Supplemental material, for *Item-fit statistic based on posterior probabilities of membership in ability groups*, by Bartosz Kondratek, in *Applied Psychological Measurement.*

This supplement provides a detailed description of simulation Study 1 and its results. The goal of this Monte Carlo experiment was to investigate relationship between the number of ability bins used for computation of χ^2_w and its asymptotic properties under true H_0 and its power to detect misfit. The Monte Carlo experiment covered items of varying marginal difficulty, different IRT models, two sample sizes, $m \in \{400, 4000\}$, and two test lengths, $n \in \{10,\ 40\}$.

Item parameters used in simulations are presented in Table 1. Under each IRT model one item of moderate difficulty ($\pi_j = 0.5$) and one easy ($\pi_j = 0.9$) item were tested for fit. For 1PL and 2PL models the case of easy item is equivalent to a hard item ($\pi_j = 0.1$) scenario because both θ and $b_{v \neq j}$ are symmetrically distributed around 0. This equivalence does not hold for the 3PLM, so a hard item condition ($\pi_j = 0.3$) was added.

It should be noted that when item j was simulated under the 3PLM, the remaining items $v \neq j$ were simulated according to the same 2PL model that was used when j was simulated from 2PLM. Also, the fixed $a_{v \neq j} = 1.7$ used when j was simulated under the 1PLM was equal to the expected median value of $a_{v \neq j}$ for remaining items under other scenarios. Such a setup was chosen to obtain a comparable precision of measurement of ability between different scenarios of IRT models for the item j.

Table 1

Item parameters used in simulations evaluating performance of χ^2_w under true H_0 in relation to the number of ability bins

Item tested for fit					Remaining items		
Model	π_j	a_j	b_j	c_j	Model	$a_{v\neq j}$	$b_{v\neq j}$
1PL	0.5	1.7	0		1PL	1.7	$N(0,1)$
	0.9	1.7	-1.84		1PL	1.7	$N(0,1)$
2PL	0.5	1.7	0		2PL	$LN(\ln 1.7, 0.4)$	$N(0,1)$
	0.9	1.7	-1.84		2PL	$LN(\ln 1.7, 0.4)$	$N(0,1)$
3PL	0.3	1.7	1.65	0.2	2PL	$LN(\ln 1.7, 0.4)$	$N(0,1)$
	0.5	1.7	0.45	0.2	2PL	$LN(\ln 1.7, 0.4)$	$N(0,1)$
	0.9	1.7	-1.65	0.2	2PL	$LN(\ln 1.7, 0.4)$	$N(0,1)$

Note. LN = log-normal distribution.

The following analysis was performed for every replicated dataset:

1. $\chi^2_{w\,j}$ was computed with all item parameters fixed at values used for simulating them. Number of bins varied $r \in \{2, 3, \ldots, 10\}$. Assumed degrees of freedom were $df = r$. These results were used to infer about Type I error and empirical distribution of $\chi^2_{w\,j}$ under true H_0 in the "known parameters" case. The results are presented in Figures 1-7.

2. All item parameters were estimated according to an IRT models that were used for simulating the data, and then $\chi^2_{w\,j}$ was computed with varying number of bins, $r \in \{q+1, q+2, \ldots, 10\}$. Assumed degrees of freedom were $df = r - q$. These results provided information on performance of $\chi^2_{w\,j}$ under true H_0 in the "estimated parameters" case. The results are presented in Figures 8-13.

3. If item j was simulated from 3PLM the parameters were also estimated under simpler models: 2PLM and 1PLM. When the 1PLM was fit to data in this step, the remaining item responses were drawn with $a_{v\neq j} = 1.7$ and $b_{v\neq j} \sim N(0,1)$. $\chi^2_{w\,j}$ was computed with varying number of bins, $r \in \{q+1, q+2, \ldots, 10\}$, and $df = r - q$ was assumed. These results provided information about power of $\chi^2_{w\,j}$ in presence of misfit. The results are presented in Figure 14.

Item parameters of 1PLM and 2PLM were estimated without imposing prior distributions on them. When parameters of item j were estimated under 3PLM following priors were used to ensure convergence: $N(0,3)$ for b_j, $N(1.7,3)$ for a_j, and $\beta(1.7,3)$ for c_j. These prior distributions were substantially less informative than typically chosen in similar research. The choice was motivated by results of preliminary simulations, which indicated that more informative prior distributions adversely affected the assumed distribution of χ^2_w. It is also noteworthy, that in studies that used more informative priors for estimating 3PLM, some item-fit statistics, that are reported to follow the nominal distribution under other IRT models, produced elevated Type I error rates specifically in case of 3PLM (Glas & Suarez-Falcon, 2003; Sinharay 2006; Chalmers & Ng, 2017).

Inspection of Type I error rates and power of χ^2_w in relation to the number of ability bins under all tested conditions leads to following observations:

1. Type I error rate was significantly affected by the number of intervals, sample size and marginal item difficulty, and not so much by test length. Also, results for the known parameters and estimated parameters case followed similar patterns.

2. In general, at fixed sample size and test length, increase of r led to departure from the nominal significance level, $\alpha = 0.05$. This effect was most prominent if sample size was small, and item was easy. When sample size was $m = 4000$ and marginal difficulty was $\pi_j = 0.5$, Type I error did not exceed the nominal level at any r. However, for items with extreme π_j under $m = 400$ only the smallest r provided Type I errors close to nominal level in the known parameters case. In the estimated parameter case this was also true except for items generated under 3PLM. For easy and difficult 3PLM item Type I error visibly exceeded nominal level even for $r = 4$.

3. For large sample size it can be observed that increasing r does not improve the power of χ^2_w. Conversely, in some cases the power visibly decreased with finer partition of

ability. In small sample size the relation is more complex, however larger number of intervals should not be considered as a viable solution in small samples because they appear to be associated with unacceptably high Type I error rates. Only the case of medium difficulty items simulated under 3PLM and then fit with 1PLM provides evidence that using $r = 3$, instead of $r = 2$, is beneficial in terms of both power and Type I error rate.

Q-Q plots that compare empirical distribution of χ_w^2 with its theoretical distribution are also presented in Figures 1-13, along the Type I error plots. For the known parameters case (Figures 1-7) all Q-Q plots are obtained for $r = 2$. For the estimated parameters case (Figures 8-13) the number of ability bins chosen for Q-Q plots depended on the model under which the item was estimated ($r = 3$ for 2PLM, and $r = 4$ for 3PLM leading to a single degree of freedom), and on the marginal item difficulty if item was estimated in 1PLM ($r = 3$ for $\pi_j = 0.5$, and $r = 2$ for $\pi_j = 0.9$). These plots provide a more detailed verification of the postulated asymptotic distribution of χ_w^2 than Type I error rates. For $m = 4000$ the empirical distribution of χ_w^2 is well aligned with theoretical χ^2 both in the known and estimated parameters case. Approximation is also well-behaved for $m = 400$ and moderate item difficulty. However, combined conditions of small sample size and extreme item difficulties result in deviation of χ_w^2 from its theoretical asymptotic distribution.

Figure 1

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 1PLM, $\pi_j = 0.5$, known parameters case

