A Modified Tucker's Congruence Coefficient for Factor Matching

Electronic Supplementary Materials

An overview of the Electronic Supplementary materials:

- A – Toy examples of TCC, mTCC and Pearson CC
- B – Details on the simulations and R code
- C – Alternative Figure 4 of the main body of the paper
- D – Simulations with different random seeds
- E – Simulations with primary loadings 0.80 (E.1), 0.60 (E.2) and 0.40 (E.3)
- F – Simulations with 2, 4 (F.1), 10 and 16 (F.2) items for 2 factors
- G – Simulations with mean cross-loading of 0 (G.1) and 0.20 (G.2)
- H – Simulation results comparing the TCC-mTCC-PCC triad
- I – Histograms of concordance of TCCs and mTCCs in the simulations

A. Toy examples comparing the Tucker's and modified congruence coefficients and the Pearson correlation coefficient

Toy examples presenting different settings where each and every one of the coefficients can be misleading if not used with caution:

Set A: For matching two factors where the factor structure is not clear, the mTCC is misleading as is the non-negligible PCC.

$a1 = \{ -0.39, \; -0.83, \; -0.79, \quad 0.66, \; -0.39, \; -0.75, \; -0.22, \quad 0.80\}$

$a2 = \{ \quad 0.77, \quad 0.77, \; -0.46, \; -0.04, \; -0.10, \; -0.52, \quad 0.57, \quad 0.81\}$

TCC=0.118   mTCC=0.854   PCC=0.307

Set B: If there are no negative loadings at all, TCC and mTCC are identically high (and misleading in matching two non-simple factors. PCC is low.

$b1 = \{0.39, \quad 0.83, \quad 0.79, \quad 0.66, \quad 0.39, \quad 0.75, \quad 0.22, \quad 0.80\}$

$b2 = \{0.77, \quad 0.77, \quad 0.46, \quad 0.04, \quad 0.10, \quad 0.52, \quad 0.57, \quad 0.81\}$

TCC=0.854   mTCC=0.854   PCC=0.189

Set C: While PCC show a very high negative linear association, both TCC and mTCC show the two factors to be incongruent.

$c1 = \{ \quad 0.80, \quad 0.70, \quad 0.66, \quad 0.49, \quad 0.12, \quad 0.08, \quad 0.01, \quad 0.05\}$

$c2 = \{ \quad 0.07, \quad 0.07, \quad 0.26, \quad 0.14, \quad 0.70, \quad 0.82, \quad 0.64, \quad 0.81\}$

TCC=0.263   mTCC=0.263   PCC=-0.945

Set D: All coefficients indicate incongruence.

$d1 = \{ -0.80, \quad 0.70, \; -0.66, \quad 0.49, \; -0.12, \quad 0.08, \; -0.01, \quad 0.05\}$

$d2 = \{ \quad 0.07, \quad 0.07, \quad 0.26, \quad 0.14, \quad 0.70, \quad 0.82, \quad 0.64, \quad 0.81\}$

TCC=-0.046   mTCC=0.263   PCC=0.020

Set E: When all loadings are identical, all coefficients indicate perfect similarity.

$e1 = \{-0.39, -0.83, -0.79, \quad 0.66, -0.39, -0.75, -0.22, \quad 0.80\}$

$e2 = \{-0.39, -0.83, -0.79, \quad 0.66, -0.39, -0.75, -0.22, \quad 0.80\}$

TCC=1   mTCC=1   PCC=1

Sets F and G: When the magnitude of the loadings is the same, but the signs are occasionnaly different, the mTCC indicates congruence while TCC and PCC indicate incongruence even though these factors should likely be matched.

$f1 = \{ -0.39, \; -0.83, \; -0.79, \quad 0.66, \; -0.39, \; -0.75, \; -0.02, \quad 0.80\}$

$f2 = \{ \quad 0.39, \quad 0.83, \quad 0.79, \quad 0.66, \quad 0.39, \quad 0.75, \quad 0.02, \quad 0.80\}$

TCC=-0.3399   mTCC=1   PCC=-0.089

$g1 = \{ -0.39, \quad 0.83, \quad 0.79, \quad 0.66, \quad 0.39, \quad 0.75, \; -0.02, \quad 0.80\}$

$g2 = \{ \quad 0.39, \quad 0.83, \quad 0.79, \quad 0.66, \quad 0.39, \quad 0.75, \quad 0.02, \quad 0.80\}$

TCC=0.906   mTCC=1   PCC=0.906

Set H: All three coefficients are misleading for factor matching when $F_1 = k * F_2, \quad k \in \mathbb{N}$.

$h1 = \{ \; -0.39,, \; -0.83, \; -0.79, \quad 0.66, \; -0.39, \; -0.75, \quad -0.22, \; 0.80\}$

$h2 = \{-0.039, \; -0.083, \; -0.079, \quad 0.066, \; -0.039, \; -0.075, \; -0.022, \; 0.080\}$

TCC=1   mTCC=1   PCC=1

B. Details on the simulations and R code

Below the code for the simulations is provided. This is the code for a 2 factor, 2 item setting for mean 0.10 and standard deviation 0.025 for the cross-loadings. Primary loadings are 0.80 and 0.70 for the first factor and 0.70 and 0.60 for the second factor.

```
#Generate 100*100 factor loading matrices where primary loadings are
    predefined and cross-loadings are normally distributed with mean and
    sd defined below
set.seed(2604)
factor=2
item=2
mean=0.10
sd=0.025
flT<-matrix(NA,4,factor*0000)
for (i in 1:10000) {
  fl <- matrix(round(rnorm(factor*item*factor, mean, sd), 3),factor*item,
     factor)
  fl[fl < -0.999] <- -0.999
  fl[fl > 0.999] <- 0.999
  fl[1,1]=0.8
  fl[2,1]=0.7
  fl[3,2]=0.7
  fl[4,2]=0.6
  flT[,(1+(i-1)*2):(2+(i-1)*2)]<-fl
}

#Identifies "datasets" with factor loadings
data = flT
data.list=NULL
list.idx=seq(1,20000,2)
for (i in 1:10000){
  data.list[[i]]=data[,list.idx[i]:(list.idx[i]+1)]
}

data1=data.list[[1]]
data2=data.list[[2]]
k=1

#define CC functions
TCC <- function(x,y){
  out=sum(x*y)/ (sqrt(sum(x^2)*sum(y^2)))
  return(out)
}

mTCC <- function(x,y){
  out=sum(abs(x*y))/ (sqrt(sum(x^2)*sum(y^2)))
  return(out)
}
```

```
#Function to calculate CCs for two datasets and find location of maxima (per row!)
temp=matrix(0, dim(data2)[2],4*dim(data1)[2])
ccc <- function(data1,data2){
  tcc.coef=matrix(0,dim(data2)[2],dim(data1)[2])
  for (i in 1:(dim(data1)[2])){
    for (j in 1:(dim(data2)[2])){
      tcc.coef[i,j]=TCC(data1[,i],data2[,j])
    }
  }
  mtcc.coef=matrix(0,dim(data2)[2],dim(data1)[2])
  for (k in 1:(dim(data1)[2])){
    for (l in 1:(dim(data2)[2])){
      mtcc.coef[k,l]=mTCC(data1[,k],data2[,l])
    }
  }
  max.value=apply(tcc.coef,2,max)
  max.location = matrix(0,dim(tcc.coef)[1],dim(tcc.coef)[2])
  for (i in 1:dim(tcc.coef)[1]){
    max.loc=which(tcc.coef[,i]==max.value[i])
    max.location[max.loc,i]=1
  }
  max.value.mtcc=apply(mtcc.coef,2,max)
  max.location.mtcc = matrix(0,dim(mtcc.coef)[1],dim(mtcc.coef)[2])
  for (i in 1:dim(mtcc.coef)[1]){
    max.loc=which(mtcc.coef[,i]==max.value.mtcc[i])
    max.location.mtcc[max.loc,i]=1
  }
  temp=cbind(tcc.coef, max.location, mtcc.coef, max.location.mtcc)
  return(temp)
}

#Calculates CC functions for each pair of factor analyses
myres=temp
myres.final=NULL
allres <- function(data.list){
  myres=NULL
  i=1
  j=i
  temp <- ccc(data.list[[i]], data.list[[j]])
  myres <-cbind(rep(i,dim(temp)[1]),rep(j,dim(temp)[1]),temp)
  for (i in 1:100 ){
    for (j in i:100){
      if (i < j) {
        temp <- ccc(data.list[[i]], data.list[[j]])
        myres.final <-rbind(myres.final, cbind(rep(i,dim(temp)[1]),rep(j, dim(temp)[1]),temp)) }
    }
  }
  return(myres.final)
}

out.allres=allres(data.list)


#Calculates the number of not wel-matched factors
count.mismatch <- function(out.allres){
  zero.ind1=rep(0,dim(out.allres)[1])
  zero.ind2=rep(0,dim(out.allres)[1])

  for (i in 1:dim(out.allres)[1]){
    zero.ind1[i]=sum(out.allres[i,5:6])
    zero.ind2[i]=sum(out.allres[i,9:10])
  }
  tcc.mismatch=sum(zero.ind1==0)
  mtcc.mismatch=sum(zero.ind2==0)
  common=length(intersect(which(zero.ind2==0),which(zero.ind1==0)))
  return(cbind(tcc.mismatch, mtcc.mismatch, common))
}

count.mismatch(out.allres)

#Calculate result for all simulations
sim=NULL
dataset=NULL
ki=NULL
idx=matrix(1:10000,100,100)
output.allres=NULL
output.count=matrix(0,100,4)
ki=matrix(0,100,8)
colnames(ki)=c('Replication_id','tcc.mismatch', 'mtcc.mismatch', 'common'
    , 'factor', 'item','mean','sd')
for (i in 1:100){
  dataset[[i]]=data.list[idx[,i]]
  output.allres[[i]]=allres(dataset[[i]])
  output.count[i,]=c(i,count.mismatch(output.allres[[i]]))
  ki=c(output.count[i,],factor,item,mean,sd)
  sim=rbind(sim, ki)
}
```

B.2 The exact primary loadings for the main simulation

The primary loadings and random seeds used for the simulation in the main body of the paper:

2 factors, 2 items – F1: 0.8, 0.7; F2: 0.7, 0.6 – seed: 2604
2 factors, 10 items – F1: 0.8 (2), 0.7 (4), 0.6 (2); F2: 0.7 (2), 0.6 (4), 0,5 (4) – seed: 2605
5 factors, 2 items – F1: 0.9, 0.8; F2: 0.8, 0.7; F3: 0.7 (2); F4: 0.7, 0.6; F5: 0.6 (2) – seed: 2606
5 factors, 10 items – F1: 0.8 (2), 0.7 (4), 0.6 (4); F2: 0.8, 0.7, 0.6 (4), 0.5 (4); F3: 0.7, 0.6 (4), 0.5 (5), 0.4; F4: 0.6 (4), 0.5 (3), 0.4 (3); F5: 0.6 (4), 0.5 (3), 0.4 (3) – seed: 2607

B.3 Random seeds for the simulations in the ESM

For the different settings, different seeds were used. To ensure reproducibility these are given in Table 1. Each setting wsa given a different random seed as simulations were run in parallel.

