0.01	2.57E+03	63.62	4.63E+03	0.27	0.25
	75	8.18	2.68	541.91	0.02	7.41E+04	136.21	1.86E+04	0.17	0.58
1,500	15	3.17	2.56	2.81	0.01	184.39	16.17	945.73	0.30	0.33
	75	3.41	2.44	39.06	0.01	5.24E+03	129.41	1.73E+04	0.21	0.82

[a]0.00001[th] quantile point of a $\chi^2(4)$ distribution; [b](1 – 0.00001)[th] quantile point of a $\chi^2(4)$ distribution

Table 24. Descriptive Statistics for LM(β) in SU High Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01[a]	28.47[b]	1.41	3.00	--	0.05
1PL										
500	15	14.20	3.95	229.16	0.02	2.67E+04	94.44	1.02E+04	0.26	0.18
	75	23.91	3.93	318.00	0.02	2.03E+04	41.94	2.11E+03	0.17	0.59
1,500	15	10.35	3.77	321.28	0.04	4.29E+04	127.70	1.69E+04	0.23	0.11
	75	35.02	3.54	3.04E+03	0.01	4.15E+05	135.99	1.86E+04	0.11	0.13
2PL										
500	15	102.43	2.94	7.55E+03	0.01	1.00E+06	125.42	1.65E+04	0.26	0.15
	75	184.65	3.04	6.83E+03	0.01	7.87E+05	89.70	9.66E+03	0.16	0.54
1,500	15	88.87	2.65	9.14E+03	0.01	1.25E+06	135.59	1.85E+04	0.26	0.21
	75	289.67	2.78	1.29E+04	0.01	1.21E+06	67.72	5.23E+03	0.15	0.44
3PL										
500	15	13.40	2.65	279.46	0.01	2.65E+04	61.67	4.92E+03	0.27	0.24
	75	87.52	2.64	2.05E+03	0.02	1.89E+05	58.53	4.50E+03	0.17	0.59
1,500	15	12.69	2.39	431.08	0.02	4.21E+04	67.71	5.57E+03	0.30	0.36
	75	87.74	2.30	3.20E+03	0.01	3.29E+05	73.21	6.60E+03	0.23	0.90

[a]0.00001[th] quantile point of a $\chi^2(4)$ distribution; [b](1 – 0.00001)[th] quantile point of a $\chi^2(4)$ distribution

Table 25. Descriptive Statistics for LM(β) in SU Low Discrimination Conditions

N	*n*	Mean	*Mdn*	*SD*	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01[a]	28.47[b]	1.41	3.00	--	0.05
1PL										
500	15	4.27	3.41	4.41	0.02	113.82	9.16	152.05	0.23	0.09
	75	4.27	3.37	4.58	0.01	183.71	11.18	253.60	0.10	0.04
1,500	15	4.07	3.40	2.93	0.03	32.80	1.56	4.28	0.23	0.09
	75	4.07	3.41	2.92	0.01	30.85	1.55	4.17	0.10	0.09
2PL										
500	15	3.77	2.75	6.07	0.01	239.08	14.34	317.15	0.27	0.23
	75	4.03	2.65	10.62	0.01	1.03E+03	51.99	4.63E+03	0.16	0.58
1,500	15	3.32	2.56	5.12	0.01	234.35	29.62	1.17E+03	0.28	0.27
	75	3.43	2.55	12.78	0.01	1.35E+03	82.26	7.88E+03	0.18	0.74
3PL										
500	15	3.29	2.59	4.56	0.01	492.66	67.24	7.08E+03	0.29	0.30
	75	4.52	2.51	124.88	0.02	1.60E+04	117.37	1.46E+04	0.19	0.73
1,500	15	2.99	2.41	2.34	0.01	22.25	1.58	3.63	0.32	0.41
	75	2.91	2.29	2.52	0.02	112.56	6.92	219.39	0.23	0.93

[a]0.00001[th] quantile point of a $\chi^2(4)$ distribution; [b](1 – 0.00001)[th] quantile point of a $\chi^2(4)$ distribution

Further inspection revealed that some of the most inflated LM($\alpha\beta$) values corresponded with cases in which the reference category contained only one number-correct (NC) score. Like the cases where LM($\alpha\beta$) was negative, this occurred only for the 1PL and 2PL in the n = 15 high discrimination conditions. Across all study conditions, there were only 25 instances in which the reference category contained only one NC score. Removal of these cases dramatically reduced the means for LM($\alpha\beta$), though the means remained grossly inflated. For example, means in the SU condition when N = 500 decreased from 7.78E+12 to 6668.10 for the 1PL and from 4.98E+10 to 15961.06 for the 2PL. After re-inspection of the LM($\alpha\beta$) with negative values, it was found that all such instances also occurred when the reference category contained only one NC score. This foreshadows a problem with the LM statistics, namely that lack of statistical information in the reference group seriously distorts LM statistics. This will become more evident below.

Relationship between LM statistics and item parameters. There was a clear relationship between LM statistics and item *b* and *a* parameters. For all models, LM statistics were more inflated for items in the lower half of the *b* range. For the 1PL and 2PL, there also appeared to be pockets of items in the higher *b*-value range that had inflated LM statistics, though the degree of inflation was less marked than for items in the lower *b* range. LM statistics were also more inflated at higher *a* for the 2PL and 3PL. Figures 25–26 and Figures 27–28 illustrate the effect of *b* and *a*, respectively, on LM statistics for the 2PL and 3PL in the $N = 500/n = 15$ SU conditions; figures for all SU conditions can be found in Appendix G (Figures G-13 to G-32). These figures contain scatterplots between item parameters and log-transformed (to visually reduce the influence of extreme LM values) LM statistics. Plots were fairly similar for the other conditions, though the tails tended to be somewhat sparser in the large sample size conditions.

Figure 25. Scatterplots Between *b* and LM Statistics for the 2PL When $N = 500$ and $n = 15$

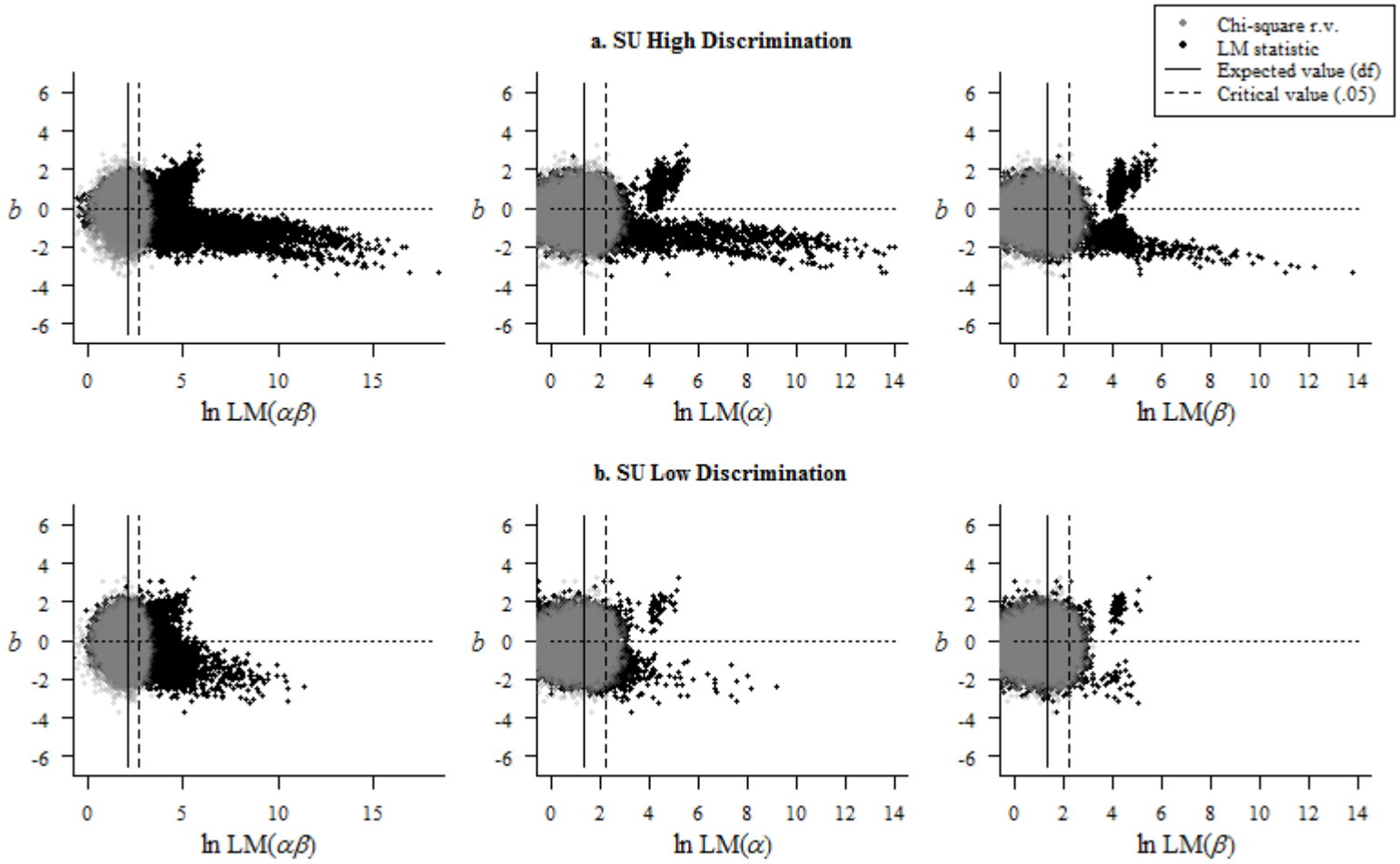


Figure 26. Scatterplots Between *b* and LM Statistics for the 3PL When $N = 500$ and $n = 15$

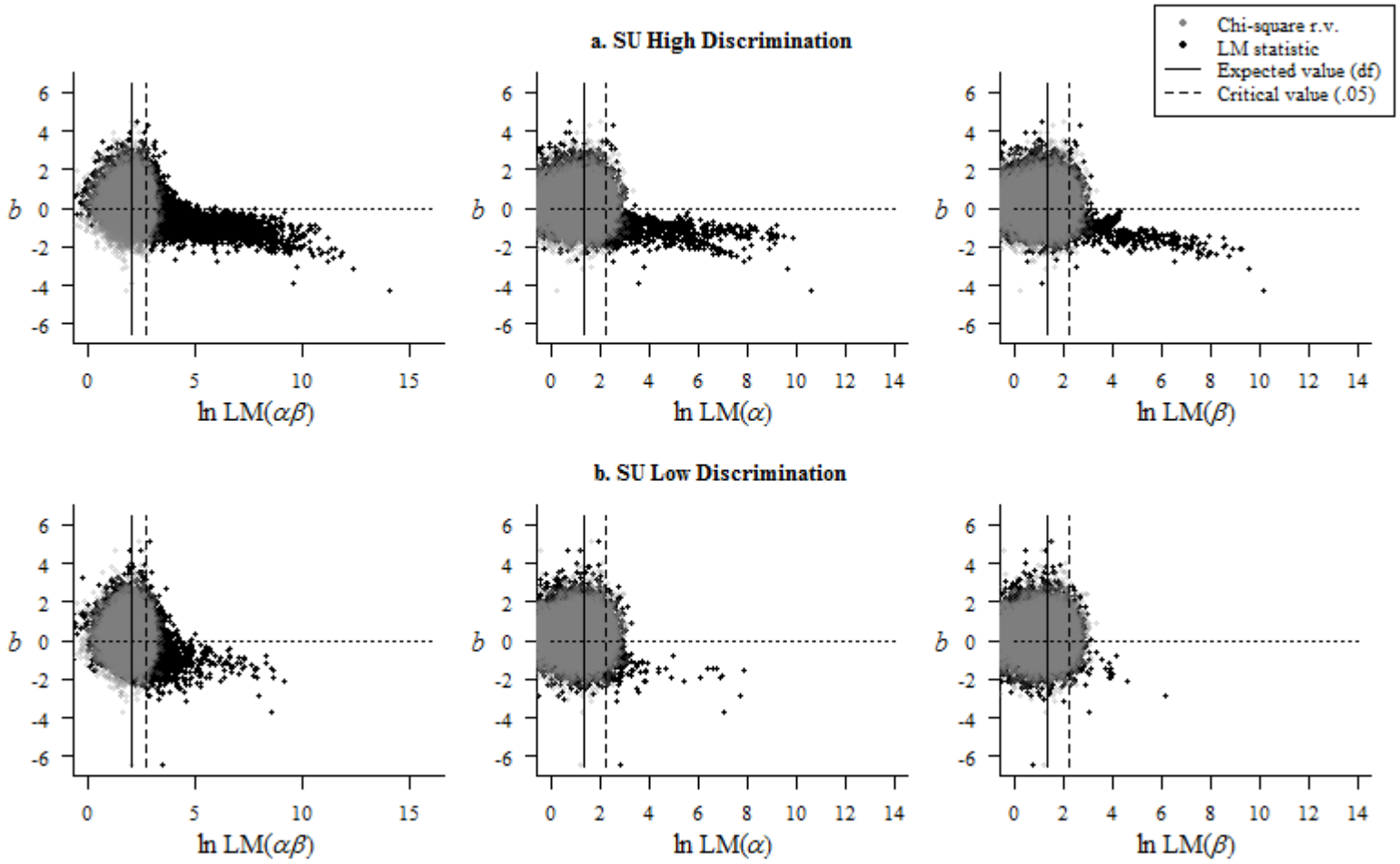


Figure 27. Scatterplots Between *a* and LM Statistics for the 2PL When $N = 500$ and $n = 15$

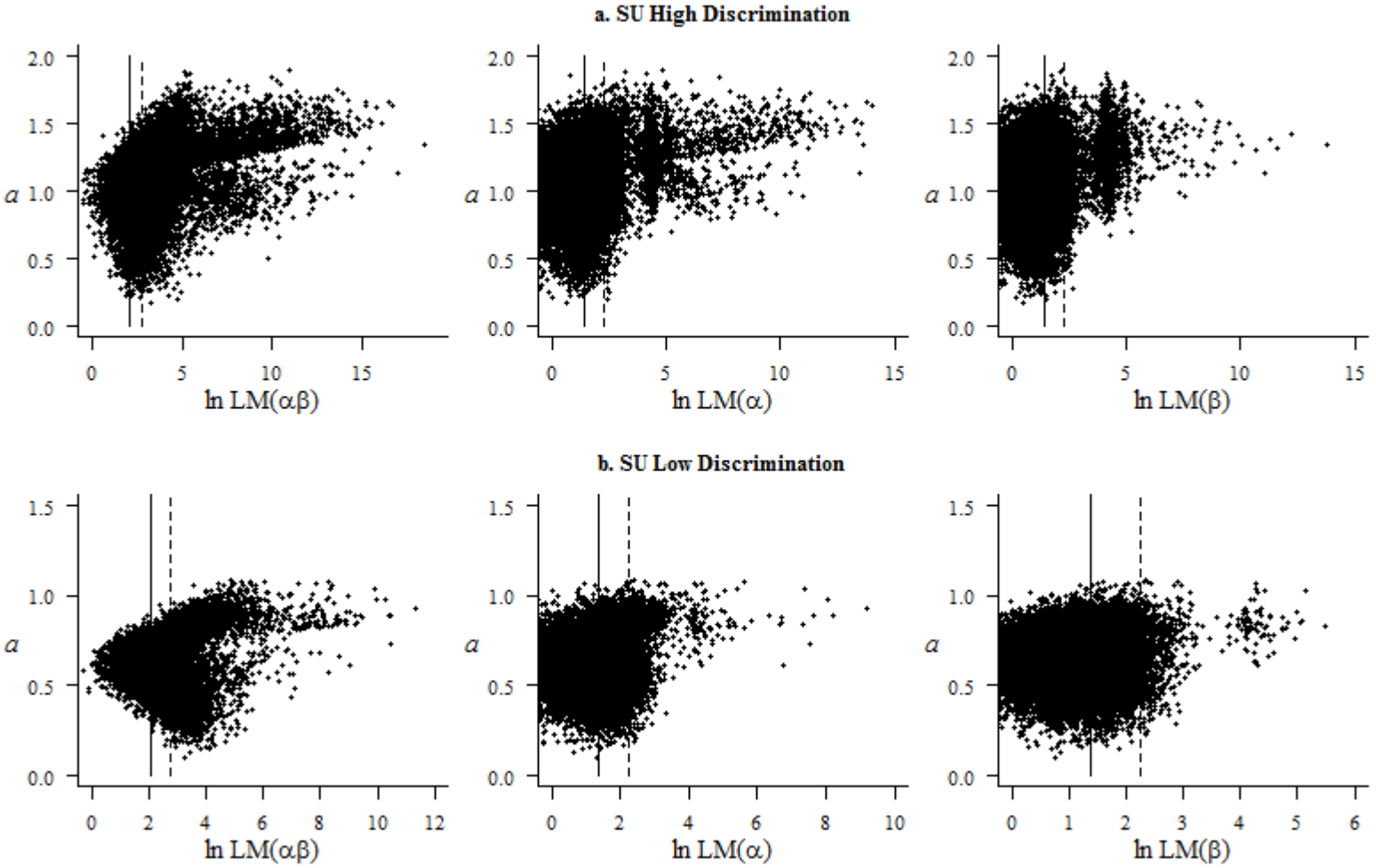
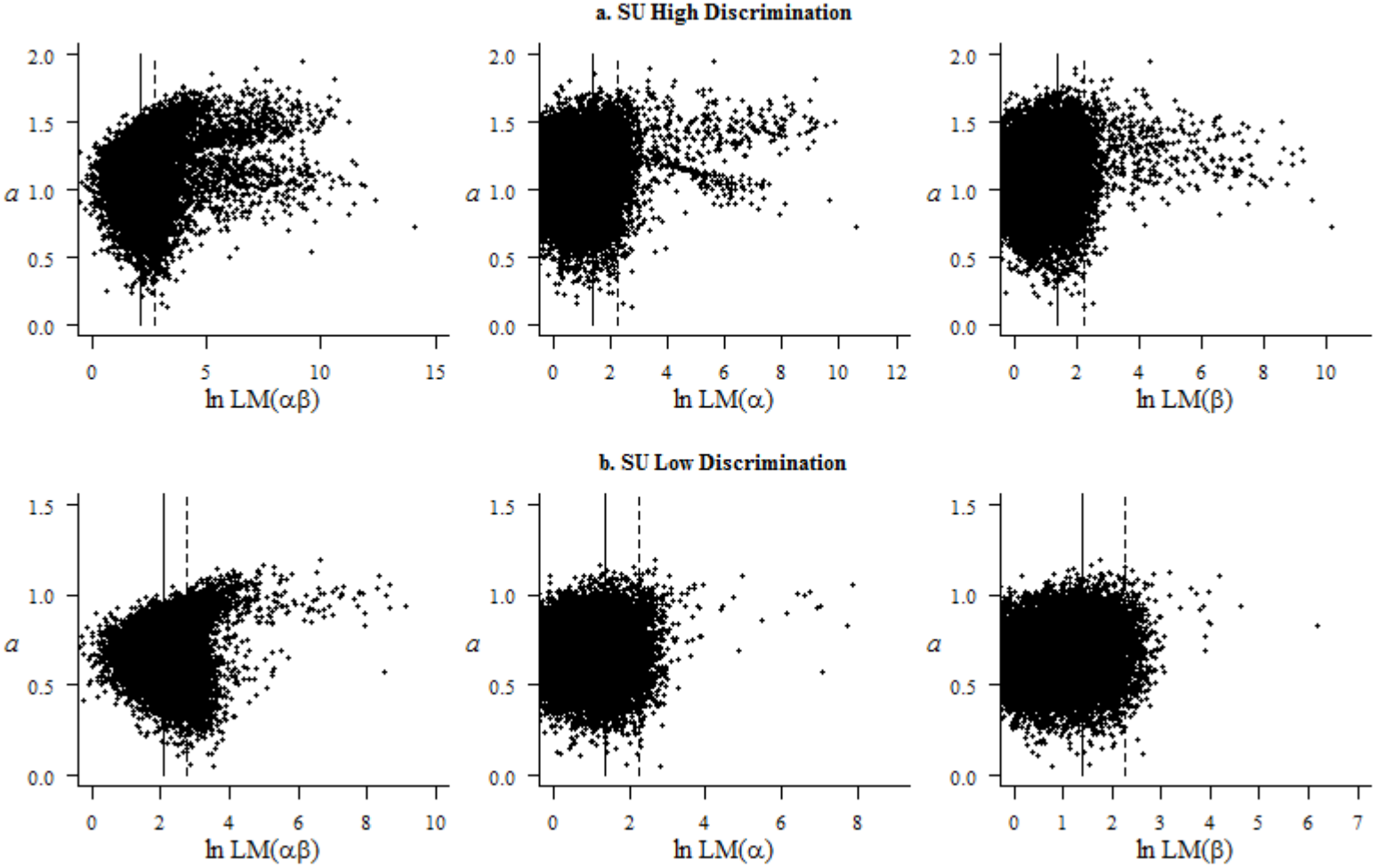


Figure 28. Scatterplots Between *a* and LM Statistics for the 3PL When $N = 500$ and $n = 15$



Because the LM statistics were computed using the highest category as the reference, when items are located in the lower *b* range there will be less statistical information in the reference category, especially when items are more highly discriminating. This lack of information appears to cause LM statistics to become very aberrant. The aberration results in $\mathbf{W}^{-1}$ being near singular in those instances.

In an attempt to fix this problem, LM statistics were re-computed using the lowest category as the reference group when $b \leq 0$ and the highest category as the reference group when $b > 0$. Tables 26–31 present comparisons between these "corrected" LM statistics and those that were originally computed. Full tables of descriptive statistics for corrected LM tests can be found in Appendix G (Tables G-1 to G-8). The correction clearly resulted in less aberrant LM values, though generally did not cause the observed distributions to be a better match to theoretical distributions, at least according to the KS tests. The average difference (across the 24 SU conditions) between the original LM π_R and the corrected LM π_R was 0.02 for LM($\alpha\beta$), 0.11 for LM(α), and −0.01 for LM(β). The correction appeared mainly to reign in extreme outlying observations.

Table 26. Comparison Between Original LM($\alpha\beta$) and Corrected LM($\alpha\beta$) in SU High Discrimination Conditions

M	N	n	Original LM($\alpha\beta$)				Corrected LM($\alpha\beta$)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	7.78E+12	5.89E+14	0.81	1.00	7.73E+11	7.68E+13	0.78	1.00
		75	1.83E+04	2.67E+05	0.80	1.00	70.69	104.17	0.73	1.00
	1,500	15	4.69E+03	8.43E+04	0.77	1.00	86.70	141.10	0.73	1.00
		75	3.62E+04	3.41E+06	0.74	1.00	56.39	86.64	0.62	1.00
2PL	500	15	4.98E+10	5.22E+12	0.53	0.93	1.98E+11	1.78E+13	0.49	0.87
		75	3.11E+05	3.33E+07	0.53	1.00	40.54	70.23	0.45	1.00
	1,500	15	5.54E+10	7.58E+12	0.36	0.52	6.96E+10	7.75E+12	0.33	0.42
		75	4.56E+04	2.51E+06	0.37	1.00	37.52	91.32	0.30	0.99
3PL	500	15	264.26	9.89E+03	0.33	0.45	20.44	31.47	0.36	0.53
		75	1.96E+03	4.04E+04	0.30	1.00	25.58	51.37	0.31	1.00
	1,500	15	115.06	3.63E+03	0.26	0.20	16.46	42.04	0.27	0.21
		75	1.57E+03	4.24E+04	0.17	0.63	20.38	55.33	0.17	0.60

Table 27. Comparison Between Original LM($\alpha\beta$) and Corrected LM($\alpha\beta$) in SU Low Discrimination Conditions

M	N	n	Original LM($\alpha\beta$)				Corrected LM($\alpha\beta$)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	83.14	627.57	0.67	0.99	48.17	73.36	0.65	0.99
		75	134.58	916.58	0.66	1.00	35.75	45.64	0.58	1.00
	1,500	15	51.34	136.69	0.65	0.99	42.77	60.95	0.64	0.99
		75	88.67	2.33E+03	0.64	1.00	31.82	38.03	0.55	1.00
2PL	500	15	53.07	876.00	0.45	0.78	18.00	28.68	0.43	0.74
		75	167.79	3.90E+03	0.40	1.00	15.56	19.20	0.31	1.00
	1,500	15	22.86	580.08	0.33	0.40	11.50	12.45	0.31	0.36
		75	29.04	619.37	0.25	0.94	10.87	15.37	0.19	0.72
3PL	500	15	15.74	129.36	0.31	0.34	12.19	12.30	0.32	0.39
		75	95.63	7.06E+03	0.20	0.82	12.11	14.24	0.21	0.84
	1,500	15	9.20	53.29	0.27	0.21	9.07	8.26	0.27	0.21
		75	14.72	522.42	0.11	0.16	8.97	9.88	0.11	0.15

Table 28. Comparison Between Original LM(α) and Corrected LM(α) in SU High Discrimination Conditions

M	N	n	Original LM(α)				Corrected LM(α)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	245.61	5.84E+03	0.37	0.57	14.18	31.65	0.28	0.25
		75	387.73	5.53E+03	0.32	1.00	17.54	35.60	0.22	0.89
	1,500	15	135.90	3.50E+03	0.34	0.48	9.76	38.63	0.25	0.14
		75	691.11	6.36E+04	0.26	0.98	13.04	50.55	0.14	0.30
2PL	500	15	717.34	1.84E+04	0.26	0.16	14.94	34.54	0.26	0.16
		75	7.21E+03	7.22E+05	0.18	0.70	19.02	39.28	0.18	0.73
	1,500	15	472.15	2.57E+04	0.24	0.13	13.98	56.93	0.24	0.14
		75	1.13E+03	4.32E+04	0.12	0.20	17.97	66.33	0.13	0.23
3PL	500	15	25.55	450.82	0.26	0.20	8.49	20.03	0.25	0.17
		75	67.10	1.20E+03	0.15	0.41	9.70	25.20	0.15	0.36
	1,500	15	18.85	584.86	0.29	0.30	7.18	30.50	0.28	0.26
		75	120.34	3.45E+03	0.19	0.74	8.81	39.32	0.19	0.73

Table 29. Comparison Between Original LM(α) and Corrected LM(α) in SU Low Discrimination Conditions

M	N	n	Original LM(α)				Corrected LM(α)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	5.95	11.63	0.26	0.19	4.68	5.45	0.25	0.15
		75	6.67	28.17	0.18	0.68	4.92	6.63	0.12	0.18
	1,500	15	5.42	5.50	0.26	0.19	4.35	3.61	0.24	0.13
		75	5.64	6.04	0.17	0.60	4.47	3.63	0.12	0.15
2PL	500	15	5.95	83.75	0.25	0.14	4.22	7.59	0.25	0.16
		75	8.79	155.63	0.12	0.18	4.84	11.49	0.13	0.28
	1,500	15	3.85	19.63	0.27	0.22	3.58	6.38	0.27	0.23
		75	4.15	29.88	0.14	0.33	3.76	9.79	0.14	0.38
3PL	500	15	4.08	30.24	0.27	0.25	3.88	6.11	0.27	0.24
		75	8.18	541.91	0.17	0.58	3.98	8.35	0.16	0.52
	1,500	15	3.17	2.81	0.30	0.33	3.27	4.09	0.29	0.29
		75	3.41	39.06	0.21	0.82	3.19	5.42	0.20	0.79

Table 30. Comparison Between Original LM(β) and Corrected LM(β) in SU High Discrimination Conditions

M	N	n	Original LM(β)				Corrected LM(β)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	14.20	229.16	0.26	0.18	12.73	28.95	0.26	0.19
		75	23.91	318.00	0.17	0.59	15.59	32.42	0.18	0.74
	1,500	15	10.35	321.28	0.23	0.11	8.88	34.75	0.24	0.13
		75	35.02	3.04E+03	0.11	0.13	11.65	45.23	0.12	0.17
2PL	500	15	102.43	7.55E+03	0.26	0.15	13.51	32.96	0.26	0.15
		75	184.65	6.83E+03	0.16	0.54	16.82	36.10	0.18	0.71
	1,500	15	88.87	9.14E+03	0.26	0.21	12.44	51.82	0.26	0.20
		75	289.67	1.29E+04	0.15	0.44	15.92	60.52	0.15	0.40
3PL	500	15	13.40	279.46	0.27	0.24	7.80	19.15	0.26	0.19
		75	87.52	2.05E+03	0.17	0.59	8.53	22.35	0.16	0.58
	1,500	15	12.69	431.08	0.30	0.36	6.51	28.16	0.30	0.33
		75	87.74	3.20E+03	0.23	0.90	7.64	33.64	0.23	0.92

Table 31. Comparison Between Original LM(β) and Corrected LM(β) in SU Low Discrimination Conditions

M	N	n	Original LM(β)				Corrected LM(β)			
			Mean	SD	KS Test		Mean	SD	KS Test	
					$\bar{D}$	π_R			$\bar{D}$	π_R
1PL	500	15	4.27	4.41	0.23	0.09	4.32	4.81	0.23	0.10
		75	4.27	4.58	0.10	0.04	4.41	5.86	0.10	0.06
	1,500	15	4.07	2.93	0.23	0.09	4.08	2.95	0.23	0.09
		75	4.07	2.92	0.10	0.09	4.09	3.01	0.10	0.10
2PL	500	15	3.77	6.07	0.27	0.23	3.87	7.17	0.27	0.23
		75	4.03	10.62	0.16	0.58	4.39	10.37	0.16	0.60
	1,500	15	3.32	5.12	0.28	0.27	3.35	5.87	0.28	0.27
		75	3.43	12.78	0.18	0.74	3.48	8.64	0.18	0.70
3PL	500	15	3.29	4.56	0.29	0.30	3.63	5.77	0.28	0.28
		75	4.52	124.88	0.19	0.73	3.72	7.70	0.18	0.68
	1,500	15	2.99	2.34	0.32	0.41	3.08	3.73	0.31	0.35
		75	2.91	2.52	0.23	0.93	2.99	4.76	0.23	0.94

LM summary. A shortcoming of the LM statistics appears to be that lack of statistical information in the reference group can result in grossly inflated statistics. This was apparent by the presence of: (1) grossly inflated values when the reference category

contained one NC score, (2) more inflation for items with $b < 0$, and a reduction of this inflation when statistics were re-computed using the lowest category as the reference group when $b \leq 0$ and the highest category as the reference group when $b > 0$, and (3) more inflation for items with higher a.

In most cases, the LM statistics did not follow their theoretical distributions and it is unclear to what degree reference group information caused this result. Re-computing the LM statistics in the manner mentioned above did reduce the presence of grossly inflated values but did not markedly improve the approximation of the statistics to their theoretical distributions, at least according to the KS test results. Overall, LM($\alpha\beta$) showed more departure from expectation than LM(α) and LM(β). LM distributions were affected to some degree by all study factors other than DN (Figures G-10 to G-12). For all three LM statistics, distributional aberrance tended to be greater when discrimination was high, n was large, or N was small. The effect of IRT model was different between the three LM statistics; LM($\alpha\beta$) distributions tended to be most aberrant for the 1PL and least aberrant for the 3PL, with the reverse being true for LM(β); LM(α) distributions tended to be somewhat more aberrant for the 1PL and 3PL than for the 2PL.

LM statistics only began to approximate their theoretical distributions for LM(α) and LM(β) in low discrimination conditions (Tables G-4 and G-6). In these conditions, means were near the expectation of 4.0, ranging from 3.19 to 4.92 for LM(α) and from 2.99 to 4.41 for LM(β). SDs still tended to be above the expectation of 2.83, but more so for the 2PL and 3PL than the 1PL; 1PL ranges for LM(α) and LM(β) were 3.61 to 6.63 and 2.95 to 5.86 respectively, whereas respective ranges for the 2PL/3PL were 4.09 to 11.49 and 3.73 to 10.37. Finally, there tended to be an over-representation of low values for the 2PL and 3PL, where medians were somewhat below the expectation of 3.33, ranging from 2.49 to 2.96 for LM(α) and 2.29 to 2.73 for LM(β); this deflation in medians was not evident for the 1PL where LM(α) and LM(β) ranges were 3.42 to 3.60 and 3.39 to 3.42 respectively.

z Statistics

Correlations between L_z, *V*I, and *V*O (which can be found in Table 32) tended to be higher in the low discrimination than high discrimination conditions. The 1PL tended to exhibit the highest correlations and the 2PL the lowest. Furthermore, correlations between the *z* statistics tended to be lower when $n = 75$ than when $n = 15$, though this was much more evident for $r(L_z,VO)$ and $r(VI,VO)$ than for $r(L_z,VI)$.

L_z and *V*I were highly negatively correlated with one another across all study conditions, with correlations ranging from −0.86 to −0.99 in the high discrimination conditions and from −0.96 to −1.00 in the low discrimination conditions. Correlations between L_z and *V*O were also negative, but weaker, ranging from −0.27 to −0.91 in the high discrimination conditions and from −0.77 to −0.98 in the low discrimination conditions. Finally, *V*I and *V*O were least similar to one another, with correlations ranging from near zero ($r = -0.03$) to 0.83 in high discrimination conditions and from 0.60 to 0.97 in low discrimination conditions.

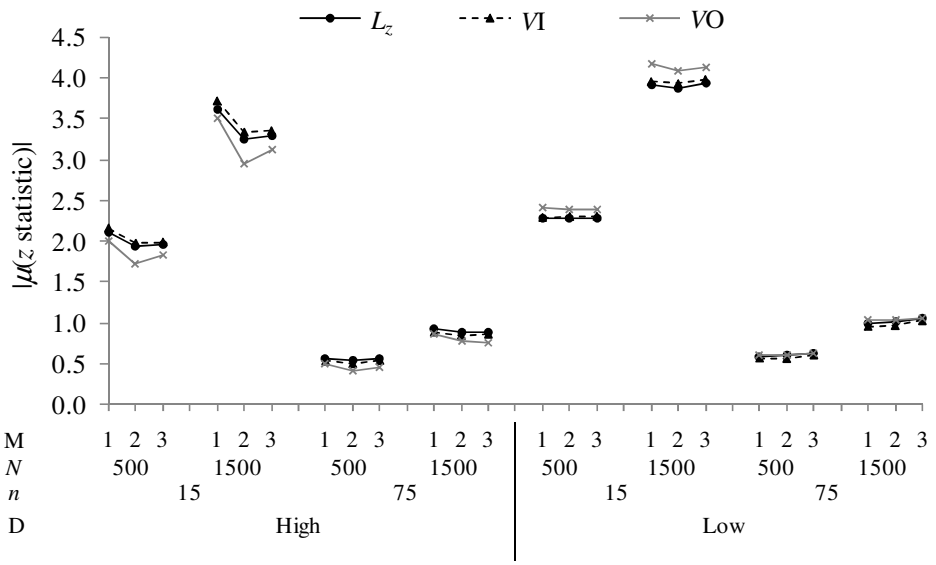
Table 32. Correlations Among *z* Statistics in SU Study Conditions

			Discrimination					
			High			Low		
Model	*N*	*n*	$r(L_z,VI)$	$r(L_z,VO)$	$r(VI,VO)$	$r(L_z,VI)$	$r(L_z,VO)$	$r(VI,VO)$
1PL	500	15	−0.99	−0.90	0.83	−1.00	−0.98	0.95
		75	−0.98	−0.80	0.68	−0.99	−0.93	0.87
	1,500	15	−0.99	−0.91	0.84	−1.00	−0.98	0.97
		75	−0.98	−0.80	0.68	−0.99	−0.93	0.88
2PL	500	15	−0.98	−0.74	0.61	−1.00	−0.97	0.95
		75	−0.86	−0.27	−0.15	−0.96	−0.78	0.60
	1,500	15	−0.99	−0.77	0.66	−1.00	−0.98	0.96
		75	−0.87	−0.36	−0.03	−0.97	−0.77	0.60
3PL	500	15	−0.99	−0.85	0.79	−1.00	−0.98	0.96
		75	−0.94	−0.45	0.19	−0.98	−0.82	0.69
	1,500	15	−0.99	−0.86	0.80	−1.00	−0.98	0.97
		75	−0.94	−0.41	0.17	−0.98	−0.85	0.75

***Empirical sampling distribution moments*.** The empirical sampling distribution means (μ) were very similar across all three *z* statistics, though they were positive for L_z

and negative for *V*I and *V*O. The absolute values of the distribution means are graphically depicted in Figure 29. The means for L_z and *V*I tended to be more similar to one another than the mean for *V*O, which is not surprising given the pattern of correlations in Table 32. The largest absolute difference between $|\mu(L_z)|$ and $|\mu(V\text{I})|$, which occurred for the 1PL when $n = 15$ and $N = 1{,}500$, was only 0.08. The largest absolute difference between $|\mu(L_z)|$ and $|\mu(V\text{O})|$ was 0.31, and that between $|\mu(V\text{I})|$ and $|\mu(V\text{O})|$ was 0.38; both of these occurred for the 2PL when $n = 15$ and $N = 1{,}500$. The differences between the means of the z statistics were minor in comparison with their magnitudes, which exceeded 3.0 when $n = 15$ and $N = 1{,}500$, were near or in excess of 2.0 when $n = 15$ and $N = 500$, and near 1.0 when $n = 75$ and $N = 1{,}500$. The means were closest to expectation (of $\mu = 0$) when $n = 75$ and $N = 500$, but still noticeably inflated, with values ranging from 0.41 to 0.63.

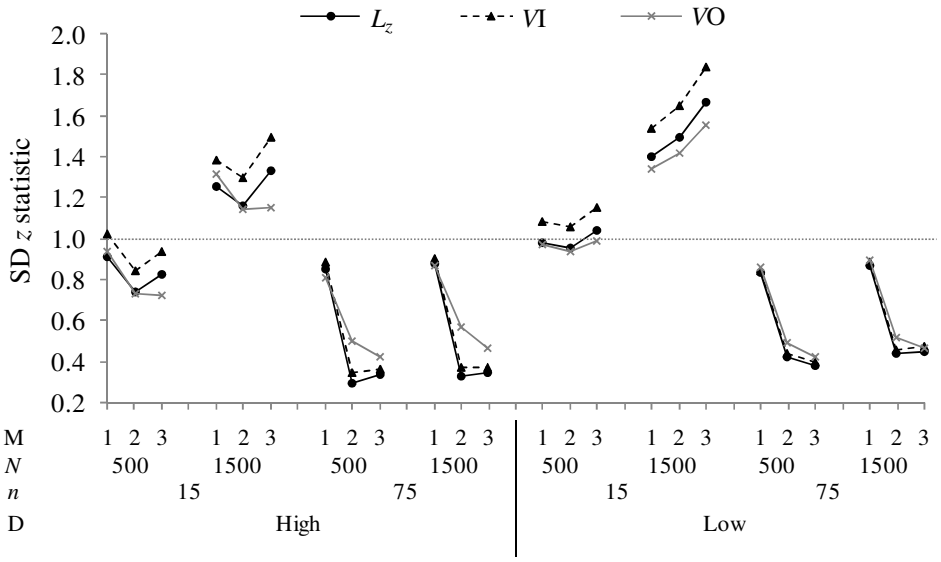
Figure 29. Absolute Values of Empirical Sampling Distribution Means for z Statistics in SU Study Conditions



The empirical sampling distribution SDs were also very similar across all three fit statistics, though they exhibited more variability than μ across fit statistics and study conditions. The SDs are graphically depicted in Figure 30. The most striking feature of the second moments is their extreme deflation for the 2PL and 3PL when $n = 75$; in these

conditions the SDs ranged from 0.30 to 0.57. When n = 75, SDs were only somewhat deflated for the 1PL, ranging from 0.81 to 0.90. SDs were inflated when n = 15 and N = 1,500, ranging from 1.15 to 1.38 in the high discrimination conditions and from 1.34 to 1.84 in the low discrimination conditions. SDs were closest to expectation when n = 15 and N = 500 where they ranged from 0.72 to 1.02 in the high discrimination conditions and from 0.93 to 1.15 in the low discrimination conditions. On average, z statistic sampling distribution SDs tended to be least extreme for the 1PL than the 2PL or 3PL; across SU study conditions the absolute difference between observed SDs and their expected values for L_z, VI, and VO averaged, respectively, (1) 0.17, 0.19, and 0.16 for the 1PL; (2) 0.43, 0.45, and 0.35 for the 2PL; and (3) 0.46, 0.49, and 0.40 for the 3PL.

Figure 30. Empirical Sampling Distribution SDs for z Statistics in SU Study Conditions



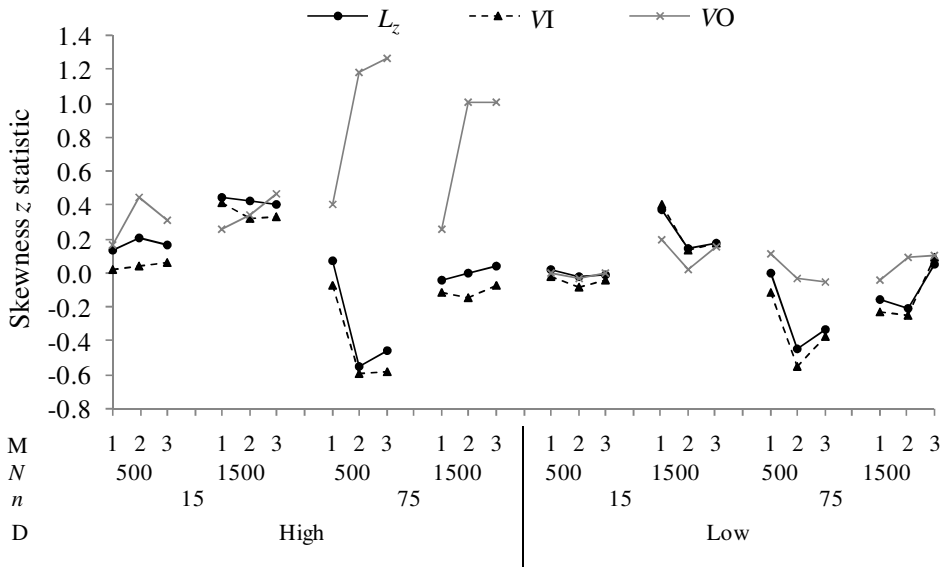
The skewness values of the empirical sampling distributions for the z statistics are graphically depicted in Figure 31. As noted above, there was symmetry between the μ for L_z (which were positive) and those for VI, and to a somewhat lesser extent VO, (which were both negative) since the magnitude of the μ was similar between the z statistics. This symmetry between L_z and VI was also evident for distributional skewness. To allow

for a better comparison between z statistic skewness values, the sign of L_z's skewness was reversed in Figure 31 to make it comparable to that for *V*I and *V*O.

The most striking feature of Figure 31 is the difference in skewness between *V*O and the other two z statistics in high discrimination $n = 75$ conditions. In these conditions *V*O distributions for the 2PL and 3PL were highly positively skewed (ranging from 1.00 to 1.26) whereas L_z and *V*I skewness in these conditions was either fairly close to expectation (ranging from −0.15 to 0.03) when $N = 500$ or negative (ranging from −0.46 to −0.59) when $N = 1{,}500$.

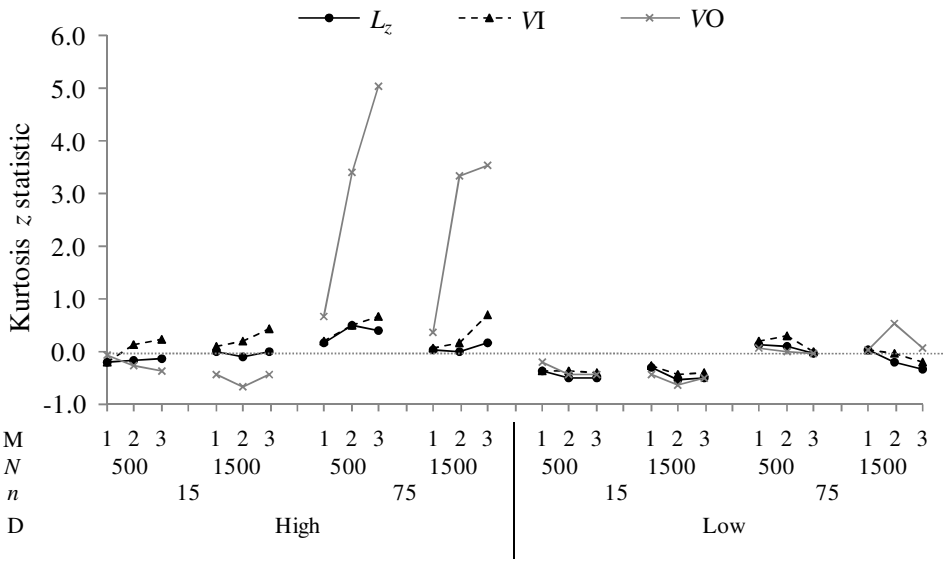
On average, z statistic sampling distribution skewness tended to be least extreme for the 1PL than the 2PL or 3PL; across SU study conditions, the absolute value of skewness for L_z, *V*I, and *V*O averaged, respectively, (1) 0.15, 0.17, and 0.18 for the 1PL; (2) 0.25, 0.27, and 0.39 for the 2PL; and (3) 0.20, 0.22, and 0.42 for the 3PL. Skewness was within 0.10 of expectation across all three IRT models for (1) all z statistics in low discrimination conditions when $n = 15$ and $N = 500$, (2) *V*I in high discrimination conditions when $n = 15$ and $N = 500$, (3) L_z in high discrimination conditions when $n = 75$ and $N = 1{,}500$, and (4) *V*O in low discrimination conditions when $n = 75$ and $N = 1{,}500$.

Figure 31. Skewness of Empirical Sampling Distributions for *z* Statistics in SU Study Conditions



The kurtosis values of the empirical sampling distributions for the *z* statistics are graphically depicted in Figure 32. The most striking feature of this figure is the extreme kurtosis values for *V*O (which ranged from 3.31 to 5.04) when used with the 2PL and 3PL in high discrimination conditions when $n = 75$. On average, *z* statistic sampling distribution kurtosis tended to be least extreme for the 1PL than the 2PL or 3PL; across SU study conditions the absolute value of kurtosis for L_z, *V*I, and *V*O averaged, respectively, (1) 0.15, 0.18, and 0.29 for the 1PL; (2) 0.27, 0.27, and 1.16 for the 2PL; and (3) 0.27, 0.38, and 1.31 for the 3PL

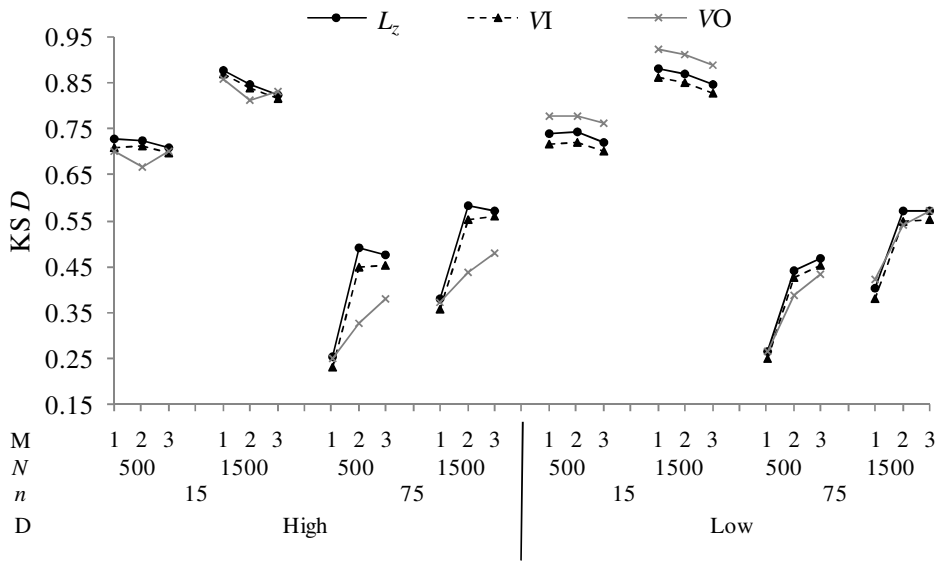
Figure 32. Kurtosis of Empirical Sampling Distributions for *z* Statistics in SU Study Conditions



***Kolmogorov-Smirnov tests*.** Not surprisingly, based on aberrancies in the empirical sampling distribution moments described above, the null hypothesis that the statistics followed a standard normal distribution was rejected for each test replication in which the KS test was conducted. In addition to being applied within each test replication, KS tests were also applied across all replications, to the entire set of 18,750 item replications in each study cell (resulting in 48 tests). In order to compare the magnitude of the departures from the standard normal distribution across *z* statistics and study conditions, the KS *D* statistics from the tests conducted across test replications were plotted and are shown in Figure 33. As a point of reference to interpret the magnitude of the KS *D* statistics in Figure 33, a KS test was conducted on a random sample of 18,750 data points that were drawn from a standard normal distribution. The test resulted in a *D* of 0.0097 with $p = 0.059$. Even the smallest KS *D* of 0.23 obtained from the *z* statistics was very extreme. There was not a marked difference between the three *z* statistics in KS *D* values, indicating that the statistics all had similar degrees of departure from a standard normal distribution. KS *D* values were least extreme for the 1PL when $n = 75$ and $N = 500$, and in fact, these were the conditions in which T1 rates tended to be closest to nominal levels. T1 rates also tended to be near nominal levels for

the 2PL and 3PL in low discrimination conditions when $n = 75$ and $N = 1{,}500$, but this was due to the combination of a somewhat inflated mean and markedly deflated variance for these fit statistics in those conditions, rather than the statistics approximating a standard normal distribution.

Figure 33. *D* Values from KS Tests Conducted Across Replicated Tests for *z* Statistics in SU Study Conditions



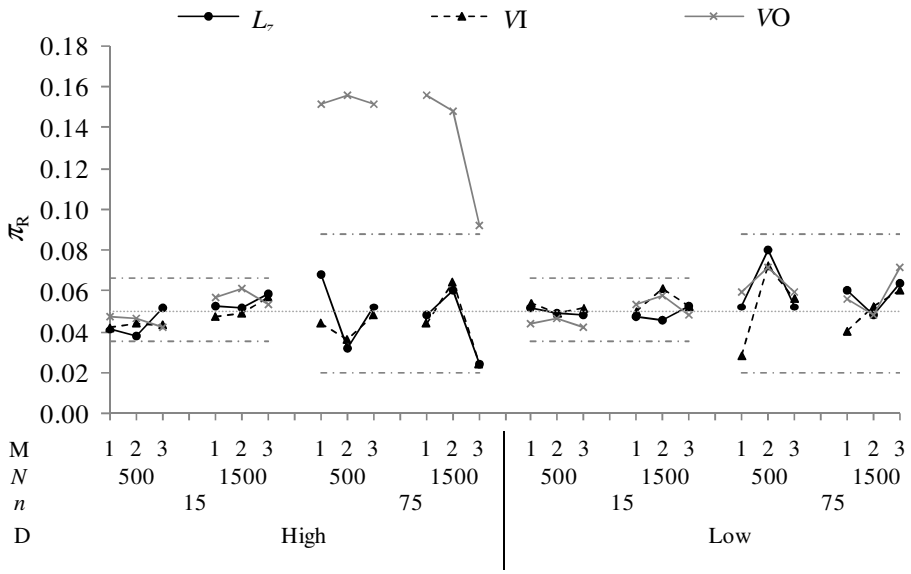
Impact of parameter estimation error. The mean and variance of the empirical sampling distributions for the *z* statistics in all SU conditions can be found in Appendix H (Tables H-1 to H-4). The mean moved closer to expectation when $\boldsymbol{\theta}$, as opposed to $\hat{\boldsymbol{\theta}}$, was used to compute the statistics, and was closest to expectation when $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ were used. For example, across all SU conditions, L_z means ranged from 0.54 to 3.94 in $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ conditions, from 0.34 to 3.93 in $\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}$ conditions, from –0.12 to 0.25 in $\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$ conditions, and from –0.01 to 0.01 in $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions.

Item parameter estimation error caused deflations in the variance of the empirical sampling distributions for all *z* statistics. For example, when $\hat{\boldsymbol{\theta}}$ was used to compute L_z, $\boldsymbol{\xi}$ condition SDs (which ranged from 0.98 to 2.06) were greater than $\hat{\boldsymbol{\xi}}$ condition SDs (which ranged from 0.30 to 1.67) by 0.10 to 0.76 units across SU study conditions, with

the differences averaging 0.44; when $\boldsymbol{\theta}$ was used to compute L_z, $\boldsymbol{\xi}$ condition SDs (which were very close to expectation, ranging from 0.99 to 1.01) were greater than $\hat{\boldsymbol{\xi}}$ condition SDs (which ranged from 0.39 to 0.95) by 0.05 to 0.60 units across SU study conditions, with the differences averaging 0.31.

Both the mean and SD were very close to expectation for the z statistics when $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ were used to compute the statistics. Thus, it appears that the z statistics followed their theoretical distributions when model parameters were known. The KS tests conducted within each replicated test in $\boldsymbol{\xi},\boldsymbol{\theta}$ conditions generally supported this conclusion. This result is summarized in Figure 34, which displays π_R in SU $\boldsymbol{\xi},\boldsymbol{\theta}$ study conditions. The π_R might be expected to exhibit greater fluctuations about 0.05 in the n = 75 than the n = 15 conditions, since the number of test replications (TR) was smaller in the former than the latter conditions. In order to establish expected bounds for π_R, its empirical sampling distribution (across 10,000 replications) was inspected when (1) n = 15 and TR = 1250 and (2) n = 75 and TR = 250; The π_R at the 0.5% and 99.5% quantiles of this distribution were used to interpret the magnitude of the π_R in Figure 34; the lower and upper bounds of this 99% CI (indicated by horizontal dashed lines in Figure 34) were, respectively, 0.0352 and 0.0664 for n = 15, and 0.0200 and 0.088 for n = 75.

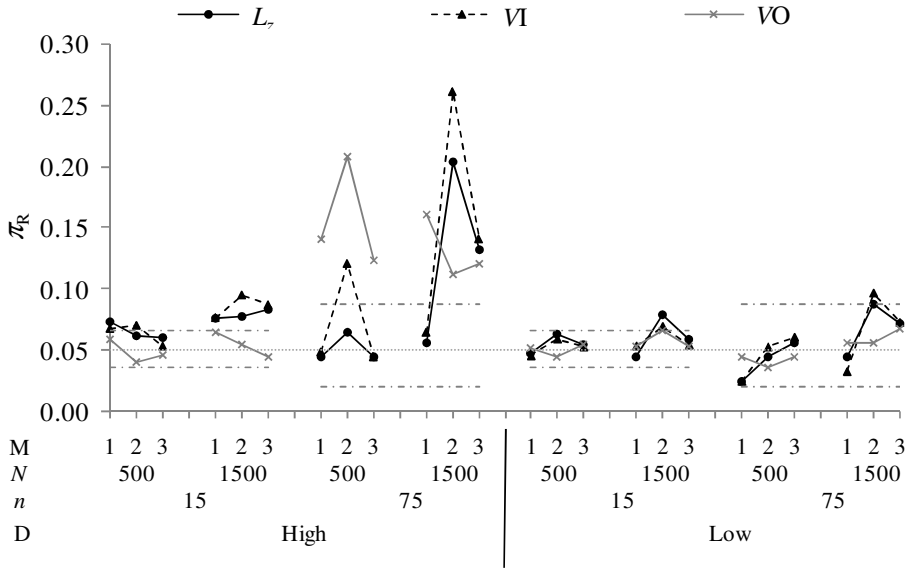
Figure 34. Proportion of Replicated Tests (π_R) in Which the KS Test Null Hypothesis Was Rejected for z Statistics in SU ξ, θ Study Conditions



The π_R were within the bounds of the 99% CI for all statistics and conditions other than *V*O in high discrimination $n = 75$ conditions. However, in these conditions *V*O means (which ranged from 0.01 to 0.05) and SDs (which ranged from 0.89 to 0.94) were not markedly distorted compared with the other conditions, where means ranged from < 0.005 to 0.03 and SDs ranged from 0.90 to 1.00. Further inspection of the SU results for *V*O revealed that skewness, and especially, kurtosis were markedly distorted for *V*O in *all* high discrimination ξ, θ conditions, where skewness ranged from 0.73 to 1.45 and kurtosis ranged from 3.32 to 11.97. For comparison, L_z and *V*I skewness in the same conditions ranged from −0.19 to −0.07 and −0.06 to 0.04 respectively; kurtosis for the two respective statistics ranged from −0.04 to 0.08 and −0.04 to 0.08. Skewness and Kurtosis for all SU ξ, θ conditions can be found in Appendix H (Table H-5). It is likely that *V*O π_R was not markedly inflated in $n = 15$ high discrimination conditions due to lack of power, since KS test power is related to sample size.

Finally, the impact of data noise in the ξ, θ conditions was examined. Figure 35 displays π_R in EU ξ, θ study conditions and, compared with Figure 34, it is clear that the presence of data noise disrupted the sampling distributions of the z statistics.

Figure 35. Proportion of Replicated Tests (π_R) in Which KS Test Null Hypothesis Was Rejected for z Statistics in EU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Study Conditions



Relationship between z statistics and item parameters. Scatterplots between z statistics and item b parameters (which can be found in figures H-1 to H-12 of Appendix H) revealed some interesting patterns. For all models, when n = 15 a V-shaped pattern was evident, indicative of a linear relationship between $|b|$ and the z statistics. This pattern was somewhat asymmetrical for the 3PL, mildly apparent for the 2PL and 3PL when N = 1,500 and n = 75, and, for all models, more pronounced when N = 1,500 as opposed to when N = 500.

For the 1PL when n = 75 a diamond-shaped pattern was evident in which items with z statistics at the positive and negative extremes had b parameters concentrated around zero while items with z statistics near zero had b parameters spanning the entire b range. *V*O exhibited less of this pattern since items with extreme positive values of *V*O spanned a wider b range. For the 2PL the diamond-shaped pattern was evident only for L_z and *V*I in the low discrimination conditions when n = 75 and N = 500. This pattern was less clear for the 3PL in these conditions.

There was a moderate to strong linear relationship, summarized by Pearson correlations in Table 33, between the z statistics and item a parameters. For the 2PL and

3PL, correlations between L_z and *a* ranged from, respectively, 0.49 to 0.87 and 0.53 to 0.78 across high discrimination conditions; across the low discrimination conditions, respective ranges for the two models were 0.52 to 0.92 and 0.53 to 0.81. For *V*I, correlations were similar in magnitude (though negative) to those for L_z, ranging from −0.53 to −0.75 for the 2PL and −0.52 to −0.68 for the 3PL across high discrimination conditions; in low discrimination conditions, respective ranges for the two models were −0.51 to −0.87 and −0.52 to −0.76. In the high discrimination conditions, correlations between *V*O and *a* parameters tended to be less extreme than for the other two *z* statistics, ranging from −0.05 to −0.19 for the 2PL and −0.22 to −0.48 for the 3PL; in low discrimination conditions, correlations were similar in magnitude to those for L_z and *V*I, ranging from −0.51 to −0.71 for the 2PL and −0.55 to −0.70 for the 3PL. For all three *z* statistics, correlations were at their maxima (ranging, in absolute value, from 0.70 to 0.92) when $N = 500$ and $n = 75$ and at (or near, for *V*O) their minima (ranging, in absolute value, from 0.07 to 0.53) when $N = 1{,}500$ and $n = 15$.

Table 33. Correlations Between *z* Statistics and Item *a* Parameters in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Study Conditions

N	*n*	*z* statistic	High Discrimination		Low Discrimination	
			2PL	3PL	2PL	3PL
500	15	L_z	0.62	0.68	0.64	0.65
		*V*I	−0.63	−0.66	−0.62	−0.63
		*V*O	−0.19	−0.48	−0.63	−0.67
	75	L_z	0.87	0.78	0.92	0.81
		*V*I	−0.75	−0.68	−0.87	−0.76
		*V*O	−0.14	−0.37	−0.71	−0.70
1,500	15	L_z	0.49	0.53	0.52	0.53
		*V*I	−0.53	−0.52	−0.51	−0.52
		*V*O	−0.07	−0.33	−0.51	−0.55
	75	L_z	0.71	0.65	0.78	0.64
		*V*I	−0.65	−0.57	−0.74	−0.59
		*V*O	−0.05	−0.22	−0.57	−0.58

When the *z* statistics were computed using $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ their relationships with *b* parameters noted above disappeared (see Figures H-13 to H-24 in Appendix H).

Furthermore, correlations between z statistics and a approached zero, ranging from −0.02 to 0.01 for L_z, from −0.01 to 0.02 for VI, and from −0.01 to 0.07 for VO (see Table 34).

Table 34. Correlations Between z Statistics and Item a Parameters in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Study Conditions

N	n	z statistic	High Discrimination		Low Discrimination	
			2PL	3PL	2PL	3PL
500	15	L_z	−0.01	0.01	−0.02	0.00
		VI	0.01	−0.01	0.02	0.00
		VO	0.07	0.01	0.02	0.00
	75	L_z	0.01	−0.01	0.00	0.00
		VI	−0.01	0.01	0.01	0.00
		VO	0.05	0.03	0.00	0.00
1,500	15	L_z	0.01	0.00	0.00	0.01
		VI	−0.01	−0.01	0.00	−0.01
		VO	0.01	0.01	0.00	−0.01
	75	L_z	0.01	0.01	0.00	0.00
		VI	0.00	0.00	0.00	0.00
		VO	0.02	0.00	0.00	0.00

z statistic summary. In the absence of item parameter estimation error and data noise, the z statistics, like QO and $Q1$, appeared to generally approximate their theoretical sampling distributions. The proportion (over all replicated tests) of KS test rejections was near expectation in all SU conditions other than high discrimination $n = 75$ conditions for VO (Figure 34). The moments for L_z and VI were very close to expectation across all SU conditions (Tables H-1 to H-5). To summarize these results, the average absolute errors (same form as ME in Equations 39 and 41) for the moments will be mentioned here. Across all SU study conditions, ME(Mean) and ME(SD) were 0.005 and 0.004, respectively, for both L_z and VI. L_z tended to be somewhat more skewed [ME(Skew) = 0.100] than VI [ME(Skew) = 0.019]. The kurtosis of L_z and VI were fairly comparable; ME(Kurt) was 0.038 for L_z and 0.032 for VI. VO moments (Tables H-1 to H-5) were closer to expectation in low discrimination conditions than in high discrimination conditions. In low discrimination conditions, ME(Mean), ME(SD), ME(Skew), and

ME(Kurt) were 0.006, 0.007, 0.180, and 0.299 respectively, whereas in high discrimination conditions these respective values were 0.017, 0.079, 1.089, and 5.974.

As for $Q1$ and QO, The introduction of data noise in the absence of parameter estimation error led to small but noticeable disruptions in the z statistics' sampling distributions. Across all statistics and EU conditions, 23/72 (32%) of KS test π_R exceeded expectation (Figure 35), which is elevated relative to SU conditions in which 4/72 (6%) of the π_R exceeded expectation. However, this effect was more evident for L_z and VI than VO; across SU and EU conditions, respective π_R averaged 0.051 and 0.069 for L_z, 0.049 and 0.075 and VI, and 0.075 and 0.076 for VO.

When the z statistics were computed with estimated model parameters, their sampling distributions were disrupted, departing significantly from expectation (according to KS tests) for all replicated tests across all study conditions. There also was no noticeable impact of data noise. Generally, the z statistics functioned very similarly to one another, though on average L_z was positive, and VI and VO were negative. In most study conditions, correlations between the statistics were quite high, ranging from −0.86 to −1.00 for $r(L_z,VI)$, −0.27 to −0.98 for $r(L_z,VO)$, and −0.03 to 0.97 for $r(VI,VO)$ (Table 32).

As was the case with $Q1$, the presence of θ estimation error caused inflations in the z statistic means; this effect was most evident in $n = 15/N = 1{,}500$ conditions (Table H-2), where means in SU $\hat{\boldsymbol{\theta}}$ conditions ranged (in absolute value) from 2.94 to 4.18, and least evident in $n = 75/N = 500$ conditions (Table H-3), where means in SU $\hat{\boldsymbol{\theta}}$ conditions ranged (in absolute value) from 0.34 to 0.63. An inflationary effect of θ estimation error on z statistic variances was also evident when $n = 15$ (Tables H-1 and H-2), especially when $N = 1{,}500$, where SDs in SU $\hat{\boldsymbol{\theta}}$ conditions ranged from 1.15 to 2.24 whereas those in $\boldsymbol{\theta}$ conditions ranged from 0.69 to 1.00. As with QO and Q1, the presence of ξ estimation error decreased sampling distribution variances, but this effect was more marked for the z statistics, for which variances fell well below expectation in some cases. Variances were most deflated (relative to expectation) in $n = 75/N = 500$ conditions,

where they ranged from 0.30 to 0.81; the most extreme deflations were observed for the 2PL and 3PL.

Finally, there were relationships between z statistics and item parameter values. For all models, when $n = 15$ a V-shaped pattern was evident, indicative of a linear relationship between $|b|$ and the z statistics (Figures H-1, H-2, H-5, H-6, H-9, H-10). This pattern was also evident in some cases, though to a lesser degree, when $n = 75$ (Figures H-8 and H-12). Another pattern, evident in some cases when $n = 75$ (Figures H-3 and H-4), was one such that items with z statistics at the positive and negative extremes had b parameters concentrated around zero while items with z statistics near zero had b parameters spanning the entire b range. There were also moderate to strong linear correlations between a and the z statistics (Table 33), ranging from 0.49 to 0.92 for L_z, −0.51 to −0.87 for VI, and −0.05 to −0.71 for VO. All of these relationships between z statistics and item parameters were not present in the absence of parameter estimation error (Figures H-13 to H-24 Table 34).

CHAPTER 4: DISCUSSION AND CONCLUSIONS

Fit Statistic Functionality in Ideal Conditions

In the absence of parameter estimation error and data noise, the fit statistics appeared to generally approximate their theoretical distributions, a result that has also been found by Smith (1991) for *V*I and *V*O, and Stone and Hansen (2000) for a variant of *Q*1 (*Q*B). Generally, in SU conditions distribution moments showed little departure from expectation other than those for *Q*1, *Q*O, and *V*O which, in some instances, were somewhat more variable than would be expected if all computed statistics followed their theoretical distribution. These mild aberrations had little impact on Type I error (T1) rates, which, for example at $\alpha = 0.05$, ranged from 0.046 to 0.056 (averaging 0.052) for *Q*O, 0.046 to 0.052 (averaging 0.049) for *Q*1, 0.047 to 0.052 (averaging 0.049) for L_z, 0.046 to 0.053 (averaging 0.050) for *V*I, and 0.029 to 0.050 (averaging 0.041) for *V*O across all SU conditions. There was a small effect for item discrimination on *V*O's distribution, which led to somewhat deflated T1 rates (averaging 0.035 at $\alpha = 0.05$) in high discrimination conditions. This deflationary effect was, at least partially, due to restriction in *V*O variances in high discrimination conditions (where SDs ranged from 0.89 to 0.96) but not in low discrimination conditions (where SDs ranged from 0.98 to 1.00).

Introduction of Data Noise

When true model parameters were used to compute fit statistics, data noise (DN) had, of all study effects, among the strongest impact on T1 rates; for example, effect sizes on T1 rates at $\alpha = 0.05$ were 0.42 for *Q*1, 0.20 for *Q*O, 0.45 for L_z, 0.47 for *V*I, and 0.16 for *V*O. For comparison, DN effects were nearly nonexistent when estimated model parameters were used to compute fit statistics, with (for example at $\alpha = 0.05$) main effect DN η^2 ranging from < 0.0001 (for *Q*1, and the *z* statistics) to 0.0027 (for *Q*O); η^2 for all interactions involving DN were of similar magnitude. These effects were seen in the distributional analyses of the fit statistics, which revealed slight disruptions relative to SU

conditions for all of the fit statistics (other than LM statistics, which were not examined when there was no parameter estimation error).

The effect of DN on fit statistic sampling distributions translated into very slightly higher (relative to nominal T1 rates) proportions of items flagged as misfitting. For example, at $\alpha = 0.05$, the proportion of items in EU conditions flagged as misfitting ranged from 0.047 to 0.072 (averaging 0.057) for $Q1$, 0.048 to 0.065 (averaging 0.056) for QO, 0.048 to 0.082 (averaging 0.062) for L_z, (0.051 to 0.078 (averaging 0.062) for VI, and 0.032 to 0.073 (averaging 0.049) for VO. For all fit statistics these rates are slightly higher, on average, than those in the SU conditions.

The fit statistics thus appeared to function reasonably well when either there was not parameter estimation error or when they were free from θ estimation error but contained a small amount of ξ estimation error, as would be the case if item parameters were estimated for an essentially unidimensional test in a very large sample (N = 10,000 in the case of this study). Of course, these conditions are very restrictive and unrealistic, especially for $Q1$ and the z statistics which are not free from θ estimation error in practice.

Impact of Parameter Estimation Error

Estimation error in θ clearly had the most impact on fit statistic functionality and led to grossly distorted distributions for $Q1$ and the z statistics. For example, $Q1$ and L_z T1 rates averaged, respectively, 0.386 and 0.378 across all $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ study conditions, 0.367 and 0.407 across all $\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$ conditions, 0.072 and 0.015 across all $\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$ conditions, and 0.053 and 0.056 across all $\boldsymbol{\xi},\boldsymbol{\theta}$ conditions. Furthermore, in the presence of θ estimation error, study factors had more impact on fit statistic functionality. For example, the percentage of total variation in $Q1$ T1 rates accounted for by all study effects combined was 47.2% in $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ conditions, 44.5% in $\boldsymbol{\xi},\hat{\boldsymbol{\theta}}$ conditions, 8.1% in $\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$ conditions, and 1.3% in $\boldsymbol{\xi},\boldsymbol{\theta}$ conditions. The z statistics had similar percentages, except for $\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$ conditions where study effects had a degree of impact similar to that in parameter estimation error (PE)

conditions involving $\hat{\boldsymbol{\theta}}$; however, the degree of overall variation in T1 rates in the $\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$ condition was very small, so, as evident in the $\boldsymbol{\xi},\boldsymbol{\theta}$ conditions, even those study factors with large effect sizes had minor impact on T1 rates.

As noted earlier, for $\boldsymbol{\xi},\boldsymbol{\theta}$ conditions, notable (but of small practical significance) effects were present for DN. However DN effects were not evident in other PE conditions where η^2 for DN ranged from < 0.0001 to 0.0027 across fit statistics and (non-$\boldsymbol{\xi},\boldsymbol{\theta}$) PE conditions; η^2 for all interactions involving DN were also near zero. In $\hat{\boldsymbol{\xi}},\boldsymbol{\theta}$ conditions, the majority of impact on $Q1$ and z statistic T1 rates was due to IRT model (M) (e.g. $\eta^2 >$ 0.70 for T1 rates at $\alpha = 0.05$) whereas in both conditions involving $\hat{\boldsymbol{\theta}}$ the majority of impact was due to n. When QO was computed with estimated item parameters M also accounted for a substantial, but lesser, proportion of T1 rate variation (e.g. $\eta^2 = 0.46$ for T1 rates at $\alpha = 0.05$). Further discussion of factors influencing fit statistics in full parameter estimation error (e.g. $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ for $Q1$ and z statistics and $\hat{\boldsymbol{\xi}}$ for QO) conditions can be found in the next section.

While estimation error in θ caused inflations in sampling distribution means and, to a much lesser degree, variances, there was also a subtle but consistent impact of ξ estimation error which caused restrictions in the variance of the fit statistic distributions. For QO, and $Q1$ with $\hat{\boldsymbol{\theta}}$ (but not $\boldsymbol{\theta}$), in nearly every study condition Bias(SD) in $\hat{\boldsymbol{\xi}}$ conditions was less than Bias(SD) in $\boldsymbol{\xi}$ conditions. z statistic SDs were also consistently lower in $\hat{\boldsymbol{\xi}}$ than $\boldsymbol{\xi}$ conditions. The reductions in variance resulted in variances falling below expectation in some cases, especially when $n = 75$ and $N = 500$, where $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}$ Bias(SD) averaged, respectively, –0.34 and 0.01 for QO, –0.06 and 0.68 for $Q1$, and SDs in the respective conditions averaged 0.52 and 1.05 for L_z, 0.54 and 1.07 for VI, and 0.58 and 0.96 for VO; for the z statistics, the lowest SDs were observed for the 2PL and 3PL. Smith (1991) also reported an inflationary impact of θ estimation error on VI and VO means and a deflationary impact of ξ estimation error on variances for the 1PL; contrary to the results of the present study, Smith reported minor restrictions in VI and VO

variance when estimated, as opposed to true, θ parameters were used. Smith's study only concerned the 1PL, but the results of the present study suggest that restrictions in z statistic variance due to ξ error are more severe for the 2PL and 3PL than the 1PL. Stone and Hansen (2000) also reported mean and variance inflation due to θ error for a variant of $Q1$ (QB) applied to a polytomous IRT model. As with the present study, they found that the impact of θ estimation error was stronger for shorter than longer tests and larger than smaller sample sizes.

The reason for the variance deflations in the presence of ξ error is not clear, but could be a result of the priors imposed on the item parameters during estimation. These deflations had little impact on T1 rates for QO and $Q1$, but caused noticeable T1 deflations for the z statistics in all $\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$ (and 2PL/3PL $n = 75$ $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$) conditions. For example, at $\alpha = 0.05$, across all $\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$ conditions, T1 rates ranged from < 0.001 to 0.042 (averaging 0.015) for L_z, < 0.001 to 0.044 (averaging 0.015) for VI, and 0.001 to 0.039 (averaging 0.039) for VO, whereas for $Q1$ and QO these ranges were, respectively, 0.049 to 0.119 (averaging 0.072) and 0.041 to 0.074 (averaging 0.052). T1 rates were somewhat elevated for $Q1$ in the $\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$ conditions due to the presence of an effect for IRT model such that Bias(Mean) increased with model complexity. This effect disappeared when $\boldsymbol{\xi}$ were used with $\boldsymbol{\theta}$.

Fit Statistic Functionality in Realistic Conditions

When estimated ξ and θ parameters were used to compute the fit statistics, as would be done in practice, there was a large degree of variation in T1 rates, most of which was attributable to fit statistic (FS). At all three α levels, all effects involving FS accounted for approximately 66% of T1 rate variation; the largest among these effects were the main effect for FS (η^2 ranging from 0.20 to 0.25) and the interaction between FS and n (η^2 ranging from 0.24 to 0.31). Small to moderate effects (ranging from 0.01 to 0.08) were also evident for FS × Discrimination, FS × M, FS × N, and FS × N × n. The only between-subjects study factor with substantial effects on T1 rates was n (η^2 ranging

from 0.22 to 0.30). Additionally, small effects (ranging from 0.01 to 0.05) were also evident for M, *N,* and $N \times n$.

Overall, T1 rates were inflated for all fit statistics other than *Q*O, and to a lesser extent, LM(β); for example, across all study conditions combined, at $\alpha = 0.05$ T1 rates were 0.05 for *Q*O, 0.07 for LM(β), 0.12 for LM(α), between 0.38 and 0.39 for *Q*1 and the *z* statistics, and 0.42 for LM($\alpha\beta$). The FS × *n* interaction was primarily due to inflated T1 rates at short test lengths for *Q*1 and the *z* statistics, but not *Q*O and the LM statistics. The FS × *N* × *n* interaction was attributable to *N* having a much stronger positive (inflationary) impact on T1 rates for *Q*1 and the *z* statistics when $n = 15$ than when $n = 75$.

Effects involving M were attributable to (1) negative relationships between model complexity and T1 rates for LM statistics, especially LM($\alpha\beta$); (2) slightly elevated 3PL T1 rates, relative to 1PL and 2PL rates, for *Q*1; and (3) slightly elevated 1PL T1 rates, relative to 2PL and 3PL rates, for the *z* statistics. Two other minor observable effects were small elevations in T1 rates for the LM statistics when $N = 500$ as opposed to when $N = 1{,}500$ and somewhat higher T1 rates in high than in low discrimination conditions for the LM statistics, with the reverse being true for *Q*1 and the *z* statistics.

*Q*O was clearly the only fit statistic that adhered to nominal T1 rates across all realistic study conditions. For *Q*1 and the *z* statistics, test length had a large degree of impact on T1 rates. Sample size, discrimination, and IRT model had smaller effects on all fit statistics (other than *Q*O), except for LM($\alpha\beta$) which was strongly impacted by IRT model. There were conditions in which T1 rates for *Q*1, LM(α), LM(β), and the *z* statistics approached nominal T1 rates, suggesting that these statistics adhere to their theoretical sampling distributions under certain conditions. This will be discussed further below, where the results of the distributional analyses for the fit statistics are summarized.

*Q*O Sampling Distribution

Though there did not appear to be any marked relationship between *K* (the number of score categories) and *Q*O sampling distribution means and variances, there did appear to be some systematic bias in the moments across *K*, especially for 3PL means which tended to be inflated. The results of the KS tests indicated that *Q*O did not strictly follow its theoretical distribution since π_K were generally greater than expected, especially for the 3PL and all models when *n* was smaller. However, the practical significance of these departures warrants attention. The minimum T1 rates (which all occurred in 2PL, $N = 500$ $n = 75$ conditions) at $\alpha = 0.01$, 0.05, and 0.10 across all study conditions were, respectively, 0.0058, 0.0412, and 0.0892. The maximum T1 rates at the three respective α levels (which occurred for the 3PL when $N = 500$ and $n = 15$) were 0.0171, 0.0737, and 0.1427 respectively; these T1 rate ranges are comparable to those reported in other studies (Glas & Falcon, 2003; Orlando & Thissen, 2000; Sinharay & Lu, 2008; Stone & Zhang, 2003). This degree of aberration in T1 rates does not markedly affect the number of items on any one test expected to be identified as misfitting by chance.

*Q*1 Sampling Distribution

*Q*1 distributions tended to become less aberrant as *K* increased for the 3PL, but not the other two models. No other relationships were observed between *K* and *Q*1 means and SDs. Across *K*, both moments were markedly inflated when $n = 15$, which caused the grossly inflated T1 rates observed for *Q*1 when $n = 15$. Not surprisingly, all KS tests were statistically significant when $n = 15$.

When $n = 75$, there tended to be less positive bias in means and SDs, especially for the 1PL and 2PL when $N = 500$; however, in this condition, where means appeared to be most on target, SDs were fairly consistently somewhat below expectation. KS test π_K were less than 1.0 for the 1PL and 2PL, but not the 3PL. However, for the 1PL and 2PL, π_K were still far from the 0.05 that would be expected if the statistics followed their

theoretical distributions; in agreement with T1 rate results and the analysis of moments, π_K were somewhat lower when $N = 500$ than when $N = 1{,}500$.

T1 rate inflation for $Q1$ (and similar fit statistics) when n is small as opposed to large has been reported by other studies (DeMars, 2005; Glas & Falcon, 2003; Orlando & Thissen, 2000; Stone & Hansen, 2000; Stone & Zhang, 2003). The closer approximation of $Q1$ to its theoretical distribution when $n = 75$ might be expected, given the impact of θ estimation error on $Q1$'s sampling distribution since, compared with the smaller sample size condition, θ should be more accurately estimated. Furthermore, the even closer approximation when $N = 500$ might also be expected since less θ error would be accumulated during $Q1$'s computation when N is smaller as opposed to larger; this effect of N has also been reported elsewhere (McKinley & Mills, 1985; Glas & Falcon, 2003). Finally, the higher than expected significance of the KS tests when $N = 500$ and $n = 75$, where $Q1$ means and T1 rates were most on target, could be due to the mild deflations in $Q1$ variance in this condition, which likely were attributable to effects of ξ estimation error.

Though $Q1$ approached, but did not strictly follow, its theoretical sampling distribution when $n = 75$, the practical significance of these departures warrants attention. For example, when $N = 500$, T1 rates were slightly deflated for the 1PL (ranging from 0.039 to 0.047 at $\alpha = 0.05$) and 2PL (ranging from 0.044 to 0.048 at $\alpha = 0.05$) but even the most extreme among these rates would not result in markedly fewer (relative to nominal α levels) items being identified as misfitting. The small amount of inflation in T1 rates for the 3PL when $N = 500$ (ranging from 0.078 to 0.087 at $\alpha = 0.05$) would result in at most three more (out of 75) items (relative to nominal α levels) being identified as misfitting. T1 rates were somewhat more inflated when $N = 1{,}500$, reaching a maximum of near 0.08 (at $\alpha = 0.05$) for the 1PL and 2PL, which would result in about two more items (relative to nominal α levels) being identified as misfitting. Finally, with the large sample size, T1 rates for the 3PL (ranging from 0.126 to 0.153 at $\alpha = 0.05$) began to become too inflated for practical use as between roughly six and eight more items (relative to nominal α levels) might be identified as misfitting.

LM Statistic Sampling Distributions

The LM statistics generally did not follow their theoretical distributions due to the presence of inordinate percentages of statistics with extremely high values and/or the overrepresentation of statistics with low values. There tended to be less distributional aberration in the low discrimination conditions and when *N* was large as opposed to small. IRT model also impacted LM distributions, but the nature of the effect differed between the three LM statistics; distributional aberrations tended to: (1) decrease with model complexity for LM($\alpha\beta$), (2) increase with model complexity for LM(β), and (3) be lower for the 2PL than 1PL and 3PL for LM(α). The aberrances in the upper tail of the distributions also were much more marked for LM($\alpha\beta$) than for LM(α) or LM(β). In fact, in the low discrimination conditions the upper tail of LM(α) and LM(β) distributions did tend to approach expectation; however, in these cases the distributions were disrupted by the overrepresentation of low values. The observed distributions were only very close to the theoretical target for LM(β) with the 1PL in the low discrimination conditions when $N = 1{,}500$.

In addition to yielding grossly inflated values, LM statistics also yielded negative values in a handful of cases. Inspection of these and the most inflated cases, as well as the over-representation of inflated values in the lower as opposed to higher difficulty ranges, indicated some computationally problematic aspects of the LM statistics— specifically, lack of statistical information in the reference group can seriously distort the LM statistics. Some of the most inflated and all of the negative LM values corresponded with cases in which the reference category contained only one number-correct score. Furthermore, the presence of grossly inflated LM statistics markedly decreased after recomputation using the lowest category as the reference group when $b \leq 0$ and the highest category as the reference group when $b > 0$.

Though LM departed from its theoretical sampling distributions, in practice it still might be useful for assessing fit, but only in some circumstances and for some IRT models. LM($\alpha\beta$) clearly would not be useful as its T1 rates were too inflated, even in the

best circumstances where, for example, the rates only began to approach nominal levels for the 3PL in low discrimination conditions when N = 1,500, where they were near 0.08 (at $\alpha = 0.05$). Across N and n, T1 rates for LM(α) tended to be at or near nominal levels for the 3PL in high discrimination conditions (where rates, at $\alpha = 0.05$, ranged from 0.03 to 0.09), and the 2PL and 3PL in low discrimination conditions (where rates ranged from 0.02 to 0.07). Finally, across N and n, LM(β) T1 rates tended to be near nominal levels for the 3PL in high discrimination conditions (where rates ranged from 0.03 to 0.07) and for all models in low discrimination conditions (where rates ranged from 0.02 to 0.06). However, given the effects of N and discrimination, T1 rates might be expected to be somewhat deflated when LM(α) and LM(β) are used to assess item fit in large samples, especially when discrimination is low (see Tables 6 and 7).

To date, only one other Monte Carlo simulation study (Glas & Falcon, 2003) could be found in which the functionality of LM statistics was examined. The study was limited to LM(β) as applied to the 3PL and used levels of a similar to high discrimination conditions in the present study, but generated data from a more restricted b-value range (−1.35 to 1.35) that was centered about zero, unlike the present study where generating b-value means for the 3PL were somewhat above zero (ranging from 0.15 to 0.25) and spanned a wider range (−4.5 to 4.0). The T1 rates reported by Glas and Falcon were comparable to, but slightly higher than, the results of the present study; Glas and Falcon's T1 rates (at $\alpha = 0.05$) ranged from 0.04 in N = 4,000 (largest sample size) conditions to 0.09 in an $N = 500/n = 40$ (smallest sample size and the longest test length) condition. In the present study, T1 rates for 3PL LM(β) ranged from 0.03 (when N = 1,500 and n = 15) to 0.07 (when N = 500 and n = 75). The subtle deflationary effect of N on T1 rates observed in this study was also evident in Glas and Falcon's results. Finally, Glas and Falcon also examined LM(β) when nine (as opposed to five) groups were used in the computation of the statistic and reported somewhat inflated T1 rates (ranging from 0.05 to 0.11), relative to computation with five groups. Inflated T1 rates might be expected with more score groups due to increased likelihood that the reference group will contain inadequate information.

z Statistic Sampling Distributions

Generally, the z statistics functioned very similarly to one another. In most study conditions, correlations between the statistics were quite high, especially between L_z and *V*I. In no study condition did the statistics appear to follow a standard normal distribution as KS test π_R was at unity in all conditions. However, distributions generally tended to be least aberrant for the 1PL compared with the other two models.

Because the z statistics rely of θ estimates, like $Q1$, their distributions became very aberrant when $n = 15$. For all models, sampling distribution means were grossly inflated (negatively for *V*I and *V*O, and positively for L_z) at the short test length, with the degree of inflation being somewhat lower in high (relative to low) discrimination and small (relative to large) N conditions. When $n = 75$ the means were still somewhat, but much less, inflated. The deflationary effects of ξ estimation error likely led to deflated SDs when $n = 75$; this effect was most marked for the 2PL and 3PL. SDs were inflated when $n = 15$ and $N = 1{,}500$, where the inflationary effect of θ estimation error likely was not outweighed by the deflationary effect of ξ estimation error; the degree of inflation was somewhat greater in low than in high discrimination conditions. Finally, when $n = 15$ and $N = 500$, SDs were closer to expectation, especially in low discrimination conditions.

A striking difference between *V*O and the other two z statistics was evident in distributional skewness and kurtosis in high discrimination $n = 75$ conditions. For the 2PL and 3PL in these conditions, *V*O was positively skewed and leptokurtic while L_z and *V*I tended to be somewhat negatively skewed and fairly mesokurtic. These results are not surprising given the tendency for (1) *V*O, but not L_z or *V*I, to be positively skewed and leptokurtic even when their computation is free from parameter estimation error; and (2) *V*O, as opposed to *V*I, to be more unduly influenced by unexpected aberrant responses (Wright et al., 1979).

The distributional aberrations, of course, led to distorted T1 rates. Due to inflated sampling distribution means, T1 rates were inflated in $n = 15$ conditions (ranging from 0.422 to 0.967 at $\alpha = 0.05$). When $n = 75$, 2PL and 3PL T1 rates were deflated (ranging

from < 0.001 to 0.035 at $\alpha = 0.05$) due to the severely restricted sampling distribution SDs (which were not overcompensated for by the mildly inflated means) in these conditions. For the 1PL, a lesser degree of SD restriction coupled with somewhat inflated means led to T1 rates near nominal levels when $n = 75$ and $N = 500$ (ranging from 0.032 to 0.057 at $\alpha = 0.05$) and somewhat inflated when $n = 75$ and $N = 1{,}500$ (ranging from 0.182 to 0.248 at $\alpha = 0.05$).

No studies could be found in which *V*I and *V*O were used with the 2PL or 3PL, but the results of the present study indicate that, relative to the 1PL, sampling distribution variances tend to become even more restricted for these models. The results of other studies involving *V*I and *V*O with the 1PL are only somewhat in agreement with the results found here. For example, Smith (1991) reported (negatively) inflated means for *V*O and *V*I at short test lengths and decreasing degrees of inflation as *n* increased. Wang and Chen (2005) also reported inflations, but to a much smaller degree and with little variation across test length conditions. Some amount of restriction in SDs was also observed in both studies across *N* and *n* conditions, which is somewhat contrary to the results of the present study where SDs were slightly inflated when *n* was small and *N* was large. Also contrary to the results of the present study, was the *degree* of mean inflation, which was much larger (ranging from −0.49 to −3.70) than that reported by Smith (ranging from −0.13 to −1.60) or Wang and Chen (ranging from −0.01 to −0.08) across similar study conditions. However, both other studies generated item response data from *b*-value distributions centered about zero while the *b*-value generating distributions used in this study (e.g. for SU conditions) were centered somewhat below zero (means ranging from −0.17 to −0.26). Smith (1991) showed that the offset between θ and *b* distributions can cause more disruptions to *V*I and *V*O distributions, and the offset in the present study (since the generating θ were centered about zero) could account for some of the differences in results.

Finally, in the only other available study involving L_z, T1 rates near nominal levels were reported (Reise, 1990). However, the results were based on 50-item tests with $N = 1{,}000$. The results of the present study suggest that L_z T1 rates (which were inflated

when n = 15 and deflated when n = 75) might be expected to be near nominal levels at moderate test lengths.

Relationship Between Fit Statistics and Item Parameters

QO was the only fit statistic that was not related to levels of item parameters. Recent study results by Sinharay and Lu (2008) partially support this result. These authors, who examined QO functionality for the 3PL, reported near zero (but positive) correlations between (averaged) QO and b but significant negative correlations between (averaged) QO and a. These correlations were generated within each of several conditions that varied in *level* of discrimination. Supporting a conclusion that QO is not affected by discrimination was the result that QO Type I error rates did not markedly vary across the discrimination conditions.

Q1 was linearly correlated with item parameters, though correlations tended to approach zero when n = 75 (especially when N = 500). Correlations between Q1 and b tended to be positive for the 1PL and 2PL and negative for the 3PL. Q1 was also positively correlated with a parameters for both the 2PL and 3PL, with correlations being stronger in low discrimination conditions than in high discrimination conditions. Dodeen (2004) also reported positive correlations between a parameters and a variant of Q1 (QB) for the 3PL but, unlike the results of this study, reported positive (though small) correlations with b. However, the correlations reported were actually correlations between *averaged* fit statistics and item parameters. On the other hand, Sinharay and Lu (2008), who replicated Dodeen's study conditions, did report negative correlations between b and the likelihood ratio version of QB (GB) when b-values on a test were generated to have higher means than θ; 3PL parameter conditions similar to (but more extreme than) those in the present study.

There was a clear relationship between LM statistics and item b and a parameters. For all models, LM statistics were more inflated for items in the lower half of the b range than for items in the upper half of the range. For the 1PL and 2PL, there also appeared to be pockets of items in the higher b-value range that had inflated LM statistics, though the

degree of inflation was less marked than for items in the lower *b* range. LM statistics were also more inflated at higher (as opposed to lower) *a* for both the 2PL and 3PL.

Finally, when $n = 15$, *z* statistics had a curvilinear relationship with *b*-values; specifically, *z* statistics became more concentrated about zero (which actually was near the tail of the *z* distributions) as *b*-values departed from zero. When $n = 75$, this pattern was much less evident and in most cases items with *z* statistics at the positive and negative extremes had *b* parameters concentrated around zero while items with *z* statistics near zero had *b* parameters spanning the entire *b* range. At both test lengths, the relationship between the *z* statistics and *b* was such that the variance of the *z* statistics became more restricted as *b* moved away from zero – this result has been reported elsewhere for *V*I and *V*O when used with the 1PL (Wang & Chen, 2005). There was also a moderate to strong linear relationship between the *z* statistics and *a* parameters; these relationships were positive for L_z and negative for the other two *z* statistics.

All of the relationships between fit statistics and item parameters were due to the presence of model parameter estimation error. The relationships noted above ceased to exist in $\boldsymbol{\xi}, \boldsymbol{\theta}$ conditions, where $Q1$ and the *z* statistics were uncorrelated with true item parameters.

Conclusions

Ability parameter estimation error was very detrimental to all fit statistics that rely on point estimates of ability. The presence of error in these parameters resulted in grossly inflated sampling distribution means, which in turn led to markedly inflated Type I error rates. Item parameter estimation error tended to deflate the variance of fit statistic sampling distributions, especially for the *z* statistics. These results indicate that when the effects of item parameter error outweigh the effects of ability parameter estimation error (e.g. with small *N* and large *n*), Type I error rates might tend to be deflated.

To date, this is the only study in which the impact of data noise on fit statistic functionality has been assessed. When computed with true model parameters, all fit statistics appeared to be slightly sensitive to the presence of minor multidimensionality. However, when computed with estimated model parameters, as would be done in

practice, the statistics lacked power to detect minor multidimensionality. Thus, the presence of data noise (at least of the kind simulated here) should have little consequence on fit statistic functionality in practice unless fit statistics are free from ability estimation error but contain small amounts of item parameter estimation error; this condition would only apply to *Q*O (and perhaps the LM statistics, though such effects were not examined for them) when used in very large samples (N = 10,000 in the case of this study).

The results of this study indicate that *Q*O appears to be the only fit statistic that can be trusted to adhere to nominal T1 rates across a variety of IRT models and test conditions. The other statistics appear to function adequately under certain circumstances, but all, other than *Q*O, were significantly affected by at least some study conditions. Therefore, the use of these statistics in practice is a form of caveat emptor. For example, *Q*1 might be useful in assessing fit if samples are small (e.g., N = 500) and tests are long (e.g., n = 75), but the limit to its functionality is unknown; the exact combination(s) of sample size and test length in which *Q*1 T1 rates will begin to depart from nominal levels is unknown. LM(α) and LM(β) might also be useful in assessing fit for the 3PL, or for the 1PL and 2PL if discrimination is low (e.g., a averaging < 0.70), but again the precise bounds of its functionality are unknown. Furthermore these statistics might lack power to detect misfit in large samples (e.g., $N \geq$ 1,500) and the question of "how large is too large?" remains unknown. Finally, the z statistics might be useful in assessing fit of items on moderate length tests (e.g., n = 50), but if a test is too short (e.g., $n \leq$ 15) T1 rates likely will be inflated and if it is too long (e.g., $n \geq$ 75) rates likely will be deflated and the statistics might lack power to detect misfit; a precise test length "balance point", at which both T1 deflation and inflation will be staved off, is unknown.

Given the results of this and other studies (Glas & Falcon, 2003; Orlando & Thissen, 2000; Sinharay & Lu, 2008; Stone & Zhang, 2003), it is reasonable to conclude that *Q*O will adhere closely to nominal T1 rates when used with common dichotomous IRT models, short to moderately long tests (e.g., $n \geq$ 10 and $\leq$ 80), small to moderately large samples (e.g., $N \geq$ 500 and $N \leq$ 2,000), and tests with low to moderate discriminating power (e.g., a averaging 0.65 to 1.00). Furthermore, several studies conducted since the completion of the present study have generalized *Q*O for use in

assessing fit of the generalized graded unfolding (GGUM; a polytomous IRT) model (Roberts, 2008) and the two-dimensional compensatory MIRT model (Zhang & Stone, 2008). The results of both studies are promising, with *Q*O Type I error rates generally remaining near nominal levels.

Limitations

The conditions in which fit statistics were examined were limited, so generalizability of the results to markedly different conditions is dubious. Sample size and test length were chosen at only the lower and upper bound of values used in common practice and the distributions from which θ and ξ were generated were intended to be fixed for each IRT model across study conditions. However, due to the manner in which SU generating parameters were obtained, the distributions for these parameters were not completely fixed and varied somewhat across SU conditions. However, these differences were very minor (see Tables B-1 to B-4) and should have had little impact. Nevertheless, θ and ξ distributions were not completely controlled across all SU conditions, and thus could have confounded the results.

The fact that KS test power is related to sample size limited the ability to compare distributional results between *n* conditions for the LM and *z* statistics. Furthermore, the category collapsing used for the *Q* statistics resulted in more difficulties assessing the adherence of these statistics to their theoretical distributions. All analyses were done conditional on *K* for the *Q* statistics but N_K varied greatly across *K*, with $N_K > 10{,}000$ in some groups and $N_K < 50$ in others. This, of course, is also a limitation inherent to these methods. Due to small sample sizes, firm conclusions could thus not be made for *Q* statistic functionality in some instances, especially *Q* statistics that required a large degree of category collapsing. Furthermore, the large sample sizes could have unduly affected KS test results, causing the test to indicate distributional departures for many groups though, in terms of practical significance, the departures were inconsequential.

Additionally, the choice to maintain an equal number of item replications across test length conditions could have affected the results since, in doing so, effects due to items being nested within tests were ignored for some of the analyses. For example, if

there is more variability of fit statistics between than within tests, this could cause less fit statistic variability across the 18,750 item replications in the long, as opposed to short, test length conditions. This could, of course, confound the analysis of fit statistic moments, which was done across item replications. However, equalizing the number of test replications between the two test length conditions would result in markedly different numbers of item replications between the two conditions, which of course would affect the stability of the moments.

Finally, several implementation choices could have confounded the results. First was the use of prior distributions in estimating item and ability parameters. These priors could have introduced some bias in model parameter estimates that might have affected the fit statistics. Secondly, examinees were classified into categories differently for $Q1$ than for QO, so it is unknown to what degree the categorization procedure itself, as opposed to the presence of θ error, accounted for the difference between $Q1$ and QO.

Future Research

In order to better understand the impact of categorization scheme on Q statistic functionality, it might be informative to study a revised $Q1$, where examinees are grouped into categories based on number-correct score, as is done with QO. Perhaps removing the dependence on θ estimates in establishing score categories might reduce $Q1$'s sampling distribution disruptions to some degree.

Because QO appears to be the only fit statistic among those examined here that adhered to expectation in all study conditions and was not correlated with item parameter values, future research should focus on this fit statistic as opposed to the others. The fact that QO remained functional across the study conditions is no guarantee that it will function adequately in other markedly different conditions. Therefore, further study is needed to establish QO's adherence to its theoretical distribution in other circumstances, for example with (1) polytomous IRT models, (2) different θ distributions, (3) extremely short (e.g., $n = 5$) or long (e.g., $n = 200$) tests, and (4) extremely small (e.g., $N = 100$) or large (e.g., $N = 10{,}000$) samples. Due to the deflationary effect of item parameter estimation error, QO might be less able to detect misfit when used with extremely long

tests in small samples; therefore, an examination of its functionality in extreme N and n conditions noted above would be useful. Though these conditions might be used little in practice, they would help establish the bounds of QO's functionality. Additionally, it would be useful to understand the impact of Bayesian model parameter estimation on QO. For example studies could be conducted to determine whether the imposition of item parameter priors during calibration accounts for the deflationary effect of ξ error on QO variances observed in the present study. Furthermore, the impact on QO of the population θ distribution imposed during MML estimation, and the degree of match between this distribution, true θ distribution, and θ distribution used in QO's quadrature approximation, warrants attention.

