《心理学报》审稿意见与作者回应

题目：基于双因子模型的测验总分和维度分的合成方法

作者：刘玥 刘红云

第一轮

审稿人 1 意见：

意见 1：在文章第五部分“5. 模拟研究结果”中多次提到“误差均方根”在不同方法中的不同，有差异，或者变化趋势，但是针对“误差均方根”的统计学比较应用的是什么分析方法，并没有介绍，也没有相应的统计量和 p 值，仅仅是从图形中可以看到这样的变化趋势或大小不同，但是针对文章中的模拟数据也是有抽样误差的，应该有具体的统计学方法，给出统计量和 p 值。

回应：谢谢审稿专家的意见。误差均方根是模拟研究中测量估计值和真实值差异的常用评价标准。针对误差均方根，本研究主要从两个角度来描述研究结果，一是对各自变量下的误差均方根结果进行方差分析，有主效应、交互作用分析的自由度、F 值、P 值和效应值，以及简单效应分析的差异值、P 值。具体请见第 9-10 页、第 12-14 页标红部分。本研究的模拟数据不涉及抽样问题，且参考了大量模型比较的模拟研究，如 Yao(2010)、Yao(2011)、Jimmy, Song 和 Hong(2011)、Wiberg(2016)的研究等。这些研究在计算参数返真性(parameter recovery)时，均采用偏差、绝对偏差、误差均方根、相关等类似指标，因此本研究也采用了上述评价标准。在模拟数据生成和分析过程中，为了避免随机偶然因素的影响，每种条件重复生成了 30 个样本，所以研究中考虑了抽样误差。

参考文献：

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. Applied

Psychological Measurement,35(1), 48-66.

Wiberg, M. (2016). Alternative linear item response theory observed-score equating methods. Applied Psychological Measurement, 40.

意见 2：图 2-a 中，一般统计方法，随着样本量的增加，模拟效果应该是越来越好的，“误差均方根”应该是越来越小的，在该图中，MIRT 残差反而随着样本量增加有升高趋势，是什么原因造成的？应加以解释，或寻找原因。

回应：谢谢审稿专家仔细的阅读。在图 2-a 中，MIRT 法所合成的总分在样本量为 500、1000、2000 的条件下的误差均方根分别为 0.376、0.386、0.381，对 MIRT 法合成总分在各样本量下的误差均方根进行方差分析，结果证明样本量的主效应不显著($F(2，42)=0.04$，$P>.001$)，说明 MIRT 法合成总分在各样本量条件下准确性的差异未达到统计学上的显著水平，不能得出“MIRT 残差反而随着样本量增加有升高趋势”的结论。另外，本研究中变化的自变量是样本量，此处考察的是能力参数估计的准确性，一般来说，增加样本量会提高题目参数估计准确性（Wang & Wilson, 2005; Park et al., 2015; Culpepper, 2015），对能力参数估计准确性的影响不大，如，Wiberg（2016）的模拟研究中，减少样本量并没有削弱局部线性等值方法对分数等值的准确性。因此，本研究中，随着样本量增加，总分估计的准确性未显著增加并未违背相关规律。

参考文献：

Culpepper, S. A. (2015). Revisiting the 4-parameter item response model: bayesian estimation and application. *Psychometrika*, 1-22.

Park, R., Pituch, K. A., Kim, J., Dodd, B. G., & Chung, H. (2015). Marginalized maximum likelihood estimation for the 1pl-ag irt model.*Applied Psychological Measurement, 39*(6).

Wang, W. C., & Wilson, M. (2005). The rasch testlet model. *Applied Psychological Measurement, 29*, 126-149.

Wiberg, M. (2016). Alternative linear item response theory observed-score equating methods. *Applied Psychological Measurement, 40*.

意见 3：模拟数据是如何产生的，来自于什么分布的数据，没有详细阐述，作为方法模拟的数据应该给出详细说明，在模拟中可以考虑用 bootstrap 等蒙特卡洛方法进行大量模拟，参数将趋于稳定。

回应：谢谢审稿专家的意见。已在“4.2 数据生成”部分增加了模拟数据的题目参数分布、能力参数分布，以及模拟数据所依据的模型公式。具体请参见第 8 页标红部分。本研究就是采用了 Monte Carlo 模拟研究的方法。关于这一点，原文叙述不太清楚，现在在文中补充了这一信息，见 4.2 的叙述（第 8 页标红部分）。关于 Monte Carlo 模拟研究的重复次数，参考了以下文献。

参考文献：

Huang, H. Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement, 39*(5).

