心理学探新 2020,Vol.40,No.3,254 -260
PSYCHOLOGICAL EXPLORATION

概化理论预算限制下最佳样本量估计*

黎光明
(华南师范大学心理学院,心理应用研究中心,广州 510631)

摘 要:概化理论广泛应用于各种心理测评实践中。当有预算限制时,概化理论需要考虑如何设计一个测量可靠性相对较高且可行性也相对较强的测量程序,这就要求通过某些途径估计最佳样本量。拉格朗日乘法是概化理论预算限制下最佳样本量估计较为成熟的方法。探讨了概化理论预算限制下最佳样本量估计的一些影响因素,如受总预算舍入的影响等,也提出了一些后续改善的建议,如推导出拉格朗日乘法的统一公式等。

关键词:概化理论;预算限制;最佳样本量估计;拉格朗日乘法;心理测评
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 -5184(2020)03 -0254 -07

1 引言

概化理论(Generalizability Theory)广泛用于各种心理测评实践中(黎光明,张敏强,张文怡,2013; Gage, Prykanowski, & Hirn, 2014; Maulana, Helms - Lorenz, & Grift, 2015; Wolbing & Riordan, 2016)。概化理论认为,研究测量必须先研究测量情境关系。概化理论提出,测量情境关系是由测量目标和测量侧面组成的(戴海崎,张锋,陈雪枫,2007)。一般来说,进行测量时增加测量侧面的观察数量,可以减小测量误差,并使测量的可靠性(如概化系数)随之增加(Lakes,2013)。然而,在实际进行测量时,概化设计中各侧面的观察数量与进行测量所需的费用密切相关,观察数量增加,测量所需的费用也随之增加。对于测量而言,这是一个两难问题。

在概化理论中,随着侧面水平数量的增加,概化系数会随之提高,直至这个潜在的增长最终达到理想的数值。然而,如果在研究过程中存在限制条件,比如受人力、物力、财力等所限(预算有限),那么就需要权衡是否需要对研究设计做出改变。在某些情况下,让概化系数提高 0.01 所要增加的某一侧面的观察数量较大,这时需要的经费可能超出预算。当出现这种情况时,就需要权衡增加侧面的观察数量的必要性(Brennan,2001)。由此看来,预算和成本是进行测量研究时不可忽略的问题,预算的高低在某种程度上会影响测量结果的正确性。在预算限制下,找到一个高可靠性的测量程序是研究者关注的主要问题之一。

一般地,增加样本量有助于增加测量的精度,提高测量的可靠性。但是,在实际测量中,由于受到时间、经费等因素的限制,无穷尽地增加样本量以提高测量的可靠性的方法是不可行的,需要找出在各种条件限制下进行测量研究的最佳样本量,以便在实际条件限制的情况下最大程度地确保测量的效用(Marcoulides,1991,1995)。那么,这就要求在设计研究程序过程中,需要同时考虑测量的可靠性及预算的问题。当有预算限制时,需要考虑如何设计一个测量可靠性相对较高且可行性也相对较强的测量程序(Marcoulides,1993,1997)。

2 概化理论预算限制下最佳样本量估计

2.1 概化理论

概化理论自提出以来日益受到关注,其最大的贡献在于可以排除严格平行测验的假设,分解测量中多种误差,使其可以在同一个分析中分别估计不同的误差来源,并且可以指导决策者选择最优测量方案。但是,当测量过程受到测验条件限制时(如需要减少侧面的水平数),概化理论测量的可靠性就可能会受到影响(Harik, Clauser, Grabovsky, Nungester, Swanson, & Nandakumar, 2009)。

概化理论将所要测量的心理特质水平称为测量目标,而构成测量条件与具体情境关系的因素称为测量侧面(Brennan,2001)。使用概化理论进行研究时,一般分为两步:概化研究(G 研究)和决策研究(D 研究)。概化研究是决策研究的基础,其主要目的是在观测全域上尽可能地辨明研究设计中各种测量的误差来源,并估计出方差分量。概化理论需要考虑测量设计中可能对测量目标产生影响的条件,并探究概化设计中各测量侧面方差分量对总方差的贡献比率。决策研究是概化研究的深化,是以概化