More power studies are also needed for QO. To date, four such studies have been found concerning dichotomous IRT models (Glas & Falcon, 2003; Orlando & Thissen, 2000, 2003; Stone & Zhang, 2003), one concerning the GGUM model (Roberts, 2008), and one concerning the compensatory MIRT model (Zhang & Stone, 2008). The results of all these studies have been encouraging. However, the conditions under which power has been examined are limited. For example, little is known regarding the impact of model parameter distributions on power and nothing is known regarding power with polytomous IRT models other than the GGUM. There is also evidence that the sampling distribution for QO is distorted for model-fitting items when there is misfit elsewhere in the data matrix (Glas & Falcon, 2003; Orlando & Thissen, 2003). This is a problem and its severity in, and consequences for, item-fit assessment with QO warrant further investigation.

Finally, in the broad scope of fit research, little attempt has been made to assess the practical implications of model-data misfit. Model fit is a continuum rather than a dichotomy. Due to the probabilistic nature of the models and errors of measurement, "perfect" fit is essentially unattainable. Even if examinees respond by processes exactly like the models used to parameterize the data, random error will always make it possible for seemingly "aberrant" responses to occur. Thus, one important question that warrants attention is how much deviation from the calibrating model is necessary before the use of the model (utility and/or validity) is compromised. This question reaches topics such as

impact of model-data misfit on accuracy of linking and equating functions, ability estimates, and decisions based on ability estimates. Such an understanding would also aid researchers in interpreting the results of fit statistic power studies. Although it is informative to compare fit statistics regarding their relative power in the search for a uniformly most powerful test, the question of how much power is *realistically* (as opposed to idealistically) needed remains open. This question could be answered only if there is some understanding of the degree to which the presence of misfitting items on tests is admissible in practice.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113-127.

Baker, F. B., & Kim, S. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Balla, J. R., & McDonald, R. P. (1985). Latent trait item analysis and facet theory: A useful combination. *Applied Psychological Measurement, 9*, 191-198.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2^n tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series 97-4). Iowa City, IA: The American College Testing Program.

De Ayala, R. J. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

DeChamplain, A. (1996, April). *Assessing the dimensionality of item response matrices using a goodness-of-fit index based on noncentrality.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

DeGroot, M. H. (1986). *Probability and statistics* (2nd ed.). Reading, MA: Addison-Wesley.

DeMars, C. E. (2004). Type I error rates for generalized graded unfolding model fit indices. *Applied Psychological Measurement, 28*, 48-71.

DeMars, C. E. (2005). Type I errors for PARSCALE's fit index. *Educational and Psychological Measurement, 65*, 42-50.

Divigi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*, 283-298.

Dodeen, H. (2004). The relationship between item parameters and item fit. *Journal of Educational Measurement, 41*, 261-270.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Glas, C. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*, 273-294.

Glas, C., & Falcon, J. (2003). A comparison of item-fit indices for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell: Kluwer -Nijhoff.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55-64.

McDonald, R. P., & Mok, M. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

McKinley, R. L., & Reckase, M. D. (1983). *Extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR 83-2). Iowa City, IA: The American College Testing Program.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75-106

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28*, 99-117.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19,* 121-129.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S - X^2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing [Computer Software]. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14*, 127-137.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19,* 213-229.

Ro, S. (2001). Characteristics of a likelihood-based person-fit index under the graded response model. (Doctoral dissertation, University of Minnesota, 2001). *Dissertation Abstracts International, 62*, B4834.

Roberts, J. S. (2003, April). *An item fit statistic based on pseudocounts from the generalized graded unfolding model: A preliminary report.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement, 32*, 407-423.

Rogers, H., & Hattie, J. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

Seo, D., & Weiss, D.J. (2009). *Detection of non-fitting person response patterns with achievement test data.* Manuscript submitted for publication.

Sinharay, S. (2005). Assessing fit in unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375-394.
Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298-321.
Sinharay, S., & Lu, Y. (2008). A further look at the correlation between item parameters and item fit statistics. *Journal of Educational Measurement, 45*, 1-15.
Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement, 51*, 541-565.
Smith, R. M. (1994). A comparison of the power of Rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement, 54*, 42-55.
Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameters. *Psychometrika, 66*, 331-342.
Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 31*, 58-75.
Stone, C. A. (2003). Empirical power and type I error rates for an IRT fit statistic that considers the precision of ability estimates. *Educational and Psychological Measurement, 63*, 566-583.
Stone, C.A., Ankenmann, R.D., Lane, S., & Liu, M. (1993, April). *Scaling QUASAR's performance assessment.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.
Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement, 60*, 974-991.
Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40*, 331-352.
Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293-325.
Tang, H. (1994, April). *Step fit analysis with polytomously scored items.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.
Van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement*, *23*, 327-345.
Wang, W., & Chen, C. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement, 65*, 376-404.
Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Wright, B. D., & Mead, R. J. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum 23). Chicago: University of Chicago, Statistical Laboratory, Department of Education.

Wright, B. D., Mead, R. J., & Bell, S. R. (1979). *BICAL: Calibrating items with the Rasch model* (Research Memorandum 23-B). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-146.

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

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement, 68*, 181-196.

GLOSSARY OF NOTATION AND ACRONYMS

Fit Statistics

$Q1$	Yen's (1981) chi-square index
QB	Bock's (1972) chi-square index
QO	Orlando and Thissen's (2000) chi-square index
LM($\alpha\beta$)	Lagrange multiplier fit index (Glas & Falcon, 1999) for item discrimination and difficulty
LM(α)	Lagrange multiplier fit index for item discrimination
LM(β)	Lagrange multiplier fit index for item difficulty
z statistics	
L_z	Reise's (1990) item fit index
VI	Infit
VO	Outfit

Other Statistics/Tests

KS	Kolmogorov-Smirnov test
ME	Mean error
T1	Type I error

Study Factors

FS	Fit statistic
PE	Parameter estimation error
DN	Data noise
SU	Strictly unidimensional
EU	Essentially unidimensional
M	IRT model
D	Item discrimination
N	Sample size
n	Test length

Notation

$\boldsymbol{\theta}$	Vector of true (generating) ability parameters
θ	Ability parameters
$\hat{\boldsymbol{\theta}}$	Vector of estimated ability parameters
$\boldsymbol{\xi}$	Vector of true (generating) item parameters
ξ	Item parameters
$\hat{\boldsymbol{\xi}}$	Vector of estimated item parameters
K	Number of score categories (applies to $Q1$, QO, and LM statistics)
π_K	Proportion of KS test rejections across K (applies to $Q1$ and QO)
π_R	Proportion of KS test rejections across replicated tests (applies to LM and the z statistics)
D	KS test statistic

APPENDIX A: R CODE

Code to Generate EU Data and Set up Files for Parameter Estimation

Code to Generate SU Data and Set up Files for Parameter Estimation

Code to Compute $Q1$

Code to Compute QO

Functions Used by $Q1$ and QO

Code to Compute LM Statistics

Code to Compute z Statistics

Code to Generate EU Data and Set up Files for Parameter Estimation

```
#LOAD BOOT PACKAGE
#N = sample size; n = test length; R = number of replications
#start = the replication (R-1) at which to start; model = IRT
model
#N_c = sample size for large-sample calibration of EU data
#a_delt = discrimination scaling factor for low disc condition

EU_Data <-
function(N,N_c,n,R,start,model,a_delt,a_m,a_sd,a_m_lo,a_sd_lo,b_m
,b_sd,c_m,c_sd) {

theta_stats <- array(rep(0, times=10*R), dim=c(R,10))
a_stats <- array(rep(0, times=9*R), dim=c(R,9))
b_stats <- array(rep(0, times=9*R), dim=c(R,9))
c_stats <- array(rep(0, times=4*R), dim=c(R,4))
x <- array(rep(0, times=R), dim=c(R))

############Set up for mlg file############
mlg_l15 <- "1"
for (i in 1:(n-1)) {
mlg_l15 <- paste(mlg_l15,"1",sep="")}
Fix_1 <- array(1:n, dim=c(n))

############Set up for bat file############
path_bat <- "C:/Program Files/multilog"
lines_bat <- array(rep("", times=R+1),dim=c(R+1))
lines_bat_lo <- array(rep("", times=R),dim=c(R))
lines_bat[1] <- paste("path",path_bat,sep=" ")
lines_bat_c <- array(rep("", times=R+1),dim=c(R+1))
lines_bat_c[1] <- paste("path",path_bat,sep=" ")
lines_bat_c_lo <- array(rep("", times=R+1),dim=c(R+1))
lines_bat_c_lo[1] <- paste("path",path_bat,sep=" ")
lines_bat_sc <- array(rep("", times=R+1),dim=c(R+1))
lines_bat_sc[1] <- paste("path",path_bat,sep=" ")
lines_bat_sc_lo <- array(rep("", times=R),dim=c(R))

############SET UP PARMS############
##Correlations between parms
r_a <- -0.29
r_b <- -0.10

##Rescale 3PL mean parms for 2PL
alpha1 <- -0.05104711
alpha2 <- (0.49/1.03)*alpha1
delta <- -0.37915156
```

```
if (model == 1 | model == 2) m_a1 <- 1.03+alpha1 else m_a1 <-
1.03
if (model == 1) sd_a1 <-0 else sd_a1 <- 0.30
if (model == 1 | model == 2) m_a2 <- 0.49+alpha2 else m_a2 <-
0.49
if (model == 1) sd_a2 <-0 else sd_a2 <- 0.10
if (model == 1 | model == 2) m_b1 <- 0.30+delta else m_b1 <- 0.30
sd_b1 <- 0.82
if (model == 1 | model == 2) m_b2 <- -0.03+delta else m_b2 <- -
0.03
sd_b2 <- 0.82

b1_a <- r_a*(sd_a2/sd_a1)
b0_a <- m_a2 - m_a1*b1_a
b1_b <- r_b*(sd_b2/sd_b1)
b0_b <- m_b2 - m_b1*b1_b

var_a2_pred <- (b1_a^2)*(sd_a1^2)
var_b2_pred <- (b1_b^2)*(sd_b1^2)

er_a2_sd <- sqrt(sd_a2^2 - var_a2_pred)
er_b2_sd <- sqrt(sd_b2^2 - var_b2_pred)

if (model == 3) h <- 1 else h <- 0

############LOOPS THROUGH REPLICATIONS############
for (f in 1:R) {
theta_1 <- rnorm(n=N,mean=0,sd=1)
theta_2 <- rnorm(n=N,mean=0,sd=1)
theta_1_lo <- rnorm(n=N,mean=0,sd=1)
theta_2_lo <- rnorm(n=N,mean=0,sd=1)
theta_c1 <- rnorm(n=N_c,mean=0,sd=1)
theta_c2 <- rnorm(n=N_c,mean=0,sd=1)
theta_c1_lo <- rnorm(n=N_c,mean=0,sd=1)
theta_c2_lo <- rnorm(n=N_c,mean=0,sd=1)
B_1 <- rnorm(n=n,mean=m_b1,sd=sd_b1)
B_2_er <- rnorm(n=n,mean=0,sd=er_b2_sd)
B_2 <- b0_b + b1_b*B_1 + B_2_er
C_p <- rlnorm(n=n,meanlog=log(0.2),sdlog=0.15)
##To sample C so that it will be equivalent in SU and EU
conditions since MLG rounds C_z
C_z <- round(log(C_p/(1-C_p)),2)
C <- 1/(1+exp(-C_z))
C <- C*h

if (model == 1) {
A_1 <- rnorm(n=n,mean=m_a1,sd=sd_a1)
A_2 <- rnorm(n=n,mean=m_a2,sd=sd_a2)}
else {
```

```
a_out <- 1
xx <- 0
while (a_out > 0) {
A_1 <- rnorm(n=n,mean=m_a1,sd=sd_a1)
A_2_er <- rnorm(n=n,mean=0,sd=er_a2_sd)
A_2 <- b0_a + b1_a*A_1 + A_2_er
xx <- xx + 1
##CHECK A PARMS
diff <- A_2 - A_1
ind1 <- A_1 < 0
ind2 <- A_2 < 0
#ind2 <- diff > 0
#ind3 <- abs(diff) <= 0.02
a_out <- sum(ind1,ind2)}
x[f] <- xx}

ind <- array(1:N, dim=c(N))
sp1 <- array(c(rep("    ", times=9),rep("   ", times=90),rep("
", times=900),rep(" ", times=9000),rep("", times=1)),
dim=c(10000))
sp <- sp1[1:N]

ind_c <- array(1:N_c, dim=c(N_c))
sp_c <- sp1[1:N_c]

##GENERATE DATA TO USE IN EU CONDITIONS
##HIGHER a
e1 <- array(runif(n=N*n,min=0,max=1),dim=c(n,N))
theta_1 <- t(theta_1)
theta_2 <- t(theta_2)
logit <- -1.7*(A_1%*%theta_1-A_1*B_1 + A_2%*%theta_2-A_2*B_2)
p <- C + (1 - C)/(1+exp(logit))
er <- p - e1
X <- t(er >= 0)*1
p <- t(p)
theta_1 <- t(theta_1)
theta_2 <- t(theta_2)

##LOWER a
e1 <- array(runif(n=N*n,min=0,max=1),dim=c(n,N))
theta_1_lo <- t(theta_1_lo)
theta_2_lo <- t(theta_2_lo)
logit_lo <- -1.7*a_delt*(A_1%*%theta_1_lo-A_1*B_1 +
A_2%*%theta_2_lo-A_2*B_2)
p_lo <- C + (1 - C)/(1+exp(logit_lo))
er <- p_lo - e1
X_lo <- t(er >= 0)*1
p_lo <- t(p_lo)
theta_1_lo <- t(theta_1_lo)
```

```
theta_2_lo <- t(theta_2_lo)

##GENERATE DATA TO USE IN CALIBRATING SU DATA
##HIGHER a
e1_c <- array(runif(n=N_c*n,min=0,max=1),dim=c(n,N_c))
theta_c1 <- t(theta_c1)
theta_c2 <- t(theta_c2)
logit_c <- -1.7*(A_1%*%theta_c1-A_1*B_1 + A_2%*%theta_c2-A_2*B_2)
p_c <- C + (1 - C)/(1+exp(logit_c))
er_c <- p_c - e1_c
X_c <- t(er_c >= 0)*1
p_c <- t(p_c)
theta_c1 <- t(theta_c1)
theta_c2 <- t(theta_c2)

##LOWER a
e1_c <- array(runif(n=N_c*n,min=0,max=1),dim=c(n,N_c))
theta_c1_lo <- t(theta_c1_lo)
theta_c2_lo <- t(theta_c2_lo)
logit_c_lo <- -1.7*a_delt*(A_1%*%theta_c1_lo-A_1*B_1 +
A_2%*%theta_c2_lo-A_2*B_2)
p_c_lo <- C + (1 - C)/(1+exp(logit_c_lo))
er_c <- p_c_lo - e1_c
X_c_lo <- t(er_c >= 0)*1
p_c_lo <- t(p_c_lo)
theta_c1_lo <- t(theta_c1_lo)
theta_c2_lo <- t(theta_c2_lo)

#SET UP LOG ODS FOR USE IN OBTAINING SU THETA
p_star <- 1/(1+exp(logit))
p_star_lo <- 1/(1+exp(logit_lo))
p_c_star <- 1/(1+exp(logit_c))
p_c_star_lo <- 1/(1+exp(logit_c_lo))

ln_ods <- log(p_star*(1/(1-p_star)))
sum_ln_ods <- t(t(apply(ln_ods,2,sum)))
ln_ods_lo <- log(p_star_lo*(1/(1-p_star_lo)))
sum_ln_ods_lo <- t(t(apply(ln_ods_lo,2,sum)))

ln_ods_c <- log(p_c_star*(1/(1-p_c_star)))
sum_ln_ods_c <- t(t(apply(ln_ods_c,2,sum)))
ln_ods_c_lo <- log(p_c_star_lo*(1/(1-p_c_star_lo)))
sum_ln_ods_c_lo <- t(t(apply(ln_ods_c_lo,2,sum)))

##SET UP FOR OUTPUT FILES
id <- paste(ind,"     ", sp,sep="")
data <- cbind(id,X)
data_lo <- cbind(id,X_lo)
id_c <- paste(ind_c,"     ", sp_c,sep="")
```

```
data_c <- cbind(id_c,X_c)
data_c_lo <- cbind(id_c,X_c_lo)

theta <- cbind(theta_1,theta_2)
theta_lo <- cbind(theta_1_lo,theta_2_lo)
theta_c <- cbind(theta_c1,theta_c2)
theta_c_lo <- cbind(theta_c1_lo,theta_c2_lo)
aparms <- cbind(A_1,A_2)
bparms <- cbind(B_1,B_2)
parms <- cbind(A_1,A_2,B_1,B_2,C)

##GET SUMMARY STATS FOR PARMS
theta_stats[f,1] <- corr(theta)
theta_stats[f,2] <- mean(theta_1)
theta_stats[f,3] <- mean(theta_2)
theta_stats[f,4] <- sd(theta_1)
theta_stats[f,5] <- sd(theta_2)
theta_stats[f,6] <- corr(theta_lo)
theta_stats[f,7] <- mean(theta_1_lo)
theta_stats[f,8] <- mean(theta_2_lo)
theta_stats[f,9] <- sd(theta_1_lo)
theta_stats[f,10] <- sd(theta_2_lo)
a_stats[f,1] <- corr(aparms)
a_stats[f,2] <- mean(A_1)
a_stats[f,3] <- mean(A_2)
a_stats[f,4] <- sd(A_1)
a_stats[f,5] <- sd(A_2)
a_stats[f,6] <- min(A_1)
a_stats[f,7] <- min(A_2)
a_stats[f,8] <- max(A_1)
a_stats[f,9] <- max(A_2)
b_stats[f,1] <- corr(bparms)
b_stats[f,2] <- mean(B_1)
b_stats[f,3] <- mean(B_2)
b_stats[f,4] <- sd(B_1)
b_stats[f,5] <- sd(B_2)
b_stats[f,6] <- min(B_1)
b_stats[f,7] <- min(B_2)
b_stats[f,8] <- max(B_1)
b_stats[f,9] <- max(B_2)
c_stats[f,1] <- mean(C)
c_stats[f,2] <- sd(C)
c_stats[f,3] <- min(C)
c_stats[f,4] <- max(C)

##WRITE PARMS TO FILE
outfile <- paste("EU_",model,"PL_N-",N,"_n-
",n,"_",start+f,sep="")
```

```
write.table(theta,file=paste(outfile,".2d.the",sep=""),quote=FALS
E,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(theta_lo,file=paste(outfile,".lo.2d.the",sep=""),quot
e=FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(sum_ln_ods,file=paste(outfile,".lods",sep=""),quote=F
ALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(sum_ln_ods_lo,file=paste(outfile,".lo.lods",sep=""),q
uote=FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(sum_ln_ods_c,file=paste(outfile,".lods.c",sep=""),quo
te=FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(sum_ln_ods_c_lo,file=paste(outfile,".lo.lods.c",sep="
"),quote=FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(parms,file=paste(outfile,".ksi",sep=""),quote=FALSE,r
ow.names=FALSE,col.names=FALSE,sep=" ")
write.table(data,file=paste(outfile,".wgr",sep=""),quote=FALSE,ro
w.names=FALSE,col.names=FALSE,sep="")
write.table(data_lo,file=paste(outfile,".lo.wgr",sep=""),quote=FA
LSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(data_c,file=paste(outfile,"_c.wgr",sep=""),quote=FALS
E,row.names=FALSE,col.names=FALSE,sep="")
write.table(data_c_lo,file=paste(outfile,"_c.lo.wgr",sep=""),quot
e=FALSE,row.names=FALSE,col.names=FALSE,sep="")

##MAKE MLG ITEM PARAMETER ESTIMATION FILE
path <- "C:/dissertation/Simulation/Test/"
lines <- array(rep("", times=19),dim=c(19))
lines[1] <- "MULTILOG syntax generated by R"
lines[2] <- date()
lines[3] <- ">PROBLEM RAMDOM,"
lines[4] <- "         INDIVIDUAL,"
lines[5] <- paste("         NITEM=",n,",",sep="")
lines[6] <- paste("         NEXAMINEES=",N,",",sep="")
lines[7] <- "         NGROUP=1,"
lines[8] <- paste("         DATA='",path,outfile,".wgr',",sep="")
lines[9] <- "         NCHARS=8;"
lines[10] <- paste(">TES AL, L",model,";",sep="")
if (model == 2 | model == 3) {
lines[11] <- paste(">PRIORS ALL, AJ,
PARAMS=(",a_m,",",a_sd,");",sep="")}
if (model == 1) {
lines[11] <- paste(">FIX ALL, AJ, VALUE=",a_m,";",sep="")}
lines[12] <- paste(">PRIORS ALL, BJ,
PARAMS=(",b_m,",",b_sd,");",sep="")
lines[13] <- ">SAVE;"
lines[14] <- ">END;"
lines[15] <- 2
lines[16] <- "01"
lines[17] <- mlg_l15
lines[18] <- "N"
```

```
lines[19] <- paste("(8A1,2X,",n,"A1)",sep="")

if (model == 3) {
lines_1_12 <- t(t(lines[1:12]))
lines_13_19 <- t(t(lines[13:19]))
Prior_c <- paste(">PRIORS ALL, DK=1,
PARAMS=(",c_m,",",c_sd,");",sep="")
lines <- rbind(lines_1_12,Prior_c,lines_13_19)}

lines_lo <- lines
lines_lo[8] <- paste("
DATA='",path,outfile,".lo.wgr',",sep="")
if (model == 2 | model == 3) {
lines_lo[11] <- paste(">PRIORS ALL, AJ,
PARAMS=(",a_m_lo,",",a_sd_lo,");",sep="")}
if (model ==1) {
lines_lo[11] <- paste(">FIX ALL, AJ, VALUE=",a_m_lo,";",sep="")}

lines_c <- lines
lines_c[6] <- paste("         NEXAMINEES=",N_c,",",sep="")
lines_c[8] <- paste("
DATA='",path,outfile,"_c.wgr',",sep="")

if (model == 3) {
Fix <- t(t(paste(">FIX IT=",Fix_1,"CJ
VALUE=",t(t(C)),";",sp="")))
lines_c_1_12 <- t(t(lines_c[1:12]))
lines_c_14_20 <- t(t(lines_c[14:20]))
lines_c <- rbind(lines_c_1_12,Fix,lines_c_14_20)}

lines_c_lo <- lines_c
lines_c_lo[8] <- paste("
DATA='",path,outfile,"_c.lo.wgr',",sep="")
if (model == 2 | model == 3) {
lines_c_lo[11] <- paste(">PRIORS ALL, AJ,
PARAMS=(",a_m_lo,",",a_sd_lo,");",sep="")}
if (model == 1) {
lines_c_lo[11] <- paste(">FIX ALL, AJ,
VALUE=",a_m_lo,";",sep="")}

write.table(lines,file=paste(outfile,".mlg",sep=""),quote=FALSE,r
ow.names=FALSE,col.names=FALSE,sep="")
write.table(lines_lo,file=paste(outfile,".lo.mlg",sep=""),quote=F
ALSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_c,file=paste(outfile,"_c.mlg",sep=""),quote=FAL
SE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_c_lo,file=paste(outfile,"_c.lo.mlg",sep=""),quo
te=FALSE,row.names=FALSE,col.names=FALSE,sep="")
```

```
##MAKE MLG THETA ESTIMATION FILE
lines_sc <- lines
lines_sc[3] <- ">PROBLEM SCORE,"
start_par1 <- ">START ALL,"
start_par2 <- paste("PARAM='",path,outfile,".PAR';",sep="")
lines_1_10 <- t(t(lines_sc[1:10]))
if (model == 3) {
lines_13_19 <- t(t(lines_sc[14:20]))}
else {lines_13_19 <- t(t(lines_sc[13:19]))}
lines_sc <- rbind(lines_1_10,start_par1,start_par2,lines_13_19)

lines_sc_lo <- lines_lo
lines_sc_lo[3] <- ">PROBLEM SCORE,"
start_par1 <- ">START ALL,"
start_par2 <- paste("PARAM='",path,outfile,".lo.PAR';",sep="")
lines_1_10 <- t(t(lines_sc_lo[1:10]))
if (model == 3) {
lines_13_19 <- t(t(lines_sc_lo[14:20]))}
else {lines_13_19 <- t(t(lines_sc_lo[13:19]))}
lines_sc_lo <-
rbind(lines_1_10,start_par1,start_par2,lines_13_19)

write.table(lines_sc,file=paste(outfile,"_sc.mlg",sep=""),quote=F
ALSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_sc_lo,file=paste(outfile,"_sc.lo.mlg",sep=""),q
uote=FALSE,row.names=FALSE,col.names=FALSE,sep="")

##MAKE BAT FILE
lines_bat[f+1] <- paste("mlg",outfile,sep=" ")
lines_bat_lo[f] <- paste("mlg ",outfile,".lo",sep="")
lines_bat_c[f+1] <- paste("mlg ",outfile,"_c",sep="")
lines_bat_c_lo[f+1] <- paste("mlg ",outfile,"_c.lo",sep="")
lines_bat_sc[f+1] <- paste("mlg ",outfile,"_sc",sep="")
lines_bat_sc_lo[f] <- paste("mlg ",outfile,"_sc.lo",sep="")}

############END OF LOOP THROUGH REPLICATIONS############
outfile0 <- paste("EU_",model,"PL_N-",N,"_n-",n,"_start-
",start,sep="")

##WRITE STATS TO FILES
write.table(theta_stats,file=paste(outfile0,".t.stat",sep=""),quo
te=FALSE,row.names=FALSE,
col.names=c("r_t1t2","mean_t1","mean_t2","sd_t1","sd_t2","r_t1t2_
lo","mean_t1_lo","mean_t2_lo","sd_t1_lo","sd_t2_lo"),
sep=" ")
write.table(a_stats,file=paste(outfile0,".a.stat",sep=""),quote=F
ALSE,row.names=FALSE,
col.names=c("r_a1a2","mean_a1","mean_a2","sd_a1","sd_a2","min_a1"
,"min_a2","max_a1","max_a2"),
```

```
sep=" ")
write.table(b_stats,file=paste(outfile0,".b.stat",sep=""),quote=F
ALSE,row.names=FALSE,
col.names=c("r_b1b2","mean_b1","mean_b2","sd_b1","sd_b2","min_b1"
,"min_b2","max_b1","max_b2"),
sep=" ")
write.table(c_stats,file=paste(outfile0,".c.stat",sep=""),quote=F
ALSE,row.names=FALSE,
col.names=c("mean","sd","min","max"),
sep=" ")

##WRITE BAT TO FILE
lines_bat <- rbind(t(t(lines_bat)),t(t(lines_bat_lo)))
lines_bat_sc <- rbind(t(t(lines_bat_sc)),t(t(lines_bat_sc_lo)))
write.table(lines_bat,file=paste(outfile0,".bat",sep=""),quote=FA
LSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_bat_c,file=paste(outfile0,"_c.bat",sep=""),quot
e=FALSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_bat_c_lo,file=paste(outfile0,"_c.lo.bat",sep=""
),quote=FALSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_bat_sc,file=paste(outfile0,"_sc.bat",sep=""),qu
ote=FALSE,row.names=FALSE,col.names=FALSE,sep="")
end_time <<-date()
}
```

Code to Generate SU Data and Set up Files for Parameter Estimation

```
SU_Data <-
function(N,n,R,start,model,hilo,a_m,a_sd,b_m,b_sd,c_m,c_sd) {
theta_stats <- array(rep(0, times=5*R), dim=c(R,5))
theta_g_stats <- array(rep(0, times=5*R), dim=c(R,5))
a_stats <- array(rep(0, times=5*R), dim=c(R,5))
b_stats <- array(rep(0, times=5*R), dim=c(R,5))
c_stats <- array(rep(0, times=4*R), dim=c(R,4))

############Set up for mlg file############
mlg_l15 <- "1"
for (i in 1:(n-1)) {
mlg_l15 <- paste(mlg_l15,"1",sep="")}

############Set up for bat file############
path_bat <- "C:/Program Files/multilog"
lines_bat <- array(rep("", times=R+1),dim=c(R+1))
lines_bat[1] <- paste("path",path_bat,sep=" ")
lines_bat_sc <- array(rep("", times=R+1),dim=c(R+1))
lines_bat_sc[1] <- paste("path",path_bat,sep=" ")

############LOOPS THROUGH REPLICATIONS############
for (f in 1:R) {
infile <- paste("EU_",model,"PL_N-",N,"_n-",n,"_",start+f,sep="")

if (model == 3) {
inp_ksi_e <- scan(paste(infile,"_c",hilo,".par",sep=""),
list(0,0,0,0))
ksi_e <-
cbind(inp_ksi_e[[1]][1:n],inp_ksi_e[[2]][1:n],inp_ksi_e[[3]][1:n]
)
ksi_e[,2] <- -(ksi_e[,2]/ksi_e[,1])
ksi_e[,1] <- ksi_e[,1]/1.7
ksi_e[,3] <- 1/(1 + exp(-ksi_e[,3]))
}

if (model == 2 | model == 1) {
inp_ksi_e <- scan(paste(infile,"_c",hilo,".par",sep=""),
list(0,0))
ksi_e_3 <- array(rep(0, times=n),dim=c(n))
ksi_e <- cbind(inp_ksi_e[[1]][1:n],inp_ksi_e[[2]][1:n],ksi_e_3)
ksi_e[,1] <- ksi_e[,1]/1.7}

##FIND SU THETAS
inp_lods <- scan(paste(infile,hilo,".lods.c",sep=""), list(0))
lods <- inp_lods[[1]]
sum_ab <- sum(1.7*ksi_e[,1]*ksi_e[,2])
sum_a <- sum(1.7*ksi_e[,1])
```

```
theta <- (lods + sum_ab)/sum_a
theta_1 <- theta[1:N]

##FIND SU thetas to use for the EU true parameter conditions
inp_lods_EU <- scan(paste(infile,hilo,".lods",sep=""), list(0))
lods_EU <- inp_lods_EU[[1]]
theta_EU <- (lods_EU + sum_ab)/sum_a

##MAKE RESPONSE DATA
p <- array(rep(0, times=N*n), dim=c(N,n))
X <- array(rep(0, times=N*n), dim=c(N,n))
ind <- array(1:N, dim=c(N))
sp <- array(c(rep("    ", times=9),rep("   ", times=90),rep("  ",
times=900),rep(" ", times=9000)), dim=c(9999))
sp <- sp[1:N]

for (i in 1:n) {
e1 <- runif(n=N,min=0,max=1)
logit <- -1.7*(ksi_e[i,1]*(theta_1-ksi_e[i,2]))
p[,i] <- ksi_e[i,3] + (1 - ksi_e[i,3])/(1+exp(logit))
er <- p[,i] - e1
X[,i] <- er >= 0}

id <- paste(ind,"     ", sp,sep="")
data <- cbind(id,X)

##GET SUMMARY STATS FOR PARMS
theta_stats[f,1] <- mean(theta_1)
theta_stats[f,2] <- sd(theta_1)
theta_stats[f,3] <- median(theta_1)
theta_stats[f,4] <- min(theta_1)
theta_stats[f,5] <- max(theta_1)
theta_g_stats[f,1] <- mean(theta)
theta_g_stats[f,2] <- sd(theta)
theta_g_stats[f,3] <- median(theta)
theta_g_stats[f,4] <- min(theta)
theta_g_stats[f,5] <- max(theta)
a_stats[f,1] <- mean(ksi_e[,1])
a_stats[f,2] <- sd(ksi_e[,1])
a_stats[f,3] <- median(ksi_e[,1])
a_stats[f,4] <- min(ksi_e[,1])
a_stats[f,5] <- max(ksi_e[,1])
b_stats[f,1] <- mean(ksi_e[,2])
b_stats[f,2] <- sd(ksi_e[,2])
b_stats[f,3] <- median(ksi_e[,2])
b_stats[f,4] <- min(ksi_e[,2])
b_stats[f,5] <- max(ksi_e[,2])
c_stats[f,1] <- mean(ksi_e[,3])
c_stats[f,2] <- sd(ksi_e[,3])
```

```
c_stats[f,3] <- min(ksi_e[,3])
c_stats[f,4] <- max(ksi_e[,3])

##WRITE PARMS TO FILE
outfile <- paste("SU_",model,"PL_N-",N,"_n-
",n,"_",start+f,sep="")
write.table(theta,file=paste(outfile,hilo,".g.the",sep=""),quote=
FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(theta_1,file=paste(outfile,hilo,".the",sep=""),quote=
FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(theta_EU,file=paste(infile,hilo,".the",sep=""),quote=
FALSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(ksi_e,file=paste(outfile,hilo,".ksi",sep=""),quote=FA
LSE,row.names=FALSE,col.names=FALSE,sep=" ")
write.table(data,file=paste(outfile,hilo,".wgr",sep=""),quote=FAL
SE,row.names=FALSE,col.names=FALSE,sep="")

##MAKE MLG ITEM PARAMETER ESTIMATION FILE
path <- "C:/dissertation/Simulation/Test/"
lines <- array(rep("", times=19),dim=c(19))
lines[1] <- "MULTILOG syntax generated by R"
lines[2] <- date()
lines[3] <- ">PROBLEM RAMDOM,"
lines[4] <- "         INDIVIDUAL,"
lines[5] <- paste("         NITEM=",n,",",sep="")
lines[6] <- paste("         NEXAMINEES=",N,",",sep="")
lines[7] <- "         NGROUP=1,"
lines[8] <- paste("
DATA='",path,outfile,hilo,".wgr',",sep="")
lines[9] <- "         NCHARS=8;"
lines[10] <- paste(">TES AL, L",model,";",sep="")
if (model == 2 | model == 3) {
lines[11] <- paste(">PRIORS ALL, AJ,
PARAMS=(",a_m,",",a_sd,");",sep="")}
if (model == 1) {
lines[11] <- paste(">FIX ALL, AJ, VALUE=",a_m,";",sep="")}
lines[12] <- paste(">PRIORS ALL, BJ,
PARAMS=(",b_m,",",b_sd,");",sep="")
lines[13] <- ">SAVE;"
lines[14] <- ">END;"
lines[15] <- 2
lines[16] <- "01"
lines[17] <- mlg_l15
lines[18] <- "N"
lines[19] <- paste("(8A1,2X,",n,"A1)",sep="")

if (model == 3) {
lines_1_12 <- t(t(lines[1:12]))
lines_13_19 <- t(t(lines[13:19]))
```

```
Prior_c <- paste(">PRIORS ALL, DK=1,
PARAMS=(",c_m,",",c_sd,");",sep="")
lines <- rbind(lines_1_12,Prior_c,lines_13_19)
}

write.table(lines,file=paste(outfile,hilo,".mlg",sep=""),quote=FA
LSE,row.names=FALSE,col.names=FALSE,sep="")

##MAKE MLG THETA ESTIMATION FILE
lines_sc <- lines
lines_sc[3] <- ">PROBLEM SCORE,"
start_par1 <- ">START ALL,"
start_par2 <- paste("PARAM='",path,outfile,hilo,".PAR';",sep="")
lines_1_10 <- t(t(lines_sc[1:10]))
if (model == 3) {
lines_13_19 <- t(t(lines_sc[14:20]))}
else {lines_13_19 <- t(t(lines_sc[13:19]))}
lines_sc <- rbind(lines_1_10,start_par1,start_par2,lines_13_19)

write.table(lines_sc,file=paste(outfile,"_sc",hilo,".mlg",sep="")
,quote=FALSE,row.names=FALSE,col.names=FALSE,sep="")

##MAKE BAT FILE
lines_bat[f+1] <- paste("mlg ",outfile,hilo,sep="")
lines_bat_sc[f+1] <- paste("mlg ",outfile,"_sc",hilo,sep="")
}

############END OF LOOP THROUGH REPLICATIONS############
outfile0 <- paste("SU_",model,"PL_N-",N,"_n-",n,"_start-
",start,sep="")

##WRITE STATS TO FILES
write.table(theta_stats,file=paste(outfile0,hilo,".t.stat",sep=""
),quote=FALSE,row.names=FALSE,
col.names=c("mean_t1","sd_t1","median_t1","min_t1","max_t1"),
sep=" ")
write.table(theta_g_stats,file=paste(outfile0,hilo,".t.g.stat",se
p=""),quote=FALSE,row.names=FALSE,
col.names=FALSE,
sep=" ")
write.table(a_stats,file=paste(outfile0,hilo,".a.stat",sep=""),qu
ote=FALSE,row.names=FALSE,
col.names=c("mean_a1","sd_a1","median_a1","min_a1","max_a1"),
sep=" ")
write.table(b_stats,file=paste(outfile0,hilo,".b.stat",sep=""),qu
ote=FALSE,row.names=FALSE,
col.names=c("mean_b1","sd_b1","median_b1","min_b1","max_b1"),
sep=" ")
```

```
write.table(c_stats,file=paste(outfile0,hilo,".c.stat",sep=""),qu
ote=FALSE,row.names=FALSE,
col.names=c("mean","sd","min","max"),
sep=" ")

##WRITE BAT TO FILE
write.table(lines_bat,file=paste(outfile0,hilo,".bat",sep=""),quo
te=FALSE,row.names=FALSE,col.names=FALSE,sep="")
write.table(lines_bat_sc,file=paste(outfile0,"_sc",hilo,".bat",se
p=""),quote=FALSE,row.names=FALSE,col.names=FALSE,sep="")
}
```

Code to Compute *Q*1

```
#Input discrimination parameters are on the normal metric
#k = number of score groups
#Requires a N*n matrix of item responses, u
#Uses the following functions: cat_theta, collapse_cells, and
calc_chisq

Q1 <- function(N,n,k,theta,ksi) {

############GET CUT POINTS FOR THETA GROUPS############
cut <- round(quantile(theta,probs=c((1/k)*1:(k-1))),11)
flag <- 0
for (c in 2:(k-1)) {
if (cut[c]==cut[c-1]) flag <- 1}
if (flag == 1) {
ans <- cat_theta(N,n,k,theta)
cut <- ans[2:k]}

############GENERATE FREQUENCIES AND EXPECTED VALUES OF
THETA########
theta_freq <- as.data.frame(table(theta))
t_val <- as.numeric(as.vector(theta_freq$theta))
t_e_freq <- cbind(t_val,theta_freq$Freq)
cut_pt <- cbind(min(theta)-0.001,t(cut),max(theta)+0.001)
grp <- cut(t_e_freq[,1],cut_pt,include.lowest=TRUE,right=FALSE)
t_e_freq <- cbind(t_e_freq,grp)
theta_c <- t(t_e_freq[,1])
logit <- -1*(ksi[,1]%*%theta_c-ksi[,1]*ksi[,2])
t_e_freq_ev <- t(ksi[,3] + (1-
ksi[,3])/(1+exp(logit)))*t_e_freq[,2]
t_e_freq_ev_inc <- t_e_freq[,2] - t_e_freq_ev

data <-
as.data.frame(rbind(t_e_freq_ev,array(rep(0,times=k*n),dim=c(k,n)
)))
data1 <-
as.data.frame(rbind(t_e_freq_ev_inc,array(rep(0,times=k*n),dim=c(
k,n))))
out <-
aggregate(data,list(rbind(t(t(t_e_freq[,3])),t(t(c(1:10))))),sum)
out1 <-
aggregate(data1,list(rbind(t(t(t_e_freq[,3])),t(t(c(1:10))))),sum
)
t_g_ev <- as.matrix(out)[,c(2:(n+1))]
t_g_ev_inc <- as.matrix(out1)[,c(2:(n+1))]

############COLLAPSE CELLS IF EV TO SMALL############
ans <- collapse_cells(N,n,k,t_g_ev,t_g_ev_inc,cut)
```

```
E_1 <- ans[1:k,]
CUT_1 <- ans[(k+1):(2*k),]
K_1 <- ans[(2*k+1),]

############COMPUTE Q1############
ans <- calc_chisq(N,n,k,theta,E_1,CUT_1,K_1)

############OUTPUT############
K_1 <<- K_1
CUT_1 <<- CUT_1
Ng_1 <<- ans[1:k,]
E_1 <<- E_1
O_1 <<- ans[(k+1):(2*k),]
Chi_1 <<- ans[(2*k+1):(3*k),]
Q_1 <<- ans[3*k+1,]
Qp_1 <<- ans[3*k+2,]
Qp_g1 <<- ans[3*k+3,]
flag <<- flag
}
```

Code to Compute *Q*O

```
#Requires Irttoys package
#Input discrimination parameters are on the normal metric
#k = number of score groups
#Requires a N*n matrix of item responses, u
#Requires a N*1 matrix of number-correct scores, nc
#Uses the following functions: collapse_cells, and calc_chisq

QO <- function(N,n,k,ksi,qpts) {
############FREQUENCIES FOR NC SCORES############
ncf <- factor(nc,level=c(0:n))
nc_freq <- table(ncf)
nc_freq <- t(nc_freq)

##NUMBER OF QUAD POINTS
AK_n <- qpts

############OBTAIN JOINT LIKELIHOOD FOR EACH NC SCORE############
quad <- normal.qu(n = AK_n, lower = -4, upper = 4, mu = 0, sigma
= 1)
AK <- quad$quad.points
W <- quad$quad.weights
S <- array(rep(0, times=(n+1)*AK_n), dim=c(AK_n,n+1))
SS <- array(rep(0, times=(n+1)*AK_n*n), dim=c(AK_n,n+1,n))
S1 <- array(rep(0, times=(n+1)*n), dim=c(n,n+1))
S2 <- array(rep(0, times=(n+1)*n), dim=c(n,n+1))
S[,1] <- 1
SS[,1,] <- 1

for (i in 1:n) {
AK_p2 <- ksi[i,3] + (1 - ksi[i,3])/(1+exp(-ksi[i,1]*(AK-
ksi[i,2])))
for (ii in 1:n) {
AK_p1 <- ksi[ii,3] + (1 - ksi[ii,3])/(1+exp(-ksi[ii,1]*(AK-
ksi[ii,2])))
if (i == 1) {
S[,2:(n+1)] <- S[,1:n]*AK_p1 + S[,2:(n+1)]*(1 - AK_p1)
S[,1] <- S[,1]*(1 - AK_p1)}
if (i != ii) {
SS[,2:n,i] <- SS[,1:(n-1),i]*AK_p1 + SS[,2:n,i]*(1 - AK_p1)
SS[,1,i] <- SS[,1,i]*(1 - AK_p1)}}
S2[i,] <- (W*AK_p2) %*% SS[,,i]}
S1 <- W %*% S

############EXPECTED VALUES FOR EACH NC GROUP############
E_O <- nc_freq[2:n]*t(S2[,1:(n-1)])/S1[2:n]
E_O_inc <- nc_freq[2:n] - E_O
```

```
############COLLAPSE CELLS IF EV TOO SMALL############
cut <- (1:(n-1))+.01
ans <- collapse_cells(N,n,k,E_O,E_O_inc,cut)
E_O <- ans[1:k,]
CUT_O <- ans[(k+1):(2*k),]
K_O <- ans[(2*k+1),]

############COMPUTE QO############
sc <- (1 - (t(t(nc))==n))*t(t(nc))
sc <- -1001*(sc == 0) + sc
ans <- calc_chisq(N,n,k,sc,E_O,CUT_O,K_O)
############OUTPUT############
K_O <<- K_O
CUT_O <<- CUT_O
Ng_O <<- ans[1:k,]
E_O <<- E_O
O_O <<- ans[(k+1):(2*k),]
Chi_O <<- ans[(2*k+1):(3*k),]
Q_O <<- ans[3*k+1,]
Qp_O <<- ans[3*k+2,]
Qp_gO <<- ans[3*k+3,]}
```

Functions Used by *Q*1 and *QO*

```
cat_theta <- function(N,n,k,theta) {

############GENERATE INITIAL FREQUENCIES FOR THETA############
t_e_freq <- array(rep(0, times=(min(2^n,N)*4)),
dim=c(min(2^n,N),4))
t_e_sort <- sort(theta)
t_e_freq[1,1] <- t_e_sort[1]
t_e_freq[1,2] <- t_e_sort[1]
t_e_freq[1,3] <- 1
t_e_freq[1,4] <- 1
cell <- 1
for (j in 2:N) {
if (t_e_sort[j] == t_e_sort[j-1]) {
cell <- cell
t_e_freq[cell,3] <- t_e_freq[cell,3] + 1}
else {
cell <- cell + 1
t_e_freq[cell,1] <- t_e_sort[j]
t_e_freq[cell,2] <- t_e_sort[j]
t_e_freq[cell,3] <- 1
t_e_freq[cell,4] <- cell}}
iterations <- cell - k

if (iterations > 0) {
for (I in 1:iterations) {

############GENERATE FREQUENCIES OF FREQUENCIES############
ff <- array(rep(0, times=cell*2), dim=c(cell,2))
f <- t_e_freq[,3]
f <- sort(f[1:cell])
ff[1,1] <- f[1]
ff[1,2] <- 1
ff_cell <- 1
for (j in 2:cell) {
if (f[j] == f[j-1]) {
ff_cell <- ff_cell
ff[ff_cell,2] <- ff[ff_cell,2] + 1}
else {
ff_cell <- ff_cell + 1
ff[ff_cell,1] <- f[j]
ff[ff_cell,2] <- 1}}

maxf <- max(ff)

############MAKE COUNTER TABLE############
counter <- array(rep(0, times=maxf*3), dim=c(maxf,3))
ccc <- 0
```

```
min_freq <- min(f)
for (cc in 1:cell) {
if (t_e_freq[cc,3] == min_freq) {
ccc <- ccc + 1
counter[ccc,1] <- cc
if (cc > 1 & cc < cell) {
if (t_e_freq[cc-1,3]==t_e_freq[cc+1,3]) {
x <- runif(1,min=-1,max=1)
counter[ccc,2] <- t_e_freq[cc-1,3]
counter[ccc,3] <- x/abs(x)}
else {
counter[ccc,2] <- min(t_e_freq[cc-1,3],t_e_freq[cc+1,3])
counter[ccc,3] <- 1
if (counter[ccc,2] == t_e_freq[cc-1,3]) {
counter[ccc,3] <- -2}}
}
else {
if (cc == 1) {
counter[ccc,2] <- t_e_freq[cc+1,3]
counter[ccc,3] <- 1}
if (cc == cell) {
counter[ccc,2] <- t_e_freq[cc-1,3]
counter[ccc,3] <- -1}
}}}

############IDENTIFY CELL TO COLLAPSE############
subcounter <- array(rep(0, times=maxf*3), dim=c(maxf,3))
a <- min((counter[,2])[1:ccc])
target <- 0
dd <- 0
for (d in 1:ccc) {
if (counter[d,2] == a) {
dd <- dd + 1
subcounter[dd,] <- counter[d,]}}
if (dd > 1) target <- trunc(runif(1,min=1,max=dd+1)) else target
<- 1

############MAKE COLLAPSED THETA FREQUENCY TABLE############
t_e_freq_c <- array(rep(0, times=cell*4), dim=c(cell,4))
cell1 <- 1
skip <- 0
for (j in 1:cell) {
if (t_e_freq[j,4] == subcounter[target,1] | skip ==1) {
if (subcounter[target,3] == -1 | skip == 1) {
cell1 <- cell1 - 1
t_e_freq_c[cell1,1] <- t_e_freq[j-1,1]
t_e_freq_c[cell1,2] <- t_e_freq[j,2]
t_e_freq_c[cell1,3] <- t_e_freq[j-1,3] + t_e_freq[j,3]
t_e_freq_c[cell1,4] <- cell1
```

```
skip <- 0}
else {
skip <- 1
}}
else {
t_e_freq_c[cell1,1] <- t_e_freq[j,1]
t_e_freq_c[cell1,2] <- t_e_freq[j,2]
t_e_freq_c[cell1,3] <- t_e_freq[j,3]
t_e_freq_c[cell1,4] <- cell1}
cell1 <- cell1 + 1}
t_e_freq <- t_e_freq_c
cell <- max(t_e_freq[,4])
}}
c1 <- t_e_freq[,1][1:cell]
c2 <- t_e_freq[,2][1:cell]
c3 <- t_e_freq[,3][1:cell]
c4 <- t_e_freq[,4][1:cell]
t_e_freq <- t(rbind(c1,c2,c3,c4))}
```

```
collapse_cells <- function(N,n,k,g_ev,g_ev_inc,cut) {

g_ev_c <- array(rep(0, times=k*n), dim=c(k,n))
g_ev_inc_c <- array(rep(0, times=k*n), dim=c(k,n))
g_ev_c[1,] <- g_ev[1,]
g_ev_inc_c[1,] <- g_ev_inc[1,]
CUT <- array(rep(0, times=k*n), dim=c(k,n))
K <- array(rep(0, times=n), dim=c(n))
c <- 1
d <- 0
dd <- -1

############COLLAPSE FROM BOTTOM############
for (i in 1:n) {
while (c + d < k) {
while (g_ev_c[c,i] < 1 | g_ev_inc_c[c,i] < 1) {
d <- d + 1
dd <- 0
g_ev_c[c,i] <- g_ev_c[c,i] + g_ev[(c+d),i]
g_ev_inc_c[c,i] <- g_ev_inc_c[c,i] + g_ev_inc[(c+d),i]
CUT[c,i] <- cut[c+d]
if (c+d == k) break}
if (c+d == k) break
c <- c + 1
g_ev_c[c,i] <- g_ev[(c+d),i]
g_ev_inc_c[c,i] <- g_ev_inc[(c+d),i]
CUT[c+dd,i] <- cut[c+d+dd]}

if (g_ev_c[c,i] < 1 | g_ev_inc_c[c,i] < 1) {
g_ev_c[(c-1),i] <- g_ev_c[(c-1),i] + g_ev_c[c,i]
g_ev_c[c,i] <- 0
CUT[(c-1),i] <- -1000
K[i] <- c - 1}
if (g_ev_c[c,i] >= 1 & g_ev_inc_c[c,i] >= 1)
{K[i] <- c}

CUT[K[i]:k,i] <- -1000

c <- 1
d <- 0
dd <- -1}
ev_data <- rbind(g_ev_c, CUT, K)}
```

```
calc_chisq <- function(N,n,k,sc,E,CUT,K) {

############GET OBSERVED VALUES############
O <- array(rep(0, times=n*k), dim=c(k,n))
Ng <- array(rep(0, times=n*k), dim=c(k,n))
sc_rep <- array(rep(sc, times=k), dim=c(N,k))
Krep <- t(array(rep(K, times=N), dim=c(n,N)))
Q <- array(rep(0, times=n), dim=c(n))

group <- array(rep(0, times=N*n), dim=c(N,n))
for (i in 1:n) {
id <- 1*t((t(sc_rep) < CUT[,i]))
for (ii in 1:k) {
group[,i] <- group[,i] + id[,ii]}}

group <- Krep-group
groupu <- u*group

for (i in 1:n) {
groupf <- factor(group[,i],level=c(1:k),exclude=0)
groupuf <- factor(groupu[,i],level=c(1:k),exclude=0)
groupt <- t(t(table(groupf)))
grouput <- t(t(table(groupuf)))
Ng[,i] <- groupt
O[,i] <- grouput}

############OBTAIN Q############
P <- O/Ng
p <- E/Ng
Chi <- Ng*((P - p)^2)/(p*(1-p))

for (i in 1:n) {
Q[i] <- sum(Chi[1:(K[i]),i])}

Q_p <- pchisq(Q,K-model,lower.tail=FALSE)
Q_p_g <- pchisq(Q,K,lower.tail=FALSE)

Q <- rbind(Ng,O,Chi,Q,Q_p,Q_p_g)}
```

Code to Compute LM Statistics

```
#Requires Irttoys package

LM_EV <- function(N,n,k,ksi,qpts) {

AK_n <- qpts
quad <- normal.qu(n = AK_n, lower = -4, upper = 4, mu = 0, sigma
= 1)
AK <- quad$quad.points
W <- quad$quad.weights
P <- array(rep(0, times=n*AK_n), dim=c(AK_n,n))
P_Star <- array(rep(0, times=n*AK_n), dim=c(AK_n,n))
L_a <- array(rep(0, times=N*n*AK_n), dim=c(n,N,AK_n))
L_b <- array(rep(0, times=N*n*AK_n), dim=c(n,N,AK_n))
L_aa <- array(rep(0, times=N*n*AK_n), dim=c(n,N,AK_n))
L_bb <- array(rep(0, times=N*n*AK_n), dim=c(n,N,AK_n))
L_ab <- array(rep(0, times=N*n*AK_n), dim=c(n,N,AK_n))
L_q <- array(rep(0, times=N*AK_n), dim=c(N,AK_n))
group <- array(rep(0, times=n*N), dim=c(N,n))
flag <- array(rep(0, times=n), dim=c(n,1))

u <- t(u)
for (q in 1:qpts) {
P[q,] <- ksi[,3] + (1 - ksi[,3])/(1+exp(-ksi[,1]*(AK[q]-
ksi[,2])))
P_Star[q,] <- 1/(1+exp(-ksi[,1]*(AK[q]-ksi[,2])))
#exp1 <- (P[q,]-ksi[,3])/(P[q,]*(1-ksi[,3]))
#exp2 <- (1-P[q,])/(P[q,]*(1-ksi[,3]))
L_a[,,q] <- (u - P[q,])*AK[q]*(P_Star[q,]/P[q,])
L_b[,,q] <- (P[q,] - u)*(P_Star[q,]/P[q,])
P[q,])*(ksi[,1]*ksi[,3]*exp1*exp2+exp1)
l <- t((P[q,]^u)*(1-P[q,])^(1-u))
L_q[,q] <- apply(l,1,prod)*W[q]}
u <- t(u)

P_q <- L_q/apply(L_q,1,sum)

for (i in 1:n) {
L_a[i,,] <- L_a[i,,]*P_q
L_b[i,,] <- L_b[i,,]*P_q
nc_adj <- nc - u[,i]
cut_pt <- round(quantile(nc_adj,(0:k)/k),1)
cut_pt[k+1] <- cut_pt[k+1] + .01
if (cut_pt[k] > (n - 2))
{flag[i] <- 1}
grp <- cut(nc_adj,cut_pt,include.lowest=TRUE,right=FALSE)
temp <- cbind(nc_adj,grp)
group[,i] <- temp[,2]
```

```
}
La <- t(apply(L_a,c(1,2),sum))
Lb <- t(apply(L_b,c(1,2),sum))
Laa <- La^2
Lbb <- Lb^2
Lab <- La*Lb
Laa_tot <- apply(Laa,2,sum)
Lbb_tot <- apply(Lbb,2,sum)
Lab_tot <- apply(Lab,2,sum)

item <- t(array(rep(1:n,times=N),dim=c(n,N)))
ind1 <- list(group,item)

La_g <- tapply(La, ind1, sum)
Lb_g <- tapply(Lb, ind1, sum)
Laa_g <- tapply(Laa, ind1, sum)
Lbb_g <- tapply(Lbb, ind1, sum)
Lab_g <- tapply(Lab, ind1, sum)

############COMPUTE LM############
LM2 <- array(rep(0, times=n), dim=c(n,1))
LM2a <- array(rep(0, times=n), dim=c(n,1))
LM2b <- array(rep(0, times=n), dim=c(n,1))
for (i in 1:n) {
h_2 <- array(c(La_g[c(1:(k-1)),i],Lb_g[c(1:(k-1)),i]),dim=c(2*(k-
1),1))
h_2a <- array(c(La_g[c(1:(k-1)),i]),dim=c((k-1),1))
h_2b <- array(c(Lb_g[c(1:(k-1)),i]),dim=c((k-1),1))
H_22 <- rbind(cbind(diag(Laa_g[c(1:(k-1)),i]),diag(Lab_g[c(1:(k-
1)),i])),cbind(diag(Lab_g[c(1:(k-1)),i]),diag(Lbb_g[c(1:(k-
1)),i])))
H_22a <- diag(Laa_g[c(1:(k-1)),i])
H_22b <- diag(Lbb_g[c(1:(k-1)),i])
H_21 <- rbind(cbind(Laa_g[c(1:(k-1)),i],Lab_g[c(1:(k-
1)),i]),cbind(Lab_g[c(1:(k-1)),i],Lbb_g[c(1:(k-1)),i]))
H_21a <- array(Laa_g[c(1:(k-1)),i],dim=c((k-1),1))
H_21b <- array(Lbb_g[c(1:(k-1)),i],dim=c((k-1),1))
H_11 <-
rbind(cbind(Laa_tot[i],Lab_tot[i]),cbind(Lab_tot[i],Lbb_tot[i]))
H_11a <- Laa_tot[i]
H_11b <- Lbb_tot[i]
W_LM <- H_22 - H_21%*%solve(H_11,t(H_21))
W_LMa <- H_22a - (H_21a/H_11a)%*%t(H_21a)
W_LMb <- H_22b - (H_21b/H_11b)%*%t(H_21b)
LM2[i] <- t(h_2)%*%solve(W_LM,h_2,tol=1e-100)
LM2a[i] <- t(h_2a)%*%solve(W_LMa,h_2a,tol=1e-100)
LM2b[i] <- t(h_2b)%*%solve(W_LMb,h_2b,tol=1e-100)
}
############OUTPUT############
```

```
LM2 <<- LM2
LM2a <<- LM2a
LM2b <<- LM2b
LM2_p <<- pchisq(LM2,2*(k-1),lower.tail=FALSE)
LM2a_p <<- pchisq(LM2a,(k-1),lower.tail=FALSE)
LM2b_p <<- pchisq(LM2b,(k-1),lower.tail=FALSE)
flag <<- flag}
```

Code to Compute *z* Statistics

```
z_stats <- function(N,n,theta,ksi) {

############COMPUTE P############
p <- array(rep(0, times=N*n), dim=c(N,n))
for (i in 1:n) {
p[,i]<-ksi[i,3] + (1-ksi[i,3])/(1+exp(-ksi[i,1]*(theta-
ksi[i,2])))
}

############COMPUTE INFIT/OUTFIT############
##MEAN SQUARES
vO <- apply(((u - p)^2)/(p*(1-p)),2,sum)/N
vI_num <- apply((u - p)^2,2,sum)
w <- apply(p*(1-p),2,sum)
w1 <- apply(1/(p*(1-p)),2,sum)
w2 <- apply((p*(1-p))^2,2,sum)
vI <- (vI_num/w)
SD_vO <- sqrt(w1 - 4*N)/N
SD_vI <- sqrt(w - 4*w2)/w
#TRANSFORM TO NORMAL
vO_z <- 3*(vO^(1/3) - 1)/SD_vO + SD_vO/3
vI_z <- 3*(vI^(1/3) - 1)/SD_vI + SD_vI/3
vO_z_p <- pnorm(abs(vO_z), mean=0, sd=1, lower.tail = FALSE)
vI_z_p <- pnorm(abs(vI_z), mean=0, sd=1, lower.tail = FALSE)

############COMPUTE lz############
lnL <- u*log(p) + (1 - u)*log(1 - p)
E_lnL <- p*log(p) + (1 - p)*log(1 - p)
Var_lnL <- p*(1 - p)*(log(p/(1 - p)))^2

lnL <- apply(lnL,2,sum)
E_lnL <- apply(E_lnL,2,sum)
Var_lnL <- apply(Var_lnL,2,sum)

lz <- (lnL - E_lnL)/sqrt(Var_lnL)
lz_p <- pnorm(abs(lz), mean=0, sd=1, lower.tail = FALSE)

############OUTPUT############
VO_Z <<- vO_z
VI_Z <<- vI_z
VO_Z_p <<- vO_z_p
VI_Z_p <<- vI_z_p
LZ_O <<- lnL
LZ_E <<- E_lnL
LZ_Var <<- Var_lnL
LZ <<- lz
LZ_p <<- lz_p}
```

APPENDIX B: MODEL PARAMETERS

Descriptive statistics for generating parameters

Descriptive statistics for estimated parameters

Relationship between estimated EU and SU parameters

Relationship between generating and estimated parameters

Table B-1. Descriptive Statistics for Generating *b* Parameters in SU Conditions

M	D	*N*	*n*	Mean	Median	SD	Skew	Kurt
1PL	High	500	15	−0.254	−0.257	0.840	0.023	0.041
			75	−0.173	−0.174	0.849	−0.007	−0.055
		1,500	15	−0.257	−0.256	0.835	−0.004	−0.003
			75	−0.170	−0.175	0.861	−0.001	−0.005
	Low	500	15	−0.241	−0.243	0.788	0.025	0.049
			75	−0.200	−0.202	0.785	0.000	−0.053
		1,500	15	−0.245	−0.245	0.783	−0.001	−0.002
			75	−0.196	−0.204	0.796	0.000	−0.003
2PL	High	500	15	−0.256	−0.251	0.805	0.013	0.005
			75	−0.150	−0.151	0.846	−0.006	0.019
		1,500	15	−0.253	−0.249	0.814	−0.008	0.006
			75	−0.155	−0.153	0.844	−0.027	0.073
	Low	500	15	−0.263	−0.258	0.807	0.010	0.005
			75	−0.198	−0.199	0.808	0.004	0.050
		1,500	15	−0.259	−0.255	0.818	−0.012	0.026
			75	−0.204	−0.201	0.805	−0.018	0.105
3PL	High	500	15	0.245	0.245	0.812	0.035	0.089
			75	0.155	0.156	0.807	0.027	0.127
		1,500	15	0.254	0.256	0.809	0.045	0.108
			75	0.151	0.153	0.800	0.027	0.013
	Low	500	15	0.245	0.246	0.811	0.020	0.046
			75	0.203	0.206	0.802	0.003	0.073
		1,500	15	0.256	0.261	0.809	0.020	0.057
			75	0.197	0.202	0.797	0.004	0.012

Table B-2. Descriptive Statistics for Generating *a* Parameters in SU Conditions

M	D	*N*	*n*	Mean	Median	SD	Skew	Kurt
2PL	High	500	15	1.069	1.078	0.259	−0.165	−0.104
			75	1.013	1.022	0.245	−0.179	−0.166
		1,500	15	1.072	1.075	0.257	−0.105	−0.112
			75	1.013	1.023	0.243	−0.173	−0.140
	Low	500	15	0.647	0.649	0.156	−0.082	−0.107
			75	0.647	0.649	0.157	−0.082	−0.127
		1,500	15	0.648	0.648	0.156	−0.025	−0.092
			75	0.647	0.649	0.156	−0.069	−0.091
3PL	High	500	15	1.121	1.129	0.253	−0.183	−0.094
			75	1.116	1.126	0.254	−0.189	−0.079
		1,500	15	1.121	1.132	0.254	−0.185	−0.060
			75	1.118	1.130	0.253	−0.221	−0.116
	Low	500	15	0.681	0.684	0.155	−0.072	−0.083
			75	0.685	0.687	0.158	−0.074	−0.112
		1,500	15	0.681	0.683	0.156	−0.063	−0.041
			75	0.685	0.689	0.157	−0.089	−0.103

NOTE. 1PL parameters were fixed at 1.000 in high discrimination condition and 0.653 in low discrimination conditions

Table B-3. Descriptive Statistics for Generating *c* Parameters in SU Conditions

M	D	*N*	*n*	Mean	Median	SD	Skew	Kurt
2PL	High	500	15	0.202	0.199	0.031	0.435	0.335
			75	0.202	0.199	0.030	0.418	0.320
		1,500	15	0.202	0.199	0.031	0.475	0.399
			75	0.202	0.199	0.030	0.427	0.241
	Low	500	15	0.202	0.199	0.031	0.435	0.335
			75	0.202	0.199	0.030	0.418	0.320
		1,500	15	0.202	0.199	0.031	0.475	0.399
			75	0.202	0.199	0.030	0.427	0.241

Table B-4. Descriptive Statistics for Generating θ Parameters in SU Conditions

M	D	*N*	*n*	Mean	Median	SD	Skew	Kurt
1PL	High	500	15	0.011	0.012	1.085	0.002	0.004
			75	0.104	0.102	1.086	−0.005	−0.014
		1,500	15	0.011	0.011	1.084	0.002	0.001
			75	0.101	0.102	1.086	−0.004	0.009
	Low	500	15	0.000	0.000	0.995	−0.002	−0.007
			75	0.051	0.052	0.995	−0.009	−0.005
		1,500	15	0.002	0.003	0.997	−0.002	0.009
			75	0.053	0.053	0.997	0.001	−0.012
2PL	High	500	15	0.013	0.012	1.017	0.001	−0.006
			75	0.123	0.122	1.075	0.008	0.010
		1,500	15	0.011	0.011	1.015	0.001	0.000
			75	0.128	0.129	1.073	0.000	−0.001
	Low	500	15	0.002	0.002	1.008	−0.001	0.001
			75	0.057	0.056	1.006	−0.009	−0.002
		1,500	15	0.002	0.003	1.009	0.001	0.008
			75	0.058	0.059	1.007	−0.007	0.001
3PL	High	500	15	−0.012	−0.012	1.020	−0.001	−0.009
			75	−0.110	−0.110	1.020	−0.003	−0.010
		1,500	15	−0.014	−0.014	1.021	0.000	0.002
			75	−0.108	−0.107	1.023	0.002	−0.003
	Low	500	15	−0.006	−0.005	1.009	0.001	−0.001
			75	−0.056	−0.056	1.000	−0.005	−0.035
		1,500	15	−0.004	−0.003	1.008	−0.002	0.012
			75	−0.051	−0.052	1.000	0.000	0.001

Table B-5. Descriptive Statistics for Estimated 1PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.00	−0.25	--	−0.01	0.65	−0.24	--	−0.01	1.00	−0.25	--	−0.01	0.65	−0.24	--	−0.01
SD	0.00	0.21	--	0.01	0.00	0.20	--	0.00	0.00	0.21	--	0.01	0.00	0.20	--	0.00
Min	1.00	−1.06	--	−0.04	0.65	−0.91	--	−0.03	1.00	−1.14	--	−0.05	0.65	−0.96	--	−0.03
Max	1.00	0.35	--	0.00	0.65	0.46	--	0.01	1.00	0.54	--	0.01	0.65	0.39	--	0.01
SD for Test																
Mean	0.00	0.81	--	0.92	0.00	0.78	--	0.84	0.00	0.83	--	0.92	0.00	0.78	--	0.84
SD	0.00	0.16	--	0.02	0.00	0.15	--	0.02	0.00	0.16	--	0.02	0.00	0.15	--	0.02
Min	0.00	0.34	--	0.84	0.00	0.34	--	0.77	0.00	0.37	--	0.84	0.00	0.30	--	0.77
Max	0.00	1.38	--	0.99	0.00	1.34	--	0.93	0.00	1.43	--	0.98	0.00	1.36	--	0.92
Mean absolute distance (MAD) Between Test Mean and Median																
	0.00	0.12	--	0.06	0.00	0.12	--	0.06	0.00	0.12	--	0.06	0.00	0.11	--	0.06

Table B-6. Descriptive Statistics for Estimated 1PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.00	−0.17	0.00	0.09	0.65	−0.20	0.00	0.04	1.00	−0.17	0.00	0.09	0.65	−0.20	0.00	0.04
SD	0.00	0.07	0.00	0.05	0.00	0.08	0.00	0.02	0.00	0.07	0.00	0.06	0.00	0.08	0.00	0.02
Min	1.00	−0.35	0.00	−0.04	0.65	−0.39	0.00	−0.02	1.00	−0.36	0.00	−0.09	0.65	−0.41	0.00	−0.03
Max	1.00	0.07	0.00	0.23	0.65	0.00	0.00	0.10	1.00	0.05	0.00	0.25	0.65	0.00	0.00	0.10
SD for Test																
Mean	0.00	0.85	0.00	1.03	0.00	0.79	0.00	0.95	0.00	0.85	0.00	1.03	0.00	0.79	0.00	0.95
SD	0.00	0.07	0.00	0.03	0.00	0.06	0.00	0.03	0.00	0.07	0.00	0.03	0.00	0.06	0.00	0.03
Min	0.00	0.67	0.00	0.95	0.00	0.63	0.00	0.88	0.00	0.67	0.00	0.95	0.00	0.62	0.00	0.85
Max	0.00	1.05	0.00	1.13	0.00	0.98	0.00	1.04	0.00	1.04	0.00	1.11	0.00	0.98	0.00	1.05
Mean absolute distance (MAD) Between Test Mean and Median																
	0.00	0.06	0.00	0.03	0.00	0.05	0.00	0.03	0.00	0.06	0.00	0.03	0.00	0.06	0.00	0.03

Table B-7. Descriptive Statistics for Estimated 1PL Model Parameters Across Test Replications: $N = 1{,}500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.00	−0.26	--	−0.01	0.65	−0.25	--	−0.01	1.00	−0.26	--	−0.01	0.65	−0.24	--	−0.01
SD	0.00	0.21	--	0.01	0.00	0.20	--	0.01	0.00	0.21	--	0.01	0.00	0.20	--	0.01
Min	1.00	−0.94	--	−0.04	0.65	−0.91	--	−0.03	1.00	−0.93	--	−0.04	0.65	−0.89	--	−0.03
Max	1.00	0.38	--	0.01	0.65	0.34	--	0.01	1.00	0.42	--	0.01	0.65	0.37	--	0.01
SD for Test																
Mean	0.00	0.81	--	0.92	0.00	0.77	--	0.84	0.00	0.82	--	0.92	0.00	0.77	--	0.84
SD	0.00	0.15	--	0.02	0.00	0.14	--	0.01	0.00	0.15	--	0.02	0.00	0.15	--	0.01
Min	0.00	0.38	--	0.86	0.00	0.35	--	0.79	0.00	0.37	--	0.85	0.00	0.36	--	0.79
Max	0.00	1.31	--	0.96	0.00	1.28	--	0.88	0.00	1.33	--	0.97	0.00	1.25	--	0.88
Mean absolute distance (MAD) Between Test Mean and Median																
	0.00	0.12	--	0.05	0.00	0.11	--	0.06	0.00	0.12	--	0.05	0.00	0.11	--	0.06

Table B-8. Descriptive Statistics for Estimated 1PL Model Parameters Across Test Replications: $N = 1{,}500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.00	−0.17	--	0.09	0.65	−0.20	--	0.05	1.00	−0.17	--	0.09	0.65	−0.20	--	0.05
SD	0.00	0.07	--	0.04	0.00	0.08	--	0.02	0.00	0.07	--	0.04	0.00	0.08	--	0.02
Min	1.00	−0.38	--	−0.05	0.65	−0.40	--	−0.03	1.00	−0.37	--	−0.02	0.65	−0.39	--	−0.02
Max	1.00	0.04	--	0.21	0.65	0.01	--	0.10	1.00	0.03	--	0.20	0.65	0.06	--	0.10
SD for Test																
Mean	0.00	0.86	--	1.03	0.00	0.80	--	0.95	0.00	0.86	--	1.03	0.00	0.80	--	0.95
SD	0.00	0.07	--	0.02	0.00	0.06	--	0.02	0.00	0.07	--	0.02	0.00	0.06	--	0.02
Min	0.00	0.70	--	0.99	0.00	0.65	--	0.92	0.00	0.70	--	0.99	0.00	0.66	--	0.91
Max	0.00	1.04	--	1.09	0.00	0.96	--	0.99	0.00	1.04	--	1.09	0.00	0.97	--	1.00
Mean absolute distance (MAD) Between Test Mean and Median																
	0.00	0.06	--	0.02	0.00	0.05	--	0.02	0.00	0.06	--	0.02	0.00	0.06	--	0.02

Table B-9. Descriptive Statistics for Estimated 2PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.07	−0.26	--	−0.01	0.64	−0.26	--	−0.01	1.05	−0.26	--	−0.01	0.64	−0.26	--	−0.01
SD	0.06	0.21	--	0.01	0.03	0.21	--	0.01	0.06	0.21	--	0.01	0.04	0.21	--	0.01
Min	0.85	−0.93	--	−0.03	0.51	−0.99	--	−0.03	0.87	−0.89	--	−0.03	0.52	−1.02	--	−0.03
Max	1.28	0.37	--	0.01	0.75	0.33	--	0.01	1.27	0.36	--	0.01	0.74	0.41	--	0.01
SD for Test																
Mean	0.25	0.79	--	0.89	0.14	0.79	--	0.85	0.24	0.80	--	0.88	0.14	0.80	--	0.84
SD	0.05	0.15	--	0.01	0.03	0.15	--	0.02	0.05	0.15	--	0.01	0.03	0.15	--	0.02
Min	0.13	0.43	--	0.84	0.06	0.39	--	0.77	0.10	0.41	--	0.84	0.04	0.39	--	0.78
Max	0.39	1.27	--	0.94	0.23	1.27	--	0.90	0.45	1.30	--	0.94	0.23	1.32	--	0.90
Mean absolute distance (MAD) Between Test Mean and Median																
	0.04	0.12	--	0.03	0.02	0.12	--	0.03	0.04	0.12	--	0.03	0.02	0.12	--	0.03

Table B-10. Descriptive Statistics for Estimated 2PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.02	−0.15	--	0.11	0.63	−0.21	--	0.04	1.01	−0.15	--	0.11	0.63	−0.21	--	0.04
SD	0.02	0.07	--	0.06	0.01	0.08	--	0.02	0.03	0.06	--	0.06	0.01	0.08	--	0.02
Min	0.94	−0.32	--	−0.05	0.60	−0.38	--	−0.01	0.94	−0.31	--	−0.04	0.60	−0.41	--	−0.02
Max	1.08	0.02	--	0.29	0.66	0.00	--	0.10	1.09	0.03	--	0.32	0.67	0.00	--	0.10
SD for Test																
Mean	0.25	0.84	--	1.02	0.14	0.82	--	0.98	0.24	0.85	--	1.01	0.14	0.82	--	0.98
SD	0.02	0.07	--	0.02	0.01	0.07	--	0.02	0.02	0.07	--	0.02	0.01	0.07	--	0.02
Min	0.20	0.64	--	0.95	0.11	0.63	--	0.92	0.18	0.63	--	0.96	0.11	0.62	--	0.92
Max	0.31	1.03	--	1.08	0.18	0.97	--	1.03	0.30	1.02	--	1.07	0.18	0.99	--	1.05
Mean absolute distance (MAD) Between Test Mean and Median																
	0.02	0.06	--	0.03	0.01	0.06	--	0.03	0.02	0.06	--	0.03	0.01	0.06	--	0.02

Table B-11. Descriptive Statistics for Estimated 2PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.08	−0.25	--	−0.01	0.65	−0.26	--	−0.01	1.07	−0.25	--	−0.01	0.65	−0.26	--	−0.01
SD	0.07	0.21	--	0.01	0.04	0.21	--	0.01	0.07	0.21	--	0.01	0.04	0.21	--	0.01
Min	0.87	−0.99	--	−0.03	0.53	−0.96	--	−0.04	0.85	−0.98	--	−0.03	0.52	−0.95	--	−0.03
Max	1.34	0.50	--	0.01	0.82	0.47	--	0.02	1.31	0.49	--	0.02	0.79	0.47	--	0.02
SD for Test																
Mean	0.25	0.79	--	0.88	0.15	0.80	--	0.84	0.25	0.80	--	0.88	0.15	0.81	--	0.84
SD	0.05	0.16	--	0.01	0.03	0.16	--	0.01	0.05	0.16	--	0.01	0.03	0.16	--	0.01
Min	0.12	0.38	--	0.84	0.06	0.33	--	0.80	0.09	0.38	--	0.84	0.07	0.34	--	0.80
Max	0.41	1.37	--	0.91	0.25	1.31	--	0.89	0.40	1.34	--	0.91	0.25	1.33	--	0.88
Mean absolute distance (MAD) Between Test Mean and Median																
	0.04	0.12	--	0.02	0.02	0.12	--	0.02	0.04	0.12	--	0.02	0.02	0.12	--	0.02

Table B-12. Descriptive Statistics for Estimated 2PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.02	−0.15	--	0.11	0.64	−0.21	--	0.05	1.01	−0.15	--	0.11	0.64	−0.21	--	0.05
SD	0.03	0.06	--	0.05	0.02	0.08	--	0.02	0.03	0.06	--	0.05	0.01	0.08	--	0.02
Min	0.95	−0.31	--	−0.06	0.60	−0.43	--	−0.02	0.94	−0.32	--	−0.07	0.60	−0.43	--	−0.01
Max	1.10	0.02	--	0.24	0.69	0.06	--	0.10	1.08	0.01	--	0.25	0.67	0.02	--	0.11
SD for Test																
Mean	0.24	0.83	--	1.01	0.15	0.81	--	0.97	0.24	0.85	--	1.01	0.15	0.81	--	0.97
SD	0.02	0.07	--	0.01	0.01	0.07	--	0.01	0.02	0.07	--	0.01	0.01	0.07	--	0.01
Min	0.19	0.64	--	0.98	0.12	0.62	--	0.94	0.19	0.64	--	0.97	0.12	0.60	--	0.94
Max	0.30	1.03	--	1.05	0.18	1.01	--	1.00	0.29	1.05	--	1.04	0.18	1.01	--	0.99
Mean absolute distance (MAD) Between Test Mean and Median																
	0.02	0.06	--	0.02	0.01	0.05	--	0.02	0.02	0.06	--	0.02	0.01	0.06	--	0.02

Table B-13. Descriptive Statistics for Estimated 3PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.11	0.24	0.20	0.05	0.68	0.24	0.20	0.05	1.09	0.23	0.20	0.05	0.68	0.24	0.20	0.05
SD	0.06	0.20	0.01	0.01	0.04	0.21	0.01	0.01	0.06	0.20	0.01	0.01	0.04	0.21	0.01	0.01
Min	0.94	−0.49	0.18	0.02	0.55	−0.39	0.18	0.01	0.91	−0.44	0.17	0.02	0.53	−0.37	0.18	0.01
Max	1.30	0.86	0.23	0.10	0.80	0.87	0.22	0.08	1.30	0.77	0.23	0.10	0.80	0.90	0.22	0.09
SD for Test																
Mean	0.23	0.79	0.03	0.81	0.14	0.81	0.02	0.78	0.22	0.80	0.03	0.81	0.14	0.81	0.02	0.77
SD	0.04	0.16	0.01	0.02	0.03	0.17	0.00	0.02	0.04	0.16	0.01	0.02	0.03	0.16	0.00	0.02
Min	0.08	0.39	0.01	0.75	0.07	0.39	0.01	0.72	0.10	0.39	0.01	0.75	0.07	0.32	0.01	0.71
Max	0.41	1.42	0.05	0.87	0.23	1.90	0.05	0.83	0.38	1.68	0.06	0.87	0.22	1.40	0.04	0.83
Mean absolute distance (MAD) Between Test Mean and Median																
	0.04	0.12	0.00	0.04	0.02	0.12	0.00	0.03	0.03	0.12	0.00	0.04	0.02	0.12	0.00	0.03

Table B-14. Descriptive Statistics for Estimated 3PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.10	0.19	0.20	−0.04	0.68	0.23	0.20	−0.01	1.09	0.18	0.20	−0.04	0.68	0.22	0.20	−0.01
SD	0.03	0.07	0.00	0.04	0.02	0.09	0.00	0.01	0.03	0.07	0.00	0.04	0.02	0.09	0.00	0.01
Min	1.02	−0.02	0.19	−0.14	0.64	0.00	0.20	−0.04	1.02	0.00	0.19	−0.14	0.63	0.01	0.20	−0.05
Max	1.19	0.38	0.21	0.07	0.73	0.47	0.21	0.03	1.17	0.37	0.21	0.06	0.72	0.50	0.21	0.02
SD for Test																
Mean	0.24	0.80	0.03	0.94	0.15	0.81	0.03	0.92	0.24	0.81	0.03	0.93	0.15	0.82	0.03	0.92
SD	0.02	0.07	0.00	0.02	0.01	0.08	0.00	0.02	0.02	0.07	0.00	0.02	0.01	0.07	0.00	0.02
Min	0.18	0.60	0.03	0.87	0.12	0.62	0.02	0.88	0.19	0.63	0.02	0.88	0.12	0.59	0.02	0.88
Max	0.29	1.00	0.04	0.99	0.18	1.04	0.04	0.98	0.30	1.08	0.04	0.99	0.19	1.05	0.03	0.96
Mean absolute distance (MAD) Between Test Mean and Median																
	0.02	0.06	0.00	0.03	0.01	0.05	0.00	0.03	0.02	0.05	0.00	0.03	0.01	0.05	0.00	0.03

Table B-15. Descriptive Statistics for Estimated 3PL Model Parameters Across Test Replications: $N = 1{,}500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.12	0.26	0.20	0.06	0.69	0.26	0.20	0.05	1.10	0.26	0.20	0.06	0.68	0.26	0.20	0.05
SD	0.06	0.21	0.01	0.02	0.04	0.21	0.01	0.02	0.06	0.21	0.01	0.02	0.04	0.21	0.01	0.02
Min	0.93	−0.37	0.18	0.00	0.56	−0.45	0.18	0.01	0.92	−0.37	0.18	0.01	0.55	−0.45	0.18	0.00
Max	1.31	0.95	0.24	0.14	0.80	0.99	0.23	0.11	1.31	0.96	0.23	0.14	0.78	0.92	0.24	0.10
SD for Test																
Mean	0.24	0.78	0.03	0.81	0.15	0.79	0.03	0.77	0.24	0.80	0.03	0.81	0.15	0.80	0.03	0.77
SD	0.05	0.16	0.01	0.02	0.03	0.16	0.01	0.01	0.05	0.16	0.01	0.02	0.03	0.16	0.01	0.01
Min	0.10	0.36	0.01	0.74	0.07	0.36	0.01	0.73	0.09	0.35	0.01	0.74	0.06	0.35	0.01	0.72
Max	0.39	1.78	0.06	0.86	0.25	1.57	0.05	0.81	0.42	1.41	0.06	0.86	0.25	1.38	0.05	0.81
Mean absolute distance (MAD) Between Test Mean and Median																
	0.04	0.12	0.00	0.03	0.02	0.12	0.00	0.03	0.04	0.12	0.00	0.03	0.02	0.12	0.00	0.03

Table B-16. Descriptive Statistics for Estimated 3PL Model Parameters Across Test Replications: $N = 1{,}500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Mean for Test																
Mean	1.11	0.20	0.20	−0.03	0.69	0.23	0.21	0.00	1.10	0.20	0.20	−0.03	0.68	0.23	0.21	0.00
SD	0.02	0.07	0.00	0.03	0.02	0.09	0.00	0.01	0.03	0.07	0.00	0.03	0.02	0.09	0.00	0.01
Min	1.04	0.01	0.19	−0.10	0.64	−0.01	0.19	−0.02	1.03	0.01	0.19	−0.11	0.63	−0.01	0.20	−0.03
Max	1.18	0.42	0.21	0.09	0.74	0.53	0.22	0.03	1.16	0.46	0.21	0.06	0.73	0.49	0.22	0.03
SD for Test																
Mean	0.25	0.79	0.03	0.94	0.15	0.80	0.03	0.92	0.25	0.81	0.03	0.93	0.15	0.80	0.03	0.92
SD	0.02	0.07	0.00	0.01	0.01	0.07	0.00	0.01	0.02	0.07	0.00	0.01	0.01	0.07	0.00	0.01
Min	0.19	0.56	0.03	0.90	0.12	0.58	0.02	0.89	0.19	0.59	0.03	0.90	0.12	0.57	0.02	0.90
Max	0.30	0.99	0.04	0.96	0.19	1.02	0.04	0.94	0.31	1.02	0.04	0.96	0.19	1.03	0.04	0.94
Mean absolute distance (MAD) Between Test Mean and Median																
	0.02	0.05	0.00	0.03	0.01	0.06	0.00	0.02	0.02	0.05	0.00	0.03	0.01	0.05	0.00	0.02

Table B-17. Relationship Between Estimated EU and SU 1PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	--	0.99	--	0.00	--	0.98	--	0.00
	SD	--	0.00	--	0.04	--	0.01	--	0.04
	Min	--	0.96	--	–0.18	--	0.91	--	–0.13
	Max	--	1.00	--	0.17	--	1.00	--	0.16
MAD	Mean	--	0.10	--	1.05	--	0.12	--	0.96
	SD	--	0.03	--	0.03	--	0.03	--	0.03
	Min	--	0.04	--	0.93	--	0.05	--	0.84
	Max	--	0.25	--	1.15	--	0.27	--	1.06
MD	Mean	--	0.00	--	0.00	--	0.00	--	0.00
	SD	--	0.07	--	0.00	--	0.07	--	0.00
	Min	--	–0.22	--	–0.01	--	–0.19	--	–0.01
	Max	--	0.24	--	0.01	--	0.27	--	0.01

MAD – Mean absolute distance; MD – Mean distance

Table B-18. Relationship Between Estimated EU and SU 1PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	--	0.99	--	0.00	--	0.98	--	0.00
	SD	--	0.00	--	0.05	--	0.00	--	0.05
	Min	--	0.99	--	–0.11	--	0.97	--	–0.12
	Max	--	1.00	--	0.15	--	0.99	--	0.14
MAD	Mean	--	0.10	--	1.17	--	0.12	--	1.08
	SD	--	0.02	--	0.04	--	0.01	--	0.04
	Min	--	0.07	--	1.08	--	0.09	--	0.98
	Max	--	0.20	--	1.30	--	0.17	--	1.19
MD	Mean	--	0.00	--	0.00	--	0.00	--	0.00
	SD	--	0.07	--	0.06	--	0.05	--	0.02
	Min	--	–0.16	--	–0.16	--	–0.16	--	–0.04
	Max	--	0.20	--	0.14	--	0.14	--	0.05

Table B-19. Relationship Between Estimated EU and SU 1PL Model Parameters Across Test Replications: N = 1,500, n = 15

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	--	1.00	--	0.00	--	0.99	--	0.00
	SD	--	0.00	--	0.03	--	0.00	--	0.03
	Min	--	0.99	--	–0.10	--	0.97	--	–0.09
	Max	--	1.00	--	0.09	--	1.00	--	0.08
MAD	Mean	--	0.06	--	1.05	--	0.07	--	0.96
	SD	--	0.02	--	0.02	--	0.02	--	0.02
	Min	--	0.02	--	0.95	--	0.03	--	0.90
	Max	--	0.14	--	1.11	--	0.12	--	1.02
MD	Mean	--	0.00	--	0.00	--	0.00	--	0.00
	SD	--	0.04	--	0.00	--	0.04	--	0.00
	Min	--	–0.12	--	–0.01	--	–0.11	--	0.00
	Max	--	0.14	--	0.01	--	0.12	--	0.00

MAD – Mean absolute distance; MD – Mean distance

Table B-20. Relationship Between Estimated EU and SU 1PL Model Parameters Across Test Replications: N = 1,500, n = 75

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	--	1.00	--	0.00	--	0.99	--	0.00
	SD	--	0.00	--	0.02	--	0.00	--	0.03
	Min	--	0.99	--	–0.05	--	0.99	--	–0.08
	Max	--	1.00	--	0.06	--	1.00	--	0.08
MAD	Mean	--	0.06	--	1.17	--	0.07	--	1.08
	SD	--	0.01	--	0.02	--	0.01	--	0.02
	Min	--	0.04	--	1.12	--	0.05	--	1.03
	Max	--	0.11	--	1.24	--	0.10	--	1.14
MD	Mean	--	0.00	--	0.00	--	0.00	--	0.00
	SD	--	0.04	--	0.03	--	0.03	--	0.01
	Min	--	–0.10	--	–0.08	--	–0.09	--	–0.02
	Max	--	0.11	--	0.09	--	0.08	--	0.02

MAD – Mean absolute distance; MD – Mean distance

Table B-21. Relationship Between Estimated EU and SU 2PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.79	0.99	--	0.00	0.71	0.98	--	0.00
	SD	0.12	0.01	--	0.04	0.14	0.01	--	0.04
	Min	0.09	0.95	--	–0.14	–0.04	0.90	--	–0.15
	Max	0.96	1.00	--	0.14	0.95	1.00	--	0.13
MAD	Mean	0.13	0.11	--	1.01	0.08	0.15	--	0.96
	SD	0.03	0.03	--	0.03	0.02	0.03	--	0.03
	Min	0.07	0.05	--	0.90	0.04	0.05	--	0.86
	Max	0.22	0.27	--	1.12	0.13	0.28	--	1.06
MD	Mean	0.02	0.00	--	0.00	0.00	0.00	--	0.00
	SD	0.05	0.07	--	0.00	0.03	0.08	--	0.00
	Min	–0.16	–0.24	--	–0.01	–0.09	–0.24	--	–0.01
	Max	0.20	0.25	--	0.01	0.09	0.22	--	0.01

MAD – Mean absolute distance; MD – Mean distance

Table B-22. Relationship Between Estimated EU and SU 2PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.84	0.99	--	0.01	0.79	0.98	--	0.00
	SD	0.04	0.00	--	0.04	0.04	0.01	--	0.05
	Min	0.66	0.97	--	–0.11	0.62	0.96	--	–0.12
	Max	0.92	0.99	--	0.11	0.89	0.99	--	0.14
MAD	Mean	0.11	0.11	--	1.15	0.07	0.14	--	1.11
	SD	0.01	0.02	--	0.03	0.01	0.01	--	0.03
	Min	0.08	0.08	--	1.06	0.06	0.10	--	1.00
	Max	0.15	0.21	--	1.23	0.09	0.19	--	1.21
MD	Mean	0.01	0.00	--	0.00	0.00	0.00	--	0.00
	SD	0.03	0.06	--	0.06	0.01	0.05	--	0.02
	Min	–0.08	–0.16	--	–0.16	–0.04	–0.13	--	–0.04
	Max	0.11	0.19	--	0.16	0.03	0.16	--	0.04

Table B-23. Relationship Between Estimated EU and SU 2PL Model Parameters Across Test Replications: N = 1,500, n = 15

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.91	1.00	--	0.00	0.87	0.99	--	0.00
	SD	0.05	0.00	--	0.03	0.07	0.01	--	0.03
	Min	0.56	0.97	--	–0.08	0.51	0.94	--	–0.08
	Max	0.99	1.00	--	0.08	0.99	1.00	--	0.08
MAD	Mean	0.09	0.07	--	1.00	0.06	0.09	--	0.96
	SD	0.02	0.02	--	0.02	0.01	0.02	--	0.02
	Min	0.04	0.03	--	0.94	0.03	0.04	--	0.89
	Max	0.15	0.16	--	1.06	0.10	0.24	--	1.02
MD	Mean	0.01	0.00	--	0.00	0.00	0.00	--	0.00
	SD	0.04	0.04	--	0.00	0.02	0.05	--	0.00
	Min	–0.10	–0.12	--	–0.01	–0.06	–0.16	--	–0.01
	Max	0.15	0.14	--	0.01	0.09	0.20	--	0.01

MAD – Mean absolute distance; MD – Mean distance

Table B-24. Relationship Between Estimated EU and SU 2PL Model Parameters Across Test Replications: N = 1,500, n = 75

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.93	1.00	--	0.00	0.91	0.99	--	0.00
	SD	0.02	0.00	--	0.03	0.02	0.00	--	0.03
	Min	0.86	0.99	--	–0.07	0.84	0.98	--	–0.07
	Max	0.96	1.00	--	0.06	0.95	1.00	--	0.07
MAD	Mean	0.07	0.07	--	1.15	0.05	0.09	--	1.09
	SD	0.01	0.01	--	0.02	0.00	0.01	--	0.02
	Min	0.05	0.04	--	1.10	0.04	0.06	--	1.04
	Max	0.11	0.11	--	1.20	0.07	0.12	--	1.15
MD	Mean	0.01	0.00	--	0.01	0.00	0.00	--	0.00
	SD	0.02	0.04	--	0.03	0.01	0.03	--	0.01
	Min	–0.05	–0.11	--	–0.09	–0.03	–0.08	--	–0.03
	Max	0.08	0.11	--	0.09	0.06	0.08	--	0.03

Table B-25. Relationship Between Estimated EU and SU 3PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.62	0.98	0.36	0.00	0.54	0.96	0.35	0.00
	SD	0.18	0.01	0.25	0.05	0.20	0.03	0.26	0.04
	Min	−0.33	0.82	−0.63	−0.16	−0.63	0.73	−0.56	−0.14
	Max	0.95	1.00	0.90	0.16	0.93	1.00	0.97	0.14
MAD	Mean	0.16	0.14	0.03	0.93	0.11	0.19	0.02	0.88
	SD	0.03	0.04	0.01	0.03	0.02	0.05	0.00	0.03
	Min	0.07	0.07	0.01	0.84	0.04	0.08	0.01	0.78
	Max	0.27	0.33	0.05	1.03	0.18	0.60	0.03	0.97
MD	Mean	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SD	0.05	0.08	0.01	0.01	0.04	0.09	0.01	0.01
	Min	−0.16	−0.22	−0.03	−0.02	−0.12	−0.30	−0.02	−0.02
	Max	0.20	0.30	0.03	0.02	0.12	0.35	0.02	0.02

MAD – Mean absolute distance; MD – Mean distance

Table B-26. Relationship Between Estimated EU and SU 3PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.68	0.98	0.33	0.00	0.63	0.96	0.27	0.00
	SD	0.06	0.01	0.10	0.05	0.07	0.01	0.12	0.04
	Min	0.47	0.95	0.02	−0.11	0.42	0.91	−0.08	−0.11
	Max	0.82	0.99	0.59	0.15	0.81	0.98	0.58	0.14
MAD	Mean	0.15	0.14	0.03	1.06	0.10	0.19	0.02	1.04
	SD	0.01	0.02	0.00	0.03	0.01	0.02	0.00	0.03
	Min	0.12	0.10	0.02	0.97	0.08	0.12	0.02	0.94
	Max	0.20	0.22	0.04	1.18	0.13	0.26	0.03	1.13
MD	Mean	0.01	0.01	0.00	0.00	0.00	0.01	0.00	0.00
	SD	0.03	0.06	0.00	0.04	0.02	0.06	0.00	0.01
	Min	−0.08	−0.15	−0.01	−0.10	−0.04	−0.15	−0.01	−0.03
	Max	0.11	0.20	0.01	0.10	0.05	0.17	0.01	0.02

Table B-27. Relationship Between Estimated EU and SU 3PL Model Parameters Across Test Replications: N = 1,500, n = 15

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.75	0.99	0.41	0.00	0.72	0.98	0.33	0.00
	SD	0.13	0.01	0.23	0.03	0.14	0.01	0.26	0.03
	Min	−0.25	0.94	−0.39	−0.10	0.02	0.89	−0.71	−0.07
	Max	0.97	1.00	0.94	0.09	0.97	1.00	0.88	0.07
MAD	Mean	0.13	0.10	0.03	0.93	0.09	0.13	0.02	0.88
	SD	0.03	0.03	0.01	0.03	0.02	0.03	0.01	0.02
	Min	0.06	0.04	0.01	0.83	0.04	0.06	0.01	0.80
	Max	0.23	0.28	0.05	1.01	0.15	0.27	0.04	0.94
MD	Mean	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SD	0.05	0.05	0.01	0.01	0.03	0.06	0.01	0.01
	Min	−0.14	−0.16	−0.03	−0.02	−0.09	−0.17	−0.03	−0.02
	Max	0.15	0.19	0.03	0.02	0.10	0.21	0.03	0.02

MAD – Mean absolute distance; MD – Mean distance

Table B-28. Relationship Between Estimated EU and SU 3PL Model Parameters Across Test Replications: N = 1,500, n = 75

		High Discrimination				Low Discrimination			
		a	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation	Mean	0.82	0.99	0.43	0.00	0.78	0.98	0.29	0.00
	SD	0.04	0.00	0.10	0.03	0.05	0.01	0.11	0.03
	Min	0.67	0.97	0.09	−0.06	0.54	0.92	−0.07	−0.08
	Max	0.90	0.99	0.68	0.07	0.88	0.99	0.61	0.07
MAD	Mean	0.12	0.10	0.03	1.06	0.08	0.14	0.03	1.04
	SD	0.01	0.01	0.00	0.02	0.01	0.01	0.00	0.02
	Min	0.09	0.07	0.02	1.01	0.06	0.10	0.02	0.99
	Max	0.15	0.14	0.04	1.13	0.10	0.18	0.04	1.09
MD	Mean	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	SD	0.02	0.04	0.00	0.03	0.02	0.04	0.00	0.00
	Min	−0.06	−0.10	−0.01	−0.08	−0.04	−0.11	−0.01	−0.01
	Max	0.07	0.10	0.01	0.08	0.04	0.09	0.01	0.01

Table B-29. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	--	1.00	--	0.92	--	0.99	--	0.87	--	0.89	--	0.84	--	0.88	--	0.79
SD	--	0.00	--	0.01	--	0.00	--	0.01	--	0.06	--	0.01	--	0.07	--	0.02
Min	--	0.98	--	0.89	--	0.94	--	0.84	--	0.49	--	0.77	--	0.55	--	0.72
Max	--	1.00	--	0.94	--	1.00	--	0.90	--	0.99	--	0.87	--	0.99	--	0.83
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	--	0.92	--	0.87	--	0.92	--	0.82
SD	--	--	--	--	--	--	--	--	--	0.05	--	0.01	--	0.05	--	0.01
Min	--	--	--	--	--	--	--	--	--	0.60	--	0.83	--	0.55	--	0.76
Max	--	--	--	--	--	--	--	--	--	0.99	--	0.90	--	0.99	--	0.86
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.07	--	0.33	0.00	0.08	--	0.39	0.02	0.34	--	0.44	0.07	0.33	--	0.49
SD	0.00	0.02	--	0.01	0.00	0.02	--	0.01	0.00	0.07	--	0.02	0.00	0.07	--	0.02
Min	0.00	0.03	--	0.29	0.00	0.04	--	0.34	0.02	0.12	--	0.39	0.07	0.13	--	0.44
Max	0.00	0.17	--	0.39	0.00	0.20	--	0.43	0.02	0.62	--	0.52	0.07	0.55	--	0.54
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.28	0.31	--	0.37	0.22	0.28	--	0.38
SD	--	--	--	--	--	--	--	--	0.00	0.06	--	0.01	0.00	0.05	--	0.01
Min	--	--	--	--	--	--	--	--	0.28	0.14	--	0.32	0.22	0.13	--	0.33
Max	--	--	--	--	--	--	--	--	0.28	0.53	--	0.41	0.22	0.48	--	0.43
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.00	--	0.02	0.00	0.00	--	0.01	–0.02	0.18	--	0.01	–0.07	0.17	--	0.01
SD	0.00	0.05	--	0.05	0.00	0.05	--	0.04	0.00	0.11	--	0.04	0.00	0.11	--	0.04
Min	0.00	–0.17	--	–0.16	0.00	–0.20	--	–0.14	–0.02	–0.16	--	–0.14	–0.07	–0.19	--	–0.16
Max	0.00	0.17	--	0.23	0.00	0.14	--	0.16	–0.02	0.56	--	0.15	–0.07	0.53	--	0.17