Jimmy, D. L. T., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: a higher-order irt model approach. *Applied Psychological Measurement, 33*(8), 620-639.

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*(2), 177-197.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement,35*(1), 48-66.

意见 4：“6.2 模型拟合结果”提到了用逻辑回归，应该给出逻辑回归分析结果的常用统计量，比如回归系数，等。并应该给出逻辑回归自变量的赋值说明。

回应：谢谢审稿专家的意见。文章中“6.2 模型拟合结果”提到的逻辑回归，其实是指单维 IRT 模型中的单维两参数 Logistic 模型，它是单维两参数 IRT 模型中的一种类型。其形式如下

$$P(x_{ij}=1|\theta_i,\beta_j)=\frac{e^{\beta_{2j}\bullet\theta_i-\beta_{1j}}}{1+e^{\beta_{2j}\bullet\theta_i-\beta_{1j}}}$$

其中，x_{ij}=0 或 1 分别表示被试 i 答错或答对题目 j。β_{2j} 表示区分度参数，β_{1j} 表示难度参数，θ_i 表示被试 i 的能力值。为了避免混淆，在文中作了修改。由于是 IRT 模型的分析结果，所以没有按照逻辑回归分析的形式给出常用统计量的结果。

意见 5：“表 7 呈现了各方法合成的实际数据总分和维度分的百分位数。从表中可以看出，除了 Bifactor-M2 法合成总分的 0.05 和 0.95 百分位数，Bifactor-M1 法合成维度分的 0.05 和 0.25 百分位数，Bifactor-M2 法合成维度分的 0.05 百分位数，Bifactor-M3 法合成维度分的 0.05 和 0.95 百分位数与 MIRT 法相差 0.1 以上，其余方法的百分位数与 MIRT 法相差都较小。Bifactor-M3 法合成总分，Bifactor-M4 法合成维度分的百分位数与 MIRT 法的最为接近，绝对差异基本未超过 0.05。”

“绝对差异基本未超过 0.05。”为什么用绝对差异，应该用相对数表达更为客观。

回应：谢谢审稿专家的意见。已经改成“相对差异基本在[-0.05,0.05]区间内”，具体请参见第 17 页标红部分。

审稿人 2 意见：

该文基于双因子模型，提出了四种合成总分和维度分的方法，分别是：原始分法，加和法，全局题目加权加和法和局部题目加权加和法，并采用模拟研究的方法，在样本量、测验长度、维度间相关变化的条件下考察了这些方法与传统多维 IRT 方法的表现。最后，通过实证研究对研究结果进行了进一步验证。该文有一定的意义。但有如下不足和作者商榷：

意见 1：实证研究的结果与模拟研究的结果一致，这是十分正常的现象。因此，本文在模拟研究之后，进行了一项实证研究，来进一步验证模拟研究的结果，似乎没有很充分的必要性。

回应：谢谢审稿专家的意见。实证研究的结果与模拟研究的结果一致，的确是十分正常的现象，就本研究主要探讨的问题来看，模拟研究部分已经可以说明问题。但是本研究之所以进行实证研究，是出于两个方面的考虑。

（1）实证研究可以为应用研究者提供这一方法应用过程中应该关注问题和方法选择上的一些建议。如模型选择的问题，通过将研究所关注的模型与实际数据拟合，进而比较拟合指标，可以对各模型与数据的拟合程度进行比较，从而选择较为适用的模型。在本研究中，根据似然值，Bi-factor 模型拟合效果最好，建议使用该模型拟合数据。因此，也进一步反映了探索基于双因子模型的测验总分和维度分合成方法的必要性。（2）通过在实证研究中运用研究所提出的各方法合成总分与维度分，并对结果进行比较，可以结合实际测验比较实际数据中各方法的差异，并对基于双因子模型的测验总分和维度分合成方法的应用提出建议。另外，