Table 1: *Seeds used for the simulations*

Factor	Item	Mean (CL)	SD (CL)	PL	Seed	Results presented in
2	2	0.10	all	Various (see above)	2604	Main paper
2	10	0.10	all	Various (see above)	2605	Main paper
5	2	0.10	all	Various (see above)	2606	Main paper
5	10	0.10	all	Various (see above)	2607	Main paper
2	2	0.10	all	0.80	1324	ESM D
2	2	0.10	all	0.80	6132	ESM D
2	2	0.10	all	0.80	8124	ESM D
2	2	0.10	all	0.80	9385	ESM D
2	2	0.10	all	0.80	2608	ESM E.1
2	10	0.10	all	0.80	2614	ESM E.1
5	2	0.10	all	0.80	2620	ESM E.1
2	2	0.10	all	0.60	2609	ESM E.2
2	10	0.10	all	0.60	2615	ESM E.2
5	2	0.10	all	0.60	2621	ESM E.2
2	2	0.10	all	0.40	2610	ESM E.3
2	10	0.20	all	0.40	2616	ESM E.3
5	2	0.10	all	0.40	2622	ESM E.3
2	4	0.10	all	0.80	2632	ESM F.1
2	4	0.10	all	0.60	2633	ESM F.1
2	4	0.10	all	0.40	2634	ESM F.1
2	16	0.10	all	0.80	2638	ESM F.2
2	16	0.10	all	0.60	2639	ESM F.2
2	16	0.10	all	0.40	2640	ESM F.2
2	2	0	all	0.80	2611	ESM G.1
2	2	0	all	0.60	2612	ESM G.1
2	2	0	all	0.40	2613	ESM G.1
2	10	0	all	0.80	2617	ESM G.1
2	10	0	all	0.60	2618	ESM G.1
2	10	0	all	0.40	2619	ESM G.1
2	2	0.20	all	0.80	2644	ESM G.2
2	2	0.20	all	0.60	2645	ESM G.2
2	2	0.20	all	0.40	2646	ESM G.2
2	10	0.20	all	0.80	2650	ESM G.2
2	10	0.20	all	0.60	2651	ESM G.2
2	10	0.20	all	0.40	2652	ESM G.2

C. Alternative figure of the simulation in the main body of the paper

Original Figure (Figure 1) with extended x axis (SD ranging from 0 to 0.40). Since, values below 0.15 are all nil, Figures 4a-4c have an abscissa starting at 0.15. However, this figure may offer a better understanding of how the coefficients change depending on the cross-loadings in general.

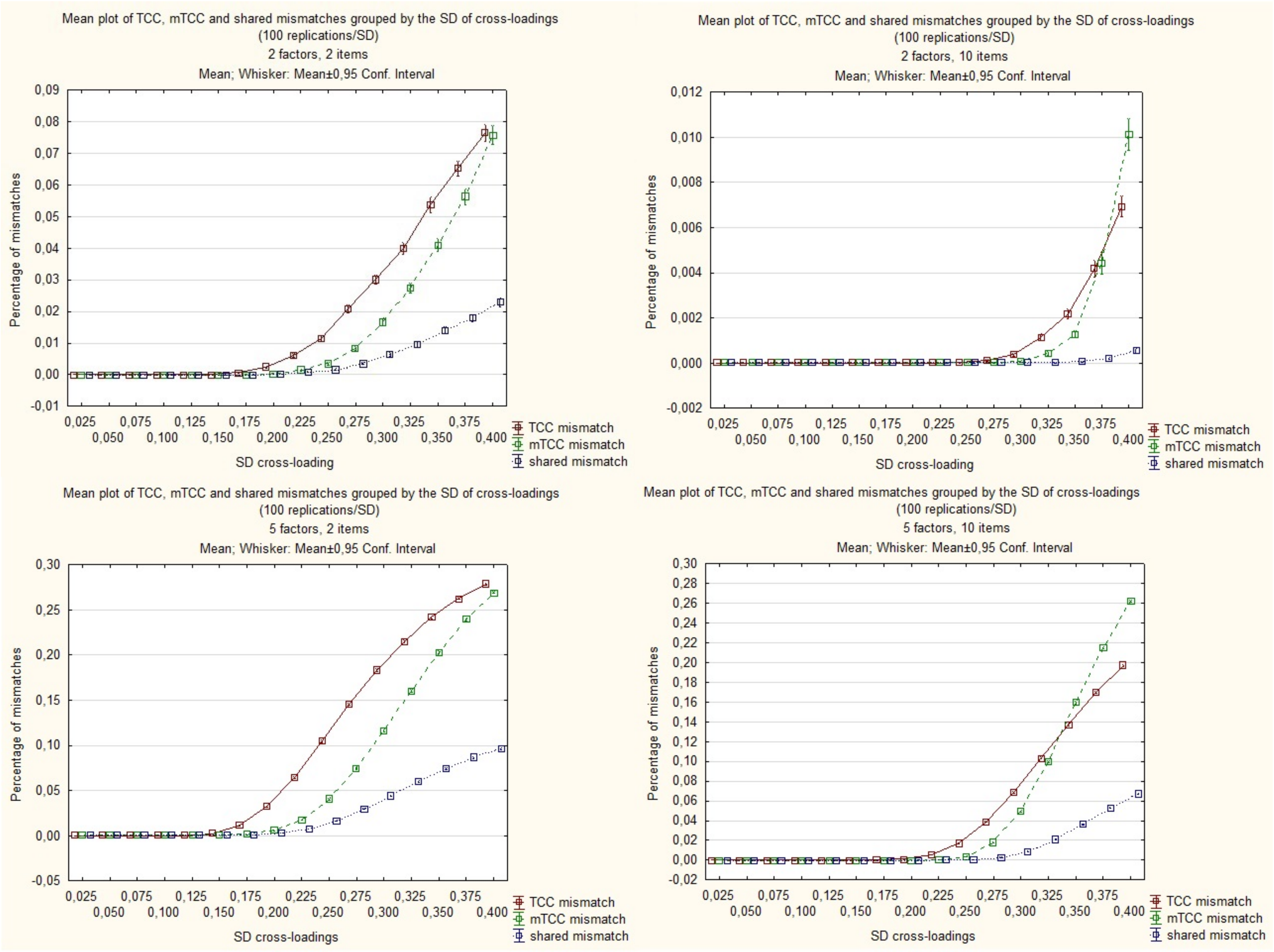

Figure 1: Alternative figure 4

D. Simulations with different random seeds

The setting is identical to what is presented in Figure 4a in the main body of the paper: various primary loadings (as listed in ESM B.2), 2 factors, 2 items. As it can be seen, the choice of the random seeds does not influence the results to an extent that could affect the interpretation of the results.

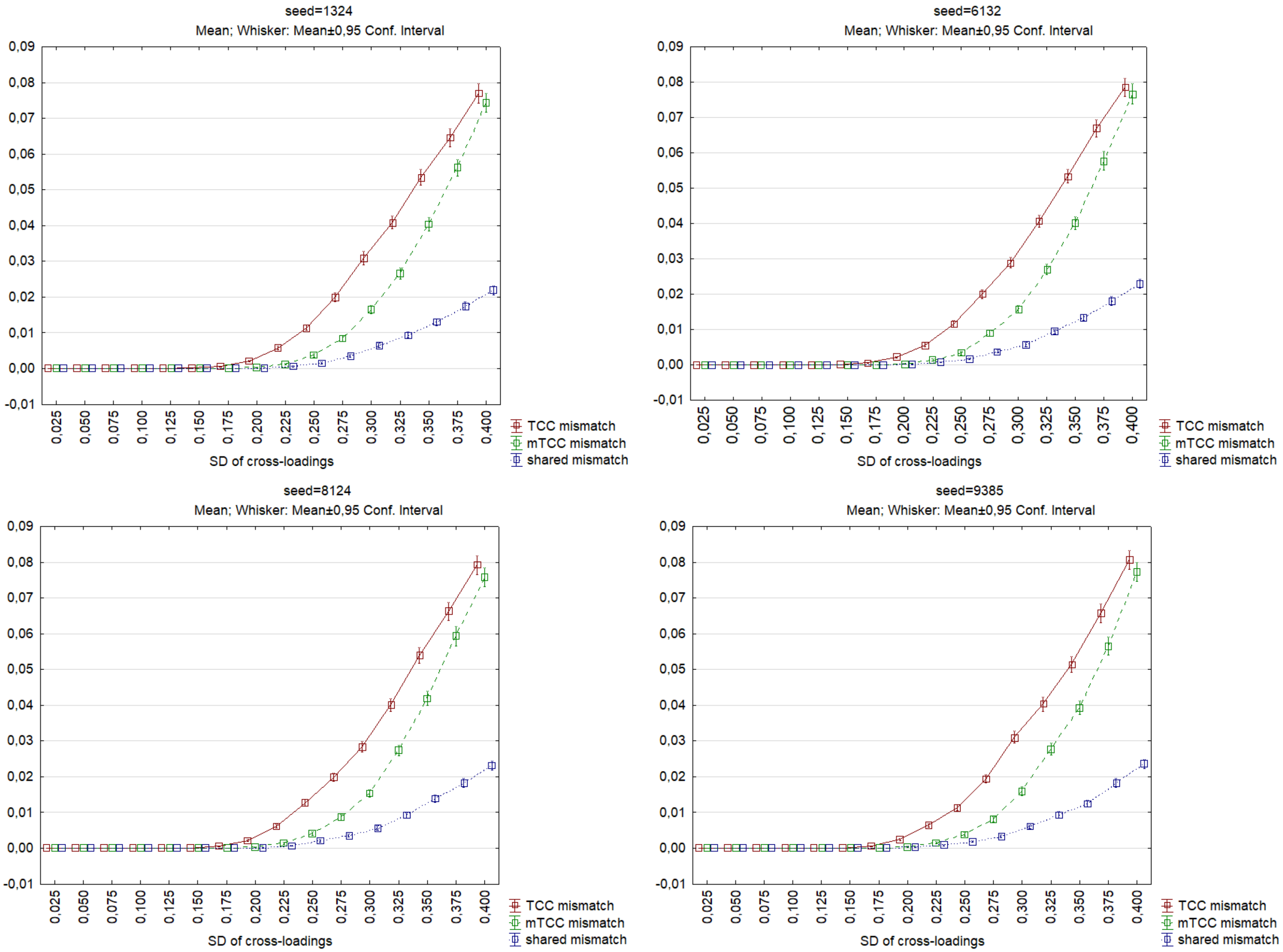

Figure 2: Simulation with different random seeds

E. Simulations with different primary loadings

The simulation in this section focus on the differences due to primary loadings. Three settings are explored: high (0.80), moderate (0.60) and low (0.40) primary loadings. Obviously, this is quite artificial as primary loadings are unlikely to be are equal across factor and items. Also, the low primary loading setting would only make sense - from the perspective of simple factor structure analyses - if the cross-loadings were nil or close to nil. Otherwise, mismatches will occur in a great many cases. We expect to have a high proportion of mismatches and that mismatches start to occur with lower SD for the cross-loadings compared to other primary loadings.