Table B-29. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.28	0.01	--	0.01	−0.22	0.00	--	0.01
SD	--	--	--	--	--	--	--	--	0.00	0.10	--	0.03	0.00	0.10	--	0.03
Min	--	--	--	--	--	--	--	--	−0.28	−0.42	--	−0.11	−0.22	−0.28	--	−0.09
Max	--	--	--	--	--	--	--	--	−0.28	0.42	--	0.11	−0.22	0.33	--	0.11

Table B-30. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	--	1.00	--	0.98	--	0.99	--	0.97	--	0.89	--	0.89	--	0.89	--	0.87
SD	--	0.00	--	0.00	--	0.00	--	0.00	--	0.02	--	0.01	--	0.02	--	0.01
Min	--	0.99	--	0.98	--	0.98	--	0.96	--	0.82	--	0.86	--	0.79	--	0.85
Max	--	1.00	--	0.99	--	1.00	--	0.98	--	0.94	--	0.91	--	0.94	--	0.90
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	--	0.93	--	0.93	--	0.92	--	0.91
SD	--	--	--	--	--	--	--	--	--	0.02	--	0.01	--	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	--	0.87	--	0.91	--	0.86	--	0.89
Max	--	--	--	--	--	--	--	--	--	0.96	--	0.94	--	0.96	--	0.93
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.07	--	0.16	0.00	0.08	--	0.20	0.02	0.32	--	0.40	0.07	0.32	--	0.39
SD	0.00	0.02	--	0.01	0.00	0.01	--	0.01	0.00	0.03	--	0.02	0.00	0.03	--	0.01
Min	0.00	0.05	--	0.14	0.00	0.06	--	0.17	0.02	0.24	--	0.36	0.07	0.25	--	0.35
Max	0.00	0.18	--	0.21	0.00	0.11	--	0.22	0.02	0.42	--	0.45	0.07	0.43	--	0.44

Table B-30. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.28	0.33	--	0.38	0.22	0.28	--	0.34
SD	--	--	--	--	--	--	--	--	0.00	0.03	--	0.01	0.00	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.28	0.26	--	0.34	0.22	0.22	--	0.30
Max	--	--	--	--	--	--	--	--	0.28	0.43	--	0.42	0.22	0.36	--	0.37
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.00	--	0.01	0.00	0.00	--	0.01	–0.02	0.08	--	–0.10	–0.07	0.11	--	–0.04
SD	0.00	0.05	--	0.05	0.00	0.04	--	0.03	0.00	0.07	--	0.06	0.00	0.06	--	0.04
Min	0.00	–0.18	--	–0.17	0.00	–0.09	--	–0.07	–0.02	–0.17	--	–0.28	–0.07	–0.07	--	–0.16
Max	0.00	0.13	--	0.14	0.00	0.10	--	0.11	–0.02	0.27	--	0.06	–0.07	0.27	--	0.09
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	–0.28	–0.08	--	–0.10	–0.22	–0.04	--	–0.04
SD	--	--	--	--	--	--	--	--	0.00	0.05	--	0.05	0.00	0.05	--	0.03
Min	--	--	--	--	--	--	--	--	–0.28	–0.24	--	–0.24	–0.22	–0.17	--	–0.11
Max	--	--	--	--	--	--	--	--	–0.28	0.05	--	0.05	–0.22	0.07	--	0.03

Table B-31. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	--	1.00	--	0.92	--	1.00	--	0.87	--	0.88	--	0.84	--	0.88	--	0.79
SD	--	0.00	--	0.00	--	0.00	--	0.01	--	0.07	--	0.01	--	0.07	--	0.01
Min	--	0.99	--	0.90	--	0.99	--	0.85	--	0.44	--	0.81	--	0.40	--	0.76
Max	--	1.00	--	0.94	--	1.00	--	0.89	--	0.99	--	0.86	--	0.98	--	0.83
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	--	0.92	--	0.87	--	0.92	--	0.82
SD	--	--	--	--	--	--	--	--	--	0.04	--	0.01	--	0.04	--	0.01
Min	--	--	--	--	--	--	--	--	--	0.62	--	0.85	--	0.66	--	0.79
Max	--	--	--	--	--	--	--	--	--	0.99	--	0.89	--	0.99	--	0.85
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.04	--	0.33	0.00	0.05	--	0.39	0.02	0.34	--	0.44	0.07	0.33	--	0.49
SD	0.00	0.01	--	0.01	0.00	0.01	--	0.01	0.00	0.07	--	0.01	0.00	0.07	--	0.01
Min	0.00	0.02	--	0.30	0.00	0.02	--	0.36	0.02	0.14	--	0.41	0.07	0.17	--	0.45
Max	0.00	0.10	--	0.37	0.00	0.10	--	0.42	0.02	0.60	--	0.48	0.07	0.57	--	0.53
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.28	0.31	--	0.37	0.22	0.27	--	0.38
SD	--	--	--	--	--	--	--	--	0.00	0.06	--	0.01	0.00	0.05	--	0.01
Min	--	--	--	--	--	--	--	--	0.28	0.14	--	0.34	0.22	0.13	--	0.35
Max	--	--	--	--	--	--	--	--	0.28	0.52	--	0.40	0.22	0.47	--	0.41
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.00	--	0.02	0.00	0.00	--	0.01	–0.02	0.18	--	0.01	–0.07	0.16	--	0.01
SD	0.00	0.03	--	0.03	0.00	0.03	--	0.03	0.00	0.10	--	0.03	0.00	0.10	--	0.03
Min	0.00	–0.10	--	–0.09	0.00	–0.08	--	–0.06	–0.02	–0.22	--	–0.08	–0.07	–0.23	--	–0.07
Max	0.00	0.09	--	0.11	0.00	0.09	--	0.09	–0.02	0.52	--	0.12	–0.07	0.50	--	0.09

Table B-31. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.28	0.01	--	0.01	−0.22	0.00	--	0.01
SD	--	--	--	--	--	--	--	--	0.00	0.09	--	0.02	0.00	0.09	--	0.02
Min	--	--	--	--	--	--	--	--	−0.28	−0.29	--	−0.06	−0.22	−0.36	--	−0.06
Max	--	--	--	--	--	--	--	--	−0.28	0.31	--	0.09	−0.22	0.26	--	0.07

Table B-32. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	--	1.00	--	0.98	--	1.00	--	0.97	--	0.90	--	0.89	--	0.90	--	0.88
SD	--	0.00	--	0.00	--	0.00	--	0.00	--	0.02	--	0.01	--	0.02	--	0.01
Min	--	1.00	--	0.98	--	0.99	--	0.96	--	0.79	--	0.87	--	0.79	--	0.86
Max	--	1.00	--	0.98	--	1.00	--	0.97	--	0.94	--	0.90	--	0.94	--	0.89
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	--	0.93	--	0.93	--	0.93	--	0.91
SD	--	--	--	--	--	--	--	--	--	0.02	--	0.00	--	0.02	--	0.00
Min	--	--	--	--	--	--	--	--	--	0.87	--	0.91	--	0.88	--	0.90
Max	--	--	--	--	--	--	--	--	--	0.97	--	0.94	--	0.97	--	0.92
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.04	--	0.16	0.00	0.05	--	0.19	0.02	0.32	--	0.39	0.07	0.31	--	0.39
SD	0.00	0.01	--	0.00	0.00	0.01	--	0.00	0.00	0.03	--	0.01	0.00	0.03	--	0.01
Min	0.00	0.03	--	0.15	0.00	0.03	--	0.18	0.02	0.25	--	0.37	0.07	0.25	--	0.37
Max	0.00	0.08	--	0.18	0.00	0.07	--	0.21	0.02	0.40	--	0.43	0.07	0.39	--	0.41

Table B-32. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 1PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.28	0.32	--	0.38	0.22	0.28	--	0.34
SD	--	--	--	--	--	--	--	--	0.00	0.02	--	0.01	0.00	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.28	0.26	--	0.35	0.22	0.22	--	0.32
Max	--	--	--	--	--	--	--	--	0.28	0.42	--	0.41	0.22	0.34	--	0.36
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.00	--	0.01	0.00	0.00	--	0.01	−0.02	0.09	--	−0.09	−0.07	0.12	--	−0.05
SD	0.00	0.03	--	0.03	0.00	0.02	--	0.02	0.00	0.06	--	0.04	0.00	0.05	--	0.03
Min	0.00	−0.07	--	−0.07	0.00	−0.05	--	−0.04	−0.02	−0.06	--	−0.21	−0.07	−0.04	--	−0.13
Max	0.00	0.08	--	0.09	0.00	0.06	--	0.07	−0.02	0.26	--	0.03	−0.07	0.26	--	0.04
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.28	−0.08	--	−0.09	−0.22	−0.05	--	−0.04
SD	--	--	--	--	--	--	--	--	0.00	0.04	--	0.04	0.00	0.03	--	0.02
Min	--	--	--	--	--	--	--	--	−0.28	−0.17	--	−0.20	−0.22	−0.15	--	−0.11
Max	--	--	--	--	--	--	--	--	−0.28	0.00	--	0.02	−0.22	0.05	--	0.02

Table B-33. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.89	0.99	--	0.92	0.84	0.99	--	0.87	0.87	0.84	--	0.83	0.83	0.84	--	0.79
SD	0.06	0.00	--	0.01	0.09	0.01	--	0.01	0.08	0.10	--	0.02	0.10	0.10	--	0.03
Min	0.48	0.98	--	0.88	0.24	0.92	--	0.81	0.31	0.16	--	0.75	-0.01	0.17	--	0.64
Max	0.99	1.00	--	0.94	0.98	1.00	--	0.90	0.99	0.99	--	0.89	0.98	0.99	--	0.86
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.87	0.91	--	0.87	0.83	0.90	--	0.82
SD	--	--	--	--	--	--	--	--	0.08	0.06	--	0.01	0.10	0.06	--	0.02
Min	--	--	--	--	--	--	--	--	0.30	0.41	--	0.81	0.01	0.43	--	0.75
Max	--	--	--	--	--	--	--	--	0.98	0.99	--	0.91	0.98	0.99	--	0.86
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.09	0.08	--	0.31	0.07	0.11	--	0.39	0.13	0.36	--	0.44	0.09	0.37	--	0.49
SD	0.02	0.02	--	0.02	0.01	0.03	--	0.02	0.03	0.08	--	0.03	0.02	0.08	--	0.03
Min	0.03	0.02	--	0.26	0.03	0.05	--	0.34	0.05	0.10	--	0.36	0.05	0.12	--	0.39
Max	0.17	0.18	--	0.38	0.11	0.20	--	0.47	0.26	0.67	--	0.54	0.16	0.73	--	0.61
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.33	0.30	--	0.35	0.21	0.31	--	0.39
SD	--	--	--	--	--	--	--	--	0.04	0.07	--	0.02	0.02	0.07	--	0.02
Min	--	--	--	--	--	--	--	--	0.20	0.13	--	0.30	0.15	0.15	--	0.32
Max	--	--	--	--	--	--	--	--	0.50	0.54	--	0.42	0.27	0.55	--	0.44
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	0.00	--	0.02	0.00	0.00	--	0.01	–0.07	0.17	--	0.01	–0.05	0.17	--	0.01
SD	0.04	0.05	--	0.05	0.02	0.06	--	0.04	0.05	0.12	--	0.05	0.03	0.12	--	0.05
Min	–0.13	–0.14	--	–0.16	–0.09	–0.19	--	–0.15	–0.21	–0.28	--	–0.15	–0.15	–0.25	--	–0.14
Max	0.12	0.17	--	0.17	0.08	0.17	--	0.14	0.07	0.64	--	0.18	0.06	0.56	--	0.17

Table B-33. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.33	0.02	--	0.01	−0.21	0.02	--	0.01
SD	--	--	--	--	--	--	--	--	0.04	0.11	--	0.03	0.02	0.11	--	0.03
Min	--	--	--	--	--	--	--	--	−0.50	−0.34	--	−0.09	−0.27	−0.34	--	−0.13
Max	--	--	--	--	--	--	--	--	−0.19	0.36	--	0.13	−0.14	0.37	--	0.10

Table B-34. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.92	0.99	--	0.98	0.89	0.99	--	0.97	0.90	0.85	--	0.89	0.88	0.85	--	0.88
SD	0.02	0.00	--	0.00	0.02	0.00	--	0.00	0.02	0.04	--	0.01	0.03	0.04	--	0.01
Min	0.84	0.98	--	0.98	0.81	0.96	--	0.96	0.81	0.56	--	0.86	0.76	0.56	--	0.84
Max	0.96	1.00	--	0.99	0.95	0.99	--	0.98	0.95	0.93	--	0.92	0.93	0.93	--	0.91
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.89	0.91	--	0.92	0.87	0.91	--	0.91
SD	--	--	--	--	--	--	--	--	0.02	0.02	--	0.01	0.02	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.82	0.84	--	0.90	0.79	0.83	--	0.88
Max	--	--	--	--	--	--	--	--	0.94	0.96	--	0.94	0.93	0.95	--	0.93
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.08	0.08	--	0.16	0.06	0.10	--	0.20	0.11	0.34	--	0.39	0.08	0.35	--	0.39
SD	0.01	0.01	--	0.01	0.01	0.01	--	0.01	0.01	0.04	--	0.02	0.01	0.04	--	0.02
Min	0.06	0.05	--	0.14	0.04	0.07	--	0.18	0.09	0.26	--	0.35	0.06	0.26	--	0.34
Max	0.10	0.15	--	0.20	0.07	0.14	--	0.23	0.15	0.44	--	0.45	0.11	0.46	--	0.45

Table B-34. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.29	0.33	--	0.38	0.20	0.31	--	0.36
SD	--	--	--	--	--	--	--	--	0.02	0.03	--	0.02	0.01	0.03	--	0.02
Min	--	--	--	--	--	--	--	--	0.22	0.26	--	0.33	0.17	0.23	--	0.31
Max	--	--	--	--	--	--	--	--	0.36	0.40	--	0.44	0.22	0.39	--	0.40
MD Between ξ_1 and $\hat{\xi}$																
Mean	−0.01	0.00	--	0.01	0.01	0.01	--	0.01	−0.03	0.07	--	−0.11	−0.04	0.13	--	−0.04
SD	0.02	0.05	--	0.05	0.01	0.04	--	0.04	0.03	0.07	--	0.06	0.02	0.06	--	0.04
Min	−0.07	−0.15	--	−0.13	−0.02	−0.09	--	−0.08	−0.13	−0.14	--	−0.26	−0.09	−0.03	--	−0.15
Max	0.06	0.11	--	0.11	0.05	0.11	--	0.11	0.05	0.30	--	0.09	0.01	0.29	--	0.05
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.29	−0.09	--	−0.11	−0.20	−0.03	--	−0.04
SD	--	--	--	--	--	--	--	--	0.02	0.05	--	0.05	0.01	0.04	--	0.03
Min	--	--	--	--	--	--	--	--	−0.36	−0.24	--	−0.28	−0.22	−0.16	--	−0.13
Max	--	--	--	--	--	--	--	--	−0.21	0.06	--	0.02	−0.17	0.08	--	0.04

Table B-35. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.96	1.00	--	0.92	0.94	1.00	--	0.87	0.94	0.85	--	0.84	0.92	0.85	--	0.79
SD	0.02	0.00	--	0.01	0.04	0.00	--	0.01	0.04	0.10	--	0.02	0.05	0.10	--	0.02
Min	0.79	0.99	--	0.89	0.73	0.97	--	0.83	0.73	0.09	--	0.74	0.63	0.13	--	0.69
Max	0.99	1.00	--	0.94	0.99	1.00	--	0.91	0.99	0.99	--	0.89	0.99	0.98	--	0.86
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.93	0.92	--	0.87	0.92	0.91	--	0.82
SD	--	--	--	--	--	--	--	--	0.04	0.06	--	0.01	0.05	0.06	--	0.01
Min	--	--	--	--	--	--	--	--	0.64	0.05	--	0.83	0.60	0.10	--	0.77
Max	--	--	--	--	--	--	--	--	0.99	0.99	--	0.90	1.00	0.99	--	0.85
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.06	0.05	--	0.31	0.04	0.06	--	0.39	0.11	0.35	--	0.44	0.07	0.35	--	0.49
SD	0.01	0.01	--	0.01	0.01	0.02	--	0.02	0.02	0.08	--	0.03	0.02	0.08	--	0.03
Min	0.03	0.02	--	0.27	0.02	0.03	--	0.34	0.05	0.14	--	0.36	0.02	0.15	--	0.41
Max	0.11	0.11	--	0.36	0.07	0.15	--	0.45	0.23	0.72	--	0.53	0.13	0.71	--	0.59
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.34	0.29	--	0.35	0.21	0.29	--	0.39
SD	--	--	--	--	--	--	--	--	0.04	0.06	--	0.02	0.02	0.06	--	0.01
Min	--	--	--	--	--	--	--	--	0.24	0.13	--	0.30	0.15	0.13	--	0.35
Max	--	--	--	--	--	--	--	--	0.47	0.55	--	0.41	0.28	0.50	--	0.43
MD Between ξ_1 and $\hat{\xi}$																
Mean	−0.01	0.00	--	0.02	0.00	0.00	--	0.01	−0.08	0.17	--	0.01	−0.06	0.18	--	0.01
SD	0.03	0.03	--	0.03	0.02	0.03	--	0.03	0.03	0.12	--	0.03	0.02	0.12	--	0.03
Min	−0.09	−0.11	--	−0.08	−0.06	−0.12	--	−0.07	−0.17	−0.21	--	−0.09	−0.12	−0.23	--	−0.09
Max	0.09	0.09	--	0.12	0.05	0.11	--	0.09	0.01	0.65	--	0.10	0.02	0.61	--	0.09

Table B-35. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.34	0.01	--	0.01	−0.21	0.01	--	0.01
SD	--	--	--	--	--	--	--	--	0.04	0.10	--	0.02	0.02	0.10	--	0.02
Min	--	--	--	--	--	--	--	--	−0.47	−0.34	--	−0.07	−0.28	−0.36	--	−0.06
Max	--	--	--	--	--	--	--	--	−0.23	0.30	--	0.10	−0.15	0.29	--	0.07

Table B-36. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.97	1.00	--	0.98	0.96	1.00	--	0.97	0.95	0.85	--	0.89	0.95	0.85	--	0.88
SD	0.01	0.00	--	0.00	0.01	0.00	--	0.00	0.01	0.04	--	0.01	0.01	0.04	--	0.01
Min	0.94	0.99	--	0.98	0.93	0.99	--	0.96	0.90	0.69	--	0.85	0.91	0.67	--	0.85
Max	0.98	1.00	--	0.98	0.98	1.00	--	0.97	0.98	0.93	--	0.91	0.97	0.93	--	0.91
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.95	0.92	--	0.92	0.94	0.91	--	0.91
SD	--	--	--	--	--	--	--	--	0.01	0.02	--	0.01	0.01	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.90	0.85	--	0.90	0.88	0.83	--	0.89
Max	--	--	--	--	--	--	--	--	0.97	0.97	--	0.94	0.97	0.97	--	0.92
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.05	0.05	--	0.16	0.04	0.06	--	0.20	0.09	0.34	--	0.39	0.07	0.34	--	0.39
SD	0.01	0.01	--	0.01	0.00	0.01	--	0.01	0.01	0.04	--	0.02	0.01	0.03	--	0.01
Min	0.04	0.03	--	0.15	0.03	0.04	--	0.18	0.06	0.25	--	0.34	0.04	0.26	--	0.34
Max	0.07	0.09	--	0.18	0.05	0.08	--	0.21	0.13	0.46	--	0.45	0.09	0.46	--	0.43

Table B-36. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 2PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.29	0.33	--	0.38	0.20	0.30	--	0.35
SD	--	--	--	--	--	--	--	--	0.02	0.03	--	0.02	0.01	0.03	--	0.01
Min	--	--	--	--	--	--	--	--	0.24	0.22	--	0.32	0.17	0.21	--	0.32
Max	--	--	--	--	--	--	--	--	0.35	0.40	--	0.42	0.23	0.38	--	0.38
MD Between ξ_1 and $\hat{\xi}$																
Mean	−0.01	0.00	--	0.01	0.01	0.00	--	0.01	−0.03	0.08	--	−0.11	−0.05	0.13	--	−0.05
SD	0.02	0.03	--	0.03	0.01	0.02	--	0.02	0.02	0.06	--	0.05	0.01	0.05	--	0.03
Min	−0.06	−0.09	--	−0.07	−0.02	−0.05	--	−0.04	−0.11	−0.11	--	−0.24	−0.09	−0.03	--	−0.14
Max	0.03	0.08	--	0.08	0.03	0.06	--	0.06	0.02	0.28	--	0.05	−0.02	0.27	--	0.05
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.29	−0.09	--	−0.11	−0.20	−0.04	--	−0.05
SD	--	--	--	--	--	--	--	--	0.02	0.04	--	0.05	0.01	0.04	--	0.02
Min	--	--	--	--	--	--	--	--	−0.35	−0.19	--	−0.24	−0.23	−0.14	--	−0.11
Max	--	--	--	--	--	--	--	--	−0.24	0.00	--	0.05	−0.17	0.06	--	0.03

Table B-37. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.70	0.99	0.40	0.88	0.69	0.97	0.27	0.82	0.66	0.85	0.39	0.79	0.66	0.83	0.26	0.74
SD	0.16	0.01	0.22	0.01	0.15	0.02	0.24	0.02	0.18	0.10	0.23	0.03	0.18	0.10	0.25	0.03
Min	-0.30	0.87	-0.29	0.80	0.01	0.76	-0.56	0.75	-0.26	0.17	-0.45	0.69	-0.24	0.12	-0.77	0.61
Max	0.97	1.00	0.90	0.91	0.97	1.00	0.84	0.87	0.95	0.98	0.88	0.86	0.96	0.98	0.86	0.83
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.66	0.91	--	0.83	0.66	0.90	--	0.77
SD	--	--	--	--	--	--	--	--	0.18	0.06	--	0.02	0.17	0.06	--	0.02
Min	--	--	--	--	--	--	--	--	-0.19	0.60	--	0.76	-0.19	0.60	--	0.70
Max	--	--	--	--	--	--	--	--	0.97	0.99	--	0.89	0.95	0.98	--	0.84
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.14	0.11	0.03	0.38	0.09	0.15	0.03	0.46	0.18	0.34	0.03	0.48	0.12	0.35	0.03	0.54
SD	0.03	0.03	0.01	0.02	0.02	0.04	0.00	0.02	0.04	0.08	0.01	0.03	0.02	0.08	0.01	0.03
Min	0.06	0.05	0.01	0.31	0.04	0.07	0.01	0.40	0.08	0.14	0.01	0.40	0.04	0.12	0.01	0.45
Max	0.24	0.30	0.05	0.45	0.19	0.50	0.04	0.55	0.36	0.63	0.05	0.60	0.19	0.71	0.05	0.66
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.34	0.30	--	0.37	0.22	0.32	--	0.41
SD	--	--	--	--	--	--	--	--	0.04	0.07	--	0.02	0.03	0.07	--	0.02
Min	--	--	--	--	--	--	--	--	0.21	0.12	--	0.31	0.13	0.12	--	0.36
Max	--	--	--	--	--	--	--	--	0.47	0.63	--	0.43	0.34	0.56	--	0.46
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.01	0.01	0.00	−0.06	0.00	0.00	0.00	−0.05	−0.06	0.06	0.00	−0.05	−0.06	0.05	0.00	−0.05
SD	0.05	0.06	0.01	0.05	0.03	0.07	0.01	0.05	0.06	0.12	0.01	0.05	0.04	0.13	0.01	0.05
Min	−0.13	−0.23	−0.02	−0.23	−0.10	−0.23	−0.03	−0.21	−0.26	−0.32	−0.03	−0.20	−0.19	−0.47	−0.03	−0.21
Max	0.19	0.18	0.03	0.10	0.12	0.34	0.03	0.13	0.15	0.42	0.03	0.11	0.05	0.57	0.03	0.11

Table B-37. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: $N = 500$, $n = 15$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.33	−0.10	--	−0.05	−0.22	−0.11	--	−0.05
SD	--	--	--	--	--	--	--	--	0.04	0.10	--	0.04	0.03	0.12	--	0.04
Min	--	--	--	--	--	--	--	--	−0.47	−0.42	--	−0.20	−0.34	−0.42	--	−0.19
Max	--	--	--	--	--	--	--	--	−0.16	0.23	--	0.07	−0.13	0.29	--	0.06

Table B-38. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.77	0.99	0.44	0.97	0.77	0.97	0.30	0.95	0.74	0.86	0.43	0.87	0.75	0.85	0.30	0.86
SD	0.05	0.00	0.09	0.00	0.05	0.01	0.11	0.01	0.07	0.04	0.10	0.01	0.06	0.04	0.10	0.01
Min	0.60	0.97	0.16	0.96	0.61	0.95	0.01	0.93	0.47	0.66	0.12	0.82	0.57	0.65	-0.02	0.81
Max	0.89	0.99	0.62	0.98	0.89	0.99	0.55	0.96	0.87	0.93	0.68	0.91	0.87	0.92	0.53	0.89
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.74	0.91	--	0.91	0.75	0.90	--	0.89
SD	--	--	--	--	--	--	--	--	0.06	0.02	--	0.01	0.05	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.51	0.83	--	0.88	0.53	0.83	--	0.87
Max	--	--	--	--	--	--	--	--	0.87	0.96	--	0.93	0.88	0.96	--	0.92
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.13	0.11	0.03	0.20	0.08	0.14	0.03	0.25	0.17	0.35	0.03	0.39	0.11	0.35	0.03	0.41
SD	0.01	0.02	0.00	0.01	0.01	0.02	0.00	0.01	0.02	0.04	0.00	0.02	0.01	0.03	0.00	0.02
Min	0.10	0.08	0.02	0.18	0.06	0.10	0.02	0.22	0.12	0.26	0.02	0.34	0.08	0.26	0.02	0.37
Max	0.17	0.16	0.03	0.24	0.11	0.19	0.03	0.29	0.21	0.46	0.03	0.45	0.14	0.45	0.03	0.48

Table B-38. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: $N = 500$, $n = 75$

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.34	0.30	--	0.33	0.22	0.32	--	0.34
SD	--	--	--	--	--	--	--	--	0.02	0.03	--	0.02	0.01	0.03	--	0.01
Min	--	--	--	--	--	--	--	--	0.28	0.21	--	0.28	0.19	0.23	--	0.31
Max	--	--	--	--	--	--	--	--	0.40	0.41	--	0.37	0.26	0.41	--	0.38
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.01	−0.04	0.00	−0.07	0.00	−0.02	0.00	−0.04	−0.06	0.12	0.00	0.05	−0.06	0.08	0.00	0.01
SD	0.03	0.04	0.00	0.04	0.02	0.04	0.00	0.04	0.03	0.07	0.00	0.05	0.02	0.06	0.00	0.04
Min	−0.08	−0.15	−0.01	−0.18	−0.04	−0.14	−0.01	−0.16	−0.17	−0.07	−0.01	−0.10	−0.11	−0.08	−0.01	−0.13
Max	0.08	0.10	0.01	0.06	0.06	0.09	0.01	0.05	0.02	0.31	0.01	0.18	−0.01	0.23	0.01	0.11
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.33	−0.05	--	0.04	−0.22	−0.09	--	0.01
SD	--	--	--	--	--	--	--	--	0.02	0.05	--	0.04	0.01	0.06	--	0.03
Min	--	--	--	--	--	--	--	--	−0.40	−0.18	--	−0.08	−0.26	−0.30	--	−0.07
Max	--	--	--	--	--	--	--	--	−0.28	0.05	--	0.16	−0.19	0.05	--	0.09

Table B-39. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.82	0.99	0.52	0.88	0.82	0.98	0.34	0.82	0.79	0.85	0.51	0.79	0.80	0.84	0.33	0.74
SD	0.10	0.01	0.22	0.01	0.11	0.01	0.24	0.02	0.13	0.10	0.21	0.02	0.12	0.10	0.24	0.03
Min	0.31	0.93	-0.41	0.82	0.05	0.87	-0.57	0.75	0.03	-0.02	-0.23	0.69	-0.16	0.03	-0.45	0.62
Max	0.98	1.00	0.94	0.91	0.98	1.00	0.91	0.86	0.98	0.98	0.93	0.85	0.98	0.99	0.89	0.81
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.79	0.91	--	0.83	0.80	0.91	--	0.77
SD	--	--	--	--	--	--	--	--	0.13	0.06	--	0.01	0.12	0.06	--	0.02
Min	--	--	--	--	--	--	--	--	0.02	0.47	--	0.76	-0.25	0.45	--	0.69
Max	--	--	--	--	--	--	--	--	0.98	0.99	--	0.87	0.98	0.99	--	0.81
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.11	0.08	0.02	0.38	0.07	0.11	0.03	0.46	0.16	0.33	0.02	0.48	0.10	0.34	0.03	0.54
SD	0.02	0.02	0.01	0.02	0.01	0.03	0.01	0.02	0.03	0.08	0.01	0.03	0.02	0.08	0.01	0.02
Min	0.05	0.03	0.01	0.32	0.03	0.05	0.01	0.40	0.07	0.12	0.01	0.41	0.04	0.12	0.01	0.47
Max	0.19	0.28	0.04	0.48	0.12	0.22	0.04	0.53	0.29	0.69	0.05	0.58	0.17	0.66	0.05	0.64
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.35	0.30	--	0.37	0.23	0.31	--	0.41
SD	--	--	--	--	--	--	--	--	0.04	0.07	--	0.02	0.02	0.07	--	0.01
Min	--	--	--	--	--	--	--	--	0.22	0.12	--	0.32	0.14	0.13	--	0.37
Max	--	--	--	--	--	--	--	--	0.48	0.57	--	0.42	0.31	0.58	--	0.46
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	–0.01	0.00	–0.07	0.00	–0.01	0.00	–0.06	–0.07	0.04	0.00	–0.06	–0.06	0.04	0.00	–0.05
SD	0.04	0.04	0.01	0.04	0.02	0.04	0.01	0.03	0.05	0.11	0.01	0.03	0.03	0.12	0.01	0.03
Min	–0.12	–0.25	–0.02	–0.24	–0.09	–0.17	–0.03	–0.17	–0.23	–0.43	–0.03	–0.17	–0.14	–0.37	–0.03	–0.16
Max	0.14	0.12	0.03	0.05	0.07	0.11	0.03	0.04	0.08	0.44	0.03	0.05	0.03	0.53	0.03	0.05

Table B-39. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: N = 1,500, n = 15

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.34	−0.12	--	−0.06	−0.22	−0.12	--	−0.05
SD	--	--	--	--	--	--	--	--	0.04	0.10	--	0.03	0.02	0.10	--	0.02
Min	--	--	--	--	--	--	--	--	−0.48	−0.45	--	−0.17	−0.31	−0.47	--	−0.14
Max	--	--	--	--	--	--	--	--	−0.20	0.22	--	0.03	−0.14	0.30	--	0.03

Table B-40. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
Correlation Between ξ_1 and $\hat{\xi}$																
Mean	0.89	0.99	0.60	0.97	0.87	0.99	0.41	0.95	0.86	0.85	0.59	0.87	0.86	0.84	0.40	0.86
SD	0.03	0.00	0.08	0.00	0.03	0.01	0.10	0.00	0.04	0.04	0.09	0.01	0.04	0.04	0.10	0.01
Min	0.78	0.98	0.35	0.96	0.75	0.93	0.10	0.94	0.67	0.70	0.26	0.83	0.66	0.65	0.13	0.82
Max	0.94	1.00	0.79	0.97	0.93	0.99	0.67	0.96	0.93	0.93	0.76	0.90	0.93	0.93	0.65	0.88
Correlation Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.86	0.92	--	0.91	0.85	0.91	--	0.89
SD	--	--	--	--	--	--	--	--	0.04	0.02	--	0.01	0.04	0.02	--	0.01
Min	--	--	--	--	--	--	--	--	0.71	0.83	--	0.89	0.70	0.82	--	0.88
Max	--	--	--	--	--	--	--	--	0.93	0.97	--	0.94	0.92	0.96	--	0.91
MAD Between ξ_1 and $\hat{\xi}$																
Mean	0.09	0.08	0.02	0.20	0.06	0.11	0.03	0.24	0.14	0.34	0.02	0.39	0.09	0.34	0.03	0.41
SD	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.02	0.04	0.00	0.02	0.01	0.03	0.00	0.01
Min	0.07	0.05	0.02	0.18	0.05	0.08	0.02	0.22	0.10	0.25	0.02	0.35	0.07	0.25	0.02	0.37
Max	0.12	0.13	0.03	0.24	0.08	0.15	0.03	0.27	0.18	0.44	0.03	0.44	0.12	0.44	0.03	0.46

Table B-40. Relationship Between Generating (ξ) and Estimated ($\hat{\xi}$) 3PL Model Parameters Across Test Replications: N = 1,500, n = 75

	SU High Discrimination				SU low Discrimination				EU High Discrimination				EU Low Discrimination			
	a	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ	*a*	*b*	*c*	θ
MAD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	0.34	0.29	--	0.33	0.23	0.31	--	0.34
SD	--	--	--	--	--	--	--	--	0.02	0.03	--	0.01	0.01	0.03	--	0.01
Min	--	--	--	--	--	--	--	--	0.29	0.23	--	0.29	0.20	0.23	--	0.31
Max	--	--	--	--	--	--	--	--	0.40	0.39	--	0.36	0.26	0.41	--	0.37
MD Between ξ_1 and $\hat{\xi}$																
Mean	0.00	−0.05	0.00	−0.08	0.00	−0.04	0.00	−0.05	−0.07	0.10	0.00	0.03	−0.07	0.06	0.00	0.00
SD	0.02	0.03	0.00	0.03	0.01	0.03	0.00	0.02	0.02	0.06	0.00	0.04	0.01	0.06	0.00	0.03
Min	−0.05	−0.12	−0.01	−0.17	−0.05	−0.11	−0.02	−0.12	−0.15	−0.07	−0.01	−0.09	−0.10	−0.09	−0.01	−0.10
Max	0.06	0.06	0.01	0.02	0.03	0.07	0.01	0.02	0.00	0.28	0.01	0.15	−0.03	0.22	0.01	0.07
MD Between $(\xi_1 + \xi_2)/2$ and $\hat{\xi}$																
Mean	--	--	--	--	--	--	--	--	−0.34	−0.06	--	0.03	−0.23	−0.10	--	0.00
SD	--	--	--	--	--	--	--	--	0.02	0.04	--	0.03	0.01	0.05	--	0.02
Min	--	--	--	--	--	--	--	--	−0.40	−0.19	--	−0.06	−0.26	−0.23	--	−0.06
Max	--	--	--	--	--	--	--	--	−0.28	0.03	--	0.12	−0.20	0.05	--	0.06

APPENDIX C: ANALYSIS OF TYPE I ERROR RATES

Sums of Squares and Effect Sizes for Study Factors on Type I Error Rates

Average Type I Error Rates by Fit Statistic and Discrimination Conditions

Average Type I Error Rates by Fit Statistic and Model Conditions

Average Type I Error Rates by Fit Statistic, *N*, and *n* Conditions

DN Bias by Study Condition for *Q*, LM, and *z* Statistics at $\alpha = 0.01$ and 0.10

Type I Error Rates at $\alpha = 0.01$, 0.05, and 0.10 for all EU Conditions

Table C-1. Sums of Squares and Effect Sizes for Study Factors on Type I Error Rates

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
	SS	η^2	SS	η^2	SS	η^2
Main Effects						
Fit Statistic (FS)	5.7779	0.1970	8.6115	0.2311	9.6266	0.2450
Discrimination (D)	0.0004	0.0000	0.0051	0.0001	0.0038	0.0001
Data Noise (DN)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Model (M)	0.6126	0.0209	0.8380	0.0225	0.9909	0.0252
Sample Size (*N*)	1.0235	0.0349	0.4761	0.0128	0.2904	0.0074
Test Length (*n*)	6.4998	0.2217	10.4955	0.2816	11.6854	0.2974
2-way Interactions						
FS × D	0.5372	0.0183	0.4272	0.0115	0.3849	0.0098
FS × DN	0.0001	0.0000	0.0001	0.0000	0.0002	0.0000
FS × M	2.2884	0.0780	2.2767	0.0611	2.1412	0.0545
FS × *N*	2.1665	0.0739	1.4552	0.0390	1.1174	0.0284
FS × *n*	7.0748	0.2413	11.0728	0.2971	12.1892	0.3102
D × DN	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
D × M	0.0318	0.0011	0.0189	0.0005	0.0181	0.0005
D × *N*	0.0001	0.0000	0.0001	0.0000	0.0011	0.0000
D × *n*	0.1081	0.0037	0.0441	0.0012	0.0107	0.0003
DN × M	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000
DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
M × *N*	0.0094	0.0003	0.0300	0.0008	0.0346	0.0009
M × *n*	0.0088	0.0003	0.0129	0.0003	0.0359	0.0009
N × *n*	1.4165	0.0483	0.5824	0.0156	0.2149	0.0055
3-way Interactions						
FS × D × DN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FS × D × M	0.0282	0.0010	0.0287	0.0008	0.0262	0.0007
FS × D × *N*	0.0403	0.0014	0.0259	0.0007	0.0173	0.0004
FS × D × *n*	0.0954	0.0033	0.0397	0.0011	0.0158	0.0004
FS × DN × M	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000
FS × DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
FS × DN × *n*	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000
FS × M × *N*	0.0439	0.0015	0.0539	0.0014	0.0448	0.0011
FS × M × *n*	0.0572	0.0020	0.0305	0.0008	0.0585	0.0015
FS × *N* × *n*	1.4155	0.0483	0.6387	0.0171	0.2749	0.0070
D × DN × M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
D × DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × M × *N*	0.0046	0.0002	0.0010	0.0000	0.0001	0.0000

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
	SS	η^2	SS	η^2	SS	η^2
D × M × *n*	0.0156	0.0005	0.0093	0.0003	0.0038	0.0001
D × *N* × *n*	0.0114	0.0004	0.0167	0.0004	0.0195	0.0005
DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
DN × M × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
M × *N* × *n*	0.0024	0.0001	0.0051	0.0001	0.0046	0.0001
4-way Interactions						
FS × D × DN × M	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
FS × D × DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
FS × D × DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FS × D × M × *N*	0.0062	0.0002	0.0028	0.0001	0.0030	0.0001
FS × D × M × *n*	0.0087	0.0003	0.0052	0.0001	0.0063	0.0002
FS × D × *N* × *n*	0.0155	0.0005	0.0141	0.0004	0.0185	0.0005
FS × DN × M × *N*	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000
FS × DN × M × *n*	0.0001	0.0000	0.0001	0.0000	0.0002	0.0000
FS × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FS × M × *N* × *n*	0.0160	0.0005	0.0494	0.0013	0.0475	0.0012
D × DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × DN × M × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × M × *N* × *n*	0.0015	0.0001	0.0001	0.0000	0.0003	0.0000
DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5-way Interactions						
FS × D × DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
FS × D × DN × M × *n*	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000
FS × D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
FS × D × M × *N* × *n*	0.0051	0.0002	0.0018	0.0000	0.0022	0.0001
FS × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
D × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000
6-way Interaction						
FS × D × DN × M × *N* × *n*	0.0000	0.0000	0.0001	0.0000	0.0001	0.0000
TOTAL WS	19.5776	0.6676	24.7350	0.6637	25.9756	0.6611
TOTAL BS	9.7467	0.3324	12.5357	0.3363	13.3146	0.3389
TOTAL	29.3243		37.2707		39.2902	

Table C-2. Average Type I Error Rates by Fit Statistic and Discrimination Conditions

α	D	Fit Statistic								Overall
		$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO	
0.01	High	0.242	0.011	0.407	0.118	0.065	0.253	0.261	0.220	0.197
	Low	0.300	0.011	0.277	0.035	0.011	0.302	0.302	0.324	0.195
0.05	High	0.358	0.051	0.476	0.165	0.098	0.363	0.363	0.334	0.276
	Low	0.414	0.053	0.361	0.080	0.040	0.394	0.388	0.420	0.269
0.10	High	0.438	0.102	0.526	0.209	0.135	0.421	0.419	0.398	0.331
	Low	0.490	0.106	0.423	0.124	0.077	0.452	0.443	0.482	0.325

Table C-3. Average Type I Error Rates by Fit Statistic and Model Conditions

α	M	Fit Statistic								Overall
		$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO	
0.01	1PL	0.238	0.011	0.660	0.146	0.050	0.302	0.307	0.301	0.252
	2PL	0.247	0.009	0.264	0.063	0.049	0.263	0.268	0.252	0.177
	3PL	0.329	0.012	0.103	0.020	0.015	0.267	0.270	0.264	0.160
0.05	1PL	0.345	0.049	0.728	0.224	0.099	0.418	0.416	0.417	0.337
	2PL	0.363	0.047	0.357	0.100	0.074	0.360	0.357	0.354	0.252
	3PL	0.452	0.059	0.170	0.044	0.035	0.356	0.353	0.360	0.229
0.10	1PL	0.419	0.097	0.768	0.285	0.151	0.490	0.483	0.491	0.398
	2PL	0.441	0.096	0.425	0.140	0.105	0.413	0.408	0.414	0.305
	3PL	0.533	0.118	0.231	0.074	0.062	0.406	0.402	0.415	0.280

Table C-4. Average Type I Error Rates by Fit Statistic, N, and n Conditions

N	n	Fit Statistic								Overall
		$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO	
$\alpha = 0.01$										
500	15	0.216	0.012	0.376	0.088	0.048	0.321	0.341	0.310	0.214
	75	0.012	0.008	0.412	0.099	0.060	0.003	0.005	0.003	0.075
1,500	15	0.837	0.012	0.277	0.058	0.021	0.773	0.768	0.764	0.439
	75	0.021	0.010	0.305	0.060	0.024	0.012	0.013	0.011	0.057
$\alpha = 0.05$										
500	15	0.454	0.056	0.458	0.136	0.081	0.579	0.576	0.569	0.364
	75	0.057	0.046	0.489	0.146	0.092	0.019	0.020	0.017	0.111
1,500	15	0.943	0.055	0.348	0.104	0.052	0.866	0.855	0.872	0.512
	75	0.092	0.051	0.378	0.104	0.052	0.049	0.050	0.052	0.103
$\alpha = 0.10$										
500	15	0.599	0.110	0.517	0.182	0.119	0.700	0.689	0.693	0.451
	75	0.117	0.096	0.543	0.191	0.129	0.040	0.042	0.038	0.150
1,500	15	0.971	0.107	0.405	0.147	0.089	0.904	0.891	0.914	0.553
	75	0.169	0.102	0.433	0.146	0.087	0.103	0.102	0.116	0.157

Figure C-1. DN Bias by Study Condition for Q Statistics at $\alpha = 0.01$

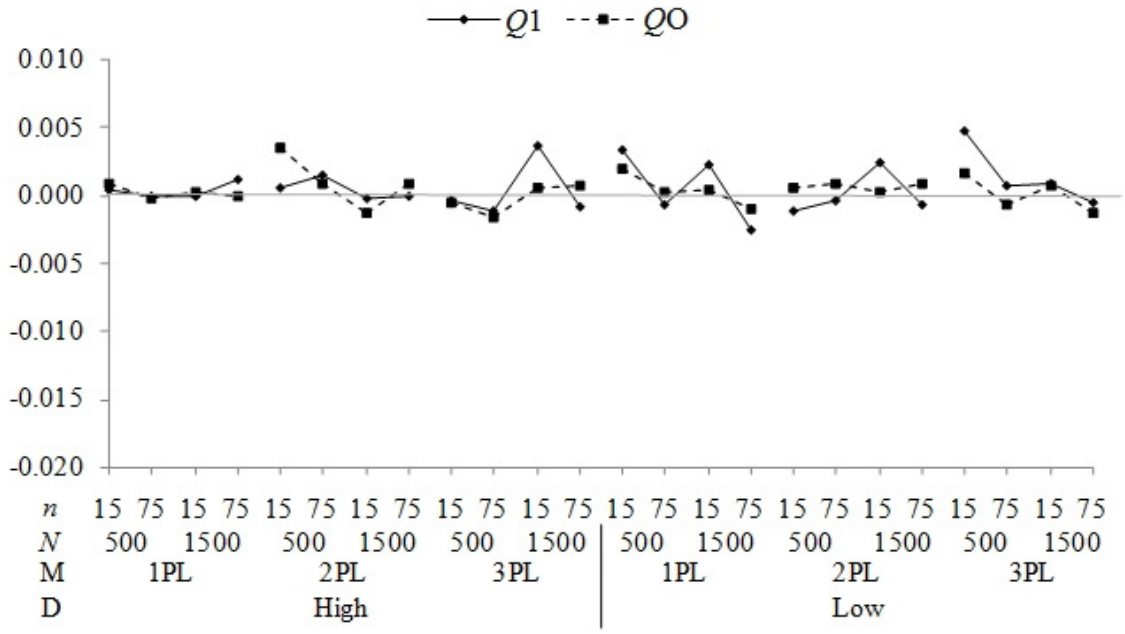


Figure C-2. DN Bias by Study Condition for LM Statistics at $\alpha = 0.01$

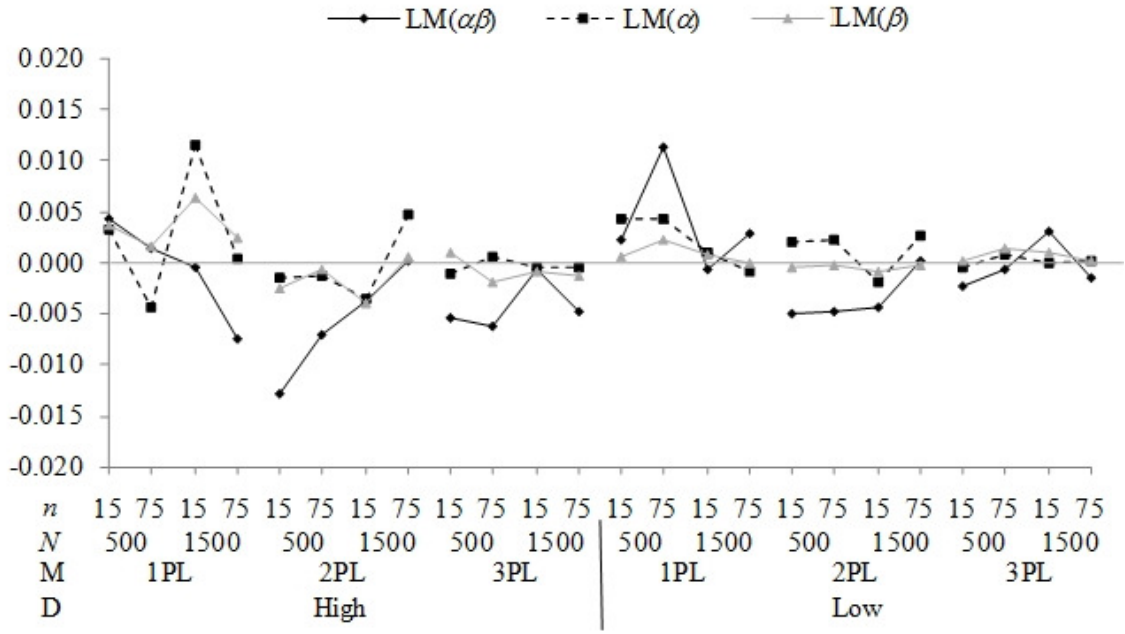


Figure C-3. DN Bias by Study Condition for z Statistics at $\alpha = 0.01$

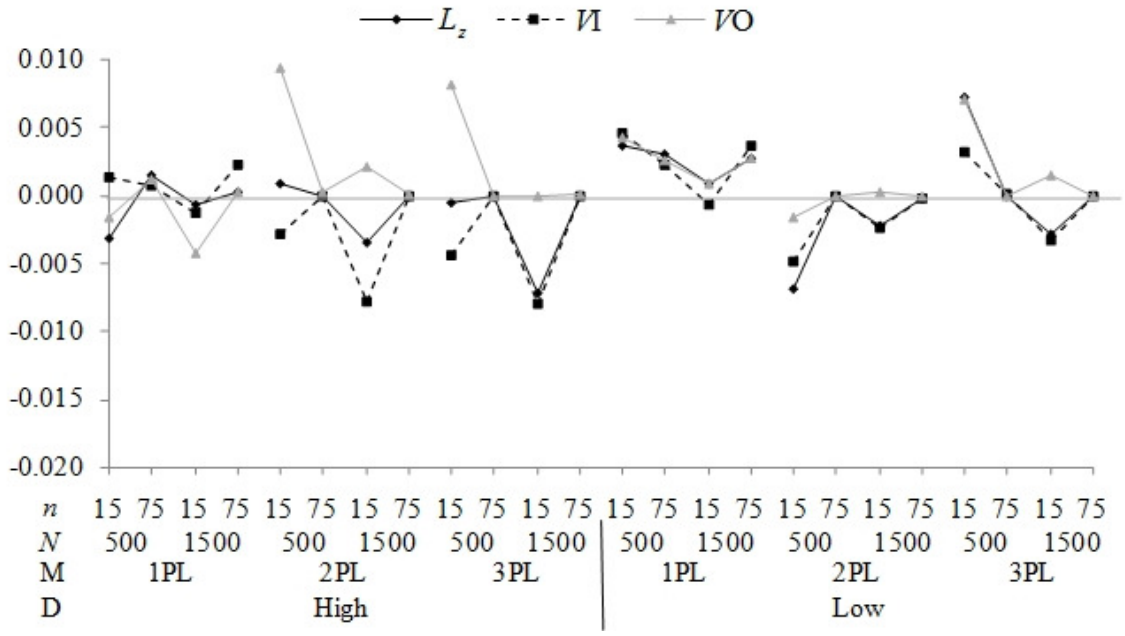


Figure C-4. DN Bias by Study Condition for Q Statistics at $\alpha = 0.10$

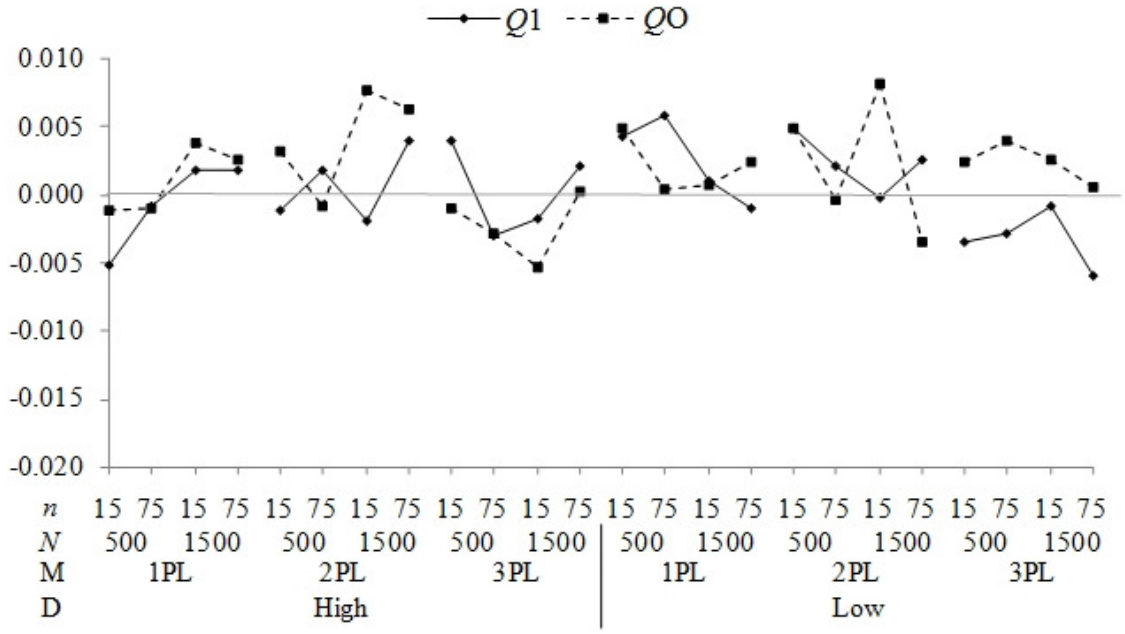


Figure C-5. DN Bias by Study Condition for LM Statistics at $\alpha = 0.10$

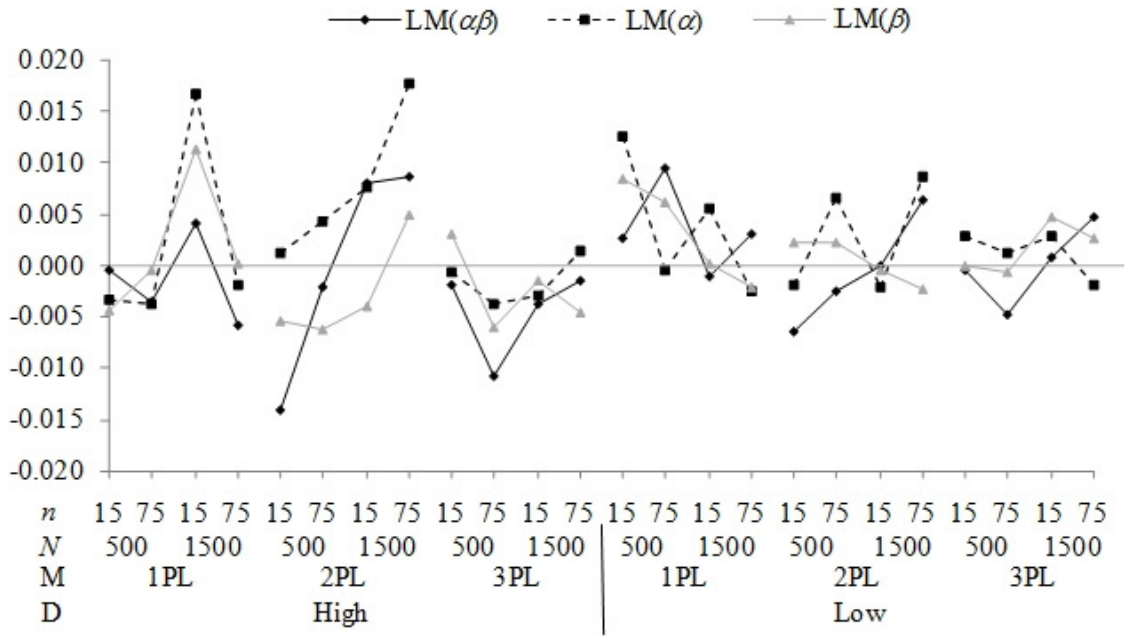


Figure C-6. DN Bias by Study Condition for z Statistics at $\alpha = 0.10$

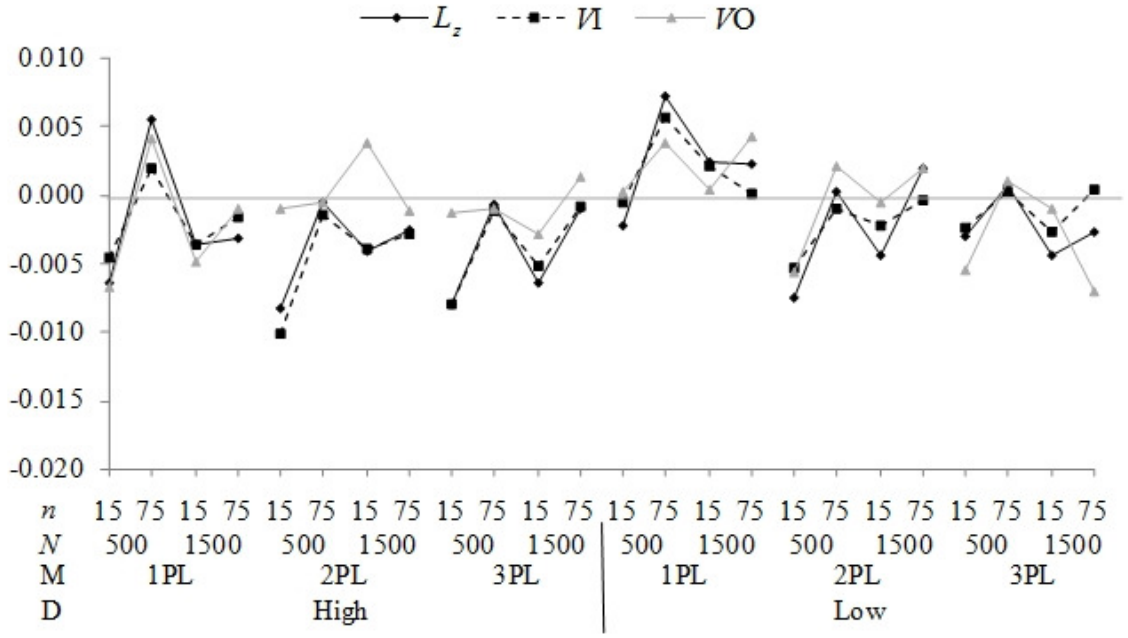


Table C-5. Type I Error Rates at $\alpha = 0.01$ for all EU Conditions

M	N	n	Fit Statistic							
			$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO
High Discrimination										
1PL	500	15	0.090	**0.011**	0.762	0.239	0.114	0.315	0.345	0.282
		75	**0.008**	**0.011**	0.779	0.249	0.131	**0.011**	**0.014**	**0.005**
	1,500	15	0.746	**0.011**	0.719	0.205	0.055	0.799	0.800	0.739
		75	**0.013**	**0.012**	0.700	0.180	0.049	0.032	0.036	**0.019**
2PL	500	15	0.113	**0.011**	0.391	0.134	0.109	0.190	0.224	0.126
		75	**0.009**	**0.007**	0.464	0.160	0.136	**0.000**	**0.000**	**0.001**
	1,500	15	0.728	**0.011**	0.194	0.051	0.038	0.734	0.735	0.632
		75	**0.013**	**0.010**	0.285	0.068	0.055	**0.000**	**0.000**	**0.001**
3PL	500	15	0.281	**0.013**	0.166	0.035	**0.026**	0.227	0.252	0.158
		75	**0.016**	**0.008**	0.228	0.059	0.047	**0.000**	**0.000**	**0.000**
	1,500	15	0.857	**0.014**	0.062	**0.012**	**0.008**	0.720	0.719	0.689
		75	0.035	**0.011**	0.115	**0.023**	**0.019**	**0.000**	**0.000**	**0.000**
Low Discrimination										
1PL	500	15	0.153	**0.012**	0.577	0.080	**0.019**	0.395	0.404	0.440
		75	**0.008**	**0.010**	0.620	0.085	**0.020**	**0.012**	**0.014**	**0.012**
	1,500	15	0.872	**0.012**	0.548	0.068	**0.013**	0.818	0.804	0.865
		75	**0.015**	**0.011**	0.581	0.072	**0.012**	0.040	0.043	0.046
2PL	500	15	0.214	**0.010**	0.264	0.032	**0.015**	0.388	0.402	0.425
		75	**0.009**	**0.007**	0.275	0.033	**0.019**	**0.000**	**0.000**	**0.000**
	1,500	15	0.880	**0.011**	0.108	**0.011**	**0.006**	0.787	0.776	0.834
		75	**0.014**	**0.009**	0.115	**0.012**	**0.007**	**0.000**	**0.001**	**0.001**
3PL	500	15	0.448	**0.017**	0.086	**0.010**	**0.006**	0.411	0.419	0.444
		75	**0.019**	**0.009**	0.100	**0.012**	**0.009**	**0.000**	**0.000**	**0.000**
	1,500	15	0.940	**0.015**	**0.027**	**0.004**	**0.004**	0.773	0.761	0.825
		75	0.037	**0.010**	**0.030**	**0.005**	**0.004**	**0.000**	**0.000**	**0.000**

NOTE. Type I error rates (T1) near nominal levels, defined by $0.00 \leq T1 \leq 0.03$, are in bold are in bold, those below nominal levels are in italics, and those above nominal levels are in regular font.

Table C-6. Type I Error Rates at $\alpha = 0.05$ for all EU Conditions

M	N	n	Fit Statistic							
			$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	VI	VO
High Discrimination										
1PL	500	15	0.278	**0.047**	0.814	0.312	0.166	0.571	0.575	0.523
		75	**0.039**	**0.046**	0.827	0.321	0.179	**0.056**	**0.062**	**0.036**
	1,500	15	0.910	**0.048**	0.775	0.293	0.119	0.887	0.883	0.854
		75	**0.062**	**0.053**	0.762	0.259	0.094	0.119	0.118	0.093
2PL	500	15	0.312	**0.050**	0.475	0.177	0.137	0.512	0.514	0.423
		75	**0.045**	**0.041**	0.544	0.201	0.163	*0.000*	*0.000*	*0.002*
	1,500	15	0.907	**0.050**	0.281	0.091	**0.064**	0.846	0.842	0.780
		75	**0.063**	**0.050**	0.372	0.108	0.082	*0.001*	*0.002*	*0.004*
3PL	500	15	0.541	**0.065**	0.248	**0.064**	**0.050**	0.522	0.520	0.471
		75	**0.078**	**0.048**	0.300	0.084	**0.068**	*0.000*	*0.001*	*0.001*
	1,500	15	0.945	**0.063**	0.119	**0.033**	**0.024**	0.823	0.817	0.817
		75	0.128	**0.053**	0.175	**0.043**	**0.035**	*0.002*	*0.004*	*0.001*
Low Discrimination										
1PL	500	15	0.390	**0.051**	0.659	0.155	**0.067**	0.622	0.614	0.674
		75	**0.047**	**0.047**	0.697	0.163	**0.067**	**0.057**	**0.059**	**0.061**
	1,500	15	0.962	**0.054**	0.631	0.145	**0.056**	0.898	0.881	0.942
		75	**0.076**	**0.052**	0.663	0.148	**0.053**	0.140	0.135	0.153
2PL	500	15	0.499	**0.052**	0.373	**0.072**	**0.041**	0.619	0.615	0.658
		75	**0.048**	**0.043**	0.383	**0.073**	**0.045**	*0.002*	*0.003*	*0.003*
	1,500	15	0.955	**0.054**	0.206	**0.043**	**0.029**	0.877	0.859	0.926
		75	**0.075**	**0.047**	0.211	**0.045**	**0.027**	*0.018*	*0.019*	**0.036**
3PL	500	15	0.707	**0.074**	0.172	**0.038**	**0.029**	0.618	0.611	0.660
		75	0.085	**0.049**	0.176	**0.037**	**0.029**	*0.001*	*0.002*	*0.001*
	1,500	15	0.979	**0.068**	**0.080**	**0.025**	**0.023**	0.856	0.837	0.905
		75	0.147	**0.055**	0.082	**0.024**	*0.019*	*0.016*	*0.019*	**0.023**

NOTE. Type I error rates (T1) near nominal levels, defined by $0.02 \leq T1 \leq 0.08$, are in bold, those below nominal levels are in italics, and those above nominal levels are in regular font.

Table C-7. Type I Error Rates at $\alpha = 0.10$ for all EU Conditions

M	N	n	Fit Statistic							
			$Q1$	QO	LM($\alpha\beta$)	LM(α)	LM(β)	L_z	*V*I	*V*O
High Discrimination										
1PL	500	15	0.423	**0.091**	0.842	0.371	0.217	0.692	0.686	0.648
		75	**0.085**	**0.091**	0.854	0.378	0.224	**0.111**	**0.114**	**0.080**
	1,500	15	0.956	**0.097**	0.811	0.359	0.182	0.919	0.913	0.902
		75	**0.121**	**0.102**	0.799	0.321	0.144	0.203	0.197	0.181
2PL	500	15	0.476	**0.099**	0.536	0.218	0.166	0.666	0.659	0.571
		75	**0.094**	**0.090**	0.598	0.240	0.191	*0.001*	*0.002*	*0.004*
	1,500	15	0.953	**0.102**	0.349	0.133	**0.094**	0.890	0.885	0.847
		75	**0.126**	**0.100**	0.440	0.152	**0.115**	*0.009*	*0.016*	*0.033*
3PL	500	15	0.686	0.131	0.315	**0.101**	**0.080**	0.658	0.651	0.624
		75	0.154	**0.101**	0.356	**0.113**	**0.097**	*0.003*	*0.005*	*0.002*
	1,500	15	0.970	**0.119**	0.175	*0.061*	*0.049*	0.867	0.860	0.872
		75	0.218	**0.104**	0.229	**0.070**	*0.056*	*0.012*	*0.020*	*0.009*
Low Discrimination										
1PL	500	15	0.543	**0.101**	0.708	0.220	**0.121**	0.728	0.711	0.780
		75	**0.100**	**0.094**	0.744	0.229	**0.121**	**0.112**	**0.113**	**0.121**
	1,500	15	0.982	**0.101**	0.681	0.206	**0.107**	0.930	0.912	0.967
		75	0.145	**0.106**	0.713	0.209	**0.103**	0.226	0.218	0.253
2PL	500	15	0.655	**0.104**	0.447	**0.114**	**0.075**	0.723	0.711	0.766
		75	**0.105**	**0.089**	0.457	**0.118**	**0.081**	*0.012*	*0.014*	*0.017*
	1,500	15	0.975	**0.105**	0.281	**0.079**	*0.058*	0.915	0.895	0.958
		75	0.147	**0.096**	0.290	**0.086**	*0.059*	**0.079**	**0.072**	**0.120**
3PL	500	15	0.814	0.143	0.245	**0.074**	*0.058*	0.714	0.701	0.758
		75	0.167	**0.110**	0.242	*0.069*	*0.058*	*0.008*	*0.009*	*0.008*
	1,500	15	0.988	**0.129**	0.137	*0.055*	*0.050*	0.891	0.872	0.936
		75	0.260	**0.110**	0.138	*0.047*	*0.044*	**0.085**	**0.086**	**0.099**

NOTE. Type I error rates (T1) near nominal levels, defined by $0.07 \leq T1 \leq 0.13$, are in bold are in bold, those below nominal levels are in italics, and those above nominal levels are in regular font.

APPENDIX D: TYPE I ERROR RATES IN PARAMETER ESTIMATION ERROR CONDITIONS

Sums of Squares and Effect Sizes for Study Factors on *Q*O, *Q*1, L_z, *V*I, and *V*O Type I Error Rates at $\alpha = 0.05$ Within Each Parameter Estimation Error Condition

Type I Error Rates for *Q*O at $\alpha = 0.01$, 0.05, and 0.10 by Parameter Estimation Error Condition

Type I Error Rates for *Q*1, L_z, *V*I, and *V*O at $\alpha = 0.10$ by Parameter Estimation Error Condition

Type I Error Rates at $\alpha = 0.01$, 0.05, and 0.10 for $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Table D-1. Sums of Squares and Effect Sizes for Study Factors on *Q*O Type I Error Rates at $\alpha = 0.05$ Within Each PE Condition

	PE Condition			
	$\hat{\xi}$		ξ	
	SS	η^2	SS	η^2
Main Effects				
Discrimination (D)	0.0001	0.0216	0.0000	0.0205
Data Noise (DN)	0.0000	0.0027	0.0002	0.1946
Model (M)	0.0013	0.4583	0.0001	0.0823
Sample Size (*N*)	0.0001	0.0240	0.0003	0.3647
Test Length (*n*)	0.0006	0.2068	0.0000	0.0303
2-way Interactions				
D × DN	0.0000	0.0003	0.0000	0.0025
D × M	0.0000	0.0053	0.0000	0.0020
D × *N*	0.0000	0.0001	0.0000	0.0004
D × *n*	0.0000	0.0121	0.0000	0.0141
DN × M	0.0000	0.0032	0.0001	0.0599
DN × *N*	0.0000	0.0006	0.0000	0.0479
DN × *n*	0.0000	0.0013	0.0000	0.0175
M × *N*	0.0000	0.0133	0.0000	0.0074
M × *n*	0.0006	0.1946	0.0000	0.0249
N × *n*	0.0001	0.0324	0.0000	0.0407
3-way Interactions				
D × DN × M	0.0000	0.0014	0.0000	0.0155
D × DN × *N*	0.0000	0.0014	0.0000	0.0057
D × DN × *n*	0.0000	0.0029	0.0000	0.0068
D × M × *N*	0.0000	0.0009	0.0000	0.0062
D × M × *n*	0.0000	0.0006	0.0000	0.0033
D × *N* × *n*	0.0000	0.0000	0.0000	0.0004
DN × M × *N*	0.0000	0.0005	0.0000	0.0042
DN × M × *n*	0.0000	0.0014	0.0000	0.0144
DN × *N* × *n*	0.0000	0.0004	0.0000	0.0036
M × *N* × *n*	0.0000	0.0056	0.0000	0.0001
4-way Interactions				
D × DN × M × *N*	0.0000	0.0006	0.0000	0.0010
D × DN × M × *n*	0.0000	0.0009	0.0000	0.0044
D × DN × *N* × *n*	0.0000	0.0018	0.0000	0.0119
D × M × *N* × *n*	0.0000	0.0002	0.0000	0.0031
DN × M × *N* × *n*	0.0000	0.0005	0.0000	0.0003
5-way Interactions				
D × DN × M × *N* × *n*	0.0000	0.0044	0.0000	0.0096

	PE Condition			
	$\hat{\xi}$		ξ	
	SS	η^2	SS	η^2
Study Effects TOTAL	0.0029	0.0221	0.0009	0.0063
Intercept	0.1301	0.9779	0.1380	0.9937
TOTAL	0.1331		0.1389	

Table D-2. Sums of Squares and Effect Sizes for Study Factors on $Q1$ Type I Error Rates at $\alpha = 0.05$ Within Each PE Condition

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Main Effects								
Discrimination (D)	0.0373	0.0058	0.0008	0.0385	0.0135	0.0026	0.0001	0.0541
Data Noise (DN)	0.0000	0.0000	0.0001	0.0024	0.0003	0.0001	0.0007	0.4209
Model (M)	0.1046	0.0164	0.0156	0.7167	0.0533	0.0103	0.0002	0.1062
Sample Size (N)	0.8222	0.1287	0.0002	0.0075	0.8520	0.1640	0.0001	0.0751
Test Length (n)	4.6636	0.7298	0.0023	0.1060	3.6408	0.7009	0.0000	0.0048
2-way Interactions								
D × DN	0.0000	0.0000	0.0000	0.0002	0.0004	0.0001	0.0000	0.0220
D × M	0.0005	0.0001	0.0000	0.0016	0.0006	0.0001	0.0000	0.0182
D × N	0.0070	0.0011	0.0001	0.0046	0.0008	0.0001	0.0000	0.0000
D × n	0.0234	0.0037	0.0000	0.0011	0.0255	0.0049	0.0000	0.0008
DN × M	0.0000	0.0000	0.0000	0.0002	0.0002	0.0000	0.0002	0.1113
DN × N	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0002	0.0968
DN × n	0.0000	0.0000	0.0000	0.0002	0.0005	0.0001	0.0000	0.0001
M × N	0.0286	0.0045	0.0001	0.0063	0.0116	0.0022	0.0000	0.0153
M × n	0.0226	0.0035	0.0020	0.0898	0.0046	0.0009	0.0000	0.0092
N × n	0.6196	0.0970	0.0000	0.0000	0.5488	0.1057	0.0000	0.0016
3-way Interactions								
D × DN × M	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0149
D × DN × N	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0013
D × DN × n	0.0000	0.0000	0.0000	0.0008	0.0001	0.0000	0.0000	0.0072
D × M × N	0.0009	0.0001	0.0002	0.0078	0.0003	0.0000	0.0000	0.0043

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
D × M × *n*	0.0007	0.0001	0.0000	0.0009	0.0008	0.0002	0.0000	0.0029
D × *N* × *n*	0.0113	0.0018	0.0001	0.0031	0.0004	0.0001	0.0000	0.0007
DN × M × *N*	0.0000	0.0000	0.0000	0.0006	0.0000	0.0000	0.0000	0.0202
DN × M × *n*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0002
DN × *N* × *n*	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0002
M × *N* × *n*	0.0468	0.0073	0.0001	0.0031	0.0394	0.0076	0.0000	0.0042
4-way Interactions								
D × DN × M × *N*	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0005
D × DN × M × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × M × *N* × *n*	0.0010	0.0002	0.0001	0.0053	0.0001	0.0000	0.0000	0.0028
DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0004	0.0000	0.0000	0.0000	0.0007
5-way Interactions								
D × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0010	0.0000	0.0000	0.0000	0.0034
Study Effects TOTAL	6.3901	0.4715	0.0218	0.0805	5.1941	0.4453	0.0017	0.0126
Intercept	7.1640	0.5285	0.2486	0.9195	6.4710	0.5547	0.1363	0.9874
TOTAL	13.5541		0.2703		11.6651		0.1380	

Table D-3. Sums of Squares and Effect Sizes for Study Factors on L_z Type I Error Rates at $\alpha = 0.05$ Within Each PE Condition

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Main Effects								
Discrimination (D)	0.0116	0.0018	0.0001	0.0076	0.0194	0.0045	0.0000	0.0008
Data Noise (DN)	0.0000	0.0000	0.0000	0.0009	0.0000	0.0000	0.0019	0.4488
Model (M)	0.0391	0.0062	0.0101	0.9206	0.0083	0.0019	0.0006	0.1372
Sample Size (*N*)	0.3034	0.0484	0.0000	0.0021	0.3642	0.0840	0.0004	0.1039
Test Length (*n*)	5.6891	0.9074	0.0004	0.0337	3.7911	0.8743	0.0000	0.0048
2-way Interactions								
D × DN	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0001
D × M	0.0008	0.0001	0.0002	0.0150	0.0013	0.0003	0.0000	0.0016
D × *N*	0.0015	0.0002	0.0000	0.0007	0.0005	0.0001	0.0000	0.0024
D × *n*	0.0060	0.0009	0.0000	0.0012	0.0067	0.0015	0.0000	0.0016
DN × M	0.0000	0.0000	0.0000	0.0002	0.0006	0.0001	0.0006	0.1312
DN × *N*	0.0000	0.0000	0.0000	0.0006	0.0000	0.0000	0.0004	0.0988
DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0003	0.0001	0.0000	0.0007
M × *N*	0.0045	0.0007	0.0000	0.0009	0.0017	0.0004	0.0001	0.0220
M × *n*	0.0076	0.0012	0.0001	0.0131	0.0155	0.0036	0.0000	0.0004
N × *n*	0.1986	0.0317	0.0000	0.0006	0.1208	0.0279	0.0000	0.0034
3-way Interactions								
D × DN × M	0.0000	0.0000	0.0000	0.0000	0.0003	0.0001	0.0000	0.0002
D × DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0008
D × M × *N*	0.0002	0.0000	0.0000	0.0000	0.0003	0.0001	0.0000	0.0007

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
D × M × *n*	0.0009	0.0001	0.0000	0.0001	0.0015	0.0003	0.0000	0.0025
D × *N* × *n*	0.0043	0.0007	0.0000	0.0004	0.0027	0.0006	0.0000	0.0006
DN × M × *N*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0214
DN × M × *n*	0.0000	0.0000	0.0000	0.0004	0.0004	0.0001	0.0000	0.0051
DN × *N* × *n*	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0025
M × *N* × *n*	0.0022	0.0004	0.0000	0.0001	0.0001	0.0000	0.0000	0.0026
4-way Interactions								
D × DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007
D × DN × M × *n*	0.0000	0.0000	0.0000	0.0001	0.0002	0.0001	0.0000	0.0008
D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
D × M × *N* × *n*	0.0001	0.0000	0.0000	0.0007	0.0000	0.0000	0.0000	0.0006
DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0025
5-way Interactions								
D × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0014
Study Effects TOTAL	6.2698	0.4774	0.0110	0.5139	4.3362	0.3531	0.0043	0.0283
Intercept	6.8638	0.5226	0.0104	0.4861	7.9436	0.6469	0.1486	0.9717
TOTAL	13.1335		0.0214		12.2798		0.1530	

Table D-4. Sums of Squares and Effect Sizes for Study Factors on *VI* Type I Error Rates at $\alpha = 0.05$ Within Each PE Condition

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Main Effects								
Discrimination (D)	0.0077	0.0013	0.0001	0.0134	0.0146	0.0035	0.0000	0.0011
Data Noise (DN)	0.0000	0.0000	0.0000	0.0013	0.0000	0.0000	0.0017	0.4723
Model (M)	0.0396	0.0065	0.0099	0.9124	0.0082	0.0020	0.0004	0.1177
Sample Size (*N*)	0.2844	0.0466	0.0000	0.0023	0.3338	0.0800	0.0004	0.1132
Test Length (*n*)	5.5546	0.9108	0.0004	0.0342	3.6717	0.8797	0.0000	0.0076
2-way Interactions								
D × DN	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0001
D × M	0.0011	0.0002	0.0002	0.0185	0.0014	0.0003	0.0000	0.0004
D × *N*	0.0019	0.0003	0.0000	0.0009	0.0004	0.0001	0.0000	0.0020
D × *n*	0.0036	0.0006	0.0000	0.0005	0.0051	0.0012	0.0000	0.0008
DN × M	0.0000	0.0000	0.0000	0.0003	0.0008	0.0002	0.0004	0.1182
DN × *N*	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0004	0.0982
DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0003	0.0001	0.0000	0.0004
M × *N*	0.0036	0.0006	0.0000	0.0013	0.0014	0.0003	0.0001	0.0185
M × *n*	0.0071	0.0012	0.0001	0.0091	0.0157	0.0038	0.0000	0.0002
N × *n*	0.1870	0.0307	0.0000	0.0007	0.1148	0.0275	0.0000	0.0009
3-way Interactions								
D × DN × M	0.0000	0.0000	0.0000	0.0000	0.0004	0.0001	0.0000	0.0004
D × DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D × DN × *n*	0.0000	0.0000	0.0000	0.0003	0.0001	0.0000	0.0000	0.0008
D × M × *N*	0.0003	0.0000	0.0000	0.0003	0.0002	0.0001	0.0000	0.0006

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
D × M × *n*	0.0009	0.0001	0.0000	0.0010	0.0015	0.0004	0.0000	0.0013
D × *N* × *n*	0.0050	0.0008	0.0000	0.0003	0.0028	0.0007	0.0000	0.0000
DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0262
DN × M × *n*	0.0000	0.0000	0.0000	0.0009	0.0004	0.0001	0.0000	0.0047
DN × *N* × *n*	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0014
M × *N* × *n*	0.0017	0.0003	0.0000	0.0000	0.0001	0.0000	0.0000	0.0006
4-way Interactions								
D × DN × M × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003
D × DN × M × *n*	0.0000	0.0000	0.0000	0.0002	0.0002	0.0000	0.0000	0.0055
D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0013
D × M × *N* × *n*	0.0002	0.0000	0.0000	0.0011	0.0000	0.0000	0.0000	0.0004
DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0040
5-way Interactions								
D × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0008
Study Effects TOTAL	6.0987	0.4744	0.0108	0.5008	4.1739	0.3457	0.0036	0.0240
Intercept	6.7570	0.5256	0.0108	0.4992	7.8988	0.6543	0.1480	0.9760
TOTAL	12.8558		0.0216		12.0728		0.1517	

Table D-5. Sums of Squares and Effect Sizes for Study Factors on *VO* Type I Error Rates at $\alpha = 0.05$ Within Each PE Condition

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Main Effects								
Discrimination (D)	0.0887	0.0138	0.0001	0.0197	0.1145	0.0238	0.0027	0.5328
Data Noise (DN)	0.0000	0.0000	0.0000	0.0013	0.0000	0.0000	0.0008	0.1573
Model (M)	0.0392	0.0061	0.0050	0.7449	0.0121	0.0025	0.0005	0.0929
Sample Size (*N*)	0.3430	0.0535	0.0000	0.0049	0.3971	0.0826	0.0003	0.0622
Test Length (*n*)	5.6490	0.8810	0.0007	0.0992	4.0973	0.8519	0.0000	0.0018
2-way Interactions								
D × DN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0091
D × M	0.0020	0.0003	0.0004	0.0621	0.0008	0.0002	0.0000	0.0070
D × *N*	0.0023	0.0004	0.0000	0.0009	0.0006	0.0001	0.0000	0.0002
D × *n*	0.0478	0.0075	0.0000	0.0056	0.0185	0.0038	0.0000	0.0000
DN × M	0.0000	0.0000	0.0000	0.0017	0.0004	0.0001	0.0002	0.0439
DN × *N*	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0397
DN × *n*	0.0000	0.0000	0.0000	0.0000	0.0002	0.0001	0.0000	0.0014
M × *N*	0.0025	0.0004	0.0000	0.0001	0.0013	0.0003	0.0000	0.0070
M × *n*	0.0038	0.0006	0.0004	0.0523	0.0149	0.0031	0.0000	0.0003
N × *n*	0.2151	0.0335	0.0000	0.0018	0.1410	0.0293	0.0000	0.0018
3-way Interactions								
D × DN × M	0.0000	0.0000	0.0000	0.0001	0.0002	0.0000	0.0000	0.0032
D × DN × *N*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0054
D × DN × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000
D × M × *N*	0.0004	0.0001	0.0000	0.0003	0.0000	0.0000	0.0000	0.0025

	PE Condition							
	$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
	SS	η^2	SS	η^2	SS	η^2	SS	η^2
D × M × *n*	0.0053	0.0008	0.0000	0.0016	0.0031	0.0006	0.0000	0.0005
D × *N* × *n*	0.0098	0.0015	0.0000	0.0001	0.0068	0.0014	0.0000	0.0001
DN × M × *N*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001	0.0220
DN × M × *n*	0.0000	0.0000	0.0000	0.0001	0.0003	0.0001	0.0000	0.0015
DN × *N* × *n*	0.0000	0.0000	0.0000	0.0009	0.0000	0.0000	0.0000	0.0020
M × *N* × *n*	0.0032	0.0005	0.0000	0.0012	0.0003	0.0001	0.0000	0.0013
4-way Interactions								
D × DN × M × *N*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0006
D × DN × M × *n*	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0001
D × DN × *N* × *n*	0.0000	0.0000	0.0000	0.0002	0.0000	0.0000	0.0000	0.0010
D × M × *N* × *n*	0.0001	0.0000	0.0000	0.0001	0.0001	0.0000	0.0000	0.0001
DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0003	0.0000	0.0000	0.0000	0.0000
5-way Interactions								
D × DN × M × *N* × *n*	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0022
Study Effects TOTAL	6.4123	0.4842	0.0068	0.2728	4.8097	0.3900	0.0050	0.0483
Intercept	6.8316	0.5158	0.0180	0.7272	7.5230	0.6100	0.0989	0.9517
TOTAL	13.2439		0.0248		12.3327		0.1040	

Figure D-1. Type I Error Rates at $\alpha = 0.01$ for QO in SU Conditions by Parameter Estimation Error Condition

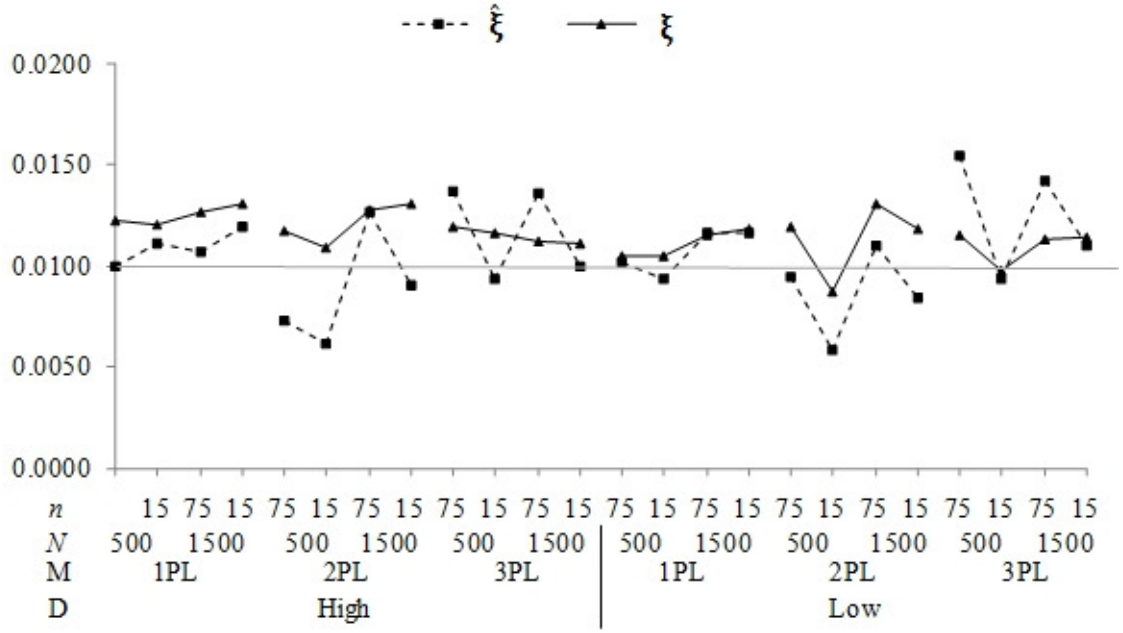


Figure D-2. Type I Error Rates at $\alpha = 0.01$ for QO in EU Conditions by Parameter Estimation Error Condition

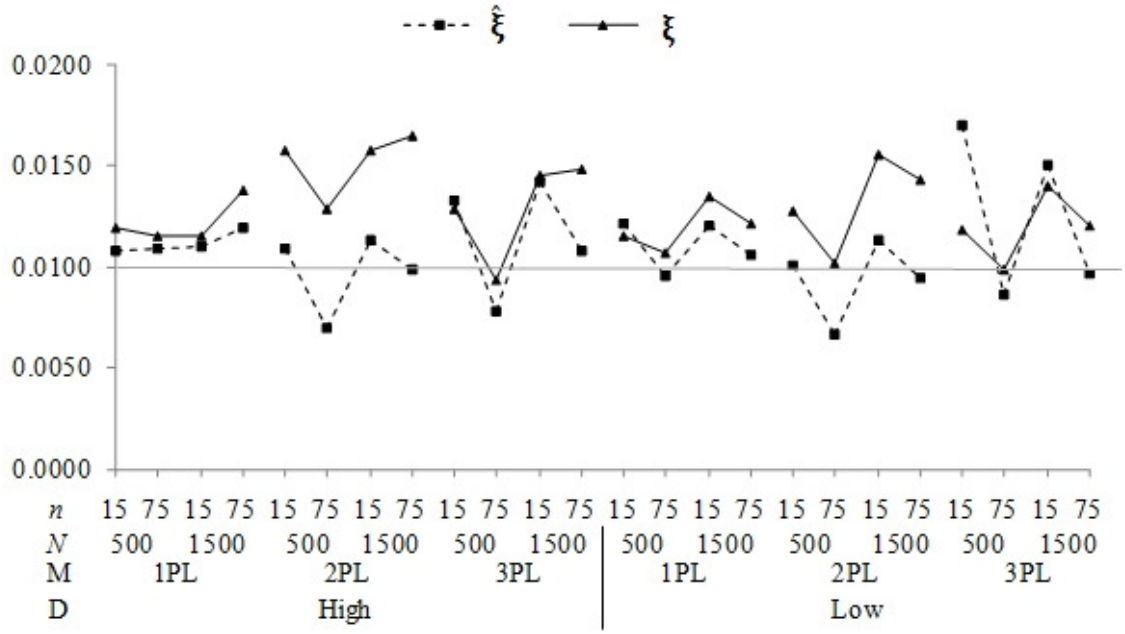


Figure D-3. Type I Error Rates at $\alpha = 0.05$ for *Q*O in SU Conditions by Parameter Estimation Error Condition

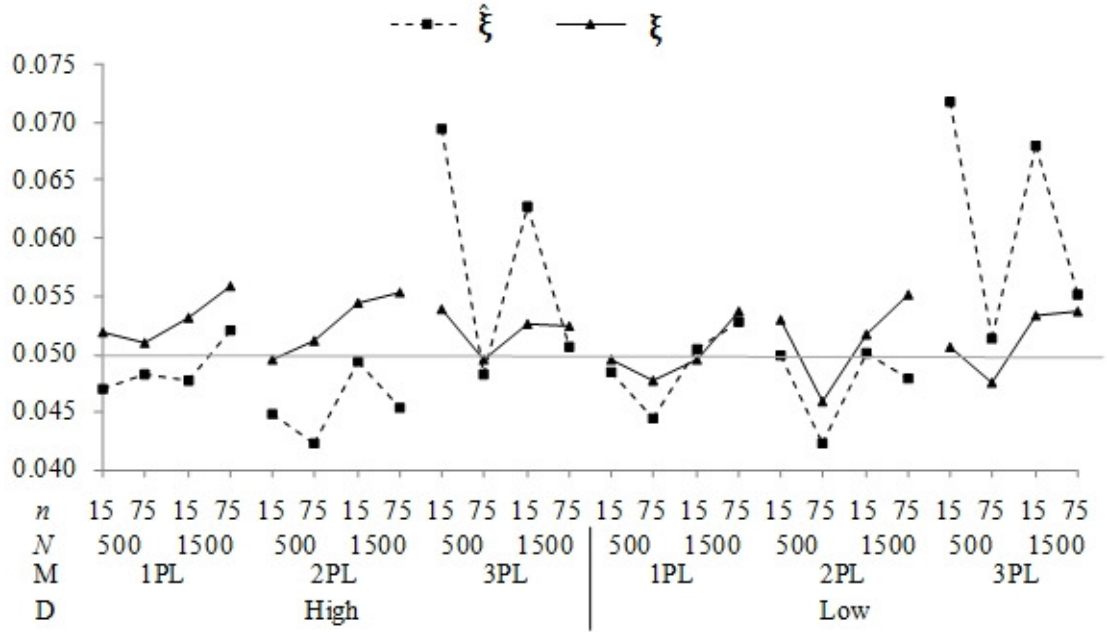


Figure D-4. Type I Error Rates at $\alpha = 0.05$ for *Q*O in EU Conditions by Parameter Estimation Error Condition

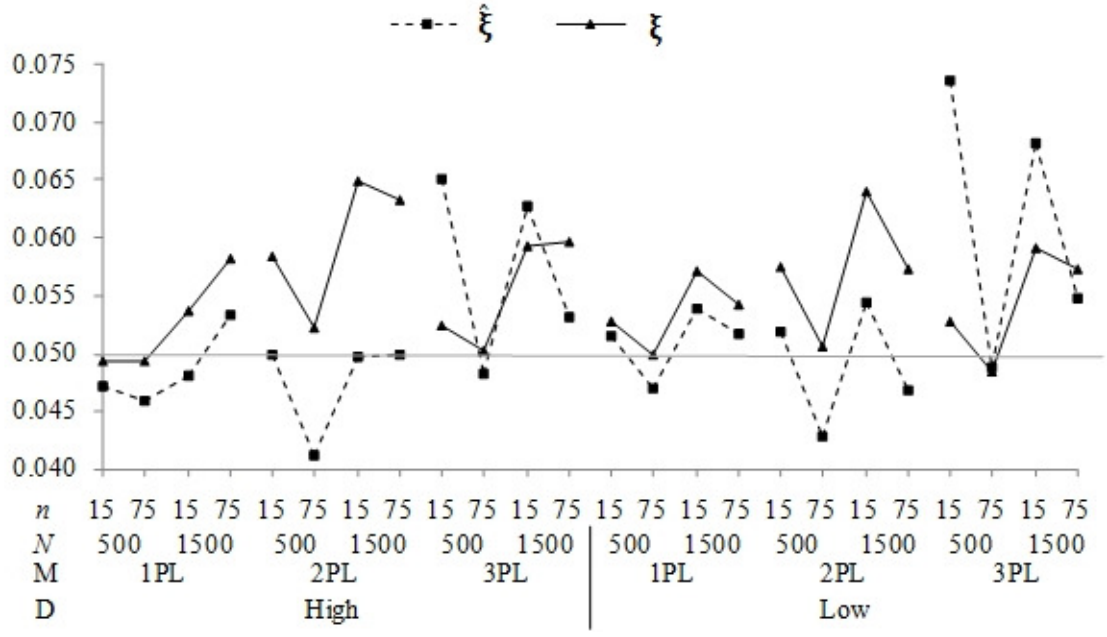


Figure D-5. Type I Error Rates at $\alpha = 0.10$ for *Q*O in SU Conditions by Parameter Estimation Error Condition

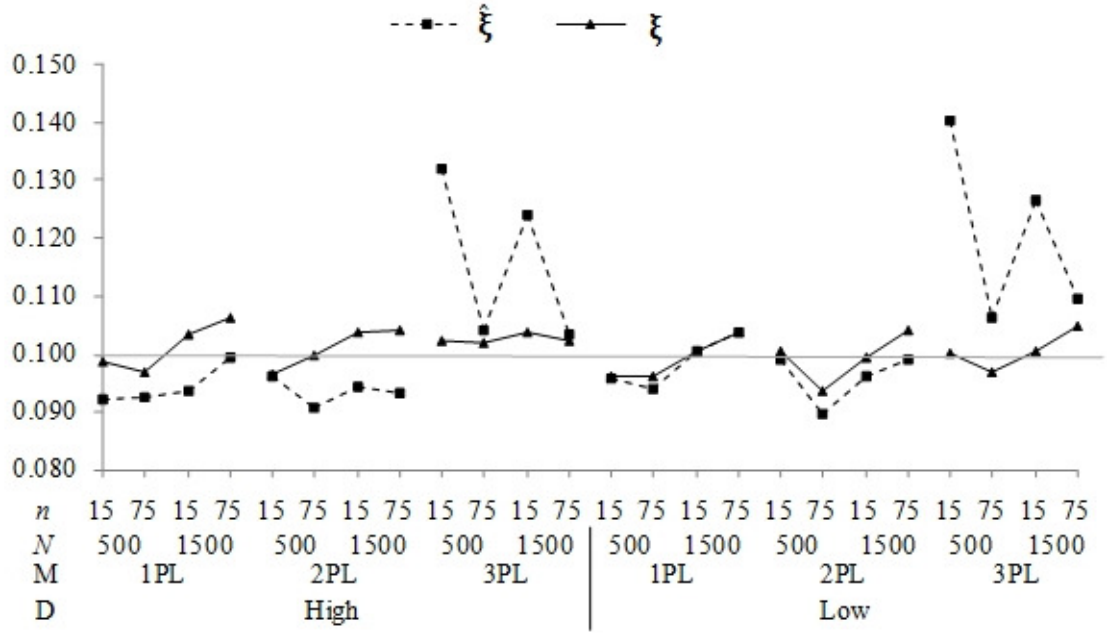


Figure D-6. Type I Error Rates at $\alpha = 0.10$ for *Q*O in EU Conditions by Parameter Estimation Error Condition

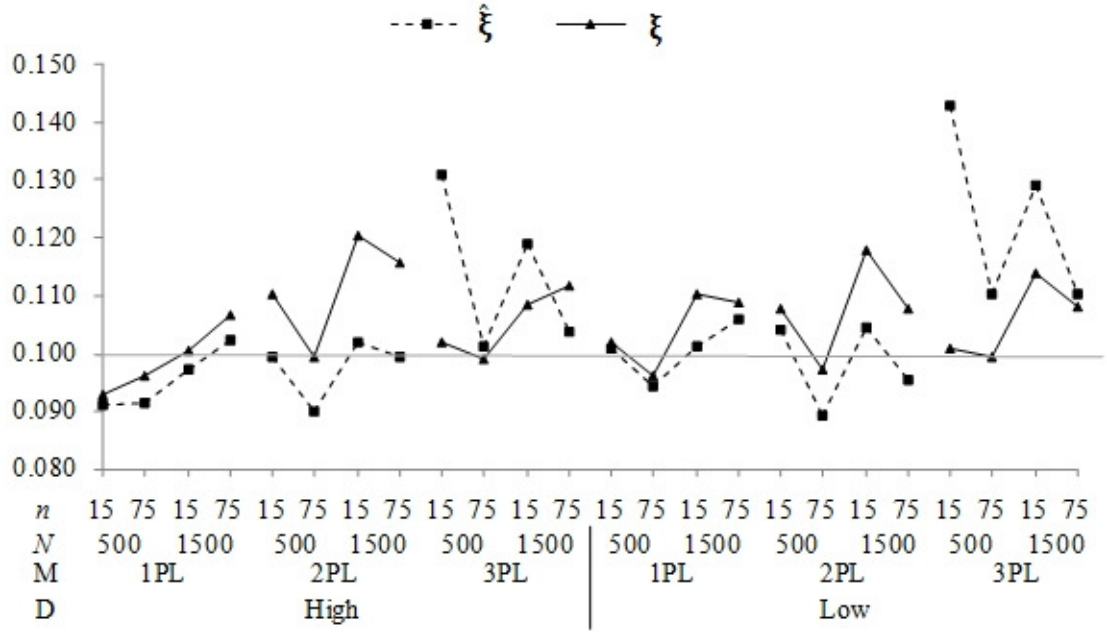


Figure D-7. Type I Error Rates at $\alpha = 0.10$ for $Q1$ in SU Conditions by Parameter Estimation Error Condition

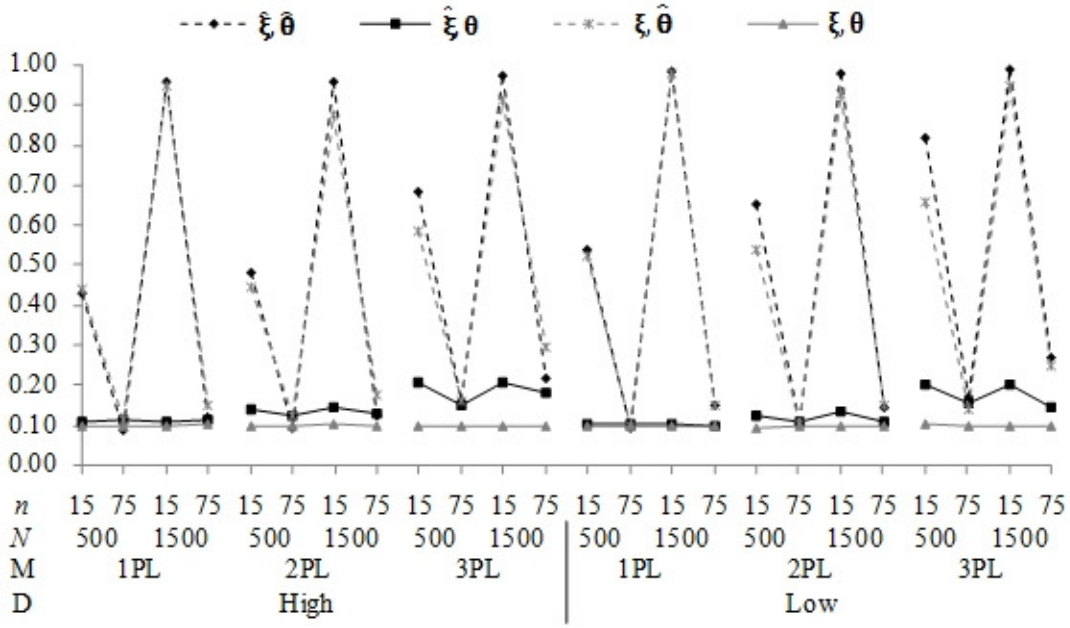


Figure D-8. Type I Error Rates at $\alpha = 0.10$ for $Q1$ in EU Conditions by Parameter Estimation Error Condition

