《心理学报》审稿意见与作者回应

题目：概化理论缺失数据方差分量估计

作者：张敏强、张文怡、黎光明、刘晓瑜、黄菲菲

第一轮

审稿人 1 意见：应用多种方法对缺失数据情形下方差分量的估计进行比较有一定的意义。

意见 1：MCMC 方法中的建议分布（proposal distribution）是什么？

回应：参考 Mislevy(2001)，本文设定模型的先验分布，这种 MCMC 方法称为有先验信息的 MCMC 方法（MCMC with informative priors，MCMC inf），先验信息服从温和的逆τ 分布，其平均数设为 2，标准差设为 1，即 $p \sim \tau(2,1)$，$i \sim \tau(2,1)$，$r \sim \tau(2,1)$，$pi \sim \tau(2,1)$，$ir \sim \tau(2,1)$，$pr \sim \tau(2,1)$，$pir \sim \tau(2,1)$，所有初始值均设定为 0.001。

意见 2：作者在对式（1）中变量进行解释时说道："$\tilde{n}_{\alpha}$ 为α 某一水平的观察值的总数量；……；$\alpha\varepsilon$ 为设计中的所有层面。"请作者针对缺失数据情形说明得更清晰些，否则显得模糊。

回应：$X_{\alpha\varepsilon}$ 为观测分数，$\alpha\varepsilon$ 为设计中的所有层面，$\tilde{n}_{\alpha}$ 为α 某一水平的观察值的总数量，ε 为除α 外的其他层面。对于测量目标 p 来说，$X_{\alpha\varepsilon}$ 为 X_{pir}，α 为 p，ε 为i 和r。于是测量目标 p 的T 值公式为$T_{(p)} = \sum_{p}[\frac{(\sum_{i}\sum_{r} X_{pir})^2}{n_p}]$。类似地，其它 6 个测量侧面的$T$ 值可以按相同的逻辑推导出来。

意见 3：在分析与讨论部分，折线图是取各种情形的平均数算得吧，比如图 2(a) 被试方差分量估计偏差中，三条折线分别代表了不同被试样本容量，其上的每个点应该是不同评分者和不同题目数下的平均数吗？需要说明，如果这样，也应该提供标准差，因为一平均就掩盖了估计结果的稳定性特征。

回应：在分析与讨论部分，折线图是取不同评分者和不同题目数下的平均数算得的。对于每个点为了反映其稳定性特征，既要考虑方差分量的平均数，也要考虑方差分量的标准差。作者行文中未将标准差罗列出来，但在实际作比较时已考虑标准差。例如，对于图 2（a）,在比较三条折线各个方差分量的偏差改变时，以 p 为例，其三组被试的方差分量平均数和标准差（$M \pm SD$）如下：-0.0516±0.1075、-0.0513±0.1056、-0.0558±0.1123，经检验差异不显著（F =0.005, p =0.995）。对于图 2（a）,还需要考虑其它六个方差分量在不同被试情况下的变化，如果都将它们的平均数和标准差（$M \pm SD$）列出，显得篇幅较长。综合七个方差分量，图 2（a）三组被试的方差分量平均数和标准差（$M \pm SD$）如下：-0.00017±0.013586、-0.00321±0.010925、0.00816±0.005706，经检验差异不显著（F =2.778, p =0.082）。

实际上，为了反映七个估计的方差分量的稳定性特征，计算四种方法估计的方差分量标准误或方差分量置信区间包含率，将更加有效和准确（Wiley, 2001; Brennan, 2001）。但这是本文后续需要进一步研究的，即概化理论缺失数据方差分量变异量估计，包括标准误估计和置信区间估计。与其它统计量一样，概化理论下估计出的方差分量受限于抽样，不同的抽样

样本，所估计的方差分量可能不一样，这就要求进行方差分量估计时需要对其变异量进行探讨。以标准误估计为例，因为是针对多个大循环（可能套有小循环）获取的多批模拟数据，所以每个样本都能估计出相应的方差分量，对每个样本估计的方差分量求标准差，实际上就是方差分量标准误。但是，公式法、REML 法、拆分法和 MCMC 法估计方差分量标准误存在少许不同，作如下分述：