在方法研究的后面加上实证研究，也是许多测量方法类的文章通常会采用的做法。如Yao(2010)、Jimmy, Song 和 Hong(2011)的研究等，在模拟研究后都加入了实证研究，来进一步验证和解释模拟研究的结果，希望使得整个研究更加完整。

参考文献：

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

意见 2：讨论部分，直到最后的建议的第二条才出现参考文献。讨论部分建议能充分的利用文献，结合前人研究进行讨论和比较。另外，请注意和前言的讨论相呼应，目前感觉讨论和前言联系不够紧密。

回应：谢谢审稿专家的意见。原文讨论部分的确存在参考文献引用不足，和前言联系不够紧密的问题，已经对原文的前言和讨论部分都进行了修改。具体请参见第 18-21 页标红部分。

意见 3：模拟研究结果部分，作者使用了方差分析，但是方差分析的结果表述存在不规范和错误的地方，请作者自查更正。

1)例如，在 5.2.1 部分，作者谈到“结果显示，维度 1 维度分合成方法的主效应显著(F(5，196)=5453.35，P<.001，η^2=0.993)，”，根据图 4 可知，作者比较了 5 种方法，方差分析的第一自由度应该是 4，不是 5。

回应：谢谢审稿专家仔细的阅读与指正，已经对相应内容进行了修改，具体请参见第 12 页标红部分。

意见 4：2)作者谈到“结果表明，维度分合成方法与样本量(F(10，196)=7.31，P<.001，η^2=0.272)、维度分合成方法与测验长度(F(10，196)=57.21，P<.001，η^2=0.745)、维度分合成方法与维度间相关(F(20，196)=423.01，P<.001，η^2=0.977)的交互作用显著。”既然交互作用显著，就应该进行简单主效应分析，此时再讨论主效应，意义不大。

回应：谢谢审稿专家的意见，交互作用显著的情况下，确实应当进行简单效应分析。已经分别对总分和维度分的结果添加了这部分内容，具体请参加第 10 页、第 13-14 页标红部分。

意见 5：3)作者的简单主效应分析也没有讲清楚，例如，作者谈到“随着样本量增大，MIRT 法、Bifactor-M2 法和 Bifactor-M3 法的误差均方根的差异相对减小。”模拟研究明确说明，样本量分 500、1000、2000 三种，简单主效应应该说，当样本量为 500 时，几种方法之间的简单主效应显不显著，谁显著大于谁。

回应：谢谢审稿专家的意见，原文中简单效应分析部分的确描述不够准确，已经在交互作用结果之后加入了简单效应分析的结果，具体请参加第 10 页、第 13-14 页标红部分。

意见 6：3.2 部分作者提出了 5 种总分维度分合成方法，但是方法 3、4、5 都没有参考文献，应该是作者自己提出的方法，但是这些方法提出的依据是什么，作者并没有充分说明。

回应：谢谢审稿专家的意见。这些方法确实是我们提出的方法，关于这些方法的依据原文中叙述的确过于简单。这些方法的依据主要来自于双因子模型对全局因子和局部因子的定义，教育测量中对总分和维度分的定义，以及一些相关参考文献。已在修改稿中进一步细化了这些方法提出的逻辑以及依据，具体请参见第 6-7 页标红部分。

意见 7：引言中，作者对已有的研究做了评述，但是有如下不足：

1)本文研究的是双因子模型的测验总分和维度分合成方法，从题目上看，引言应更多立足于介绍和评述双因子模型已有的测验总分和维度分合成方法，但是引言却花了大量篇幅来评述 IRT，只在前言的最后两段话才谈到双因子模型及其总分和维度分的合成方法。

回应：谢谢审稿专家的意见。原文中的引言的确对双因子模型及其总分和维度分的合成方法论述不足，已经查阅文献并对引言进行了修改。一是删减或略写了其他分数合成方法的介绍，二是增加了双因子模型的应用，双因子模型分数合成方法的介绍及评述等内容。具体请参见第 1-3 页标红部分。