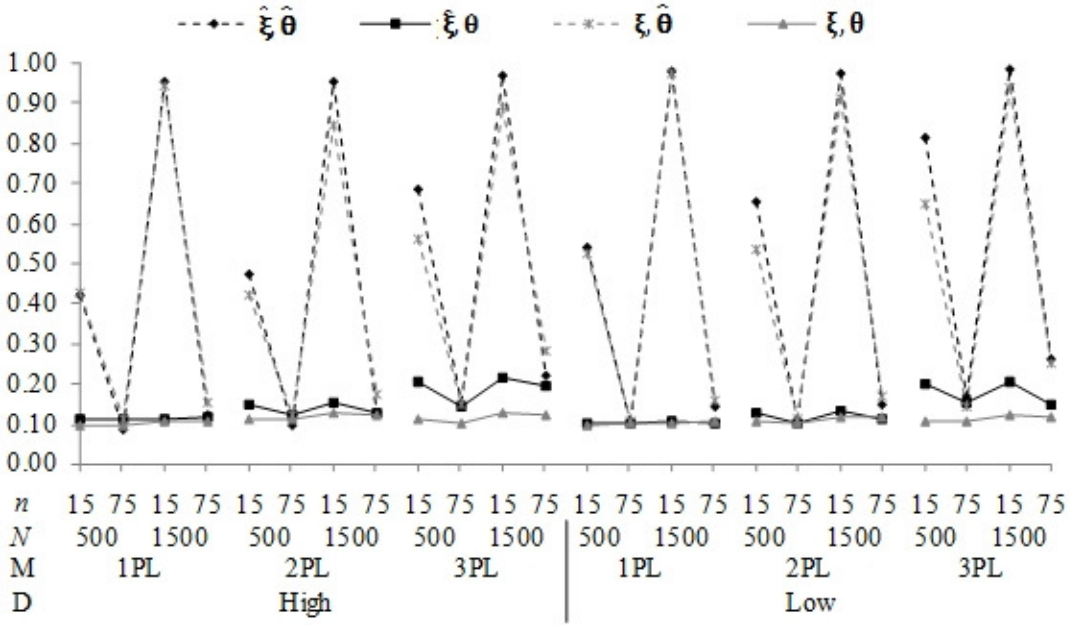


Figure D-9. Type I Error Rates at $\alpha = 0.10$ for L_z in SU Conditions by Parameter Estimation Error Condition

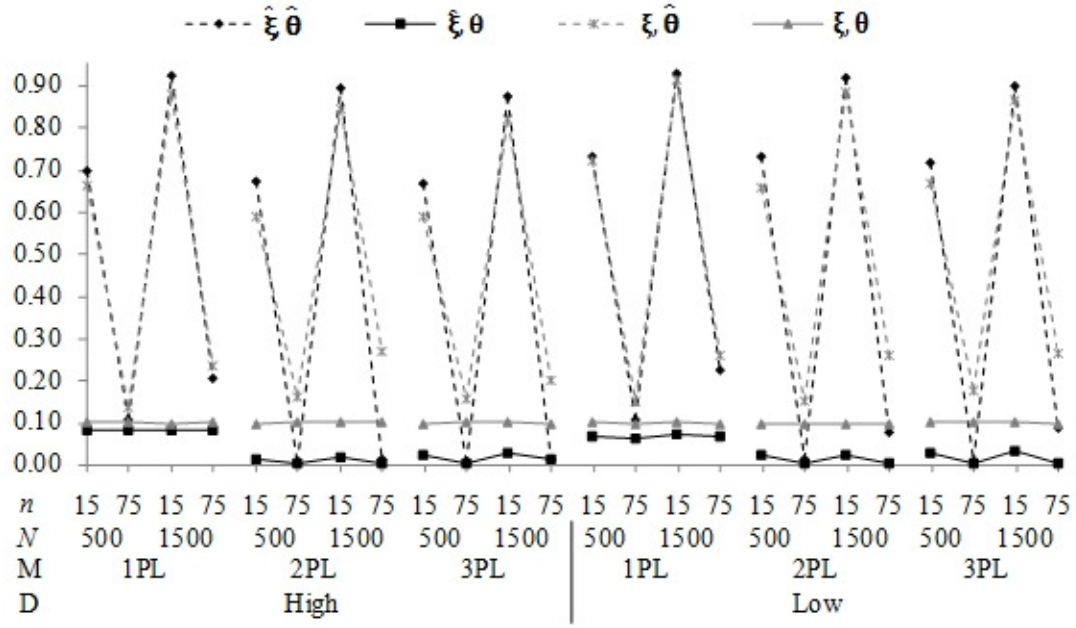


Figure D-10. Type I Error Rates at $\alpha = 0.10$ for L_z in EU Conditions by Parameter Estimation Error Condition

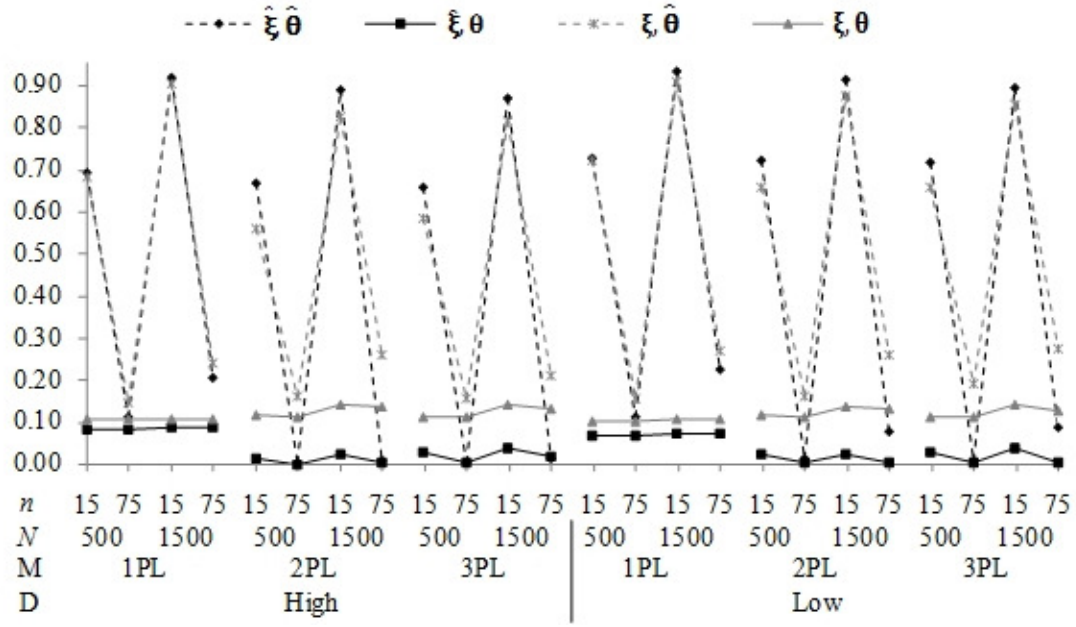


Figure D-11. Type I Error Rates at a = 0.10 for *V*O in SU Conditions by Parameter Estimation Error Condition

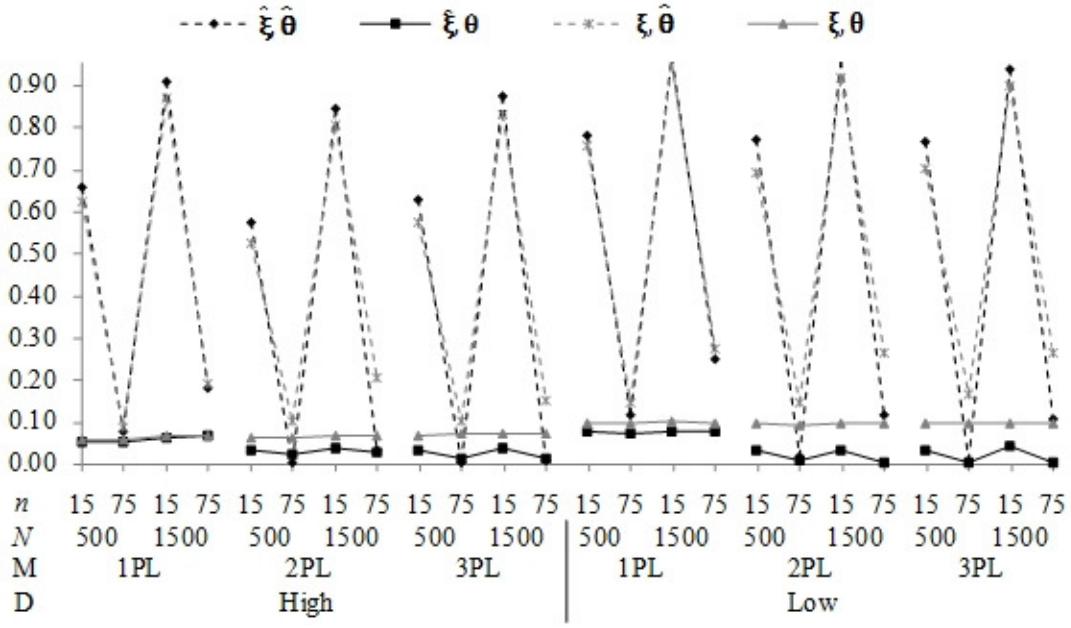


Figure D-12. Type I Error Rates at α = 0.10 for *V*O in EU Conditions by Parameter Estimation Error Condition

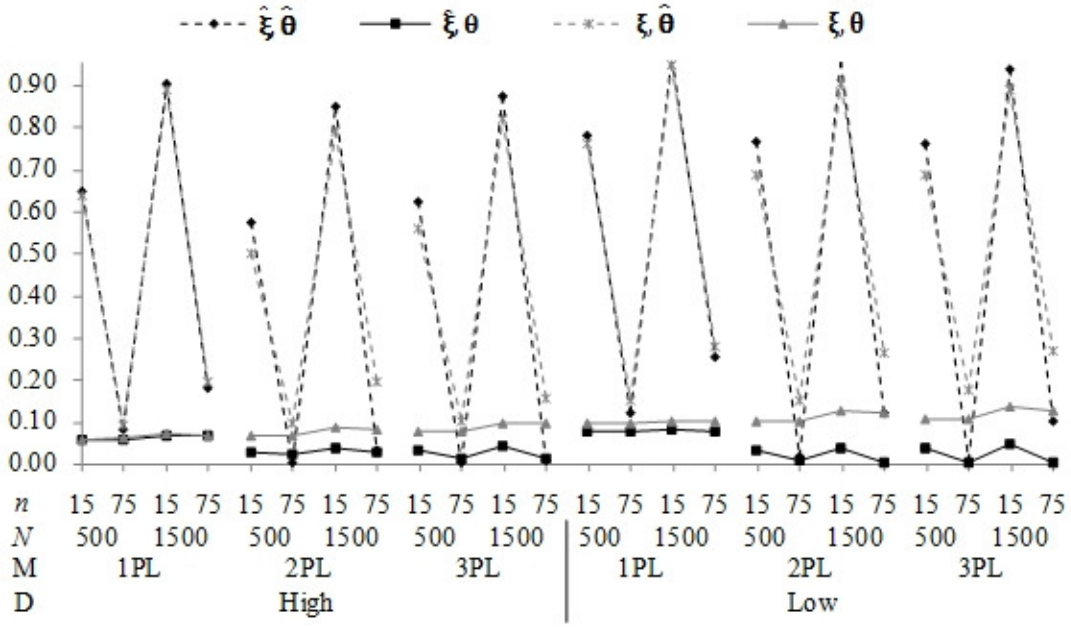


Table D-6. Type I Error Rates at $\alpha = 0.01$ for all SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.011	0.012	0.009	0.009	0.012
			75	0.013	0.012	0.010	0.010	0.011
		1,500	15	0.010	0.013	0.011	0.011	0.012
			75	0.011	0.013	0.010	0.010	0.013
	2PL	500	15	0.013	0.012	0.010	0.010	0.012
			75	0.013	0.011	0.010	0.010	0.011
		1,500	15	0.012	0.013	0.010	0.010	0.012
			75	0.012	0.013	0.010	0.010	0.011
	3PL	500	15	0.010	0.012	0.009	0.010	0.011
			75	0.010	0.012	0.010	0.010	0.012
		1,500	15	0.010	0.011	0.011	0.011	0.012
			75	0.010	0.011	0.010	0.010	0.013
Low	1PL	500	15	0.009	0.011	0.011	0.011	0.010
			75	0.011	0.011	0.009	0.009	0.010
		1,500	15	0.010	0.012	0.010	0.010	0.011
			75	0.011	0.012	0.010	0.010	0.010
	2PL	500	15	0.010	0.012	0.010	0.009	0.011
			75	0.011	0.009	0.011	0.011	0.011
		1,500	15	0.011	0.013	0.009	0.009	0.010
			75	0.010	0.012	0.010	0.010	0.010
	3PL	500	15	0.010	0.012	0.011	0.011	0.012
			75	0.009	0.010	0.009	0.010	0.010
		1,500	15	0.011	0.011	0.010	0.010	0.011
			75	0.010	0.011	0.010	0.011	0.010

Table D-7. Type I Error Rates at $\alpha = 0.01$ for all EU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.012	0.012	0.009	0.011	0.012
			75	0.012	0.012	0.011	0.011	0.014
		1,500	15	0.012	0.012	0.011	0.012	0.011
			75	0.012	0.014	0.012	0.013	0.012
	2PL	500	15	0.017	0.016	0.015	0.014	0.013
			75	0.017	0.013	0.014	0.013	0.013
		1,500	15	0.020	0.016	0.024	0.021	0.017
			75	0.016	0.017	0.021	0.020	0.015
	3PL	500	15	0.013	0.013	0.013	0.013	0.014
			75	0.012	0.009	0.013	0.012	0.015
		1,500	15	0.016	0.015	0.020	0.019	0.018
			75	0.016	0.015	0.019	0.018	0.018
Low	1PL	500	15	0.012	0.012	0.010	0.010	0.010
			75	0.012	0.011	0.011	0.009	0.011
		1,500	15	0.011	0.014	0.011	0.012	0.011
			75	0.011	0.012	0.011	0.011	0.012
	2PL	500	15	0.012	0.013	0.013	0.013	0.012
			75	0.011	0.010	0.014	0.013	0.014
		1,500	15	0.013	0.016	0.021	0.019	0.019
			75	0.013	0.014	0.019	0.017	0.018
	3PL	500	15	0.011	0.012	0.012	0.012	0.012
			75	0.012	0.010	0.013	0.012	0.014
		1,500	15	0.015	0.014	0.020	0.019	0.018
			75	0.015	0.012	0.019	0.019	0.017

Table D-8. Type I Error Rates at $\alpha = 0.05$ for all SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.051	0.052	0.048	0.048	0.031
			75	0.051	0.051	0.049	0.049	0.029
		1,500	15	0.049	0.053	0.049	0.050	0.035
			75	0.051	0.056	0.049	0.051	0.036
	2PL	500	15	0.051	0.050	0.048	0.050	0.032
			75	0.050	0.051	0.049	0.048	0.035
		1,500	15	0.052	0.054	0.051	0.051	0.036
			75	0.049	0.055	0.051	0.052	0.034
	3PL	500	15	0.048	0.054	0.049	0.051	0.036
			75	0.050	0.050	0.049	0.049	0.037
		1,500	15	0.048	0.053	0.051	0.050	0.039
			75	0.048	0.052	0.050	0.050	0.038
Low	1PL	500	15	0.047	0.050	0.052	0.051	0.048
			75	0.049	0.048	0.047	0.047	0.047
		1,500	15	0.050	0.049	0.052	0.053	0.048
			75	0.048	0.054	0.047	0.046	0.048
	2PL	500	15	0.049	0.053	0.049	0.048	0.048
			75	0.049	0.046	0.050	0.049	0.047
		1,500	15	0.050	0.052	0.049	0.048	0.047
			75	0.046	0.055	0.049	0.048	0.048
	3PL	500	15	0.051	0.051	0.051	0.051	0.050
			75	0.048	0.048	0.048	0.049	0.049
		1,500	15	0.050	0.053	0.048	0.049	0.047
			75	0.049	0.054	0.048	0.048	0.047

Table D-9. Type I Error Rates at $\alpha = 0.05$ for all EU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.052	0.049	0.048	0.051	0.032
			75	0.051	0.049	0.051	0.051	0.033
		1,500	15	0.054	0.054	0.055	0.057	0.037
			75	0.053	0.058	0.055	0.055	0.035
	2PL	500	15	0.061	0.058	0.059	0.059	0.038
			75	0.059	0.052	0.059	0.058	0.035
		1,500	15	0.072	0.065	0.082	0.078	0.048
			75	0.065	0.063	0.073	0.071	0.044
	3PL	500	15	0.058	0.052	0.059	0.057	0.042
			75	0.053	0.050	0.060	0.058	0.042
		1,500	15	0.068	0.059	0.076	0.074	0.054
			75	0.067	0.060	0.072	0.071	0.053
Low	1PL	500	15	0.047	0.053	0.052	0.052	0.049
			75	0.050	0.050	0.050	0.051	0.050
		1,500	15	0.052	0.057	0.053	0.052	0.051
			75	0.053	0.054	0.055	0.056	0.052
	2PL	500	15	0.053	0.057	0.059	0.060	0.054
			75	0.054	0.051	0.059	0.059	0.055
		1,500	15	0.059	0.064	0.077	0.075	0.071
			75	0.058	0.057	0.072	0.070	0.064
	3PL	500	15	0.053	0.053	0.058	0.059	0.053
			75	0.052	0.048	0.058	0.056	0.055
		1,500	15	0.064	0.059	0.077	0.076	0.073
			75	0.062	0.057	0.069	0.070	0.067

Table D-10. Type I Error Rates at $\alpha = 0.10$ for all SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.098	0.099	0.099	0.098	0.057
			75	0.098	0.097	0.099	0.101	0.055
		1,500	15	0.095	0.103	0.097	0.100	0.065
			75	0.100	0.106	0.100	0.101	0.069
	2PL	500	15	0.096	0.097	0.097	0.100	0.061
			75	0.096	0.100	0.100	0.100	0.061
		1,500	15	0.100	0.104	0.099	0.098	0.065
			75	0.095	0.104	0.103	0.102	0.067
	3PL	500	15	0.097	0.102	0.097	0.098	0.067
			75	0.097	0.102	0.100	0.099	0.070
		1,500	15	0.098	0.104	0.102	0.103	0.074
			75	0.094	0.102	0.098	0.099	0.072
Low	1PL	500	15	0.096	0.096	0.100	0.101	0.096
			75	0.099	0.096	0.096	0.097	0.094
		1,500	15	0.098	0.100	0.100	0.100	0.100
			75	0.094	0.104	0.096	0.097	0.095
	2PL	500	15	0.093	0.100	0.098	0.100	0.095
			75	0.099	0.094	0.096	0.097	0.092
		1,500	15	0.097	0.099	0.098	0.096	0.097
			75	0.096	0.104	0.098	0.097	0.095
	3PL	500	15	0.100	0.100	0.103	0.104	0.096
			75	0.098	0.097	0.102	0.102	0.096
		1,500	15	0.098	0.100	0.101	0.100	0.096
			75	0.099	0.105	0.098	0.098	0.096

Table D-11. Type I Error Rates at $\alpha = 0.10$ for all EU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	Model	N	n	Fit Statistic				
				$Q1$	QO	L_z	VI	VO
High	1PL	500	15	0.097	0.093	0.105	0.104	0.059
			75	0.098	0.096	0.104	0.103	0.060
		1,500	15	0.104	0.100	0.107	0.109	0.070
			75	0.105	0.107	0.107	0.107	0.068
	2PL	500	15	0.113	0.110	0.115	0.115	0.069
			75	0.112	0.099	0.112	0.111	0.065
		1,500	15	0.129	0.121	0.143	0.143	0.084
			75	0.121	0.116	0.134	0.134	0.079
	3PL	500	15	0.111	0.102	0.109	0.112	0.076
			75	0.104	0.099	0.110	0.112	0.075
		1,500	15	0.125	0.108	0.138	0.134	0.096
			75	0.122	0.112	0.132	0.130	0.094
Low	1PL	500	15	0.096	0.102	0.101	0.103	0.096
			75	0.101	0.096	0.101	0.102	0.097
		1,500	15	0.104	0.110	0.107	0.108	0.101
			75	0.105	0.109	0.108	0.109	0.103
	2PL	500	15	0.106	0.108	0.114	0.115	0.103
			75	0.102	0.097	0.113	0.114	0.102
		1,500	15	0.115	0.118	0.137	0.136	0.126
			75	0.116	0.108	0.129	0.128	0.119
	3PL	500	15	0.105	0.101	0.110	0.110	0.106
			75	0.105	0.099	0.113	0.113	0.108
		1,500	15	0.120	0.114	0.140	0.139	0.134
			75	0.118	0.108	0.128	0.128	0.124

APPENDIX E: ANALYSIS OF *Q*O DISTRIBUTION

Number of Items Within Each of the *K* *Q*O Groups

Relationship Between *K* and Item Parameters

Detrended *Q*O Sample Means by *K* With 95% Confidence Intervals

Detrended *Q*O Sample Variancess by *K* With 95% Confidence Intervals

Frequency of KS Test Rejections in Parameter Estimation Error Conditions Aggregated Across Discrimination Conditions

Bias of *Q*O Sampling Distribution Means and SDs

Mean Error for *Q*O Sampling Distribution Means and SDs

Estimates of ME(SD) and 95% CIs About the Estimates for *Q*O

Table E-1. Number of Items Within Each K by Model and DN for High Discrimination, $N = 500$, and $n = 15$ Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	2	1	0	0
2	2	1	3	2	0	0
3	5	4	6	11	1	1
4	5	4	20	16	0	0
5	6	14	32	33	1	1
6	38	41	97	89	3	3
7	91	103	169	156	17	14
8	174	211	311	316	104	113
9	485	501	639	683	347	381
10	973	1076	1627	1596	1130	1229
11	2776	2947	3690	3896	3485	3673
12	7386	7607	5995	5899	7821	7832
13	6297	5811	4743	4727	5800	5471
14	512	430	1416	1325	41	32
TOTAL	18750	18750	18750	18750	18750	18750

Table E-2. Number of Items Within Each K by Model and DN for Low Discrimination, $N = 500$, and $n = 15$ Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
5	0	0	1	0	0	0
6	0	0	0	2	0	2
7	1	1	1	2	2	0
8	4	3	5	8	2	1
9	7	8	38	41	34	31
10	78	67	274	287	624	594
11	1194	1172	2765	2910	6334	6527
12	10475	10411	9423	9531	9960	9891
13	6902	7028	5751	5493	1794	1704
14	89	60	492	476	0	0
TOTAL	18750	18750	18750	18750	18750	18750

Table E-3. Number of Items Within Each K by Model and DN for High Discrimination, N = 1,500, and n = 15 Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	2	1	0	0
4	1	0	2	1	0	0
5	0	2	5	9	0	0
6	1	2	18	17	0	0
7	7	11	34	37	2	3
8	19	26	88	71	3	4
9	52	56	187	165	24	27
10	145	172	419	453	149	139
11	523	580	1147	1126	633	713
12	2316	2469	3347	3482	3008	3095
13	8343	8500	7231	7266	11156	11464
14	7343	6932	6270	6122	3775	3305
TOTAL	18750	18750	18750	18750	18750	18750

Table E-4. Number of Items Within Each K by Model and DN for Low Discrimination, N = 1,500, and n = 15 Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	2	0
8	0	0	0	0	3	0
9	0	0	4	2	24	0
10	0	0	19	18	149	6
11	26	24	158	171	633	206
12	713	668	1984	1985	3008	4427
13	9693	9685	9642	9835	11156	13697
14	8318	8373	6943	6739	3775	414
TOTAL	18750	18750	18750	18750	18750	18750

Table E-5. Number of Items Within Each K by Model and DN for High Discrimination, N = 500, and n = 75 Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
2	0	0	1	0	0	0
3	0	0	0	1	1	1
4	0	0	0	0	0	0
5	0	2	3	2	0	1
6	1	1	3	2	0	0
7	5	2	7	4	2	1
8	2	3	4	9	1	1
9	4	7	11	10	0	3
10	5	5	10	6	1	2
11	8	12	14	20	2	5
12	14	9	19	17	7	7
13	11	11	14	19	4	4
14	14	15	28	26	6	5
15	14	27	32	22	7	12
16	21	22	29	22	9	9
17	27	33	35	49	14	12
18	35	41	47	42	14	12
19	42	46	46	63	20	18
20	51	40	71	58	20	22
21	52	57	62	70	22	29
22	68	63	85	81	36	35
23	80	66	72	82	42	46
24	71	93	98	105	61	48
25	112	84	117	103	64	77
26	94	110	122	126	79	75
27	132	104	154	163	98	78
28	119	164	167	193	107	114
29	168	158	197	188	131	124
30	172	181	196	214	161	162
31	204	172	290	246	156	189
32	220	236	306	267	172	211
33	259	249	343	364	244	227
34	269	260	367	384	303	272
35	338	320	441	463	330	327
36	372	374	532	554	338	332

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
37	416	443	565	571	356	332
38	470	477	682	701	428	414
39	519	586	789	754	486	511
40	629	654	854	853	536	563
41	751	784	1013	990	611	609
42	947	883	976	1024	634	733
43	1232	1197	994	959	752	756
44	1430	1484	1020	1058	776	899
45	1873	1903	937	982	850	986
46	2344	2122	906	948	1044	974
47	2102	1980	901	861	1127	1075
48	1591	1640	838	742	1254	1188
49	935	1042	706	693	1229	1214
50	358	427	635	608	1185	1306
51	131	136	546	552	1202	1337
52	33	19	443	495	1148	1080
53	5	5	419	442	1065	1004
54	0	1	364	323	763	674
55	0	0	269	280	490	354
56	0	0	206	232	227	190
57	0	0	190	168	98	73
58	0	0	159	163	26	13
59	0	0	115	113	10	4
60	0	0	98	92	1	0
61	0	0	62	66	0	0
62	0	0	57	39	0	0
63	0	0	18	32	0	0
64	0	0	26	19	0	0
65	0	0	17	9	0	0
66	0	0	12	5	0	0
67	0	0	7	1	0	0
68	0	0	3	0	0	0
TOTAL	18750	18750	18750	18750	18750	18750

Table E-6. Number of Items Within Each K by Model and DN for Low Discrimination, $N = 500$, and $n = 75$ Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
4	0	0	0	0	1	1
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	1
12	0	0	1	0	0	1
13	0	0	1	0	1	0
14	0	0	0	2	1	1
15	0	1	1	1	0	0
16	0	0	0	1	0	1
17	0	0	1	2	2	0
18	0	0	2	3	0	1
19	0	1	5	2	0	4
20	2	2	7	4	2	4
21	3	2	8	10	5	1
22	3	5	6	10	1	4
23	2	3	11	6	5	3
24	3	3	10	8	6	9
25	6	2	12	9	6	5
26	6	8	13	11	13	8
27	6	4	11	19	16	16
28	12	8	17	18	14	17
29	5	12	14	21	22	29
30	15	16	29	31	54	39
31	22	20	46	45	49	43
32	20	21	47	44	56	64
33	35	42	61	64	94	112
34	35	29	67	68	118	157
35	36	58	100	102	167	165
36	66	71	111	139	247	255
37	75	79	161	192	380	341
38	90	122	239	193	511	522

K	Study Conditions					
	1PL		2PL		3PL	
	SU	EU	SU	EU	SU	EU
	N_K	N_K	N_K	N_K	N_K	N_K
39	146	143	287	282	666	663
40	183	172	441	398	853	934
41	266	211	637	601	1195	1241
42	374	354	805	829	1554	1625
43	486	510	1089	1115	1916	1983
44	800	797	1362	1441	2310	2215
45	1296	1178	1615	1669	2318	2478
46	2187	1993	1763	1753	2116	2151
47	2920	3300	1792	1843	1826	1776
48	3616	3794	1788	1718	1228	1060
49	3350	3161	1582	1590	627	511
50	1697	1721	1377	1330	276	220
51	766	729	1033	1006	73	73
52	203	160	827	813	21	14
53	18	18	568	544	0	2
54	0	0	392	355	0	0
55	0	0	203	243	0	0
56	0	0	109	108	0	0
57	0	0	60	66	0	0
58	0	0	26	25	0	0
59	0	0	10	12	0	0
60	0	0	2	3	0	0
61	0	0	1	0	0	0
62	0	0	0	1	0	0
TOTAL	18750	18750	18750	18750	18750	18750

Table E-7. Number of Items Within Each K by Model and DN for High Discrimination, N = 1,500, and n = 75 Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
2	0	0	0	1	0	0
3	0	0	0	0	0	0
4	1	0	0	0	0	0
5	0	1	1	0	0	0
6	0	0	2	0	0	0
7	0	0	1	2	0	0
8	1	0	0	1	0	0
9	0	0	0	1	0	0
10	0	0	2	1	0	0
11	0	0	0	0	0	0
12	0	1	1	1	0	0
13	1	3	2	3	0	0
14	1	0	4	3	0	0
15	1	2	2	2	0	0
16	0	0	4	8	0	0
17	0	3	2	2	1	1
18	1	2	7	3	1	1
19	3	4	9	7	0	1
20	3	1	8	10	1	1
21	1	1	2	4	1	0
22	4	3	9	6	3	4
23	5	4	7	13	4	6
24	5	5	8	6	6	4
25	5	4	14	8	5	8
26	4	10	9	11	8	9
27	7	8	9	15	10	9
28	9	11	17	20	13	4
29	10	11	24	16	12	24
30	14	12	29	24	26	17
31	18	10	35	26	27	26
32	14	26	30	32	27	25
33	31	26	29	35	44	38
34	21	26	48	34	45	47
35	34	32	53	33	50	45
36	38	36	49	64	64	59

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
37	30	33	67	59	65	62
38	40	34	83	81	82	91
39	55	45	85	94	91	102
40	57	66	109	117	113	122
41	54	78	106	108	160	148
42	92	90	102	129	172	160
43	94	103	164	147	165	190
44	137	112	166	147	225	225
45	145	126	174	186	279	259
46	145	177	252	208	346	330
47	180	212	282	278	335	392
48	230	200	351	348	447	423
49	260	260	386	435	510	563
50	270	281	417	466	629	659
51	366	328	506	571	685	722
52	377	400	625	679	814	886
53	453	470	758	794	1076	973
54	586	598	871	865	1099	1192
55	709	706	963	970	1286	1391
56	877	937	1062	1011	1571	1557
57	1157	1114	1078	1118	1749	1802
58	1535	1602	1129	1178	1931	1972
59	2080	2026	1115	1126	1802	1880
60	2443	2674	1052	1024	1586	1433
61	2632	2650	1086	1032	900	681
62	1993	1918	954	1028	260	184
63	1114	924	922	856	24	20
64	349	308	814	775	0	2
65	54	36	674	703	0	0
66	4	0	545	550	0	0
67	0	0	501	414	0	0
68	0	0	355	363	0	0
69	0	0	238	232	0	0
70	0	0	188	147	0	0
71	0	0	113	75	0	0
72	0	0	34	39	0	0
73	0	0	6	5	0	0

K	Study Conditions					
	1PL		2PL		3PL	
	SU	EU	SU	EU	SU	EU
	N_K	N_K	N_K	N_K	N_K	N_K
TOTAL	18750	18750	18750	18750	18750	18750

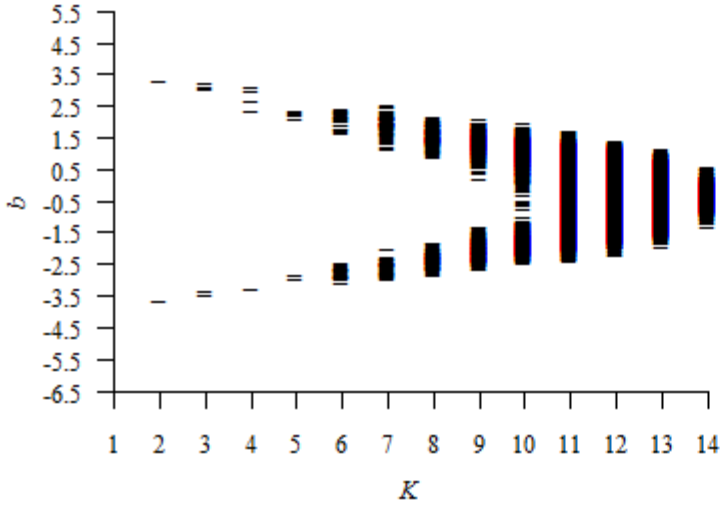
Table E-8. Number of Items Within Each K by Model and DN for Low Discrimination, N = 1,500, and n = 75 Conditions

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
17	0	0	1	0	0	0
18	0	0	0	0	0	0
19	0	0	0	0	0	0
20	0	0	0	0	0	0
21	0	0	1	0	0	0
22	1	0	0	0	0	0
23	0	0	0	0	0	0
24	0	0	0	1	0	0
25	0	0	0	0	0	0
26	0	0	0	2	0	0
27	0	0	0	0	0	0
28	0	1	1	0	0	1
29	0	0	0	0	1	1
30	0	0	0	2	0	0
31	0	0	1	1	0	2
32	0	0	3	1	0	1
33	0	1	2	2	1	0
34	1	0	3	1	6	0
35	0	0	0	2	5	5
36	0	0	1	3	4	6
37	0	0	3	2	12	8
38	1	0	4	8	11	17
39	0	0	5	5	9	19
40	0	0	7	7	37	31
41	1	0	8	9	45	47
42	0	1	10	16	76	64
43	0	4	21	23	95	121
44	3	4	24	22	188	161
45	3	8	30	24	264	266
46	7	5	55	52	330	385
47	6	9	74	81	552	591
48	23	18	108	113	848	850
49	27	25	191	184	1283	1431
50	47	49	291	265	1951	1920
51	71	64	448	424	2542	2722

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
52	121	144	641	672	3024	3046
53	219	208	908	1012	3054	3022
54	360	385	1290	1323	2384	2329
55	724	762	1704	1702	1336	1152
56	1469	1542	1965	2060	530	446
57	2831	2751	2239	2248	143	92
58	4855	4851	2153	2146	18	14
59	4915	4646	1899	1984	1	0
60	2562	2533	1663	1563	0	0
61	460	672	1250	1192	0	0
62	43	67	836	747	0	0
63	0	0	490	492	0	0
64	0	0	272	247	0	0
65	0	0	101	82	0	0
66	0	0	46	26	0	0
67	0	0	1	3	0	0
68	0	0	0	1	0	0
TOTAL	18750	18750	18750	18750	18750	18750

Figure E-1. Relationship Between *K* and Item *b* Parameters for the 1PL (*N* = 500 *n* = 15)

a. SU High Discrimination



b. SU Low Discrimination

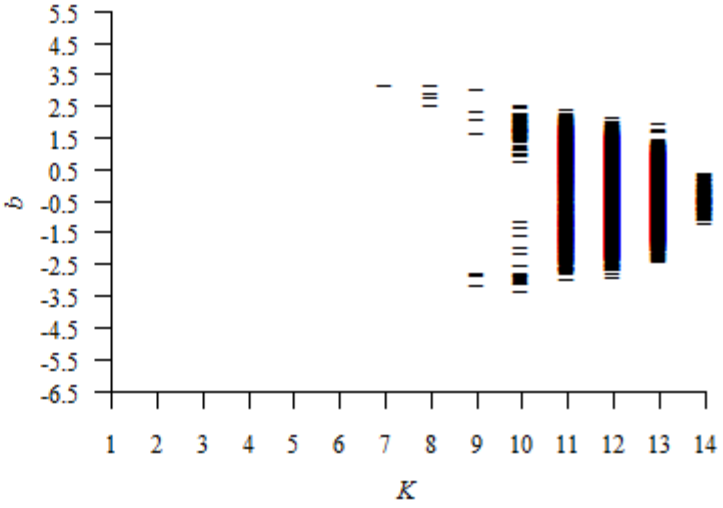


Figure E-2. Relationship Between *K* and Item *b* and *a* Parameters for the 2PL (*N* = 500 *n* = 15)

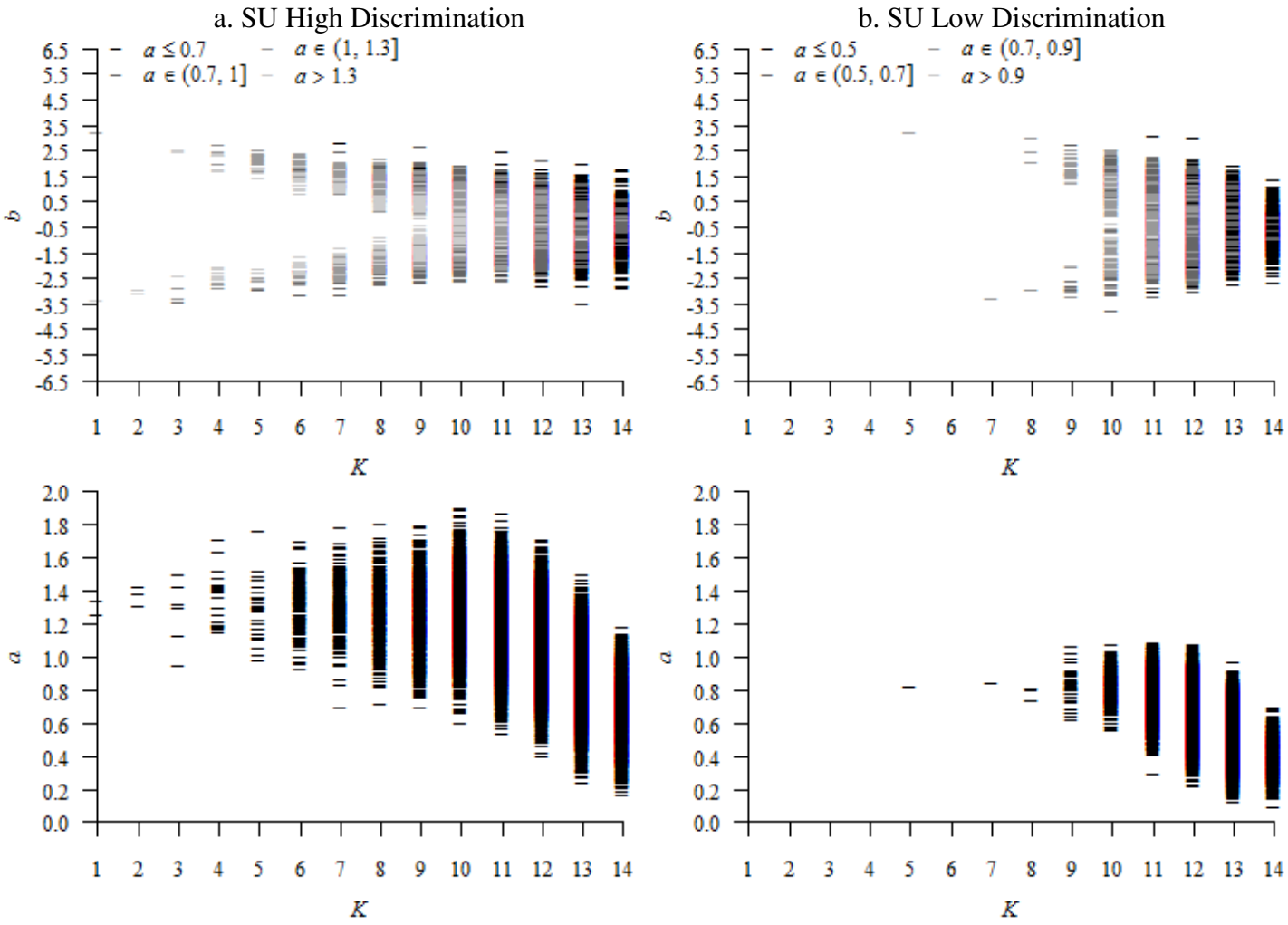


Figure E-3. Relationship Between *K* and Item *b* and *a* Parameters for the 3PL (*N* = 500 *n* = 15)

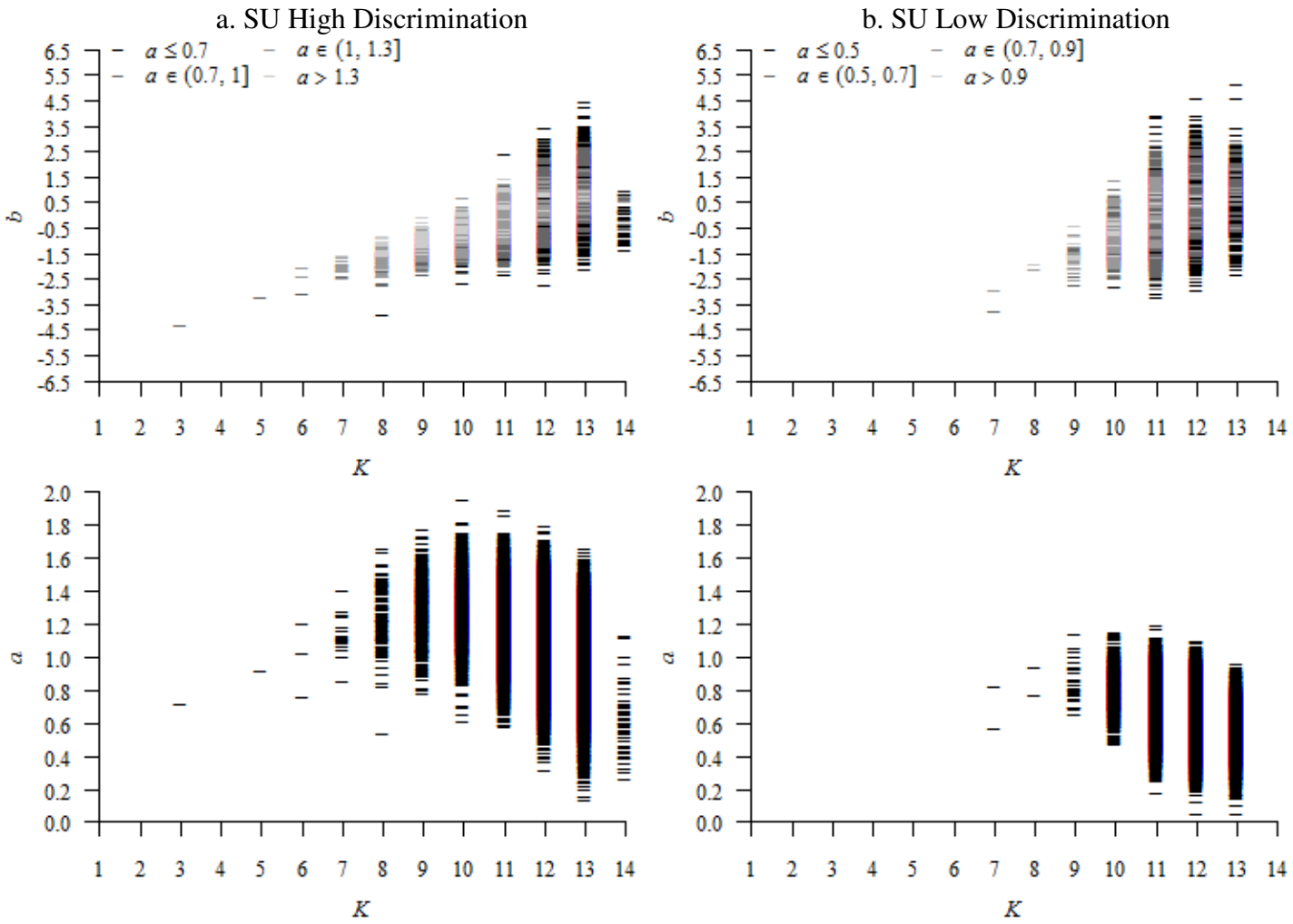
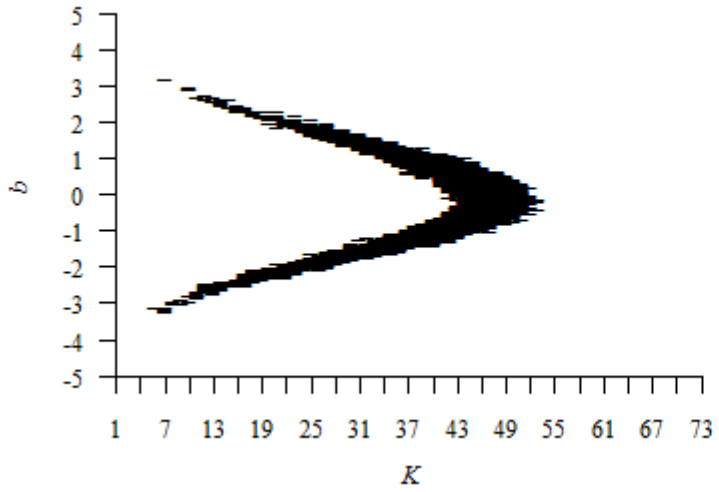


Figure E-4. Relationship Between *K* and Item *b* Parameters for the 1PL (*N* = 500 *n* = 75)

a. SU High Discrimination



b. SU Low Discrimination

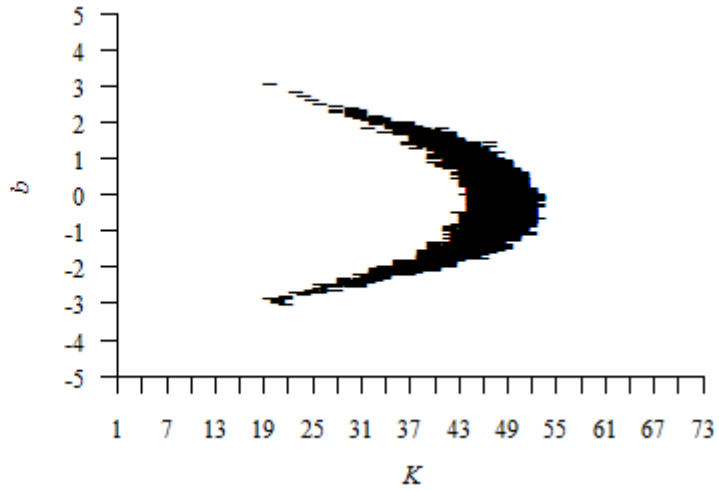


Figure E-5. Relationship Between *K* and Item *b* and *a* Parameters for the 2PL (*N* = 500 *n* = 75)

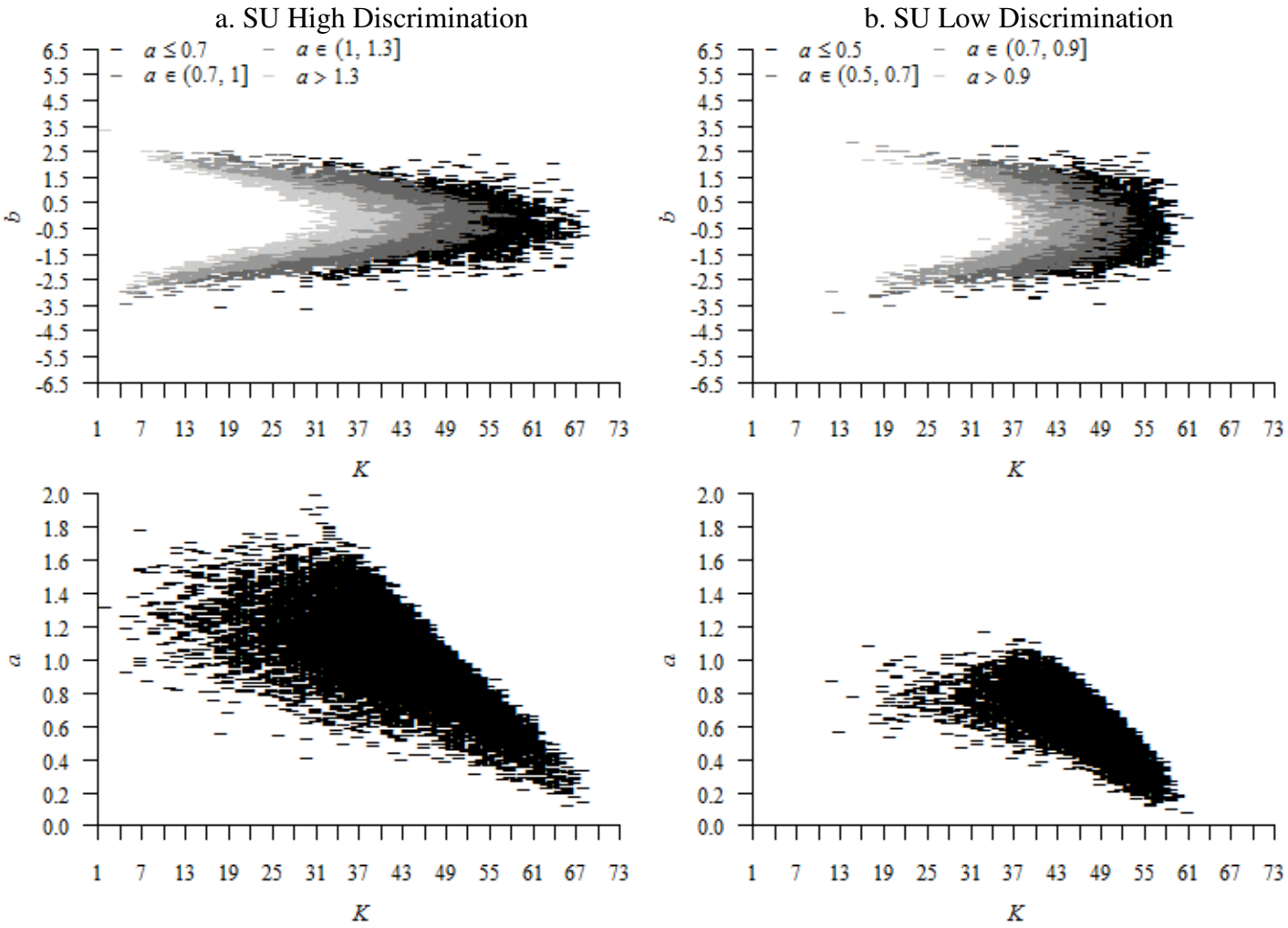


Figure E-6. Relationship Between *K* and Item *b* and *a* Parameters for the 3PL (*N* = 500 *n* = 75)

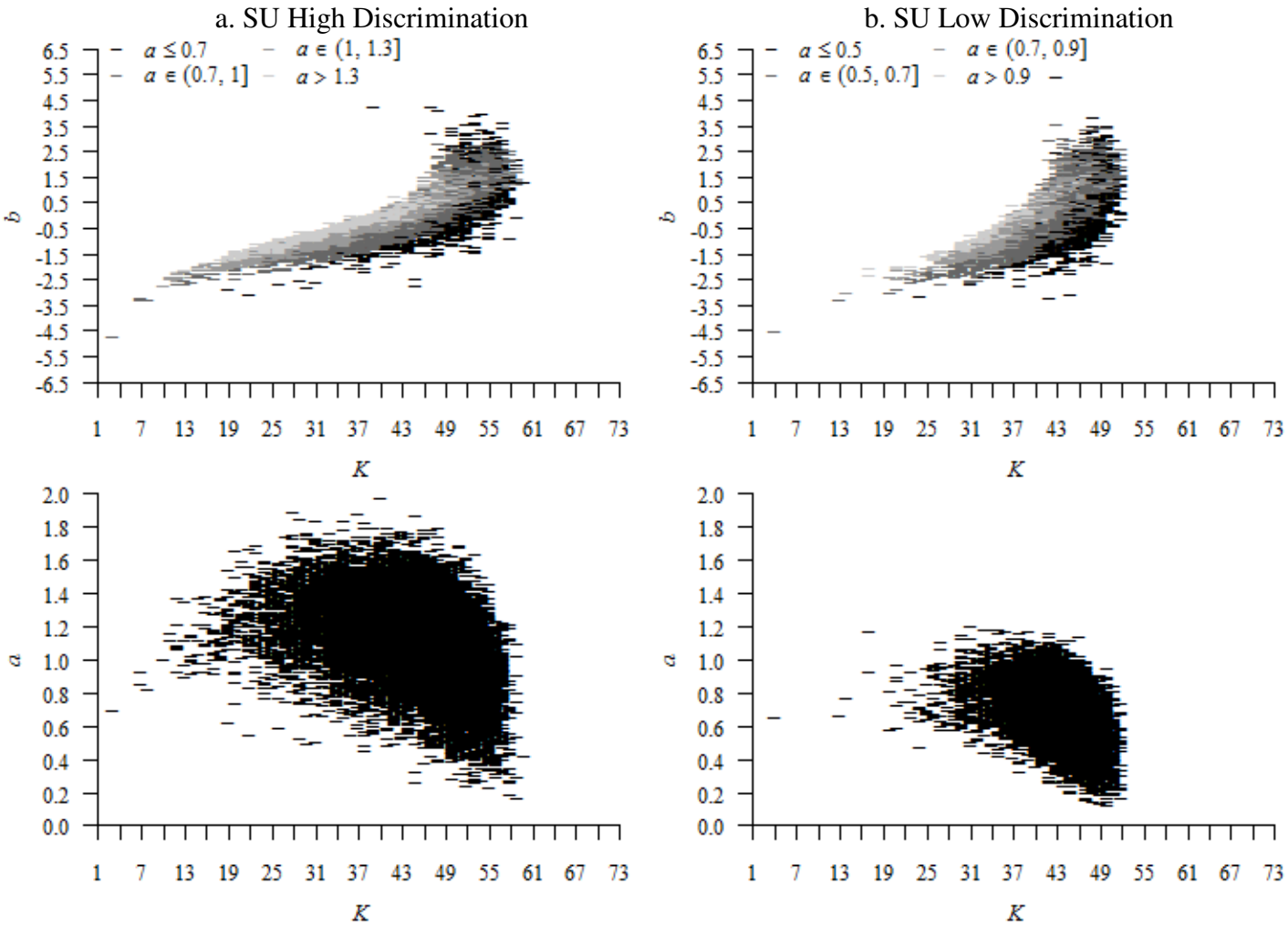
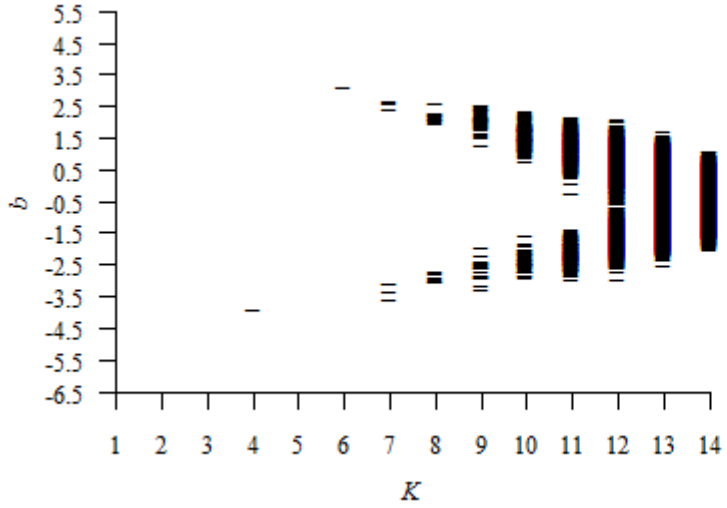


Figure E-7. Relationship Between *K* and Item *b* Parameters for the 1PL (*N* = 1,500 *n* = 15)

a. SU High Discrimination



b. SU Low Discrimination

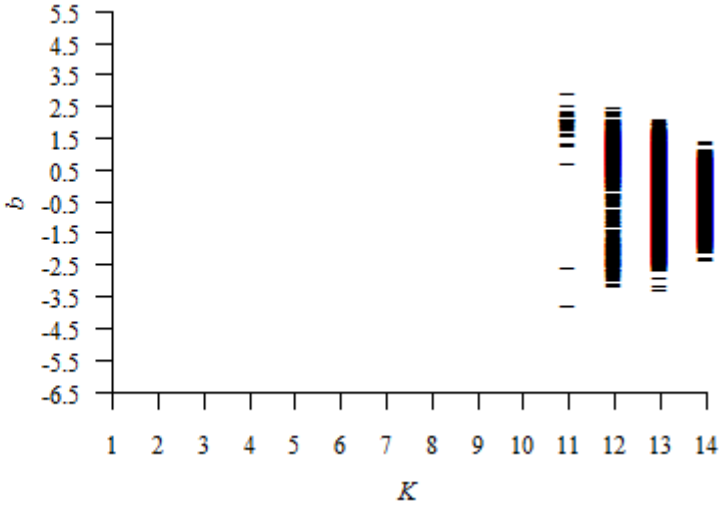


Figure E-8. Relationship Between *K* and Item *b* and *a* Parameters for the 2PL (*N* = 1,500 *n* = 15)

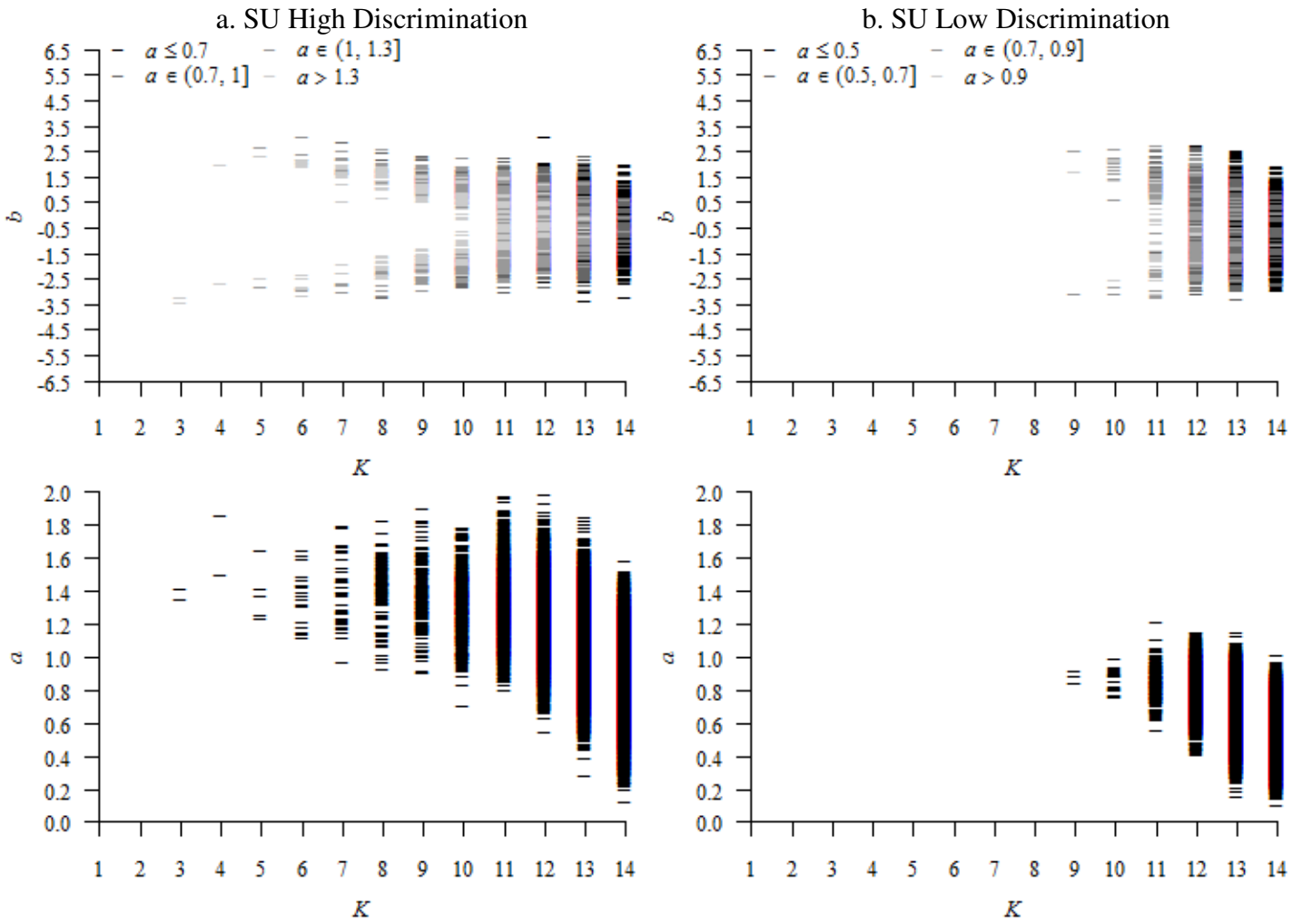


Figure E-9. Relationship Between *K* and Item *b* and *a* Parameters for the 3PL (*N* = 1,500 *n* = 15)

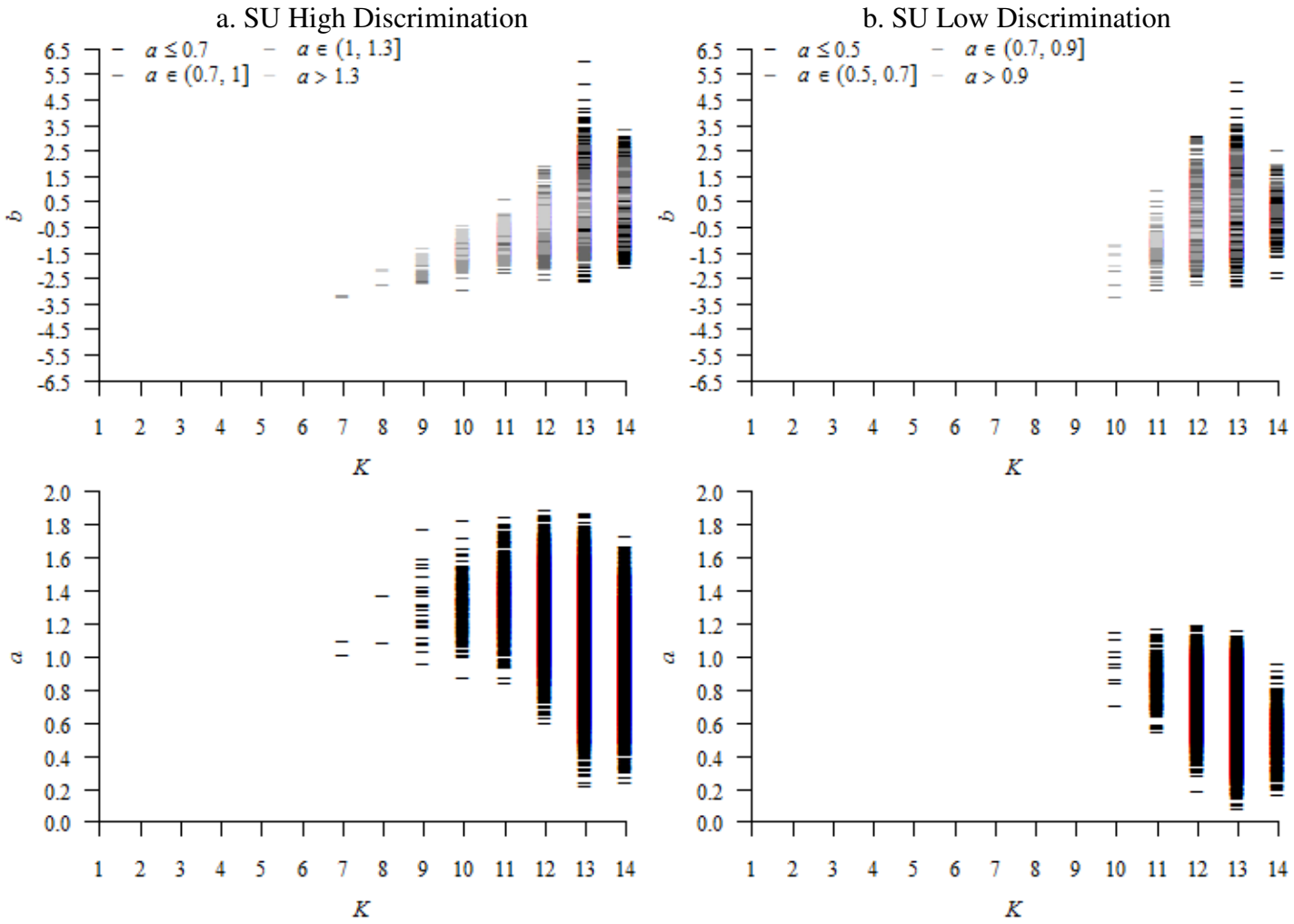
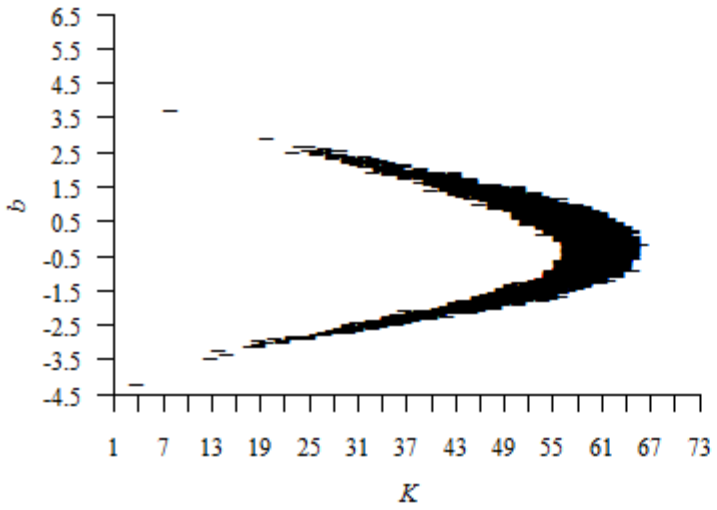


Figure E-10. Relationship Between *K* and Item *b* Parameters for the 1PL (*N* = 1,500 *n* = 75)

a. SU High Discrimination



b. SU Low Discrimination

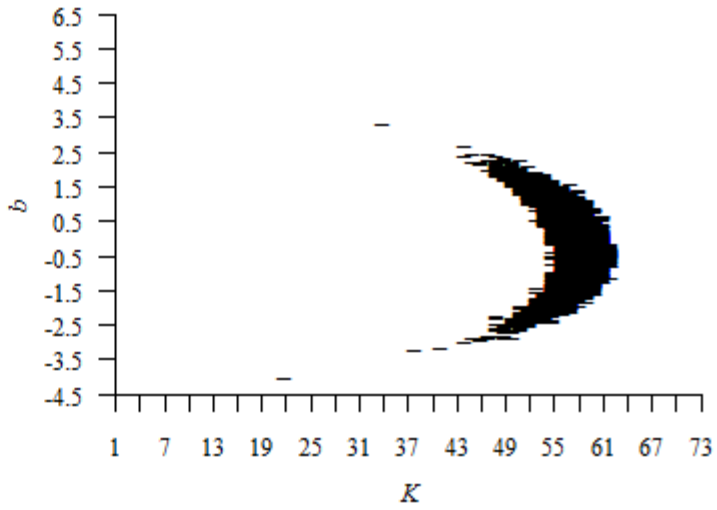


Figure E-11. Relationship Between *K* and Item *b* and *a* Parameters for the 2PL (*N* = 1,500 *n* = 75)

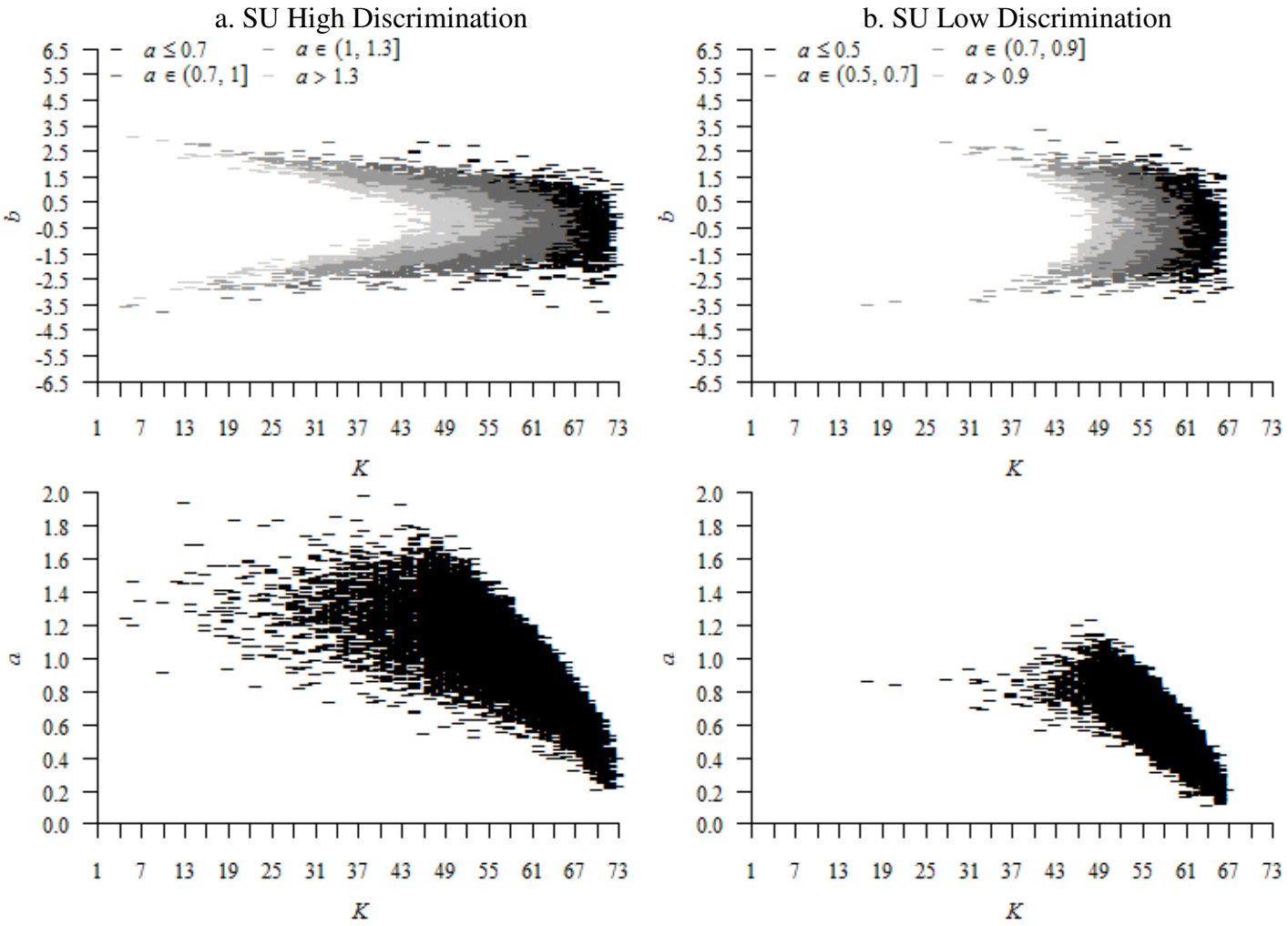


Figure E-12. Relationship Between *K* and Item *b* and *a* Parameters for the 3PL (*N* = 1,500 *n* = 75)

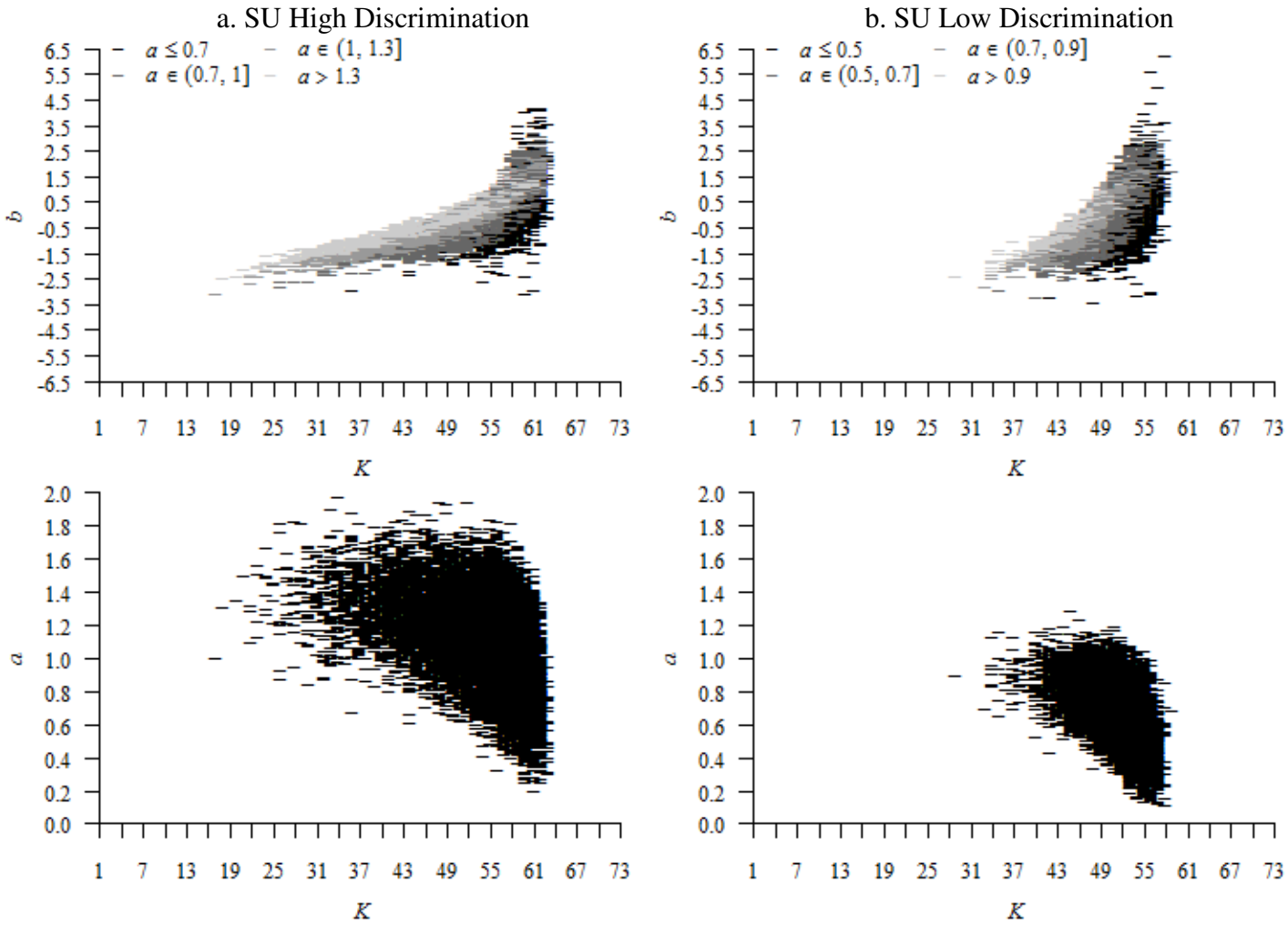


Figure E-13. De-trended *QO* Means by *K* for SU Low Discrimination *N* = 500 *n* = 15 Condition

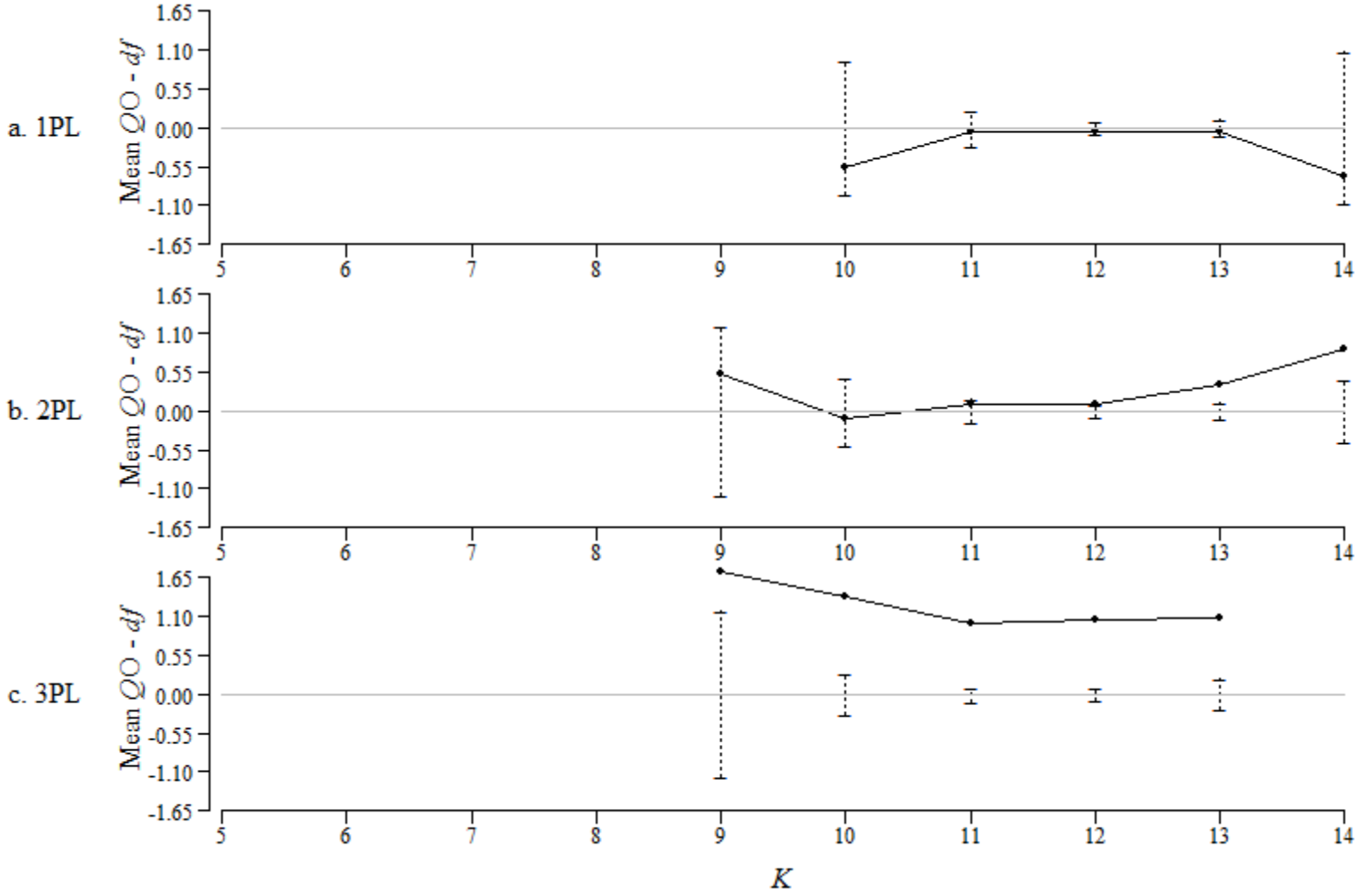


Figure E-14. De-trended *Q*O Means by *K* for SU Low Discrimination *N* = 1,500 *n* = 15 Condition

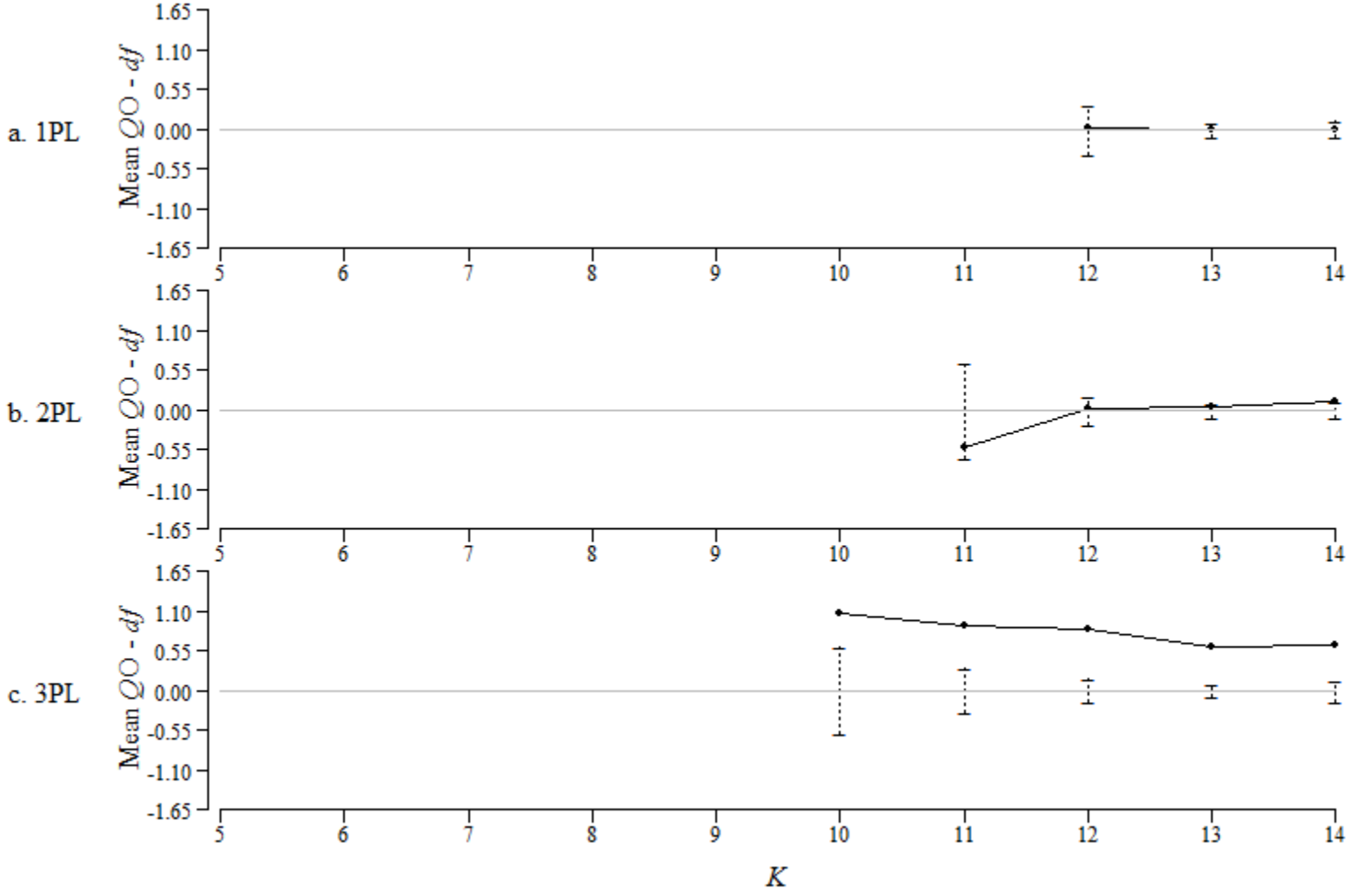


Figure E-15. De-trended *Q*O Means by *K* for SU Low Discrimination *N* = 500 *n* = 75 Condition

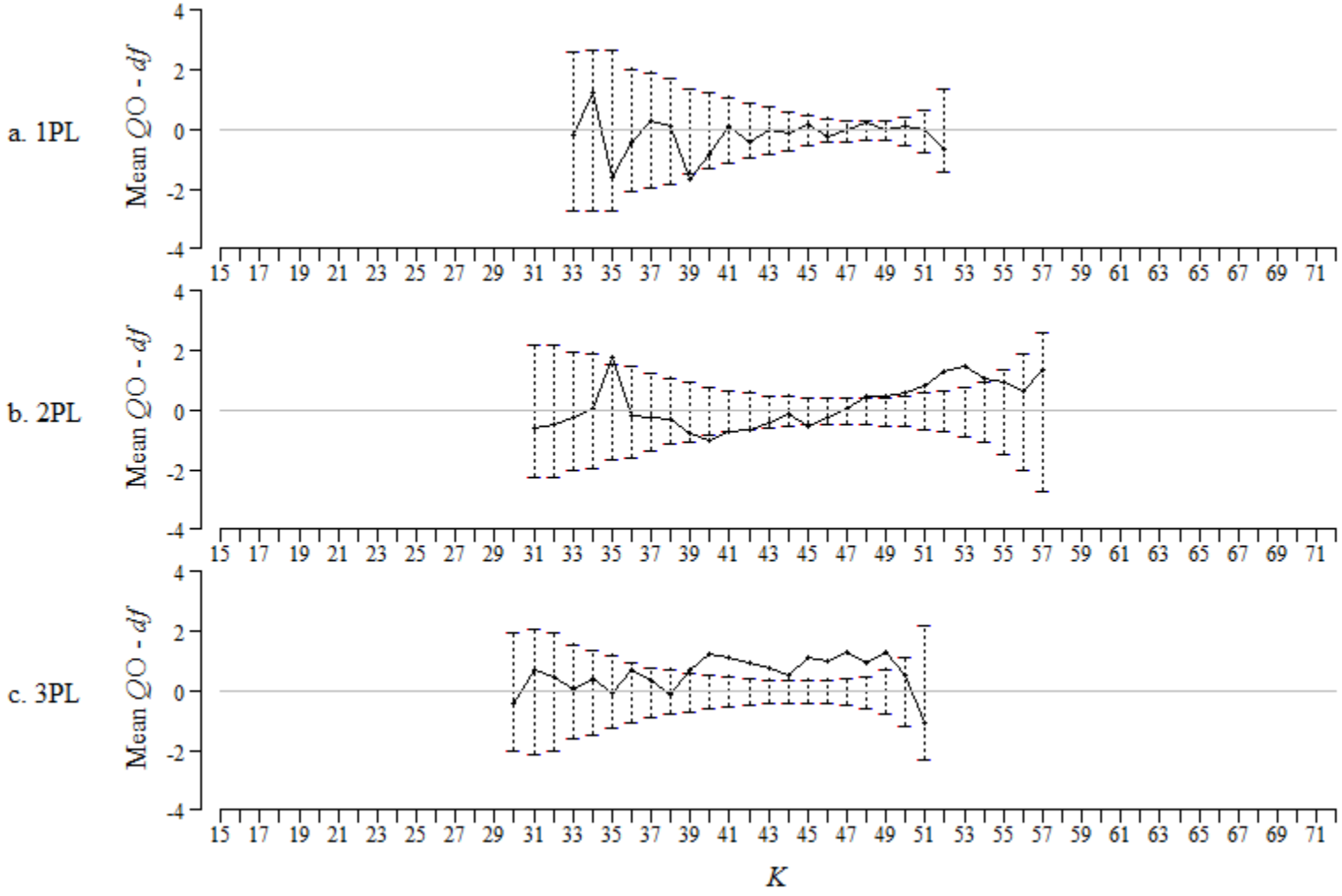


Figure E-16. De-trended *QO* Means by *K* for SU Low Discrimination *N* = 1,500 *n* = 75 Condition

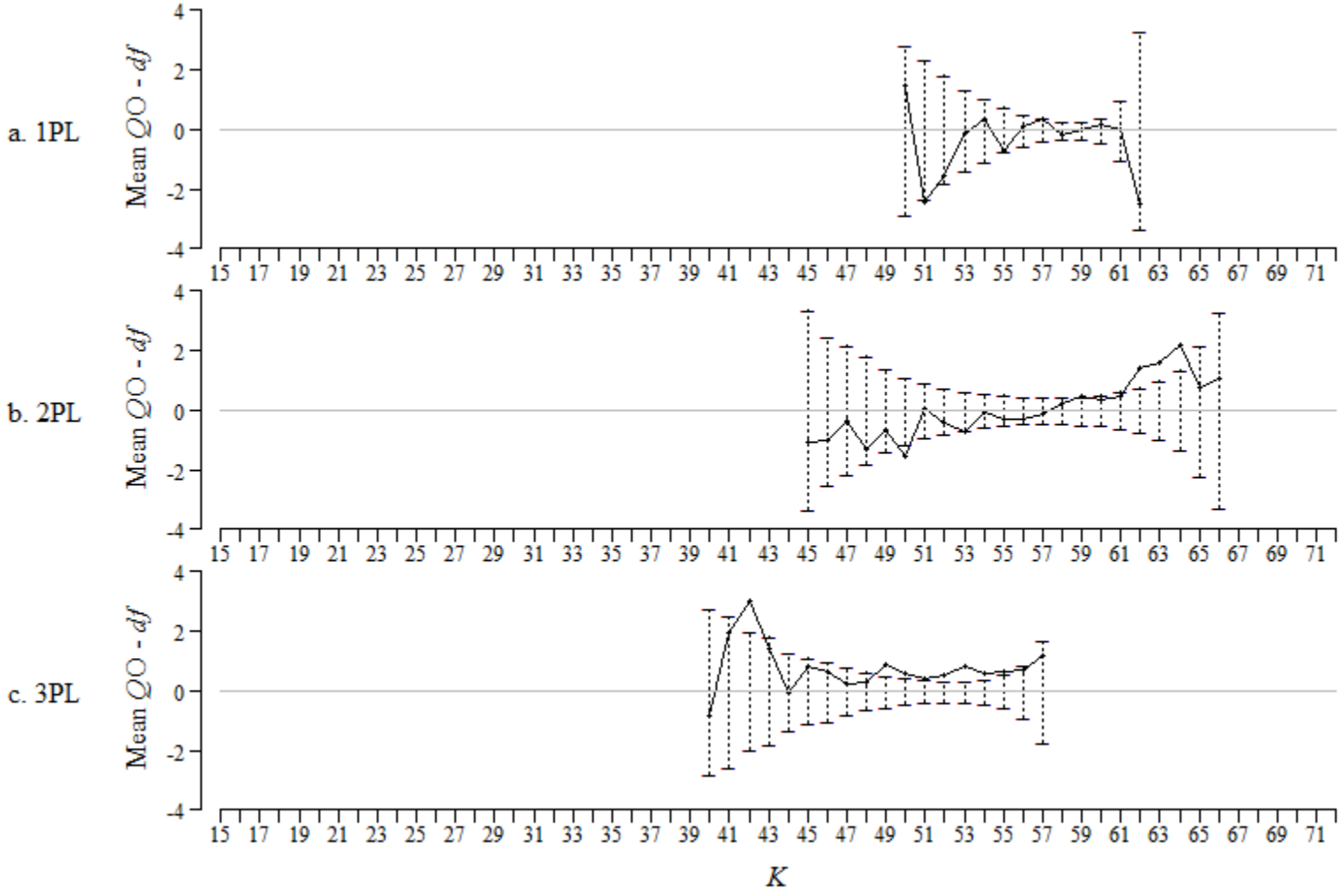


Figure E-17. De-trended *Q*O Means by *K* for EU High Discrimination *N* = 500 *n* = 15 Condition

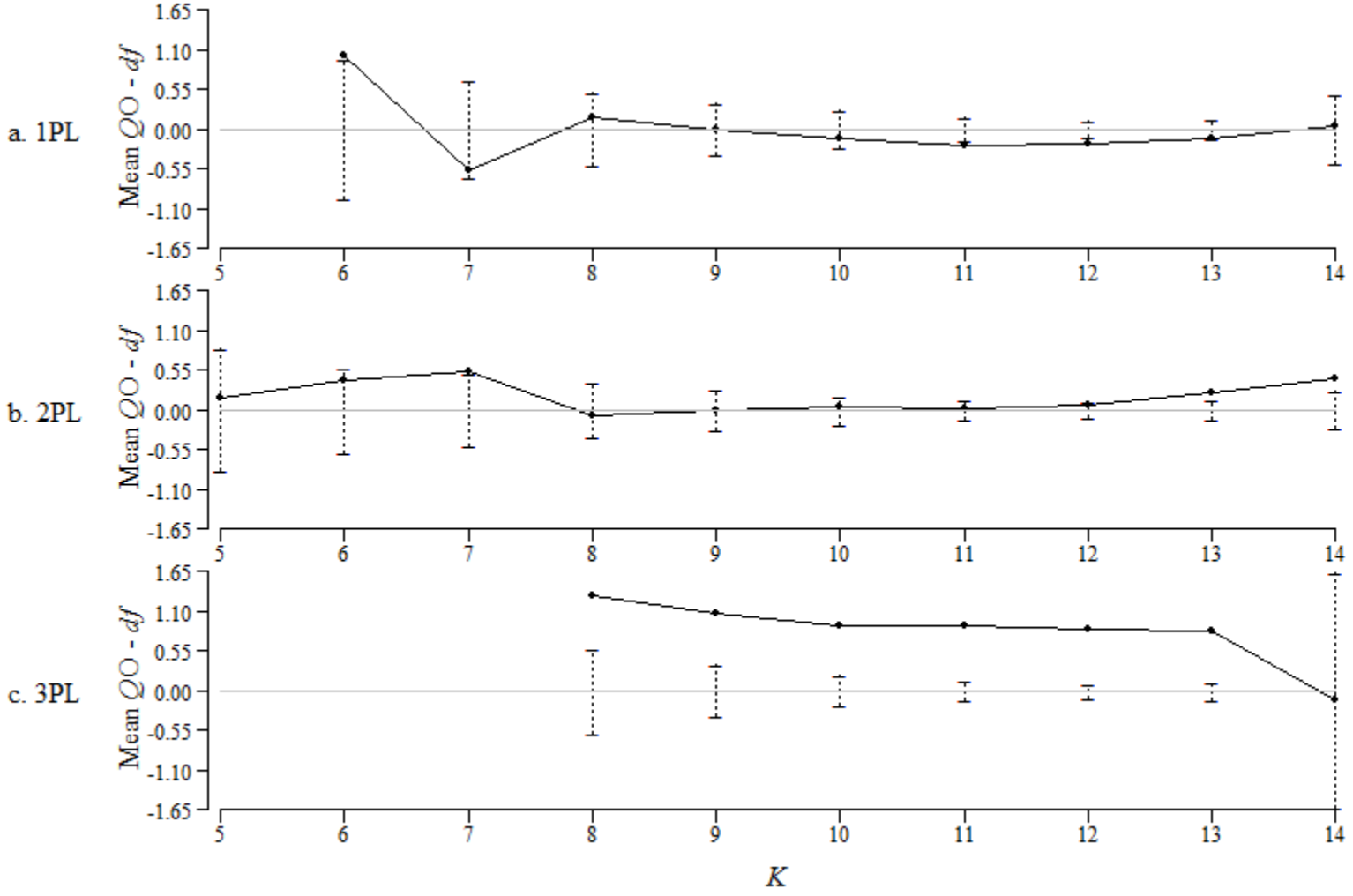


Figure E-18. De-trended *QO* Means by *K* for EU Low Discrimination *N* = 500 *n* = 15 Condition

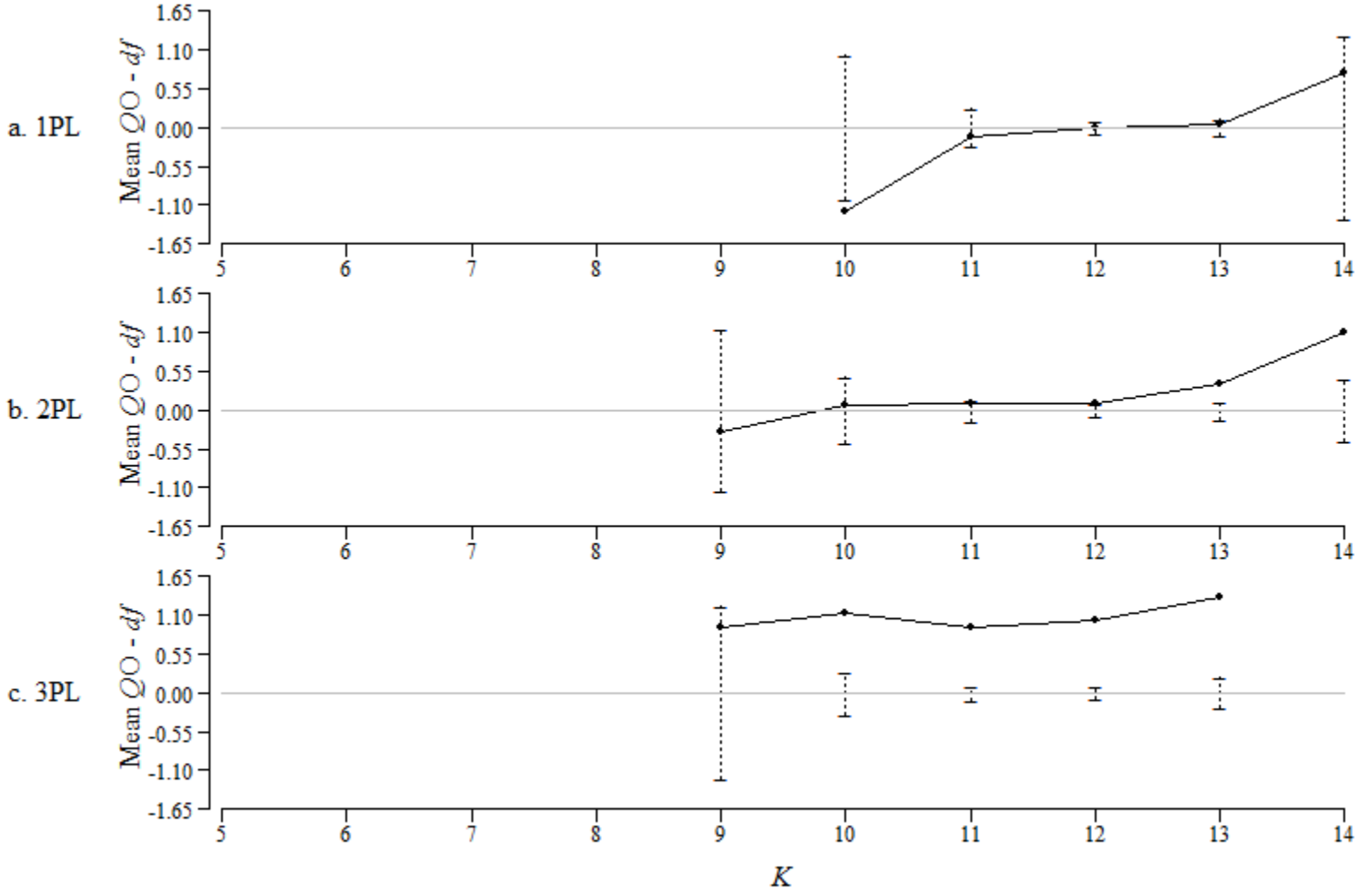


Figure E-19. De-trended *QO* Means by *K* for EU High Discrimination *N* = 1,500 *n* = 15 Condition