E.1 Simulations with high (0.8) primary loadings

Figure 3 shows the simulation results of a limited simulation where all primary loadings are set to 0.80. The four results are for 2, 4, 10 and 16 items and, as it can be seen, there is no crossing between the TCC and mTCC lines anymore. TCC always produces more mismatches, but the amount of mismatches is very low.

Figure 4 has 0 as average for the cross-loadings on the left hand side and 0.10 as average for the cross-loadings on the right hand side. As explained in the main body of the text, mean=0.10 shows the difference between the two coefficients better. However, both there are also more mismatches in these settings. It is also quite clear, that with very high primary loadings mTCC produces less mismatches (with the exception of Items=10, mean=0.00, but the number of mismatches is extremely low so even one or two mismatches jump out). Remarkably, the shared mismatches is quite low, so when mismatches occur for either coefficient, is is likely not present for the other coefficient, which underlies the argument for using the two coefficients together as suggested in the Discussion of our paper.

E.2 Simulations with moderate (0.6) primary loadings

Figure 5 shows the simulation results of a limited simulation where all primary loadings are set to 0.60. The four results are for 2, 4, 10 and 16 items and, as it can be seen, there is a crossing on each figure around SD=0.35 on all four figures. TCC produces more mismatches in mid range SDs, but this disadvantage disappears when the number of items increases. The proportion of mismatches is much higher when the number of items is low and compared to the high primary loadings settings.

E.3 Simulations with low (0.4) primary loadings

Figure 6 shows the simulation results of a limited simulation where all primary loadings are set to 0.40. The four results are for 2, 4, 10 and 16 items and, apparently there is an important difference depending on the number of items; for 2 and 4 items there is still a mid-range advantage for the mTCC, albeit rather small, but for 10 and 16 items the TCC outperforms the mTCC from the moment mismatches start to occur. As mentioned previously, this is a rather unlikely setting implying very weak correlations between the items, where the usefulness of EFA is quite dubious. One could argue that when the primary loadings are set to 0.40 and 16% of the cross-loadings exceeds this value (at SD=0.20), it is not possible to assume a simple factor structure, which is a requirement for the mTCC (and TCC) to be used for factor matching.