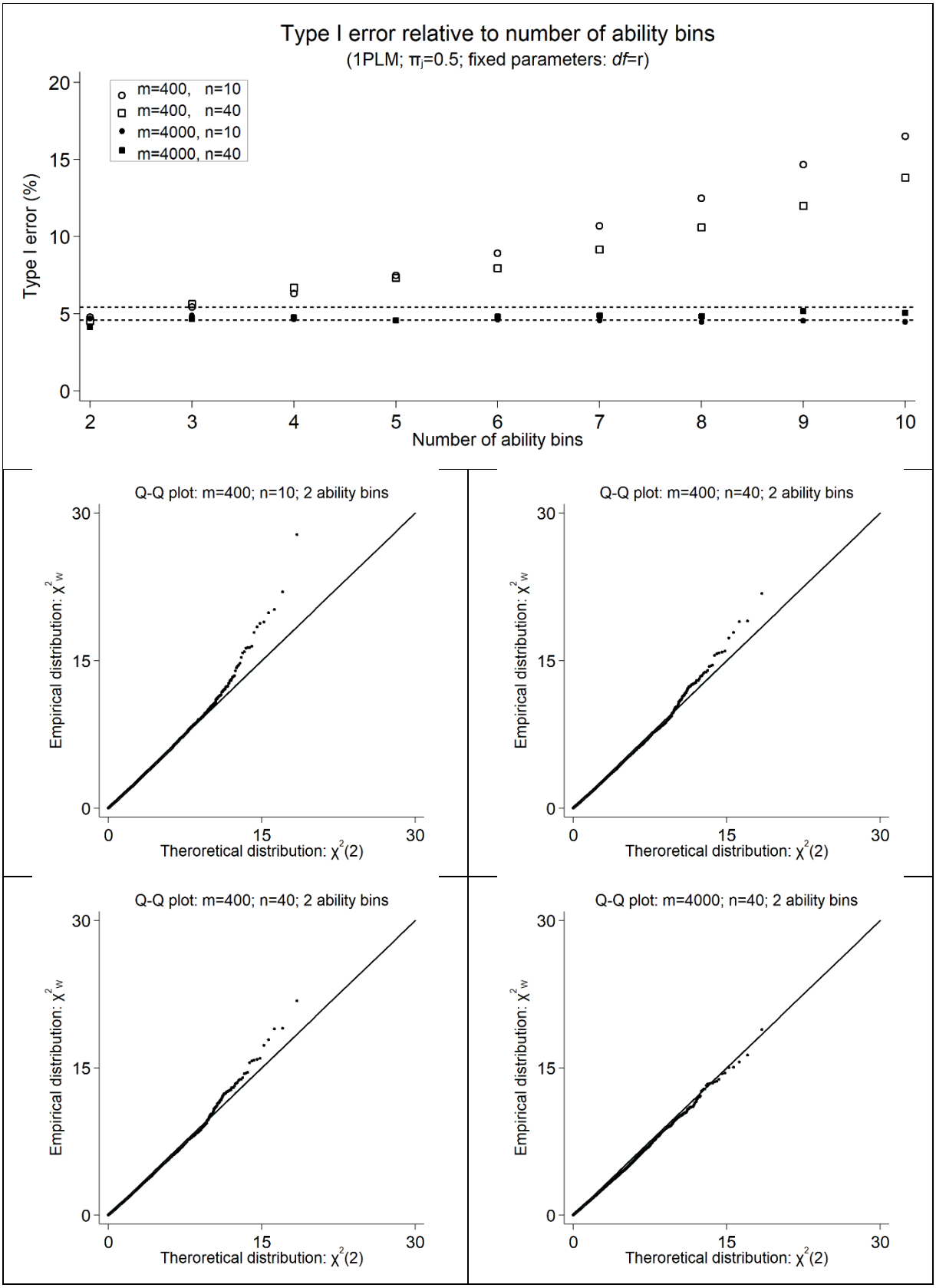


Figure 2

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 1PLM, $\pi_j = 0.9$, known parameters case

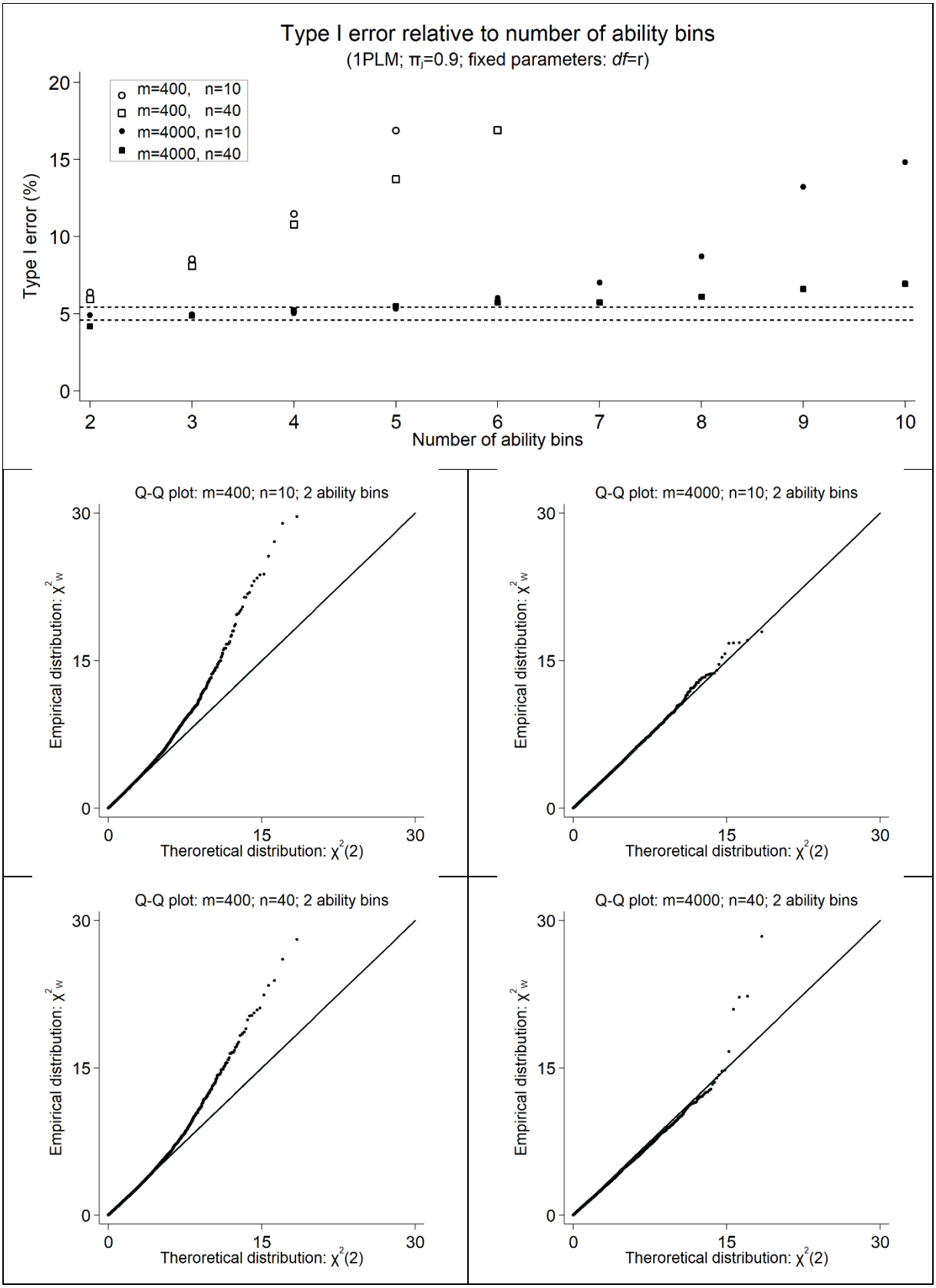


Figure 3

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 2PLM, $\pi_j = 0.5$, known parameters case