意见 8：2)有些地方论述不清楚，例如“Yao(2010)在模拟和实证研究中将双因子模型得到的总分和维度分结果和其他估计方法进行了比较，结果证明该方法合成的总分和维度分准确性远不如 MIRT 和 HO-IRT 模型的结果”，其他估计方法是哪些方法？该方法是指什么方法？

回应：谢谢审稿专家的意见，原文中的确没有阐述清楚 Yao(2010)研究中的估计方法，已在修改稿中详细说明，具体请参见第 3 页标红部分。

意见 9：文中有小错误，作者谈到“其中, x_{ji}=0 或 1 分别表示被试 i 答错或答对题目 j”，角标

写反了。

回应：谢谢审稿专家仔细的阅读与指正，原文中的角标确实写反了，但在修改稿中已经删除了这部分内容。

意见 10：2016 年关于双因子模型的总分和维度分合成，有几篇最新文献，作者并没有引用，建议作者及时更新文献。

回应：谢谢审稿专家的意见，已经重新查阅了文献，更新了几篇双因子模型分数合成的较新文献。具体请参见参考文献标红部分。

第二轮

审稿人 2 意见：

文章经过修改，较好的回答了审稿人提出的问题，但感觉还略有瑕疵，与作者商榷：

意见 1：作者在前言第 2 面用了一段话来简介双因子模型，然后在 3.1 双因子模型中又介绍了双因子模型。为什么要分开介绍？而且还有部分重复。例如前言谈到：“双因子模型的思想最早见于 Spearman(1927) 的能力二因素说，他根据人们的智力与成绩的相关程度，将能力分为一般能力和特殊能力。Holzinger 和 Swineford(1937) 正式提出了双因子模型，该模型包括了一个全局因子，反映了测验中所有题目测量的核心能力；多个局部因子，表示控制了全局因子的情况下，一组题目的共同变异，即分维度效应或方法效应。使用双因子模型拟合数据，每道题目的变异可以分解为三个独立的部分：全局因子、局部因子和残差。”，在 3.1 中作者又谈到“双因子模型又称为全局-局部因子模型(general-specific factor model)，它假定存在一个全局因子可以解释所有题目的共同变异，存在多个局部因子，控制了全局因子的影响后，每个局部因子可以额外解释部分题目的共同变异。其结构如下图所示”，因此建议是否可以整合呢？

回应：谢谢审稿专家仔细的阅读。前言中介绍双因子模型的部分和 3.1 中对模型的介绍的确存在重复的问题。因此在修改稿中将前言中介绍双因子模型源起、评价的部分内容整合到 3.1 双因子模型中，前言中只保留了基于双因子模型合成测验总分和维度分的研究综述。具体请参见修改稿第 4-6 页标红部分。

意见 2：交互效应显著后，作者进行了简单效应分析，值得肯定。但是补充的不完整。例如，

在 5.1.1 误差均方根部分，作者谈到“总分合成方法与样本量(F(6，124)=14.30，P<.001，η2=0.409)的交互作用显著。简单主效应分析结果表明，样本量为 500 时，Bifactor-M1 法的误差均方根小于 Bifactor-M2 法(d =-.021, P<.001)；样本量为 2000 时，Bifactor-M1 法的误差均方根大于 Bifactor-M2 法(d=-.016, P<.001)。”，简单主效应分析，作者分析了当样本量一定时，各种方法的误差均方根比较；还有一个方面，当方法一定时，样本量之间的误差均方根比较；但是这方面作者并没有分析，为什么不分析完整呢？请作者自查全文。

回应：谢谢审稿专家仔细的阅读。文章中的简单效应分析的确不够完整，现将包含分数合成方法的完整简单效应分析补充如下。

总分合成的简单效应分析

（1）总分合成方法与样本量(*F*(6, 124)=14.30，*P*<.001，*η*2=0.409)的交互作用显著。简单主效应分析结果表明，样本量为 500 时，Bifactor-M1 法的误差均方根小于 Bifactor-M2 法(*d* =-.021, *P*<.001)；样本量为 2000 时，Bifactor-M1 法的误差均方根大于 Bifactor-M2 法(*d*=-.016, *P*<.001)。对于 Bifactor-M2 法，样本量为 2000 时的误差均方根显著小于样本量为 1000 时的误差均方根(*d* =-.017, *P*<.001)，样本量为 1000 时的误差均方根显著小于样本量为 500 时的误差均方根(*d* =-.020, *P*<.001)，对于其他方法，不同样本量条件下的误差均方根差异不显著。