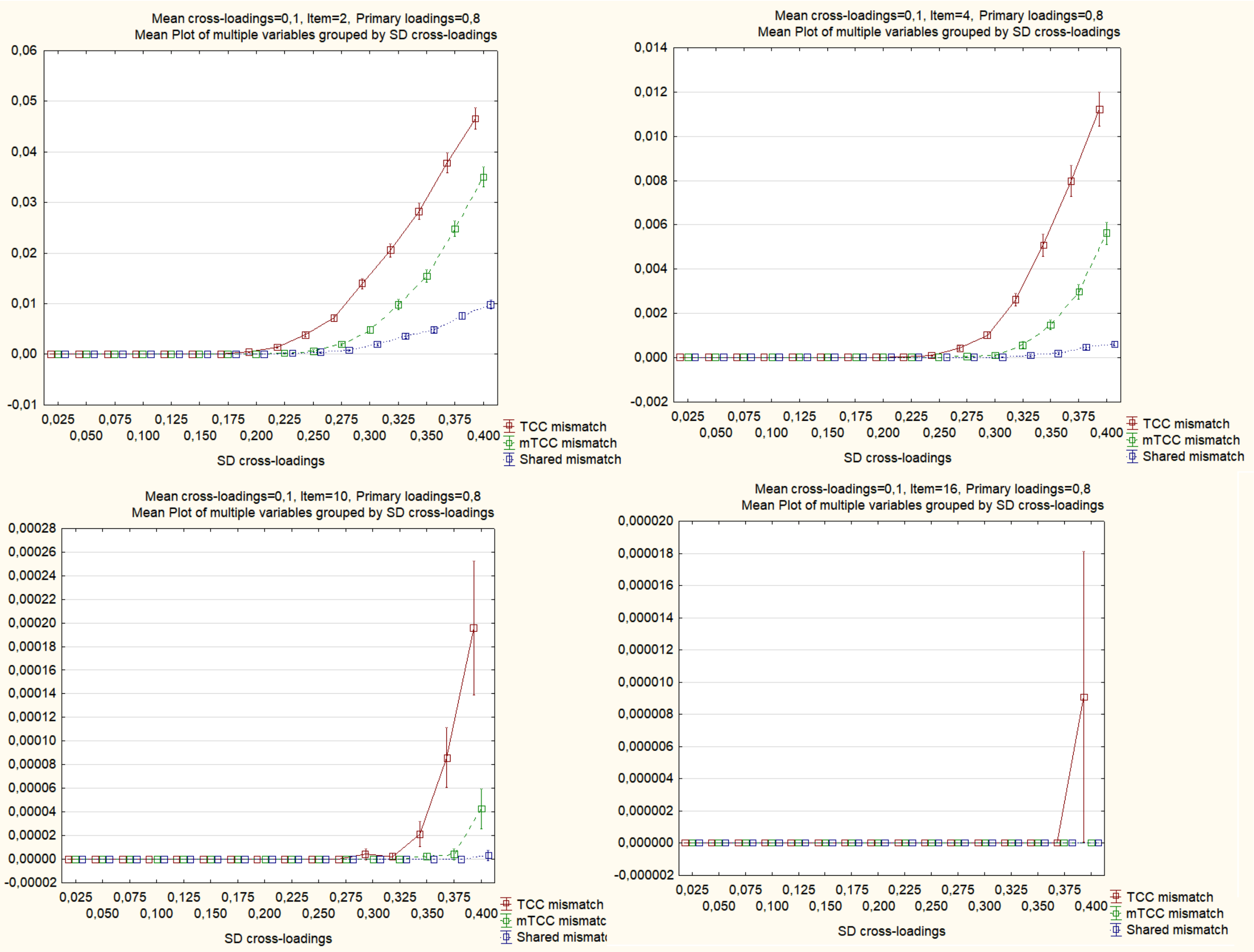

Figure 3: Simulations with high (0.8) primary loadings

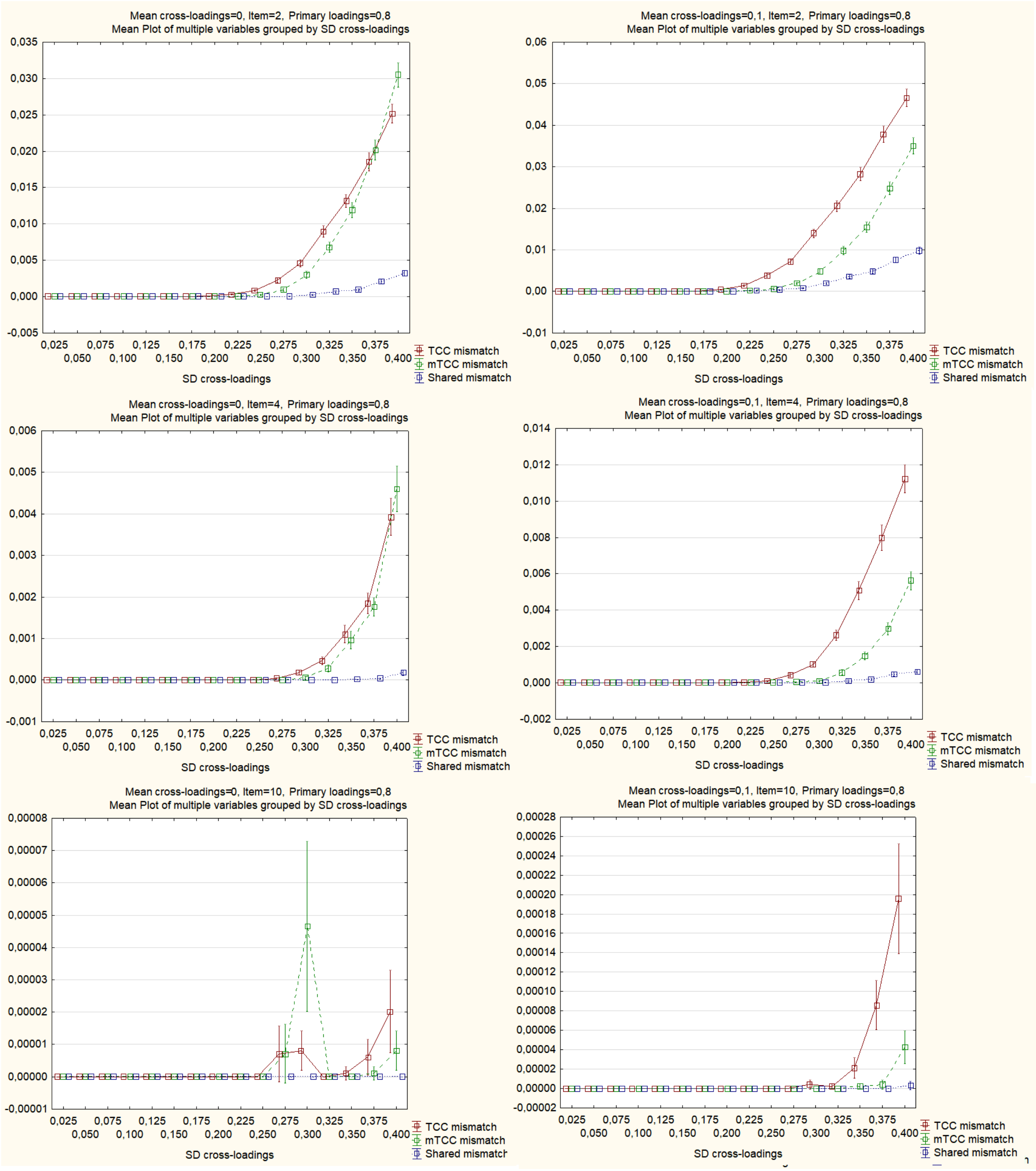

Figure 4: Simulations with high (0.8) primary loadings - also comparing mean CL

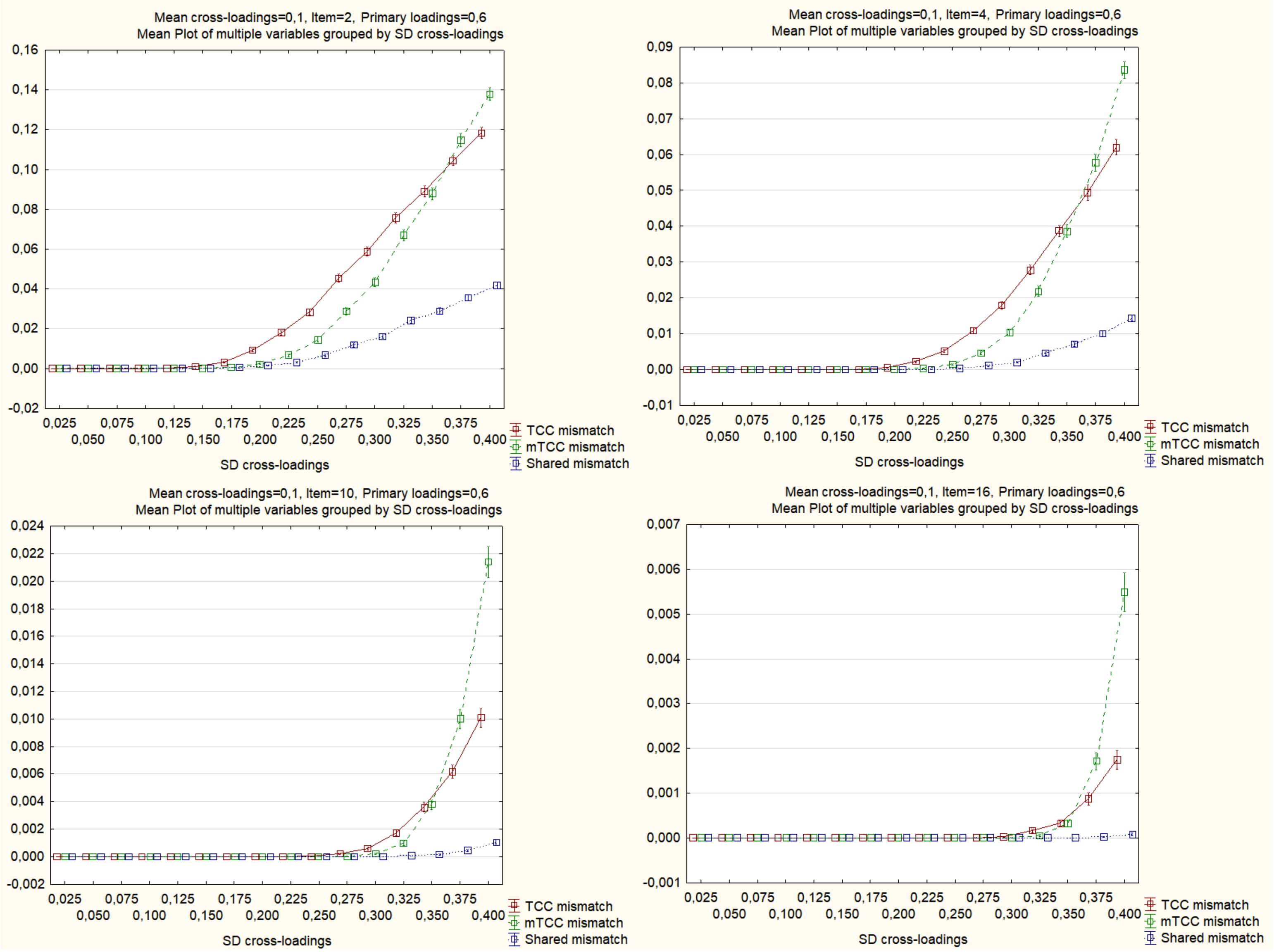

Figure 5: Simulations with moderate (0.6) primary loadings

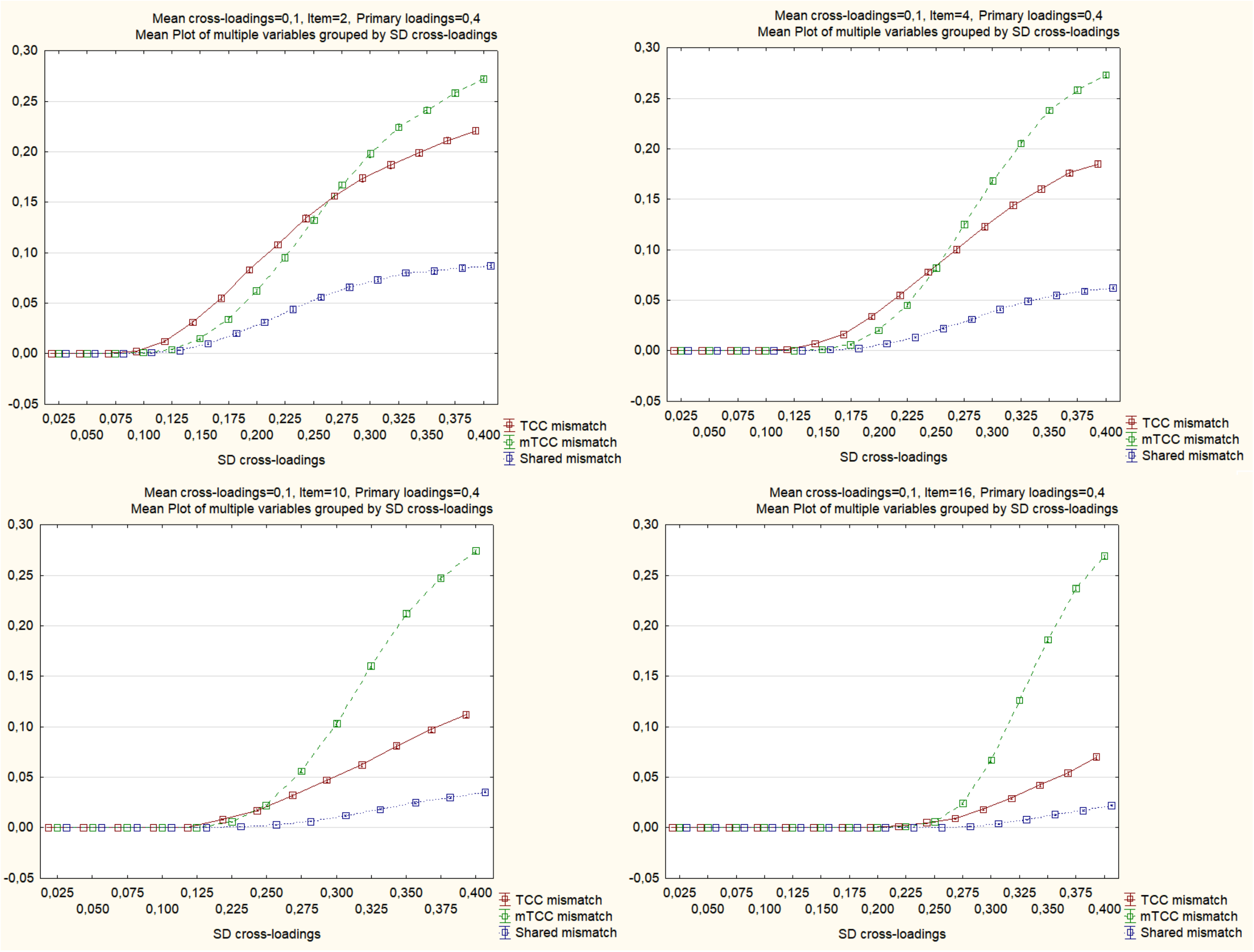

Figure 6: Simulations with low (0.4) primary loadings

F. Simulations with 2, 4, 10 and 16 items for 2 factors

Previous figures such as 3, 5and 6 can be used to compare 2, 4, 10 and 16 items per factor settings. In the Figures in this section, 4 and 16 items are presented for different primary loadings.

F.1 Simulations with 4 items per factor

Figure 7 presents first all results with 4 items categorised by primary loadings (0.8 (green), 0.6 (red) and 0.4 (blue)), then plots mismatches for each setting of primary loadings separately. As expected, the most mismatches occur with the low (0.4) primary loadings and the advantage of mTCC over TCC disappears nearly immediately after mismatches start to occur (top right), while with very high primary loadings the mTCC is clearly resulting in much less mismatches.

F.2 Simulations with 16 items per factor

Figure 8 presents first all results with 4 items categorised by primary loadings (0.8 (green), 0.6 (red) and 0.4 (blue)), then plots mismatches for each setting of primary loadings separately. As expected, many mismatches occur with the low (0.4) primary loadings. Surprisingly, these occur more often with the mTCC. However, there are nearly no mismatches (less than 0.5% 0.0% for primary loadings 0.6 and 0.8, respectively).