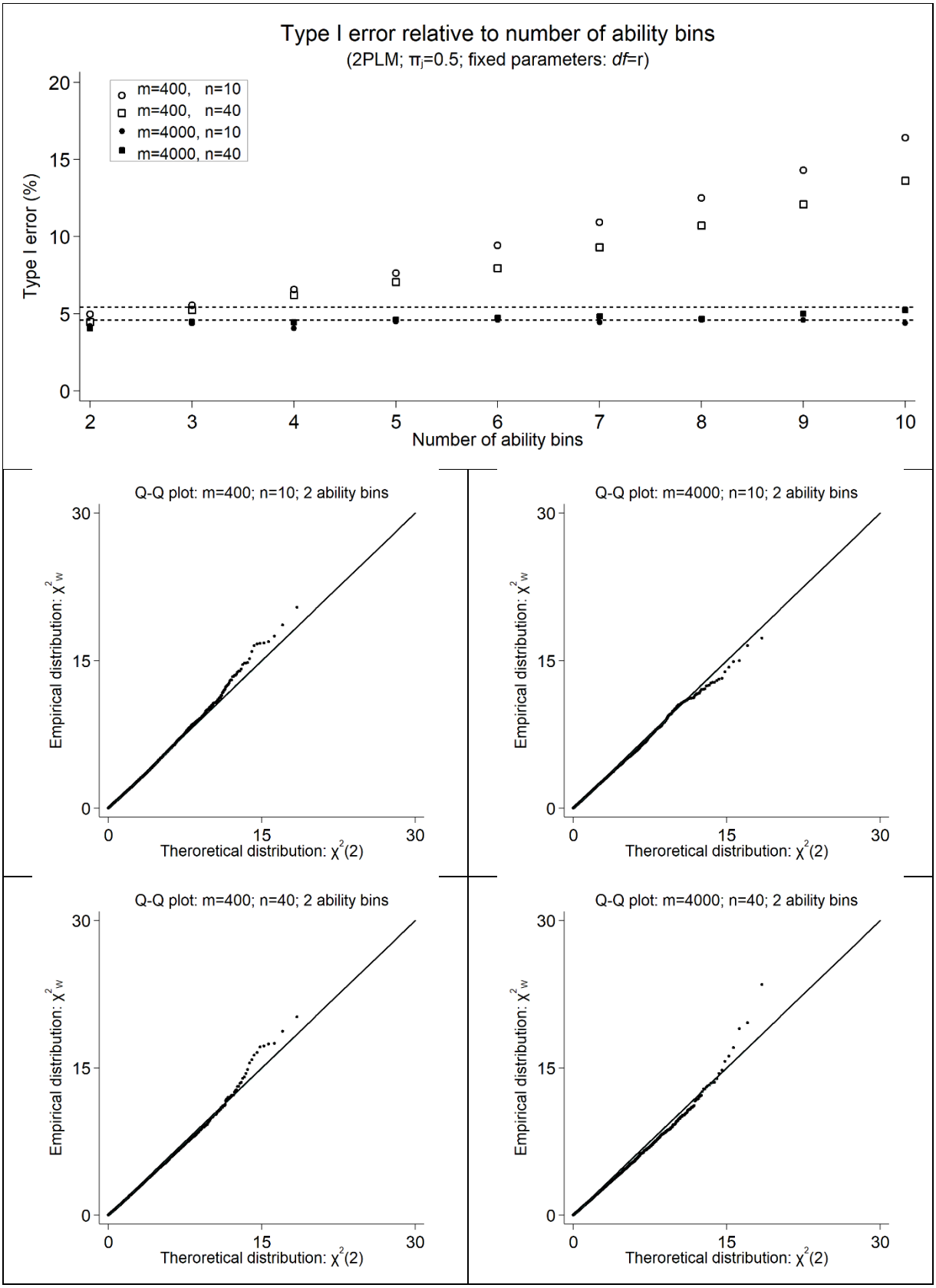


Figure 4

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 2PLM, $\pi_j = 0.9$, known parameters case

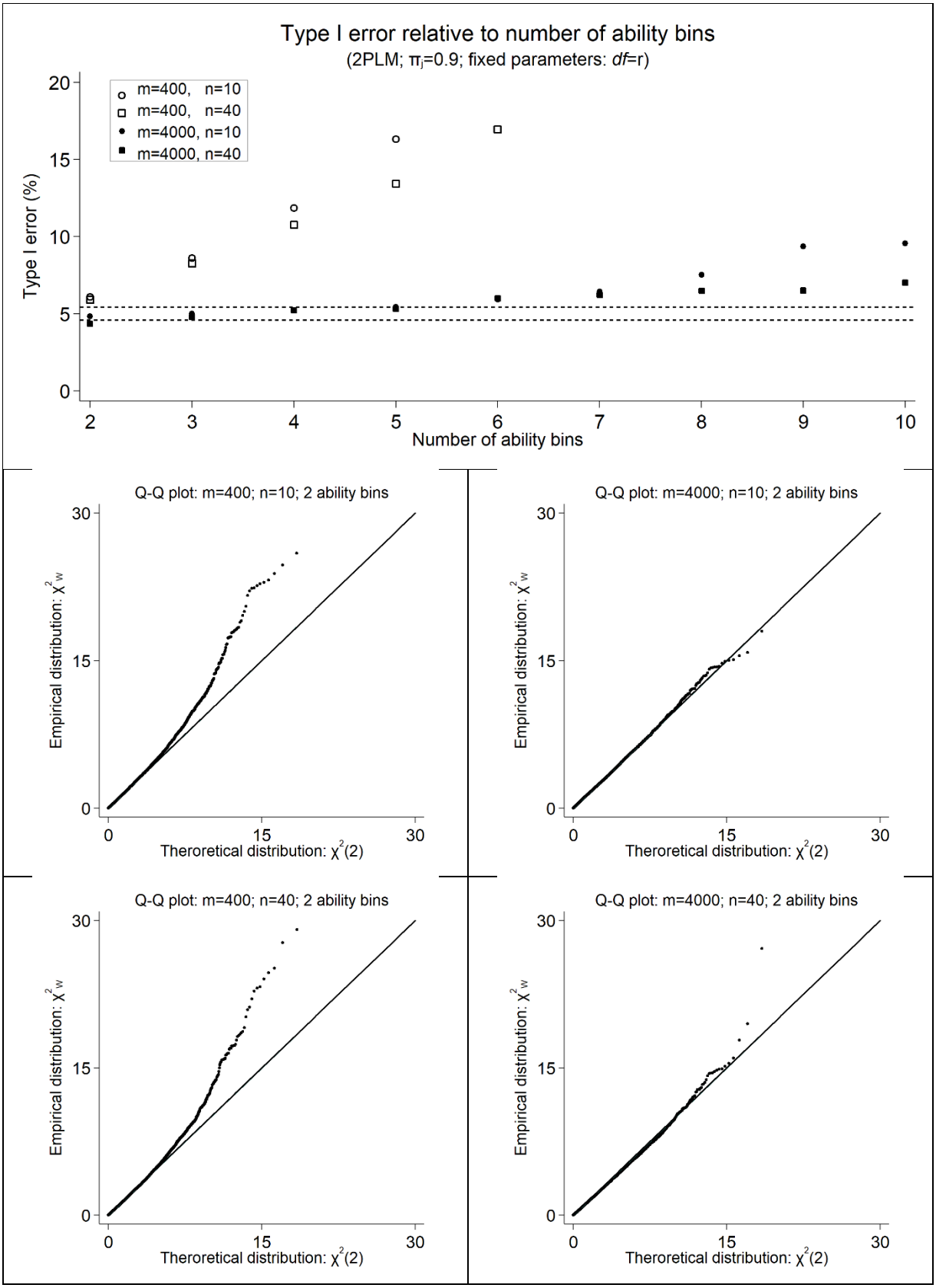


Figure 5

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 3PLM, $\pi_j = 0.3$, known parameters case