第一，对于公式法。本文规定模拟次数为 1000 次（大循环），为了获得公式法的方差分量估计值，需要对每一批大循环数据估计相应的 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir 共七个方差分量，这样 1000 批数据就有 1000 个 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir，然后分别求取 1000 批数据 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir 的平均，即为公式法得到的七个方差分量（最后）估计值。为了估计公式法的方差分量标准误，需要使用七个方差分量（最后）的平均估计值，再运用 Brennan(2001，p. 181) 标准误公式求出七个方差分量对应的标准误估计值。

第二，对于 REML 法和拆分法。对于这两种方法，因为存在模拟多批数据，且每个大循环（共 1000 次）套有小循环。对于大循环，1000 次小循环基于一定的再抽样，可获得 1000 个 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir 共七个方差分量，然后获得小循环的 SE（VC.p）、SE（VC.i）、SE（VC.r）、SE（VC.pi）、SE（VC.ir）、SE（VC.pr）和 SE（VC.pir）。对于每个大循环，都是如此进行。因为大循环为 1000 次，这样就有 1000 个小循环后的七个方差分量估计值和 1000 个小循环后的标准误估计值。分别求平均，即为最后的七个方差分量估计值和方差分量标准误估计值。对于 REML 法和拆分法，需要在大循环中构建小循环，先求出小循环的方差分量及方差分量的标准差，待大循环结束后，再对方差分量及方差分量的标准差求平均，即为最后估计出的方差分量及方差分量的标准误。

第三，对于 MCMC 方法。对于 MCMC 方法估计方差分量及方差分量的标准误，也是先求出小循环的方差分量及方差分量的标准差，待大循环结束后，再对方差分量及方差分量的标准差求平均，最后获得估计的方差分量及方差分量的标准误。与 REML 法和拆分法不同的是，MCMC 方法估计方差分量及标准误，不是设定 1000 次小循环，而是预先设定了三条马尔可夫链，每条链循环 10000 次（其中丢弃了 5000 次），通过 WinBUGS 软件计算每次大循环（共 1000 次）的方差分量及其标准差。对于每次大循环，可获得三条马尔可夫链的 5001～10000 次的 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir，分别对它们求平均，即为每次大循环的 VC.p、VC.i、VC.r、VC.pi、VC. ir、VC.pr 和 VC.pir，再分别对它们求标准差，即为每次大循环的 SE（VC.p）、SE（VC.i）、SE（VC.r）、SE（VC.pi）、SE（VC.ir）、SE（VC.pr）和 SE（VC.pir）。在 1000 次大循环结束后，最后分别求取 1000 个七个方差分量估计值的平均值，即为 MCMC 方法最后的方差分量估计值，分别求取 1000 个方差分量标准误的平均，即为 MCMC 方法最后的七个方差分量标准误估计值。

意见 4：分析与讨论中图的标题，比如“图 2(a) 被试方差分量估计偏差”是否存在错误表述？因为同一个结果图中应该包括了各个方差分量。

回应：确实存在表述错误，已作出修改。

意见 5：第二种方法中方差分量估计中的不收敛情形下的数据最好用平均数代替，或者忽略它，否则带来其他影响因素。

回应：第二种方法中方差分量估计中的不收敛情形下的数据的方差分量是忽略不收敛数据后用平均数求取的。作者在行文中已作出强调：第二种方法模拟数据迭代部分收敛，仅是一部分数据加权结果。

意见 6：结论中的第二条，题目和评分者的影响应该分方法进行分析，因为比如 MCMC 方法中这两者影响好像不大，虽然作者在分析中说有不同趋势，但这个应该有检验结果来支持。好像不同方差分量之间的差异倒是比较明显。

回应：本研究发现，增加题目数量与评分者数量都能在一定程度上提高方差估计的准确率，但增加题目数量比增加评分者数量更有效。以 MCMC 法为例，在保持题目数量为 2 不变的情况下，评分者总量增加一倍，从 5 位上升到 10 位，七个方差分量的偏差下降 0.1960。然而，保持评分者数量为 5 不变的情况下，题目数量增加一倍，从 2 题上升到 4 题，七个方差分量的偏差下降 0.3108，比增加评分者数量减少的偏差多 1 倍多。这说明增加题目数量比增加评分者数量对于提高测量准确度更为有效。进一步分析其它方法，也存在类似情形或大致趋势。