（2）总分合成方法与测验长度(*F*(6, 124)=38.24，*P*<.001，*η*2=0.649)的交互作用显著。简单效应分析结果表明，测验长度为 18 题时，Bifactor-M1 法误差均方根大于 Bifactor-M2 法(*d*=-.030, *P*<.001)；测验长度为 36 题和 60 题时，Bifactor-M1 法误差均方根小于 Bifactor-M2 法，且只有在测验长度为 60 题的情况下差异显著(*d*=-.029, *P*<.001)。MIRT 法（如测验长度 18 题与 36 题: *d*=.116, *P*<.001；测验长度 18 题与 60 题: *d*=.185, *P*<.001）、Bifactor-M1 法（如测验长度 18 题与 36 题: *d*=.116, *P*<.001；测验长度 18 题与 60 题: *d*=.166, *P*<.001）误差均方根在测验长度不同条件下的差异要大于 Bifactor-M2 法（如测验长度 18 题与 36 题: *d*=.080, *P*<.001；测验长度 18 题与 60 题: *d*=.107, *P*<.001）、Bifactor-M3 法（如测验长度 18 题与 36 题: *d*=.091, *P*<.001；测验长度 18 题与 60 题: *d*=.150, *P*<.001）。

（3）总分合成方法与维度间相关(*F*(12, 124)=198.99，*P*<.001，*η*2=0.951)的交互作用显著。简单效应分析结果表明，维度间相关较小时(相关为 0.0)，MIRT 法、Bifactor-M1 法、Bifactor-M2 法的误差均方根差异较大(*P*<.001)，随着维度间相关的增加(相关为 0.9)，MIRT 法、Bifactor-M1 法、Bifactor-M3 法的误差均方根较为接近(*P*>.001)，且小于 Bifactor-M2 法(*d*=-.123, *P*<.001, *d*=-.117, *P*<.001, *d*=-.123, *P*<.001)。对于 MIRT 法（如维度相关 0.0 和维度

相关 0.9: d=.158, P<.001）、Bifactor-M1 法（如维度相关 0.0 和维度相关 0.9: d=.416, P<.001）和 Bifactor-M3 法（如维度相关 0.0 和维度相关 0.9: d=.253, P<.001），随着维度间相关增加，误差均方根显著减小，对于 Bifactor-M2 法，维度相关为 0.9 和维度相关为 0.7(d=.011, P>.001)，维度相关为 0.9 和维度相关为 0.5(d=-.008, P>.001)条件下的误差均方根差异不显著。

维度分合成的简单效应分析（以维度 1 为例）

（1）维度分合成方法与样本量(F(8, 160)=3.20，P<.05，η2=0.138))的交互作用显著。简单效应分析结果表明，样本量为 500 时，MIRT 法和 Bifactor-M3 法的误差均方根的差异相对较大(MIRT 法和 Bifactor-M3 法，d=-.046, P<.001)；随着样本量增大，MIRT 法和 Bifactor-M3 法的误差均方根的差异减小(样本量为 2000，d=-.031, P<.001)。对于 Bifactor-M1 法，样本量为 500 条件下的误差均方根显著大于样本量为 1000 条件下的误差均方根(d=.023, P<.001)，显著大于样本量为 2000 条件下的误差均方根(d=.026, P<.001)，对于其他方法，各样本量条件下误差均方根差异不显著。