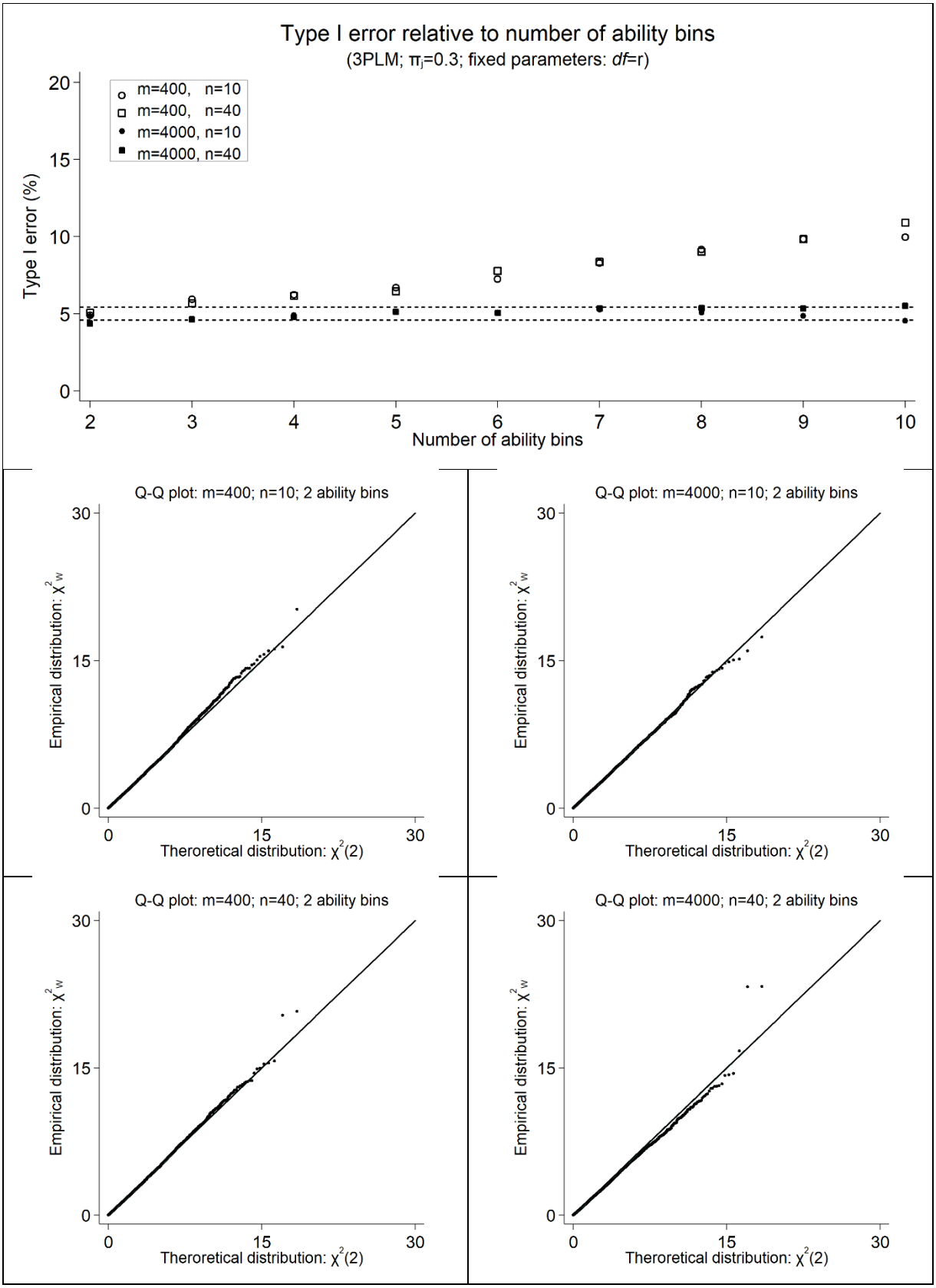


Figure 6

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 3PLM, $\pi_j = 0.5$, known parameters case

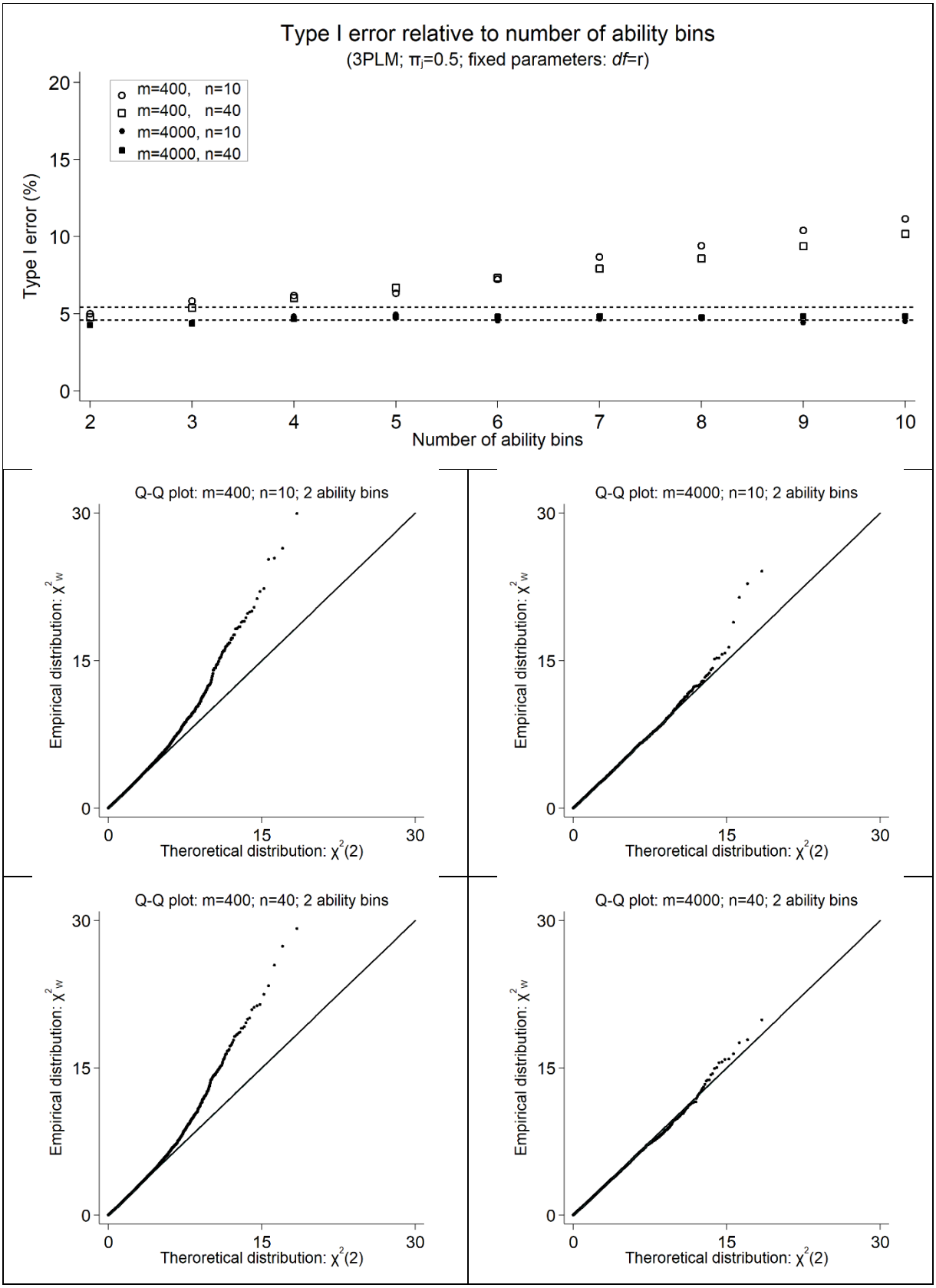


Figure 7

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 3PLM, $\pi_j = 0.9$, known parameters case

