1. 通过各种IRT或non-IRT的方法找出有DIF的题；

2. 把没有DIF的题 (至少一个) 作为anchor item, 然后估reference和focal groups的parameter. 如果放在一起估把有DIF的题看作unique items, 比如8道题最后两道DIF，就1-6 common items, 7-8作为Reference的两道题，然后9-10作为focal的两道题。分开估的话就找两组anchor item parameters之间的关系，然后把一组的参数通过加减乘除掰到另一组的scale上。

3. 用在同一scale的两组参数算DIF effect size.

你上一封邮件总结下来就是MLG能处理少一个category的题，GGUM不行。先说个没什么灵活度的，如果GGUM就是算不了少一个category的，但要是非得比reference和focal groups的参数，那明显就是比不了了。。。如果因为这个去collapse另一个sample的话得看看另一个sample每个category的频率或者百分比吧 如果你说另一个组该题fit的不错 那好像就确实说明很有DIF的可能 这样的题自然就排除在anchor item之外了 也就是你没用以上第一步但通过其他方法找到了一些DIF items。接下来第二三步，受GGUM的限制，如果想看参数就必须collapse了。情况无非两种：另一个sample的某个category选的人也很少，但还不至于是0，这种也许DIF测出来可能性小些？另一个种可能就是虽然是少数但还是有一定人数选这个category的 这种估计DIF也就容易被找出来

MLG的话比较灵活 倾向于不collapse 两个sample估同样数量的参数 一组的很极端一组的还凑合 DIF也出来了 如果你要MLG和GGUM一致的话也可以collapse MLG的数据 然后情况估计也会和GGUM的差不多

1. Scale intended to have 4 response categories: our item didn’t intend to have 3 response categories.
2. These items do not have a lot of people endorsing option 1 in the other group either; less than 10 out of 1000 or 900 people -- collapsing? Dichotomization?
3. Show him the MODFIT results using SGR – local independence.