* **基金项目**:国家自然科学基金面上项目(31470050),教育部人文社会科学研究规划基金项目(18YJA190006),广东省哲学社会科学"十三五"规划一般项目(GD17CXL01),广州市哲学社会科学"十三五"规划一般项目(2017GZYB111)。
通讯作者:黎光明,E - mail:lgm2004100@ sina. com。

研究所得的方差分量为基础，决策如何舍弃设计中不必要的侧面，调整测量过程中各侧面的关系等。相比于经典测验理论，概化理论适用性较为广泛，不受正态分布、分数效应独立等要求的限制，可以同时考虑多个因素，能将因素试验设计及其分析、方差分量模型等统计工具应用到心理测评中，并结合测量情境关系对经典测验理论给出的笼统误差进行有效分解。类似于经典测验理论的信度系数，概化理论用可靠性的概念替代了传统信度的概念。概化理论的可靠性是指从一次测量（如行为观察、意见调查等）的被试得分拓广到施测同等程度的所有可能条件下的被试得分均分的精确性。概化理论的可靠性可用概化系数 $E\rho^2$ 来表示，被定义为全域分数方差 $\sigma^2(\tau)$ 与其和相对误差方差 $\sigma^2(\delta)$ 两者之和的比率，即：

$$E\rho^2 = \frac{\sigma(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \quad (1)$$

对于公式(1)，$\sigma^2(\tau)$ 为全域分数方差，表示测量目标的分数变异；$\sigma^2(\delta)$ 为相对误差方差，表示与测量目标有关的测量侧面交互作用的变异之和。

经典测验理论使用信度来衡量测验的一致性，即测量结果的可靠性。在经典测验理论中，信度系数越大，在一定程度上说明测量的可靠性越高，这就表明在使用经典测验理论进行测量研究时，应尽可能减少误差，使信度系数尽可能大。同样地，使用概化理论进行测量研究时，为保证测量的可靠性，也应使概化系数尽可能大。

当全域分数方差 $\sigma^2(\tau)$ 增大或者相对误差方差 $\sigma^2(\delta)$ 减小时，概化系数 $E\rho^2$ 增大。这就是说，要使概化系数 $E\rho^2$ 增大，那么就要使相对误差方差尽可能小。在概化理论中，减小误差最简单的解决方法是增加样本量以平衡误差。但是，在实际研究中，由于受到研究时间、研究经费等因素的限制，无穷尽地增加样本量来提高测量的可靠性的做法是不可行的。研究者需要找出在各种条件限制下的最佳样本量。

2.2　概化理论预算限制下最佳样本量估计

2.2.1　单元概化理论预算限制下最佳样本量估计

关于在预算限制下找到最优测量程序的问题早在20世纪70年代就有研究者涉足。Cleary 和 Linn (1969)提出，在单侧面概化研究中选择最大题目数量即可在预算限制下获得最大概化系数。但是，这种方法在增加一个侧面时（即双侧面设计）变得极为复杂，表现出低可行性。Woodward 和 Joe(1973)使用约束优化的方法计算测量研究的样本大小，并推导出计算样本大小的方程，即在预算限制下可以计算最大限度地提高测量可靠性的样本大小。不足的是，该方法只能应用于两侧面或三侧面的交叉设计，不能推广至超过三侧面及以上的交叉设计或嵌套设计中。在三侧面及以上的交叉设计或嵌套设计中要解决最佳样本量估计问题，就不得不使用其他方法，如柯西－希瓦兹不等式方法(Saunders, Theunissen, & Baas, 1989)、拉格朗日乘法(Marcoulides & Goldstein, 1990)等。

Marcoulides 和 Goldstein(1990)在 Woodward 和 Joe(1973)的基础上提出了一种在预算限制下求解最佳样本量问题的方法——拉格朗日乘法，呈现了一个两侧面设计案例和一个三侧面设计案例，说明了拉格朗日乘法如何有效估计最佳样本量。然而，Marcoulides 和 Goldstein(1990)只考虑了随机模型下的两侧面和三侧面设计，没有说明此方法在其他概化设计中的应用。此外，Marcoulides 和 Goldstein (1990)在其研究中未曾提到拉格朗日乘法用于解决问题的原理，在某种程度上导致所提出的方法欠缺说服力。Goldstein 和 Marcoulide(1991a, 1991b)通过一个三侧面设计案例论证了拉格朗日乘法在解决预算限制下使概化系数最大化问题中的合理性，并呈现了一个用拉格朗日乘法解决预算限制下最佳样本量问题的四侧面方案。同时，Goldstein 和 Marcoulide(1991a, 1991b)也阐述了二分法在解决受条件限制测量设计的最优化问题，并使用一个三侧面设计来反映二分法解决最优化问题的迭代效果，论证了二分法在决策研究中的适用性。