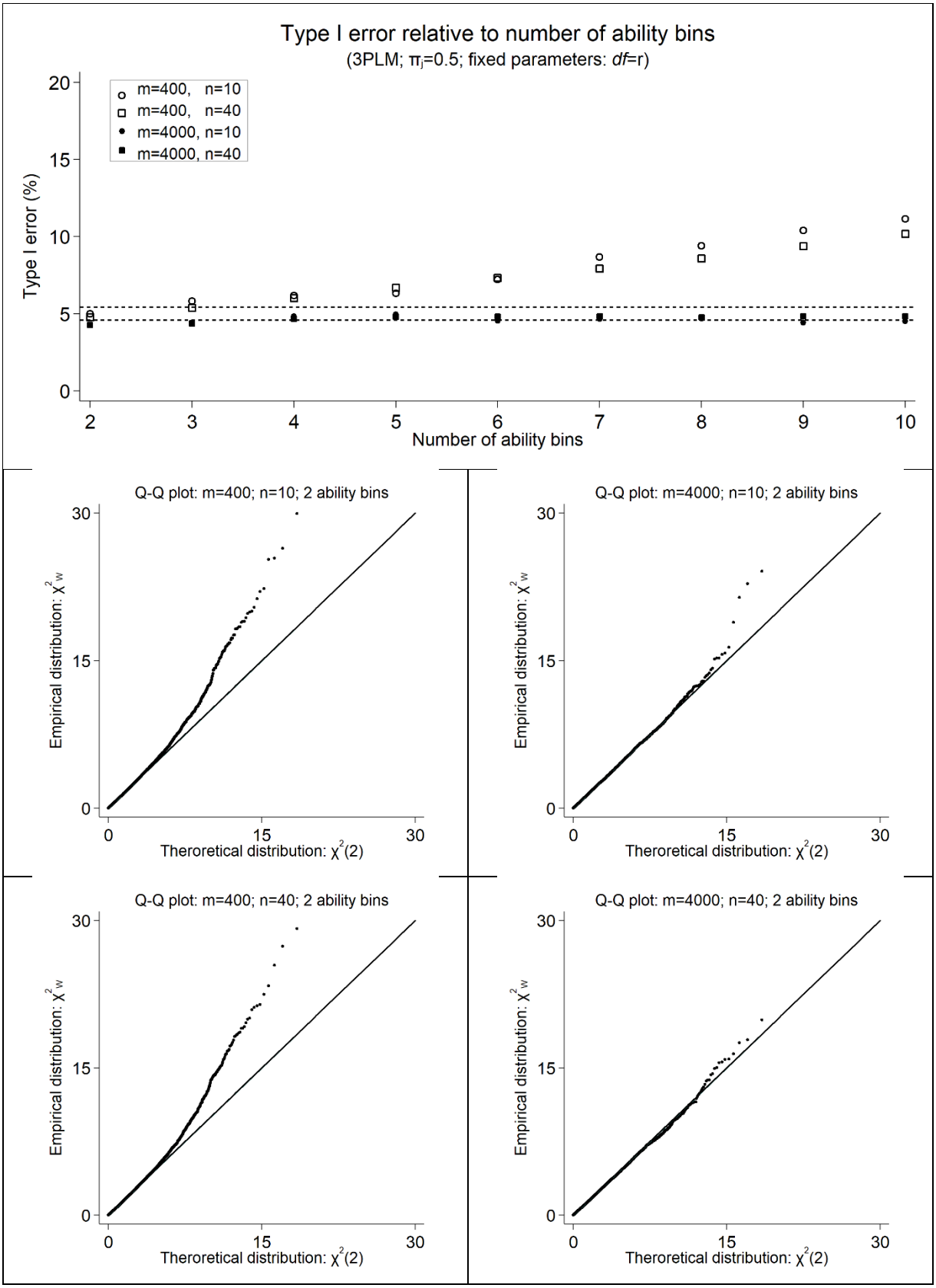


Figure 8

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 3 ability bins; 1PLM, $\pi_j = 0.5$, estimated parameters case

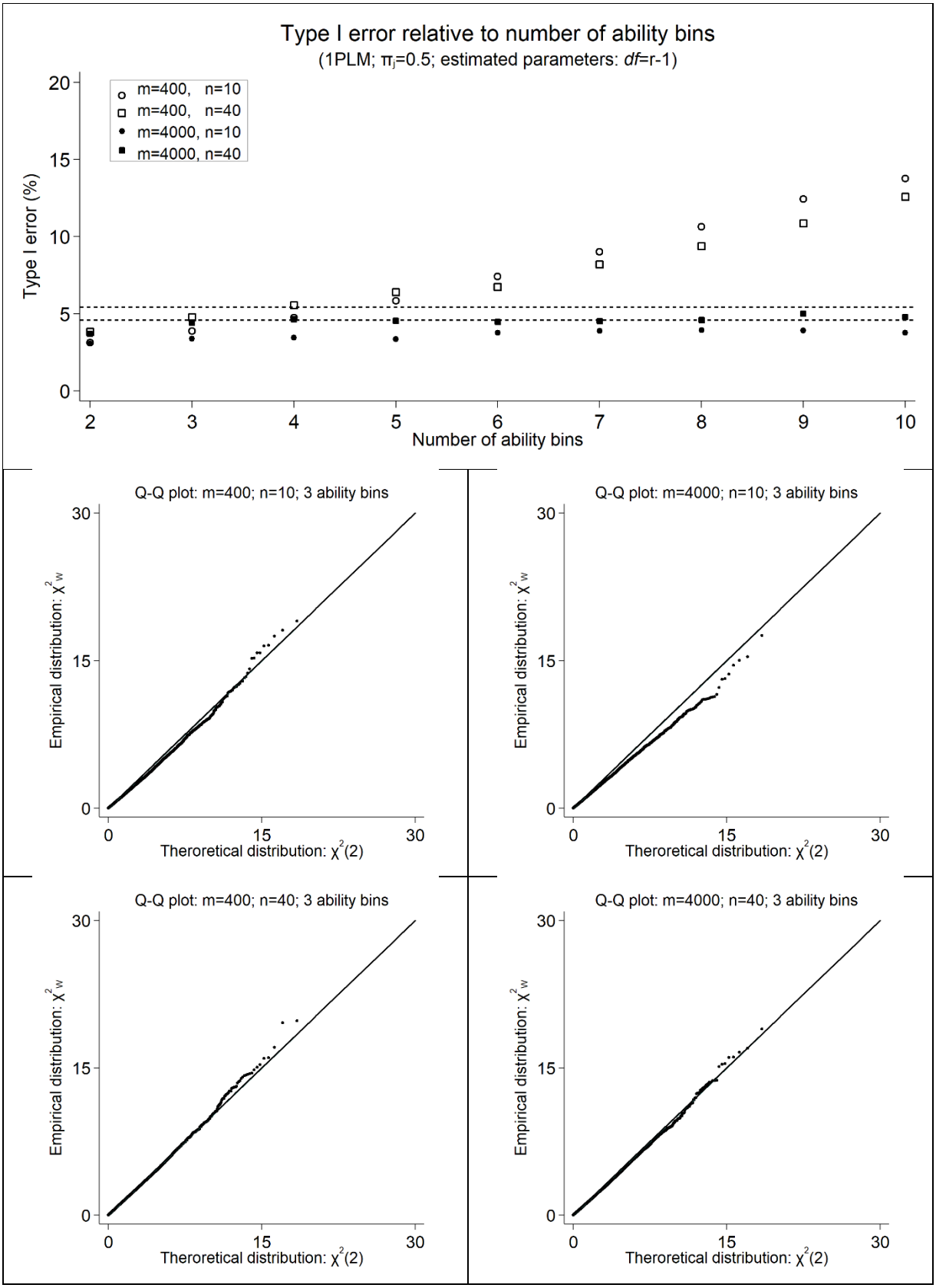


Figure 9

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 2 ability bins; 1PLM, $\pi_j = 0.9$, estimated parameters case

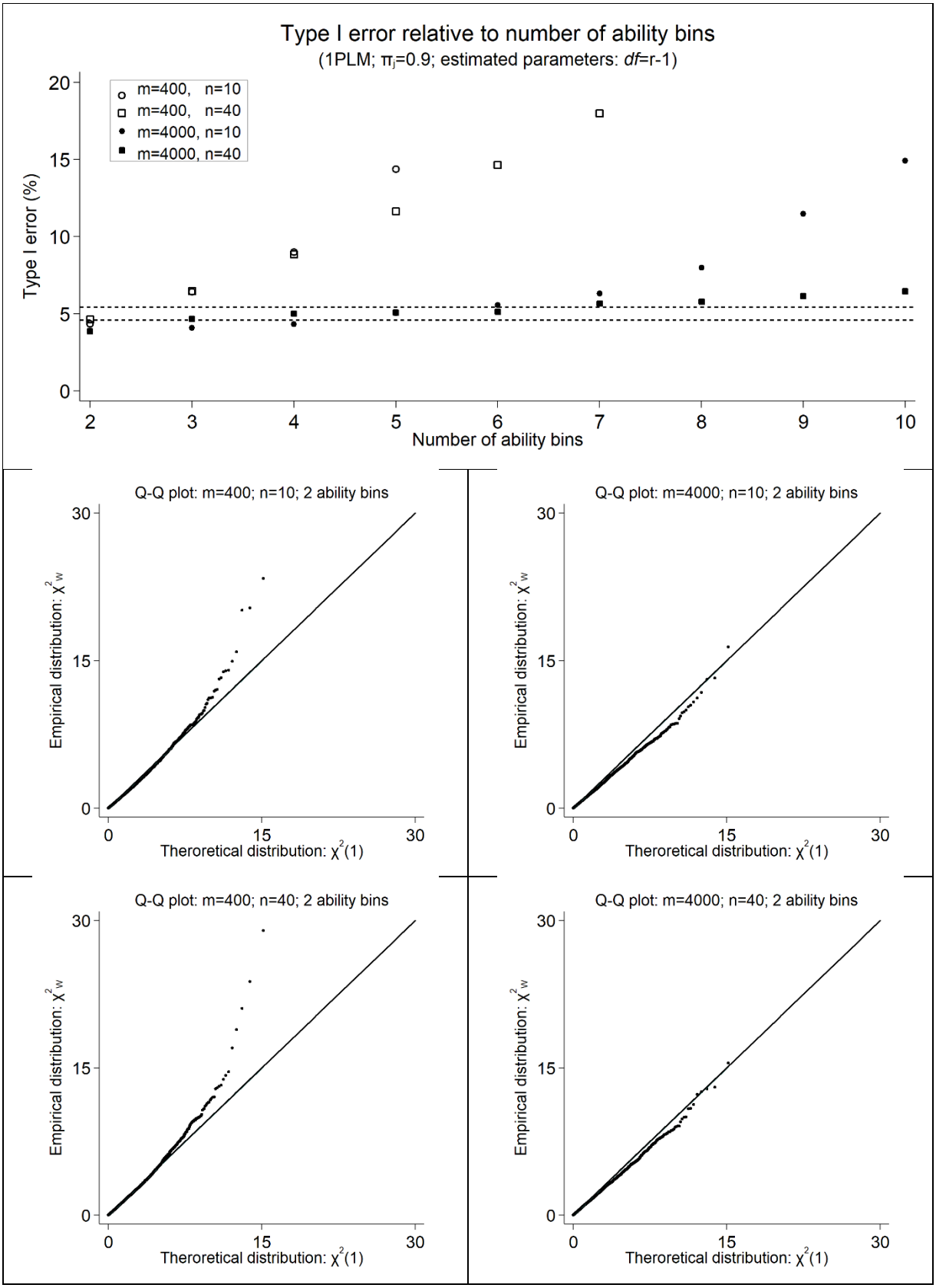


Figure 10

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 3 ability bins; 2PLM, $\pi_j = 0.5$, estimated parameters case