这些结果均提示，在考试测量中，如果受到人力、物力、财力等影响，要控制测评成本，可以优先考虑增加题目数量来提高测量准确度。题目数量仅需要少量上升，测量准确度就能有较大增幅。如果不能增加题目数量，可采取增加评分者数量的方法，以换取更准确的测量结果。一些研究者往往把关注点集中在评分者的信度上，以至于忽略题目变量，这是不足取的。实际上，人们可以多关注题目侧面在提高测量准确度方面的贡献。

另外，本文主要的目的是从整体上，比较和探讨公式法、REML 法、拆分法和 MCMC 法在估计各个方差分量上的总体性能优劣。同理，在比较题目和评分者影响效应时，也是综合了七个方差分量的结果，没有针对方法和方差分量作“交互作用”分析。因为是从整体上进行分析，所以没有过多地针对每个方差分量分别地详细地进行展开讨论。

审稿人 2 意见：

意见 1：模拟产生却是数据矩阵的第二步中，提到根据“高考评分规则”，这里的高考评分规则究竟是高考什么题型或内容的评分规则，可以说得更清楚一点。

回应：模拟产生却是数据矩阵的第二步中，提到的高考评分规则为：将评分者与试卷随机分组，系统随机抽取一份试卷分配给某位评分者，评分者进行评分后系统回收，再随机分配给另外一位评分者进行批改。

一般地，高考试卷评分中，客观题评分比较容易把握，如选择题等，多采用机器评分，评分不大具有争议性。人工评分多为主观题，如作文题、简答题和问答题等，评分者评分存在误差。这里的高考评分规则主要针对主观题题型，概化理论在分析评分者评定主观题分数一致性时具有优势，当然这依赖于概化理论方差分量的准确估计，特别是对于缺失数据的情形，这也是研究本文的初衷。

意见 2：图 2、4、5 的图名有歧义，最好改一下（见文中修订处）。

回应：已参考文中修订标识，对图 2、4、5 的图名已作修改。

意见 3：除图 2 外，图 3、4、5 的纵坐标偏差刻度可否小一点，这样不同情况的折线可以分开一点，读者更容易看出差异。

回应：感谢专家的意见。为了更为有效地估价方差分量估计的准确性，统一了“度量系统”，使得所有偏差值定义在一个量纲系统上。在本研究中，方差分量估计的偏差度量系统定义在[-1，+1]。之所以是[-1，+1]，是因为考虑到所有方差分量的总和接近于 1，偏差可能是+1 或-1 或两者之间。参照此度量系统（或量纲系统），为了便于统一比较公式法、REML 法、拆分法和 MCMC 法等四种方法估计的精度，所有的偏差值都定义在[-1，+1]的标刻中，而不管其采取何种方法，也不管其偏差是否非常接近，或分得不太开。另一种解释是，图 4 和图 5 的折线相对矮小或紧凑，这表明拆分法和 MCMC 法估计的方差分量差分量偏差相对较小，估计较为准确，或这两种方法估计七个方差分量相对较好。因此，作者为了便于比较各种方法的优劣，把偏差值都定义在[-1，+1]的标刻中，有两个好处：一是统一量纲的评价标准；二是真实反映偏差的相对大小。

意见 4：从论文看，其假想的应用对象主要是高考缺少数据矩阵情况。对于高考一般是分组阅卷，每组阅一部分试题，一般不超过 6 道，因此论文对试题数的选择是可以接受的。但一

个阅卷组评阅的试卷数可能远超过 5000，阅卷教师可能 20 人，甚至更多。由图 5（a）可见，当被试数量到 5000 时，p、ir、pir 的方差分量偏差明显增大，若被试到 5 万，甚至 10 万时，这些方差分量偏差以及其他方差分量偏差是否会大得多？当然，这一问题不影响本文目前的成果，但对成果的应用价值可能会有影响。