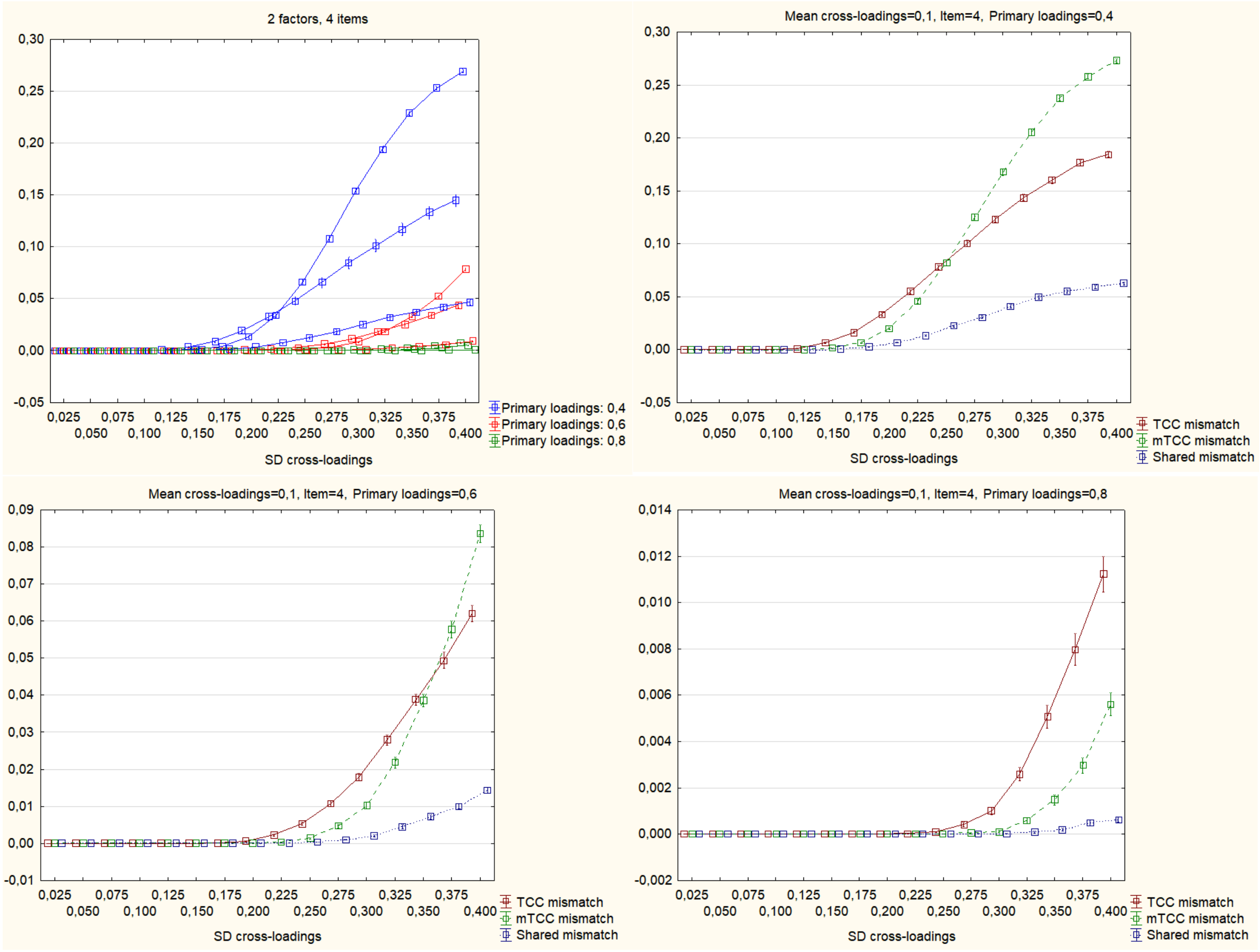

Figure 7: Simulations with 4 items per factor

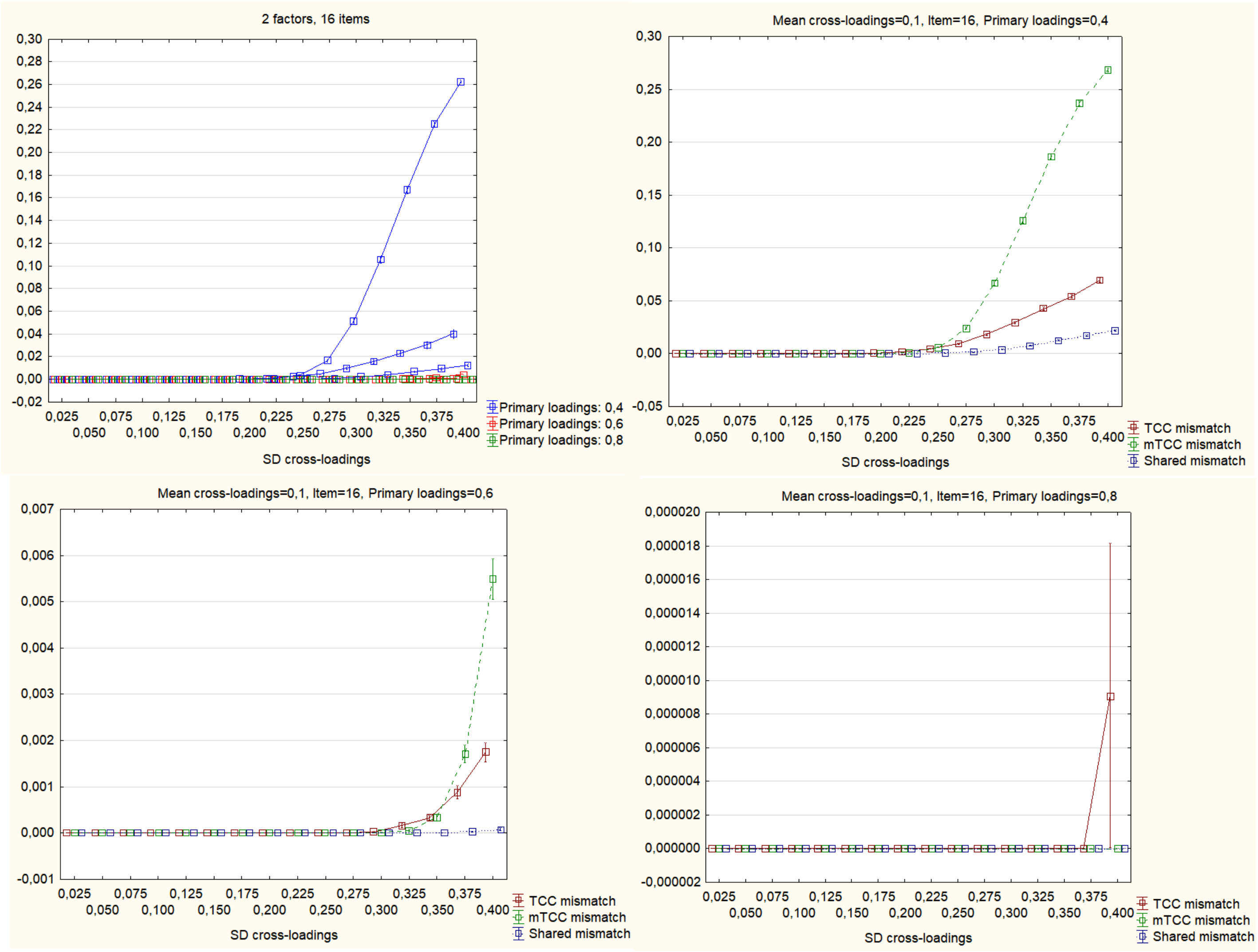

Figure 8: Simulations with 16 items per factor

G. Simulations with mean cross-loading of 0 and 0.20

Due to the choices made during EFA, to be precise, due to the chosen rotations, factor loadings tend to be positive, especially in the first factors. For this reason, we chose for most of our analyses a mean cross-loading of 0.10. However, with different mean CLs, similar patterns for TCC and mTCC mismatches can be observed.

To show the changes due to different mean cross-loadings, 0, 0.1 and 0.2 mean CL figures are presented next to each other for 2 (top row) and 10 factors. This is done for the high (Figure 9), moderate (Figure 10) and low (Figure 11) primary loadings settings separately. The tendency for an advantage of mTCC over TCC is more visible where the proportion of negative loadings is smaller (corresponding to the settings with higher mean CL).

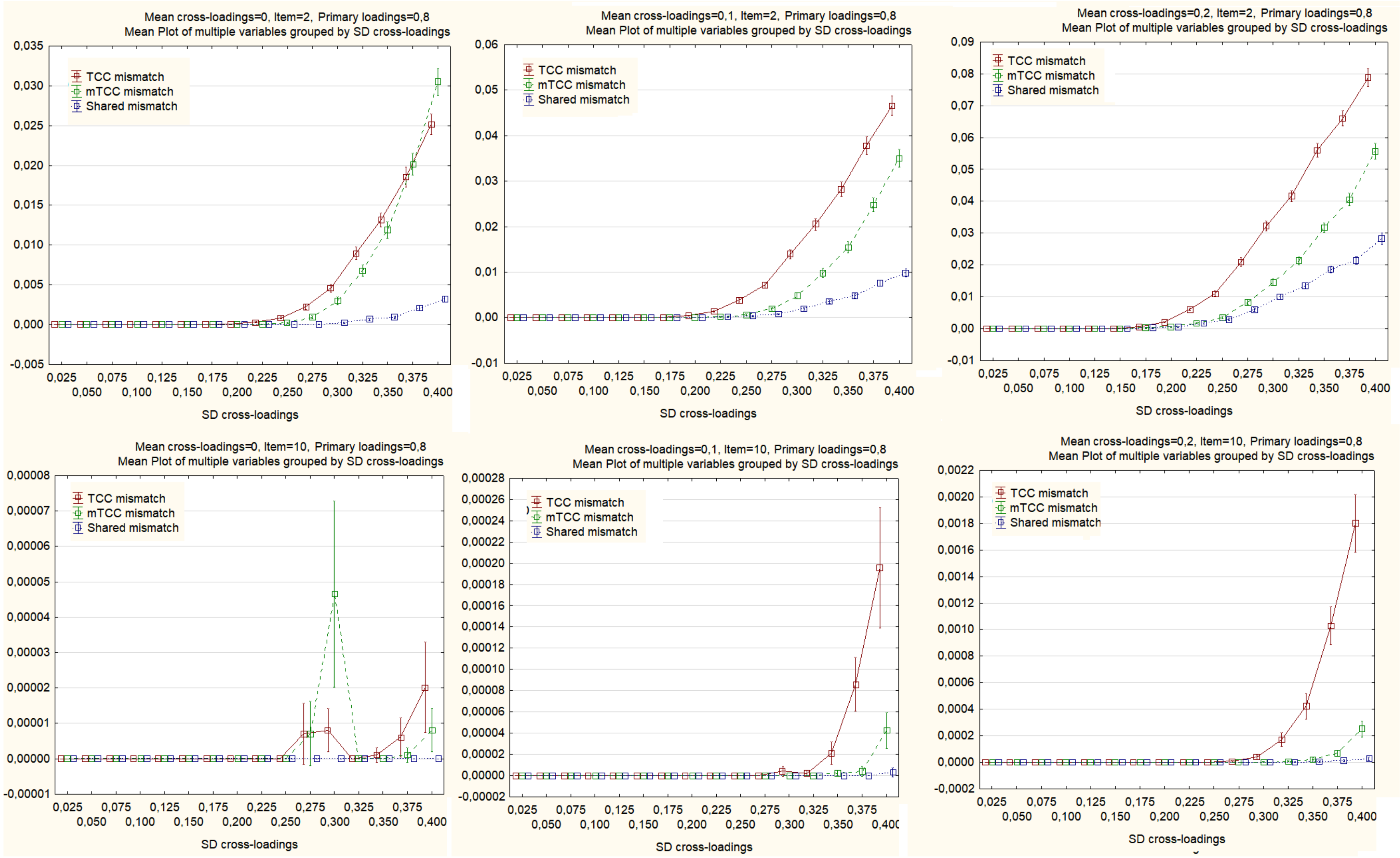

Figure 9: Simulations with different mean cross-loadings for 2 and 10 factors with high (0.80) primary loadings