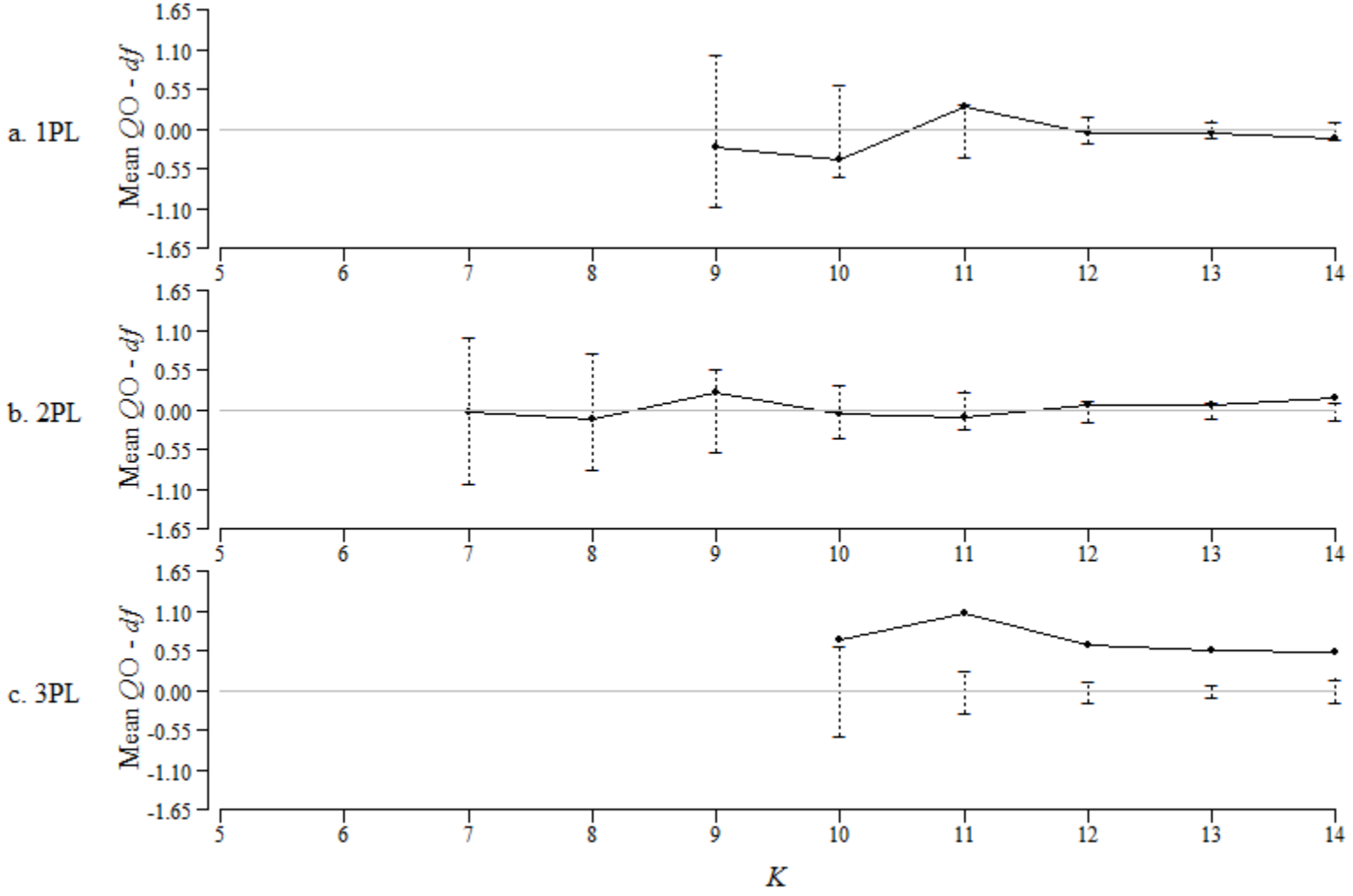


Figure E-20. De-trended *Q*O Means by *K* for EU Low Discrimination *N* = 1,500 *n* = 15 Condition

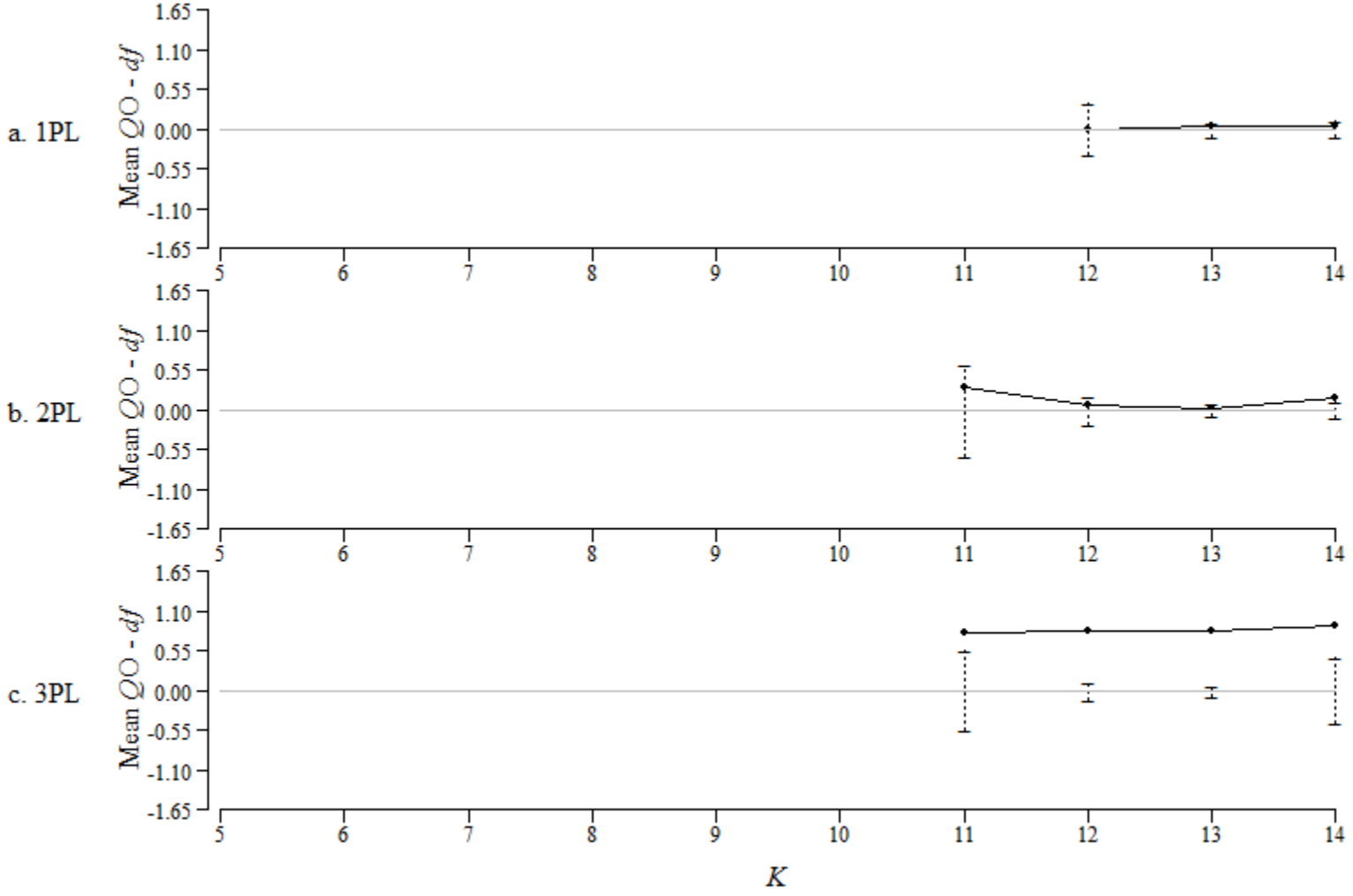


Figure E-21. De-trended *QO* Means by *K* for EU High Discrimination *N* = 500 *n* = 75 Condition

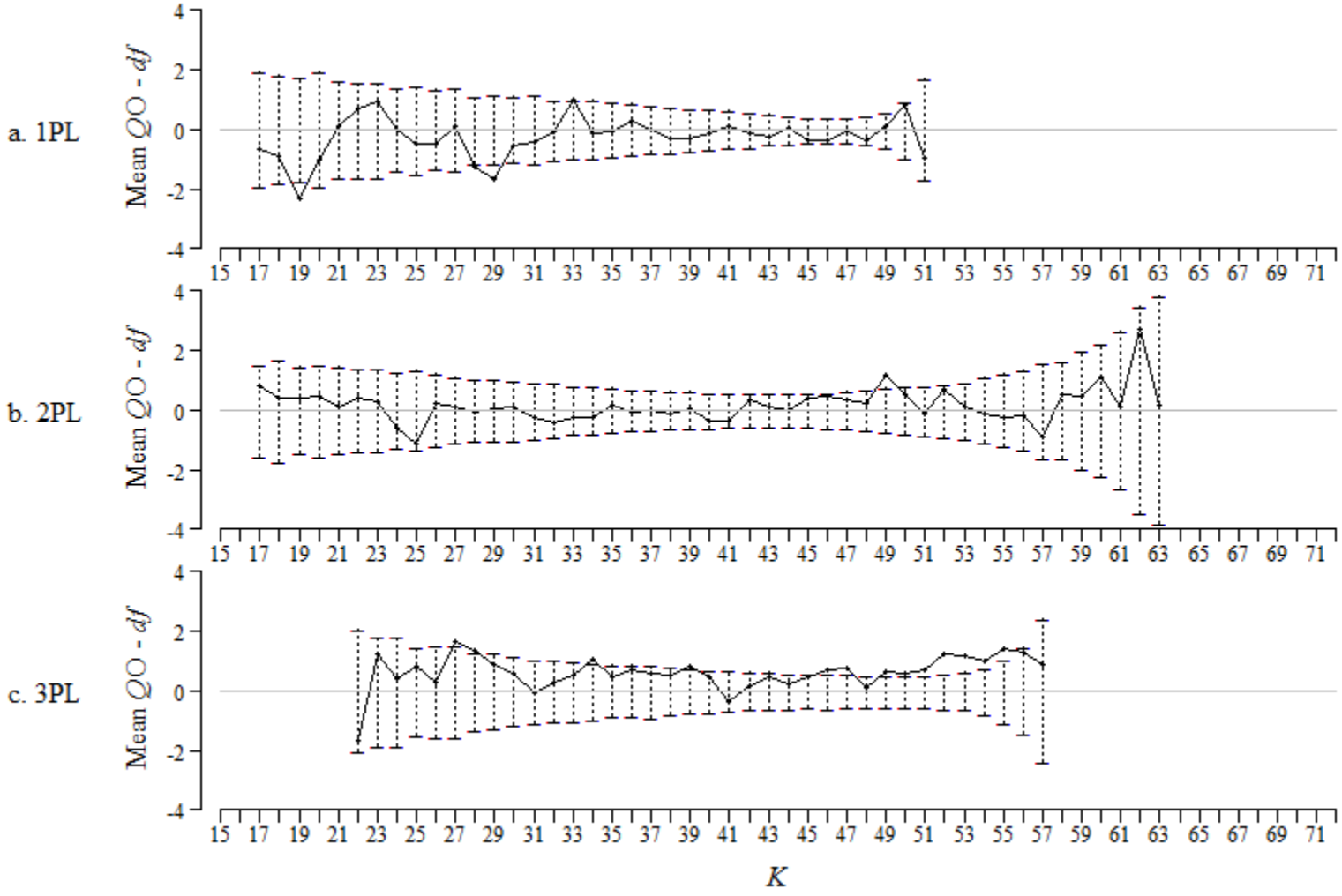


Figure E-22. De-trended *Q*O Means by *K* for EU Low Discrimination *N* = 500 *n* = 75 Condition

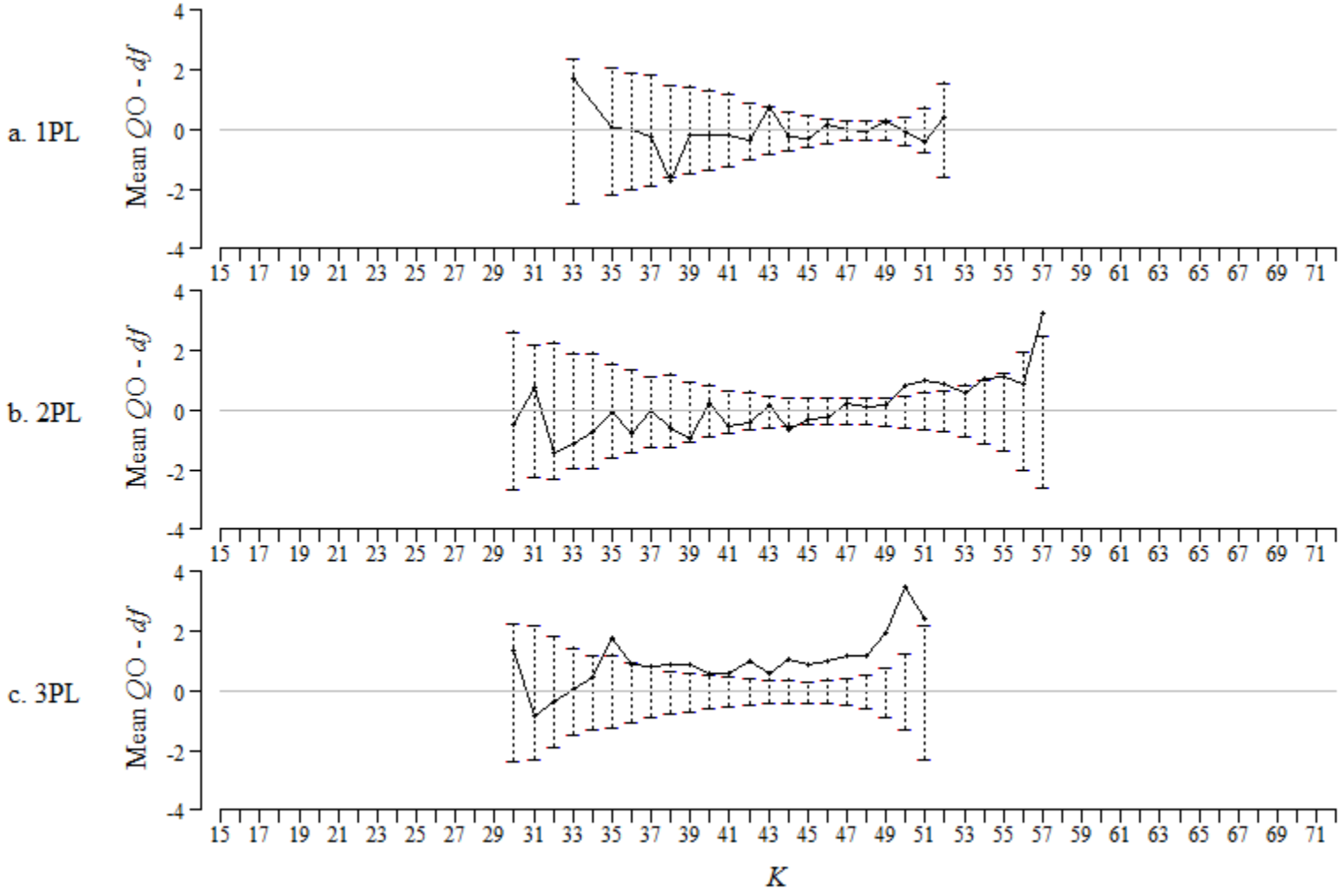


Figure E-23. De-trended *QO* Means by *K* for EU High Discrimination *N* = 1,500 *n* = 75 Condition

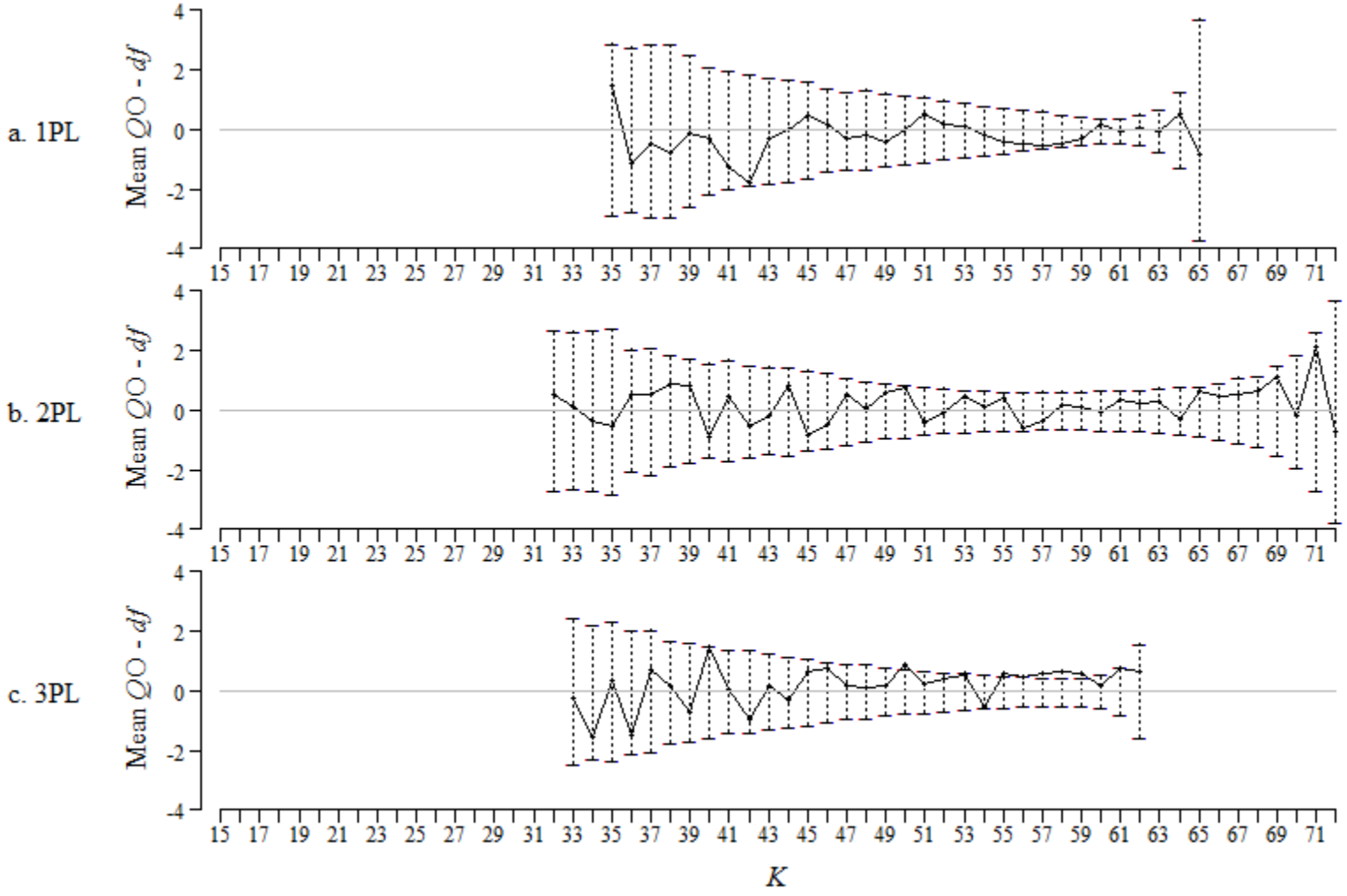


Figure E-24. De-trended *QO* Means by *K* for EU Low Discrimination *N* = 1,500 *n* = 75 Condition

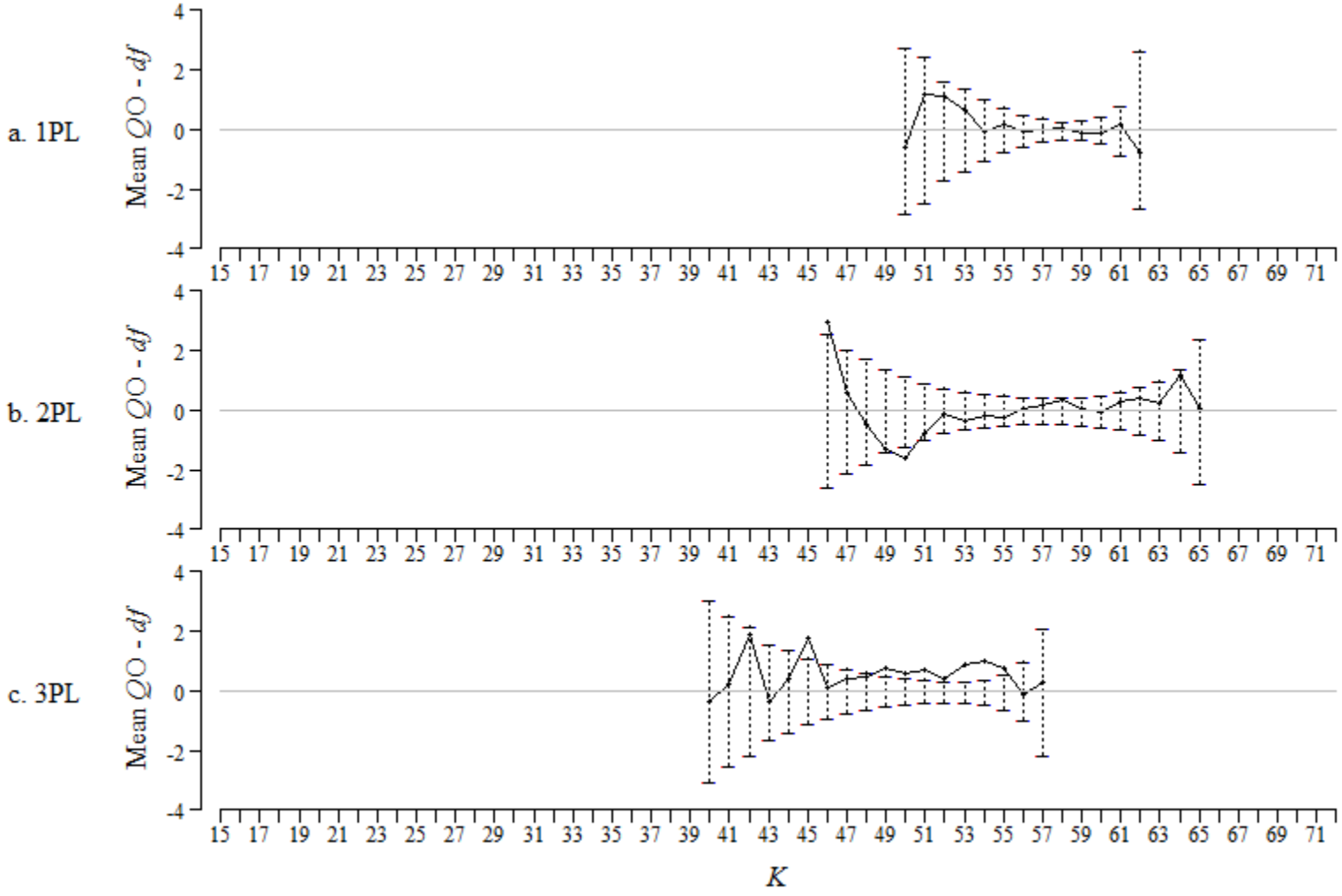


Figure E-25. De-trended *Q*O Variances by *K* for SU Low Discrimination *N* = 500 *n* = 15 Condition

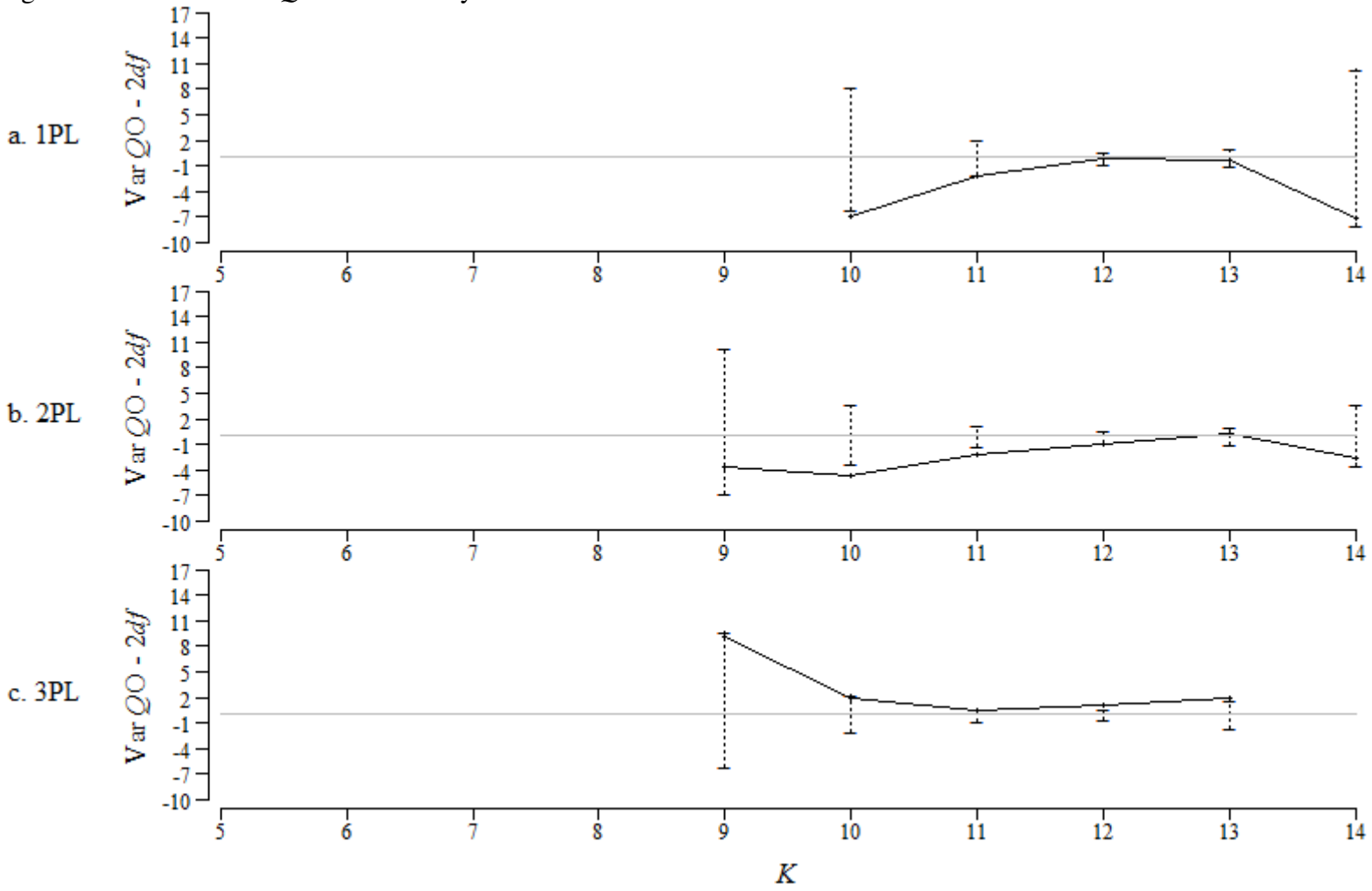


Figure E-26. De-trended *Q*O Variances by *K* for SU Low Discrimination *N* = 1,500 *n* = 15 Condition

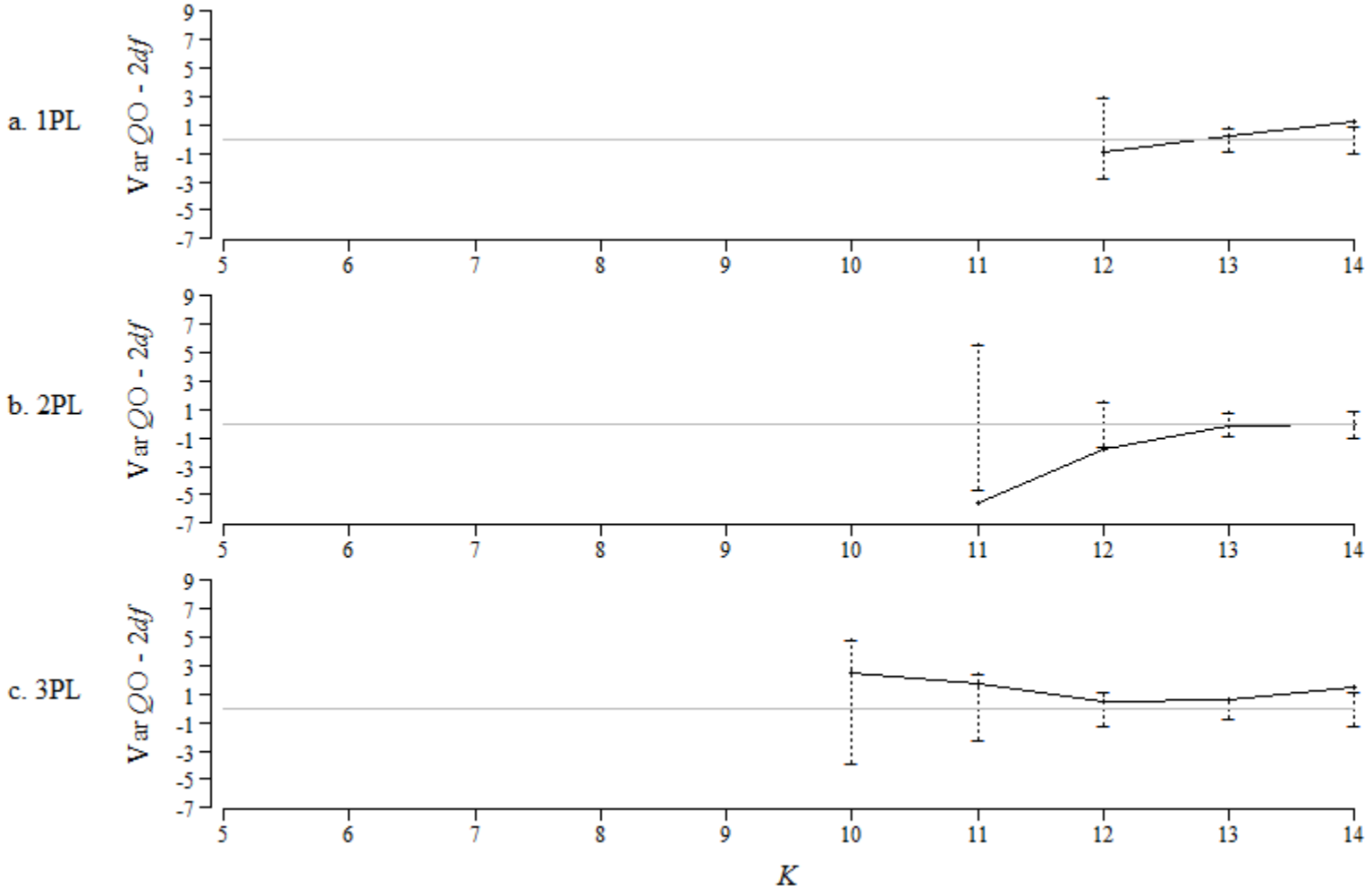


Figure E-27. De-trended *Q*O Variances by *K* for SU Low Discrimination *N* = 500 *n* = 75 Condition

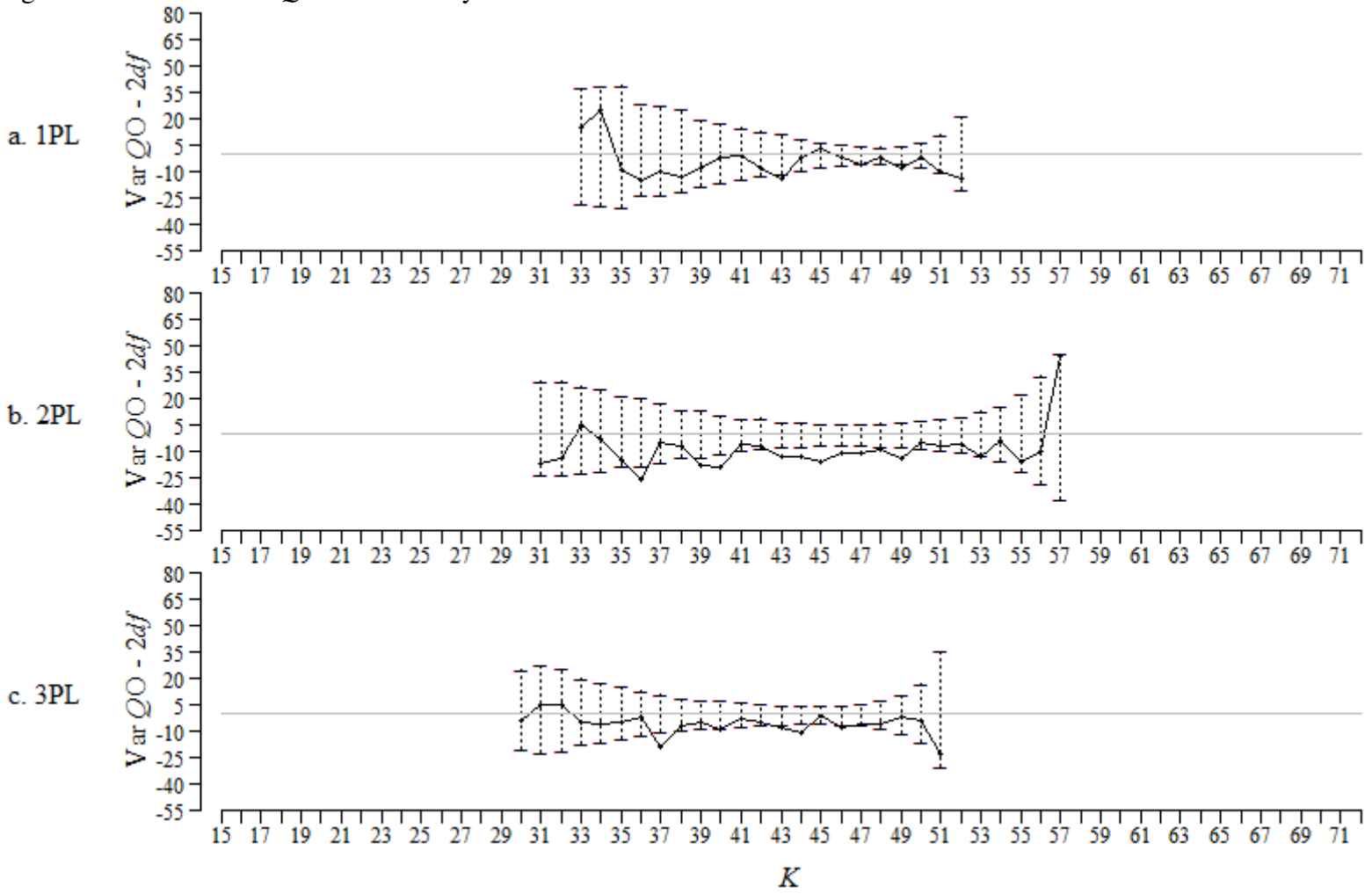


Figure E-28. De-trended *Q*O Variances by *K* for SU Low Discrimination *N* = 1,500 *n* = 75 Condition

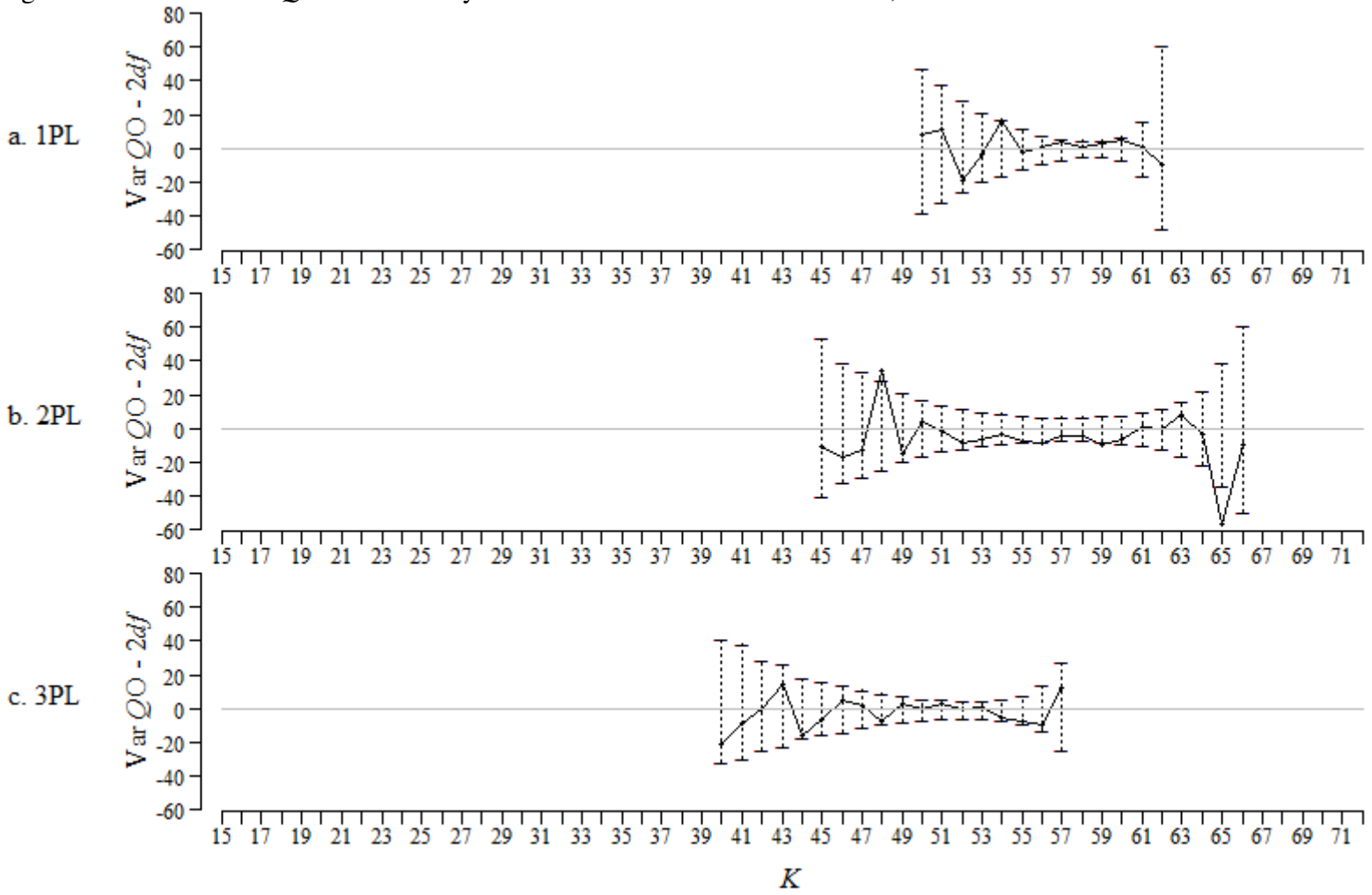


Figure E-29. De-trended *Q*O Variances by *K* for EU High Discrimination *N* = 500 *n* = 15 Condition

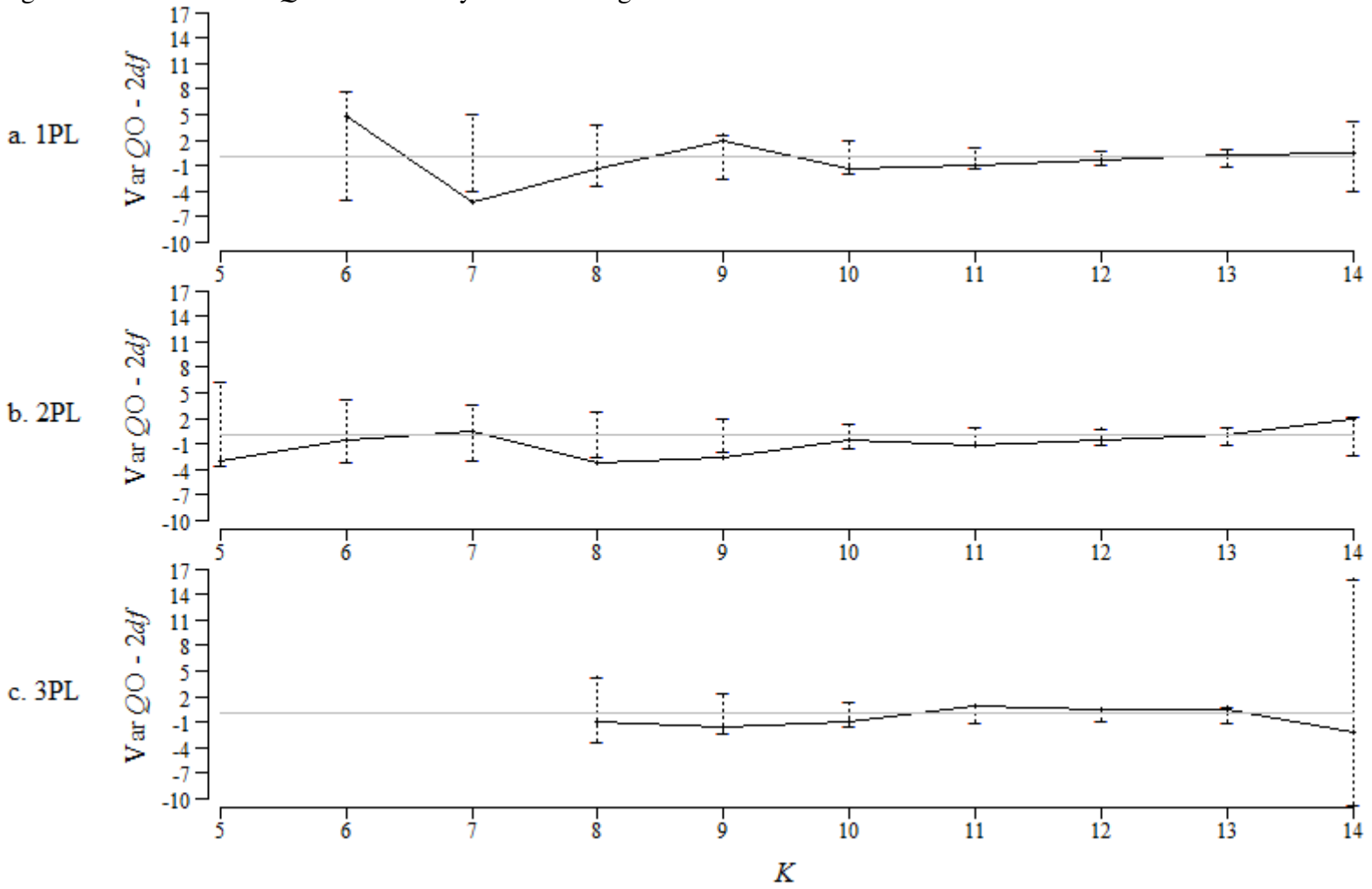


Figure E-30. De-trended *Q*O Variances by *K* for EU Low Discrimination *N* = 500 *n* = 15 Condition

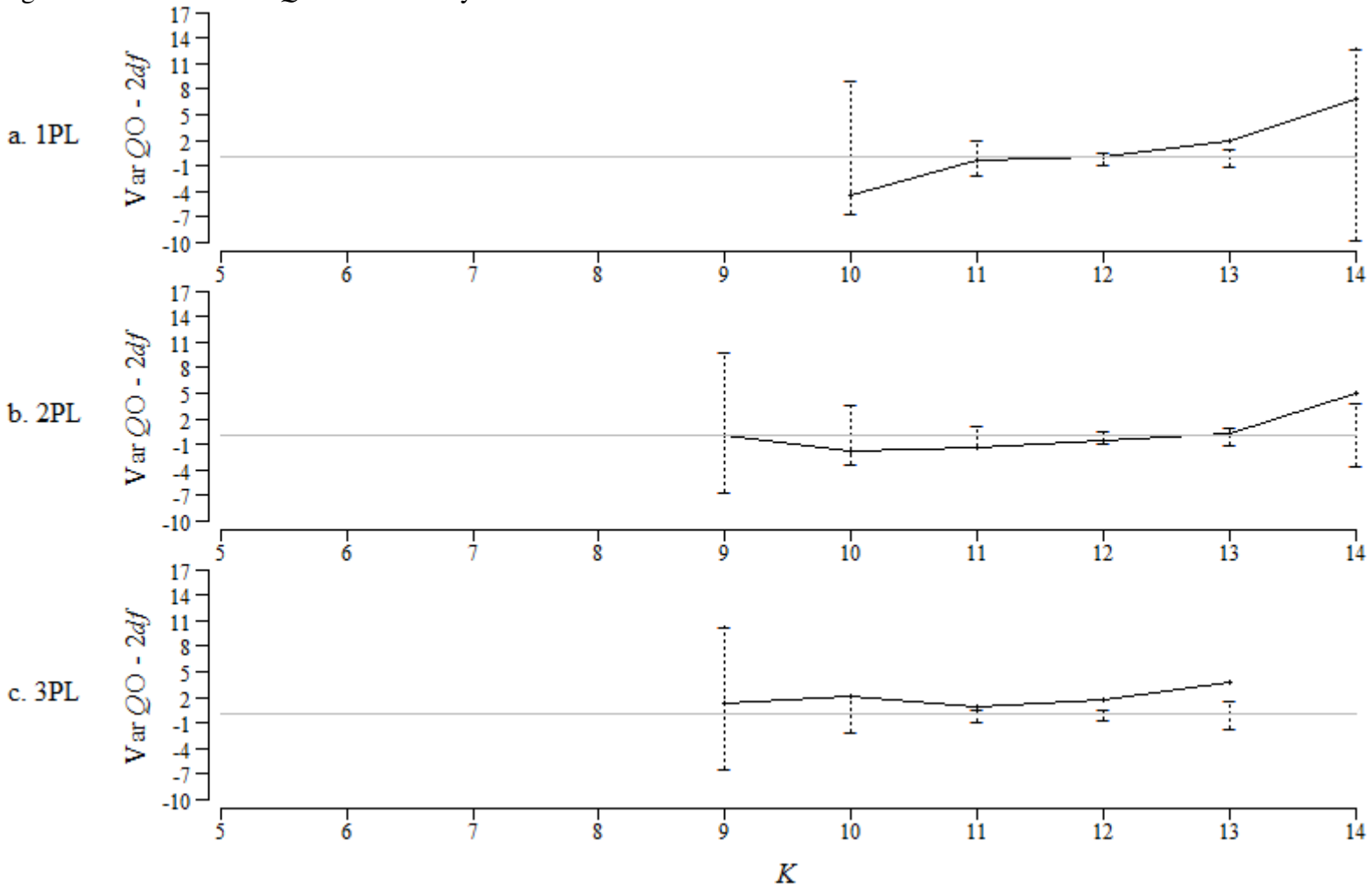


Figure E-31. De-trended *QO* Variances by *K* for EU High Discrimination *N* = 1,500 *n* = 15 Condition

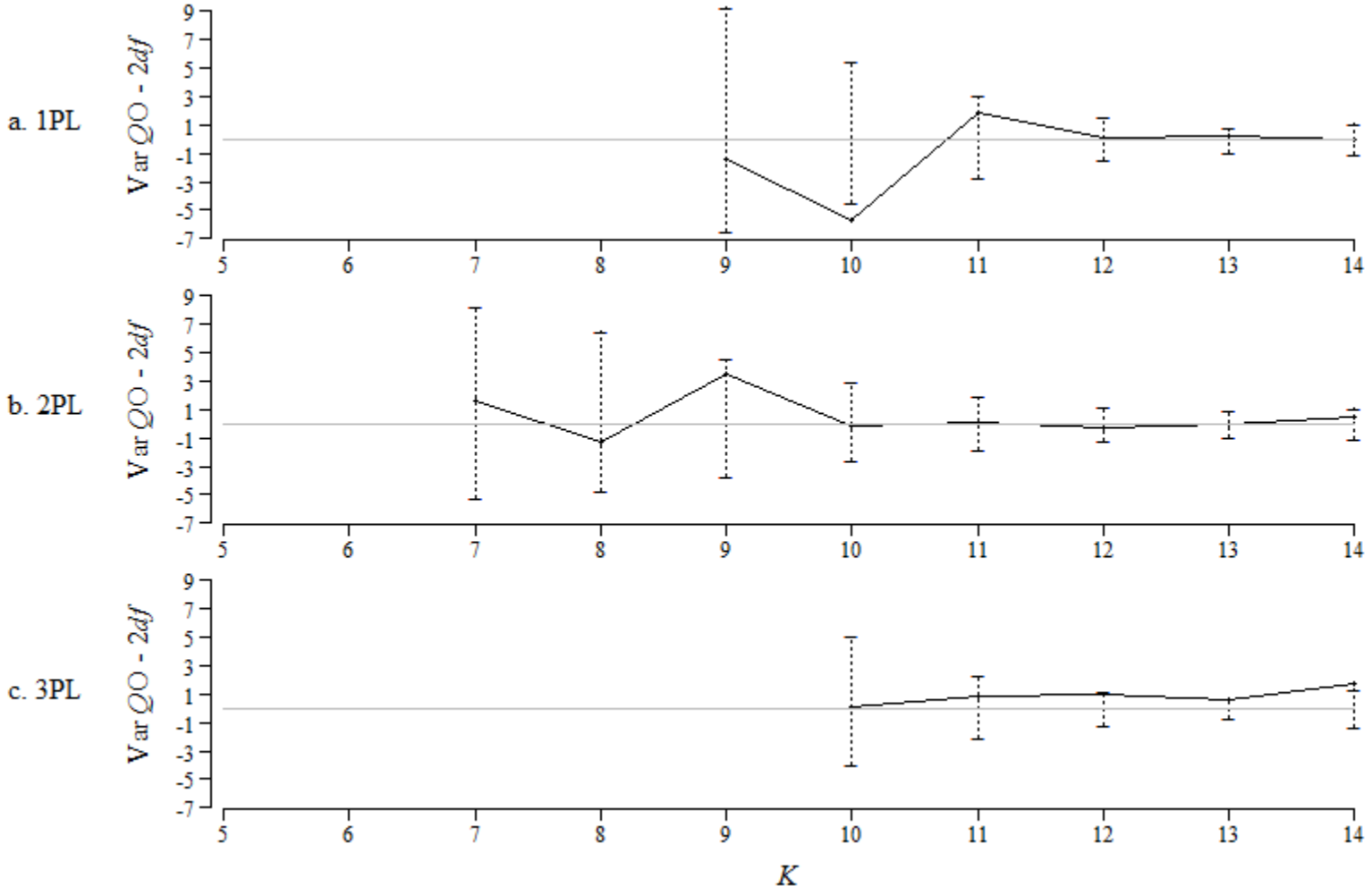


Figure E-32. De-trended *Q*O Variances by *K* for EU Low Discrimination *N* = 1,500 *n* = 15 Condition

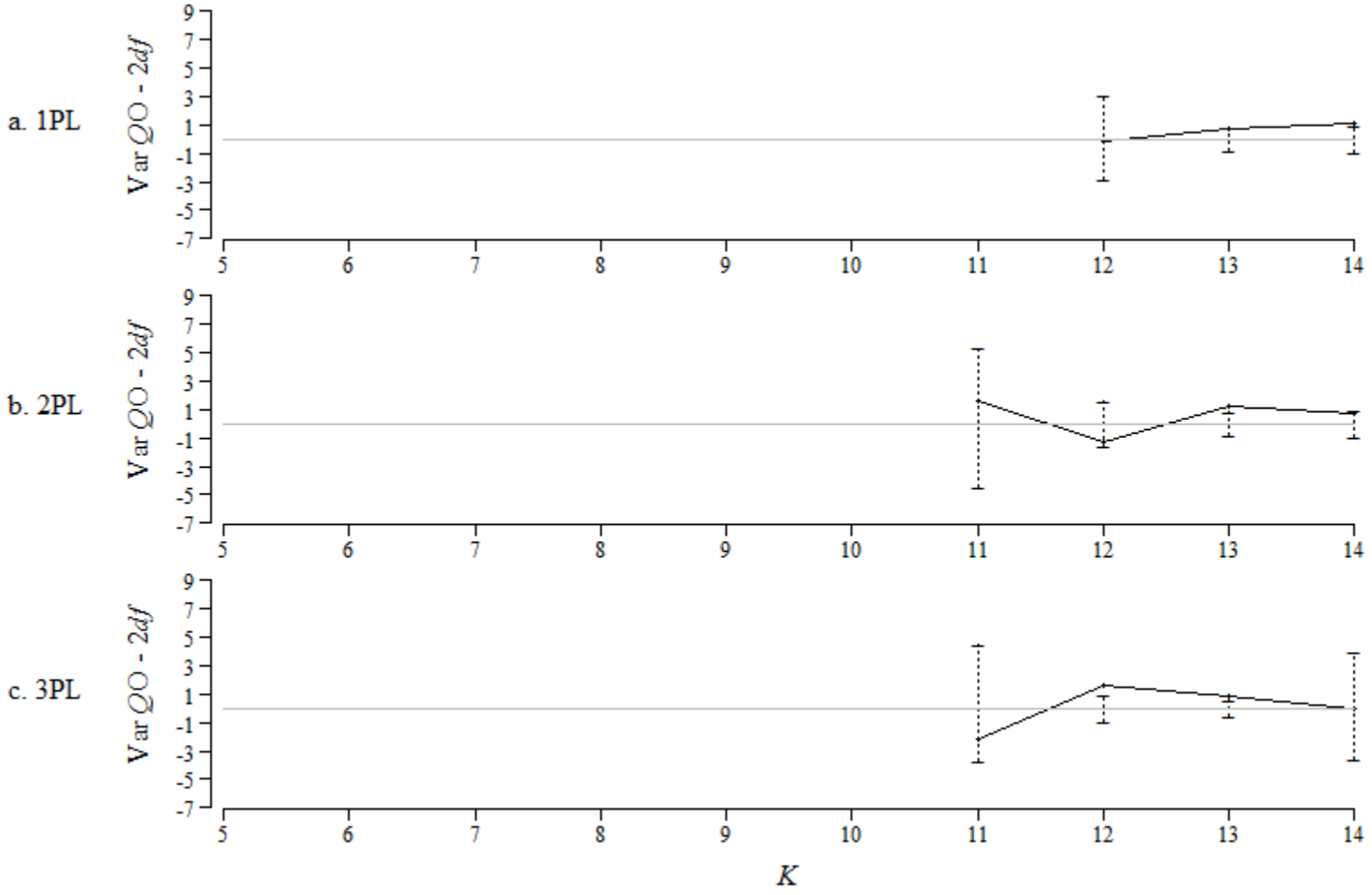


Figure E-33. De-trended *Q*O Variances by *K* for EU High Discrimination *N* = 500 *n* = 75 Condition

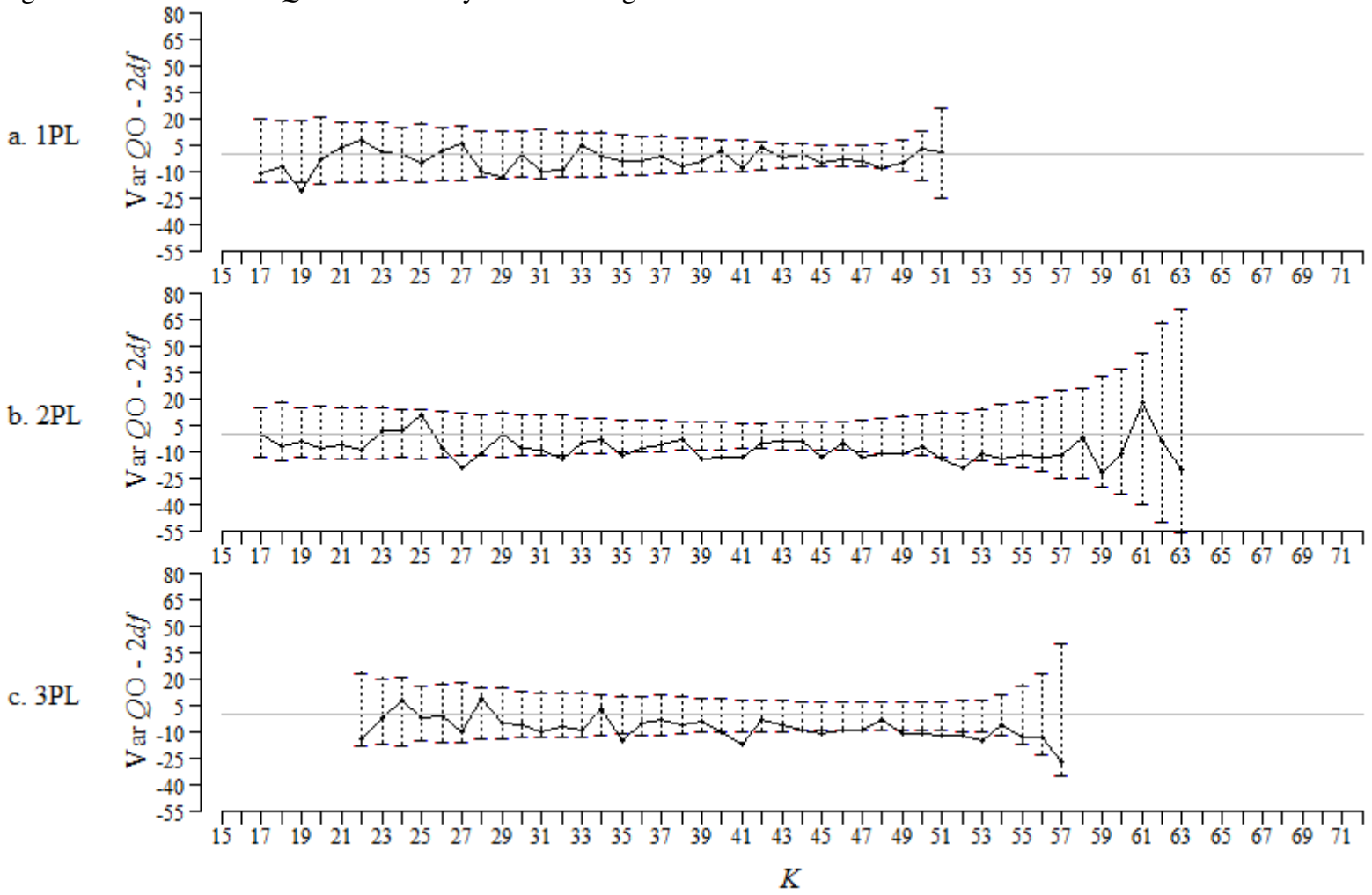


Figure E-34. De-trended *Q*O Variances by *K* for EU Low Discrimination *N* = 500 *n* = 75 Condition

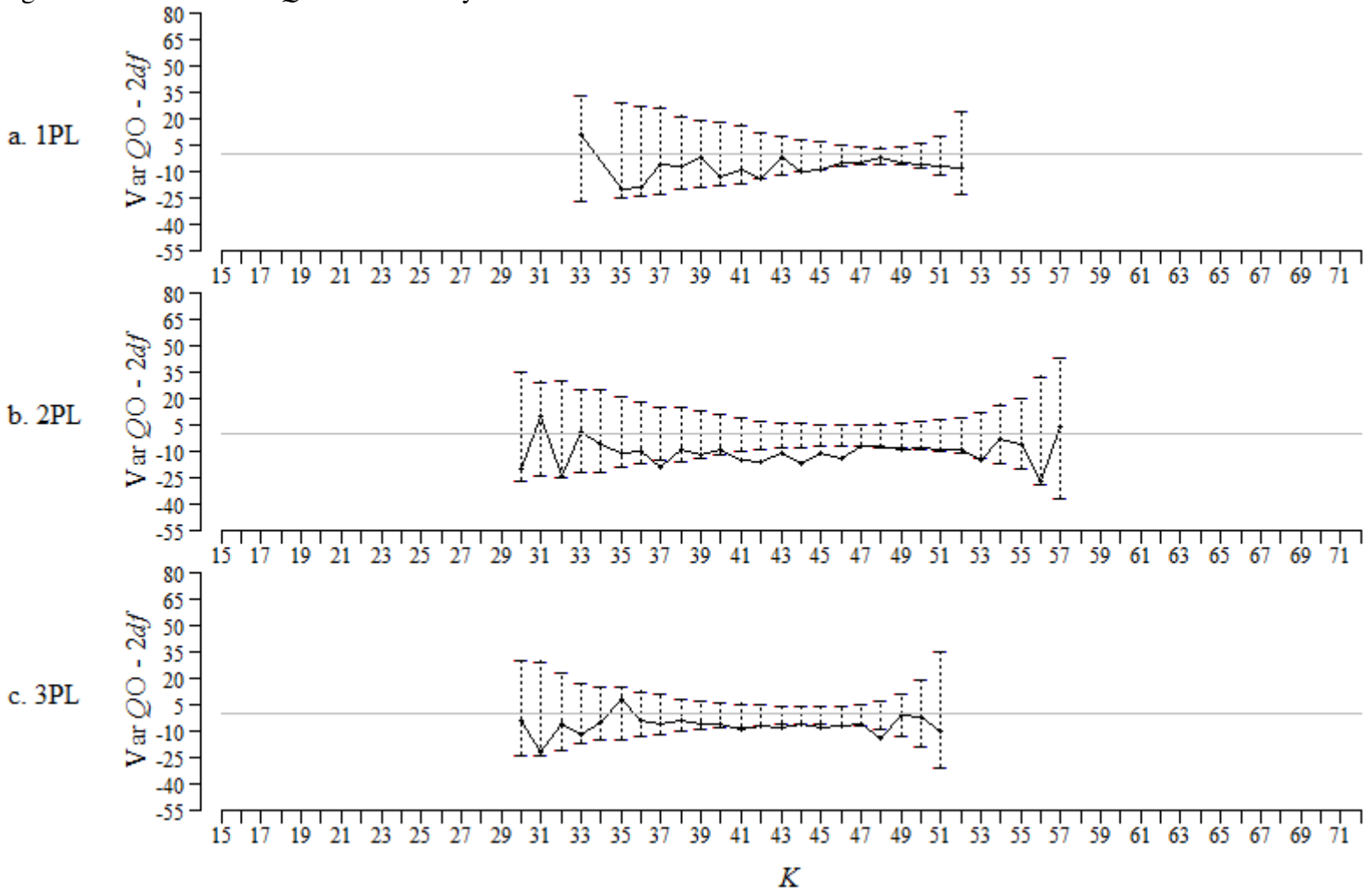


Figure E-35. De-trended *QO* Variances by *K* for EU High Discrimination *N* = 1,500 *n* = 75 Condition

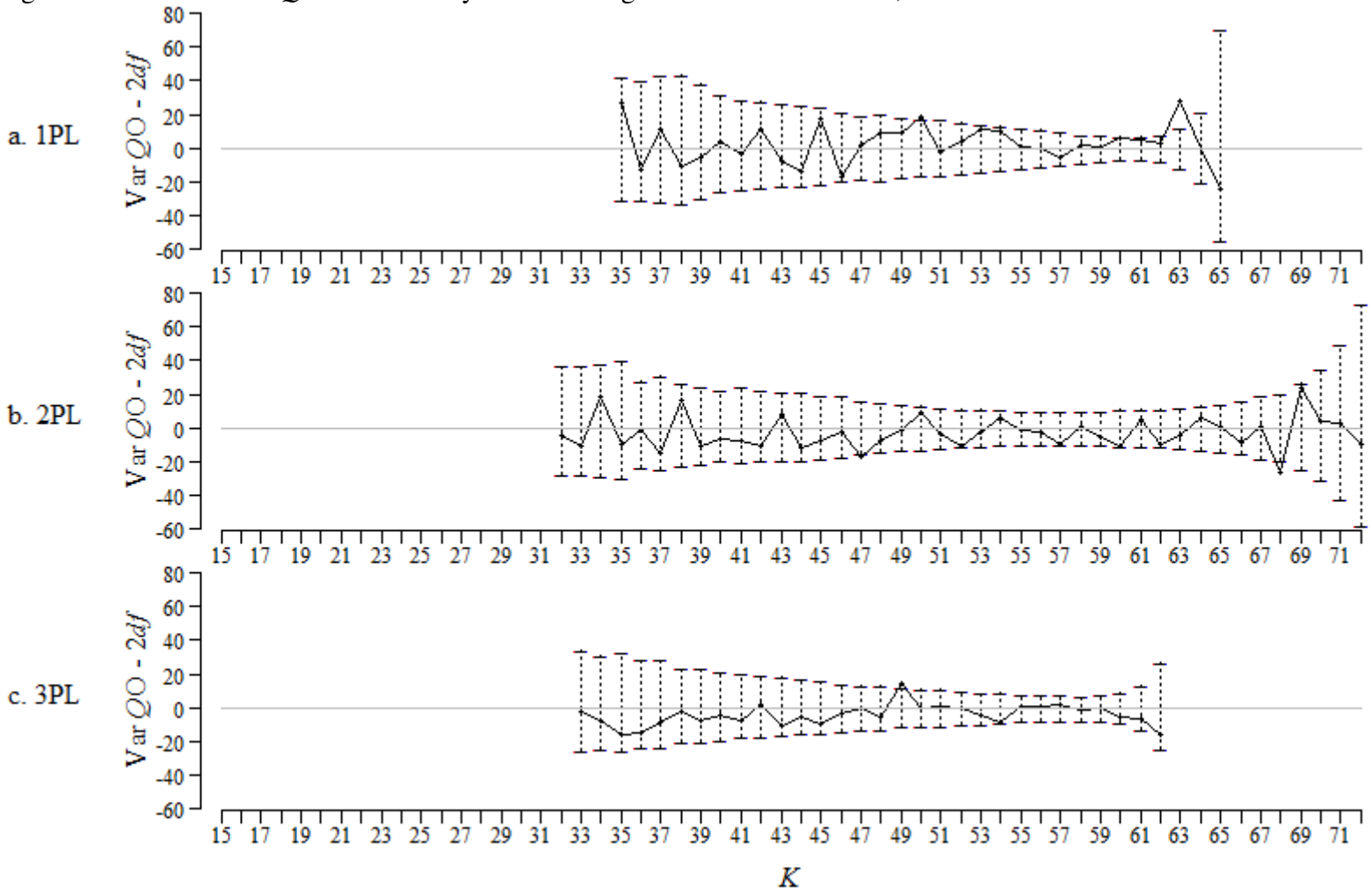


Figure E-36. De-trended *Q*O Variances by *K* for EU Low Discrimination *N* = 1,500 *n* = 75 Condition

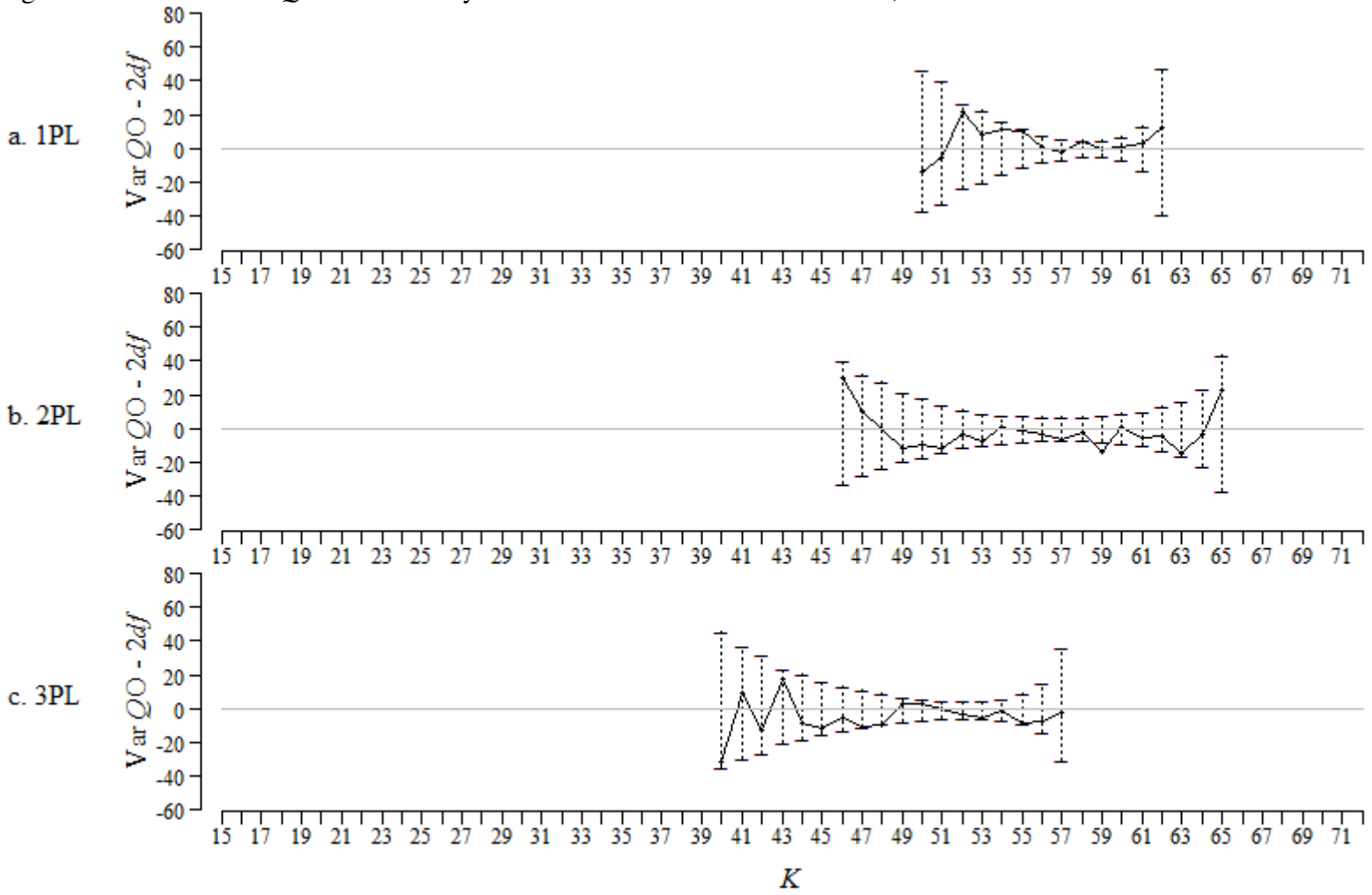


Table E-9. Frequency of KS Test Rejections in PE Conditions Aggregated Across D

n	N	DN	1PL			2PL			3PL		
			F_K	π_K	No. $N_K \geq 15$	F_K	π_K	No. $N_K \geq 15$	F_K	π_K	No. $N_K \geq 15$
						$\hat{\xi}$					
15	500	EU	4	0.29	14	7	0.41	17	10	0.83	12
		SU	5	0.36	14	9	0.53	17	11	0.85	13
	1,500	EU	2	0.18	11	1	0.07	14	8	0.80	10
		SU	2	0.18	11	2	0.14	14	12	1.00	12
75	500	EU	6	0.10	63	12	0.14	86	27	0.42	64
		SU	2	0.03	61	20	0.24	83	28	0.43	65
	1,500	EU	2	0.04	49	4	0.06	71	10	0.18	55
		SU	4	0.08	49	8	0.12	69	10	0.19	53
						ξ					
15	500	EU	4	0.27	15	3	0.19	16	1	0.08	13
		SU	1	0.07	15	1	0.06	16	0	0.00	13
	1,500	EU	2	0.18	11	7	0.50	14	4	0.40	10
		SU	0	0.00	11	3	0.21	14	0	0.00	10
75	500	EU	5	0.08	61	6	0.07	86	4	0.06	67
		SU	5	0.08	60	6	0.07	86	3	0.05	66
	1,500	EU	2	0.04	50	7	0.10	70	7	0.13	56
		SU	2	0.04	50	2	0.03	70	3	0.05	55

Table E-10. Bias of QO Sampling Distribution Means and SDs (N = 500 n = 15)

DN	D	PE	Bias(Mean)			Bias(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	−0.29	0.15	1.00	−0.23	−0.18	−0.17
		ξ	−0.16	0.01	−0.21	0.06	0.26	−0.03
	Low	$\hat{\xi}$	−0.29	0.30	1.23	−0.40	−0.30	0.35
		ξ	−0.50	0.00	0.10	−0.31	−0.22	0.09
EU	High	$\hat{\xi}$	−0.01	0.18	0.81	−0.05	−0.17	−0.08
		ξ	−0.10	0.19	0.01	0.09	0.28	0.06
	Low	$\hat{\xi}$	−0.10	0.24	1.06	0.04	0.00	0.23
		ξ	−0.19	0.08	−0.07	0.34	0.17	0.00

Table E-11. Bias of *QO* Sampling Distribution Means and SDs (N = 1,500 n = 15)

DN	D	PE	Bias(Mean)			Bias(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	−0.08	0.08	0.80	−0.09	−0.06	0.16
		ξ	0.01	−0.19	0.09	0.37	0.05	0.27
	Low	$\hat{\xi}$	0.00	−0.08	0.80	0.02	−0.23	0.16
		ξ	−0.04	−0.19	0.01	0.02	−0.13	0.17
EU	High	$\hat{\xi}$	−0.10	0.02	0.69	−0.12	0.06	0.09
		ξ	0.13	0.58	0.01	0.15	0.54	0.07
	Low	$\hat{\xi}$	0.02	0.14	0.84	0.05	0.06	0.00
		ξ	0.15	0.39	0.35	0.19	0.31	0.35

Table E-12. Bias of *QO* Sampling Distribution Means and SDs (N = 500 n = 75)

DN	D	PE	Bias(Mean)			Bias(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	0.09	0.18	0.65	−0.06	−0.53	−0.37
		ξ	0.26	0.12	0.08	0.17	0.03	0.05
	Low	$\hat{\xi}$	−0.21	0.15	0.55	−0.22	−0.52	−0.33
		ξ	−0.08	−0.06	−0.16	0.03	−0.15	−0.06
EU	High	$\hat{\xi}$	−0.27	0.15	0.61	−0.22	−0.47	−0.44
		ξ	−0.16	0.18	0.04	0.11	0.07	−0.03
	Low	$\hat{\xi}$	−0.05	0.10	0.97	−0.41	−0.60	−0.39
		ξ	0.00	0.00	0.32	0.02	−0.05	0.03

Table E-13. Bias of *QO* Sampling Distribution Means and SDs (N = 1,500 n = 75)

DN	D	PE	Bias(Mean)			Bias(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	−0.24	−0.24	0.22	0.05	−0.30	−0.13
		ξ	0.25	−0.18	−0.05	0.44	0.23	0.28
	Low	$\hat{\xi}$	−0.40	0.01	0.75	0.05	−0.32	−0.15
		ξ	−0.27	0.10	0.27	0.30	0.08	0.16
EU	High	$\hat{\xi}$	−0.23	0.18	0.17	0.11	−0.17	−0.25
		ξ	0.07	0.73	0.41	0.34	0.64	0.46
	Low	$\hat{\xi}$	0.11	0.04	0.53	0.18	−0.09	−0.28
		ξ	0.45	0.31	0.29	0.54	0.28	0.10

Table E-14. Mean Error for *Q*O Sampling Distribution Means and SDs (*N* = 500 *n* = 15)

DN	D	PE	ME(Mean)			ME(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	0.29	0.21	1.00	0.29	0.20	0.24
		ξ	0.24	0.13	0.30	0.15	0.29	0.32
	Low	$\hat{\xi}$	0.29	0.34	1.23	0.40	0.31	0.35
		ξ	0.50	0.11	0.12	0.31	0.27	0.09
EU	High	$\hat{\xi}$	0.28	0.19	0.85	0.27	0.23	0.14
		ξ	0.24	0.26	0.18	0.34	0.28	0.15
	Low	$\hat{\xi}$	0.43	0.34	1.06	0.29	0.17	0.23
		ξ	0.23	0.37	0.23	0.34	0.29	0.18

Table E-15. Mean Error for *Q*O Sampling Distribution Means and SDs (*N* = 1,500 *n* = 15)

DN	D	PE	ME(Mean)			ME(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	0.27	0.23	0.80	0.21	0.17	0.16
		ξ	0.19	0.31	0.18	0.37	0.29	0.27
	Low	$\hat{\xi}$	0.02	0.17	0.80	0.08	0.23	0.16
		ξ	0.04	0.22	0.03	0.09	0.18	0.18
EU	High	$\hat{\xi}$	0.20	0.11	0.69	0.19	0.12	0.09
		ξ	0.21	0.58	0.32	0.31	0.54	0.34
	Low	$\hat{\xi}$	0.03	0.14	0.84	0.06	0.13	0.14
		ξ	0.16	0.39	0.35	0.19	0.31	0.35

Table E-16. Mean Error for *Q*O Sampling Distribution Means and SDs (*N* = 500 *n* = 75)

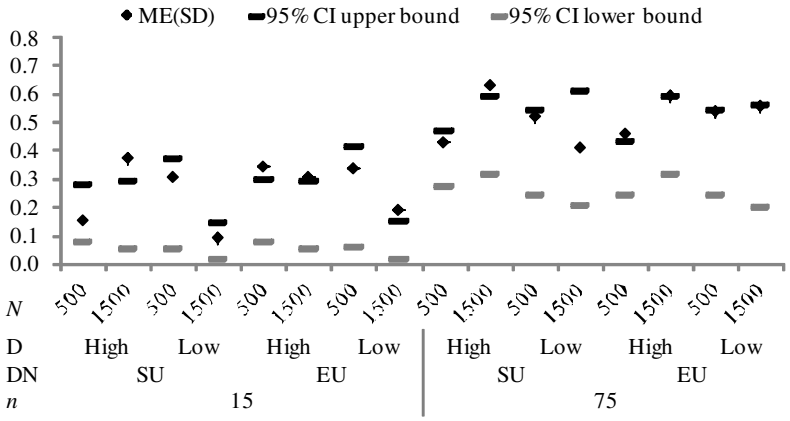
DN	D	PE	ME(Mean)			ME(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	0.43	0.66	0.69	0.29	0.56	0.49
		ξ	0.61	0.51	0.41	0.43	0.41	0.45
	Low	$\hat{\xi}$	0.42	0.65	0.71	0.47	0.69	0.40
		ξ	0.54	0.47	0.38	0.52	0.43	0.49
EU	High	$\hat{\xi}$	0.50	0.39	0.73	0.37	0.55	0.52
		ξ	0.50	0.51	0.36	0.46	0.50	0.30
	Low	$\hat{\xi}$	0.39	0.70	1.08	0.48	0.67	0.44
		ξ	0.70	0.52	0.44	0.54	0.41	0.54

Table E-17. Mean Error for *Q*O Sampling Distribution Means and SDs (*N* = 1,500 *n* = 75)

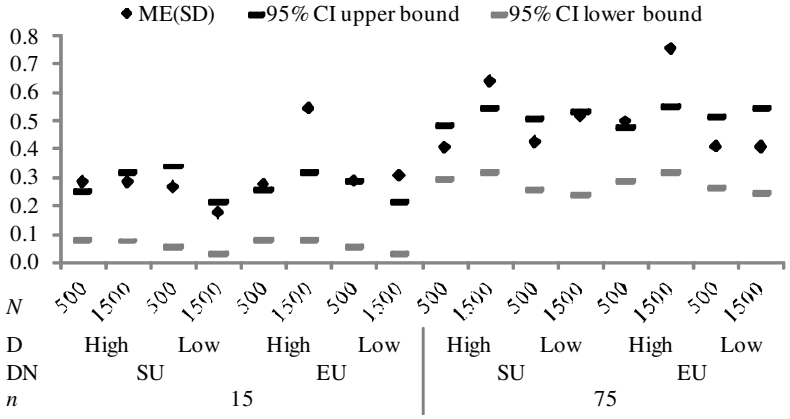
DN	D	PE	ME(Mean)			ME(SD)		
			1PL	2PL	3PL	1PL	2PL	3PL
SU	High	$\hat{\xi}$	0.72	0.61	0.49	0.59	0.50	0.42
		ξ	0.63	0.69	0.35	0.63	0.64	0.43
	Low	$\hat{\xi}$	0.77	0.75	0.85	0.31	0.52	0.37
		ξ	0.64	0.60	0.39	0.41	0.52	0.29
EU	High	$\hat{\xi}$	0.46	0.50	0.56	0.47	0.42	0.32
		ξ	0.70	0.91	0.64	0.59	0.75	0.51
	Low	$\hat{\xi}$	0.39	0.56	0.63	0.35	0.40	0.48
		ξ	0.58	0.69	0.70	0.55	0.41	0.50

Figure E-37. Estimates of ME(SD) and 95% CIs About the Estimates for *Q*O and all ξ Study Conditions

a. 1PL



b. 2PL



c. 3PL

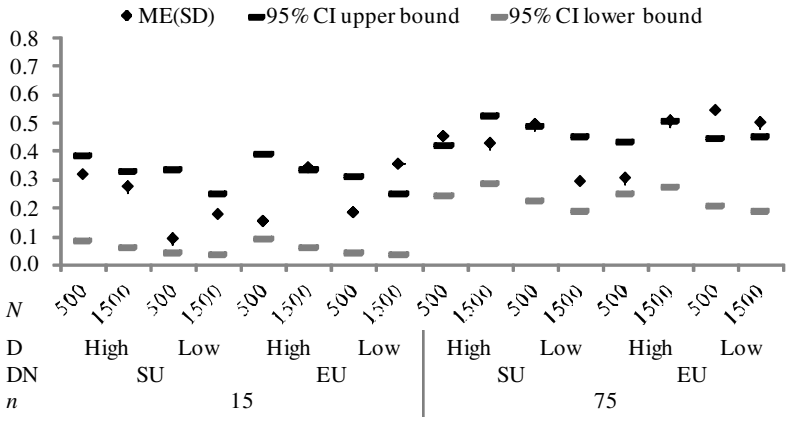
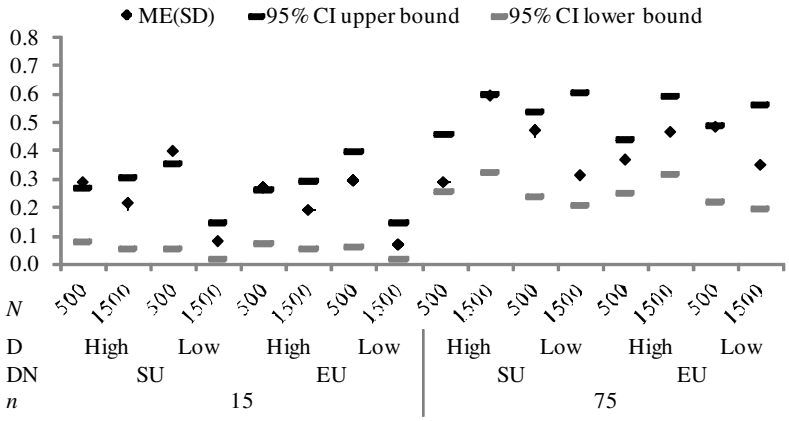
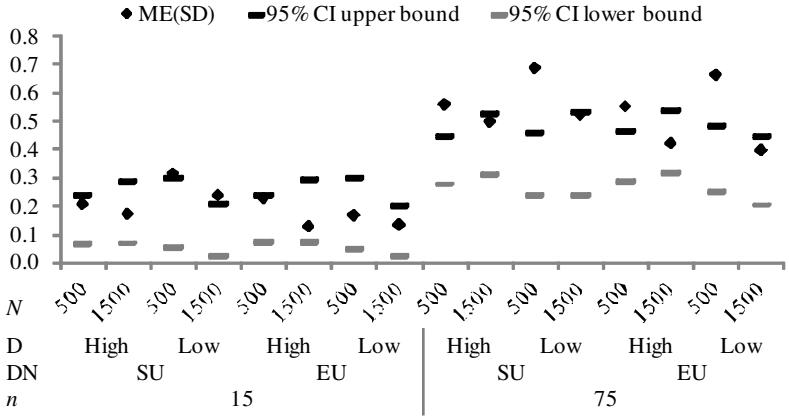


Figure E-38. Estimates of ME(SD) and 95% CIs About the Estimates for *Q*O and all $\hat{\xi}$ Study Conditions

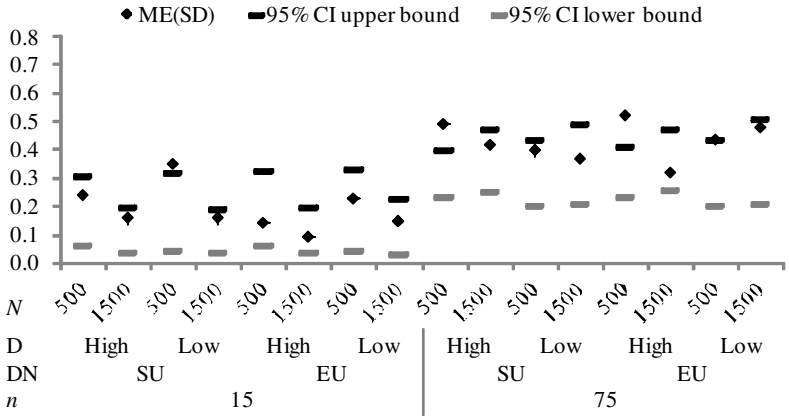
a. 1PL



b. 2PL



c. 3PL



APPENDIX F: ANALYSIS OF *Q*1 DISTRIBUTION

Number of Items Within Each of the *K* *Q*1 Groups

Detrended *Q*1 Sample Means by *K* With 95% Confidence Intervals

Detrended *Q*1 Sample SDs by *K* With 95% Confidence Intervals

Bias of *Q*1 Sampling Distribution Means and SDs in SU Conditions

Mean Error for *Q*1 Sampling Distribution Means and SDs in SU Conditions

Frequency of Cases in Which KS Test Rejected the Null Hypothesis that *Q*1 Followed its Theoretical Distribution

Correlations Between *Q*1 and *b* Within *K* in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Correlations Between *Q*1 and *a* Within *K* in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Correlations Between *Q*1 and *c* Within *K* in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Table F-1. Number of Items Within Each K by Model and DN for High Discrimination Conditions (N = 500 n = 15)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	3	1	0	0
2	4	1	4	10	1	1
3	6	7	33	22	1	1
4	19	20	54	62	8	3
5	53	65	149	127	24	35
6	168	177	257	246	75	60
7	399	425	494	548	203	208
8	1110	1176	1596	1457	638	670
9	5431	5496	5247	5068	2707	2552
10	11560	11383	10913	11209	15093	15220
TOTAL	18750	18750	18750	18750	18750	18750

Table F-2. Number of Items Within Each K by Model and DN for Low Discrimination Conditions (N = 500 n = 15)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
4	0	0	0	0	0	1
5	0	0	2	4	1	1
6	0	0	10	8	0	0
7	4	4	8	13	2	2
8	29	57	42	32	14	10
9	735	1071	504	517	264	260
10	17982	17618	18184	18176	18469	18476
TOTAL	18750	18750	18750	18750	18750	18750

Table F-3. Number of Items Within Each K by Model and DN for High Discrimination Conditions (N = 1,500 n = 15)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	1	0	0	0
2	0	0	0	2	0	0
3	1	2	5	6	0	0
4	1	0	11	9	2	1
5	4	4	27	34	4	6
6	16	11	68	54	11	6
7	44	63	180	160	34	36
8	161	176	533	549	186	188
9	1564	1635	2356	2288	1034	988
10	16959	16859	15569	15648	17479	17525
TOTAL	18750	18750	18750	18750	18750	18750

Table F-4. Number of Items Within Each K by Model and DN for Low Discrimination Conditions (N = 1,500 n = 15)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
6	0	0	1	1	0	0
7	0	0	1	1	0	0
8	0	0	4	3	1	0
9	258	198	59	64	22	23
10	18492	18552	18685	18681	18727	18727
TOTAL	18750	18750	18750	18750	18750	18750

Table F-5. Number of Items Within Each K by Model and DN for High Discrimination Conditions (N = 500 n = 75)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	1	1	0	0
2	0	1	6	3	1	1
3	8	7	25	25	1	4
4	23	26	71	64	10	14
5	69	75	116	140	24	28
6	164	191	281	267	87	82
7	416	425	704	641	225	210
8	1549	1486	2067	2030	870	811
9	7213	7158	6550	6479	3504	3377
10	9308	9381	8929	9100	14028	14223
TOTAL	18750	18750	18750	18750	18750	18750

Table F-6. Number of Items Within Each K by Model and DN for Low Discrimination Conditions (N = 500 n = 75)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
3	0	0	0	0	1	0
4	0	0	0	0	0	1
5	0	0	2	0	0	1
6	0	0	2	6	3	3
7	1	1	18	17	4	3
8	6	9	67	70	28	34
9	323	324	965	979	484	496
10	18420	18416	17696	17678	18230	18212
TOTAL	18750	18750	18750	18750	18750	18750

Table F-7. Number of Items Within Each *K* by Model and DN for High Discrimination Conditions (*N* = 1,500 *n* = 75)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
1	0	0	0	1	0	0
2	1	1	4	3	0	0
3	0	0	4	5	0	0
4	2	5	18	15	0	0
5	6	9	28	28	7	4
6	18	17	53	61	14	13
7	69	63	219	180	69	66
8	298	312	667	644	281	282
9	2525	2588	3192	3053	1542	1475
10	15831	15755	14565	14760	16837	16910
TOTAL	18750	18750	18750	18750	18750	18750

Table F-8. Number of Items Within Each *K* by Model and DN for Low Discrimination Conditions (N = 1,500 n = 75)

K	Study Conditions					
	1PL		2PL		3PL	
	SU N_K	EU N_K	SU N_K	EU N_K	SU N_K	EU N_K
6	0	0	2	0	0	0
7	0	0	0	2	0	0
8	1	0	5	9	1	5
9	5	8	118	122	52	51
10	18744	18742	18625	18617	18697	18694
TOTAL	18750	18750	18750	18750	18750	18750

Figure F-1. De-trended $Q1$ Means by K for SU Conditions (N = 500 n = 15)

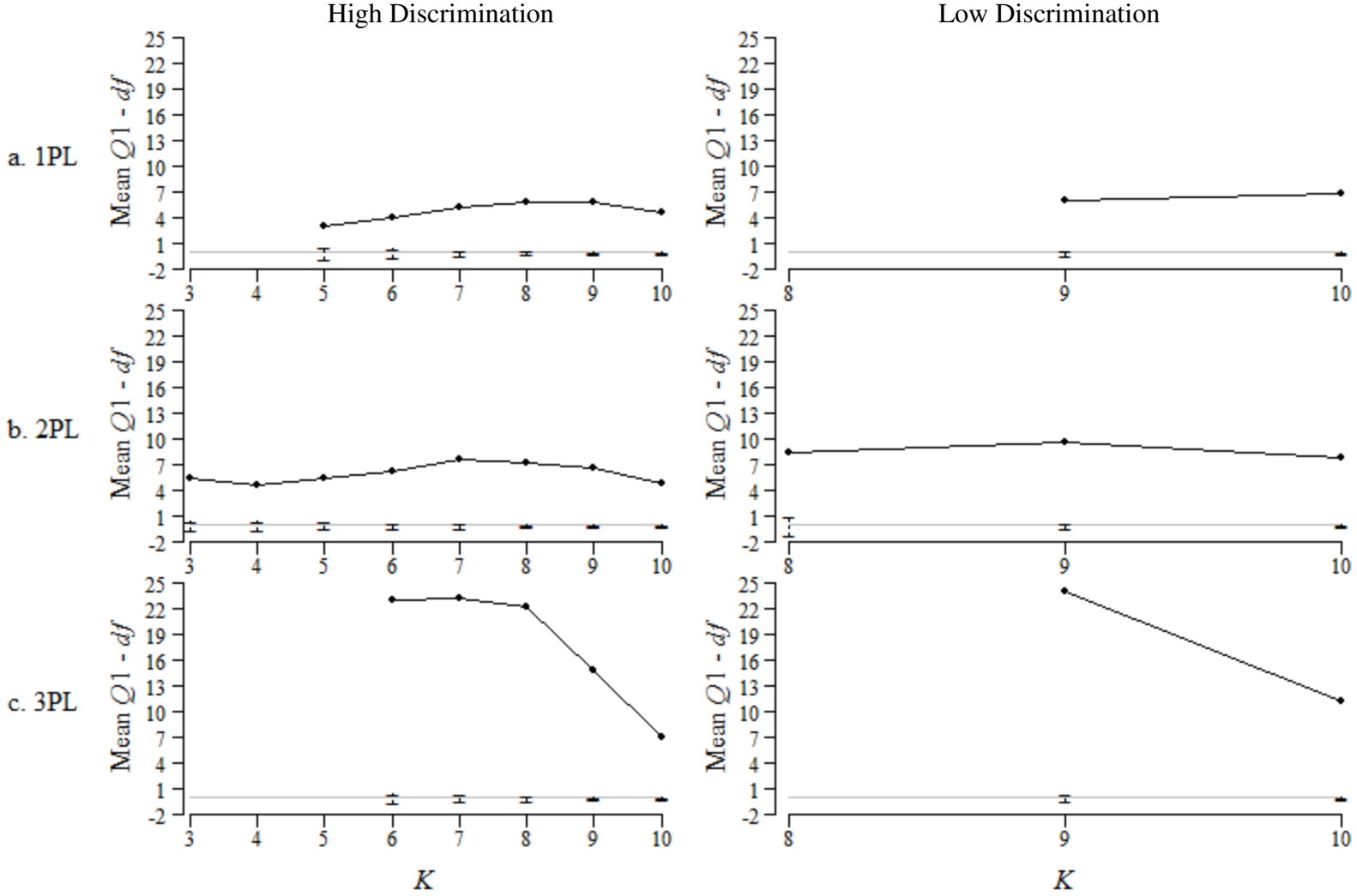


Figure F-2. De-trended *Q*1 Means by *K* for EU Conditions (*N* = 500 *n* = 15)

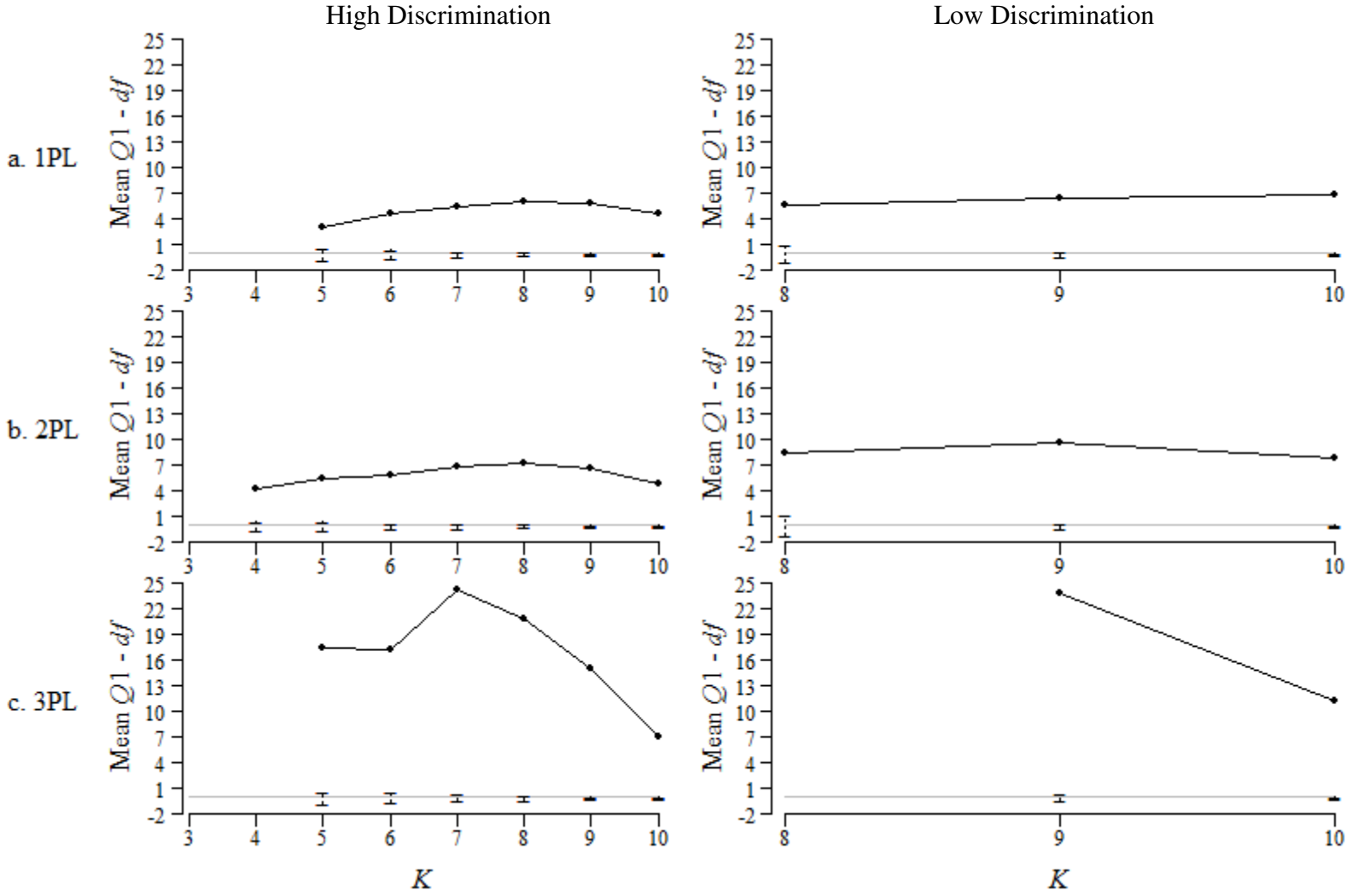


Figure F-3. De-trended *Q*1 Means by *K* for SU Conditions (*N* = 1,500 *n* = 15)