（2）维度分合成方法与测验长度(F(8, 160)=62.91，P<.001，η2=0.759)的交互作用显著。简单效应分析结果表明，测验长度为 18 题时，MIRT 法和 Bifactor-M2 法(d=.017, P>.001)、MIRT 法和 Bifactor-M3 法(d=.006, P>.001)、Bifactor-M2 法和 Bifactor-M3 法(d=-.011, P>.001)、Bifactor-M2 法和 Bifactor-M4 法(d=.019, P>.001)的误差均方根没有显著差异；测验长度为 60 题时，MIRT 法误差均方根显著小于 Bifactor-M2 法(d=-.028, P<.001)、小于 Bifactor-M3 法(d=-.084, P<.001)，Bifactor-M2 法的误差均方根显著小于 Bifactor-M3 法(d=-.056, P<.001)，大于 Bifactor-M4 法(d=-.039, P<.001)。对于 MIRT 法、Bifactor-M2 法、Bifactor-M3 法和 Bifactor-M4 法，随着测验长度增长，误差均方根显著减小（如 MIRT 法: 测验长度 18 题与 36 题: d=.153, P<.001；测验长度 36 题与 60 题: d=.078, P<.001），对于 Bifactor-M1 法，测验长度为 36 题和 60 题条件下的误差均方根差异不显著(d=.006, P>.001)。

（3）维度分合成方法与维度间相关(F(16, 160)=450.18，P<.001，η2=0.978)的交互作用显著。简单效应分析结果表明，随着维度间相关的增加(相关为 0.9)，MIRT 法(d=.073, P<.001)、Bifactor-M2 法(d=.099, P<.001)的误差均方根显著大于 Bifactor-M4 法，Bifactor-M3 法的误差均方根与 Bifactor-M4 法没有显著差异(d=.004, P>.001)。仅对于 Bifactor-M1 法，随着维度间相关增加，误差均方根显著增加（如维度相关 0.0 和维度相关 0.9: d=-.517, P<.001），其他方法在维度间相关各条件下的误差均方根差异存在不显著的情况（如 Bifactor-M1 法，维度相关 0.0 和维度相关 0.9: d=-.016, P>.001）。

由于在模拟设计下，存在的交互作用、简单效应较多，而文章的重点在于关注各合成方法表现之间的比较（文章前言部分提到“采用模拟研究的方法对所提出的各方法以及基于 MIRT 的估计方法进行了比较，期望能够对所提出的方法有更加深入和细致的认识，同时也能为实际使用者提供应用建议”），且考虑到文章篇幅限制的因素，在结果部分仅介绍以合成方法为关注点的交互作用和简单效应。在修改稿中以脚注的形式加以说明，具体请参考修改稿第 9 页标红的脚注。

意见 3：在 5.1.1 误差均方根部分，作者谈到“结果显示，不同总分合成方法的主效应显著(F(3，124)=873.60，P<.001，η2=0.955)，多重比较结果表明，MIRT 法的误差均方根显著小于其他方法，其次是 Bifactor-M3 法，然后是 Bifactor-M1 法和 Bifactor-M2 法，二者没有显著差异。交互作用分析结果如图 2 所示。”，这里有两个问题。问题一，请作者指明是用的哪一种多重比较方法比较的，更为规范。问题二，作者在这里讲到了主效应和交互作用分析，请问，作者的主效应检验是单独做了一个单因素四水平的方差分析得到“不同总分合成方法的主效应显著(F(3，124)=873.60，P<.001，η2=0.955)，多重比较结果表明，MIRT 法的误差均方根显著小于其他方法，其次是 Bifactor-M3 法，然后是 Bifactor-M1 法和 Bifactor-M2 法，二者没有显著差异”结论的吗？如果是，请说清楚。然后交互作用分析，是做的 4*3 的方差分析来分析方法与样本量，方法与测验长度的交互效应吗？如果是这样，请说清楚。否则，读者会误认为，作者是做了一个方差分析，然后即介绍了主效应，然后又介绍了交互效应，这样就不合理了。因为交互效应显著，主效应就没有介绍的意义了。

