Comments on Page 4 (Figure 2 and Table 1)

The sub-figure on the left hand side in Figure 2 shows the relationship between scaled a-parameters using Mean and Sigma method and the ones from item bank for the 10 common items. We can tell that the regression line deviates from the 45 degree line, and the deviance increases as the a-parameters go up (i.e. the deviance is larger for more discriminating items). In our case, for 9 out of the 10 common items, the original a-parameters from the item bank are overestimated after scaling.

The original a-parameters and scaled a-parameters of common items are highly but not perfectly correlated, with r being 0.836, according to Table 1.

Compared with scaling the a-parameters, the Mean and Sigma method does a better job scaling the b-parameters, which we can tell by looking at the regression line and the 45 degree line that have almost none deviance (the sub-figure on the right hand side in Figure 2), as well as the correlation that almost equals 1 (*r* = 0.989), based on Table 1.

Comments on Page 4-5 (Figure 3 and Table 2).

The sub-figure on the left hand side in Figure 3 shows the relationship between scaled a-parameters using the regression method and the ones from item bank for the 10 common items. We can tell that the regression line deviates from the 45 degree line, and the deviance increases as the a-parameters go up (i.e. the deviance is larger for more discriminating items). In our case, For 9 out of the 10 common items, the original a-parameters from the item bank are overestimated after scaling.

The original a-parameters and scaled a-parameters of common items are highly but not perfectly correlated, with *r* being 0.839, as is shown in Table 2.

Therefore, the regression method is not perfect for scaling the a-parameters.

Compared with scaling the a-parameters, the regression method does a better job scaling the b-parameters, which we can tell by looking at the regression line and the 45 degree line that have almost none deviance (the sub-figure on the right hand side in Figure 3), as well as the correlation that almost equals 1 (*r* = 0.990), based on Table 2.

Comments on Page 8 (Figure 4)

The sub-figure on the left hand side in Figure 4 displays the relationship between scaled a-parameters using the Stocking-Lord method and the ones from item bank for the 10 common items. We can see that the regression line still deviates from the 45 degree line, indicating that the scaled a-parameters are still not perfect. To be more specific, in our case, for less discriminating items (items with smaller a-parameters), the original a-parameters tend to be underestimated, while for more discriminating items (items with relatively larger a-parameters), the original a-parameters tend to be overestimated.

The same conclusion that estimates for the a-parameters are biased can be drawn from the correlation between the scaled a-parameters and the original ones (*r* = 0.838).

Scaling for the b-parameters is better than that for the a-parameters, in that there’s less deviance, but still, the correlation is not perfect. However, generally the estimates are pretty accurate, with the correlation between scaled and original b-parameters being 0.99.

Comments on Page 8 (Table 3)

Table 3 shows the summary of our main findings, and we can see that in terms of accuracy, by looking at the correlations between scaled item parameters and the ones in the item bank for common items, the 3 methods do not differ greatly, and they all do a better job scaling the b-parameters compared with the a-parameters. However, the specific pattern of the relationships between scaled and original parameters varies for each method, and can be observed by looking at the plots (Figure 2, 3, and 4).

In terms of scaling constants, the ones are very similar generated by the Mean and Sigma method and the regression method. However, when using Stocking-Lord method, we obtained a much biggerthat is very close to 1, indicating that there will be almost zero change from estimated a-parameters of common items to scaled a-parameters. The value of  is also different, compared with the other 2 methods.

Generally speaking, after scaling, discrimination levels of the common items tend to change more greatly than the b-parameters, compared with the original parameters.