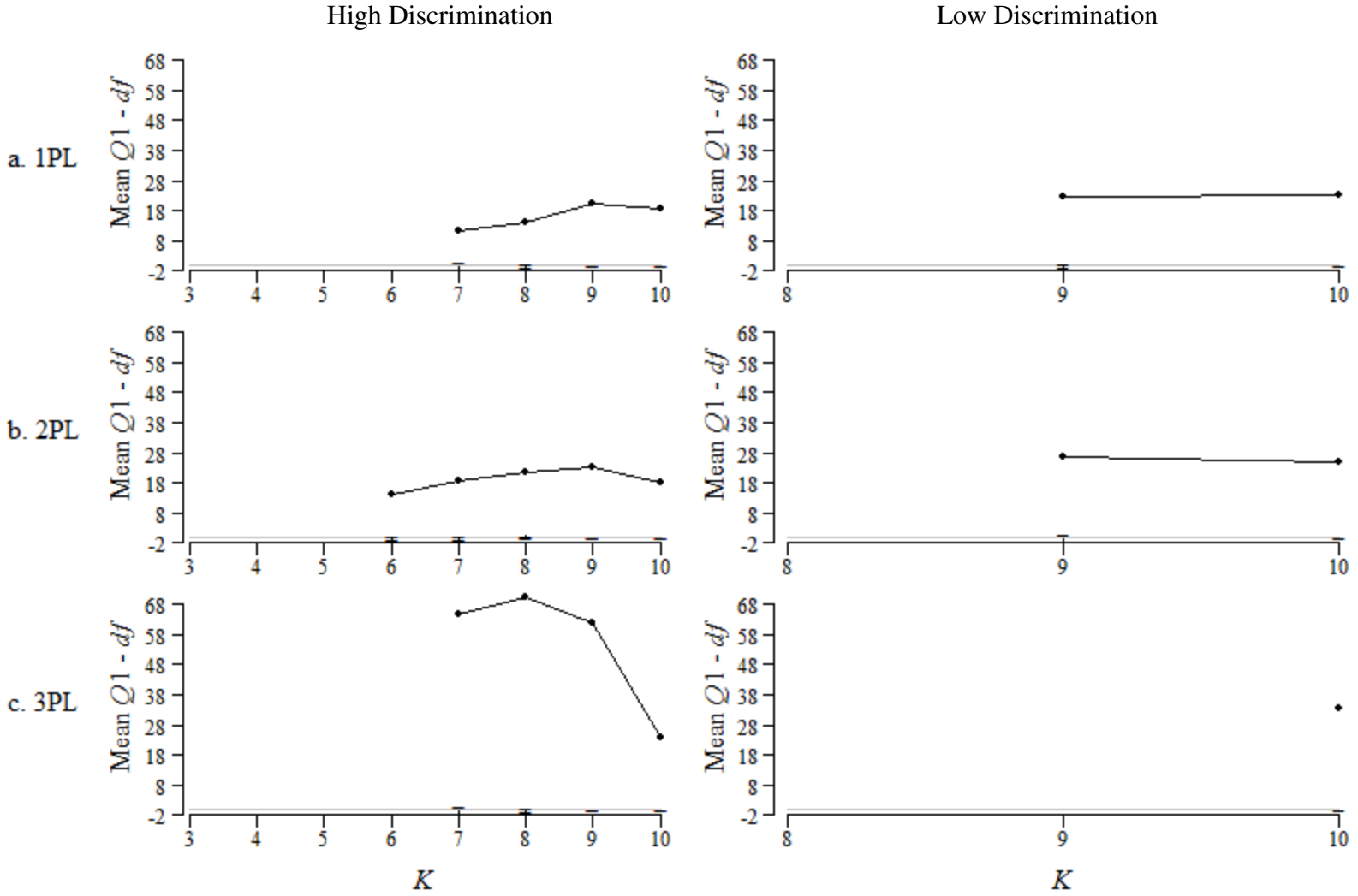


Figure F-4. De-trended *Q*1 Means by *K* for EU Conditions (*N* = 1,500 *n* = 15)

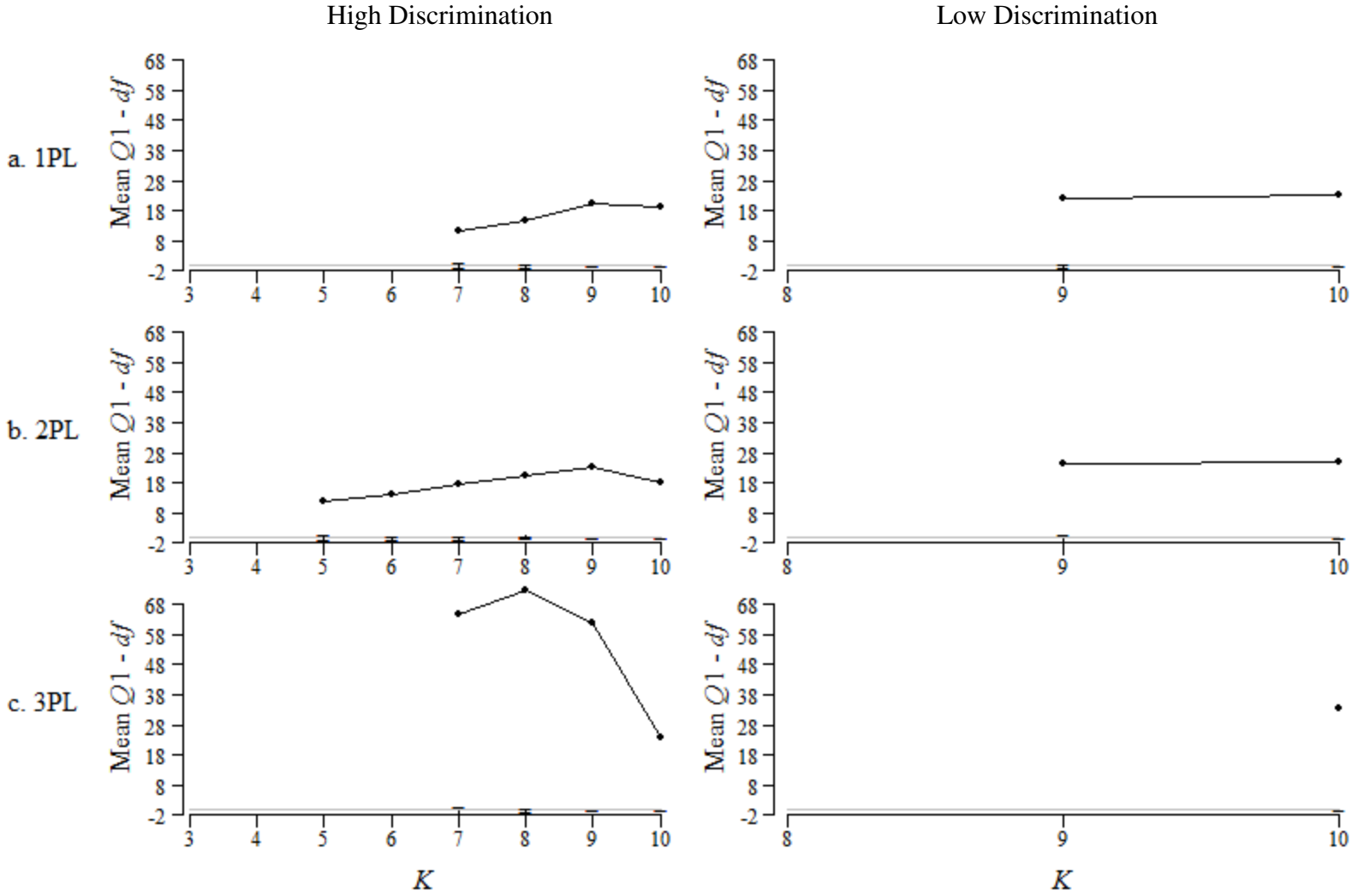


Figure F-5. De-trended $Q1$ Means by K for SU Conditions (N = 500 n = 75)

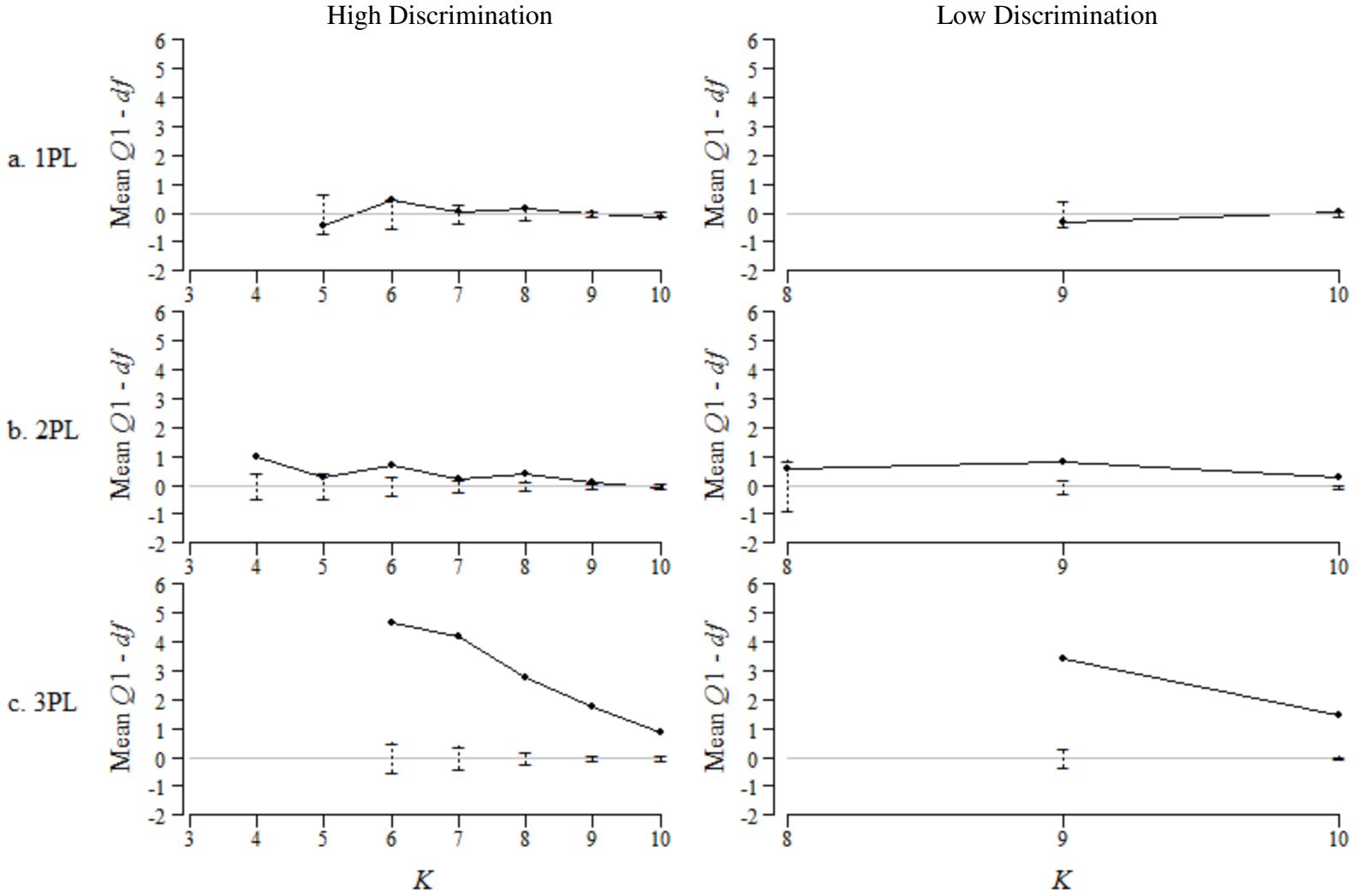


Figure F-6. De-trended *Q*1 Means by *K* for EU Conditions (*N* = 500 *n* = 75)

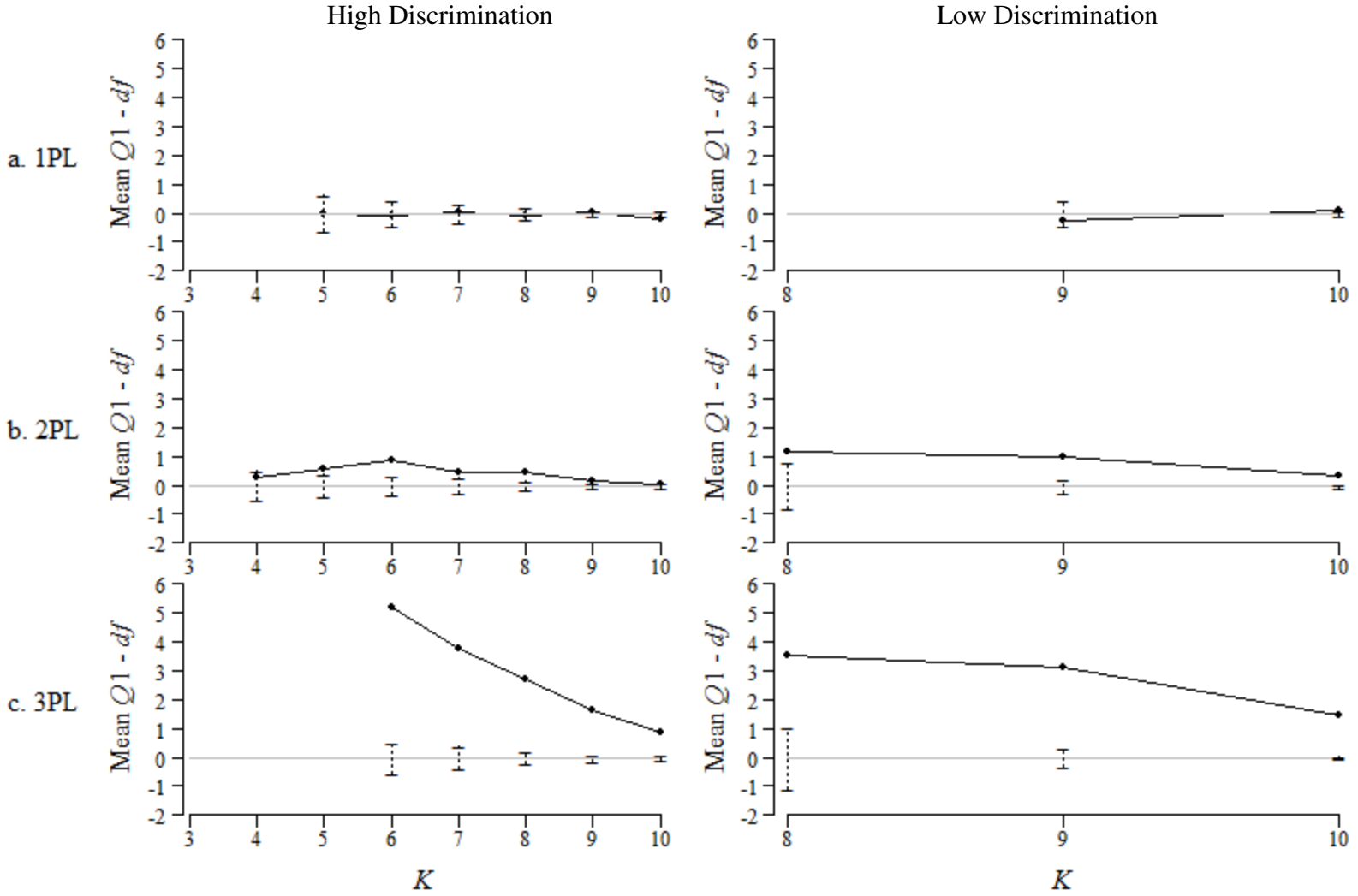


Figure F-7. De-trended *Q*1 Means by *K* for SU Conditions (*N* = 1,500 *n* = 75)

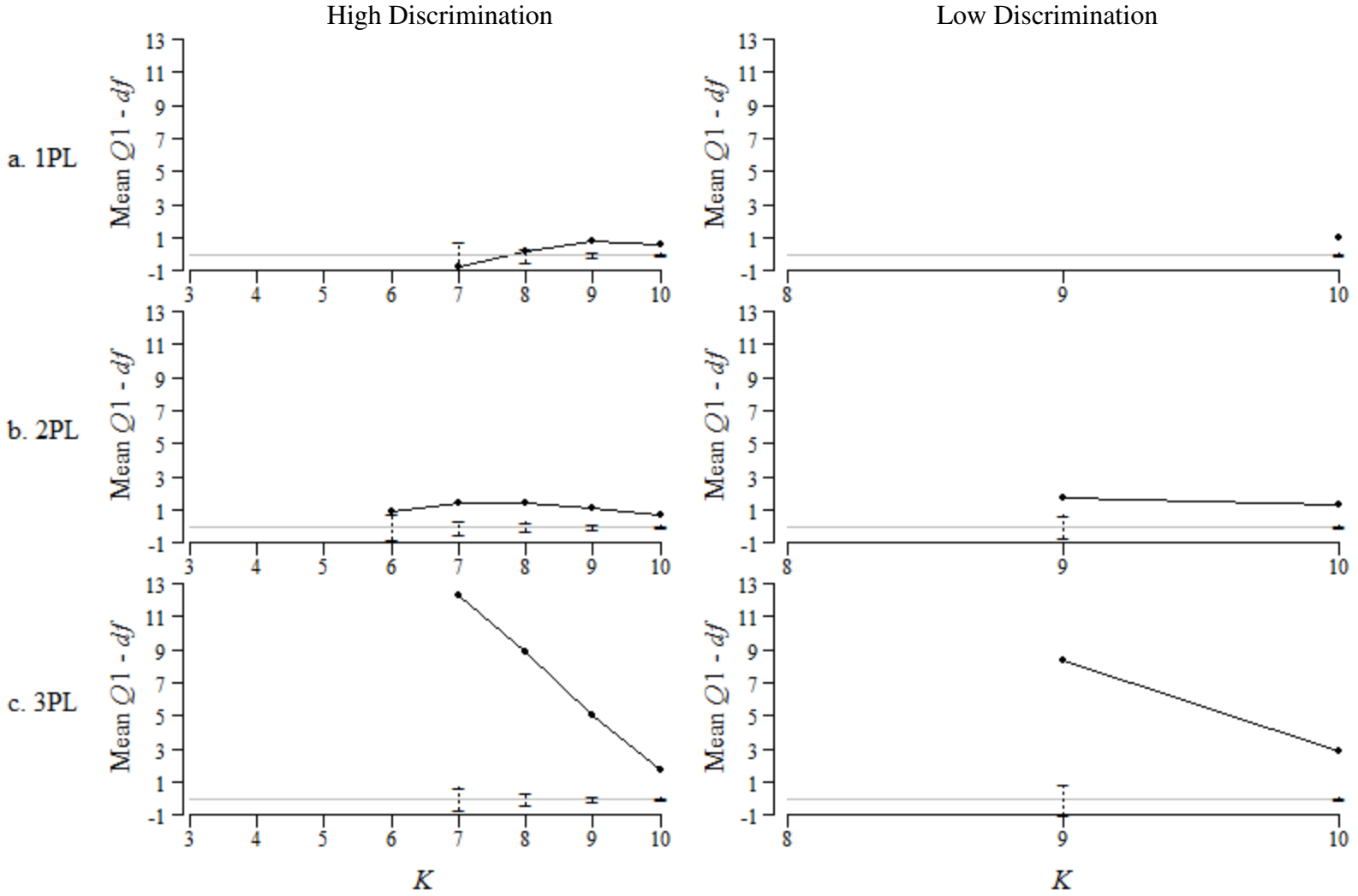


Figure F-8. De-trended *Q*1 Means by *K* for EU Conditions (*N* = 1,500 *n* = 75)

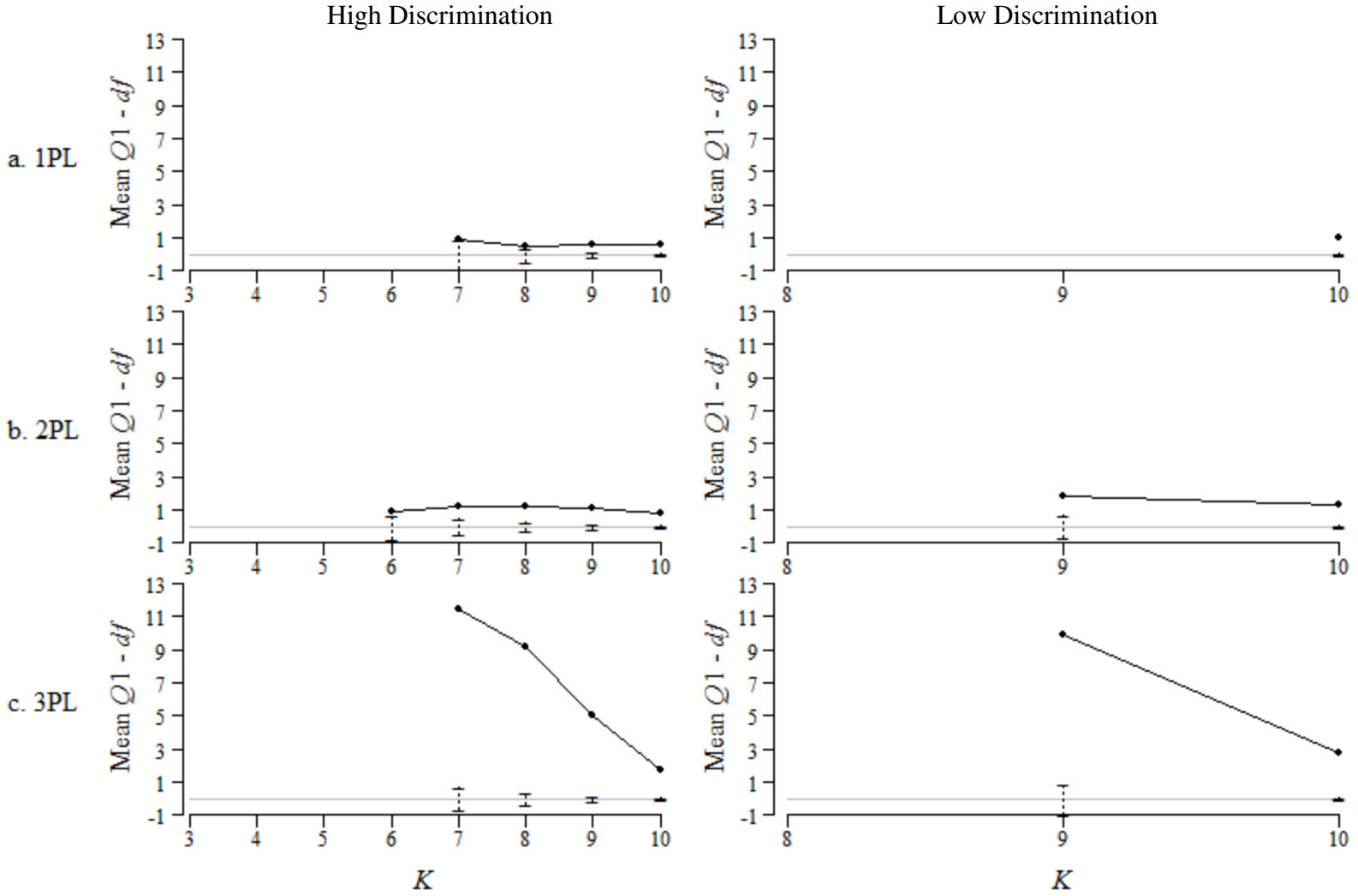


Figure F-9. De-trended $Q1$ SDs by K for SU Conditions (N = 500 n = 15)

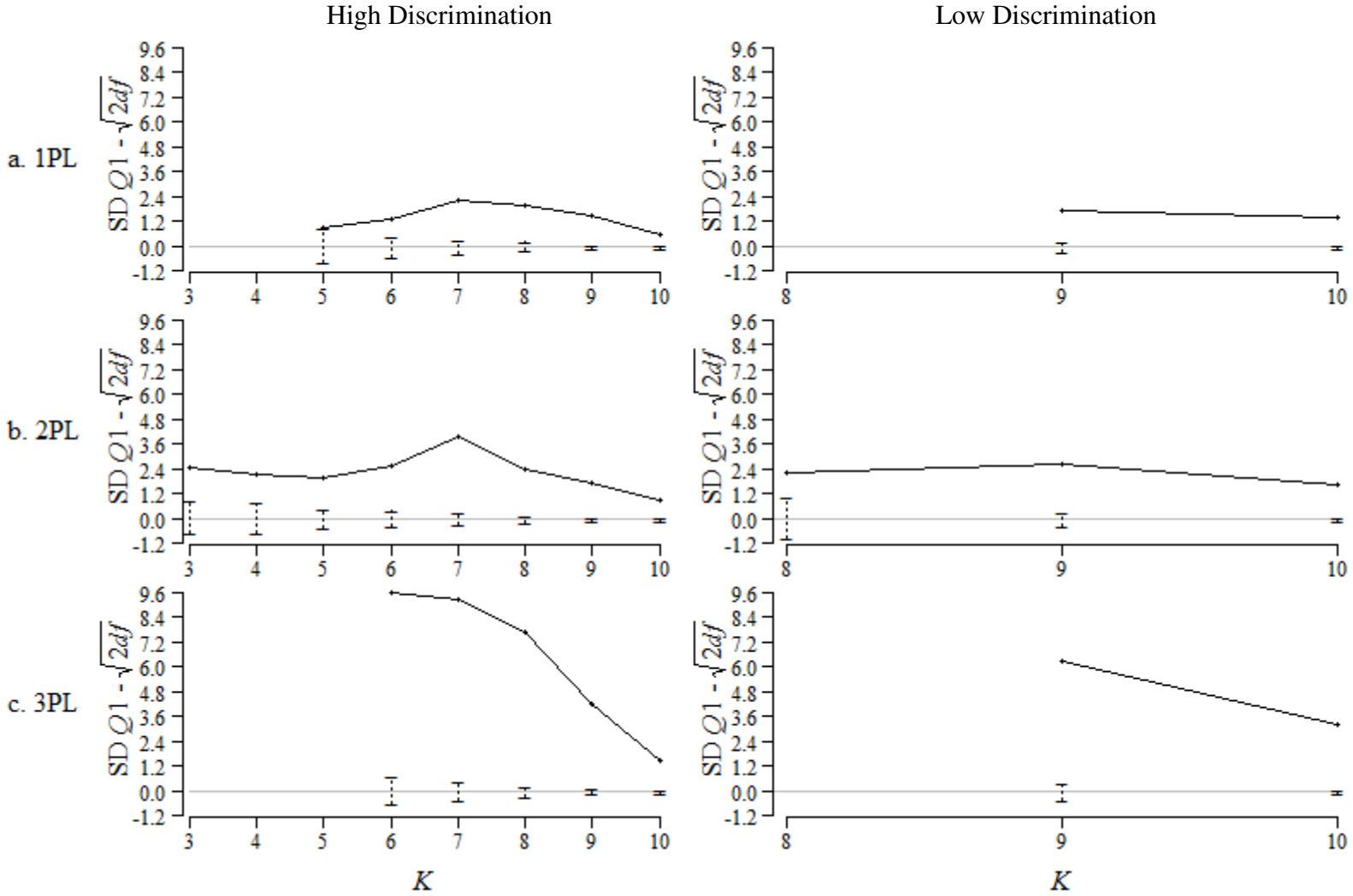


Figure F-10. De-trended $Q1$ SDs by K for EU Conditions (N = 500 n = 15)

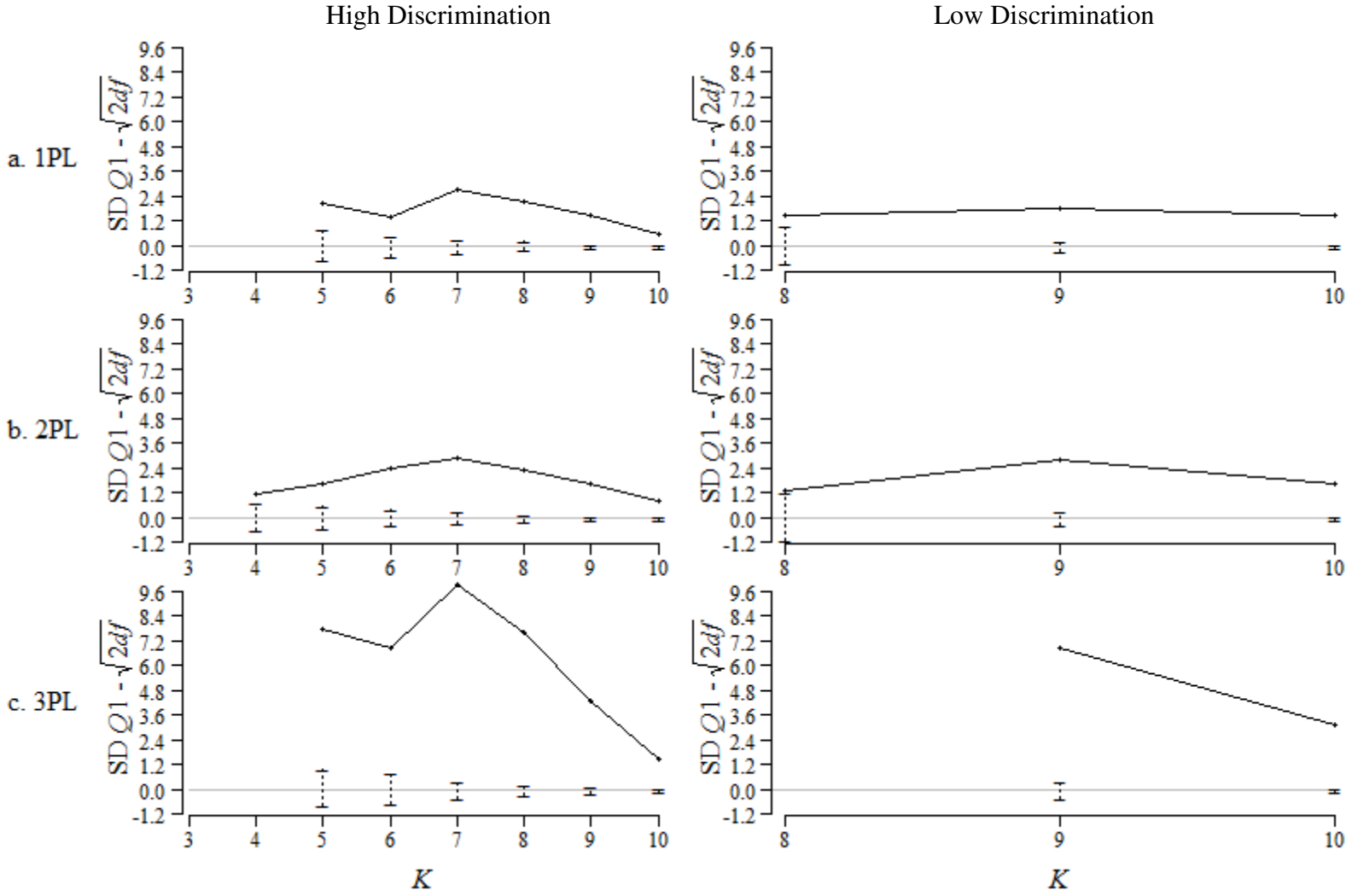


Figure F-11. De-trended $Q1$ SDs by K for SU Conditions (N = 1,500 n = 15)

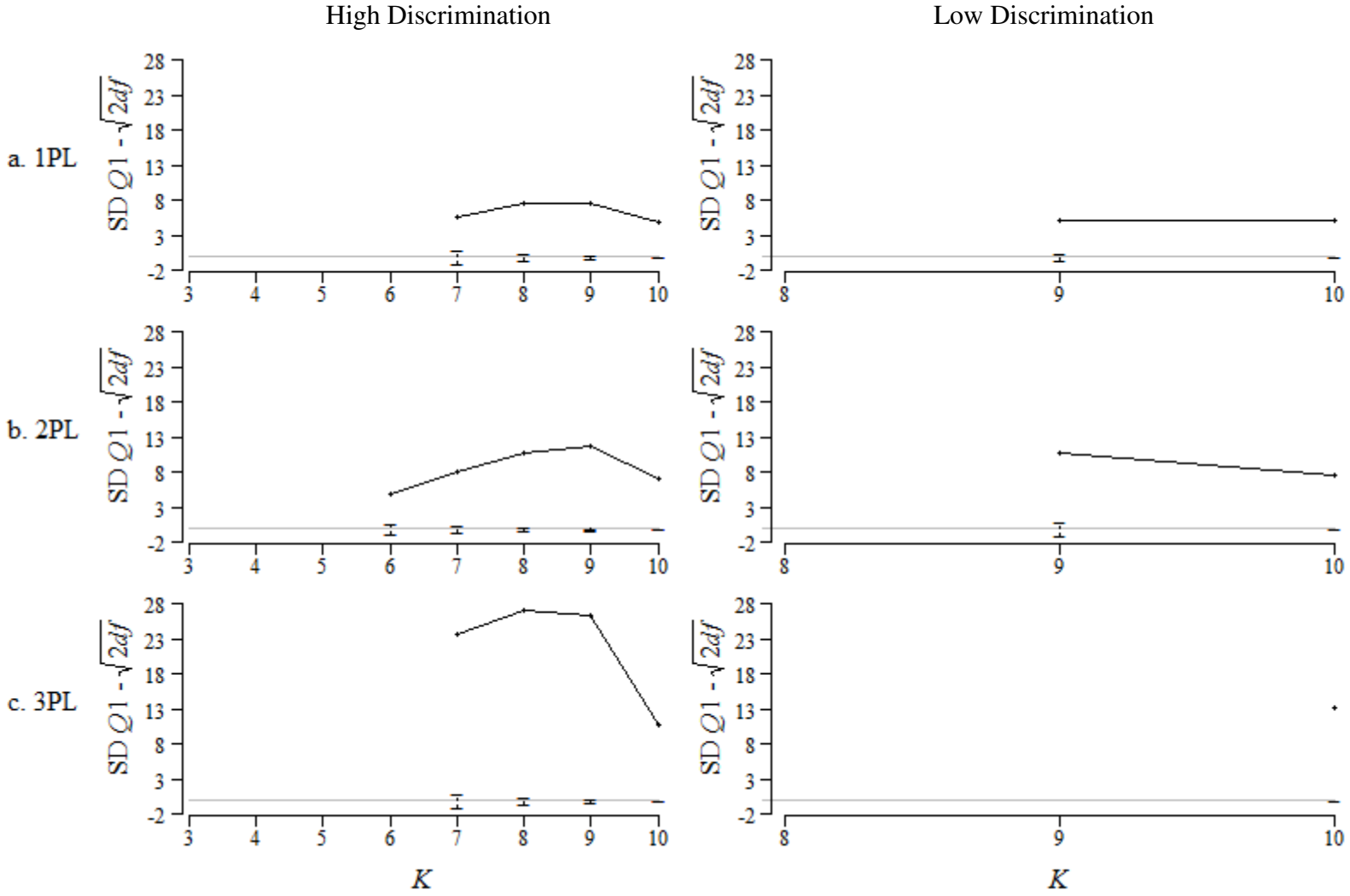


Figure F-12. De-trended $Q1$ SDs by K for EU Conditions (N = 1,500 n = 15)

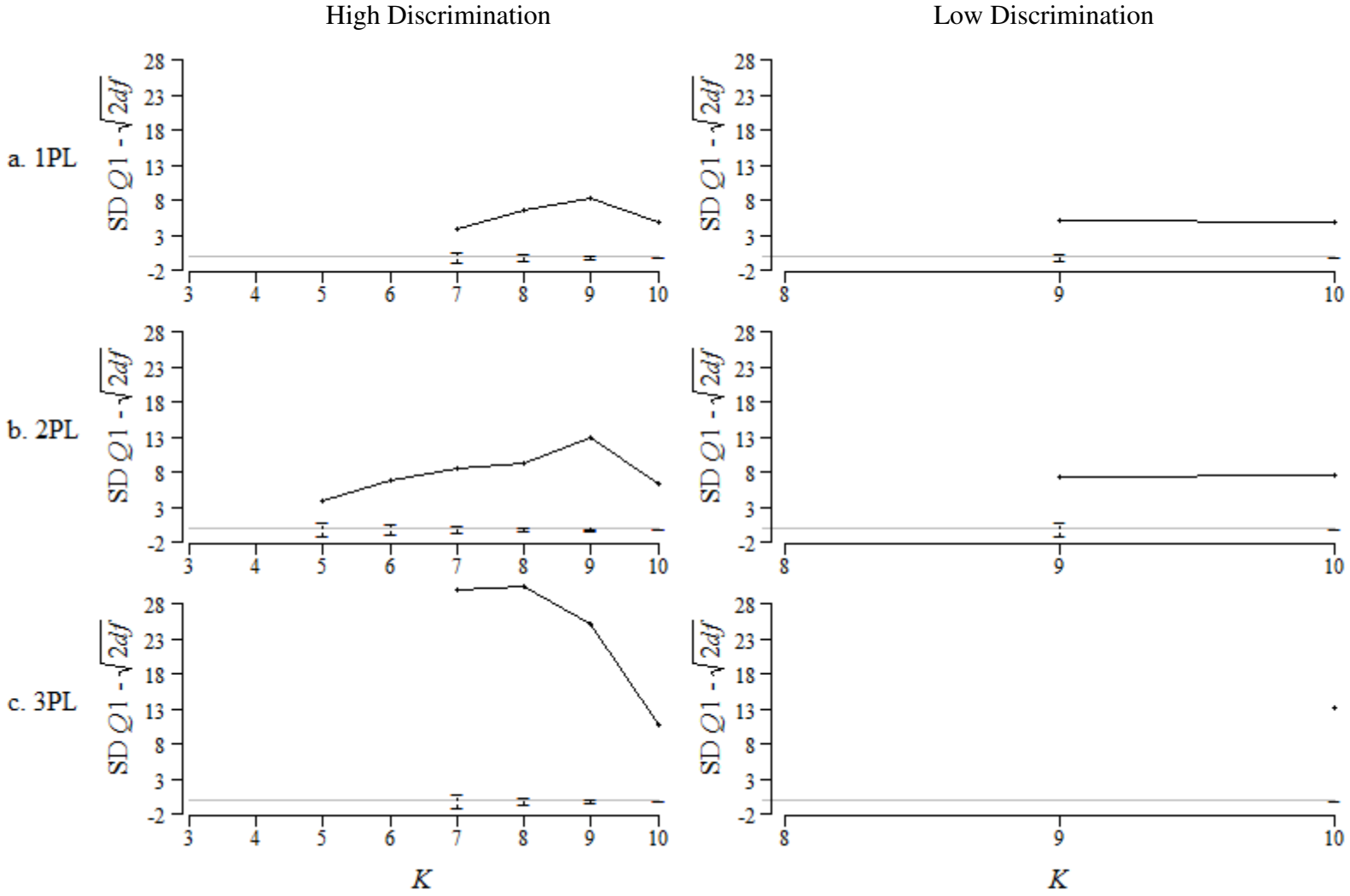


Figure F-13. De-trended $Q1$ SDs by K for SU Conditions (N = 500 n = 75)

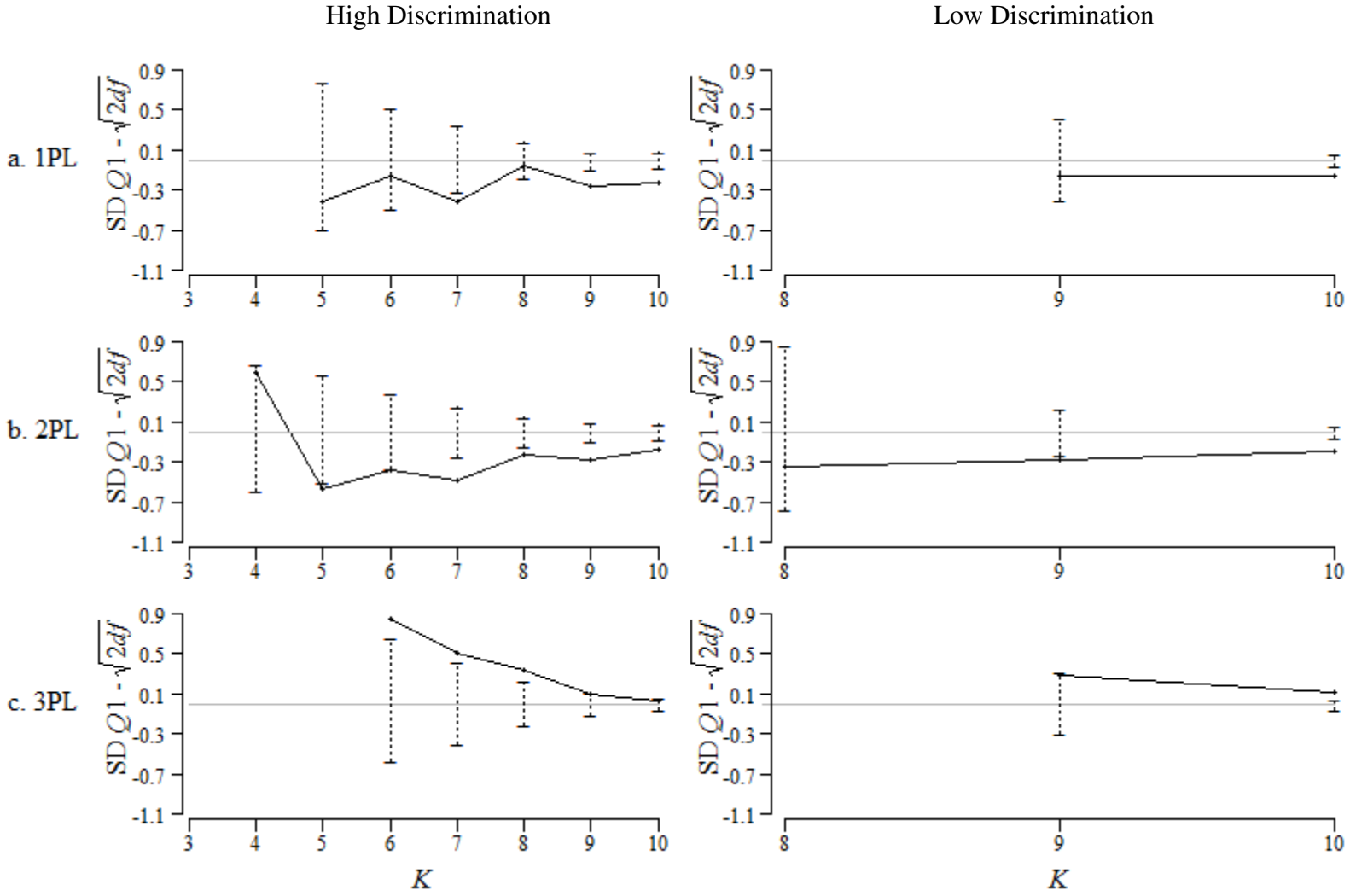


Figure F-14. De-trended $Q1$ SDs by K for EU Conditions (N = 500 n = 75)

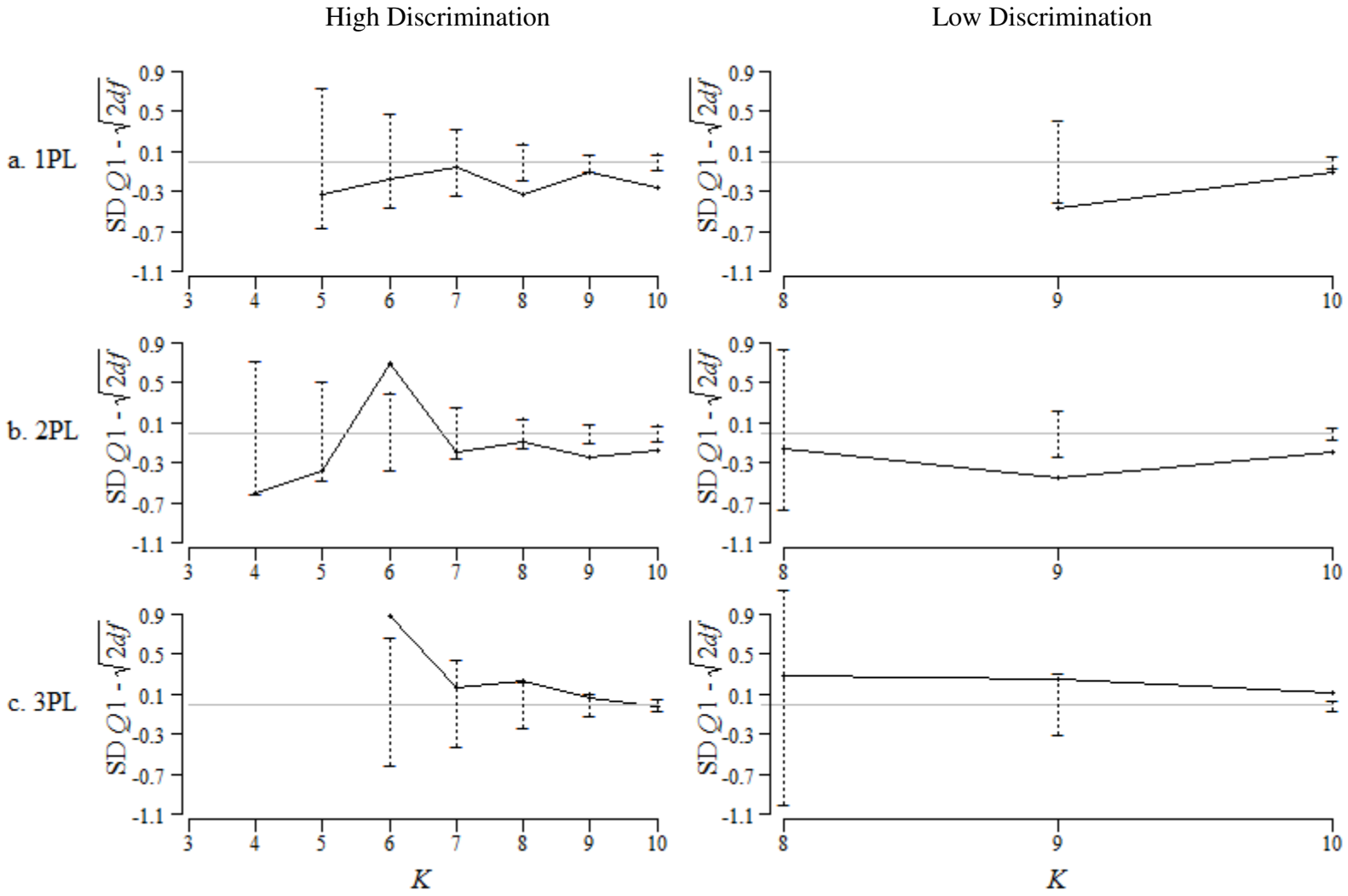


Figure F-15. De-trended $Q1$ SDs by K for SU Conditions (N = 1,500 n = 75)

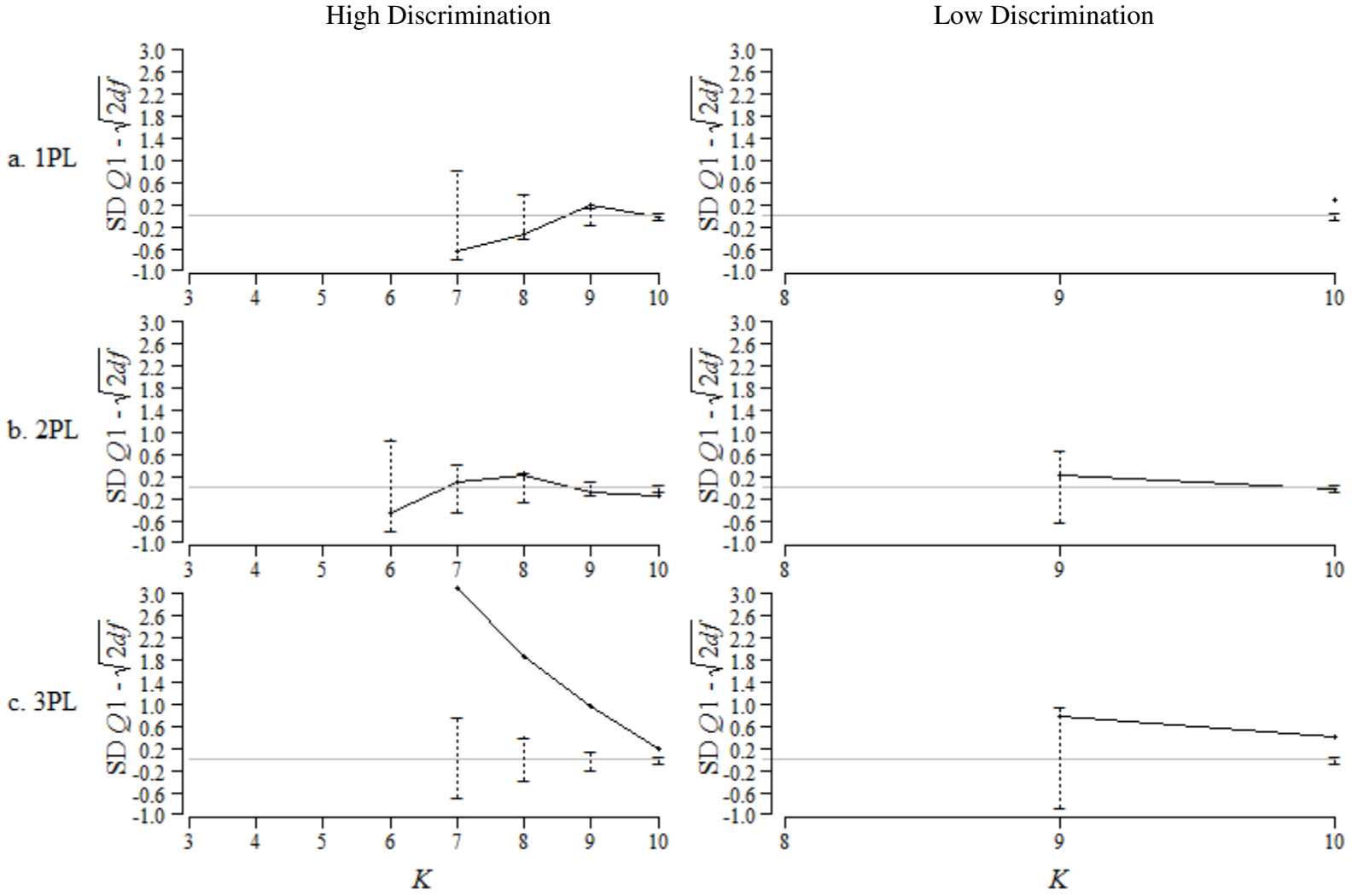


Figure F-16. De-trended $Q1$ SDs by K for EU Conditions (N = 1,500 n = 75)

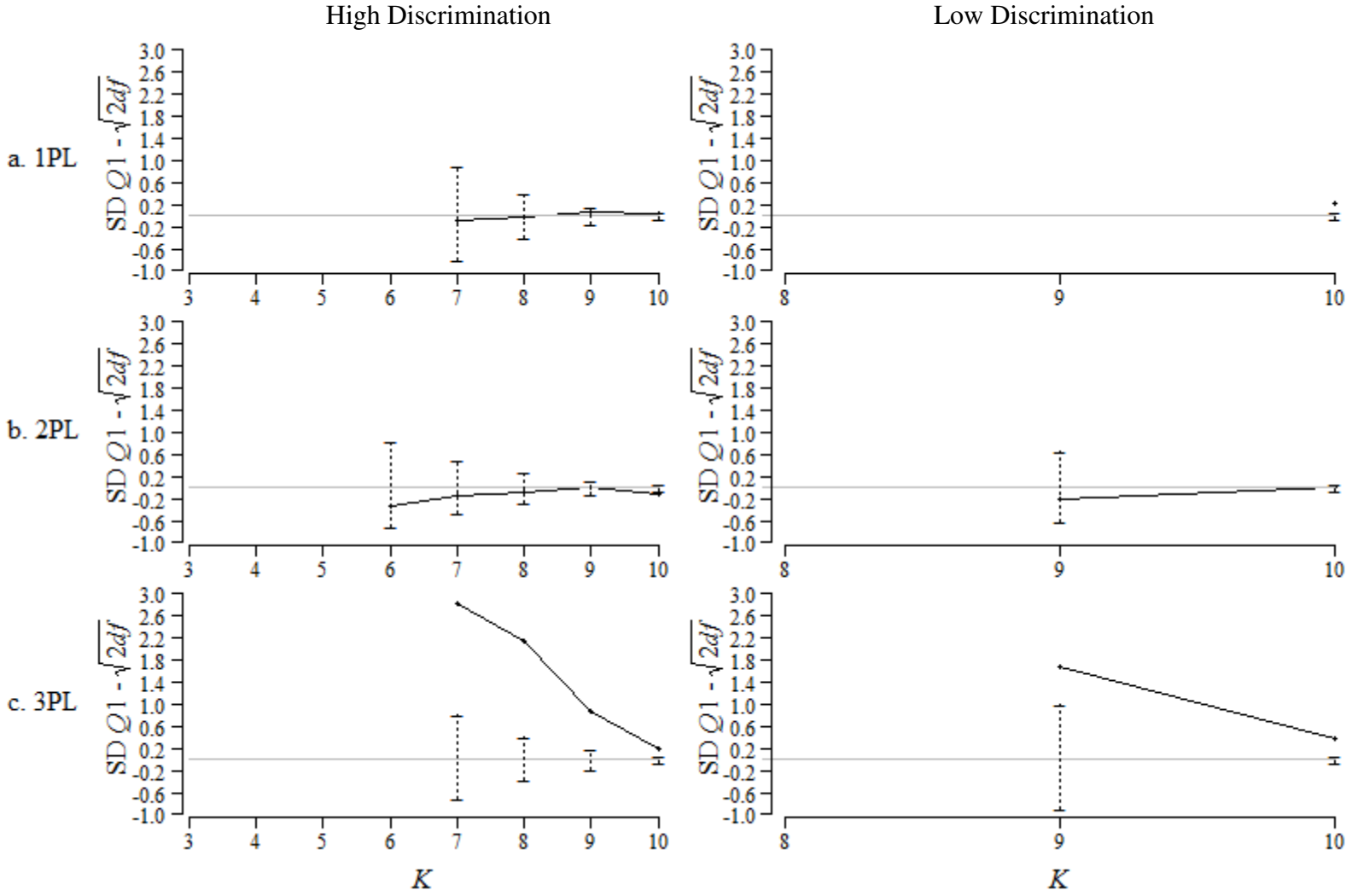


Table F-9. Bias of $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 500 n = 15)

D	PE		Bias(Mean)			Bias(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	4.86	6.02	18.07	1.40	2.26	6.43
		$\boldsymbol{\xi}$	6.15	5.81	17.73	2.82	4.01	10.09
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.86	1.88	–0.24	0.06	0.28
		$\boldsymbol{\xi}$	–0.26	–0.06	–0.13	–0.08	0.28	–0.14
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	6.45	8.65	17.57	1.58	2.16	4.73
		$\boldsymbol{\xi}$	6.52	6.50	14.57	1.83	3.04	6.82
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.02	0.57	2.02	–0.07	–0.17	0.23
		$\boldsymbol{\xi}$	–0.21	–0.30	–0.14	–0.03	–0.15	–0.22

Table F-10. Bias of $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 1,500 n = 15)

D	PE		Bias(Mean)			Bias(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	16.18	18.96	55.14	6.37	8.57	21.94
		$\boldsymbol{\xi}$	19.14	19.25	57.06	8.52	11.42	25.65
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.16	0.80	2.07	–0.24	–0.04	0.29
		$\boldsymbol{\xi}$	–0.07	–0.21	0.07	–0.13	0.00	0.23
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	22.94	25.70	33.36	5.16	9.11	13.29
		$\boldsymbol{\xi}$	23.00	22.95	34.51	5.29	10.20	17.95
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.16	0.72	1.97	0.00	0.03	0.52
		$\boldsymbol{\xi}$	–0.04	–0.09	0.16	0.01	–0.02	0.14

Table F-11. Bias of $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 500 n = 75)

D	PE		Bias(Mean)			Bias(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.37	2.82	–0.25	–0.22	0.36
		$\boldsymbol{\xi}$	0.66	0.43	3.24	0.51	0.33	1.92
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.12	0.54	1.31	0.01	0.01	–0.09
		$\boldsymbol{\xi}$	–0.05	–0.08	–0.05	0.21	0.16	0.08
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.14	0.55	2.40	–0.15	–0.27	0.19
		$\boldsymbol{\xi}$	0.02	0.27	1.23	0.15	0.08	1.07
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.06	0.47	1.00	0.13	–0.29	–0.28
		$\boldsymbol{\xi}$	–0.10	0.15	–0.19	–0.01	0.11	–0.21

Table F-12. Bias of $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 1,500 n = 75)

D	PE		Bias(Mean)			Bias(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.20	1.11	6.95	–0.20	–0.07	1.53
		$\boldsymbol{\xi}$	0.78	1.89	10.29	0.53	1.19	5.03
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.04	0.60	2.22	–0.11	–0.08	0.18
		$\boldsymbol{\xi}$	–0.33	–0.09	–0.28	–0.05	0.10	–0.07
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.99	1.52	5.60	0.28	0.10	0.59
		$\boldsymbol{\xi}$	1.09	0.84	5.52	0.31	0.39	2.33
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.07	0.25	0.71	–0.01	–0.15	–0.21
		$\boldsymbol{\xi}$	–0.07	–0.25	–0.36	–0.01	–0.08	–0.09

Table F-13. Mean Error for $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 500 n = 15)

D	PE		ME(Mean)			ME(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	4.86	6.02	18.07	1.40	2.26	6.43
		$\boldsymbol{\xi}$	7.15	7.81	20.73	3.10	4.62	11.00
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	1.01	1.14	1.12	0.51	0.55	0.58
		$\boldsymbol{\xi}$	0.26	0.18	0.16	0.15	0.35	0.19
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	6.45	8.65	17.57	1.58	2.16	4.73
		$\boldsymbol{\xi}$	7.52	8.50	17.57	2.06	3.54	7.57
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.98	1.43	0.98	0.31	0.67	0.55
		$\boldsymbol{\xi}$	0.21	0.30	0.29	0.03	0.17	0.35

Table F-14. Mean Error for $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 1,500 n = 15)

D	PE		ME(Mean)			ME(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	16.18	18.96	55.14	6.37	8.57	21.94
		$\boldsymbol{\xi}$	20.14	21.25	60.06	8.78	11.96	26.47
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.84	1.20	0.93	0.49	0.61	0.53
		$\boldsymbol{\xi}$	0.08	0.22	0.11	0.15	0.19	0.26
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	22.94	25.70	33.36	5.16	9.11	13.29
		$\boldsymbol{\xi}$	24.00	24.95	37.51	5.52	10.68	18.68
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.84	1.28	1.03	0.23	0.45	0.24
		$\boldsymbol{\xi}$	0.04	0.09	0.17	0.01	0.04	0.14

Table F-15. Mean Error for $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 500 n = 75)

D	PE		ME(Mean)			ME(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.20	0.39	2.82	0.25	0.38	0.36
		$\boldsymbol{\xi}$	1.66	2.43	6.24	0.78	0.94	2.82
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.88	1.46	1.69	0.26	0.60	0.95
		$\boldsymbol{\xi}$	0.06	0.11	0.09	0.21	0.21	0.21
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.18	0.55	2.40	0.15	0.27	0.19
		$\boldsymbol{\xi}$	1.02	2.27	4.23	0.39	0.59	1.86
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.94	1.53	2.00	0.11	0.80	1.06
		$\boldsymbol{\xi}$	0.10	0.25	0.19	0.02	0.11	0.21

Table F-16. Mean Error for $Q1$ Sampling Distribution Means and SDs in SU Conditions (N = 1,500 n = 75)

D	PE		ME(Mean)			ME(SD)		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.56	1.11	6.95	0.30	0.20	1.53
		$\boldsymbol{\xi}$	1.78	3.89	13.29	0.78	1.73	5.85
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	1.04	1.40	0.78	0.36	0.62	0.64
		$\boldsymbol{\xi}$	0.35	0.33	0.28	0.17	0.13	0.23
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.99	1.52	5.60	0.28	0.12	0.59
		$\boldsymbol{\xi}$	2.09	2.84	8.52	0.54	0.88	3.09
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.93	1.75	2.29	0.24	0.64	0.97
		$\boldsymbol{\xi}$	0.07	0.25	0.37	0.01	0.08	0.09

Table F-17. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that $Q1$ Followed its Theoretical Distribution ($N = 500$ $n = 15$)

M	DN	D	PE Condition							
			$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
			F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$
1PL	EU	High	7	7	4	7	7	7	2	7
		Low	3	3	1	2	3	3	1	2
	SU	High	7	7	2	7	7	7	2	7
		Low	3	3	1	2	3	3	1	2
	Model Total		**20**	**20**	**8**	**18**	**20**	**20**	**6**	**18**
2PL	EU	High	8	8	8	8	8	8	1	8
		Low	3	3	3	4	4	4	1	4
	SU	High	8	8	8	8	8	8	3	8
		Low	3	3	2	3	4	4	2	4
	Model Total		**22**	**22**	**21**	**23**	**24**	**24**	**7**	**24**
3PL	EU	High	6	6	6	6	7	7	1	7
		Low	2	2	2	3	3	3	1	3
	SU	High	6	6	6	6	7	7	0	6
		Low	2	2	3	3	3	3	0	3
	Model Total		**16**	**16**	**17**	**18**	**20**	**20**	**2**	**19**

Table F-18. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that $Q1$ Followed its Theoretical Distribution (N = 1,500 n = 15)

M	DN	D	PE Condition $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$		$\hat{\boldsymbol{\xi}}, \boldsymbol{\theta}$		$\boldsymbol{\xi}, \hat{\boldsymbol{\theta}}$		$\boldsymbol{\xi}, \boldsymbol{\theta}$	
			F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$
1PL	EU	High	4	4	2	4	5	5	1	5
		Low	2	2	1	1	2	2	1	1
	SU	High	5	5	3	4	5	5	3	5
		Low	2	2	1	1	2	2	0	1
	Model Total		**13**	**13**	**7**	**10**	**14**	**14**	**5**	**12**
2PL	EU	High	6	6	6	6	6	6	2	6
		Low	2	2	2	2	2	2	1	2
	SU	High	6	6	5	6	6	6	2	6
		Low	2	2	1	2	2	2	0	2
	Model Total		**16**	**16**	**14**	**16**	**16**	**16**	**5**	**16**
3PL	EU	High	4	4	4	4	4	4	3	4
		Low	2	2	2	2	2	2	1	2
	SU	High	4	4	4	4	4	4	1	5
		Low	2	2	2	2	2	2	0	2
	Model Total		**12**	**12**	**12**	**12**	**12**	**12**	**5**	**13**

Table F-19. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that $Q1$ Followed its Theoretical Distribution ($N = 500$ $n = 75$)

M	DN	D	PE Condition							
			$\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
			F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$
1PL	EU	High	3	7	3	7	3	7	2	7
		Low	1	2	1	2	1	2	0	2
	SU	High	4	7	3	7	5	7	1	7
		Low	1	2	1	2	1	2	1	2
	Model Total		**9**	**18**	**8**	**18**	**10**	**18**	**4**	**18**
2PL	EU	High	8	8	8	8	4	8	2	8
		Low	3	4	2	4	1	4	0	4
	SU	High	8	8	8	8	4	8	2	8
		Low	2	4	4	4	1	4	1	4
	Model Total		**21**	**24**	**22**	**24**	**10**	**24**	**5**	**24**
3PL	EU	High	6	6	6	6	6	7	0	7
		Low	3	3	3	3	3	3	1	3
	SU	High	6	6	6	6	7	7	0	7
		Low	3	3	3	3	2	3	1	3
	Model Total		**18**	**18**	**18**	**18**	**18**	**20**	**2**	**20**

Table F-20. Frequency of Cases in Which KS Test Rejected the Null Hypothesis that $Q1$ Followed its Theoretical Distribution (N = 1,500 n = 75)

M	DN	D	PE Condition: $\hat{\xi}, \hat{\theta}$		$\hat{\xi}, \theta$		$\xi, \hat{\theta}$		ξ, θ	
			F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$	F_K	No. $N_K \geq 15$
1PL	EU	High	3	5	2	5	4	5	0	5
		Low	1	1	2	2	1	1	1	1
	SU	High	3	5	2	5	3	5	3	5
		Low	1	1	1	1	1	1	1	1
	Model Total		**8**	**12**	**7**	**13**	**9**	**12**	**5**	**12**
2PL	EU	High	7	7	5	6	4	7	3	7
		Low	2	2	2	2	2	2	1	2
	SU	High	6	7	7	7	6	7	4	7
		Low	2	2	1	2	1	2	0	2
	Model Total		**17**	**18**	**15**	**17**	**13**	**18**	**8**	**18**
3PL	EU	High	4	4	4	4	5	5	4	5
		Low	2	2	2	2	2	2	1	2
	SU	High	4	4	4	4	5	5	3	5
		Low	2	2	1	2	2	2	0	2
	Model Total		**12**	**12**	**11**	**12**	**14**	**14**	**8**	**14**

Table F-21. Correlations Between $Q1$ and b Within K in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Model	K	n = 15				n = 75			
		N = 500		N = 1,500		N = 500		N = 1,500	
		N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination									
1PL	5	72	−0.01			67	−0.06		
	6	161	−0.07			215	0.20		
	7	408	0.00	58	0.16	423	−0.01	66	−0.22
	8	1734	0.04	347	−0.03	1846	0.01	401	−0.01
	9	8215	−0.01	3057	−0.01	8130	0.01	3222	−0.01
	10	8123	0.00	15265	−0.01	8037	0.01	15028	0.01
2PL	4	54	−0.07			68	0.03		
	5	147	−0.07	33	−0.17	115	0.11		
	6	323	−0.08	78	−0.03	302	0.01	68	−0.11
	7	672	−0.01	231	−0.12	730	0.04	222	−0.03
	8	2311	−0.02	799	−0.03	2313	0.03	724	−0.06
	9	6981	0.01	3690	−0.01	7028	−0.03	3614	−0.01
	10	8228	0.01	13904	−0.01	8162	0.02	14072	0.01
3PL	5	42	−0.20			50	0.02		
	6	115	0.08			91	0.10		
	7	266	0.02	60	−0.10	275	−0.04	90	0.05
	8	922	−0.08	280	0.04	903	0.08	299	0.07
	9	3587	0.01	1629	0.03	3444	0.00	1610	0.00
	10	13801	0.01	16756	0.00	13962	−0.01	16727	−0.01
Low Discrimination									
1PL	9	486	0.03			492	0.05		
	10	18250	−0.01	18743	0.00	18251	0.00	18737	−0.01
2PL	8	94	−0.14			96	−0.14		
	9	1286	0.04	192	−0.03	1229	0.05	160	0.22
	10	17336	0.00	18547	0.01	17389	0.00	18575	−0.01
3PL	8	39	−0.20			42	0.17		
	9	588	0.01	58	0.22	583	−0.04	71	−0.05
	10	18114	−0.01	18687	0.00	18111	0.01	18674	0.02

Table F-22. Correlations Between $Q1$ and a Within K in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

Model	K	n = 15				n = 75			
		N = 500		N = 1,500		N = 500		N = 1,500	
		N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination									
2PL	4	54	−0.19			68	0.01		
	5	147	−0.08	33	0.12	115	0.01		
	6	323	−0.02	78	−0.10	302	0.00	68	0.15
	7	672	−0.02	231	0.04	730	0.02	222	−0.01
	8	2311	−0.03	799	0.04	2313	0.00	724	−0.03
	9	6981	0.00	3690	−0.01	7028	−0.01	3614	−0.01
	10	8228	0.00	13904	0.02	8162	−0.01	14072	−0.01
3PL	5	42	−0.23			50	0.03		
	6	115	0.06			91	0.10		
	7	266	0.02	60	−0.12	275	−0.01	90	0.02
	8	922	−0.10	280	0.04	903	0.06	299	0.08
	9	3587	0.01	1629	0.02	3444	0.01	1610	0.01
	10	13801	−0.01	16756	0.00	13962	−0.01	16727	−0.01
Low Discrimination									
	8	94	0.07			96	0.00		
2PL	9	1286	0.00	192	0.11	1229	−0.03	160	−0.08
	10	17336	0.00	18547	−0.01	17389	−0.01	18575	0.00
3PL	8	39	−0.21			42	0.11		
	9	588	0.00	58	0.23	583	−0.01	71	0.03
	10	18114	0.00	18687	−0.01	18111	−0.01	18674	0.01

Table F-23. Correlations Between $Q1$ and c Within K in $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

K	n = 15				n = 75			
	N = 500		N = 1,500		N = 500		N = 1,500	
	N_K	r	N_K	r	N_K	r	N_K	r
High Discrimination								
5	42	0.01			50	0.05		
6	115	0.02			91	−0.19		
7	266	0.11	60	0.04	275	0.05	90	0.00
8	922	−0.02	280	0.08	903	−0.02	299	0.03
9	3587	0.02	1629	0.00	3444	−0.03	1610	0.02
10	13801	0.00	16756	0.00	13962	0.00	16727	−0.01
Low Discrimination								
8	39	−0.24			42	−0.02		
9	588	−0.07	58	−0.18	583	−0.03	71	−0.13
10	18114	0.01	18687	0.00	18111	0.01	18674	0.01

APPENDIX G: ANALYSIS OF LM STATISTICS

LM Distribution Histograms

KS Test π_R in SU Study Conditions

Scatterplots Between *b* and LM Statistics

Scatterplots Between *a* and LM Statistics

Descriptive Statistics for Corrected LM Statistics in SU Conditions

Figure G-1. LM($\alpha\beta$) Distributions for the 1PL

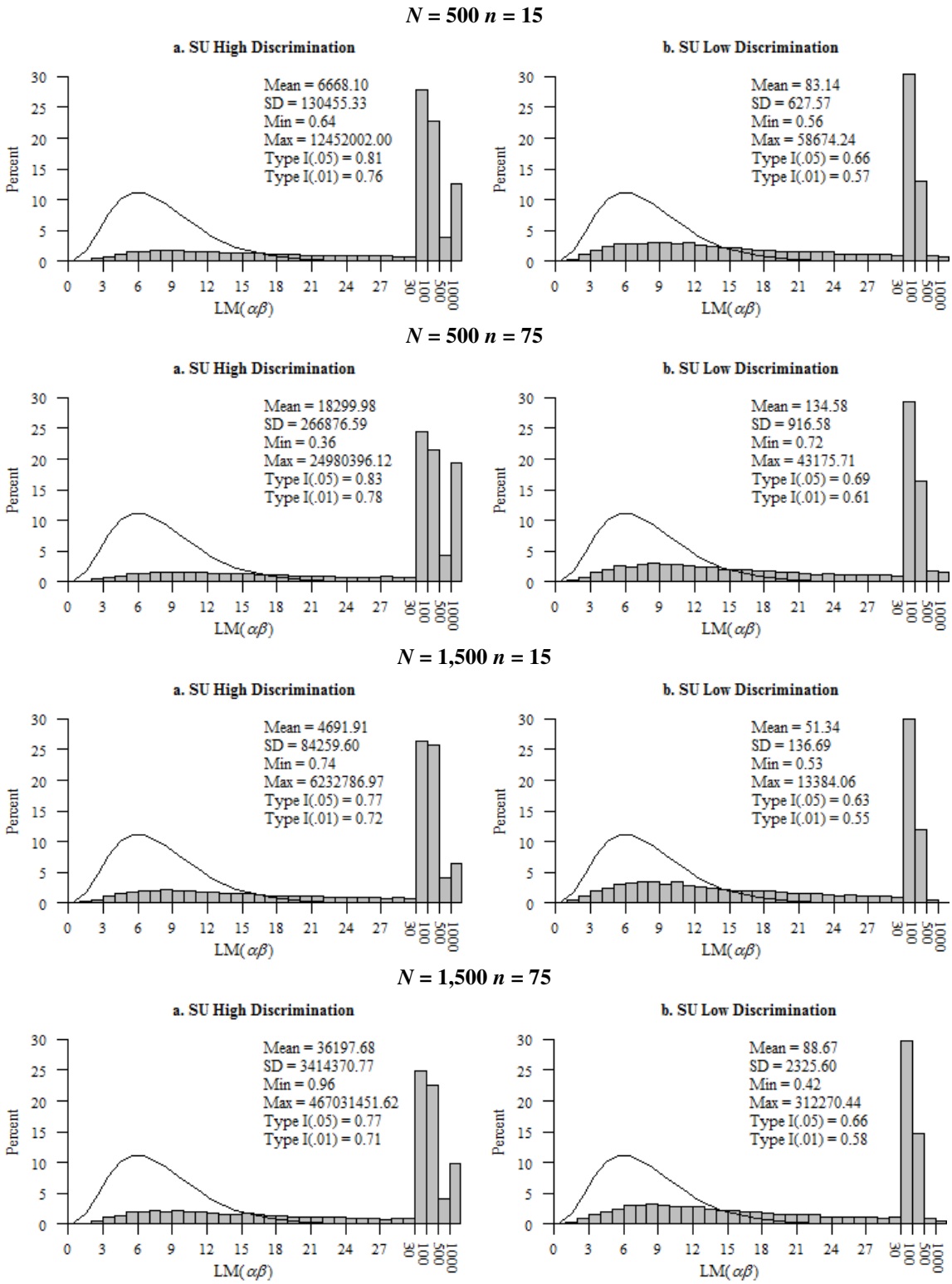


Figure G-2. LM(α) Distributions for the 1PL

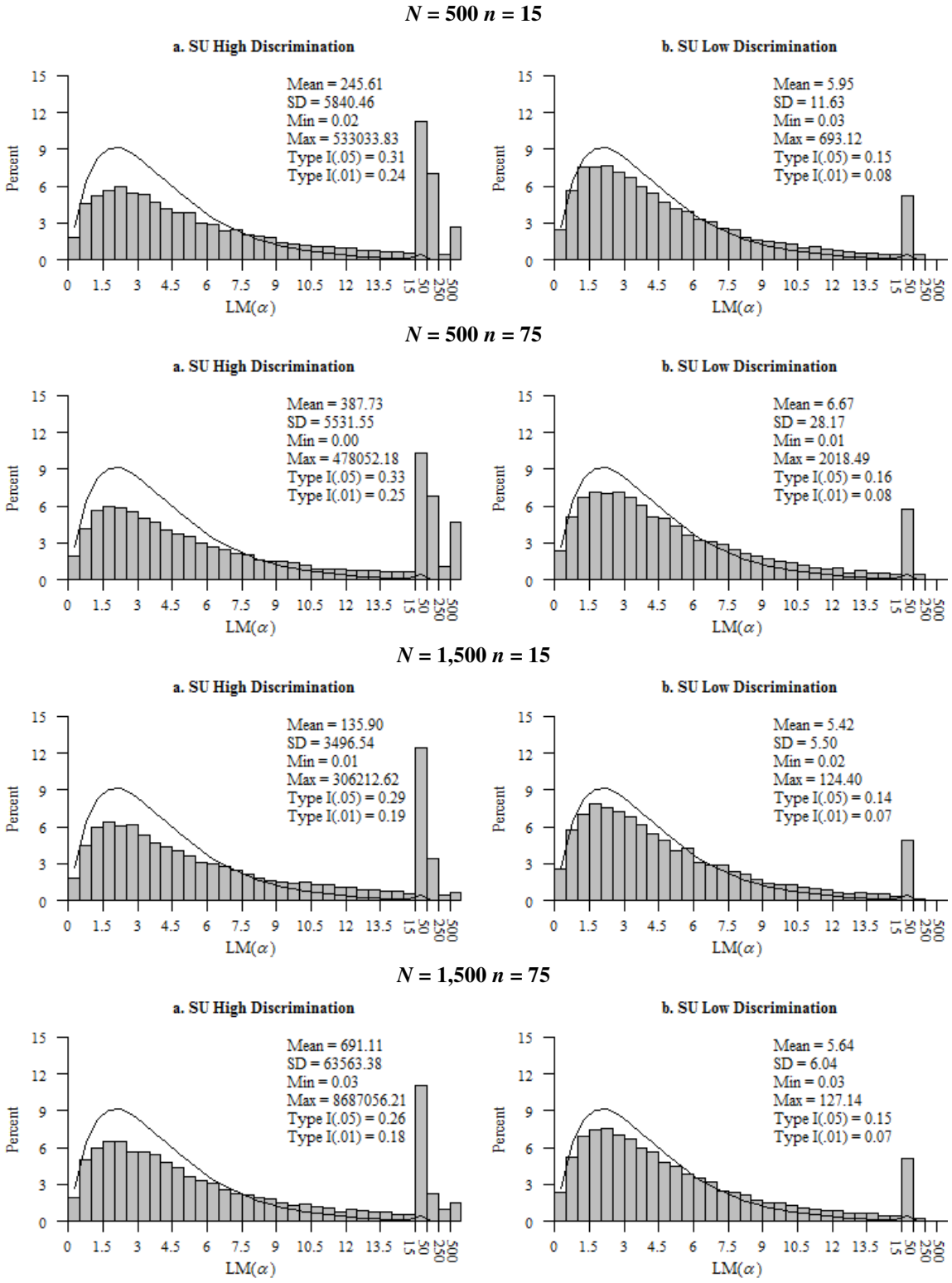


Figure G-3. LM(β) Distributions for the 1PL

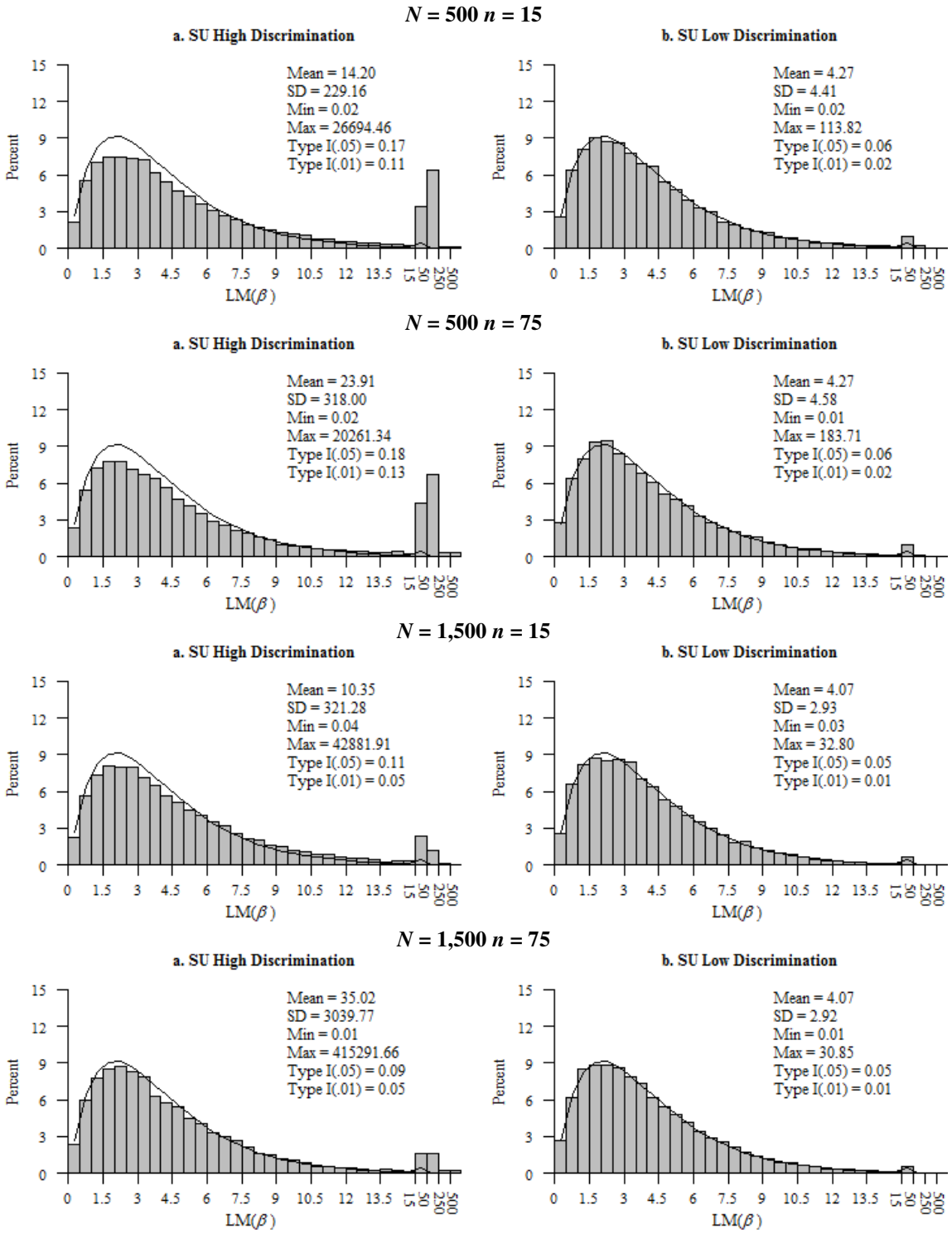


Figure G-4. LM($\alpha\beta$) Distributions for the 2PL

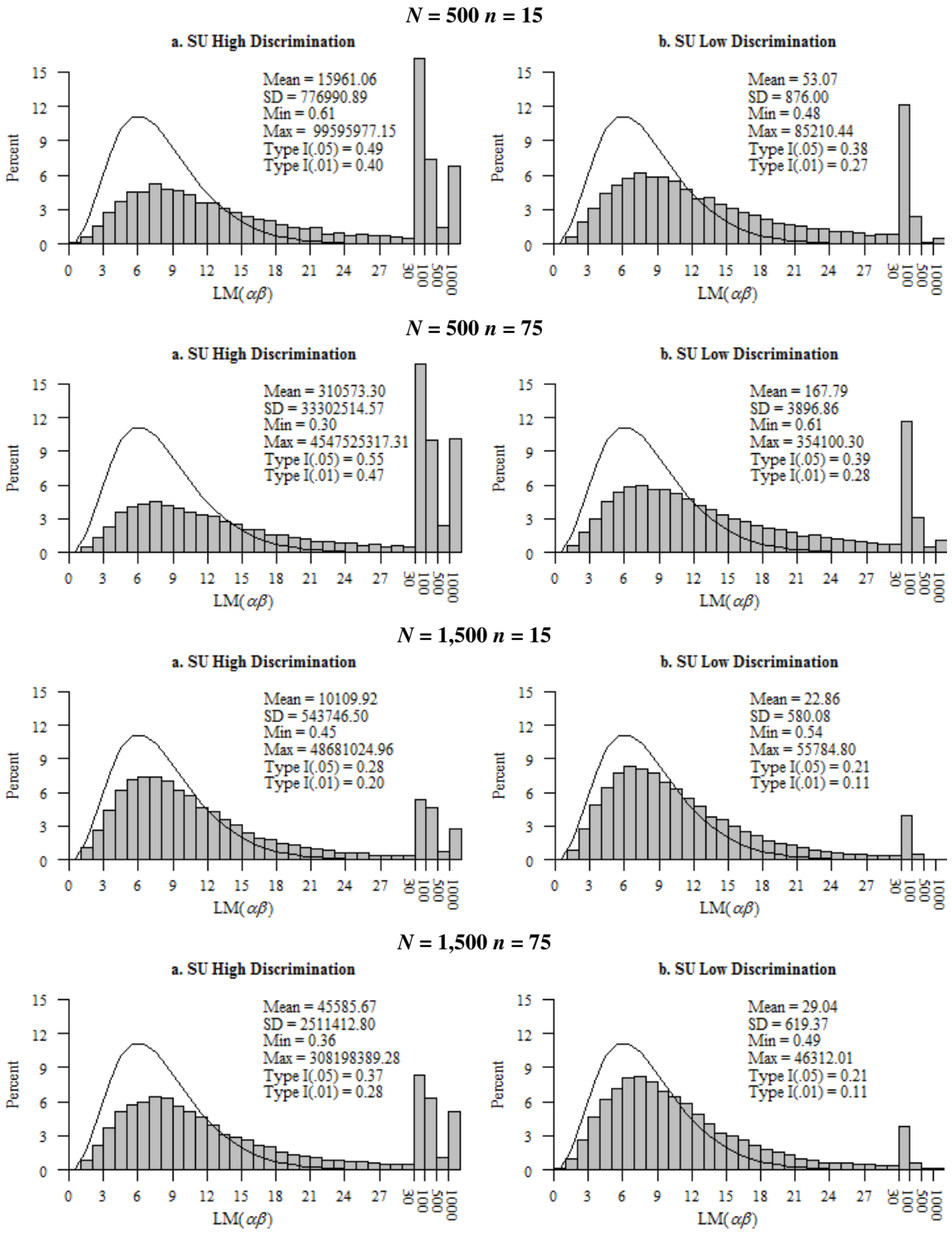


Figure G-5. LM(α) Distributions for the 2PL

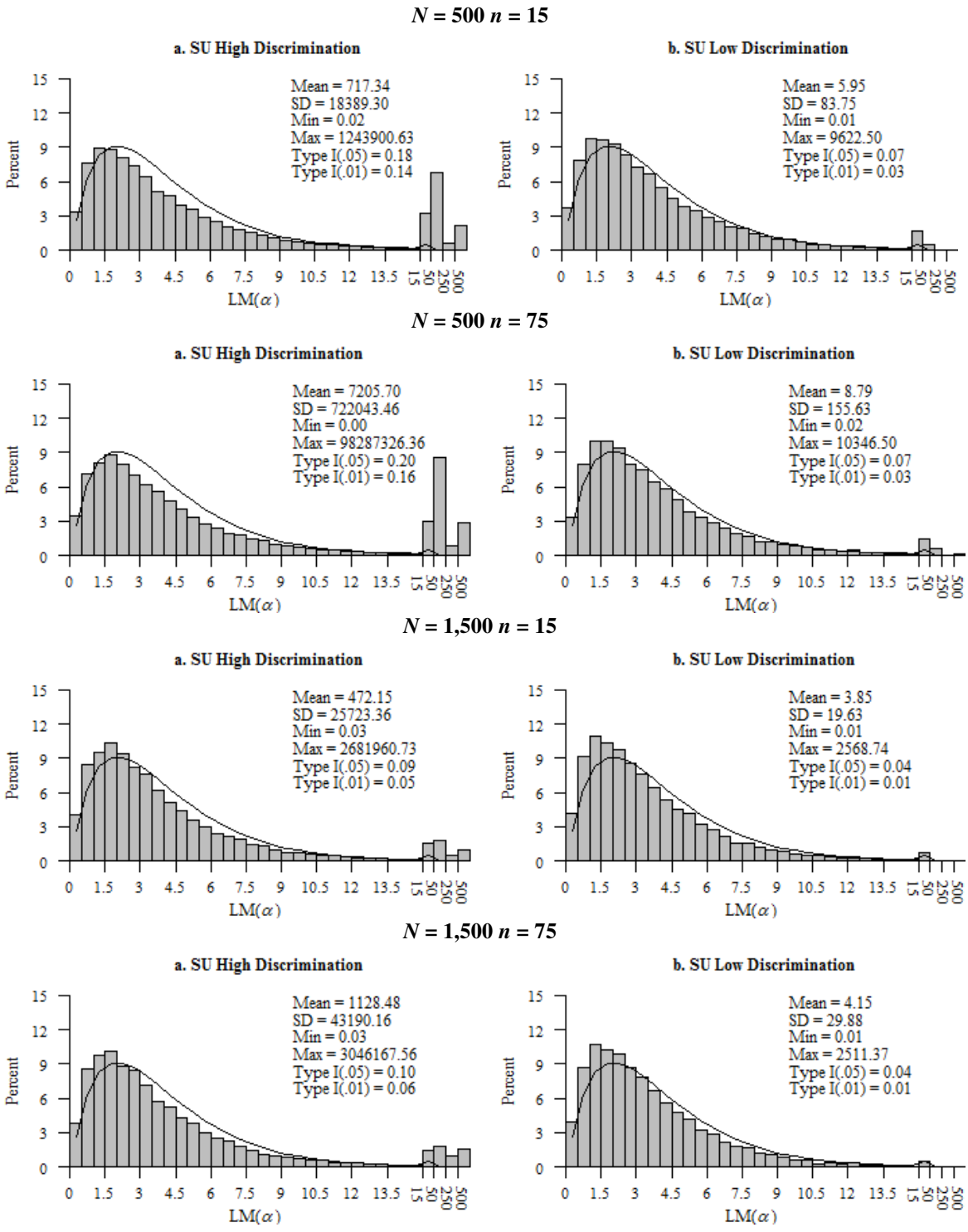


Figure G-6. LM(β) Distributions for the 2PL

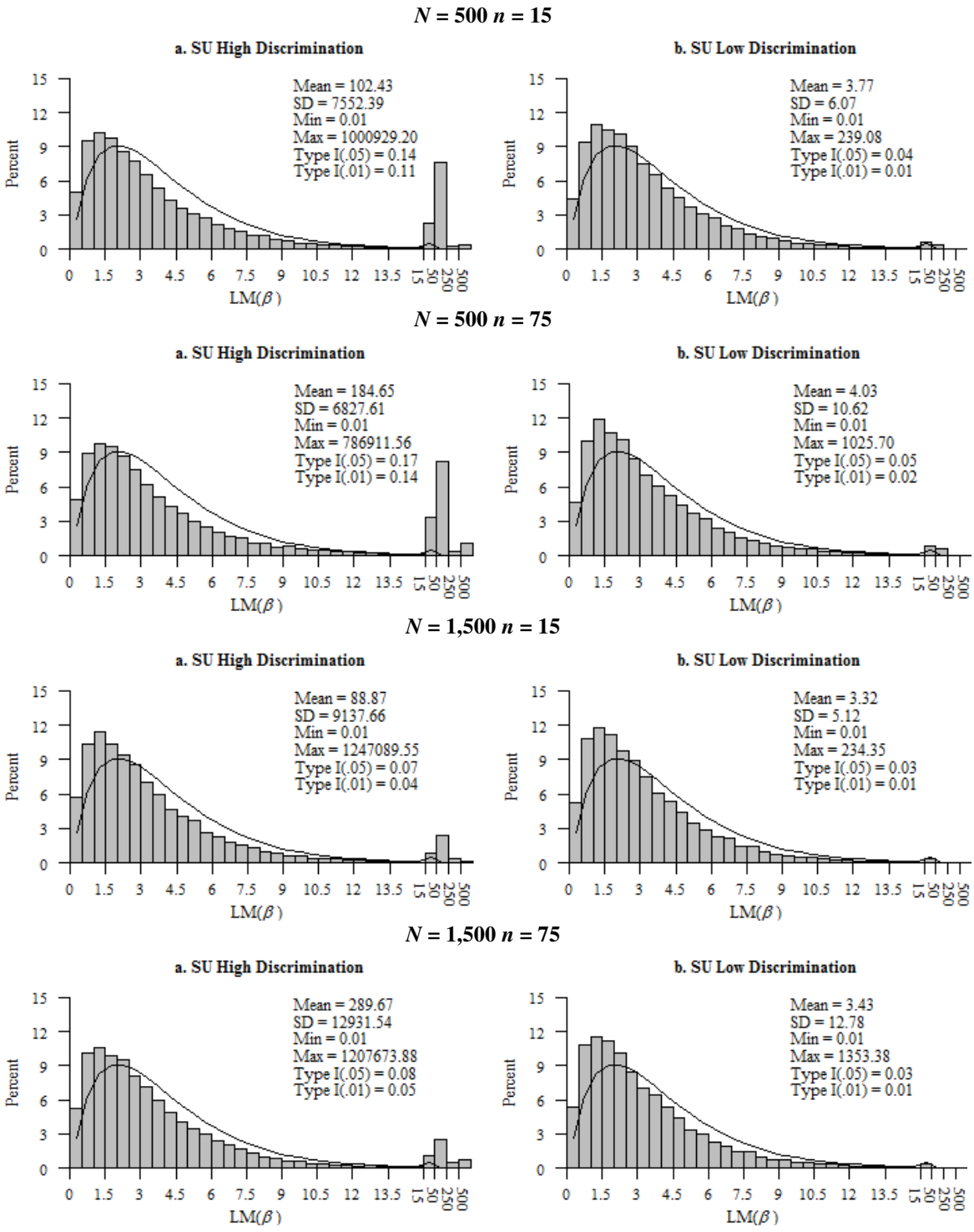


Figure G-7. LM($\alpha\beta$) Distributions for the 3PL

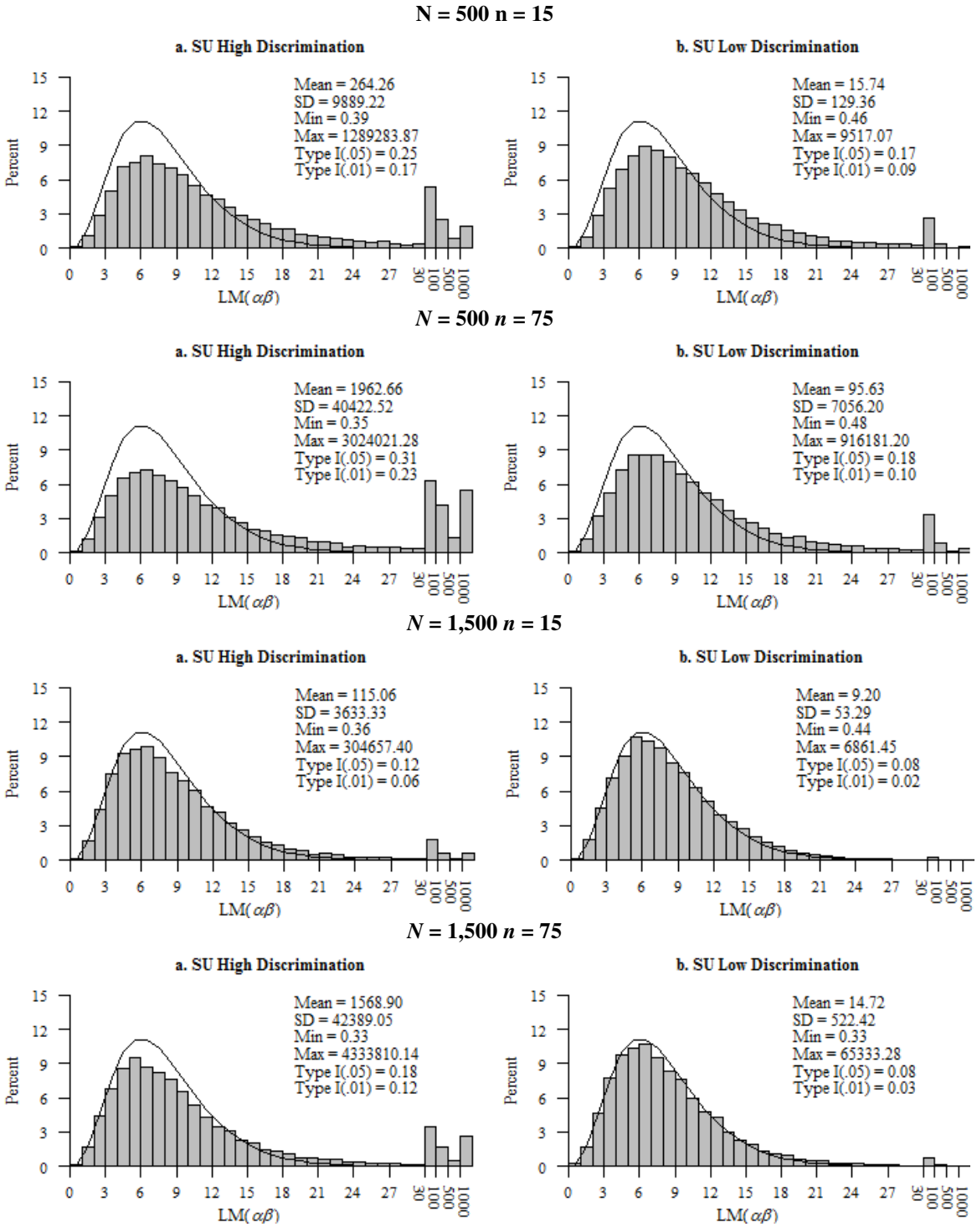


Figure G-8. LM(α) Distributions for the 3PL

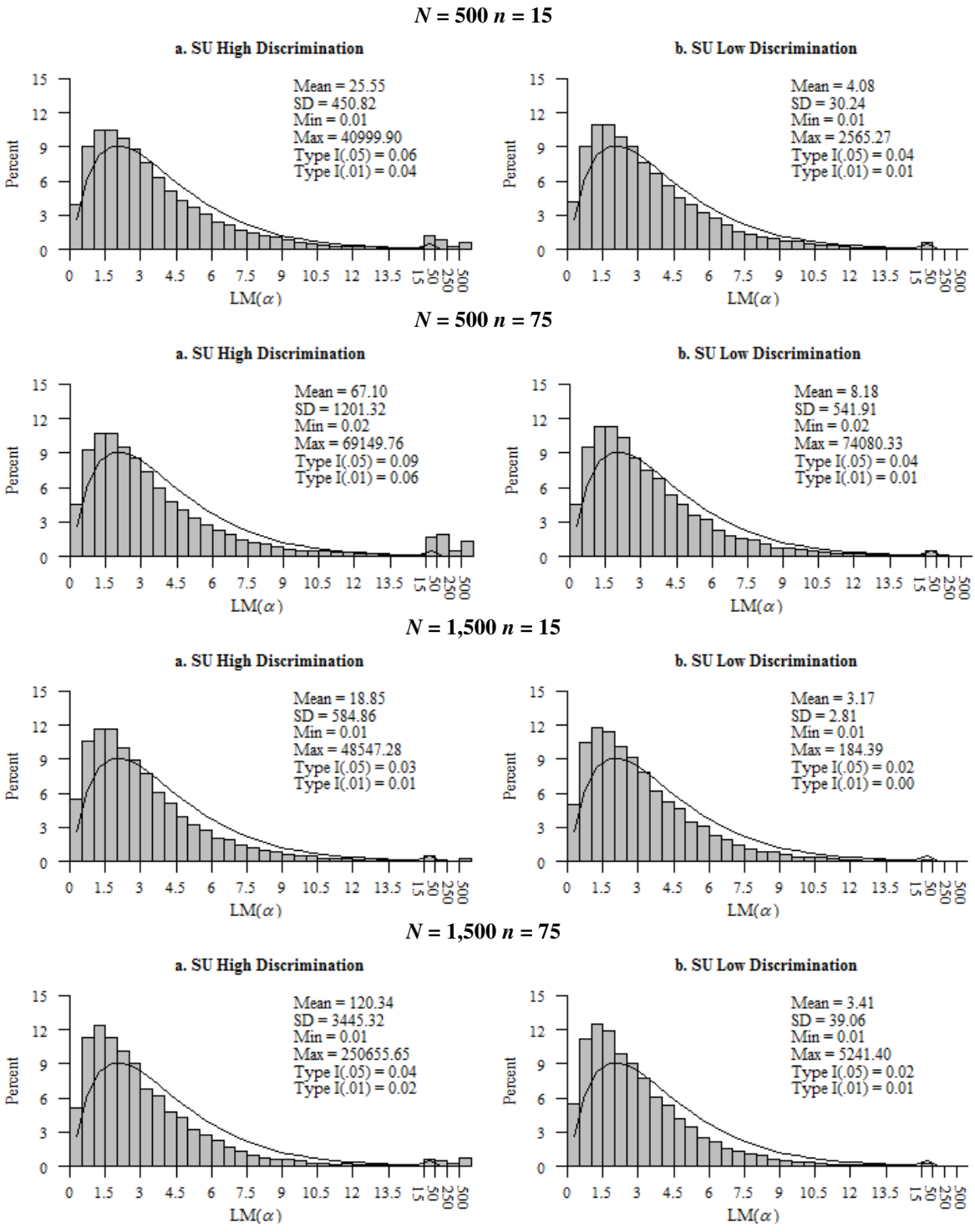


Figure G-9. LM(β) Distributions for the 3PL

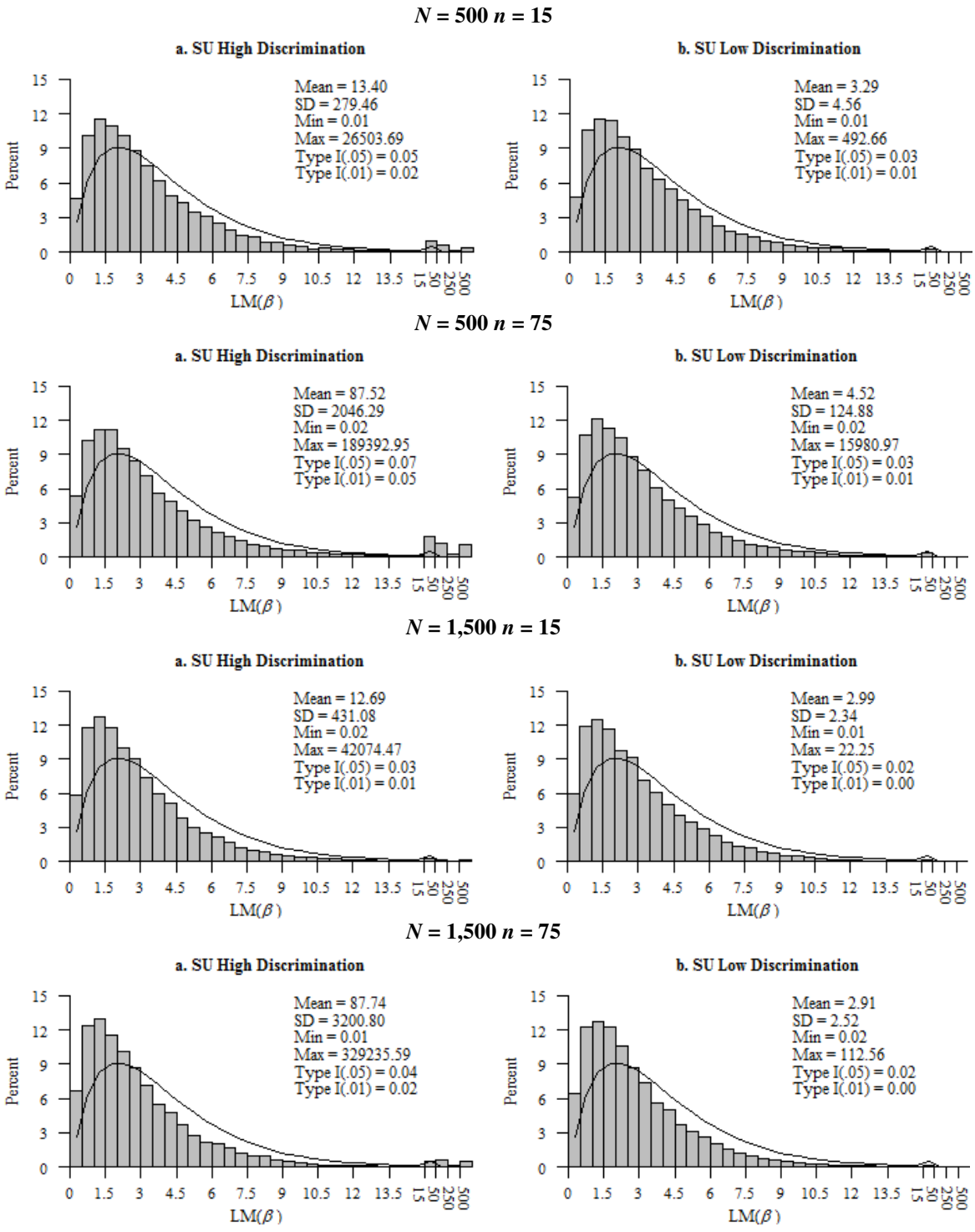


Figure G-10. KS Test π_R in SU Study Conditions for LM($\alpha\beta$)