回应：谢谢审稿专家的意见。对于问题一，确实应当指出使用的多重比较的方法较为规范，因此在修改稿中进行了补充，研究使用的是多重比较的 LSD 法、bonferroni 法，具体请参见修改稿第 11 页、第 14 页标红部分。对于问题二，方差分析的确应当先看交互作用的结果，但是正如文章前言中提到的，研究主要关注“各方法对总分和维度分估计的准确性”，即使交互作用显著，作者也期望回答方法之间是否存在显著差异的问题，因此方法的主效应是研究关注的重点，也对此进行了报告。根据审稿专家的意见，修改稿中对结果报告的顺序进行了调整，先报告交互作用结果，再报告主效应结果，以回答研究关注的主要问题。研究在借鉴类似模拟研究和使用了方差分析的实验研究的基础上，建立了一个同时包含主效应和交互作用的完整模型进行方差分析，并同时报告主效应和交互作用的结果。例如，Yao(2011)、Jimmy, Song & Hong (2011)的模拟研究结果部分既介绍了模拟条件的交互作用又明确指出不同方法之间存在显著差异。又例如《心理学报》2016 年的两篇文章（尹华站等，2016；王爱君等，

2016），都使用了方差分析的方法，在结果部分均先报告了主效应结果，再报告交互作用结果。

参考文献：

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

王爱君, 沈路, 迟莹莹, 刘晓乐, 陈骐, & 张明. (2016). 听障和听力正常人群空间主导性和空间参照框架的交互作用. *心理学报, 48*(2), 153-162.

尹华站, 李丹, 陈盈羽, & 黄希庭. (2016). 1～6 秒时距认知分段性特征. *心理学报*(9).

第三轮

审稿人 2 意见：

意见 1：文章中存在一些小错误，例如，参考文献不符合 APA 格式，正文中序号不连贯，图 1 中指标下标错误等等，请作者自查，注意细节。

回应：谢谢审稿专家仔细的阅读和指正。已经对文章进行了反复阅读和修改。重点修改了参考文献的格式，正文中的序号（详见文中标红的序号）。另外，对图 1 中的下标含义进行了详细说明，具体请参见第 4 页标红部分。

意见 2：在 4.2 数据生成部分，作者谈到，每种条件下重复模拟 30 次，重复模拟的次数会不会太少？请作者指明重复模拟 30 次的依据和参考文献。

回应：谢谢审稿专家仔细的阅读。在第一次审稿中，审稿专家也曾提出过关于 Monte Carlo 模拟研究的问题（见第一次审稿意见，审稿意见一，第 3 个问题）。本研究每种条件下数据重复模拟 30 次，参考了多个类似模拟研究的重复次数设置，如 Huang(2015)，刘玥，刘红云(2013, 2012), 詹沛达等 (2016)的模拟研究中每种条件下重复模拟 30 次，Yao 和 Boughton(2009)、Yao(2010)、Yao(2011)的模拟研究中每种条件下重复模拟 20 次，Jimmy 和 Song(2009)，Jimmy, Song 和 Hong(2011)的模拟研究中每种条件下甚至只重复模拟 1 次。因此，作者认为，每种实验条件下数据重复模拟 30 次应该能够忽略模拟的随机误差的影响，从而较好地反映估计结果的真实情况。

参考文献：

Huang, H. Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement, 39*(5),362-372.

Jimmy, D. L. T., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: a higher-order irt model approach. *Applied Psychological Measurement, 33*(8), 620-639.

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L., & Boughton, K. (2009). Multidimensional linking for tests with mixed item types. *Journal of Educational Measurement, 46*(2), 177-197.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement,35*(1), 48-66.

刘玥, 刘红云. (2012). 贝叶斯题组随机效应模型的必要性及影响因素. *心理学报, 44*(2), 263-275.

刘玥, 刘红云. (2013). 不同铆测验设计下多维 irt 等值方法的比较. *心理学报, 45*(4), 466-480.

詹沛达, 陈平, 边玉芳. (2016). 使用验证性补偿多维 irt 模型进行认知诊断评估. *心理学报 48*(10), 1347-1356.