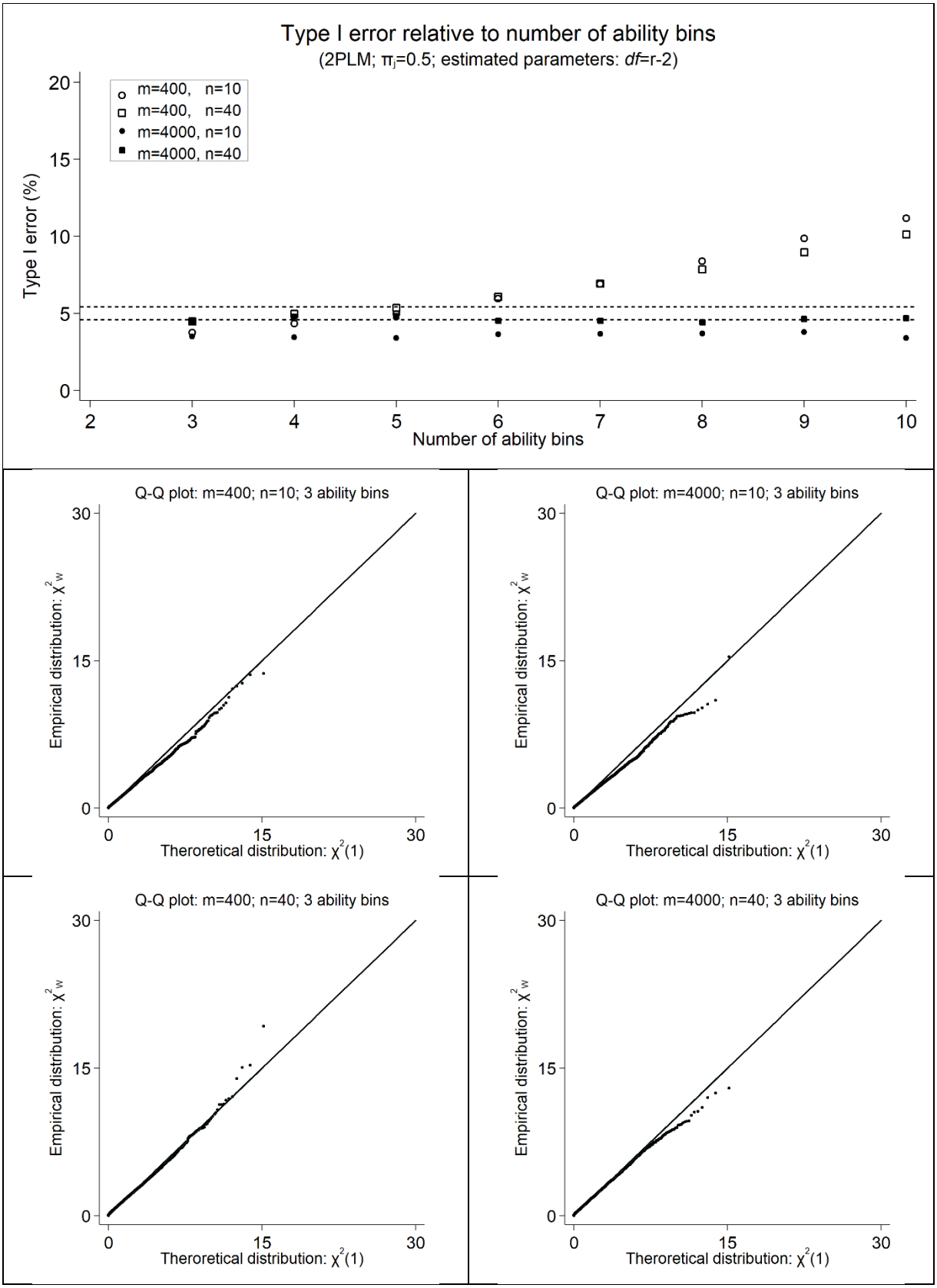


Figure 11

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 3 ability bins; 2PLM, $\pi_j = 0.9$, estimated parameters case

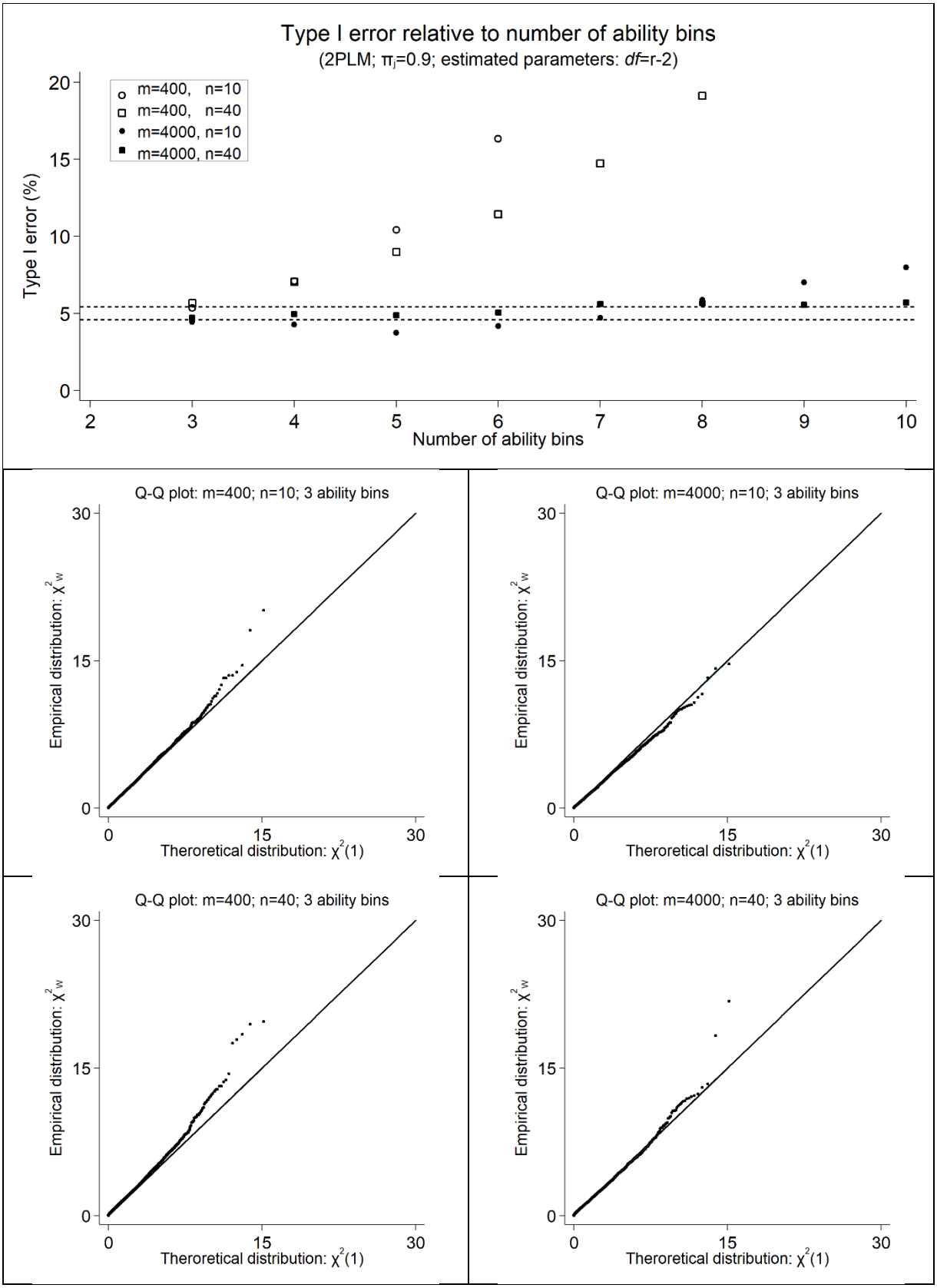


Figure 12

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 4 ability bins; 3PLM, $\pi_j = 0.3$, estimated parameters case