回应：由图 5（a）可见，相比于 200 和 1000 人，当被试数量为 5000 人时，p、ir 和 pir 的方差分量偏差有增大趋势。但是，p、ir 和 pir 的方差分量偏差未见明显的增大。采用非参数检验 Kruskal Wallis Test 对 200、1000 和 5000 人三组被试 p、ir 和 pir 的方差分量偏差进行检验，结果如下：$\chi^2(2)=1.819, p=0.403$；$\chi^2(2)=0.112, p=0.946$；$\chi^2(2)=1.149, p=0.563$，三组被试 p、ir 和 pir 的方差分量偏差差异都不显著（$p>.05$）。虽然 5000 人偏差有些增大，但增大的并不明显，且这种偏差相对于传统法仍是较小的，相对于 REML 法和拆分法也是可以接受的。

本文假想的应用对象主要是高考评卷，其数据存在实际缺失的情况。造成这种窘境的原因是,高考评分有其特殊性：考生众多，科目众多，阅卷评委也相对较多，不得不区分出一个个阅卷小组来“分而治之”。以某省 2013 年高考评分为例，共约 40 万考生参加了多门科目的考试，每门考试科目约 400 个教师参与阅卷评分，每小组阅一部分试题，一般不超过 6 道，每个阅卷小组约 5 人(约 80 个小组，10 个大组)，即对于一个阅卷小组而言，约有 5000 个被试，组成了一个缺失数据的矩阵。因为参考了高考评分的模式且区分出一个个阅卷小组，所以对于一个阅卷小组而言被试实际达到 5 万或 10 万的情况相对较少。若被试到 5 万，甚至 10 万时，这些方差分量偏差以及其它方差分量偏差是否会大得多？本文目前尚没有进行过研究，仍无法预知。

意见 5：英文摘要存在一些明显的语言错误，标注了一部分，需要认真修改。

回应：已参考文中的修订标识。对英文摘要作了修改，也对其它未标识之处进行了详细的检查并完善，详见正文。

审稿人 3 意见：

意见 1：公式法、REML 法、拆分法和 MCMC 法的结果是基于不同的软件或程序估计出来的，这本身会不会造成估计值上的差异？

回应：其实作者也很想找到一个四种方法都能使用的软件或程序，但目前尚不存在。这的确引发了一些问题，如专家指出的：方法所负载的软件或程序本身会不会造成估计值上的差异？虽然作者尚难以确认差异有多大，但是对于这种差异作者认为不必有过分的担心，理由有三:一是各种不同的方法相应的软件都是公开出版的，一些相关的文献都引用过，并得到了业内人士的普遍认可，其软件程序的算法是可信可靠的；二是本研究模拟的数据是相同的，只是估计的方法存在不同，若方法一样，只要公式使用得当，不同的软件或程序估计出来的结果差异将相对较小；三是国外学者有类似的做法，文献如下：

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modeling. Journal of Educational Measurement, 44(2), 131-155.

Journal of Educational Measurement （JEM）是比较权威的测量杂志，在 JEM 的这篇文章中，为了对比 GT（公式法）和 GIRM（MCMC 法），Briggs 和 Wilson 使用了以下软件：WinBUGS、R 软件、R2WinBUGS、GENOVA 和 urGENOVA 等。在该篇文章中没有细究不同软件本身带来的估计差异。

作者接下来又查了一些相关的国外关于 GT 的文献，如 Brennan,Yin, & Kane （2003）、Lee（2006）和 Lee & Lewis（2008）等，使用多个软件或程序估计方差分量的做法非常普遍。

这些文献均显示，多个软件一并使用的做法是可行的，这些文章并没有讨论不同软件或程序带来的差异问题，其原因是：目前尚无办法也找不到在同一软件实现所有方法的方差分量估计，只能选择不同的软件来实现，而这些软件的算法是得到公认的，是可靠可信的。

意见 2：在讨论部分，对传统法、拆分法和 MCMC 法的分析，呈现的都是被试人数、试题数、评分者数这三个变量在 7 个或 5 个方差分量估计上的差异，用的是平均值。而在 REML 上呈现的结果却是被试人数、试题数和评分者数交叉的单个结果。为什么？呈现的结果是基于不同条件的，不在同一层面上的东西能比较吗？就好比方差分析中，主效应的结果与简单效应的结果能相互比较吗？