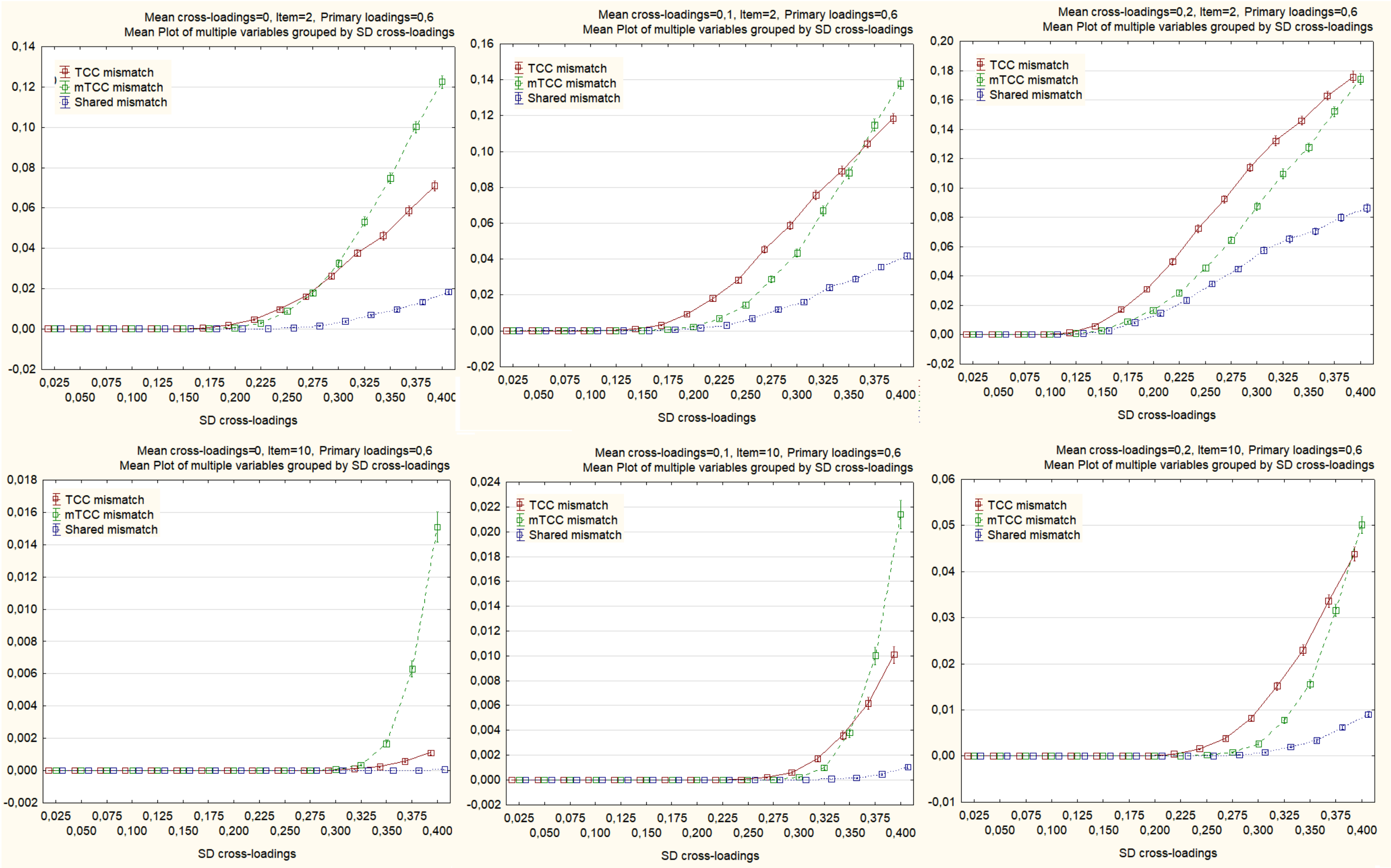

Figure 10: Simulations with different mean cross-loadings for 2 and 10 factors with moderate (0.60) primary loadings

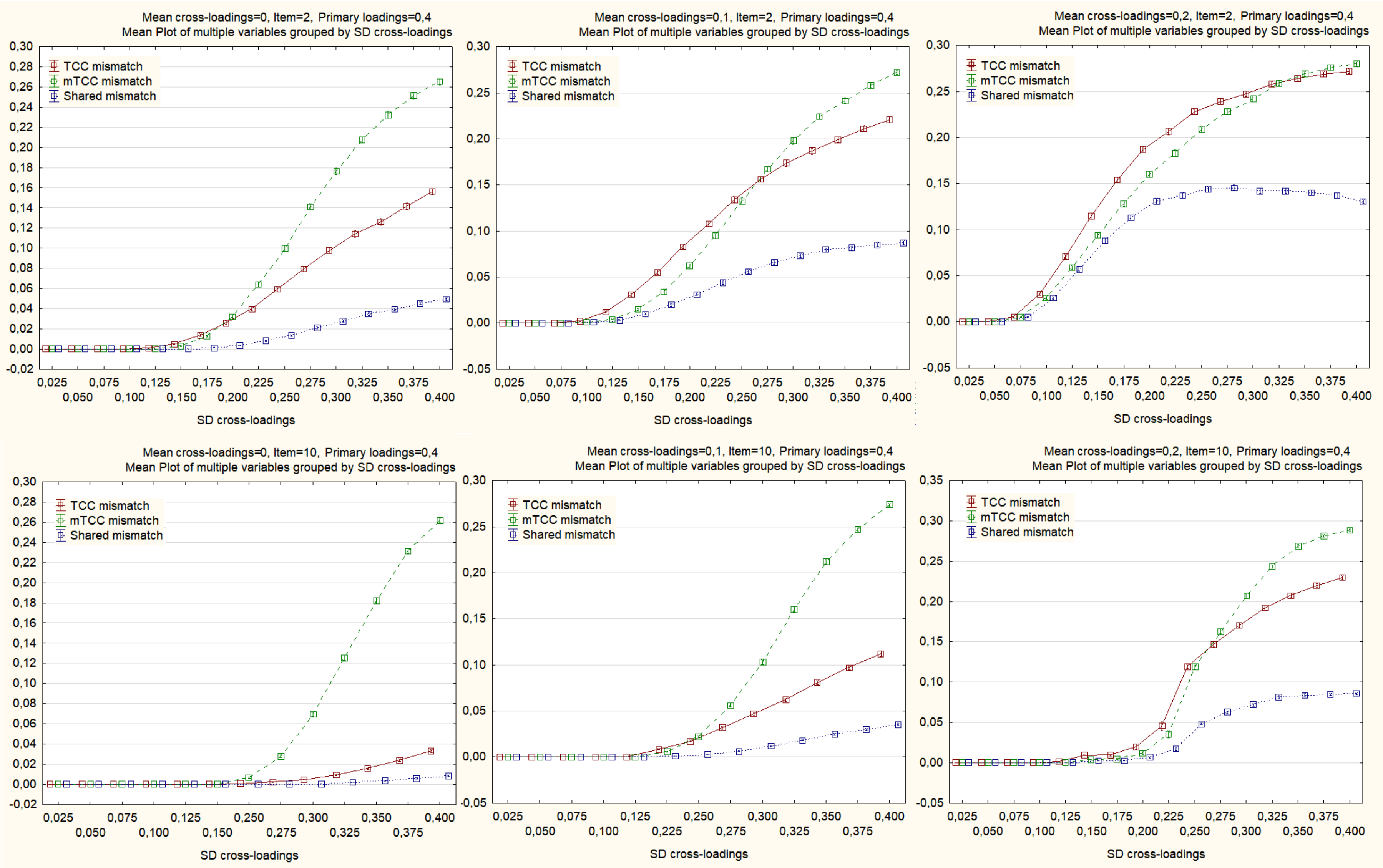

Figure 11: Simulations with different mean cross-loadings for 2 and 10 factors with low (0.40) primary loadings

H. Simulation results comparing the TCC-mTCC-PCC triad

A reviewer pointed out to us, that it woould be important to also look at the Pearson correlation coefficient (PCC) as a potential reference point for the mTCC. Below we show the results for our case study and also the same simulation as in the body of the paper but with the Pearson correlation added as a baseline criterion as requested by a reviewer of the paper.

In our motivating case study, the DiF study, the number of mismatches are 99, 97 and 1, for the TCC, PCC and mTCC, respectively. Note, that this is a five factor setting with 8-10 items per factor, main primary loadings are decreasing steadily and are quite low (range: $0.24 - -0.68$). There is at least one item per factor with a cross-loading on another factor above 0.30, but the factors are still distinguishable. Based on the results above, one can argue that despite the PCC correcting for the mean value of the factor and the TCC being uncentered, they are similar in matching factors.

The simulations are presented in Figure 12, which provides the same setting as can be seen in the main body of the paper but with the PCC added as a reference point. As it can be seen in these simulations, the PCCs and TCCs can produce very different results (this is also pointed out in ESM A with the toy examples). In the original simulation, included in the body of our manuscript, we chose a specific setting, where a) primary loadings were fixed across all factor analyses, b) primary loadings were often the same for several items and c) the number of items was equal for all factors. (The exact loadings are presented in ESM B).

In the above described situations, the simulations show that the PCC matches the factors better than the other two coefficients in some cases and not in others. In the two factor settings (F2I2 and F2I10, top row), especially where the number of items is only two (left hand side). This effect disappears completely in the 5 factor, 2 item setting (bottom left) and is less apparent in the F5I10 setting (bottom right), where the mTCC has the fewest mismatches in the range of $0.225 - 0.325$, but the PCC outperforms the mTCC in the range of $SD > 0.325$. As we have pointed out previously, this is a range where simple factor structure is debatable with our chosen primary loadings.

There are two important conclusions that can be drawn from the case study and the simulations. First, all three coefficients are good at factor matching. Taking into account the number of matches, the percentage of mismatches is low (even in the most extreme situations it never exceeds 30%, and if two coefficients are used together, it never exceeds 10% - see Figure 4a-4d in the manuscript). Second, there are settings where any of the three coefficients outperforms the other two.

However, which coefficient works best depends on several different components including but not limited to the number of factors, the number of items, the magnitude of the primary loadings, the magnitude and variability of the cross-loadings, whether the number of items is equal for all factors, how different the factor analyses are, etc. Studying each of these aspects, even if possible, is out of scope of our paper, which is to show that the mTCC, alone or together with the other coefficients, can be useful for factor matching.