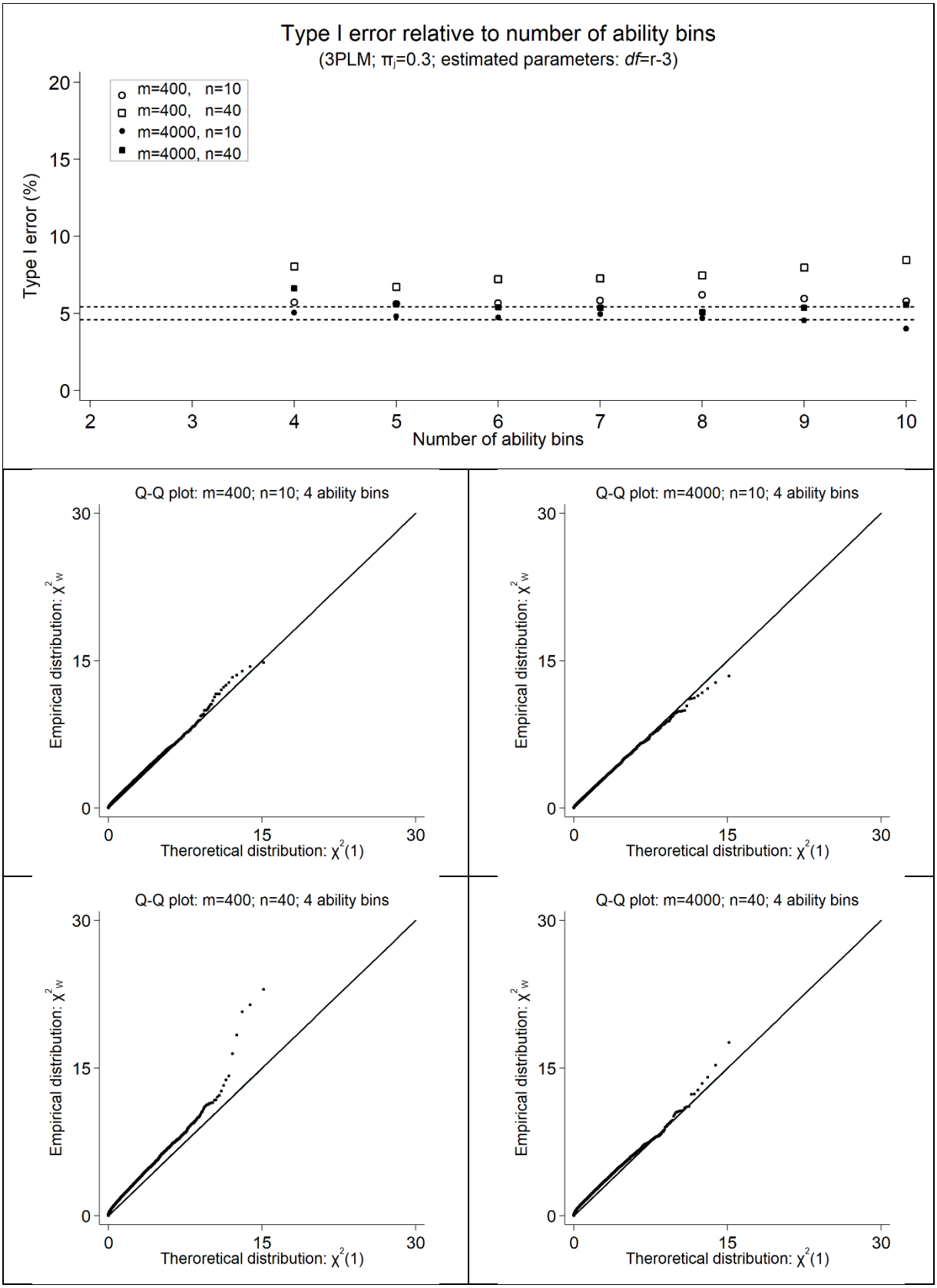


Figure 12

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 4 ability bins; 3PLM, $\pi_j = 0.5$, estimated parameters case

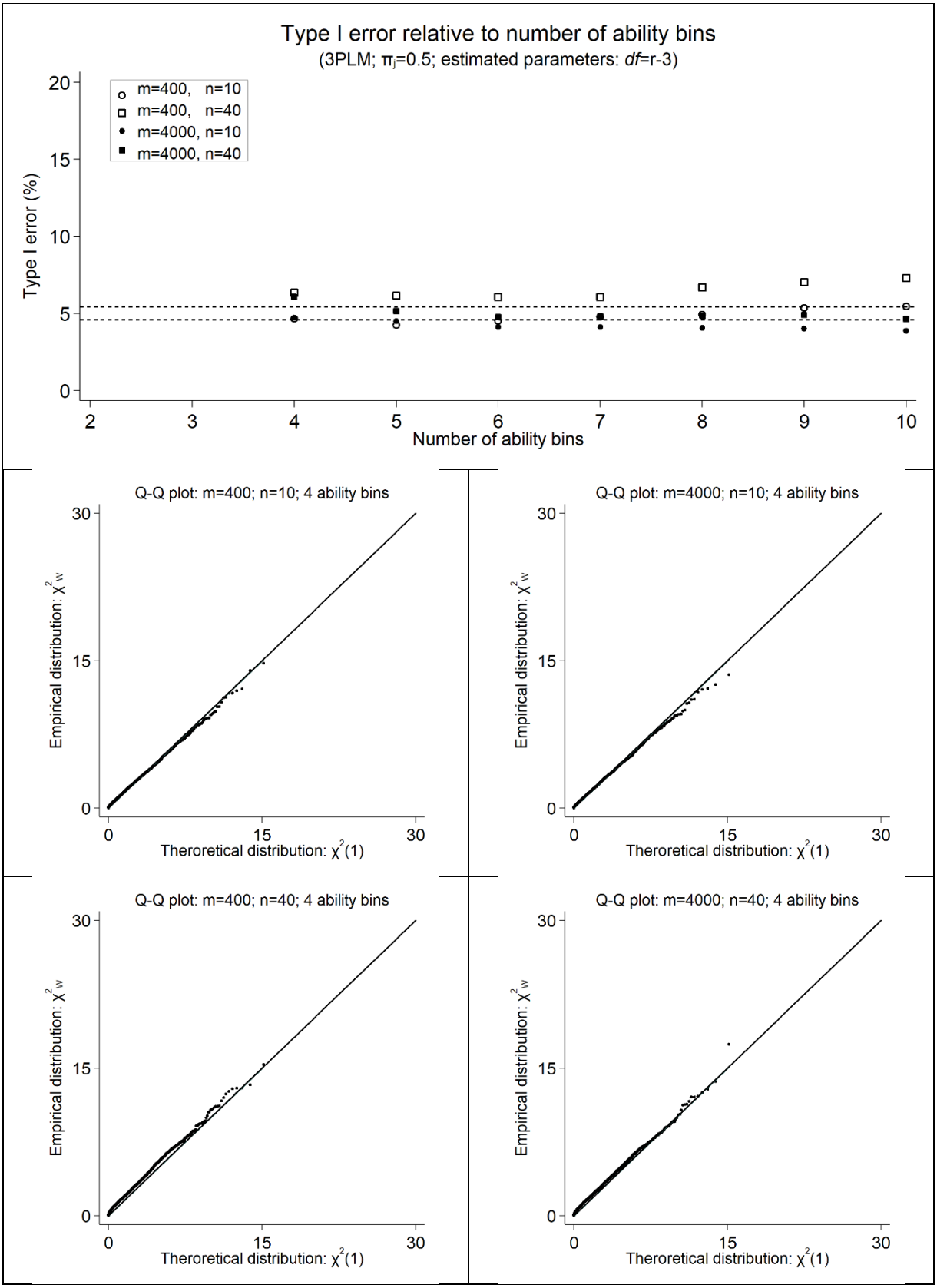


Figure 13

Type I error rates of $\chi^2_{w\,j}$ in relation to the number of ability bins and Q-Q plots for 4 ability bins; 3PLM, $\pi_j = 0.9$, estimated parameters case