Goldstein 和 Marcoulide(1991a, 1991b)的研究只涉及完全交叉设计，并没有说明用于求解预算限制下最佳样本量的拉格朗日乘法在非完全交叉设计中的适用性。Marcoulides(1993)通过一个单侧面设计，完整地呈现了如何通过概化设计获取方差分量，更为深入地说明了如何在预算限制下使用拉格朗日乘法求解最佳样本量的过程。Marcoulides(1993)更加明确了拉格朗日乘法在预算限制下求解最佳样本量的使用方法，很大程度地帮助了之后的研究者更加详细地了解该方法，以便他们在实际研究中能够更好地应用该方法。

2.2.2　多元概化理论预算限制下最佳样本量估计

在心理测评中，一个测验经常会涉及多个分数，比如韦氏成人智力量表第3版(Wechsler Adult Intelligence Scale— III, WAIS－III)使用14个分测验来测量被试四个方面的能力，即言语理解能力、知觉组织能力、工作记忆能力和加工速度能力。对于类似于 WAIS－III 在一个测验中包含有多个分测验分数的情况，Goldstein 和 Marcoulides (1990, 1991a, 1991b)提出的求解预算限制下单元设计(univariate design)的最佳样本量的方法，就变得不再适用。针对此种情况，Marcoulides 和 Goldstein(1992)对单元设计预算限制下有关最佳样本量的求解方法进行了

拓展,发展出了多元设计(multivariate design)预算限制下如何求解最佳样本量的程序。相比于单元设计,多元设计预算限制下求解最佳样本量,需要涉及不同侧面观测数量的权重问题,这让问题变得更为复杂。

Marcoulides 和 Goldstein(1991,1992)的研究并没有解决如何在特定的预算限制下选择权重方案的问题。其实,多元概化理论最佳样本量估计权重方案的选择,可以根据经验来确定,但这种方式对于决策者的经验要求较高。Marcoulides 和 Goldstein(1991,1992)认为,即使是根据经验来确定多元概化理论最佳样本量估计的权重方案,也需要在理论研究结果的指导下完成。Marcoulides(1994)提出了预算限制下估算多元概化理论最佳样本量的方法,进一步探讨了多元概化理论研究中侧面最佳样本量估计不同权重方案的选择对预算限制下多元概化设计研究结果的影响。通过比较,Marcoulides(1994)认为,使用 Marcoulides 和 Goldstein(1990,1991)提出的拉格朗日乘法求得的最佳样本量,不论是选择哪种权重方案,所求得的概化系数结果都极为相似。Marcoulides(1994)的研究论证了,预算限制下估算概化设计侧面最佳样本量以获得最大测量可靠性的拉格朗日乘法,同样也适用于多元概化理论的研究中。

Marcoulides(1995)拓展了 Marcoulides 和 Goldstein(1990,1991)的有关研究,将他们提出的应用于单元概化设计预算限制下估算最佳样本量的方法,发展至多元概化设计中。Marcoulides(1995)提出,预算限制下可以使用平均误差方差——协方差分量最小的方法,将所得的最佳样本量应用于公式中。以 p × i 单侧面多元概化设计为例,其平均误差方差——协方差可表示为

$$a'\sum\nolimits_{\bar{X}}a = \frac{a'\sum\nolimits_{p}a}{n_p} + \frac{a'\sum\nolimits_{i}a}{n_i} + \frac{a'\sum\nolimits_{pi}a}{n_p n_i} \quad (2)$$