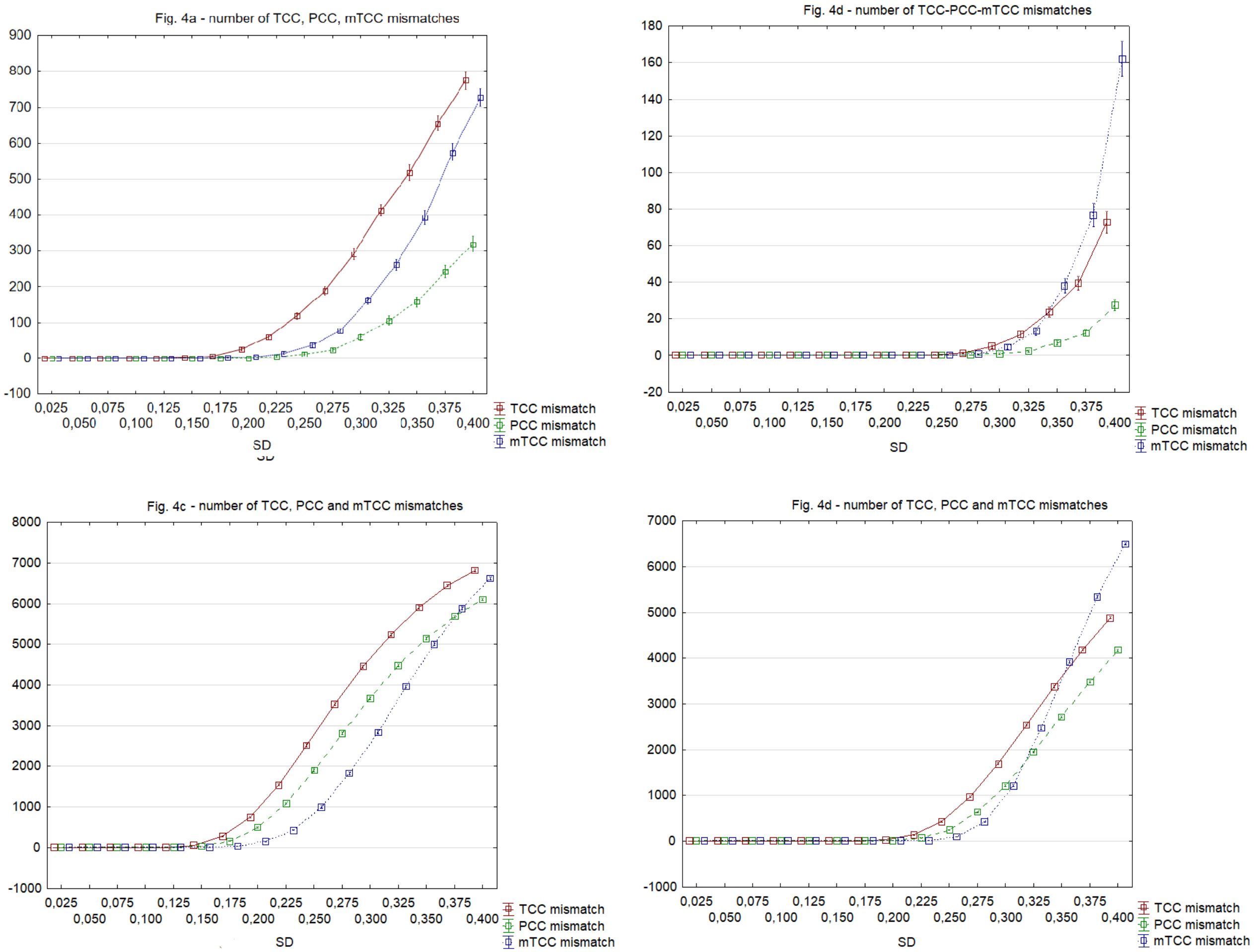

Figure 12: TCC-PCC-mTCC mismatches for the main simulation study

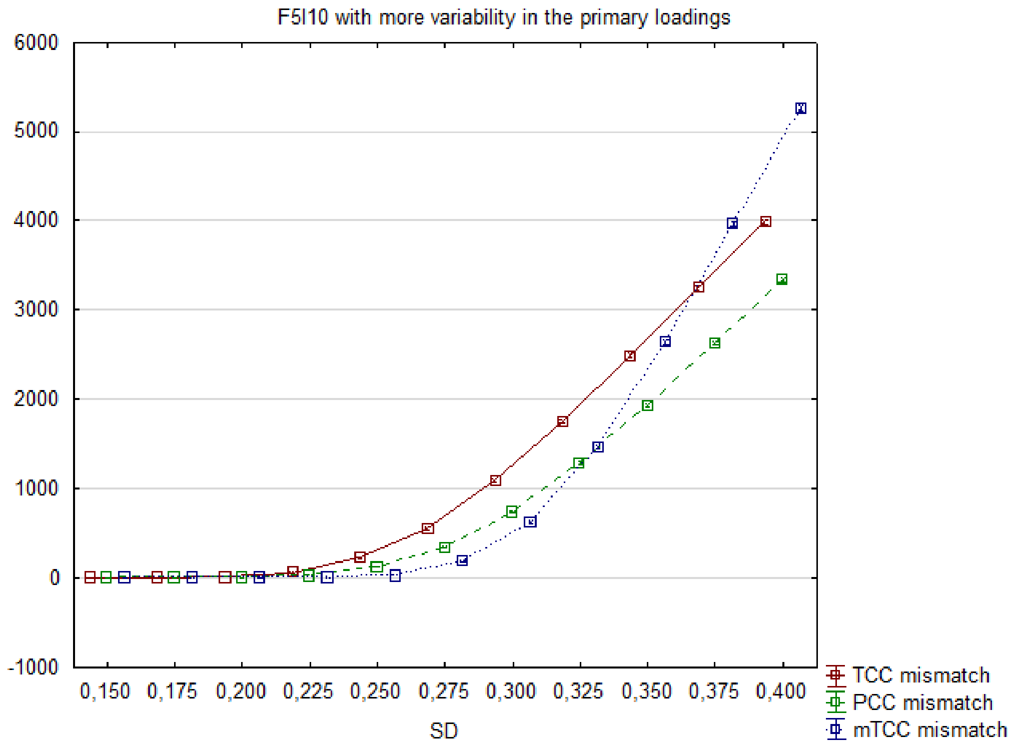

Figure 13: TCC-PCC-mTCC mismatches for the F5I10 setting with varying primary loadings

I. Histograms and scatterplots of concordance of TCCs and mTCCs in the simulations

Similar figures presenting the scatterplots of TCCs and mTCCs categorised by congruency (Figures 1 and 2 in the Main body of the paper) and the histogram of the mTCC categorised by congruency (Fig. 3 in the main body of the paper) is showi in Figures 14, 15, 16 and 17.

It is visible on the figures that fixing the primary loadings and assuming a normal distribution for the cross-loadings influences and, perhaps limits, the values of TCC and mTCC that can be observed. The figures also summarise all TCCs and mTCCs for the given setting, independently from the cross-loadings, which is not realistic in real-life data. This may be the reason why the line between congruent and incongruent pairs is slightly blurred for the 5 factor settings (where the later factors had lower primary loadings), but are quite well separated for 2 factors.