回应：作者之所以这样做，是为了能够针对 REML 法被试情况探讨得更加充分。对于 REML 法，本文重在分析当被试量为 5000 时，试题数与评分者变化对方差分量估计的影响，更加强调当被试数量较大时，REML 法迭代存在部分数据不收敛且导致估计不准确的情形。

当然，对于 REML 法的结果，也可以呈现类似于图 2、图 4 和图 5 的结果。但是考虑到 REML 法存在不收敛，特别是当被试数为 5000 时方差分量估计值在各种情况下的变化情况较为特殊，重点想探讨这种情况，即重点想深挖随着被试量的增加，方差分量估计值在各种情况变化趋势。实际上，如果将 REML 法的三个图，即图 3(a)～(c)整合起来，就相当于类似于图 2（a）的情形，都是考虑随着被试数量的增加各种情况的变化，只不过图 2（a）是将 200、1000 和 5000 人平均化了，而图 3(a)～(c)是将类似于图 2（a）的情况作出了进一步的细化。

追根溯源，本文中的图是产生于对应的表的。虽然图 3 与图 2、图 4 和图 5 有一些差异，但对应的表 2～表 5 呈现形式却是相同的。源头是一样的，呈现的结果自然也是一样的，只不过在比较各种方法性能时，可以从不同的角度抓住不同的重点且考虑不同的情况来将信息有选择性地反映在图形中。

针对 REML 法的特殊性，即需要重点探讨当被试量为 5000 时，REML 法迭代存在部分数据不收敛的情况。因此，在分析中详细地展开了这部分。然而，对于题目（i）和评分者（r）则由于篇幅，没有过多地展开。但这并不表示作者不能类似于图 2、图 4 和图 5 那样进行详述地展开。但是，如果这样做，则不但增加篇幅，而且不能有效说明 REML 法的重点之处。另外，需要说明的是，其实在图 3(a)～(c)中已经包含了估计的题目（i）和评分者（r）方差分量信息，只是未重点类似于图 2(b)～(c)那样，针对题目（i）和评分者（r）进行不同情况细究详述，因为这不是 REML 法重点想探究的。

意见 3：既然比较的是四种方法的优劣，那为什么不直接比较四种方法呢？就是把方法作为一个变量，看方法在各个方差分量上的主效应以及跟其它变量的交互效应呢？

回应：本文是基于随机双面交叉设计 p×i×r 缺失数据，探讨四种方法在各种条件下估计方差分量的性能优劣，基本条件包括 P=200，1000、5000，i=2、4、6, r=5、10、20，方法包括公式法、REML 法、拆分法和 MCMC 法等四种方法。专家建议：宜探讨方法在各个方差分量上的主效应以及跟其它变量的交互效应，这个建议是合理的，在理论上是有意义的，通常也强调这一点。

但是，本文主要的目的是从整体上，比较和探讨公式法、REML 法、拆分法和 MCMC 法在估计各个方差分量上的总体性能优劣。因为是从整体上，所以更加关心其中某种方法（或某些方法）在各个方差分量估计上总体胜于其它方法，或相对较好，可以推荐给读者。由此看来，因为是从整体上考虑，所以是为从总体上选取某种合适方法而探讨的，不分局部或各个方差分量，所以未探讨方法与各个方差分量的“交互效应”。

除了需要“整体”考虑之外，之所以难以进行“交互效应”分析，还有另外一个原因，那就是在使用软件的实际中难以做到“分段”或对“各个分方差分量”人为地使用不同的方法，因为软件程序负载了方法，往往都是“固定”化了，尚难以做到“按需而设”，除非使用者从头至尾

针对每个方差分量按估计方法的“交互效应”要求，自行编制或自行开发源程序，但这是不方便的，代价是高昂的，一般的使用者是难以做到的。因此，通常更为现实的是，从四种方法中选择出整体上较佳的可以为读者推荐的方法来，且这种方法在常用的软件中可以实现。