在公式(2)中,a 表示决策者选择的满足条件的加权系数;a' 表示加权系数列向量(特征向量);n_p 表示被试数量;n_i 表示题目数量。

Marcoulides(1995)的研究简化了其之前提出的方法,使得最佳样本量估计的表达式更加简单明了。

在一些研究中,可能会出现同一侧面的不同水平所需费用不一致的情况。在多元概化理论中,当固定某一侧面的预算时,受预算限制的表达式会相应改变。以 $p \times i$ 多元概化设计为例,如果固定项目 i 这一侧面的样本预算,那么预算限制的表达式将由 $cn_p n_i \leqslant B$ 变换为 $cn_i + cn_p n_i \leqslant B$。同样地,这种简洁的表达式也可以推广至双侧面设计中。在先前的这些研究中,研究者都是假定研究中同一个侧面的每个水平的预定费用是相同的。但是,当同一个侧面不同条件所需的成本不同时(如为了测量某种行为,不同的被试所需被试费用不同),那么之前所提出的方法均不可适用,因为当出现同一侧面不同水平所设定的预算不同时,相应地,预算限制的表达式也会发生改变。针对这种情况,于是 Marcoulides(1997)又提出了,当研究中同一个侧面内每个水平的预定费用是可变时获得测量最大可靠性的方法。以 $(o:p) \times r$ 为例,如果两个场合 o 的费用不同,那么其预算限制表达式可表示为 $c_1 n_{o1} n_{r1} + c_2 n_{o2} n_{r2} \leqslant B$。Meyer,Liu 和 Mashburn(2014)在前人研究的基础上,通过一项教师教学水平评价的实证研究,将 Marcoulides(1997)先前提出的关于预算限制下估算概化理论研究最佳样本量的方法应用于实际测量中。在实际应用中,除了简单的交叉设计外,Meyer,Liu 和 Mashburn(2014)还加入了一些相对较为复杂的嵌套设计,用以说明该方法在概化理论研究中的广泛适用性。

随着研究的推进,关于预算限制下使用拉格朗日乘法求解概化理论研究中侧面最佳样本量的方法表达越来越简便,决策者在决策最佳样本量的过程中也更为方便。然而,关于这方面的研究更多地探讨了交叉设计,无法说明拉格朗日乘法在预算限制下求解其它设计最佳样本量的适用性。

3 概化理论预算限制下最佳样本量估计拉格朗日乘法

在概化理论预算限制最佳样本量估计方法中,拉格朗日乘法发展最晚,也是相对最为成熟的最佳样本量估计方法。拉格朗日乘法是数学最优问题中一种寻找变量受一个或多个条件所限制的多元函数求极值的方法。通过引入拉格朗日乘子(新的标量未知数),拉格朗日乘法能够解决等式约束的优化问题(王政民,吴阔华,1999)。使用拉格朗日乘法求解概化设计的侧面最佳样本量,其推导过程如下:

第一步,定义不同概化设计的拉格朗日函数。例如,在 $s \times i$ 设计中,即为 $\min F(n_s, n_i, \lambda) = \sigma_\delta^2 - \lambda(cn_s n_i - B)$。

第二步,写出拉格朗日函数的偏导函数。例如,在 s × i 设计中,n_s、n_i 和 λ 为未知量,偏导函数为 $\frac{\partial F}{\partial n_s}$、$\frac{\partial F}{\partial n_i}$ 和 $\frac{\partial F}{\partial \lambda}$。

第三步,使偏导函数的值为 0。例如,在 s × i 设计中,使偏导函数的值为 0,那么可得 $\frac{\partial F}{\partial n_s} = -\frac{\sigma_s^2}{n_s^2} - \frac{\sigma_{si}^2}{n_s^2 n_i} - \lambda cn_i = 0$,$\frac{\partial F}{\partial n_i} = -\frac{\sigma_i^2}{n_i^2} - \frac{\sigma_{si}^2}{n_s n_i^2} - \lambda cn_s = 0$,$\frac{\partial F}{\partial \lambda} = -cn_s n_i + B = 0$。

第四步，求解最佳样本量。例如，在 s × i 设计中，$n_s = \sqrt{\frac{\sigma_s^2}{\sigma_i^2}\frac{B}{c}}$，$n_i = \sqrt{\frac{\sigma_i^2}{\sigma_s^2}\frac{B}{c}}$。

接下来，仍以 s × i 设计为例，来说明如何具体应用拉格朗日乘法求解最佳样本量。

Marcoulides(1993)提出了在预算限制下计算各个侧面最佳样本量的方法。对于 $s \times i$ 设计，如果 X_{si} 表示被试完成题目所得的观察分数，那么被试 s 的观察分数的期望值为 $\mu_s \equiv \underset{i}{E} X_{si}$。同样地，题目 i 的观察分数的期望值 $\mu_i \equiv \underset{s}{E} X_{si}$，所有被试完成所有题目的观察分数的期望值为 $\mu \equiv \underset{s}{E}\underset{i}{E} X_{si}$。被试完成题目所得的观察分数可表示为：

$$X_{si} = \mu + (\mu_s - \mu) + (\mu_i - \mu) + (X_{si} - \mu_s - \mu_i + \mu) \quad (3)$$