意见 3：诚如作者所言，该研究的一个局限性是没有考虑维度数的影响。因为研究的对象本身是多维测验，那维度数的影响应该会很大才对。作者在摘要中明确说明 “双因子模型可以同时包含一个全局因子和多个局部因子，在描述多维测验结构时有其独特的优势”。请作者给出明确的理由，为什么这么重要的一个因素，本文没有考虑呢？

回应：谢谢审稿专家仔细的阅读。研究暂没有考虑维度数的影响，主要是出于以下两个方面的考虑。

第一，首先，根据研究提出的几种总分和维度分合成方法的定义，总分的合成会使用各维度的信息，而维度分的合成则更多依赖于各维度自身的信息，其他维度的信息仅有较为间接的影响。因此，作者推测维度数对维度分合成的影响相对较小，在本次模拟研究中暂不考虑。其次，基于以往类似研究的结论，在 Jimmy, Song 和 Hong(2011)的模拟研究中，虽然考虑了维度数的影响，但是在他们两种维度数（2 个维度，5 个维度）的条件下，各种方法的

表现都有相似的规律。考虑到本研究的重点和结合以往研究的结论，研究设计中固定维度数进行模拟。

第二，研究的模拟部分是一个基于实际数据的模拟，正如文中所述“题目参数选自某地区高考理综测验题目参数库，……包含 3 个维度”。目前我国的高考理综考试即包含物理、化学、生物三个分维度，研究想先通过这种基于现实数据的模拟，更接近多维测验的实际情况，为实际的教育测验提供参考。因此模拟研究中，作者将维度数固定为 3 个。

研究首次提出了总分和维度分合成的几种方法，因而将研究的重点放在方法的比较上，在此基础上结合相关文献的模拟研究设计，选择了样本量、测验长度、维度间相关这几个因素。最后，在研究局限和展望中我们进一步指出，可以在后续研究中考察维度数不同条件下各方法的表现，使得对总分和维度分合成方法的研究结果更加完善和丰富。

参考文献：

Jimmy, D. L. T., Song, H., & Hong, Y. (2011). A comparison of four methods of irt subscoring. *Applied Psychological Measurement, 35*(4), 296-316.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores.*Journal of Educational Measurement, 47*(3), 339-360.

第四轮

审稿人 2 意见：

意见 1：请将第三次审稿意见的第二个问题的回答反映在正文中。目前作者的解释“研究每种条件下数据重复模拟 30 次，参考了多个类似模拟研究的重复次数设置，如 Huang(2015)，刘玥，刘红云(2013, 2012), 詹沛达等 (2016)，Yao 和 Boughton(2009)、Yao(2010)、Yao(2011)，Jimmy 和 Song(2009)，Jimmy, Song 和 Hong(2011)”没有反映在正文中。

回应：谢谢审稿专家仔细的阅读和指正。已将关于模拟重复次数的解释反映在正文中，详见第 8 页和参考文献标红部分。

意见 2：请将第三次审稿意见的第三个问题的回答反映在正文中，目前作者的解释“研究暂没有考虑维度数的影响，主要是出于两个方面的考虑”没有反映在正文中。

回应：谢谢审稿专家仔细的阅读和指正。已将研究暂未考虑维度数影响的原因反映在正文中，详见第 8 页标红部分。

意见 3：既然作者综合考虑后认为“维度数对维度分合成的影响相对较小”，“维度数（2 个维度，5 个维度）的条件下，各种方法的表现都有相似的规律”，那么文中，这样的说法“其次，维度数也是可能影响分数合成方法的重要因素(Jimmy, Song & Hong, 2011)，在研究中维度数是固定的，今后的研究可以通过将维度数设为自变量对其影响情况进行探索。”就要商榷了。还有没有必要这样讲？

回应：谢谢审稿专家仔细的阅读和指正。在综合考虑后，推测维度数对维度分合成的影响相对较小，与文中“其次，维度数也是可能影响分数合成方法的重要因素(Jimmy, Song & Hong, 2011)”的说法确有矛盾之处。已将这段话删除，详见第 22 页标红部分。

意见 4：在数据生成部分，作者说“所有条件下的测验都为简单结构，包含 3 个维度”，固定 3 个维度的原因也有必要说明。

回应：谢谢审稿专家仔细的阅读和指正。已将所有条件下测验固定为 3 个维度的原因反映在正文中，详见第 8 页标红部分。

第五轮

编委审稿意见：I have polished the English abstract and suggesting removing 2 characters in the Chinese title. Please recommend the author to remove 2000 words from the article (it is now 13000 words a bit on the long side).

回应：谢谢编委专家细致的修改。已经接受了英文摘要和中文题目的修改，并对全文字数进行了删减，删减了约 2000 字。