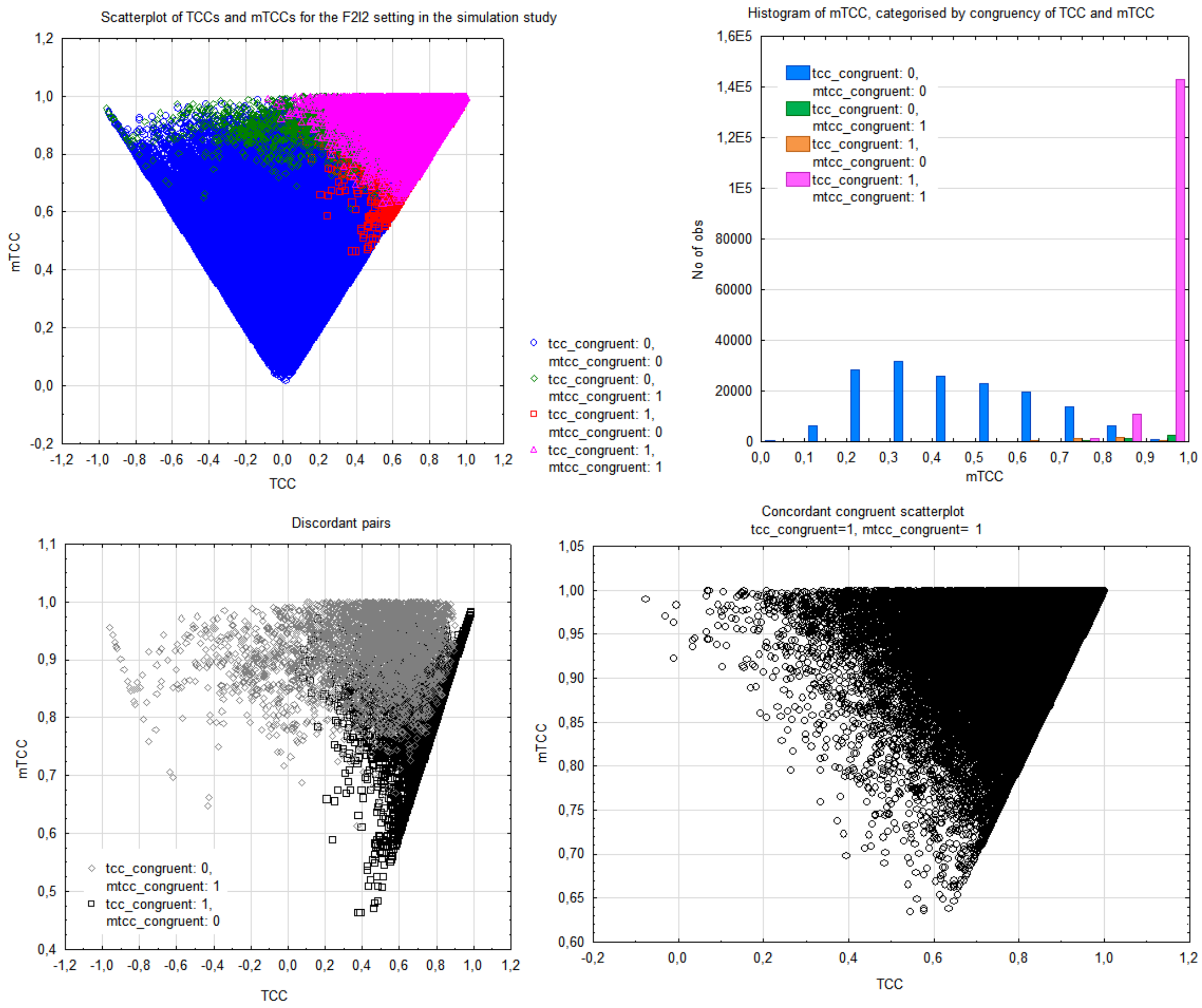

Figure 14: Figure 1-3 for the 2 factor, 2 item simulation setting

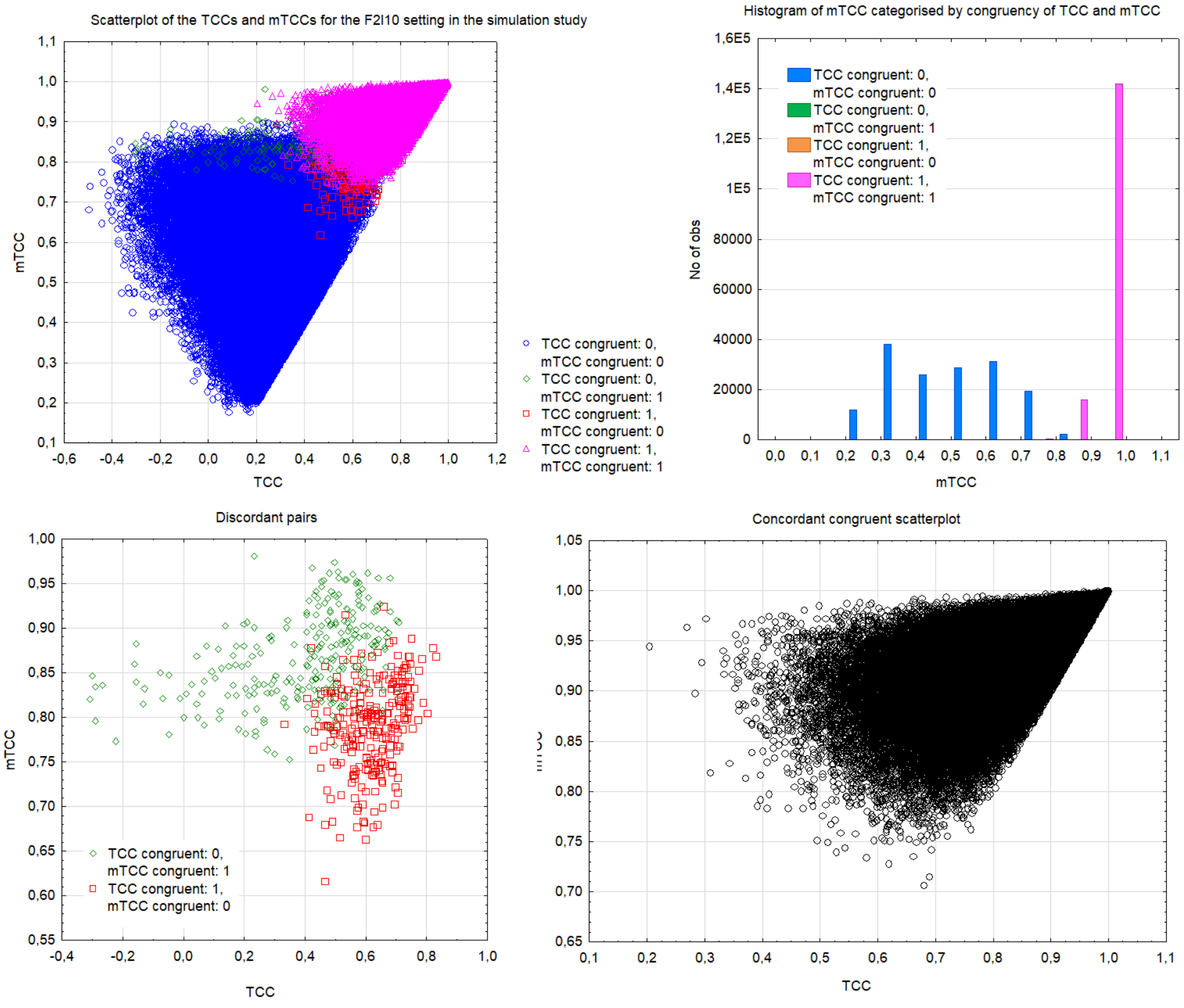

Figure 15: Figure 1-3 for the 2 factor, 10 item simulation setting

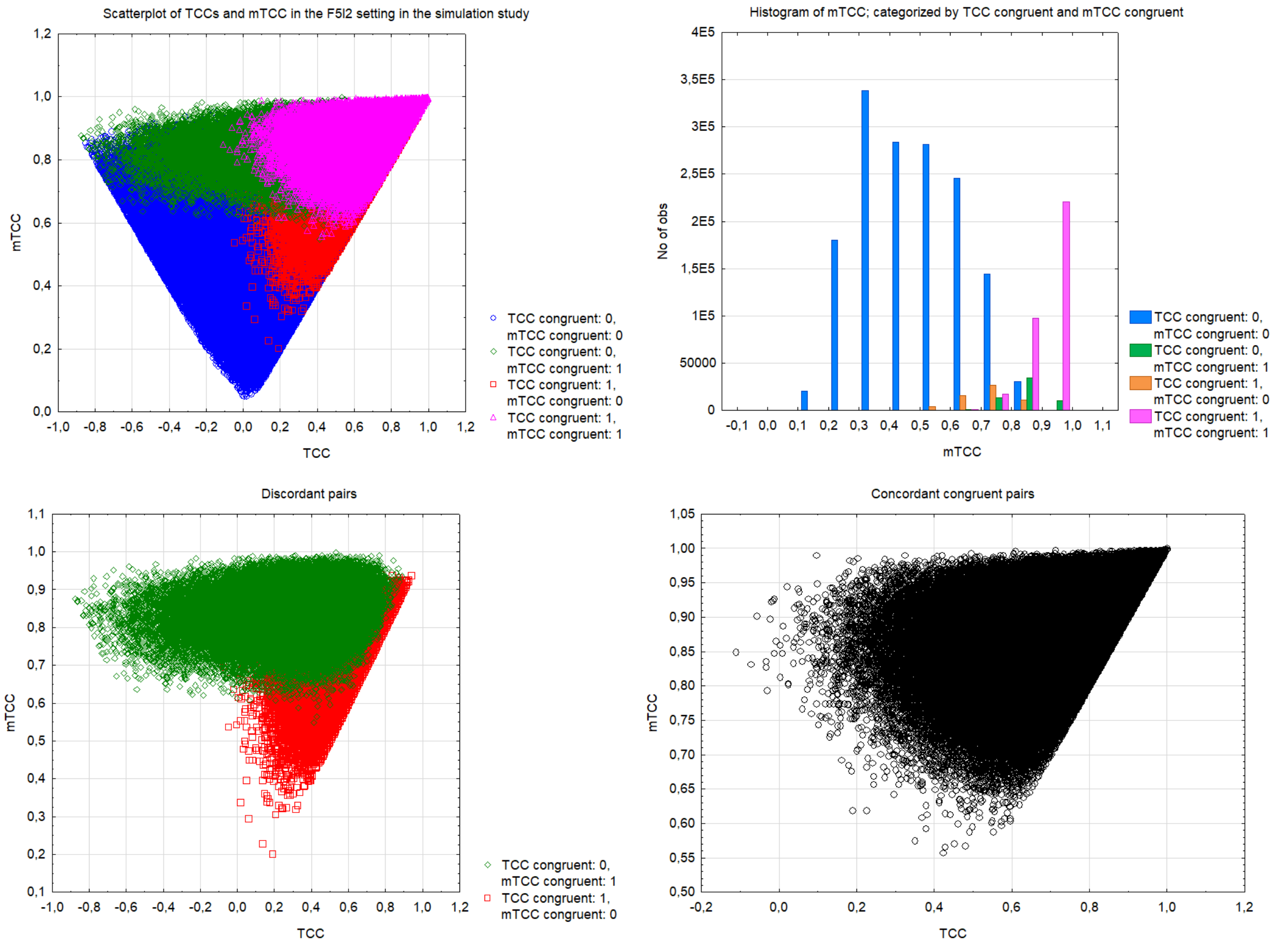

Figure 16: Figure 1-3 for the 5 factor, 2 item simulation setting

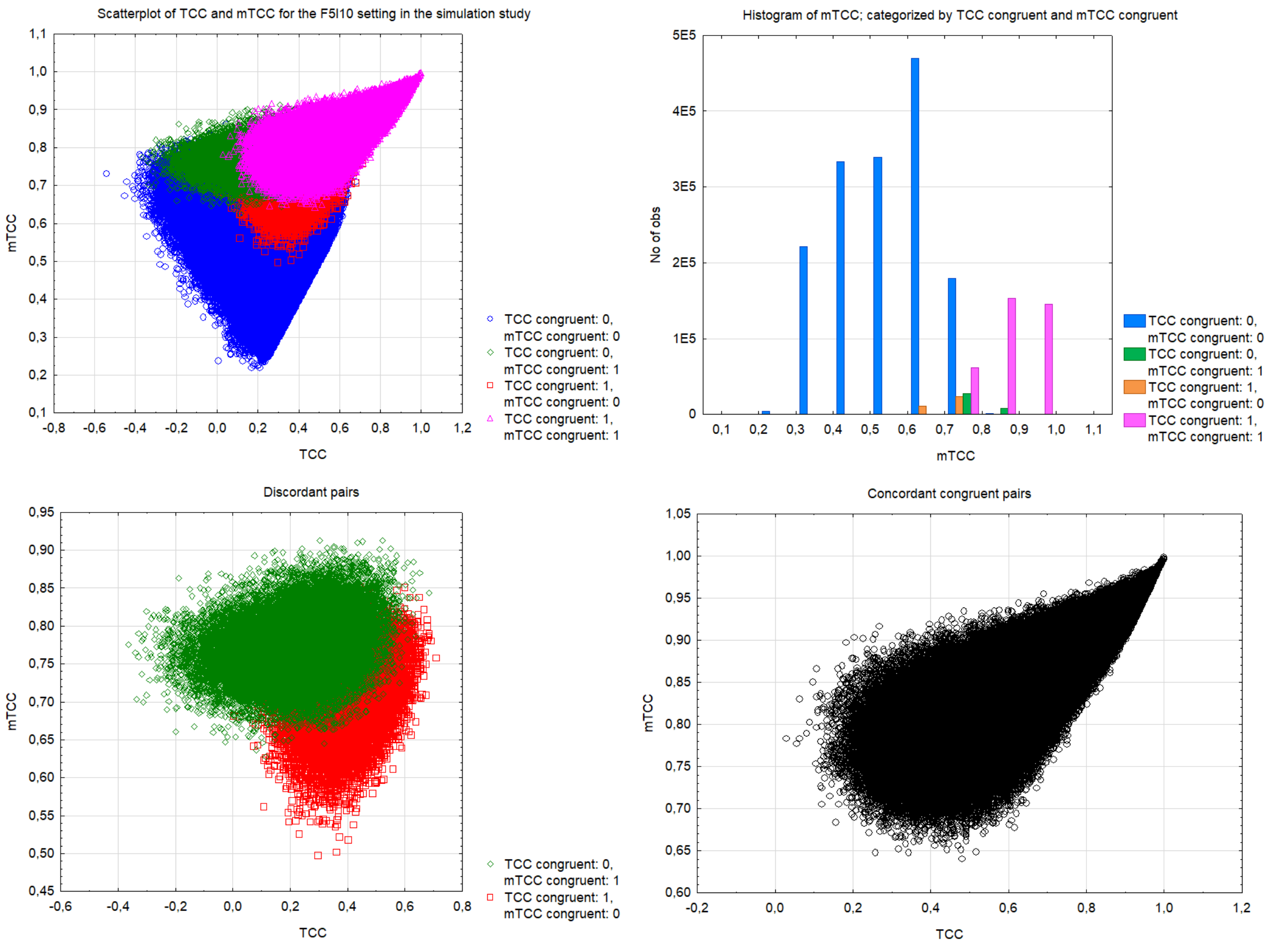

Figure 17: Figure 1-3 for the 5 factor, 10 item simulation setting