在公式(3)中，μ 表示总平均分；$(\mu_s - \mu)$ 表示被试的效应；$(\mu_i - \mu)$ 表示题目的效应；$(X_{si} - \mu_s - \mu_i + \mu)$ 表示被试与题目交互(包括残差)的效应。

从公式(3)可知，观察分数由总平均分、被试的效应、题目的效应和被试与题目交互(包括残差)的效应四部分组成。基于公式(3)，可以推导出观察分数的方差分量等于被试的方差分量、题目的方差分量、被试与题目交互(包括残差)的方差分量之和，即 $\sigma^2(X_{si}) = \underset{s}{E}\underset{i}{E}(X_{si} - \mu)^2 = \sigma_s^2 + \sigma_i^2 + \sigma_{si,e}^2$。其中，被试的方差分量为 $\sigma_s^2 = \underset{s}{E}(\mu_s - \mu)^2$，题目的方差分量为 $\sigma_i^2 = \underset{i}{E}(\mu_i - \mu)^2$。

在 $s \times i$ 设计中，n_s 个被试完成 n_i 个题目的测量误差均值可用相对误差方差来表示，其公式如下。

$$\sigma_\delta^2 = \frac{\sigma_{si,e}^2}{n_s n_i} \quad (4)$$

在公式(4)中，σ_δ^2 表示相对误差方差；σ_s^2、σ_i^2、$\sigma_{si,e}^2$ 分别表示被试的方差分量、题目的方差分量、被试与题目交互(包括残差)的方差分量；n_s、n_i 分别表示被试的样本量、题目的样本量。

在预算限制下，$s \times i$ 设计只有一个确切的最佳的样本量可选择。此时，相对误差方差的最小值其实是决策变量——题目数 n_i 和被试数 n_s 的一个非线性最优化问题，且受预算限制 $cn_i n_s \leqslant B$ 约束。可以使用拉格朗日函数 $F(n_i, n_s, \lambda) = \sigma_\delta^2 - \lambda(cn_i n_s - B)$，来对这个非线性最优化问题进行运算，得

$$n_s = \sqrt{\frac{\sigma_s^2}{\sigma_i^2}\frac{B}{c}} \quad (5)$$

$$n_i = \sqrt{\frac{\sigma_i^2}{\sigma_s^2}\frac{B}{c}} \quad (6)$$

在公式(5)中，n_s 表示最佳被试数量；σ_s^2 表示估计的被试的方差分量；σ_i^2 表示估计的题目的方差分量；c 表示单题成本；B 表示完成一次评价的预算。在公式(6)中，n_i 表示最佳题目数量，其它表示符号的意义同公式(5)。

4　问题与展望

4.1　概化理论预算限制下最佳样本量估计存在的问题

第一，受总预算舍入的影响。使用拉格朗日乘法估算所得最佳样本量结果多为非整数。在实际研究过程中，一般要求观察数量结果为整数。通常，通过将结果四舍五入至与其最接近的整数的方法来获得结果的整数解。但是，数值的改变可能会使最终的测验成本改变，舍入得到的结果也可能会导致最终的测验成本超过或低于总预算。例如，在某概化理论研究设计中，舍入取整后，完成一次测验所需成本为 2640 元，比使用未舍入前的成本 2500 元多花了 140 元。虽然这在每一次单独的测验中显得微不足道，但是在某些情况下，因舍入可能导致实际花费超出预算，且超出的费用不容小视。避免因舍入导致测量费用与预算不符的一个相对简便的操作方法是估算舍入后样本可能的所有排列组合，并从中选择最佳的观察数量。另外，Saunders，Theunissen 和 Baas(1989)提出了整数规划方法，也可用于获得最佳样本量。整数规划方法是基于整数得到样本大小，其优点是所得结果不需要进行舍入，其缺点是在计算过程中所使用的算法必须适合于每个决策研究设计，因此对算法要求较高。相比之下，综合考虑，仍然是拉格朗日乘法的适用性更强。对于结果舍入造成的预算差值问题，目前最常用的解决方案还是检查结果，估算舍入侧面样本量可能的所有排列组合，以确保费用与预算相符。