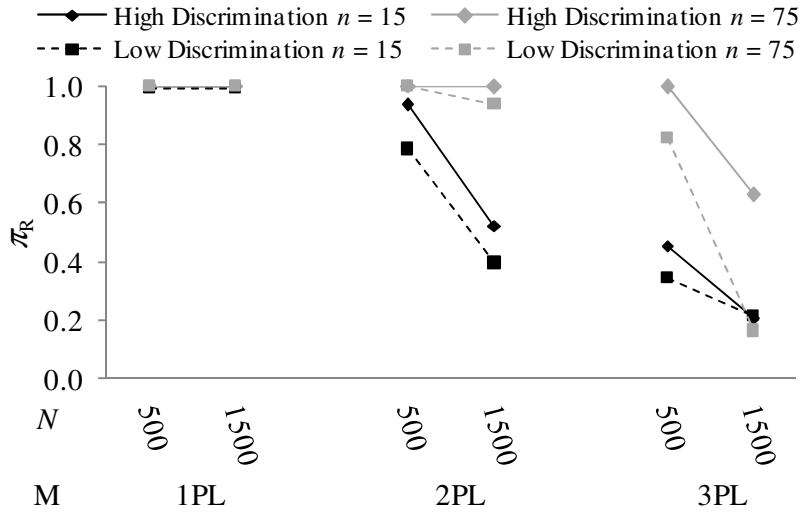


Figure G-11. KS Test π_R in SU Study Conditions for LM(α)

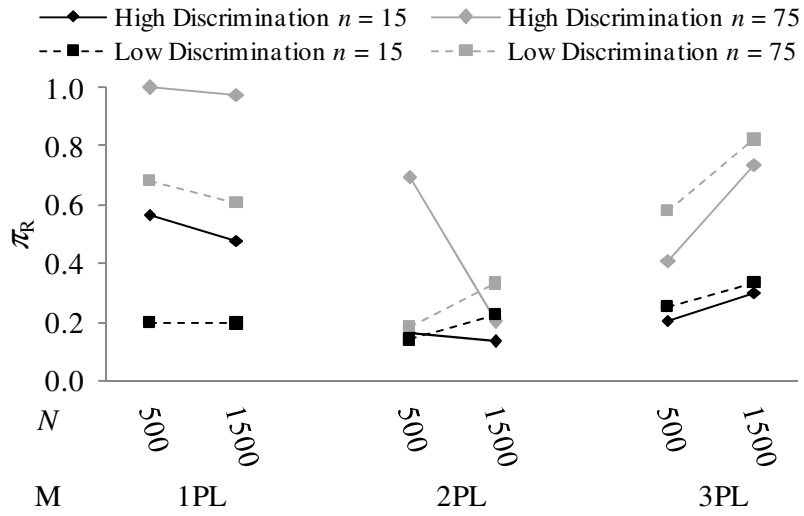


Figure G-12. KS Test π_R in SU Study Conditions for LM(β)

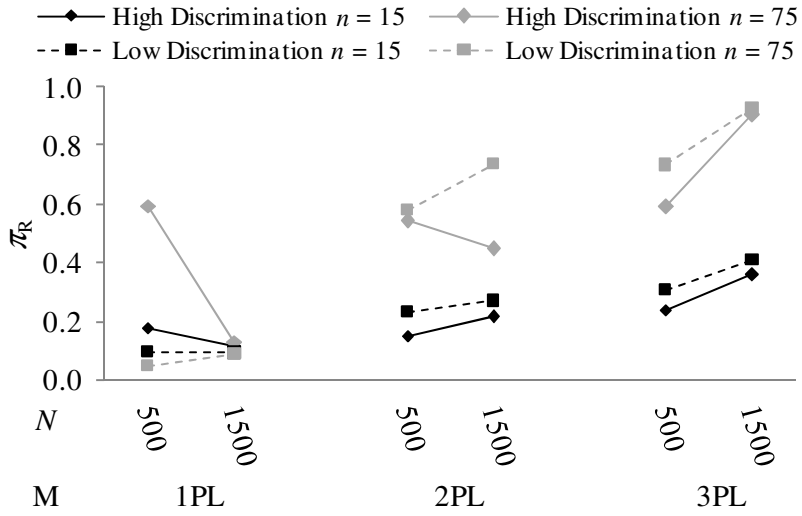


Figure G-13. Scatterplots Between *b* and LM Statistics for the 1PL When $N = 500$ $n = 15$

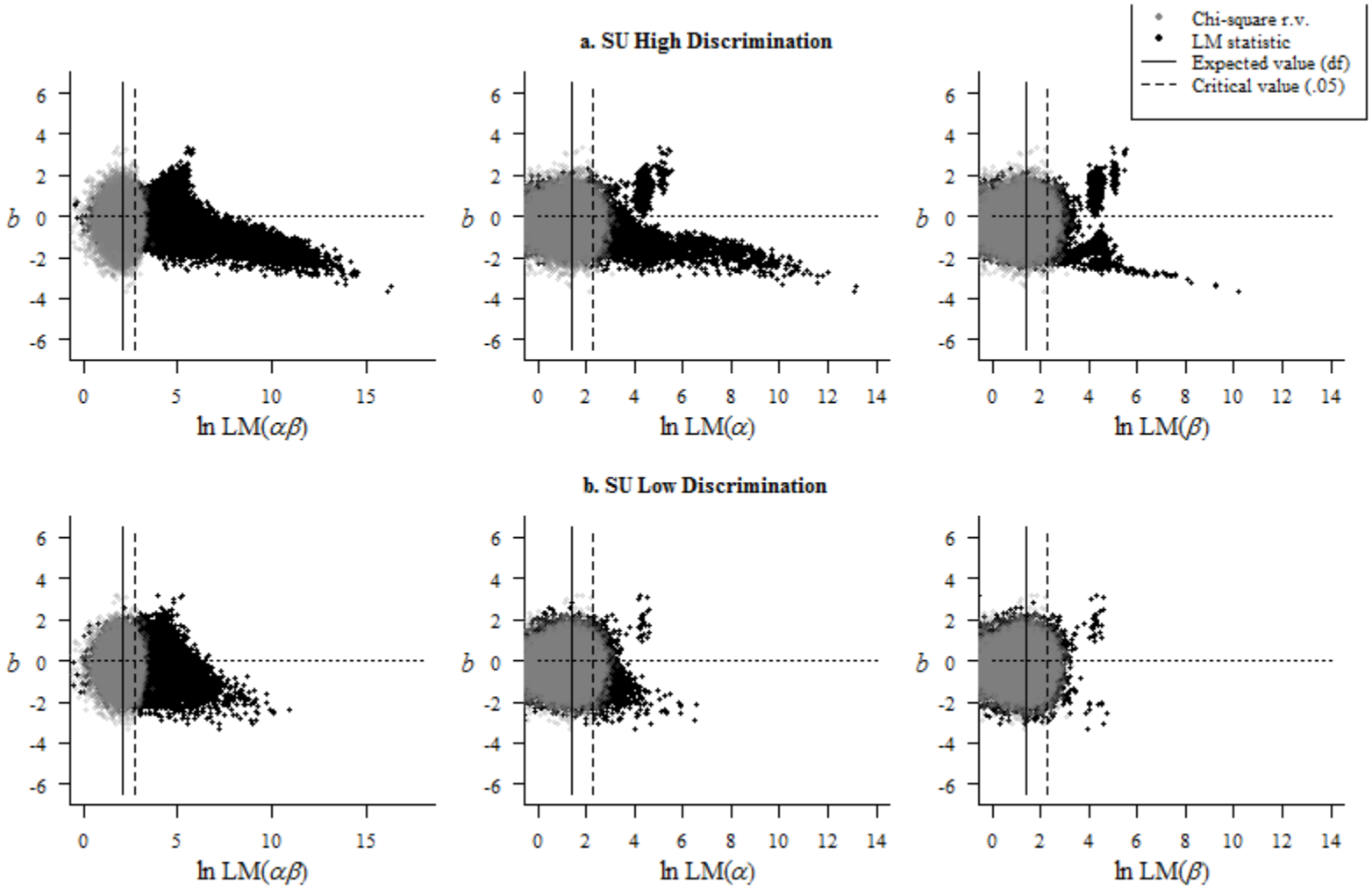


Figure G-14. Scatterplots Between *b* and LM Statistics for the 1PL When $N = 500$ $n = 75$

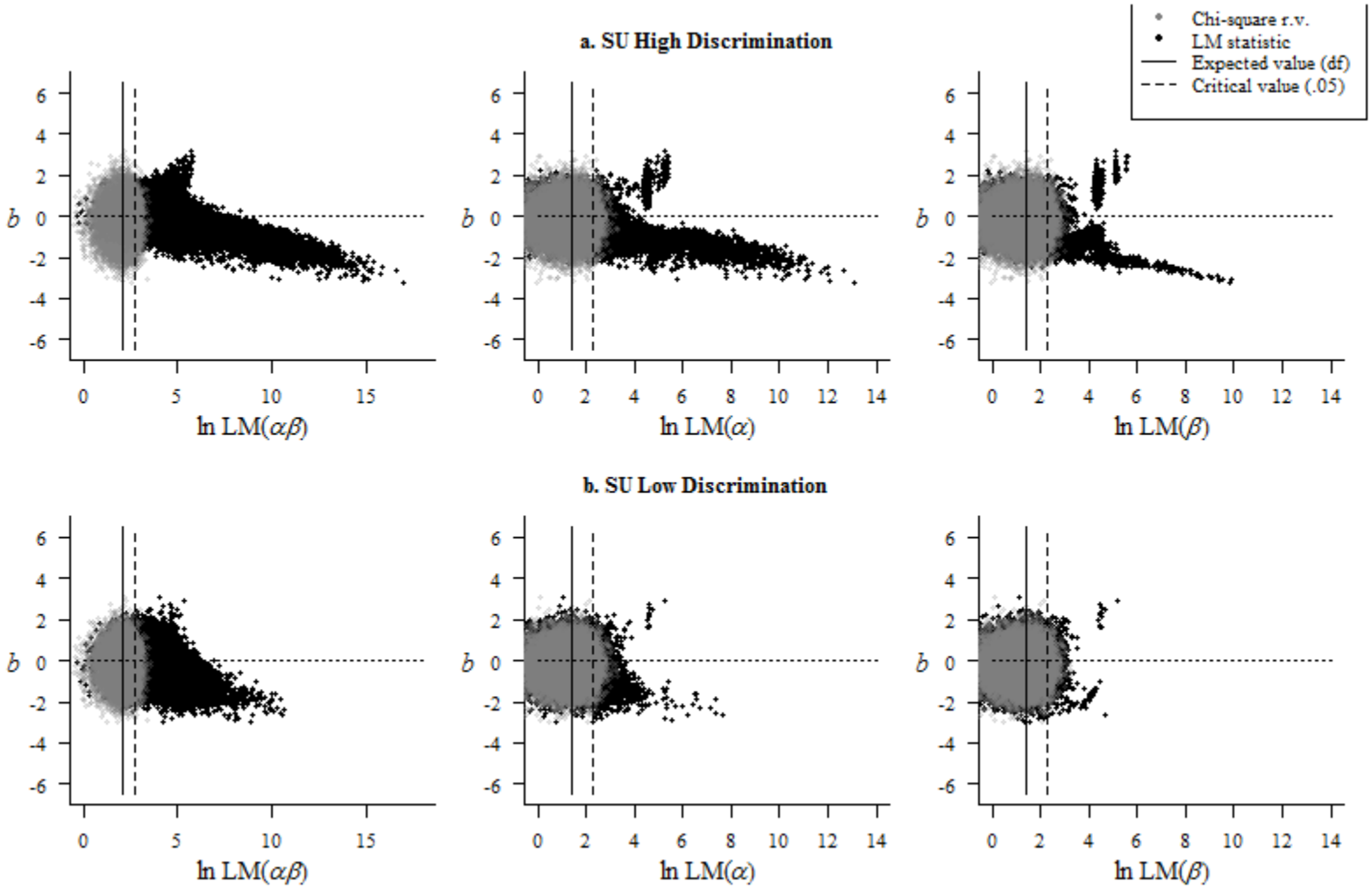


Figure G-15. Scatterplots Between *b* and LM Statistics for the 1PL When N = 1,500 n = 15

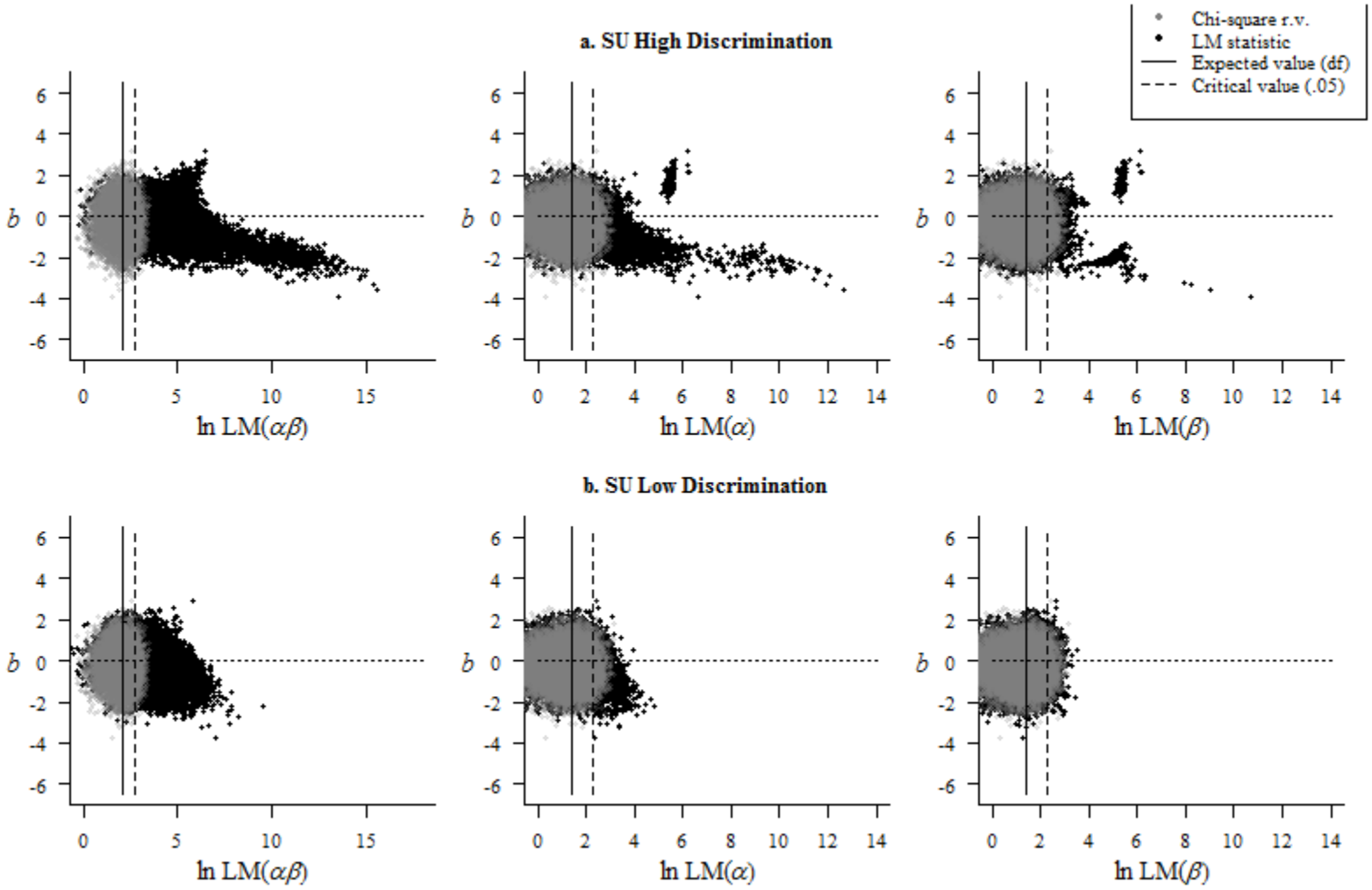


Figure G-16. Scatterplots Between *b* and LM Statistics for the 1PL When N = 1,500 n = 75

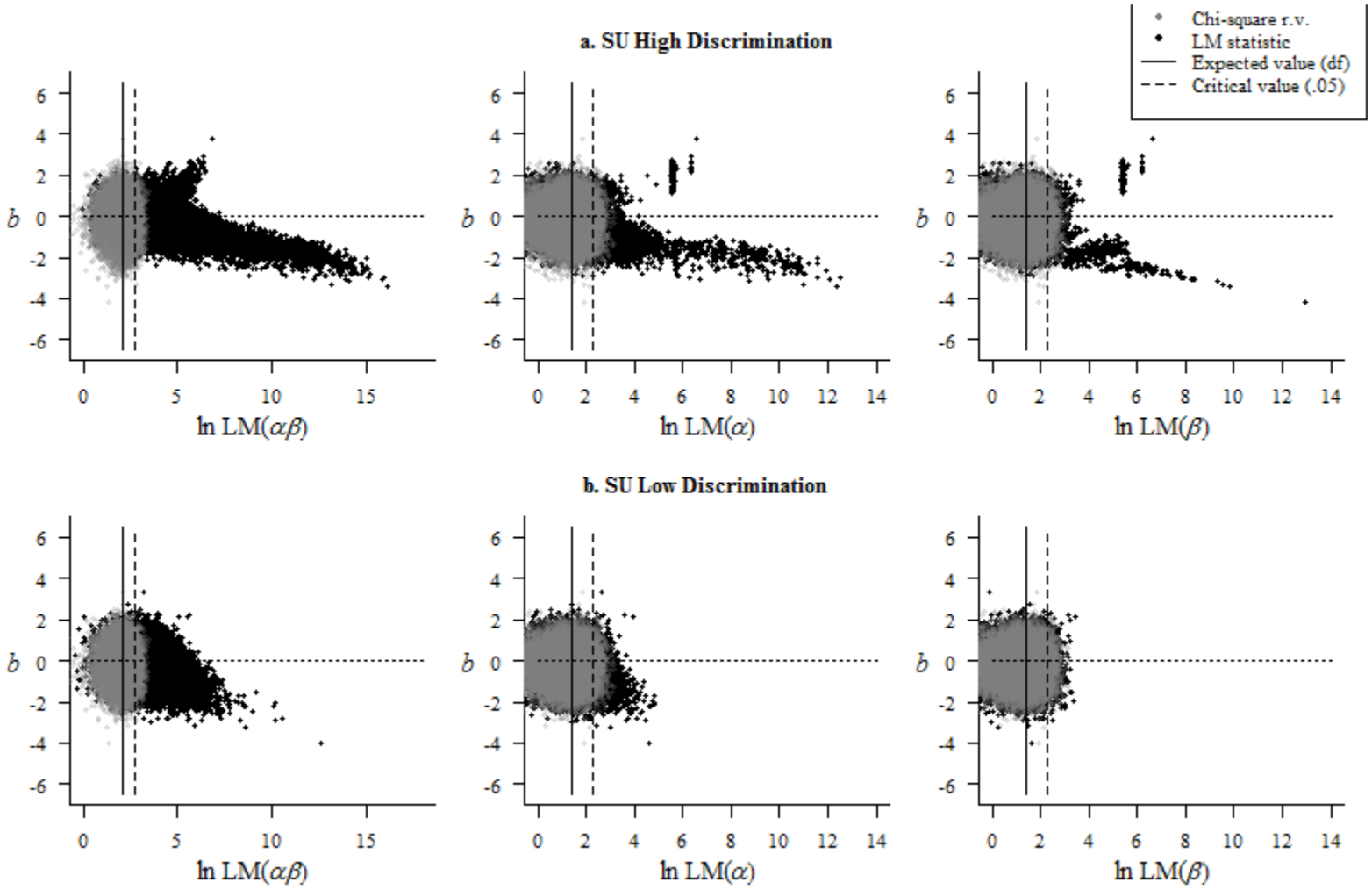


Figure G-17. Scatterplots Between *b* and LM Statistics for the 2PL When $N = 500$ $n = 15$

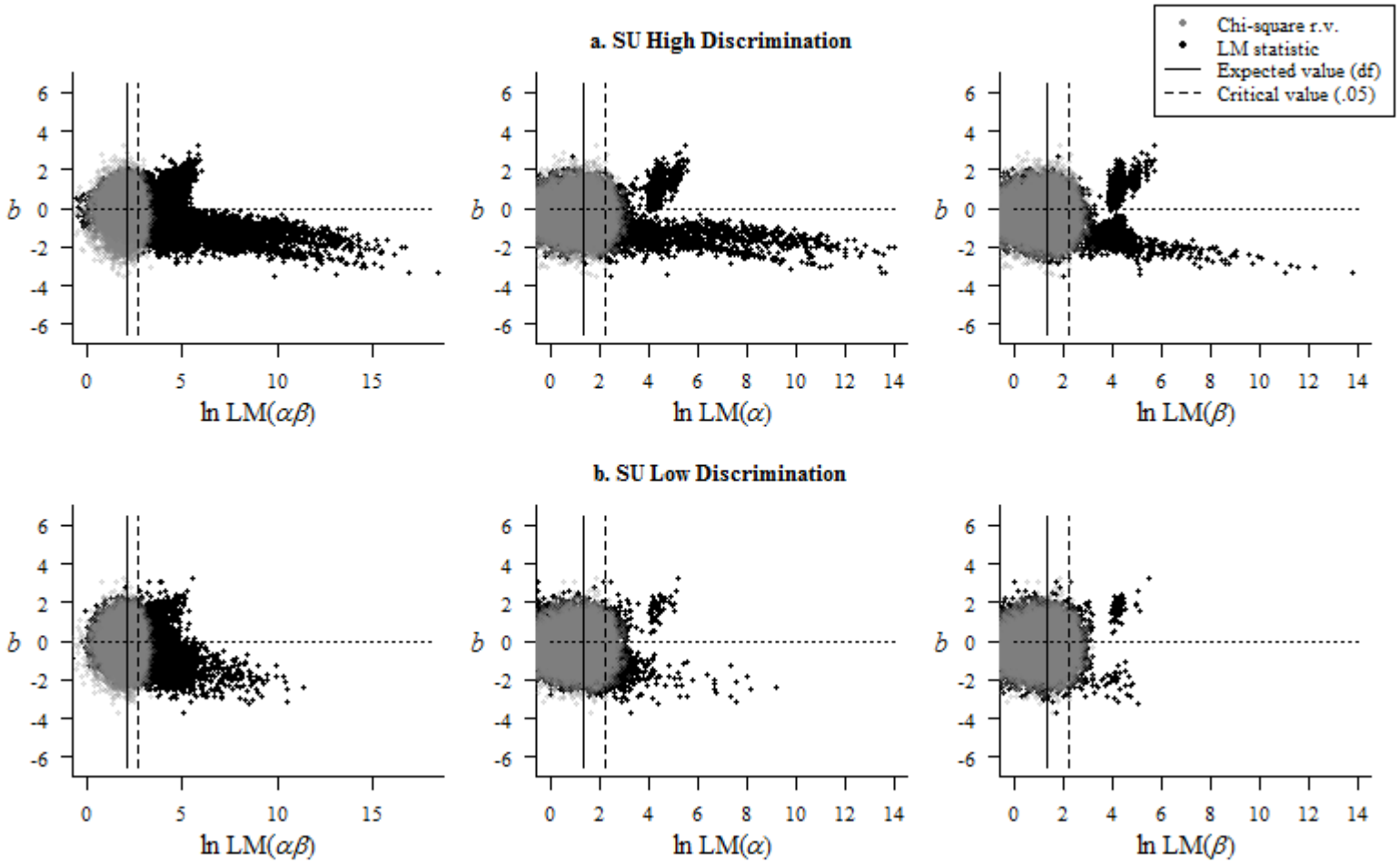


Figure G-18. Scatterplots Between *b* and LM Statistics for the 2PL When $N = 500$ $n = 75$

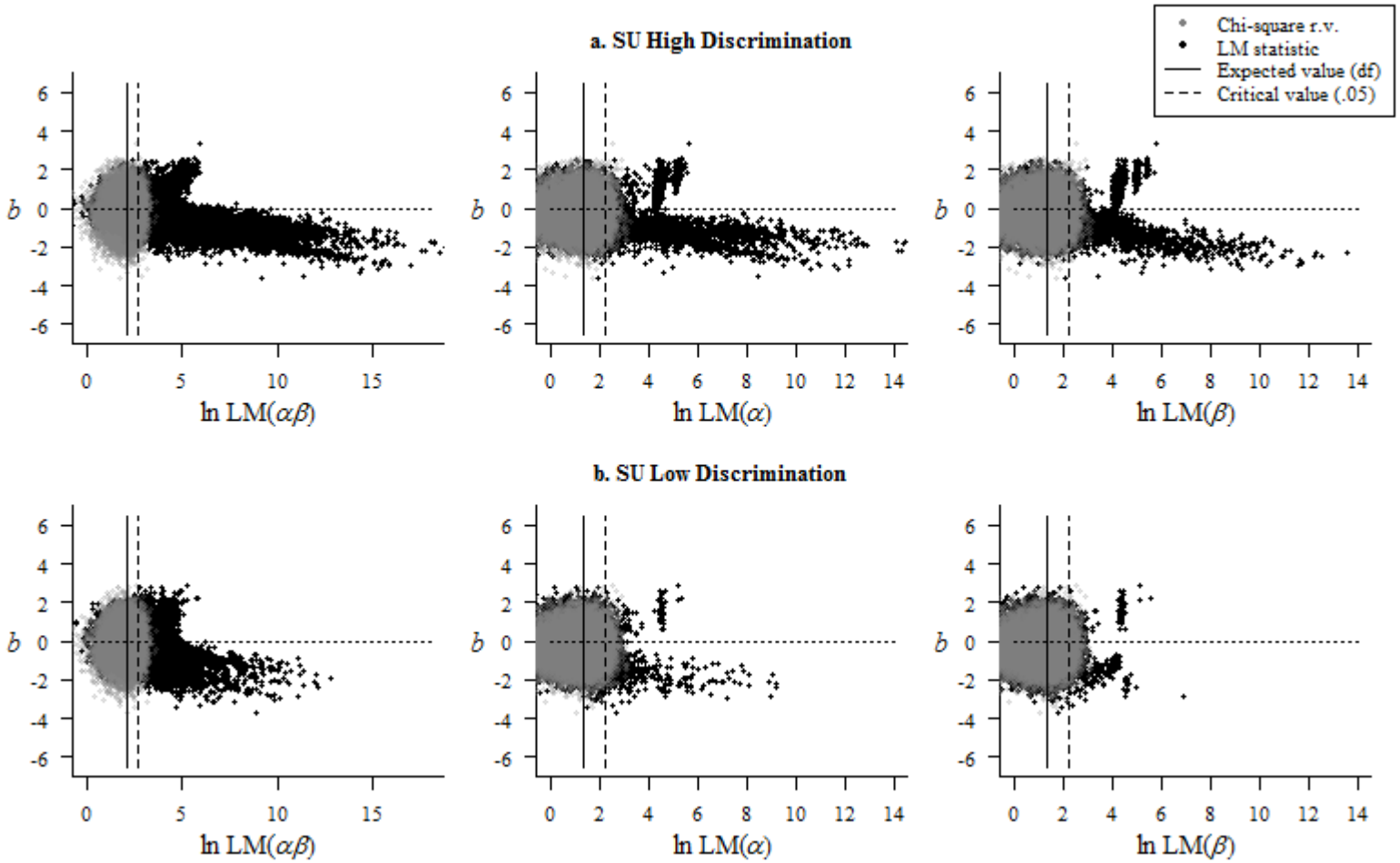


Figure G-19. Scatterplots Between *b* and LM Statistics for the 2PL When N = 1,500 n = 15

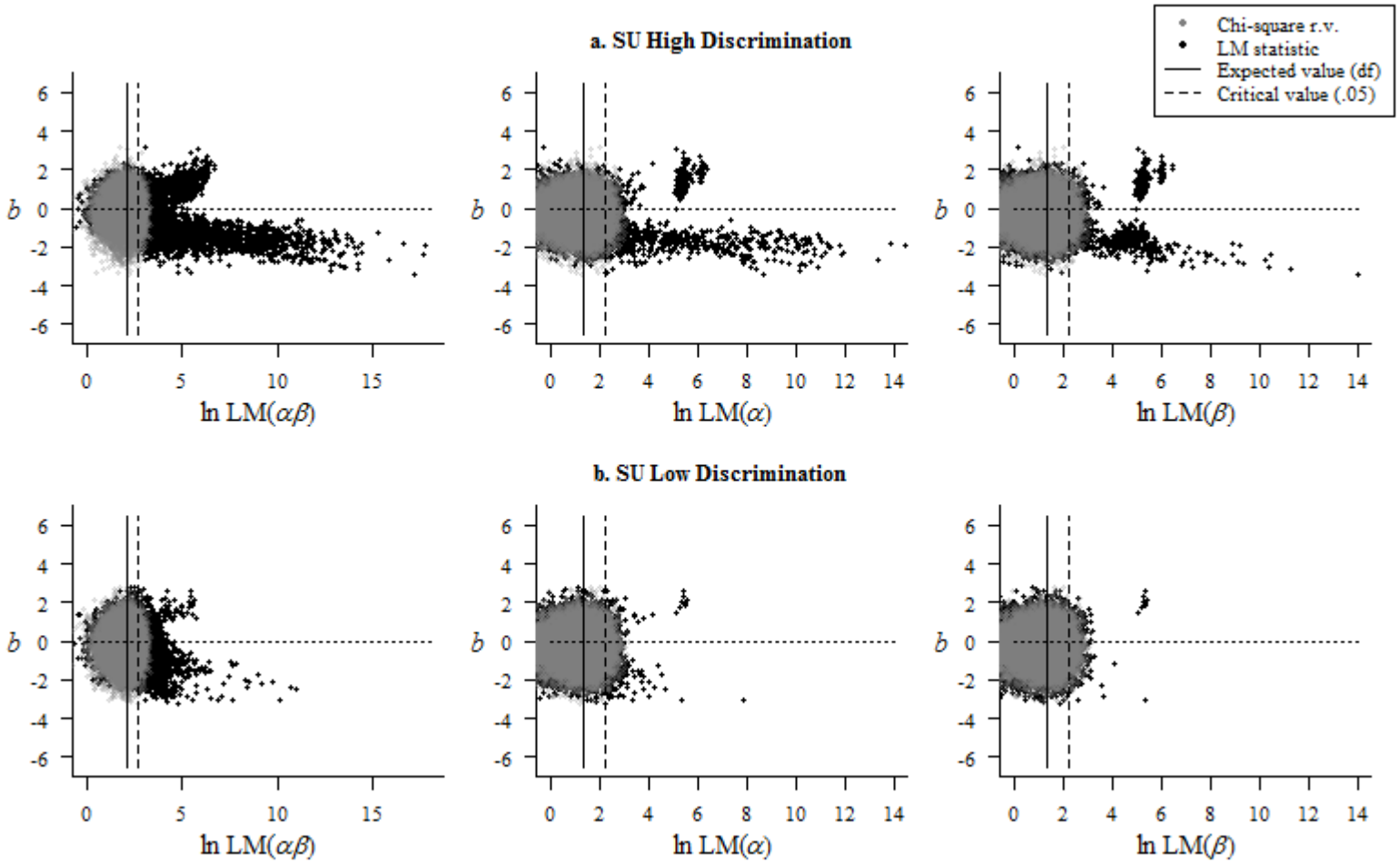


Figure G-20. Scatterplots Between *b* and LM Statistics for the 2PL When N = 1,500 n = 75

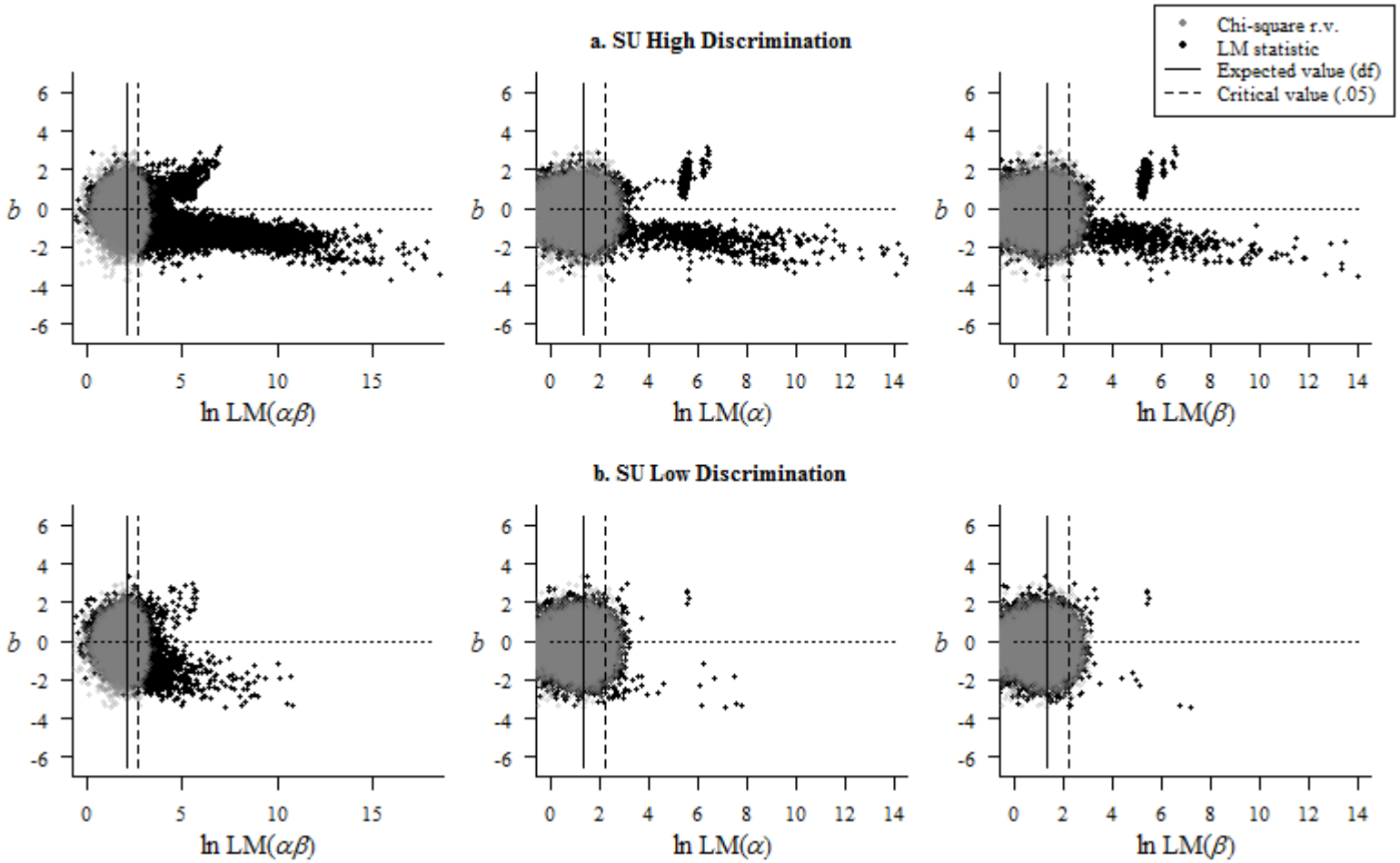


Figure G-21. Scatterplots Between *a* and LM Statistics for the 2PL When $N = 500$ $n = 15$

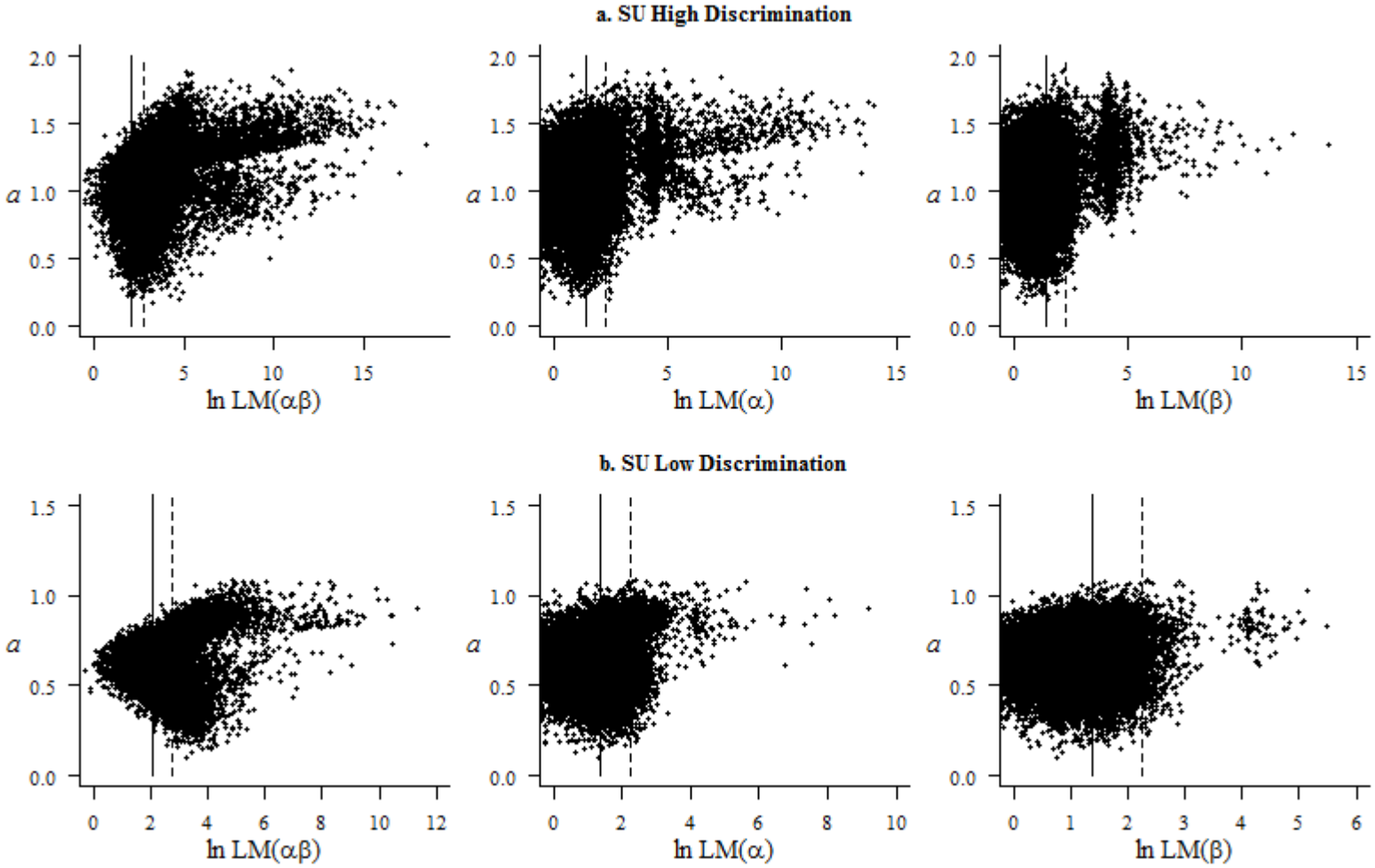


Figure G-22. Scatterplots Between *a* and LM Statistics for the 2PL When $N = 500$ $n = 75$

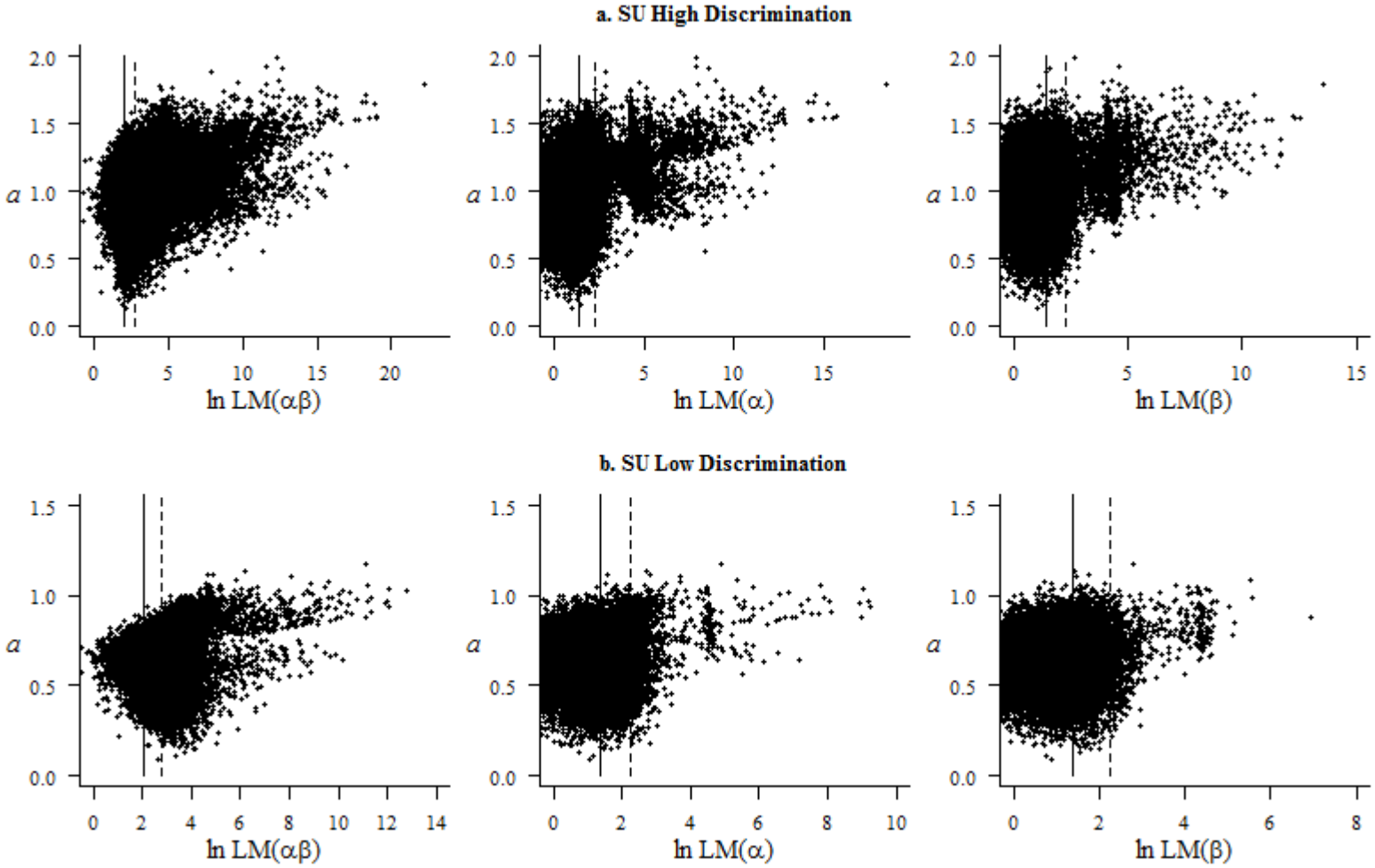


Figure G-23. Scatterplots Between *a* and LM Statistics for the 2PL When N = 1,500 n = 15

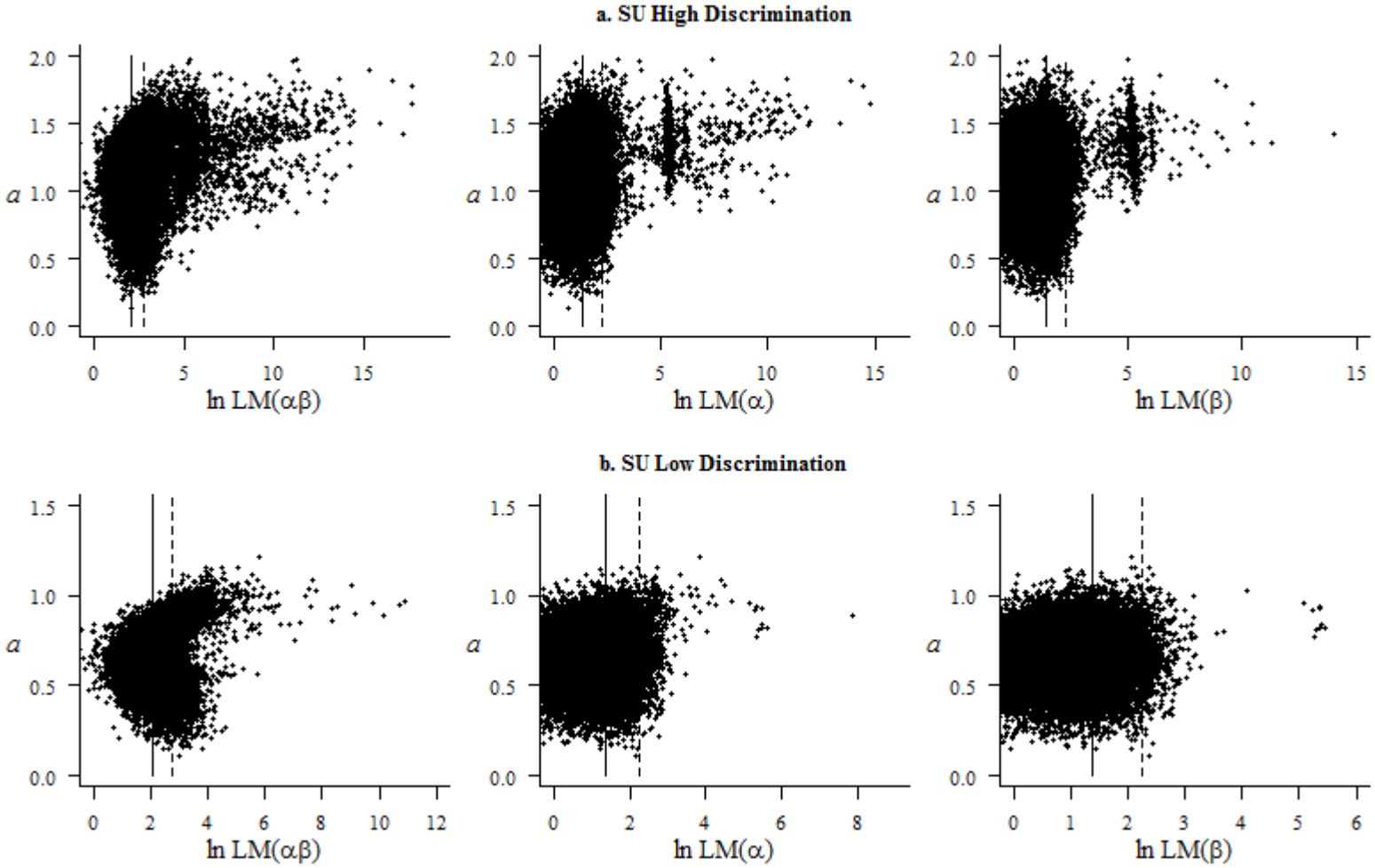


Figure G-24. Scatterplots Between *a* and LM Statistics for the 2PL When N = 1,500 n = 75

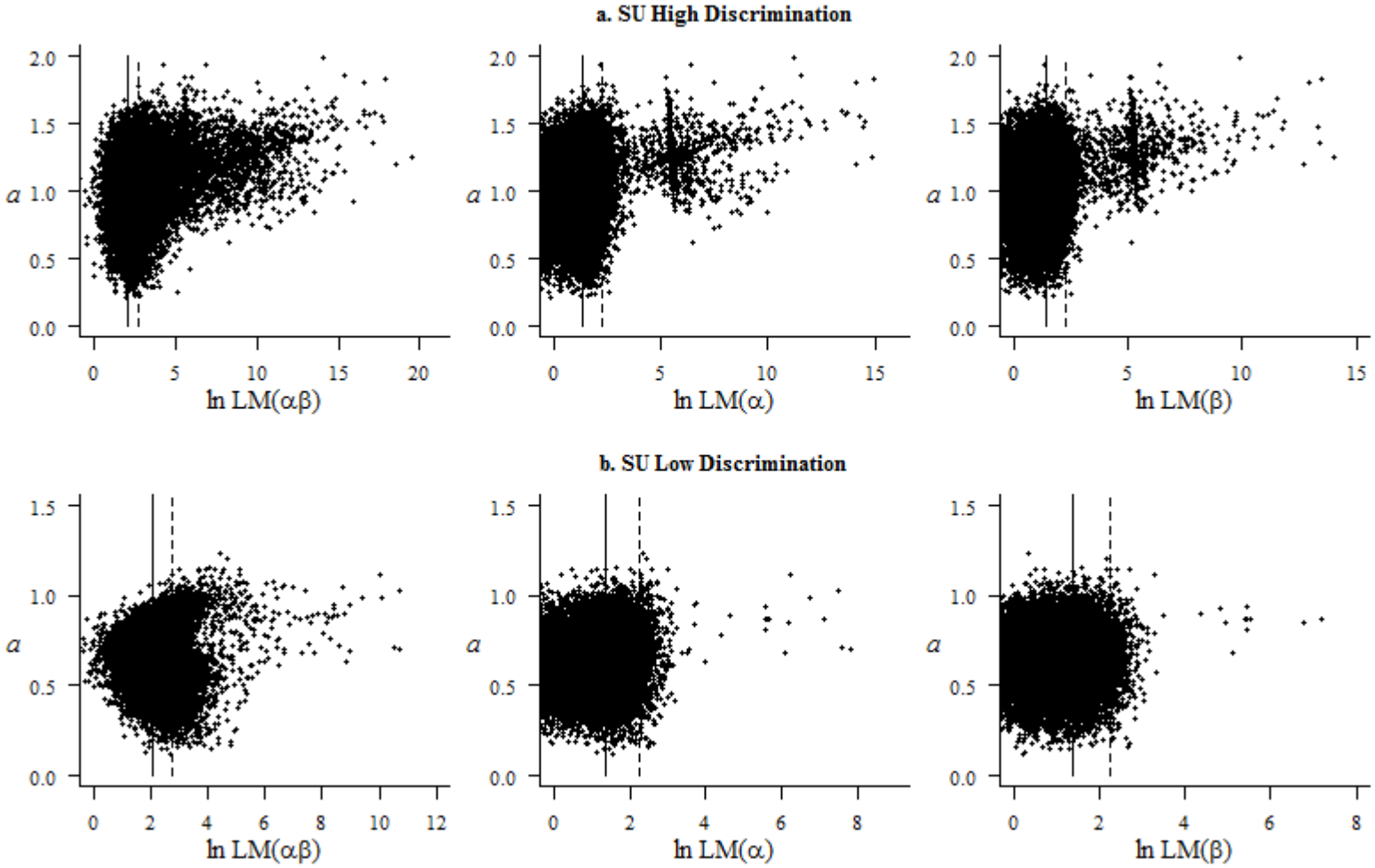


Figure G-25. Scatterplots Between *b* and LM Statistics for the 3PL When $N = 500$ $n = 15$

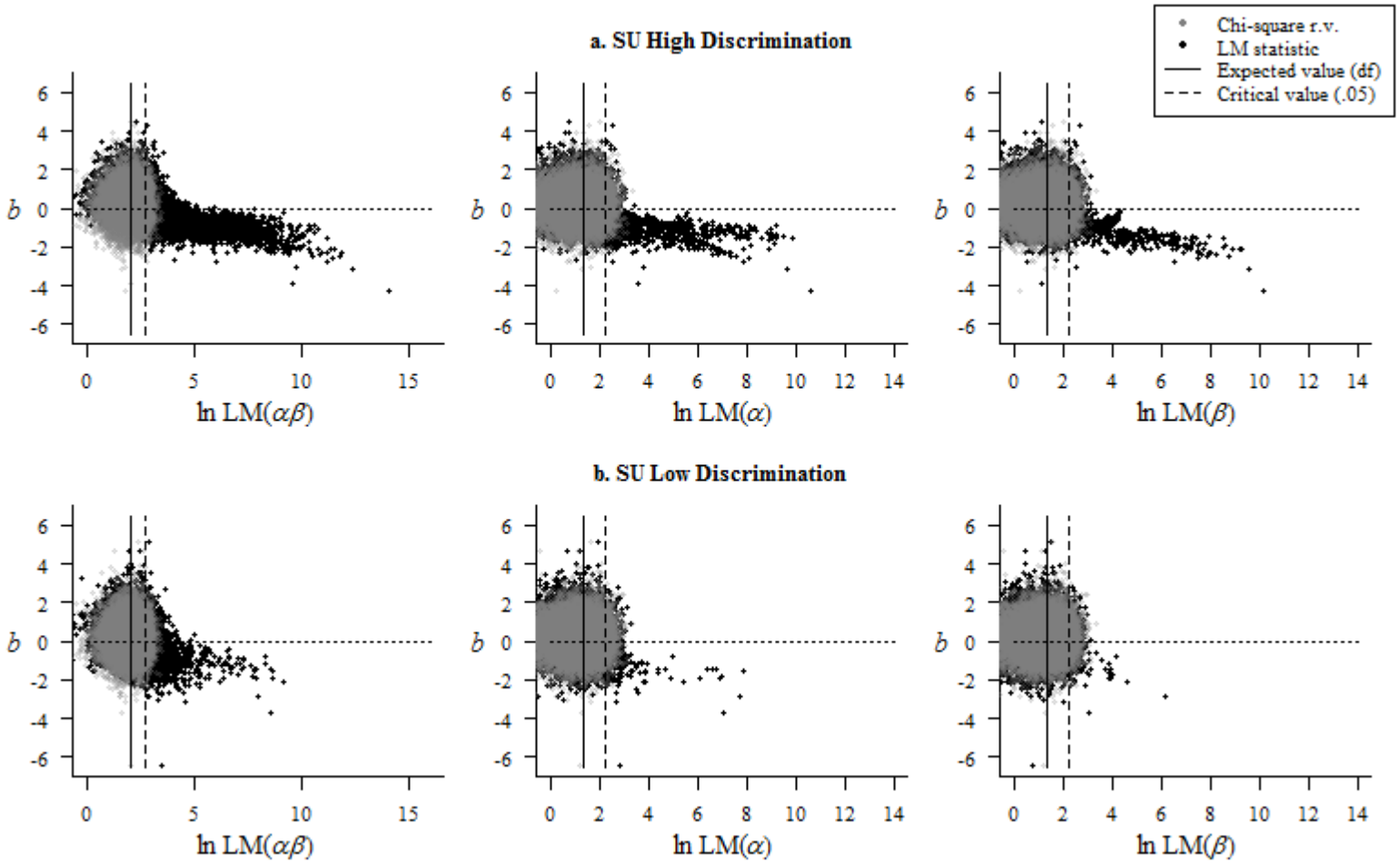


Figure G-26. Scatterplots Between *b* and LM Statistics for the 3PL When $N = 500$ $n = 75$

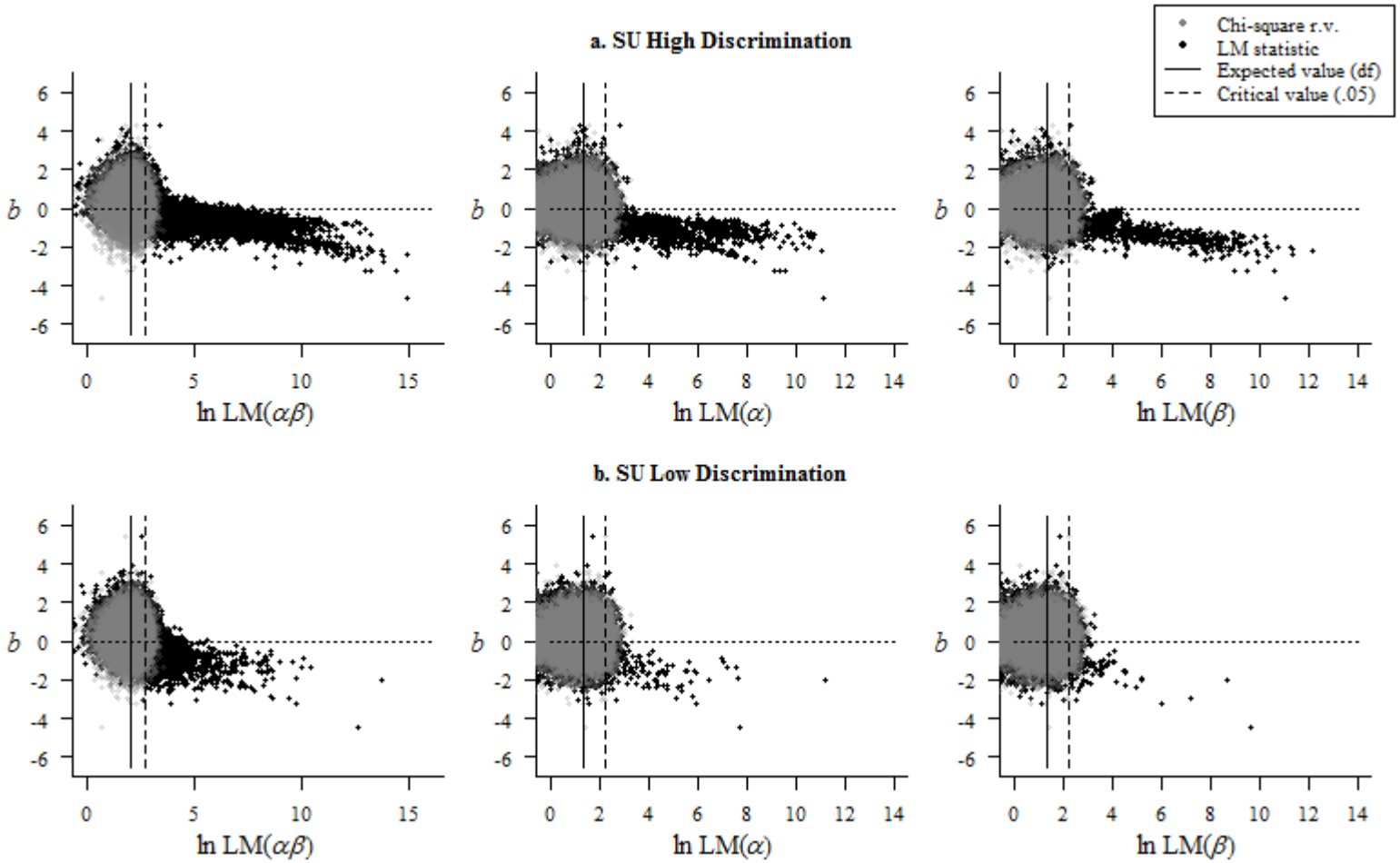


Figure G-27. Scatterplots Between *b* and LM Statistics for the 3PL When N = 1,500 n = 15

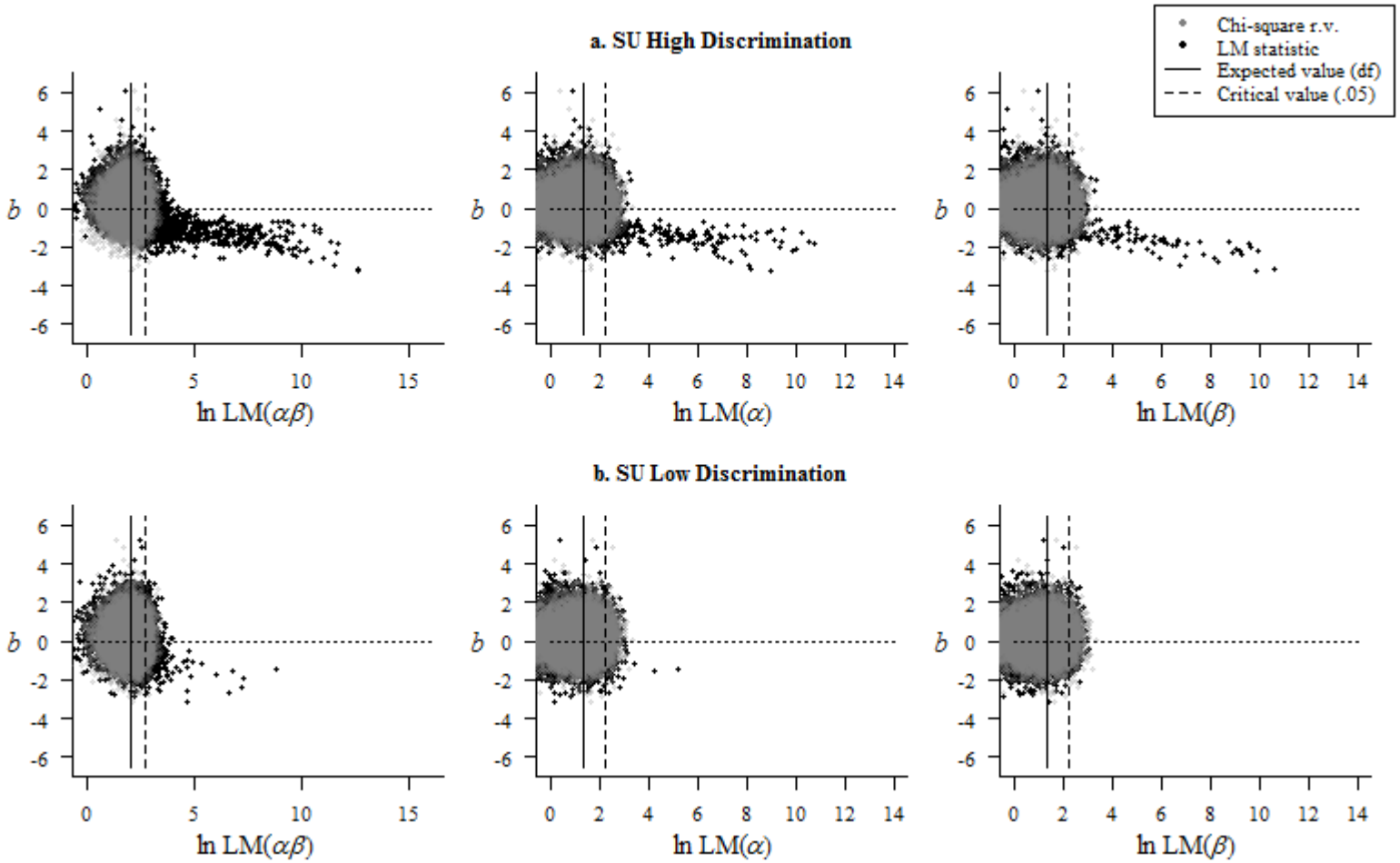


Figure G-28. Scatterplots Between *b* and LM Statistics for the 3PL When N = 1,500 n = 75

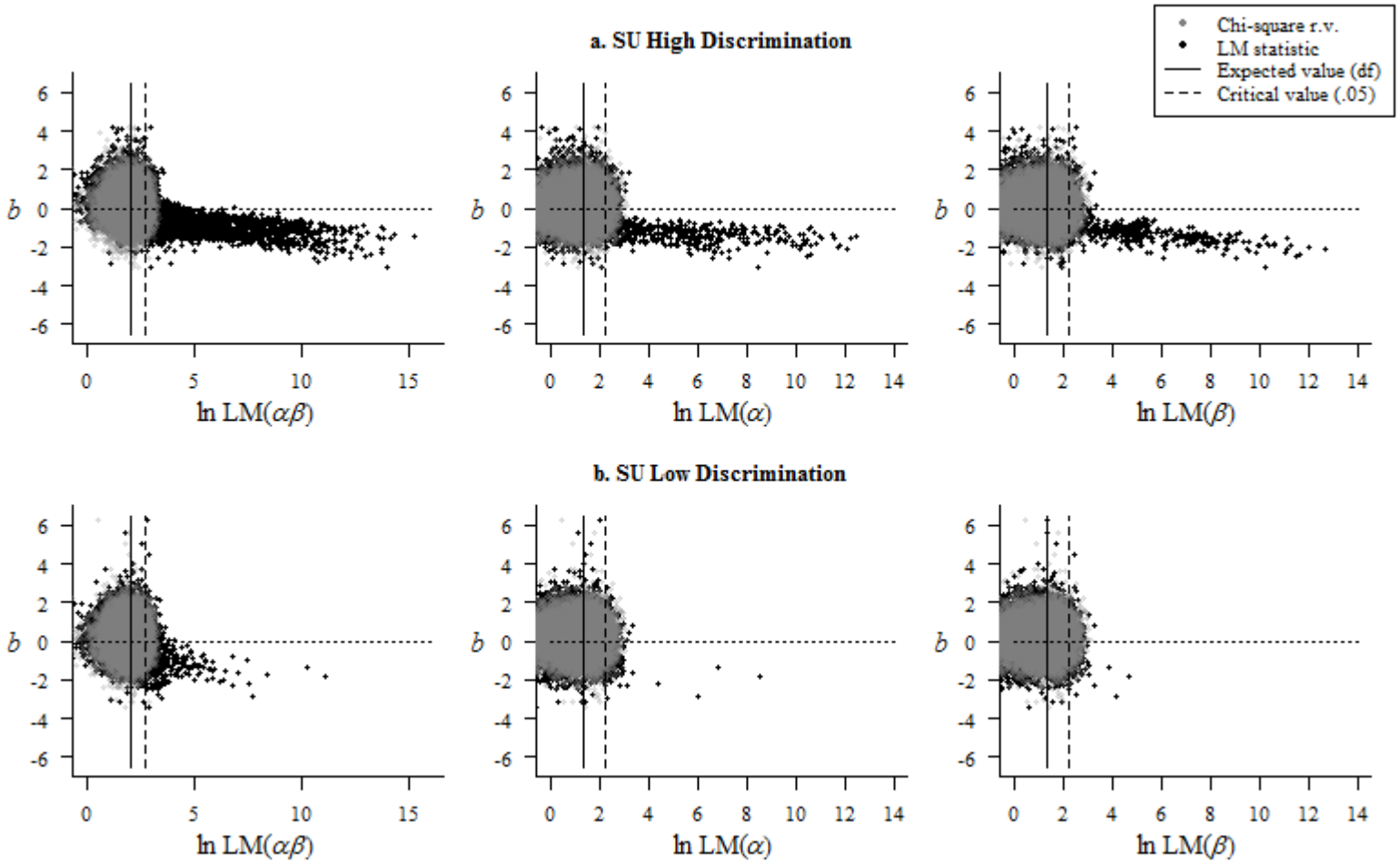


Figure G-29. Scatterplots Between *a* and LM Statistics for the 3PL When $N = 500$ $n = 15$

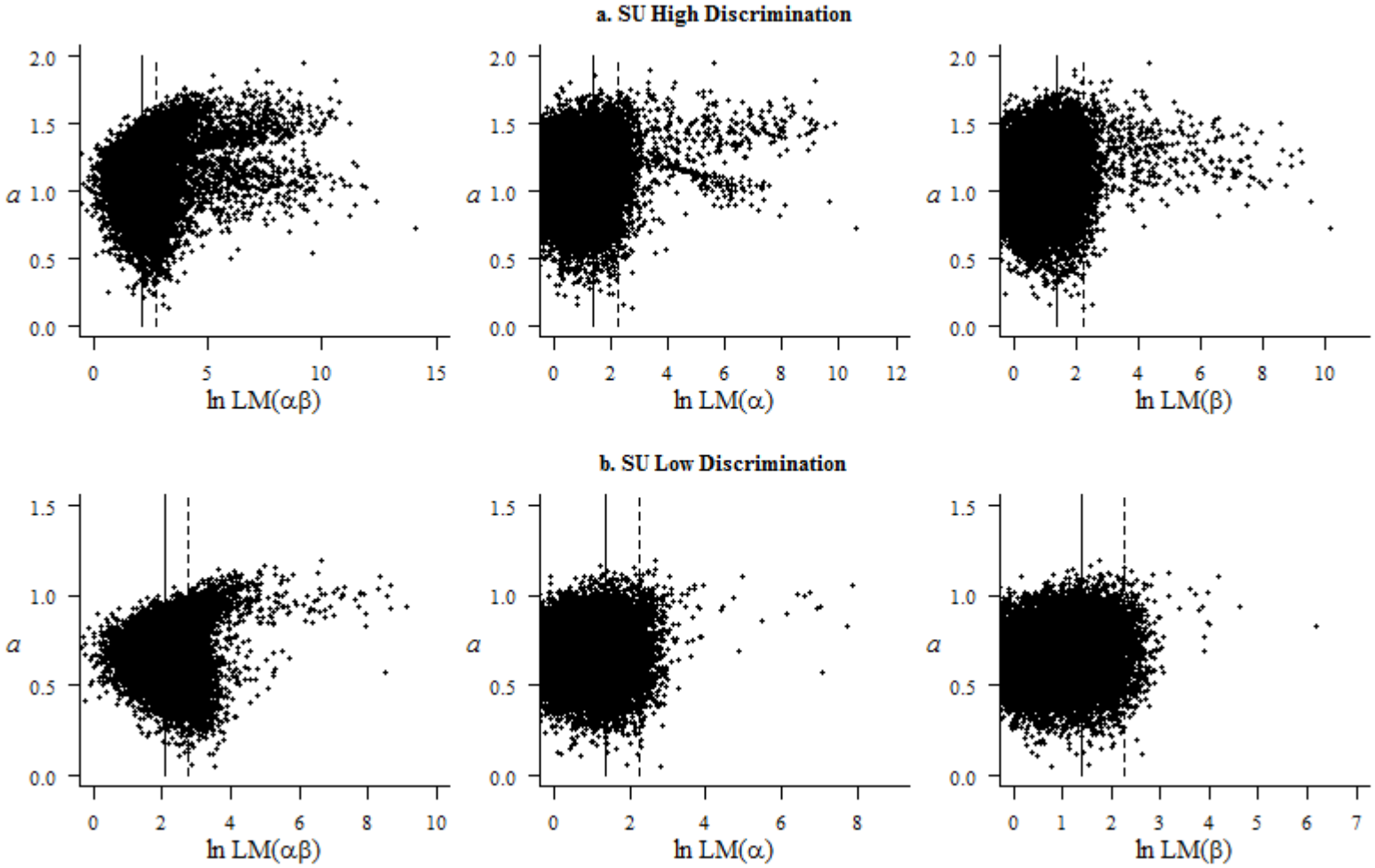


Figure G-30. Scatterplots Between *a* and LM Statistics for the 3PL When $N = 500$ $n = 75$

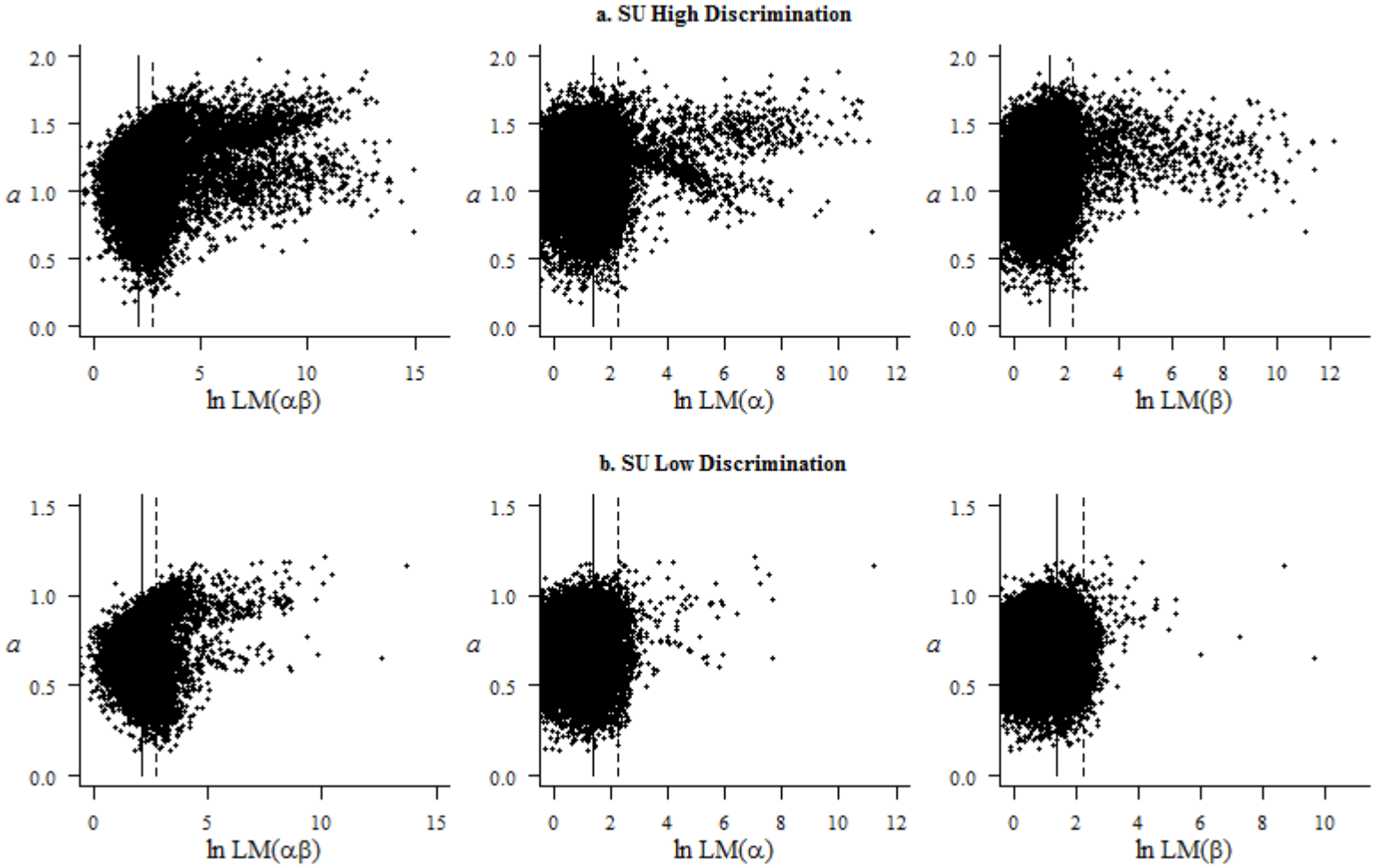


Figure G-31. Scatterplots Between *a* and LM Statistics for the 3PL When N = 1,500 n = 15

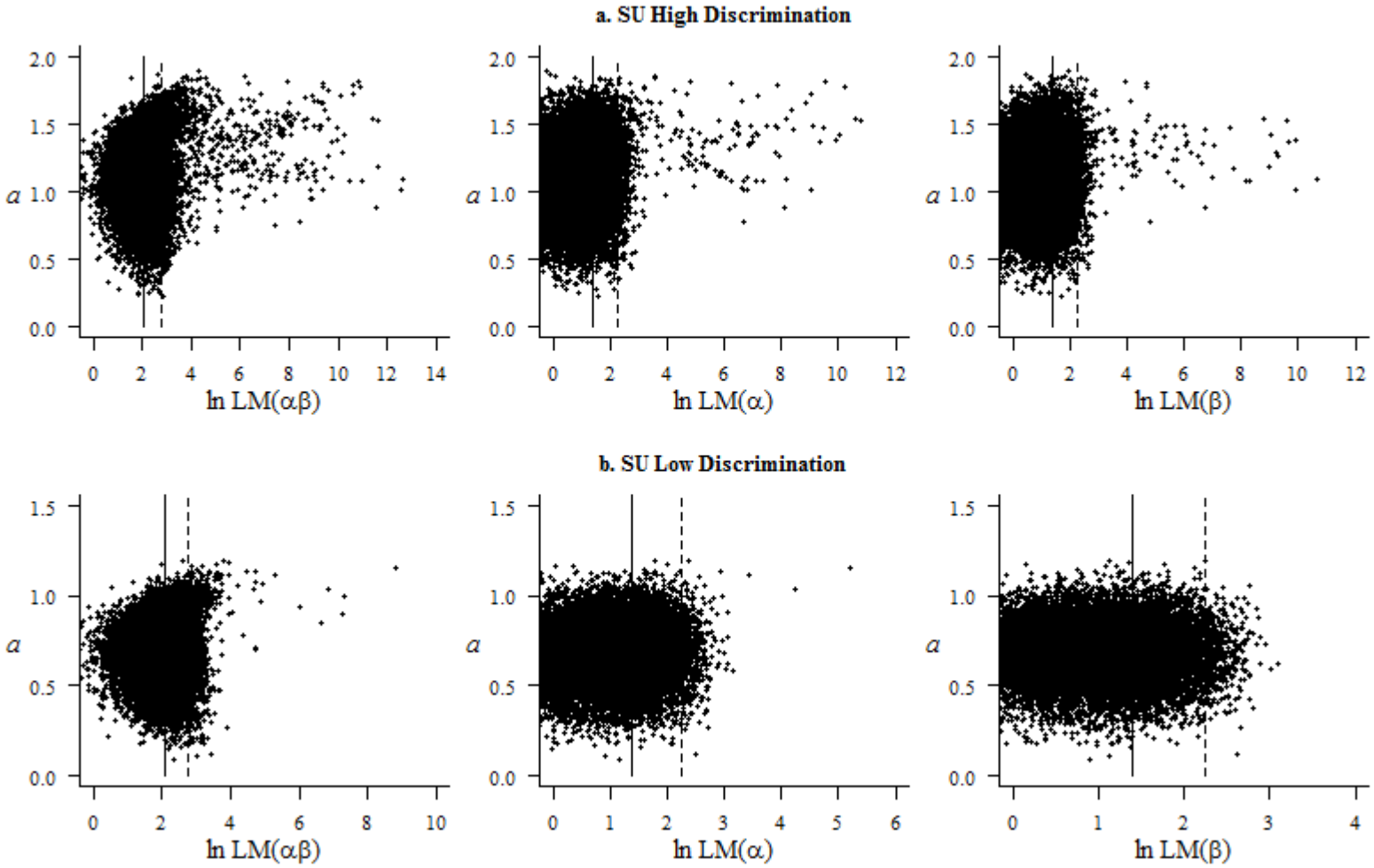


Figure G-32. Scatterplots Between *a* and LM Statistics for the 3PL When N = 1,500 n = 75

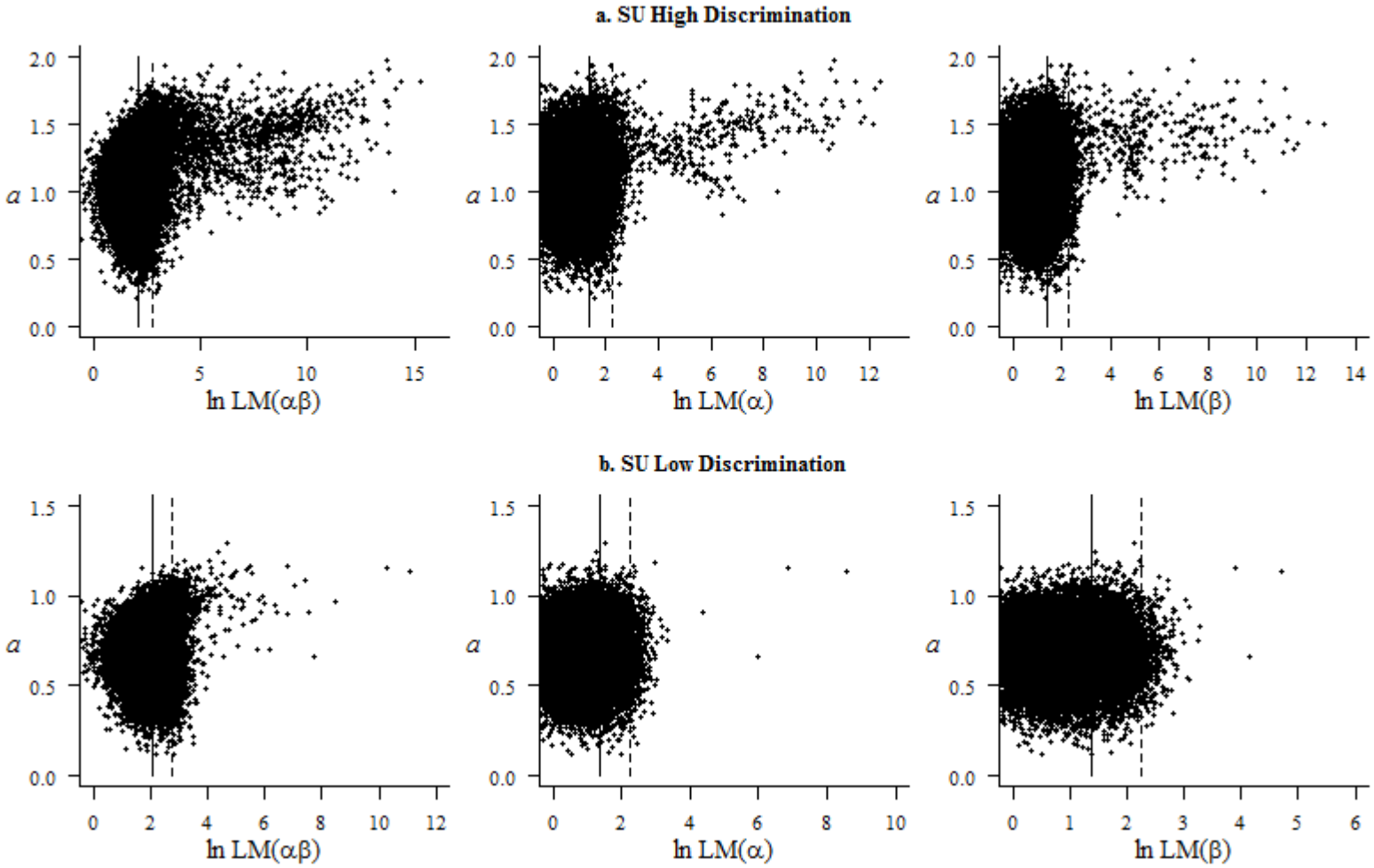


Table G-1. Descriptive Statistics for Corrected LM($\alpha\beta$) in SU High Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		8.00	7.33	4.00	0.26	37.33	1.00	1.50	--	0.05
1PL										
500	15	7.73E+11	48.81	7.68E+13	0.64	9.89E+15	118.41	14826.59	0.78	1.00
	75	70.69	38.91	104.17	0.36	3993.08	9.22	200.22	0.73	1.00
1,500	15	86.70	35.87	141.10	0.83	2827.02	5.12	47.18	0.73	1.00
	75	56.39	23.71	86.64	0.86	1302.76	3.59	19.46	0.62	1.00
2PL										
500	15	1.98E+11	13.33	1.78E+13	0.57	2.27E+15	115.57	14388.07	0.49	0.87
	75	40.54	14.06	70.23	0.30	3386.86	13.95	478.03	0.45	1.00
1,500	15	6.96E+10	9.51	7.75E+12	0.52	1.04E+15	129.53	17246.94	0.33	0.42
	75	37.52	10.28	91.32	0.36	1188.42	4.60	27.74	0.30	0.99
3PL										
500	15	20.44	10.07	31.47	0.39	572.07	4.00	23.29	0.36	0.53
	75	25.58	10.27	51.37	0.54	2191.04	11.76	331.57	0.31	1.00
1,500	15	16.46	8.03	42.04	0.36	915.94	7.59	78.22	0.27	0.21
	75	20.38	8.29	55.33	0.33	935.02	6.40	52.58	0.17	0.60

Table G-2. Descriptive Statistics for Corrected LM($\alpha\beta$) in SU Low Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		8.00	7.33	4.00	0.26	37.33	1.00	1.50	--	0.05
1PL										
500	15	48.17	23.29	73.36	0.28	1560.44	5.23	48.50	0.65	0.99
	75	35.75	20.20	45.64	0.86	1031.17	4.82	50.05	0.58	1.00
1,500	15	42.77	21.72	60.95	0.53	1118.82	4.72	40.26	0.64	0.99
	75	31.82	18.27	38.03	0.42	494.63	3.33	16.77	0.55	1.00
2PL										
500	15	18.00	11.79	28.68	0.73	2131.86	26.83	1659.78	0.43	0.74
	75	15.56	10.60	19.20	0.48	461.50	5.53	51.50	0.31	1.00
1,500	15	11.50	9.20	12.45	0.47	357.13	11.94	233.28	0.31	0.36
	75	10.87	8.81	15.37	0.49	609.27	15.86	350.44	0.19	0.72
3PL										
500	15	12.19	9.34	12.30	0.46	298.63	5.75	56.70	0.32	0.39
	75	12.11	9.00	14.24	0.48	448.85	7.81	115.40	0.21	0.84
1,500	15	9.07	7.83	8.26	0.44	441.03	18.44	671.03	0.27	0.21
	75	8.97	7.62	9.88	0.26	310.74	17.39	430.70	0.11	0.15

Table G-3. Descriptive Statistics for Corrected LM(α) in SU High Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01	28.47	1.41	3.00	--	0.05
1PL										
500	15	14.18	4.14	31.65	0.01	280.06	3.67	14.51	0.28	0.25
	75	17.54	4.17	35.60	0.00	277.81	2.89	8.70	0.22	0.89
1,500	15	9.76	3.79	38.63	0.03	813.80	9.01	92.04	0.25	0.14
	75	13.04	3.70	50.55	0.02	690.22	6.64	49.81	0.14	0.30
2PL										
500	15	14.94	3.23	34.54	0.02	325.69	3.31	11.49	0.26	0.16
	75	19.02	3.47	39.28	0.02	294.96	2.73	7.61	0.18	0.73
1,500	15	13.98	2.90	56.93	0.02	726.15	5.93	37.49	0.24	0.14
	75	17.97	2.99	66.33	0.02	765.08	5.26	31.42	0.13	0.23
3PL										
500	15	8.49	2.92	20.03	0.01	239.12	4.05	17.67	0.25	0.17
	75	9.70	2.85	25.20	0.01	288.91	4.44	22.54	0.15	0.36
1,500	15	7.18	2.57	30.50	0.01	537.08	8.20	74.73	0.28	0.26
	75	8.81	2.51	39.32	0.01	626.96	7.81	71.40	0.19	0.73

Table G-4. Descriptive Statistics for Corrected LM(α) in SU Low Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01	28.47	1.41	3.00	--	0.05
1PL										
500	15	4.68	3.49	5.45	0.01	140.71	8.98	147.30	0.25	0.15
	75	4.92	3.60	6.63	0.01	202.96	11.40	199.95	0.12	0.18
1,500	15	4.35	3.42	3.61	0.02	52.68	2.33	11.09	0.24	0.13
	75	4.47	3.56	3.63	0.01	81.87	2.53	19.01	0.12	0.15
2PL										
500	15	4.22	2.95	7.59	0.01	182.57	12.35	194.50	0.25	0.16
	75	4.84	2.96	11.49	0.00	212.67	9.54	108.51	0.13	0.28
1,500	15	3.58	2.73	6.38	0.01	338.64	32.10	1319.18	0.27	0.23
	75	3.76	2.82	9.79	0.02	593.39	32.95	1322.52	0.14	0.38
3PL										
500	15	3.88	2.76	6.11	0.01	139.33	10.09	133.53	0.27	0.24
	75	3.98	2.70	8.35	0.02	307.49	13.42	255.80	0.16	0.52
1,500	15	3.27	2.59	4.09	0.01	262.60	33.76	1842.65	0.29	0.29
	75	3.19	2.49	5.42	0.03	295.10	39.47	1958.13	0.20	0.79

Table G-5. Descriptive Statistics for Corrected LM(β) in SU High Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01	28.47	1.41	3.00	--	0.05
1PL										
500	15	12.73	4.01	28.95	0.02	303.90	4.07	19.78	0.26	0.19
	75	15.59	3.98	32.42	0.02	356.32	3.25	12.71	0.18	0.74
1,500	15	8.88	3.88	34.75	0.02	905.29	9.84	122.09	0.24	0.13
	75	11.65	3.58	45.23	0.03	808.48	7.23	64.43	0.12	0.17
2PL										
500	15	13.51	2.99	32.96	0.01	361.99	3.98	19.44	0.26	0.15
	75	16.82	3.10	36.10	0.01	338.90	3.22	12.58	0.18	0.71
1,500	15	12.44	2.68	51.82	0.00	987.36	6.83	57.67	0.26	0.20
	75	15.92	2.80	60.52	0.01	1009.57	6.28	52.16	0.15	0.40
3PL										
500	15	7.80	2.74	19.15	0.01	330.49	4.97	32.11	0.26	0.19
	75	8.53	2.67	22.35	0.03	361.02	5.19	35.38	0.16	0.58
1,500	15	6.51	2.42	28.16	0.02	729.68	10.03	134.39	0.30	0.33
	75	7.64	2.29	33.64	0.02	710.46	8.32	86.18	0.23	0.92

Table G-6. Descriptive Statistics for Corrected LM(β) in SU Low Discrimination Conditions

N	*n*	Mean	Med	SD	Min	Max	Skew	Kurt	KS Test $\overline{D}$	KS Test π_R
Expected		4.00	3.33	2.83	0.01	28.47	1.41	3.00	--	0.05
1PL										
500	15	4.32	3.42	4.81	0.03	133.16	11.01	207.74	0.23	0.10
	75	4.41	3.39	5.86	0.02	193.69	12.78	245.35	0.10	0.06
1,500	15	4.08	3.40	2.95	0.03	35.43	1.61	4.71	0.23	0.09
	75	4.09	3.40	3.01	0.01	46.79	1.75	6.68	0.10	0.10
2PL										
500	15	3.87	2.73	7.17	0.01	239.08	14.16	272.70	0.27	0.23
	75	4.39	2.65	10.37	0.02	259.10	10.23	134.98	0.16	0.60
1,500	15	3.35	2.57	5.87	0.02	295.05	30.92	1212.99	0.28	0.27
	75	3.48	2.55	8.64	0.02	544.19	33.22	1395.59	0.18	0.70
3PL										
500	15	3.63	2.60	5.77	0.01	224.02	12.14	239.46	0.28	0.28
	75	3.72	2.55	7.70	0.02	367.84	16.28	450.07	0.18	0.68
1,500	15	3.08	2.42	3.73	0.01	245.61	31.65	1725.59	0.31	0.35
	75	2.99	2.29	4.76	0.02	262.33	36.97	1800.74	0.23	0.94

APPENDIX H: ANALYSIS OF L_z, *VI*, AND *VO* STATISTICS

z Statistic Sampling Distribution Means and SDs in SU Conditions

Scatterplots Between b and z Statistics in SU $\hat{\boldsymbol{\xi}},\hat{\boldsymbol{\theta}}$ Conditions

Scatterplots Between b and z Statistics in SU $\boldsymbol{\xi},\boldsymbol{\theta}$ Conditions

z Statistic Sampling Distribution Skewness and Kurtosis in SU $\boldsymbol{\xi},\boldsymbol{\theta}$ Conditions

Table H-1. *z* Statistic Sampling Distribution Means and SDs in SU Conditions (N = 500 n = 15)

D	PE		Mean			SD		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
					L_z			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	2.11	1.94	1.96	0.91	0.73	0.83
		$\boldsymbol{\xi}$	1.98	1.85	1.89	1.09	1.25	1.40
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.07	0.02	–0.07	0.93	0.65	0.71
		$\boldsymbol{\xi}$	0.00	0.01	0.01	0.99	0.99	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	2.27	2.27	2.29	0.98	0.95	1.04
		$\boldsymbol{\xi}$	2.23	2.19	2.26	1.13	1.35	1.51
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.01	0.03	–0.07	0.90	0.71	0.72
		$\boldsymbol{\xi}$	0.00	–0.01	0.01	1.00	1.00	1.01
					*V*I			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–2.15	–1.97	–1.98	1.02	0.84	0.94
		$\boldsymbol{\xi}$	–2.01	–1.90	–1.93	1.22	1.39	1.55
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.07	0.00	0.08	0.93	0.67	0.70
		$\boldsymbol{\xi}$	0.00	–0.01	–0.01	0.99	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–2.28	–2.29	–2.30	1.08	1.06	1.15
		$\boldsymbol{\xi}$	–2.24	–2.21	–2.28	1.24	1.48	1.64
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	–0.02	0.07	0.90	0.71	0.72
		$\boldsymbol{\xi}$	0.00	0.01	–0.01	1.00	1.00	1.01
					*V*O			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–2.01	–1.72	–1.83	0.94	0.73	0.72
		$\boldsymbol{\xi}$	–1.92	–1.70	–1.84	1.02	1.12	1.23
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.05	0.10	0.89	0.74	0.74
		$\boldsymbol{\xi}$	0.02	0.03	0.01	0.90	0.90	0.91
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–2.42	–2.39	–2.40	0.97	0.93	0.99
		$\boldsymbol{\xi}$	–2.38	–2.32	–2.40	1.08	1.33	1.47
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.00	0.07	0.93	0.77	0.77
		$\boldsymbol{\xi}$	0.00	0.01	0.00	0.99	0.99	1.00

Table H-2. *z* Statistic Sampling Distribution Means and SDs in SU Conditions (*N* = 1,500 *n* = 15)

D	PE		Mean			SD		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
					L_z			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	3.62	3.25	3.29	1.26	1.16	1.33
		$\boldsymbol{\xi}$	3.45	3.26	3.26	1.46	1.59	1.83
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.12	–0.09	–0.12	0.95	0.67	0.72
		$\boldsymbol{\xi}$	0.00	0.01	0.00	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	3.92	3.88	3.94	1.40	1.49	1.67
		$\boldsymbol{\xi}$	3.90	3.85	3.93	1.50	1.81	2.06
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	–0.02	–0.09	0.90	0.69	0.75
		$\boldsymbol{\xi}$	0.00	0.01	0.00	1.00	1.00	1.00
					*V*I			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–3.70	–3.32	–3.34	1.38	1.29	1.50
		$\boldsymbol{\xi}$	–3.50	–3.35	–3.30	1.61	1.77	2.03
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.12	0.10	0.11	0.95	0.69	0.72
		$\boldsymbol{\xi}$	0.00	0.00	0.00	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–3.94	–3.91	–3.96	1.54	1.65	1.84
		$\boldsymbol{\xi}$	–3.92	–3.90	–3.96	1.65	1.99	2.24
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	0.03	0.09	0.89	0.69	0.74
		$\boldsymbol{\xi}$	0.00	–0.01	–0.01	1.00	1.00	1.00
					*V*O			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–3.50	–2.94	–3.12	1.31	1.15	1.15
		$\boldsymbol{\xi}$	–3.38	–3.00	–3.18	1.41	1.47	1.59
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.06	0.10	0.13	0.93	0.77	0.80
		$\boldsymbol{\xi}$	0.00	0.01	0.00	0.94	0.92	0.96
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–4.18	–4.08	–4.14	1.34	1.42	1.55
		$\boldsymbol{\xi}$	–4.16	–4.08	–4.15	1.42	1.75	1.96
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	0.04	0.09	0.93	0.76	0.80
		$\boldsymbol{\xi}$	0.00	0.00	0.00	1.00	0.99	1.00

Table H-3. *z* Statistic Sampling Distribution Means and SDs in SU Conditions (*N* = 500 *n* = 75)