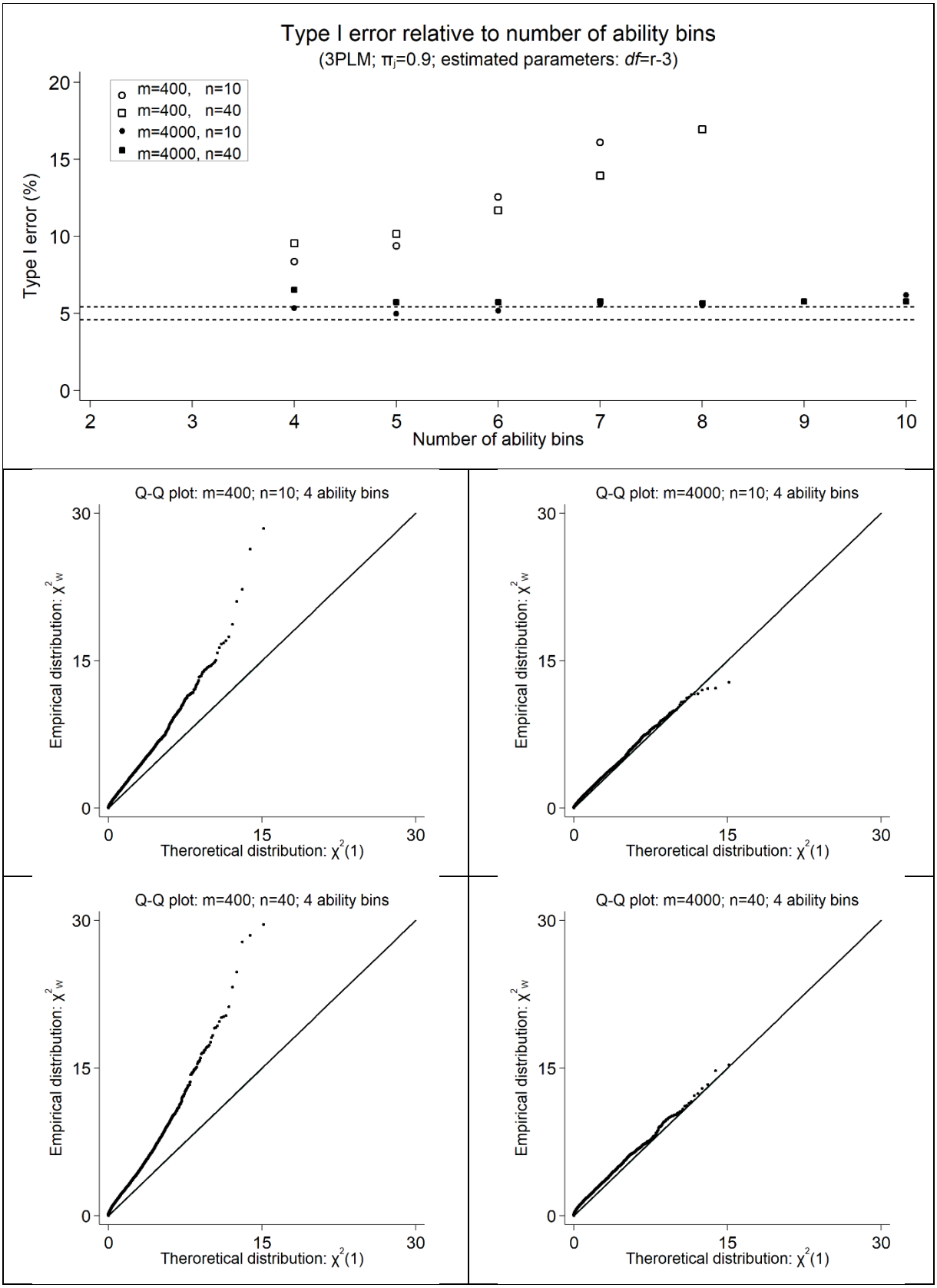
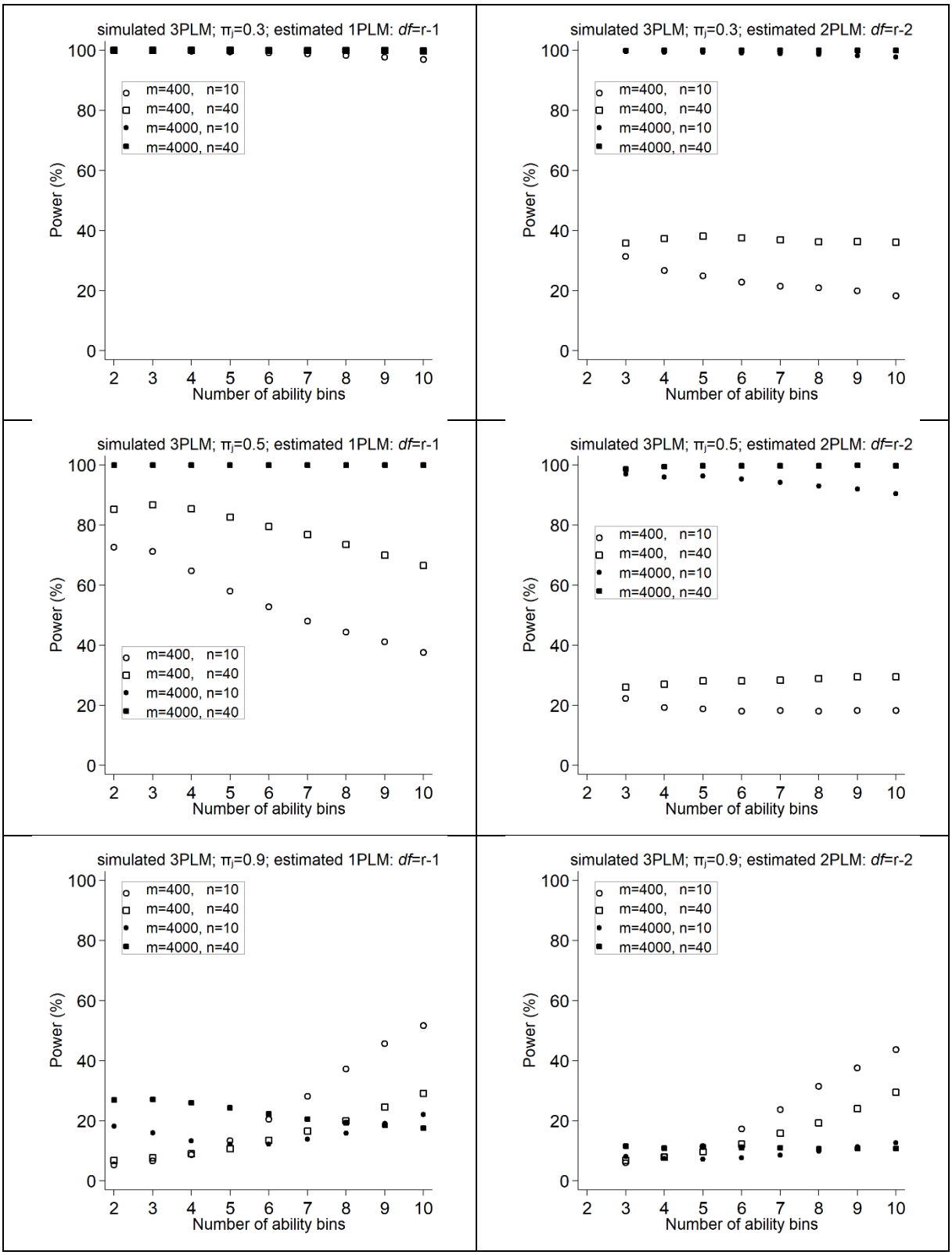


Figure 14

Power of $\chi^2_{w\,j}$ in relation to the number of ability bins



References

Chalmers, R. P., & Ng, V. (2017). Plausible-Value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, 41(5), 372–387.

Glas, C. A. W., & Suarez-Falcon, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, 27(2), 87–106.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, 59, 429–449.