第二，受方差分量为负的影响。在某些概化理论研究设计中，所估计的方差分量可能为负值，这可能源于数据本身的问题。Brennan(2001)认为，进行概化研究时，方差分量可能会出现负值，当方差分量为负值时，可将此负值的方差分量看成 0 处理。实际上，当出现负值方差分量，且此负值方差分量接近于 0 时，将负值方差分量的数值处理为 0，所得结果与未将负值方差分量的数值处理为 0 的结果并无很大差异，在这种情况下，两种结果任选其一对整个研究结果影响不大。但是，当方差分量出现的负值较大且影响到结果的代入计算时，可考虑是数据本身的问题，这反映出在收集数据的过程中是否出现了较大的偏差。方差分量出现负值时也可能源于方差分量估计方法的选择。例如，使用 urGENOVA 软件进行方差分量的估计，其估算原理通常是方差分析(ANOVA)技术，用样本平均数来估计总体均值，这种方法容易受抽样的影响，存在一定的误差，结果出现负值是较为常见的。

第三，受固定侧面的影响。在概化理论研究中，如果测量侧面水平数是可变的，那么这个侧面为随机侧面。反之，如果测量侧面水平数是不可变的，仅采用一定的侧面水平数，那么这个侧面就成了固定侧面。如果一个侧面被固定，那么该侧面将被视为测量目标的一部分，已固定的侧面不再属于随机误差来源。当固定侧面增多时，测量误差来源变少，测量的信度提高，测量的可靠性变大。但是，需要注意的是，通过固定侧面来减少测量误差是以缩小测量目标的范围为代价的，这是因为当测量目标范围不断变小，测量结果的推广意义也随之变小。例如，在教师教学水平评价的概化设计中(Bergsmann, Schultes, Winter, Schober, & Spiel, 2015; Iqbal, Lee, Pearson, & Albon, 2016)，$(s:t)\times(i:v)$ 设计固定了侧面 v，$(s:t)\times(i:v)\times o$ 设计固定了侧面 v 和侧面 o，当固定侧面不再属于误差来源时，这两个设计的随机侧面是相同的。考虑到数据的相同性，比较 $(s:t)\times i$ 设计和 $(s:t)\times(i:v)$ 设计，这两个设计具有相同的随机侧面。然而，即使固定某一个侧面使两个设计的随机侧面相同，但由于设计不同，所得结果也仍然存在不同。

第四，受不平衡设计的影响。由于在概化理论研究中，出现缺失数据会对概化分析中的方差分量估计造成影响。因此，如果在概化理论研究中出现缺失数据，那么就需要对缺失数据进行调整。概化理论一般处理平衡设计，但如果概化设计的测量侧面的观察数量存在不等，那么该概化设计可以认为是不平衡设计。如果需要对不平衡设计进行决策研究，那么就需要对设计中的不等的观察数量进行缺失值处理以使观察数量保持一致，变不平衡设计为平衡设计。例如，在教师教学水平评价的概化设计中，$(s:t)\times i$ 设计表示学生 s 嵌套于教师 t 中，学生的观察数量可以为 22、25、25、31、60、19、25、29、35、17、22、64、27、26、20、21、21、22、19。很显然，评价每位教师的学生数量是不一样的，属于不平衡设计。如果需要将不平衡设计变为平衡设计时，那么就需要将 s 的观察数量同时变为这一系列数值中的最大值 64，其它小于 64 的观察数量，当作缺失值来处理，比如第一个观察数量为 22，即缺失 42 个数值。但是，如果采用此种处理方法，那么将会使得用于分析的整个数据的总量发生巨大变化，有时得不偿失。