第二轮

审稿人 1 意见：作者已经对提到的疑问进行了解释，但本人仍然有一个建议：

意见 1：作者本研究是模拟高考评分情形的，而关于高考评分中题目的合适数量和评分者合适人数的问题已经在很久以前有了相对比较好的研究结果。本研究的焦点只是‘缺失数据情形下’方差分量估计方法的问题，因此，即使题目或评分者数量会影响估计结果，也应该是在已有研究发现合适数量题目和合适数量评分者的基础上找到一种更好的估计方法，而不是为了估计结果好而增加题目或评分者（而这又是由于缺失数据情形下）。因此，在作者的结论部分的第二点：“题目和评分者是缺失数据方差分量估计重要的影响因素。在人力物力有限的情况下，可优先考虑增加题目数量，以获取更准确的方差分量估计。评分者数量影响次之，但不是越多越好，评分者数量保持在 4—6 人，有利于方差分量的准确估计。”建议改为“题目和评分者是影响缺失数据方差分量估计结果的重要因素。”后面的内容均可以删去。因为关于合适题数和合适评分者数的问题已经有了比较合适的结论。摘要也就需要相应进行修改。

回应：作者在文中认为：题目和评分者是缺失数据方差分量估计重要的影响因素。在人力物力有限的情况下，可优先考虑增加题目数量，以获取更准确的方差分量估计。评分者数量影响次之，但不是越多越好，评分者数量保持在 4—6 人，有利于方差分量的准确估计。作者得出这个结论是基于文中缺失数据分析的结果，是基于一个事实，即题目和评分者对各种方法估计缺失数据方差分量都有影响，判断孰重孰轻其实是参考各种方法结果所得，特别是基于 MCMC 方法结果所得。当然，正如专家所言，本文的焦点只是“缺失数据情形下”方差分量估计方法的问题。在估计方法及影响因素之间需要权衡研究的重心。因此，为了把研究问题重点化，作者采纳专家的建议，把比较公式法、REML 法、拆分法和 MCMC 法估计方差分量性能优劣作为重点来提，而把评分者和题目仅仅作为重要的影响因素来考虑，不再提及合适题数和合适评分者数的问题。

原文摘要仅有这句话：题目和评分者是缺失数据方差分量估计重要的影响因素。原文摘要没有出现专家建议需要删除的内容，即在人力物力有限的情况下，可优先考虑增加题目数量，以获取更准确的方差分量估计。评分者数量影响次之，但不是越多越好，评分者数量保持在 4—6 人，有利于方差分量的准确估计。因此，对于摘要作者就没有再作修改。

意见 2：在文章的结果呈现部分，多位审稿专家均提到一些问题，其中“4.2 REML 法估计的方差分量偏差分析”部分的结果呈现模式与其他三种结果的呈现模式不一致，建议作者改为与其他三种结果一致的呈现方式，这样更加直观。当然，这样结果分析也需要修改。

回应：对“4.2 REML 法估计的方差分量偏差分析"部分的结果呈现模式进行了修改，整篇文章图形与表皆统一起来了。修改后的图 3（a），实际上包括了修改前的图 3(a)～(c)，但因为有修改后图 3（a）详细的文字说明，仍然可以结合表 3，作出对于 200、1000 和 5000 人时 REML 法估计方差分量的偏差分析，并不影响结果讨论。

为了与其它三种方法相统一，REML 法也分析了不同题目数量方差分量估计偏差和不同评分者数量方差分量估计偏差，即图 3(b)和图 3(c)，详见正文。

意见 3：建议作者把分析与讨论部分的图形与前面的结果部分的表组合到一起去，因为图和表均是研究的直接结果，成为“结果与分析”部分，这样读者阅读起来更方便，“讨论”部分只要从整体上讨论四种方法的结果比较。

回应：作者认为，按结果、分析与讨论来写也是可以的，其原因是：结果与分析讨论分开来，

对于每一种方法的讨论将更加充分。专家所提意见具有一定的合理性与建设性，但强调的是形式上的改变，并无实质性内容上的改变。考虑到重新组织对原文的结构改变较大，但内容却是一样，也考虑到按结果、分析与讨论来写并不会影响读者整体上的阅读。因此，这部分暂未作改变，请理解。