D	PE		Mean			SD		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
					L_z			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.55	0.54	0.56	0.85	0.30	0.33
		$\boldsymbol{\xi}$	0.47	0.48	0.34	1.00	1.06	1.10
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.02	–0.01	0.15	0.94	0.44	0.48
		$\boldsymbol{\xi}$	–0.01	0.00	0.01	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.59	0.59	0.63	0.83	0.42	0.38
		$\boldsymbol{\xi}$	0.55	0.44	0.51	0.98	1.05	1.09
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.02	0.13	0.07	0.88	0.46	0.42
		$\boldsymbol{\xi}$	0.00	0.00	–0.01	0.99	1.00	1.00
					*V*I			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.53	–0.50	–0.54	0.88	0.34	0.36
		$\boldsymbol{\xi}$	–0.45	–0.46	–0.34	1.03	1.09	1.11
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.03	–0.13	0.94	0.48	0.49
		$\boldsymbol{\xi}$	0.01	0.00	–0.01	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.57	–0.56	–0.60	0.85	0.43	0.40
		$\boldsymbol{\xi}$	–0.53	–0.42	–0.50	1.00	1.07	1.11
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	–0.11	–0.06	0.87	0.47	0.42
		$\boldsymbol{\xi}$	0.00	0.00	0.01	0.99	1.00	1.00
					*V*O			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.49	–0.41	–0.44	0.81	0.50	0.42
		$\boldsymbol{\xi}$	–0.45	–0.41	–0.34	0.86	0.91	0.95
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.01	0.09	–0.05	0.87	0.63	0.56
		$\boldsymbol{\xi}$	0.02	0.05	0.02	0.89	0.89	0.93
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.60	–0.59	–0.62	0.86	0.49	0.42
		$\boldsymbol{\xi}$	–0.58	–0.47	–0.53	0.96	1.02	1.05
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	–0.10	–0.06	0.92	0.57	0.51
		$\boldsymbol{\xi}$	0.01	0.00	0.01	0.98	0.99	0.99

Table H-4. *z* Statistic Sampling Distribution Means and SDs in SU Conditions (*N* = 1,500 *n* = 75)

D	PE		Mean			SD		
	θ	ξ	1PL	2PL	3PL	1PL	2PL	3PL
					L_z			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.92	0.89	0.87	0.88	0.33	0.34
		$\boldsymbol{\xi}$	0.86	0.92	0.49	1.03	1.10	1.16
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.03	–0.09	0.25	0.94	0.43	0.51
		$\boldsymbol{\xi}$	0.00	–0.01	–0.01	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	0.99	1.00	1.05	0.87	0.44	0.45
		$\boldsymbol{\xi}$	0.98	0.90	0.86	1.00	1.08	1.14
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	0.09	0.13	0.88	0.39	0.41
		$\boldsymbol{\xi}$	0.00	0.00	–0.01	0.99	1.00	1.00
					*V*I			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.88	–0.84	–0.86	0.90	0.37	0.37
		$\boldsymbol{\xi}$	–0.82	–0.88	–0.48	1.05	1.13	1.19
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.03	0.10	–0.24	0.94	0.46	0.53
		$\boldsymbol{\xi}$	0.01	0.01	0.00	1.00	1.00	1.00
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.95	–0.96	–1.02	0.88	0.46	0.47
		$\boldsymbol{\xi}$	–0.94	–0.88	–0.84	1.03	1.10	1.17
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	0.00	–0.08	–0.12	0.88	0.40	0.42
		$\boldsymbol{\xi}$	0.00	0.00	0.01	0.99	1.00	1.00
					*V*O			
High	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–0.85	–0.77	–0.75	0.87	0.57	0.46
		$\boldsymbol{\xi}$	–0.82	–0.83	–0.55	0.92	0.96	1.00
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	0.11	–0.11	0.93	0.70	0.61
		$\boldsymbol{\xi}$	0.01	0.02	0.02	0.94	0.92	0.94
Low	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\xi}}$	–1.04	–1.02	–1.05	0.89	0.52	0.47
		$\boldsymbol{\xi}$	–1.03	–0.94	–0.91	0.98	1.05	1.10
	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\xi}}$	–0.01	–0.07	–0.10	0.92	0.55	0.50
		$\boldsymbol{\xi}$	–0.01	0.00	0.01	0.98	1.00	0.99

Figure H-1. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\hat{\xi}, \hat{\theta}$ Conditions ($N = 500$ $n = 15$)

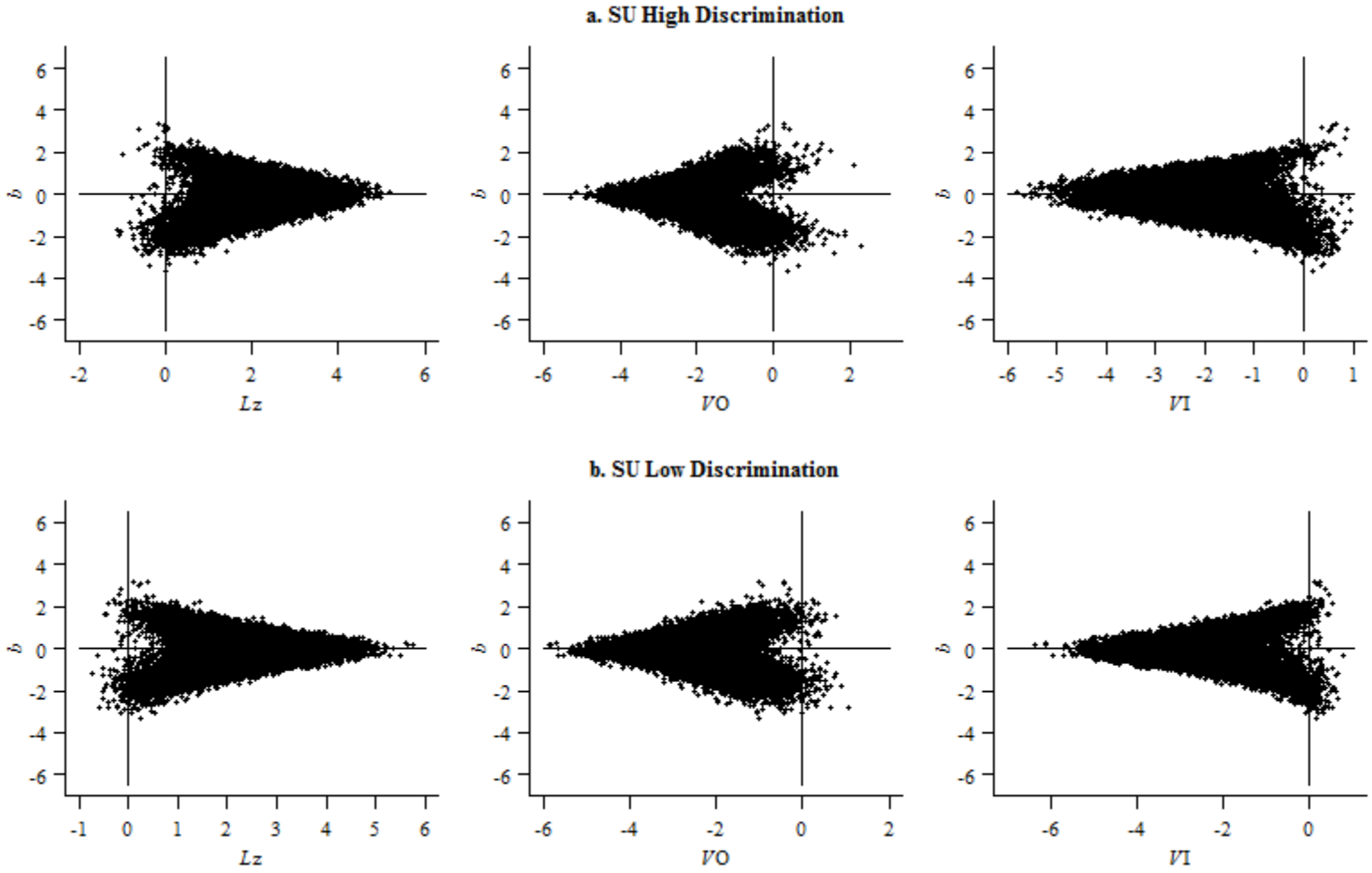


Figure H-2. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\hat{\xi}, \hat{\theta}$ Conditions (N = 1,500 n = 15)

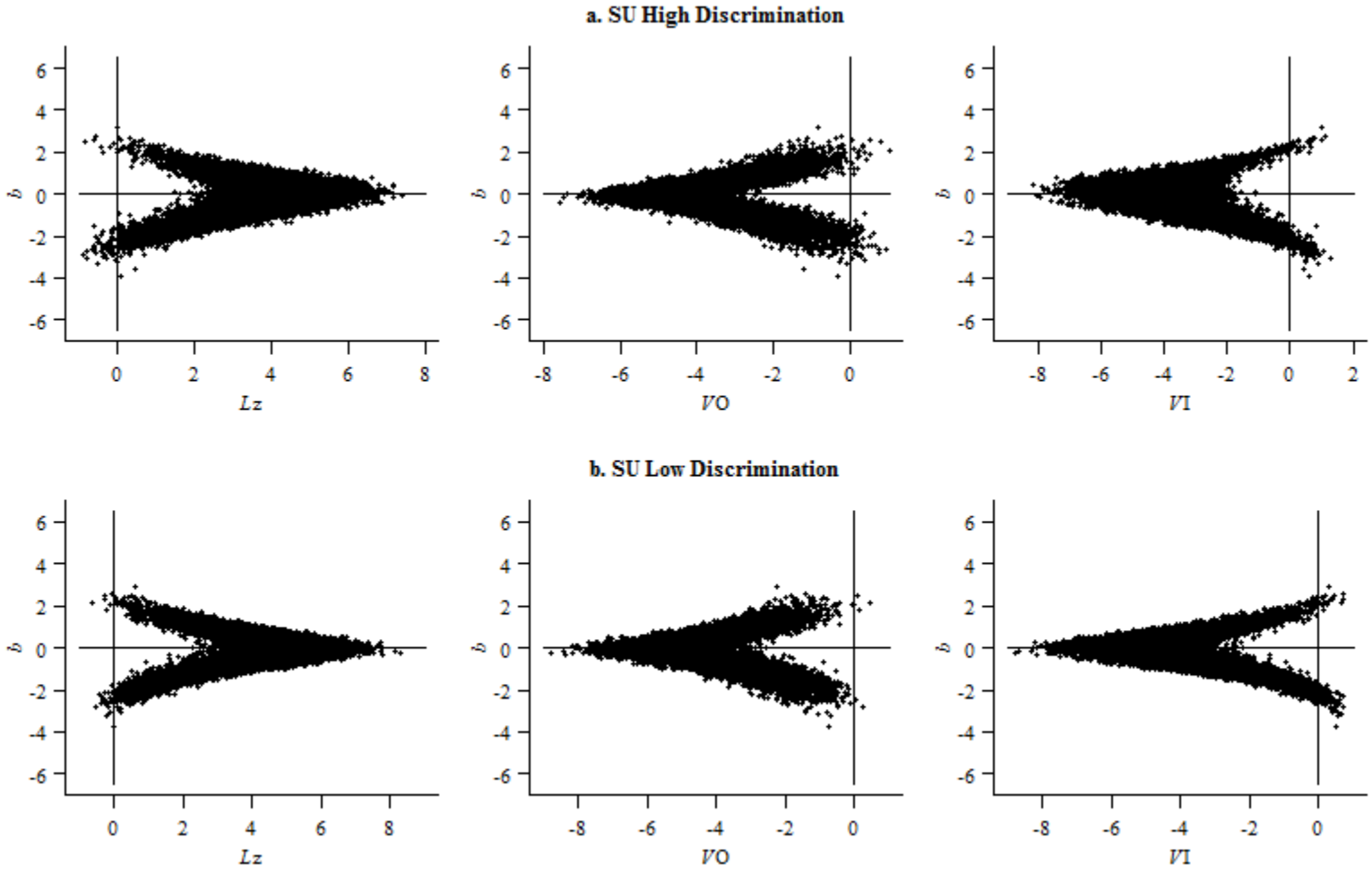


Figure H-3. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions ($N = 500$ $n = 75$)

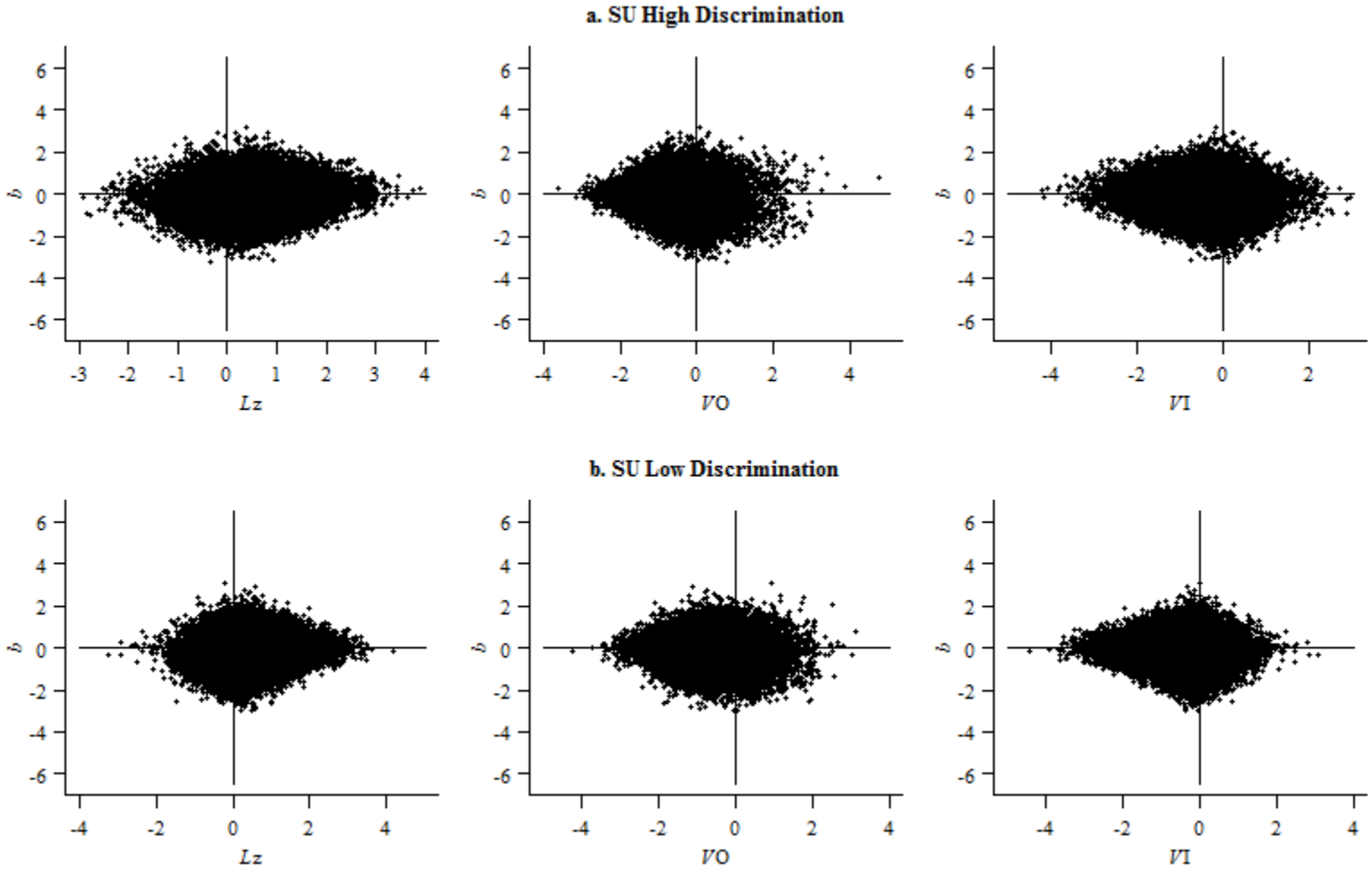


Figure H-4. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\hat{\xi}, \hat{\theta}$ Conditions (N = 1,500 n = 75)

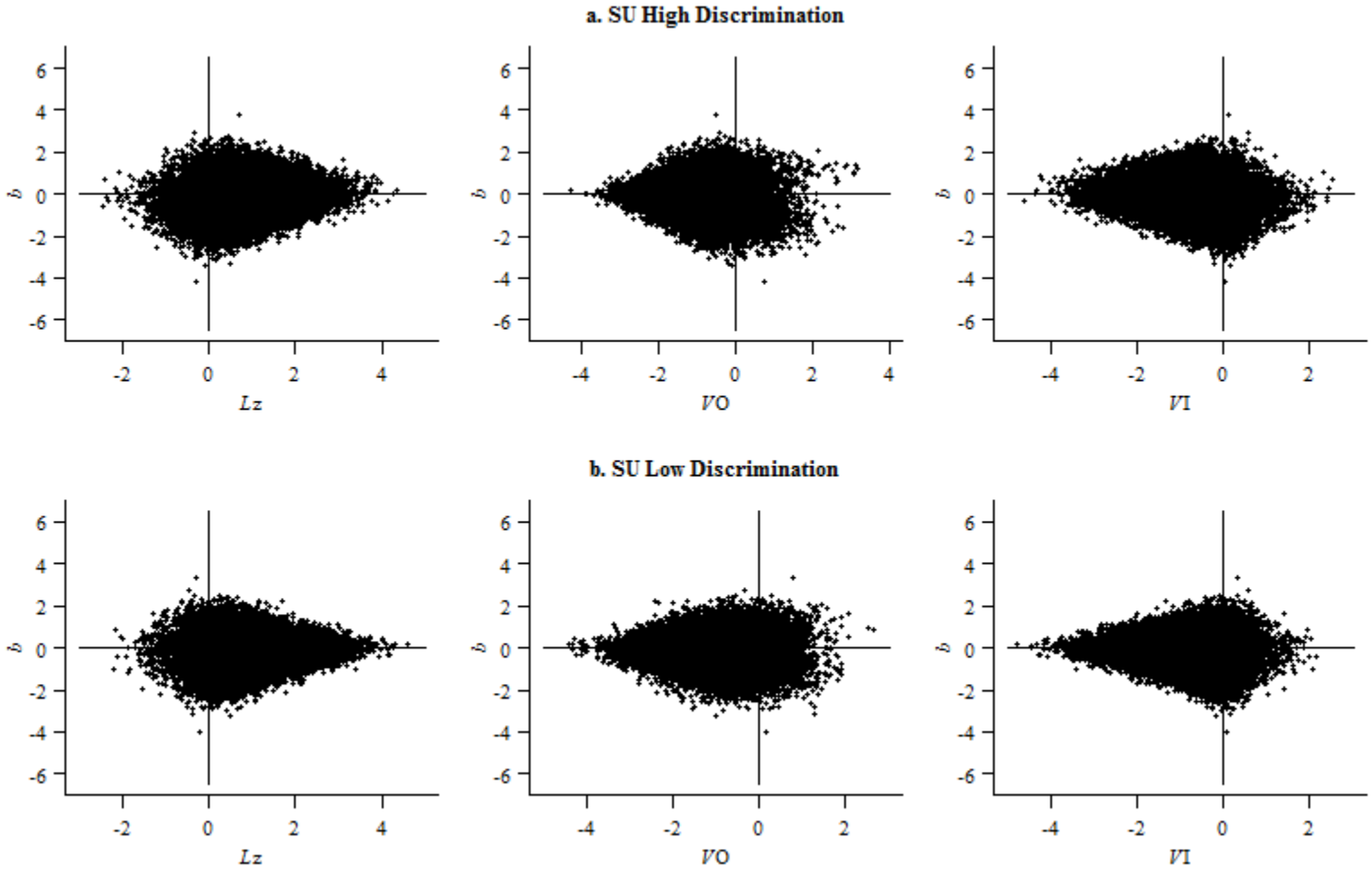


Figure H-5. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions ($N = 500$ $n = 15$)

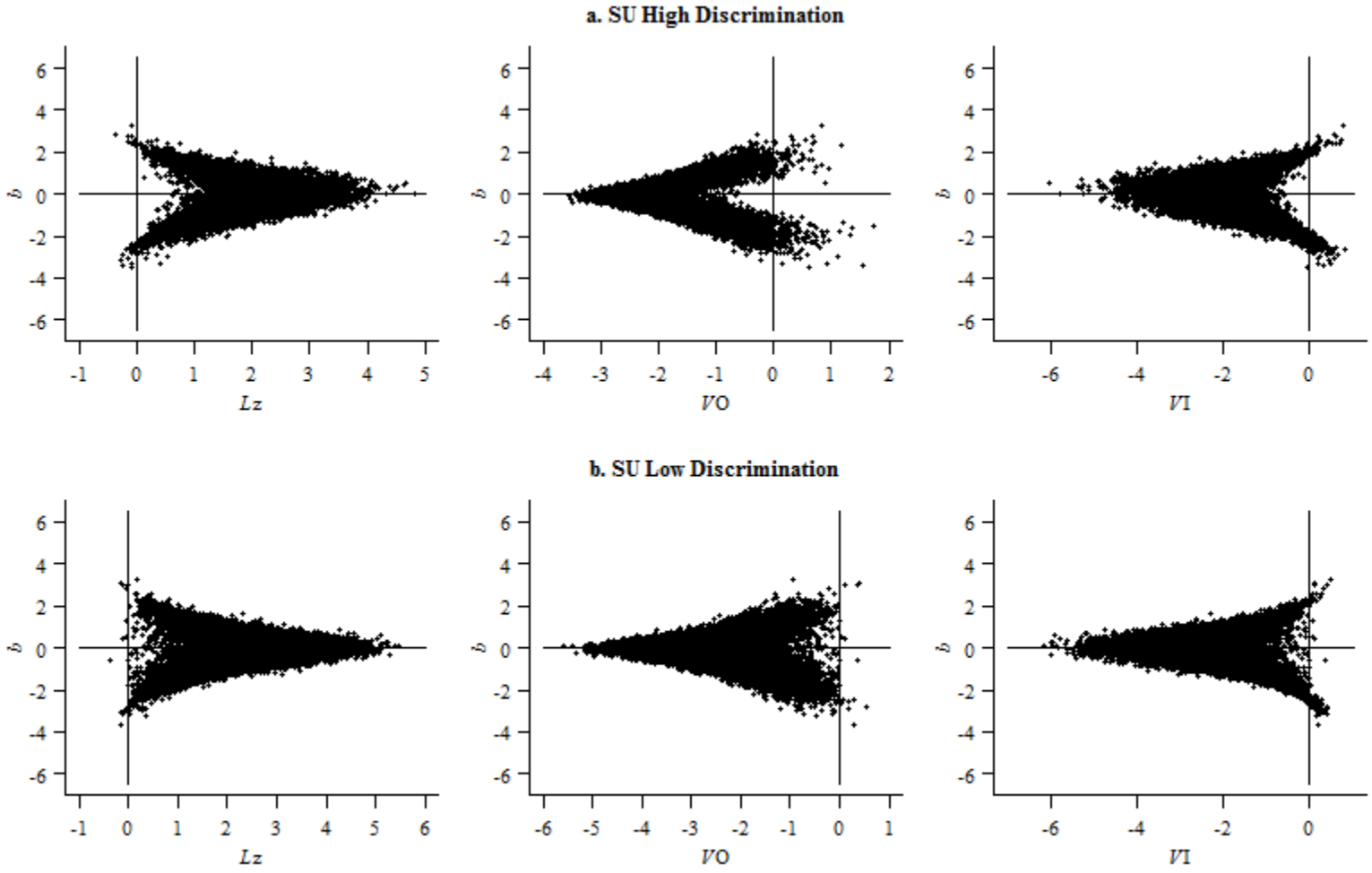


Figure H-6. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions (N = 1,500 n = 15)

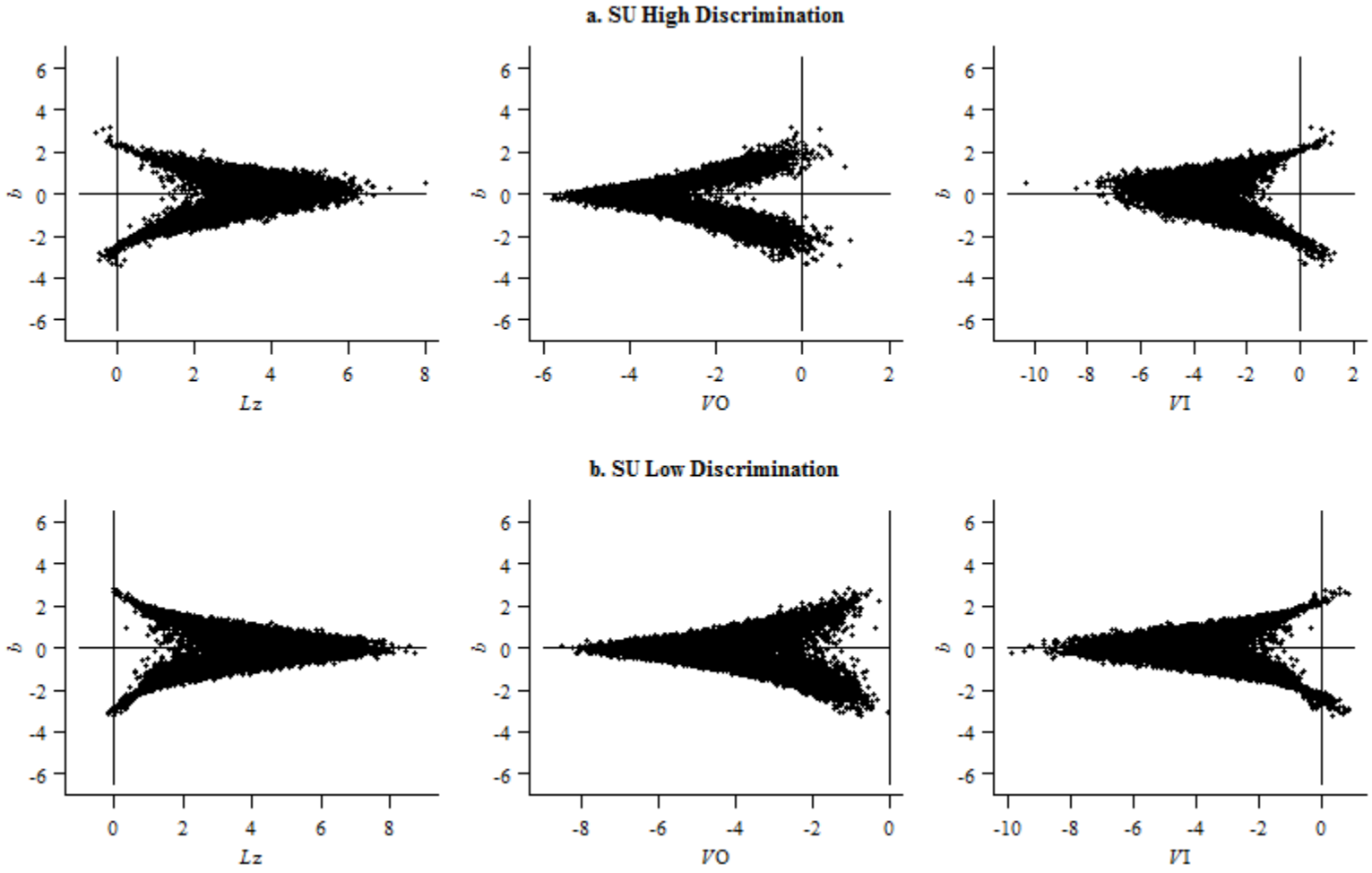


Figure H-7. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions ($N = 500$ $n = 75$)

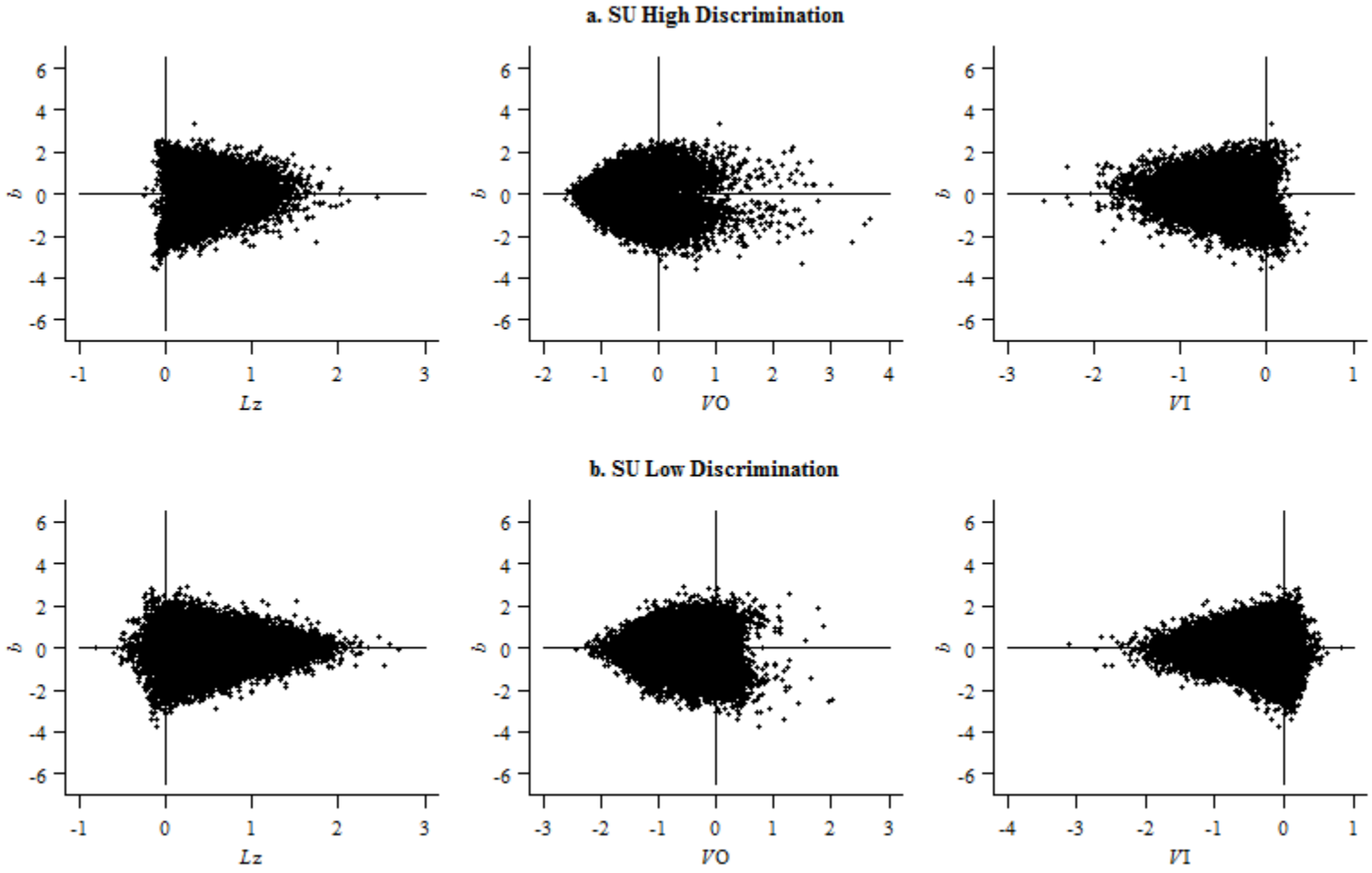


Figure H-8. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions (N = 1,500 n = 75)

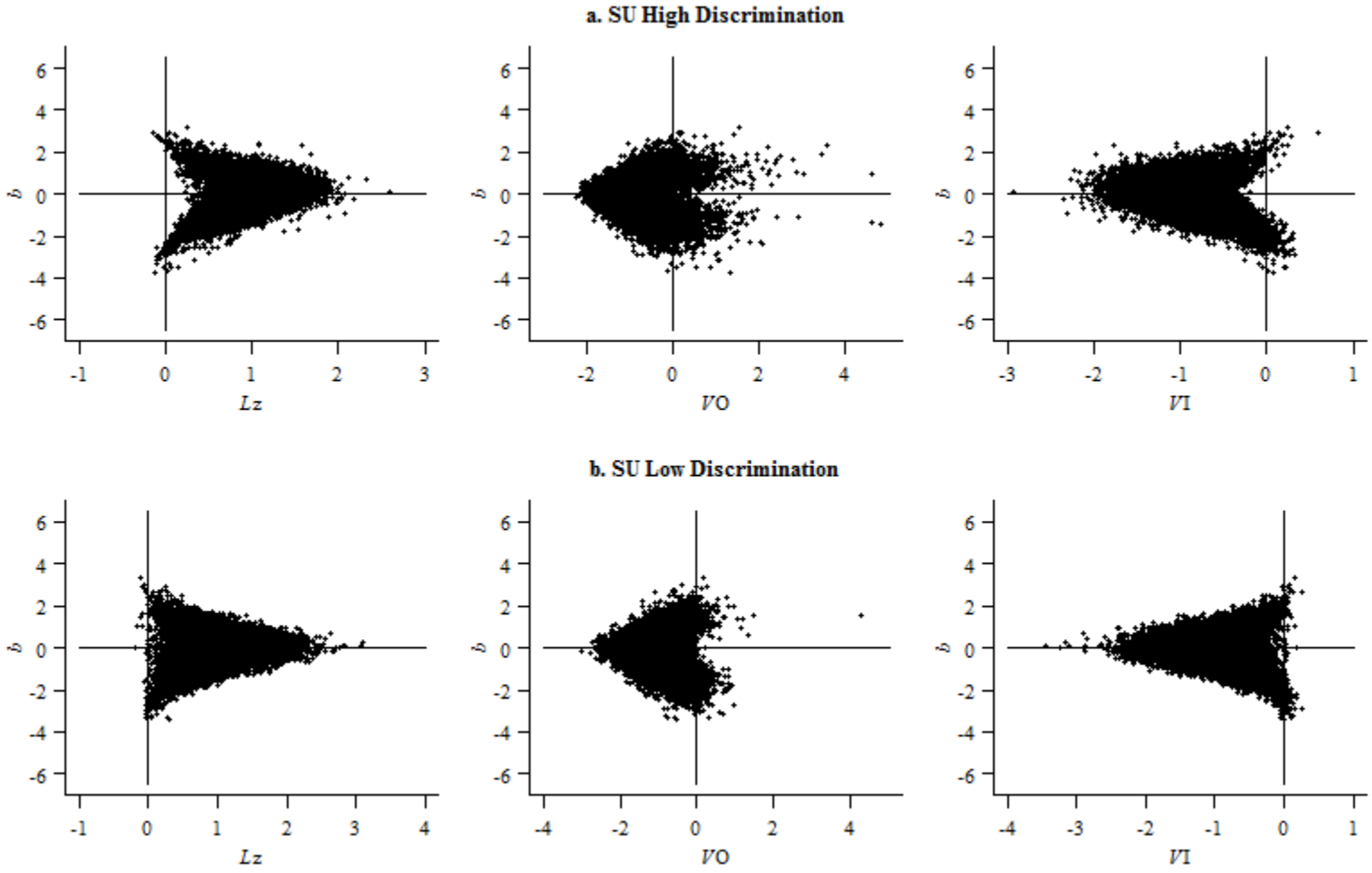


Figure H-9. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions ($N = 500$ $n = 15$)

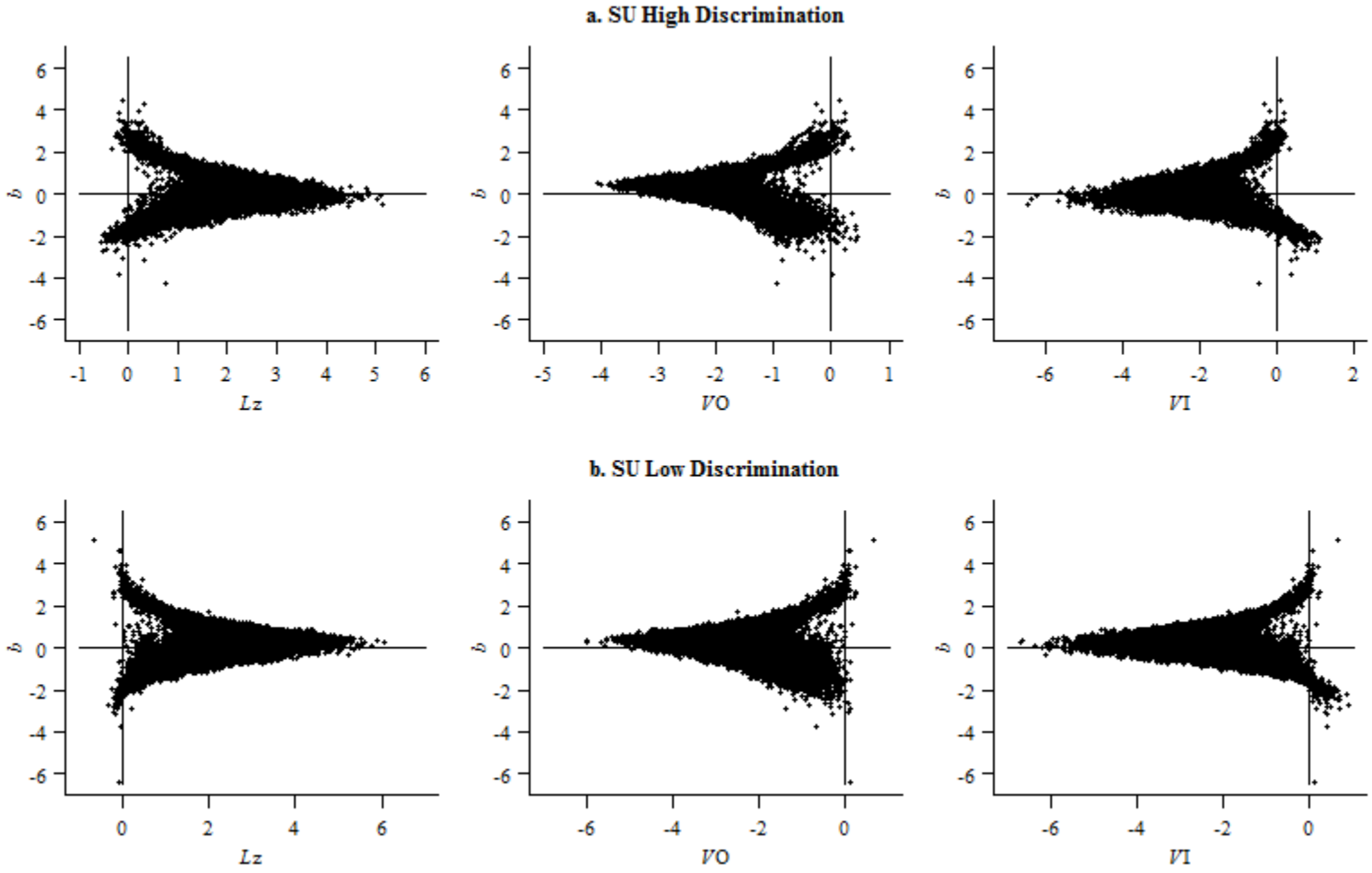


Figure H-10. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions (N = 1,500 n = 15)

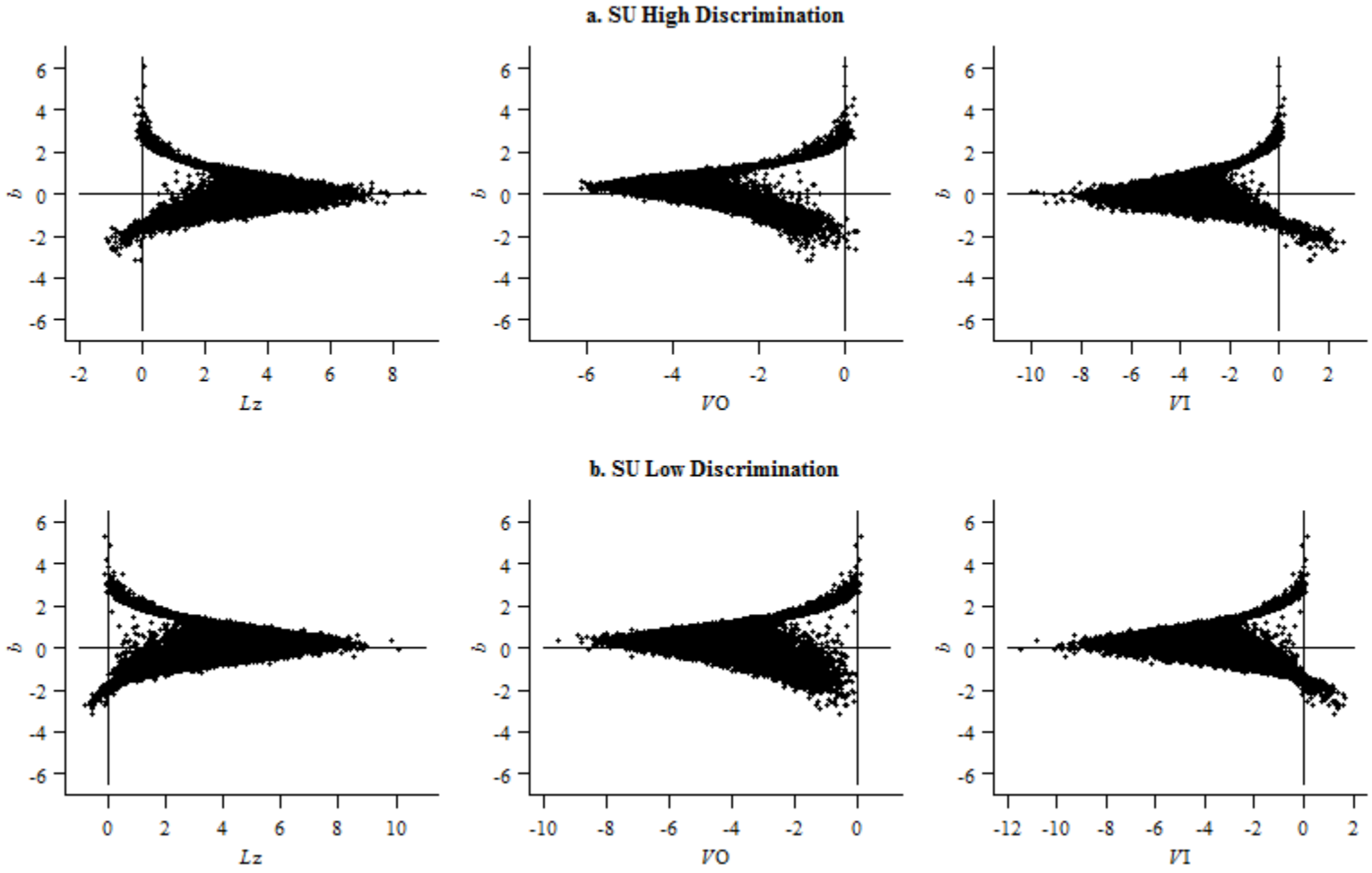


Figure H-11. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions ($N = 500$ $n = 75$)

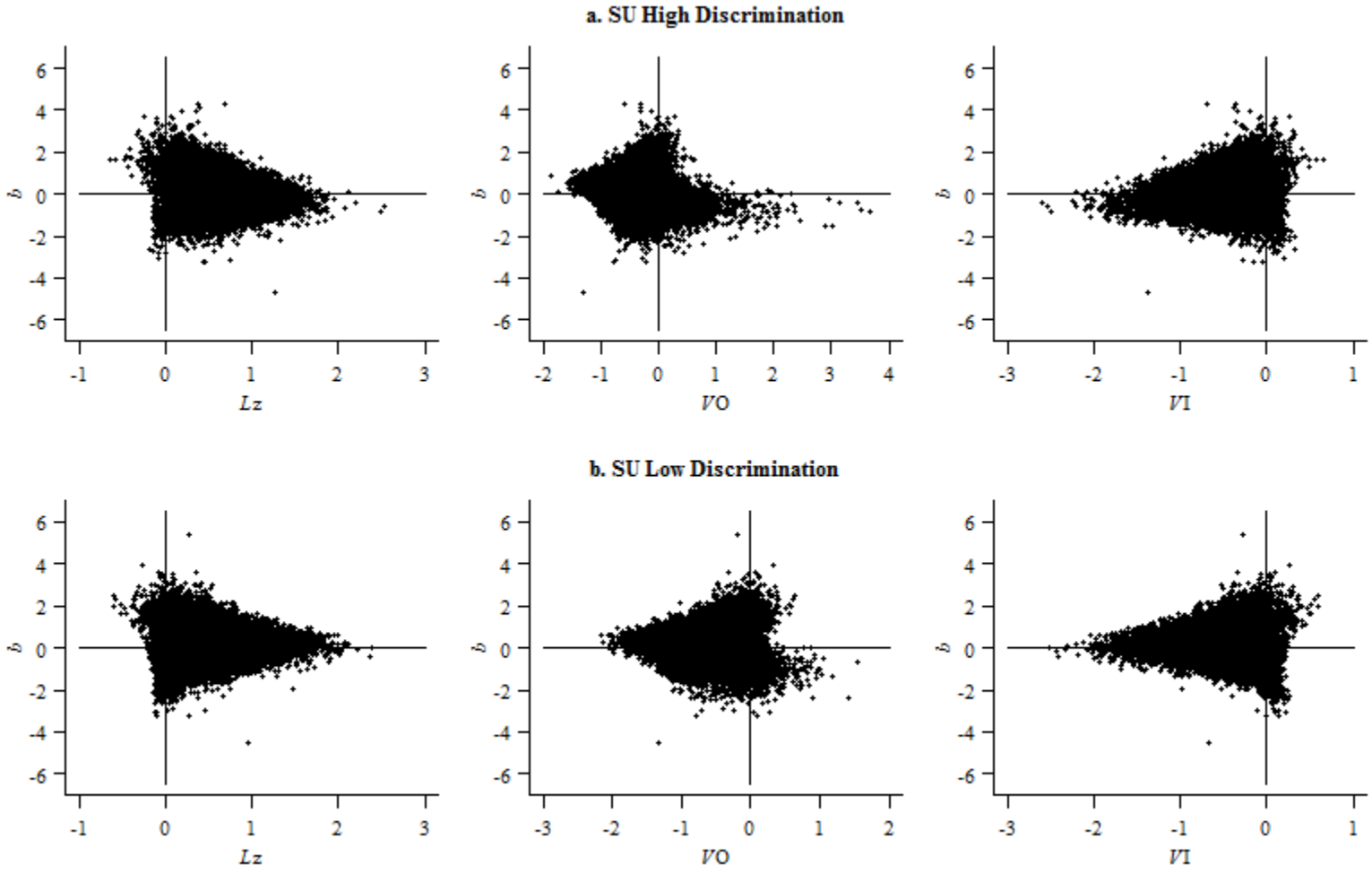


Figure H-12. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\hat{\boldsymbol{\xi}}, \hat{\boldsymbol{\theta}}$ Conditions (N = 1,500 n = 75)

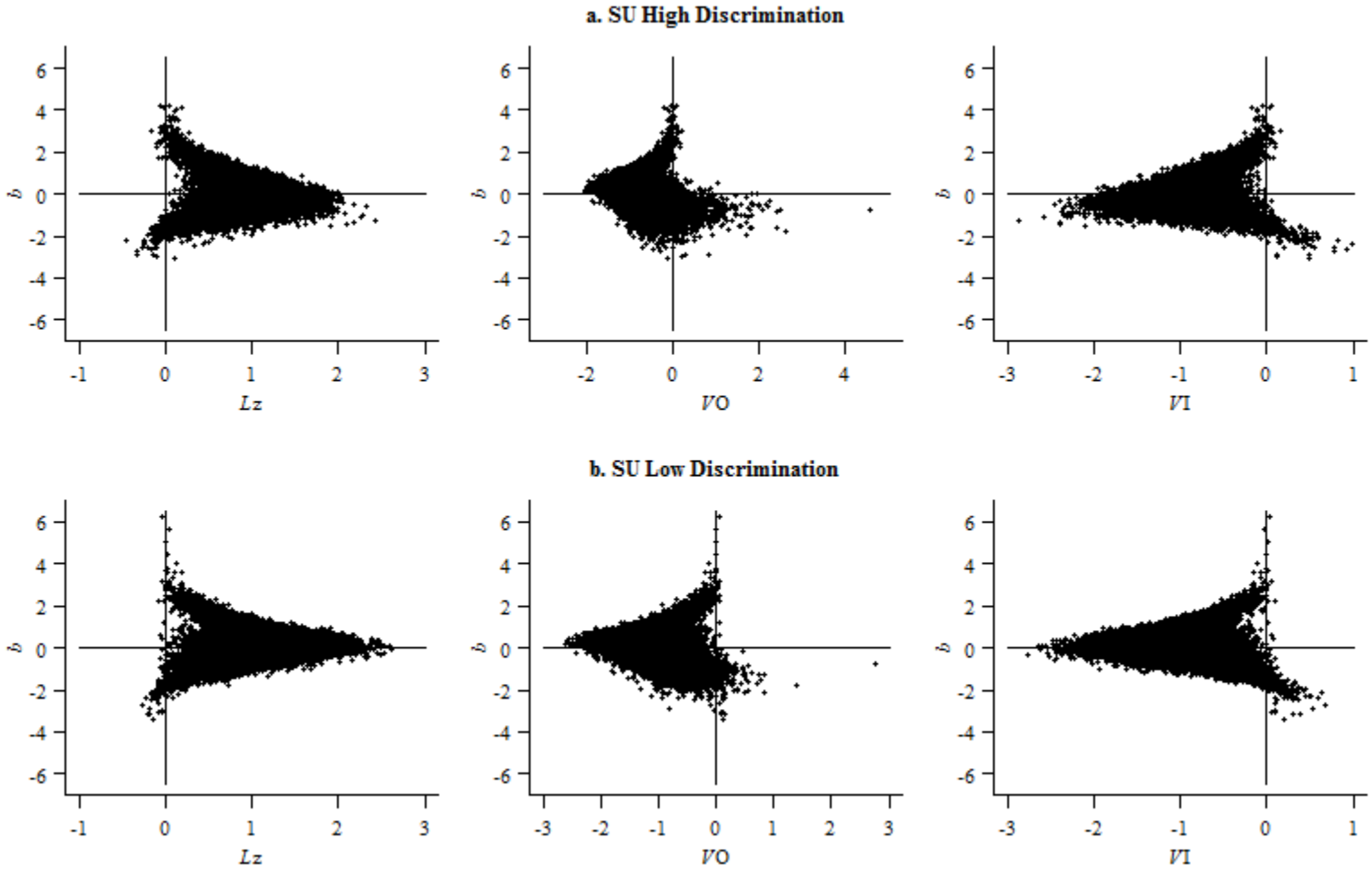


Figure H-13. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 15$)

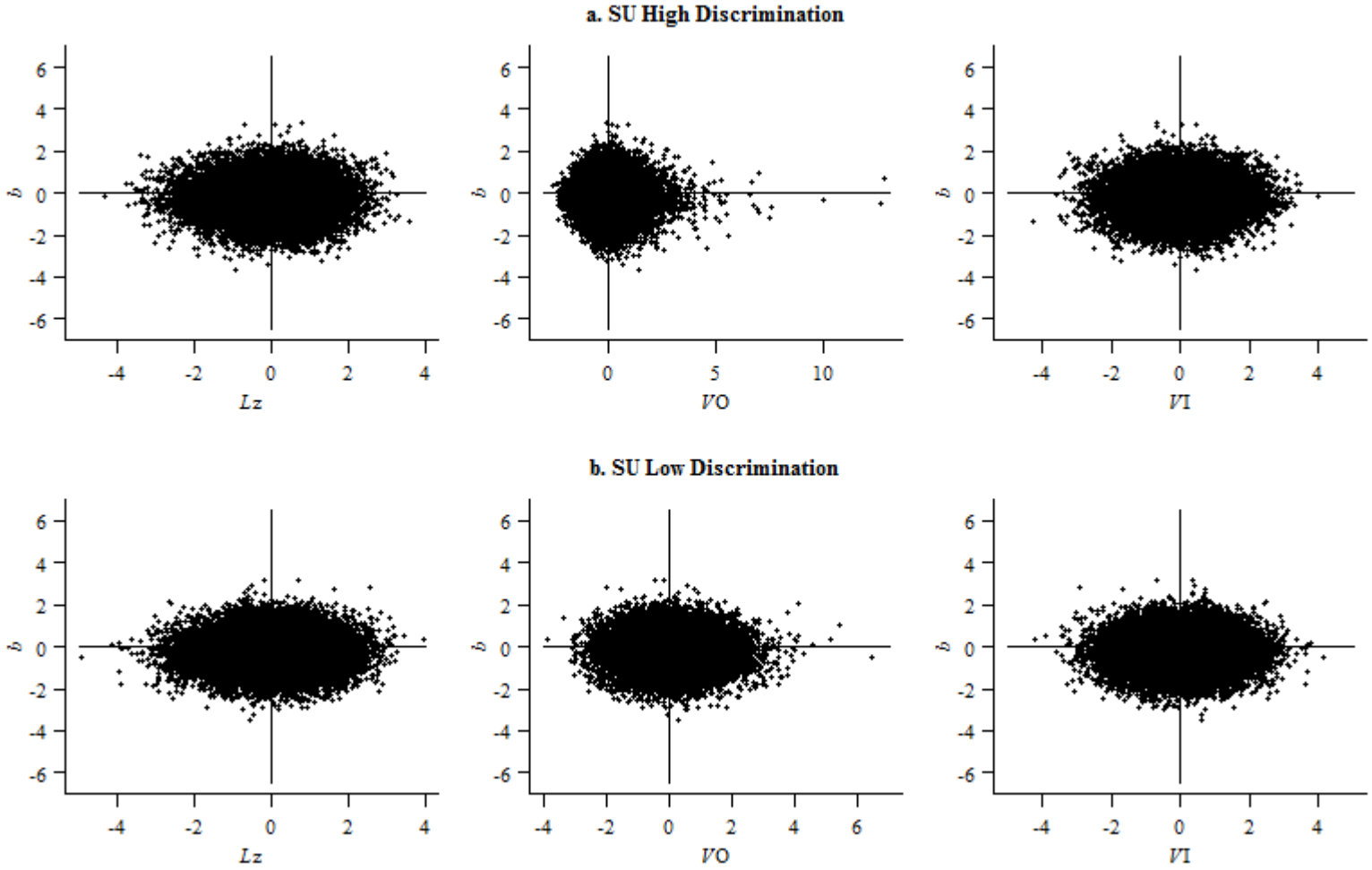


Figure H-14. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 15)

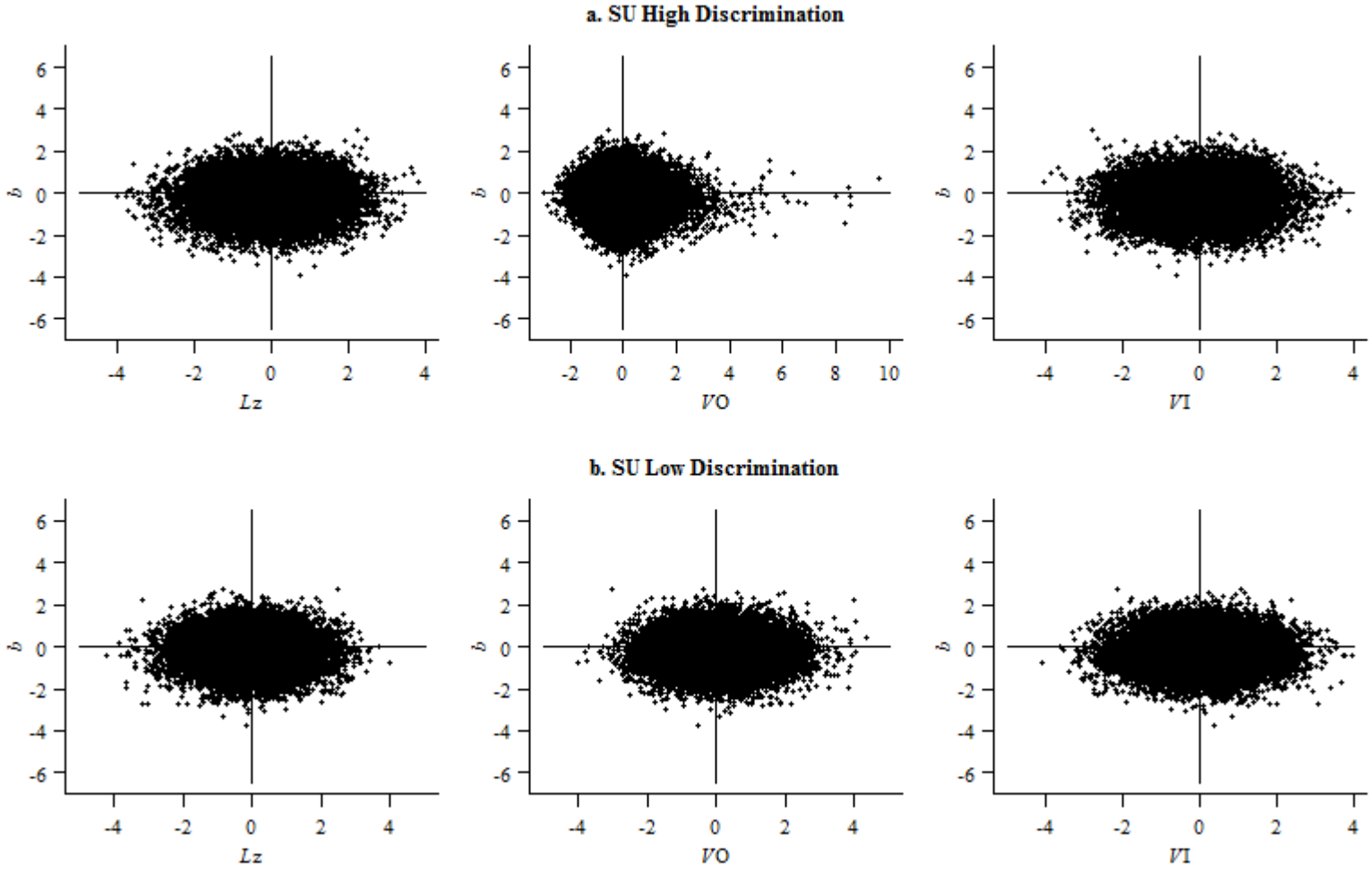


Figure H-15. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 75$)

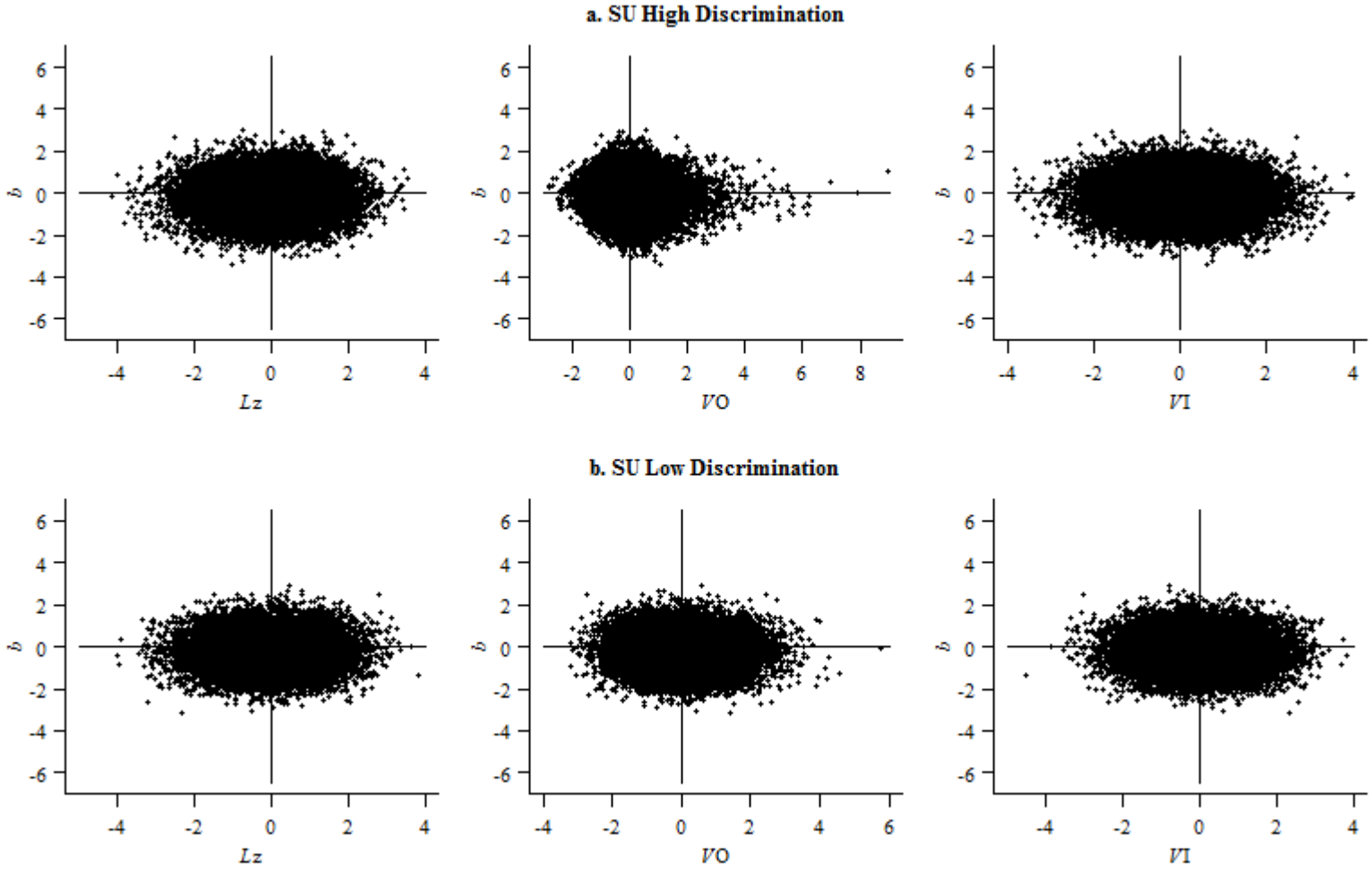


Figure H-16. Scatterplots Between *b* and *z* Statistics for the 1PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 75)

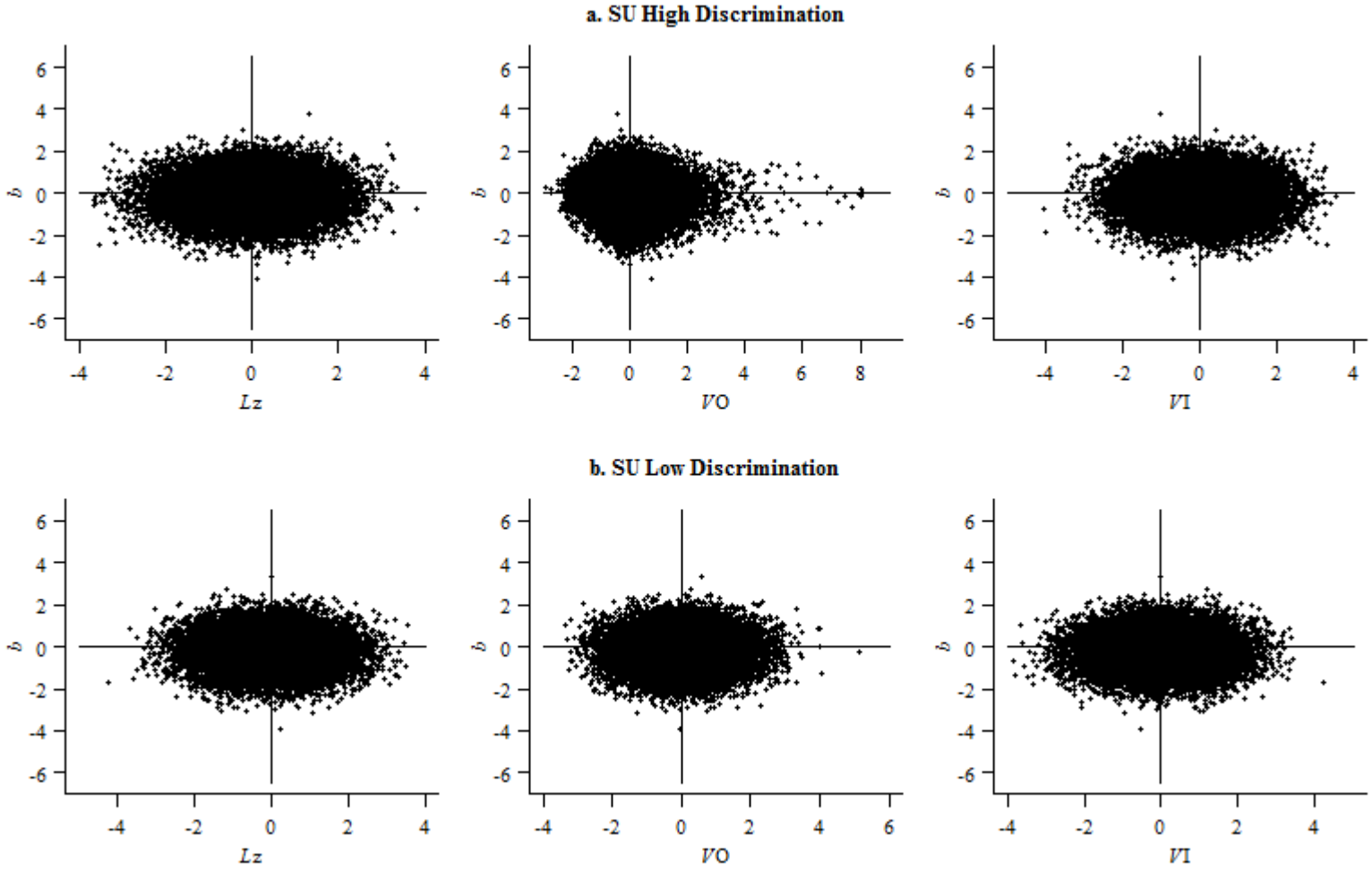


Figure H-17. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 15$)

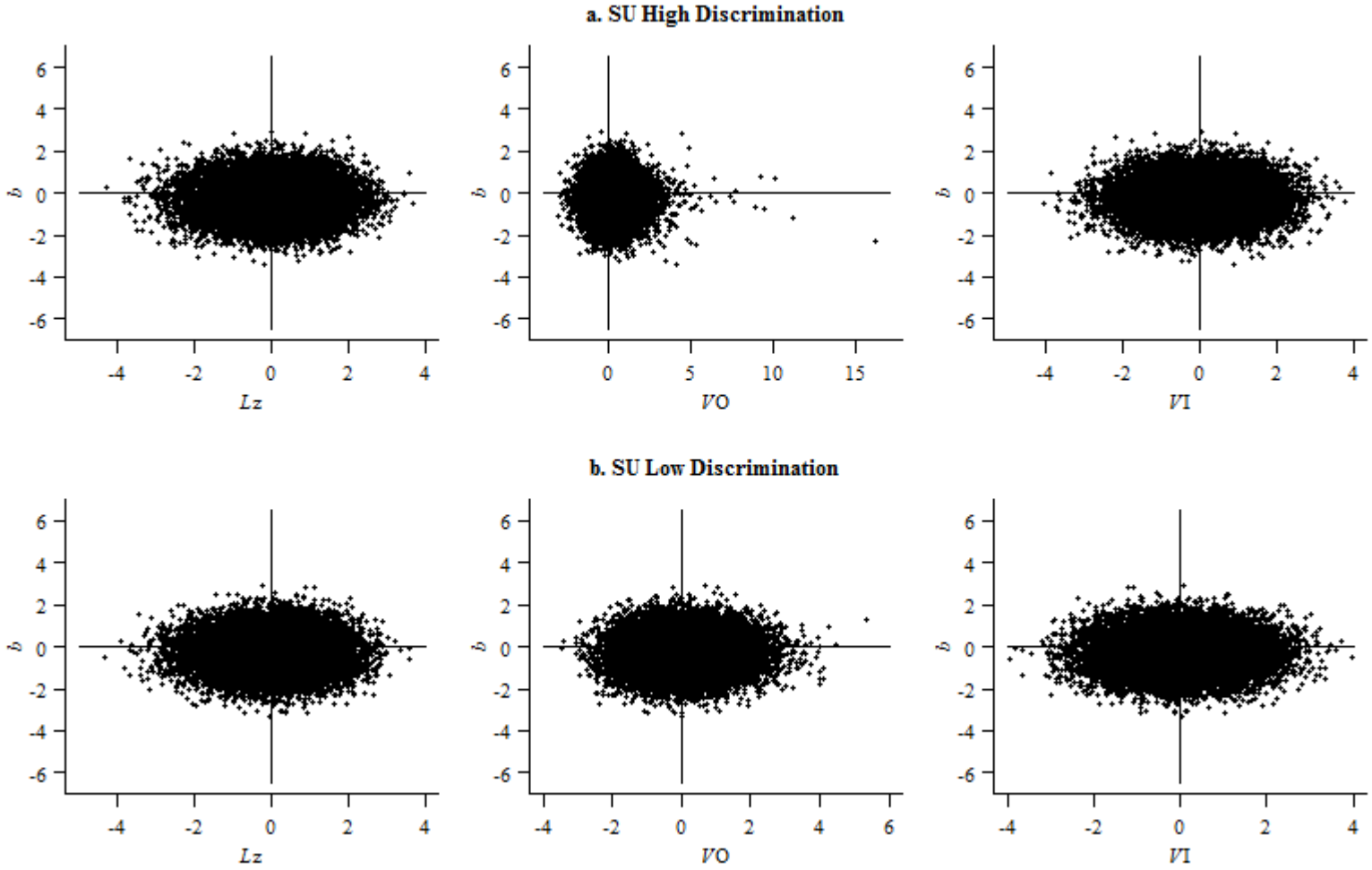


Figure H-18. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 15)

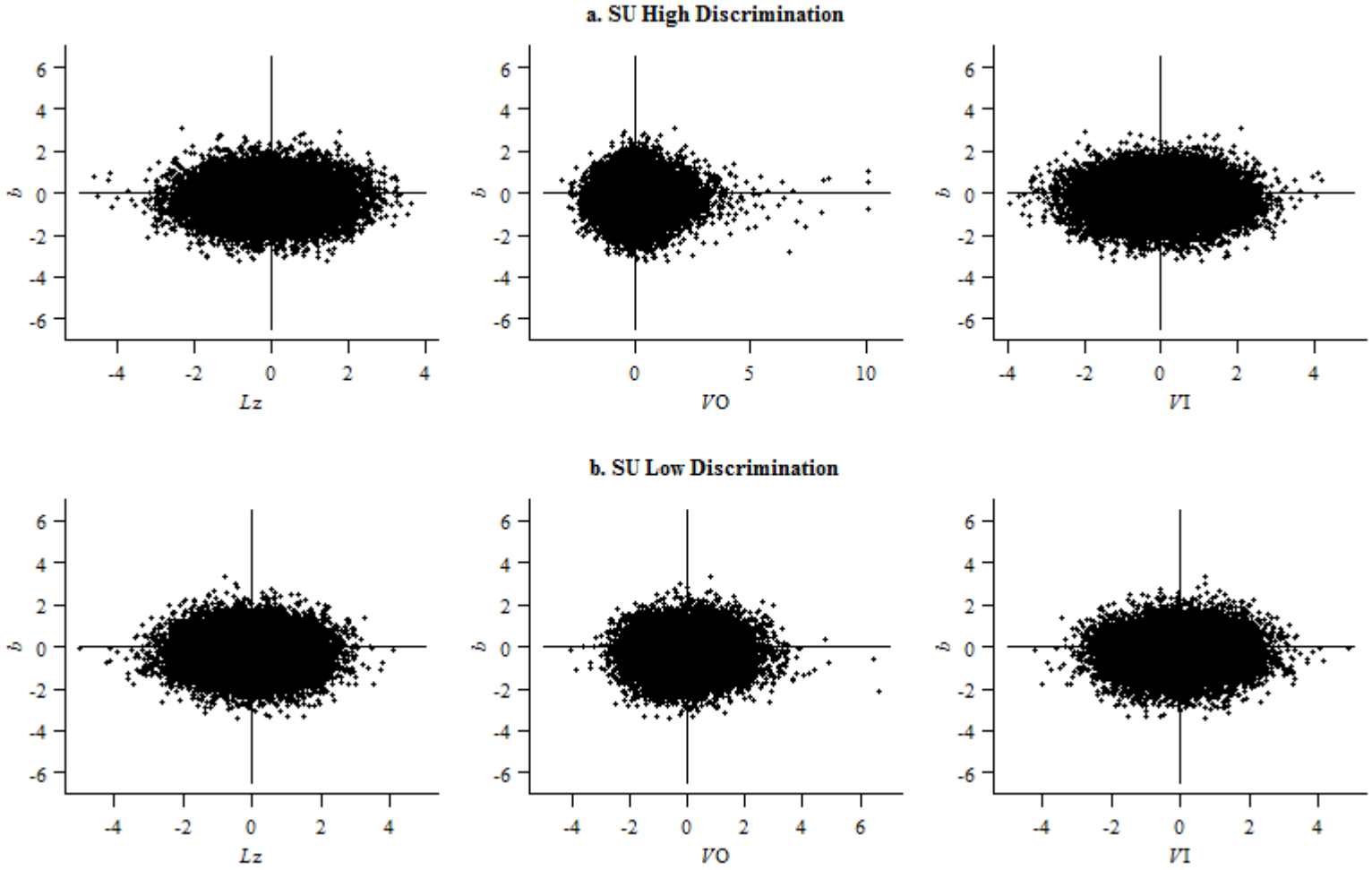


Figure H-19. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 75$)

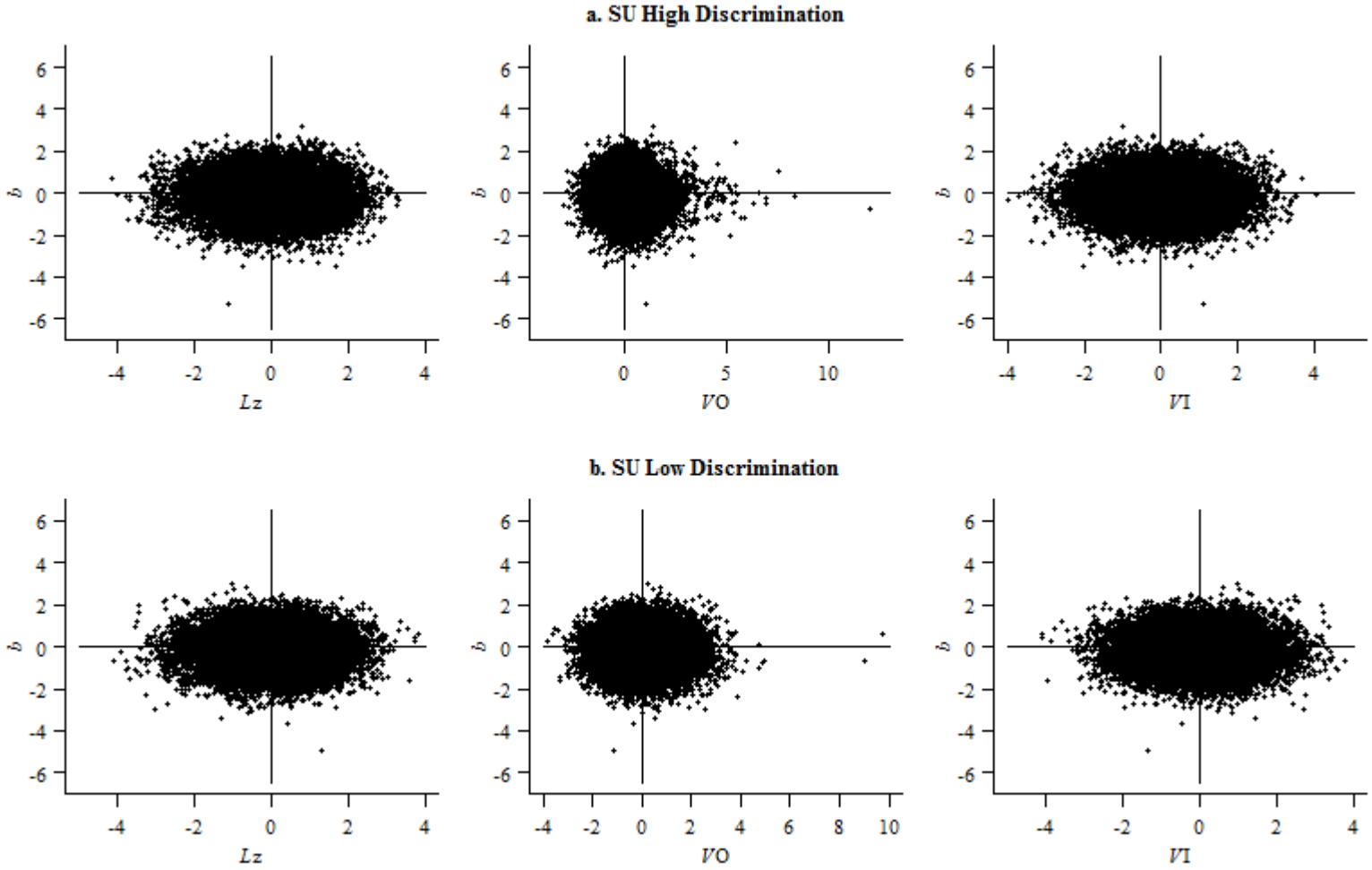


Figure H-20. Scatterplots Between *b* and *z* Statistics for the 2PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 75)

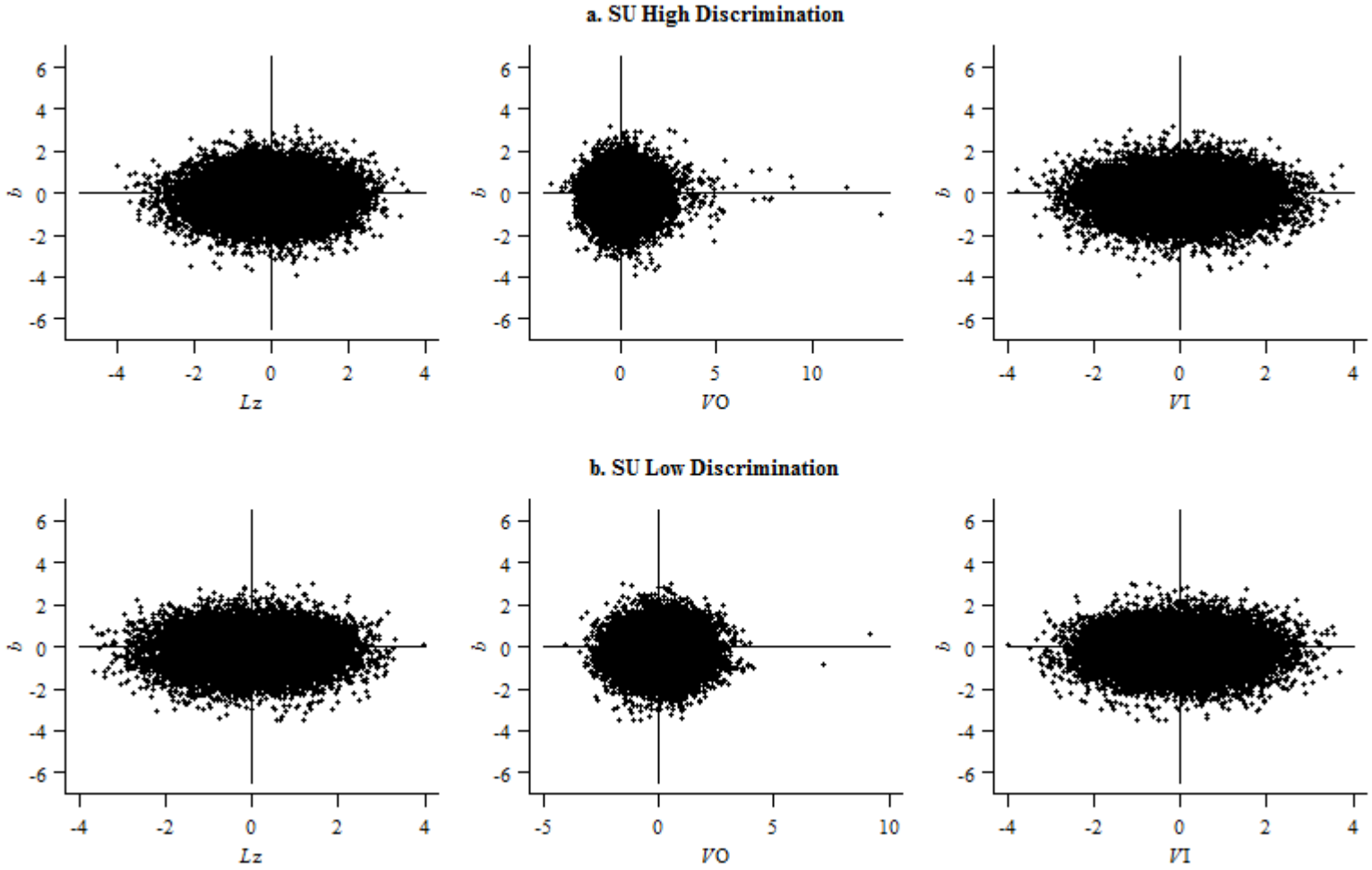


Figure H-21. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 15$)

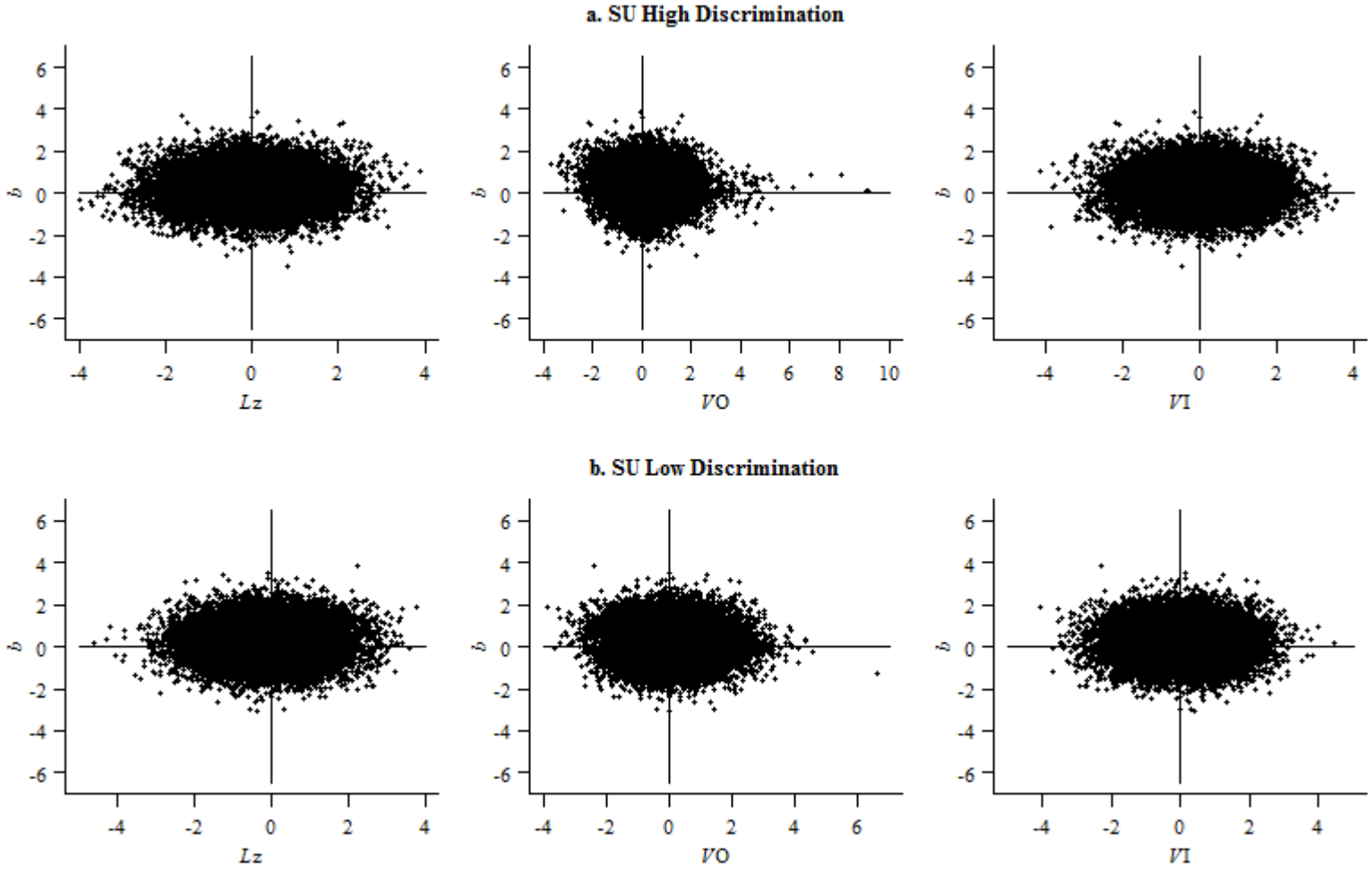


Figure H-22. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 15)

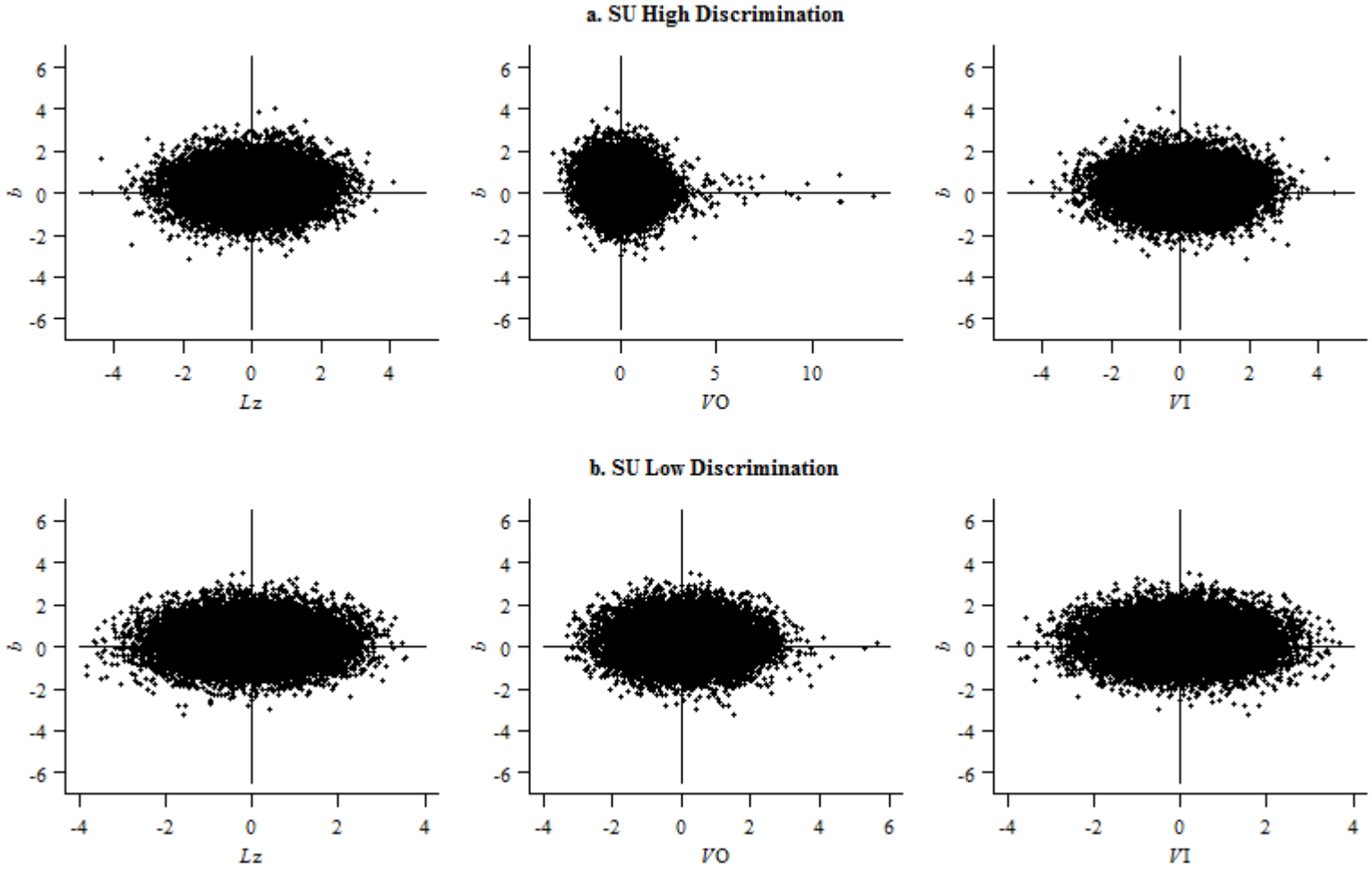


Figure H-23. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions ($N = 500$ $n = 75$)

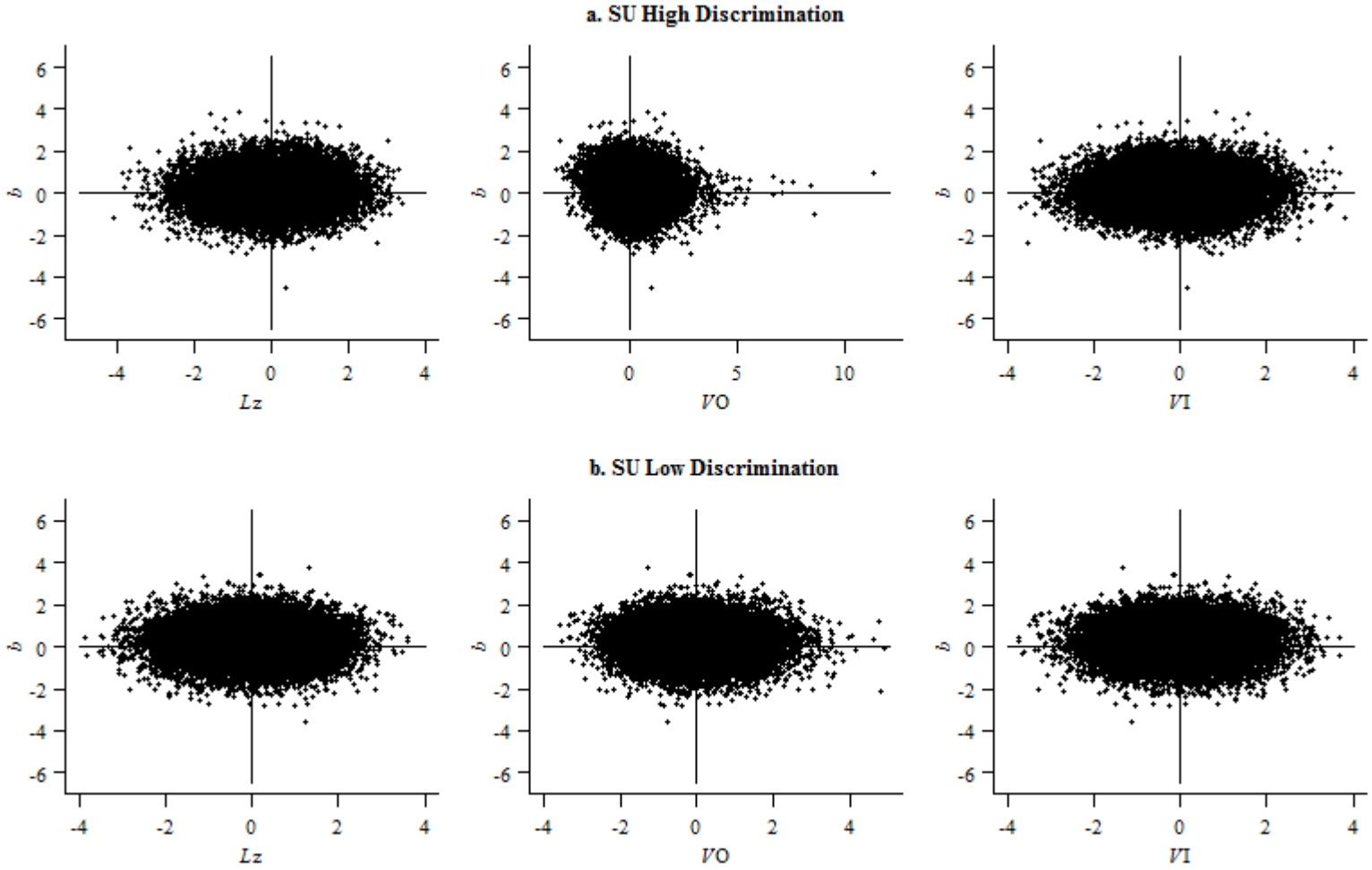


Figure H-24. Scatterplots Between *b* and *z* Statistics for the 3PL in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions (N = 1,500 n = 75)

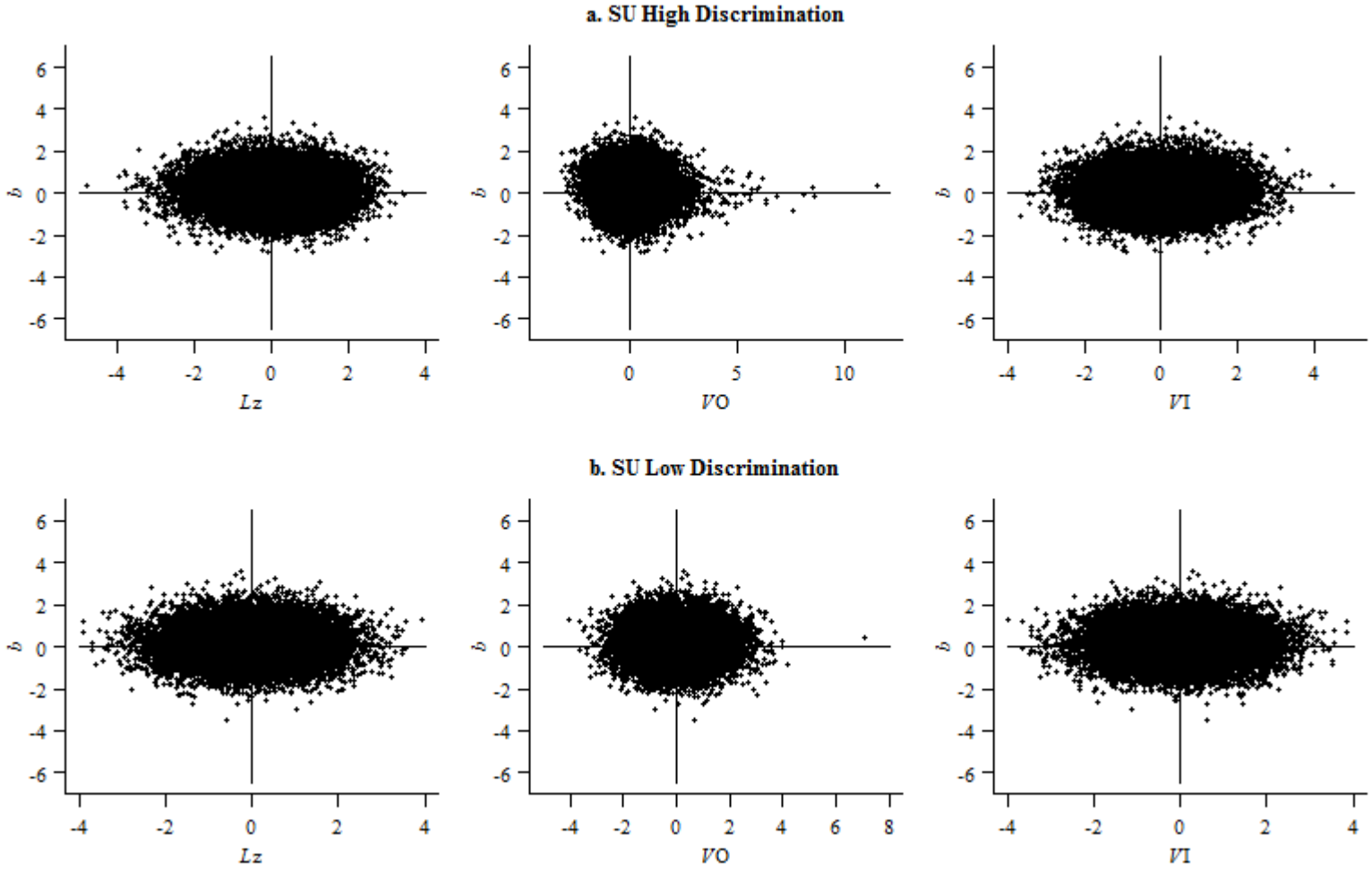


Table H-5. z Statistic Sampling Distribution Skewness and Kurtosis in SU $\boldsymbol{\xi}, \boldsymbol{\theta}$ Conditions

D	N	n	Skewness			Kurtosis		
			1PL	2PL	3PL	1PL	2PL	3PL
					L_z			
High	500	15	−0.19	−0.13	−0.07	0.01	0.08	0.01
		75	−0.15	−0.16	−0.11	0.05	0.02	0.04
	1,500	15	−0.08	−0.10	−0.07	0.03	0.03	0.04
		75	−0.11	−0.14	−0.10	−0.03	−0.04	0.02
Low	500	15	−0.16	−0.11	−0.11	0.09	−0.05	0.05
		75	−0.11	−0.10	−0.05	0.00	0.08	−0.06
	1,500	15	−0.09	−0.08	−0.06	0.05	0.02	−0.05
		75	−0.05	−0.06	−0.02	0.03	−0.04	−0.01
					*V*I			
High	500	15	0.01	−0.03	−0.06	−0.03	0.04	0.01
		75	−0.02	−0.01	−0.03	0.03	−0.01	0.02
	1,500	15	−0.02	−0.01	0.01	0.03	0.00	0.03
		75	0.00	0.04	0.03	−0.04	−0.04	0.00
Low	500	15	0.04	0.01	0.01	0.05	−0.08	0.03
		75	0.00	−0.02	−0.03	0.00	0.06	−0.05
	1,500	15	0.02	0.02	0.01	0.04	0.01	−0.05
		75	−0.01	0.00	−0.03	0.03	−0.06	−0.01
					*V*O			
High	500	15	1.40	1.45	0.73	8.25	11.97	3.32
		75	1.04	1.06	0.87	3.78	5.25	4.11
	1,500	15	1.04	1.15	1.22	4.34	6.02	8.92
		75	1.06	1.17	0.88	4.00	7.50	4.23
Low	500	15	0.24	0.23	0.22	0.33	0.18	0.28
		75	0.20	0.30	0.14	0.20	1.11	0.16
	1,500	15	0.15	0.18	0.13	0.14	0.29	0.11
		75	0.09	0.19	0.10	0.06	0.55	0.18