第五，受最优概化设计的影响。例如，使用拉格朗日乘法来探讨教师教学水平评价，影响因素是多方面的(Casabianca, Lockwood, & McCaffrey, 2015; Oghazi, 2015; Pleschová & Mcalpine, 2016)，其概化设计可以为 $t\times i$、$(s:t)\times i$、$(s:t)\times(i:v)$ 和 $(s:t)\times(i:v)\times o$ 设计。其中，t 表示测量目标(object of measurement)，代表被评教师；i、s、v 和 o 表示测量侧面(facet of measurement)，i 表示评教项目，s 表示评教学生数量，v 表示评教项目的维度，o 表示评教的场合(评教的次数)。那么，哪个设计是最优的呢？其实，可以将拉格朗日乘法应用于教师教学水平评价之中，通过比较四个概化设计的结果来获得最优概化设计。诚然，在知晓最优概化设计的基础上，概化理论预算限制下最佳样本量可以通过最优概化设计来进行估计。当然，不同的最优概化设计所得到的最佳样本量估计是不同的。因此，可以认为，最优概化设计的选择对最佳样本量估计是有较大影响的。

4.2 概化理论预算限制下最佳样本量估计展望

Marcoulide 和 Goldstein(1990)提出的拉格朗日乘法为进一步解决预算限制下多侧面设计中最佳样本量估计问题提供了崭新视角。随后，一些学者将单侧面拉格朗日乘法推广至多侧面情境中(Marcoulides, 1993, 1997; Meyer, Liu, & Mashburn, 2014)。然而，这些学者的研究是依据不同的概化设计对拉格朗日乘法进行“分而治之”，研究的关注点多在简单的交叉设计，没有将研究焦点拓广至更为复杂的混合设计之中(混合设计既包含交叉设计又包含嵌套设计，大量的概化设计属于多侧面混合设计)。这导致当前的一些研究在讨论预算限制下最佳样本量估计时，表现出一些不足。为此，概化理论预算限制下最佳样本量估计可以从以下两个方面进行改善。

第一，推导出拉格朗日乘法的统一公式。概化理论预算限制下最佳样本量估计受总预算舍入、方差分量为负、固定侧面、不平衡设计、最优概化设计等方面因素的影响，目前大多数研究没有更进一步地对拉格朗日乘法的基本原理进行系统地归纳与总结，仍然依据的是对不同的概化设计进行“分而治之”，也仍然没有提出一个相对统一的拉格朗日乘法公式来克服这些因素所带来的各种影响(Marcoulides, 1991, 1993, 1994, 1995, 1997; Meyer, Liu, & Mashburn, 2014)。为此，有必要推导出拉格朗日乘法的统一公式，这是后续研究可以弥补之处。

第二，将拉格朗日乘法应用于更为复杂的概化理论研究设计中。Meyer, Liu 和 Mashburn(2014)在前人研究的基础上，将 Marcoulides 和 Goldstein(1990, 1991, 1992)先前关于预算限制下估算概化理论不同设计最佳样本量的拉格朗日乘法应用于实际测量研究中，这具有重要意义，表明拉格朗日乘法具有广泛的适用性。但是，Meyer, Liu 和 Mashburn(2014)的研究的概化设计较之前同类研究的概化设计并没有显得更为复杂，也没有将拉格朗日乘法应用至多元概化理论的复杂设计之中。特别地，对于更为复杂的多元概化混合设计(Gitomer, Bell, Qi, McCaffrey, Hamre, & Pianta, 2014; Spooren, Mortel-

mans,& Christiaens,2014;Chen,Hsieh,& Do,2015),讨论甚少,这是未来可以继续进一步作深入探讨的热点问题之一。

参考文献

戴海崎,张锋,陈雪枫.(2011).*心理与教育测量(第三版)*.广州:暨南大学出版社.

黎光明,张敏强,张文怡.(2013).人事测评中的概化理论应用.*心理科学进展*,*21*(1),166-174.

王政民,吴阔华.(1999).*高等数学(第三版)*.南昌:江西高校出版社.

Bergsmann, E., Schultes, M. T., Winter, P., Schober, B., & Spiel, C. (2015). Evaluation of competence-based teaching in higher education: From theory to practice. *Evaluation & Program Planning*, *52*(10), 1-9.

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Chen, J. F., Hsieh, H. N., & Do, Q. H. (2015). Evaluating teaching performance based on fuzzy AHP and comprehensive evaluation approach. *Applied Soft Computing*, *28*(C), 100-108.

Cleary, T. A., & Linn, R. L. (1969). Error of measurement and the power of a statistical test. *The British Journal of Mathematical and Statistical Psychology*, *22*, 49-55.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311-337.

Gage, N. A., Prykanowski, D., & Hirn, R. (2014). Increasing reliability of direct observation measurement approaches. *Behavioral Disorders*, *39*(4), 228-224.

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1-32.

Goldstein, Z., & Marcoulides, G. A. (1991a). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement*, *51*(1), 79-88.

Goldstein, Z., & Marcoulides, G. A. (1991b). Selecting the number of observations in multivariate measurement studies under budget constraints. *Educational and Psychological Measurement*, *51*(3), 574-584.

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58.

Iqbal, I., Lee, J. D., Pearson, M. L., & Albon, S. P. (2016). Student and faculty perceptions of student evaluations of teaching in a Canadian pharmacy school. *Currents in Pharmacy Teaching & Learning*, *8*(2), 191-199.

Lakes, K. D. (2013). Restricted sample variance reduces generalizability. *Psychological Assessment*, *25*(2), 643-650.

Marcoulides, G. A. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement*, *51*(1), 79-88.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics*, *18*(2), 197-206.

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement*, *54*(1), 3-7.

Marcoulides, G. A. (1995). Designing Measurement studies under budget constraints controlling error of measurement and power. *Educational and Psychological Measurement*, *55*(3), 423-428.

Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement*, *57*(5), 808-812.

Marcoulides, G. A., & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement*, *50*(4), 761-768.

Marcoulides, G. A., & Goldstein, Z. (1991). Selecting weighting schemes in multivariate generalizability studies under budget constraints. *Educational and Psychological Measurement*, *51*(3), 573-584.

Marcoulides, G. A., & Goldstein, Z. (1992). The optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement*, *52*(2), 301-308.

Maulana, R., Helms-Lorenz, M., & Grift, W. V. D. (2015). A longitudinal study of induction on the acceleration of growth in teaching quality of beginning teachers through the eyes of their students. *American Biology Teacher*, *51*(1), 225-245.

Meyer, P. J., Liu, X., & Mashburn, A. J. (2014). A practical solution to optimizing the reliability of teaching observation measures under budget constraints. *Educational and Psychological Measurement*, *74*(2), 280-291.

Oghazi, P. (2015). Beautiful teaching and good performance. *Journal of Business Research*, *69*(5), 1887-1891.

Pleschová, G., & Mcalpine, L. (2016). Helping teachers to focus on learning and reflect on their teaching: What role does teaching context play? *Studies in Educational Evaluation*, *48*(3), 1-9.

Saunders, P. F., Theunissen, T. J., & Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika*, *54*(4), 587-589.

Spooren, P., Mortelmans, D., & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation*, *43*(11), 88-94.

Wolbing, T., & Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social Science Research*, *57*(5), 253-272.

Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multifacet decision studies. *Psychometrika*, *38*, 173 - 181.

Estimating the Best Sample Size under Budget Constraints for Generalizability Theory

Li Guangming
(School of Psychology, Center for Studies of Psychological Application, South China Normal University, Guangzhou 510631)

Abstract: Generalization theory is widely used in various psychological evaluation practices. Budget and cost are the problems that cannot be neglected in the research of measurement. When there is a budget constraint, the generalization theory needs to consider how to design a measurement program with relatively high reliability and feasibility, which requires the optimal sample size to be estimated by some means. Lagrange multiplication is a more mature method for estimating the optimal sample size under the budget constraints in generalization theory. Some influencing factors of optimal sample size estimation under the budget constraints in generalization theory, such as the influence of total budget rounding, are discussed, and some follow - up solutions are also proposed, such as deducing the unified formula of Lagrange multiplication.

Key words: generalizability theory; budget constrain; estimating the best sample size; LaGrange multiplier method; psychological evaluation

(上接第 233 页)

Natalie, A., & Phillips, S. K. (2012). Ageing and bilingualism: Absence of a "bilingual advantage" in Stroop interference in anonimmigrant sample. *The Quarterly Journalof Experimental Psychology*, *65*(2), 356 - 369.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643 - 662.

The Influence of Language Proficiency on Stroop Effect
——Task State fMRI Study Based on Unbalanced Bilinguals

Qian Qiaoyun Yi Yinqiao Dan Yibo Liu Shen Ning Ruipeng
(Shanghai Key Laboratory of Magnetic Resonance, School of Physics and Materials Science, East China Normal University, Shanghai 200062)

Abstract: Objective: Exploring the influence of language proficiency on Stroop effect and the brain mechanism behind it. Materials and Methods: 17 native Chinese subjects, whose second language was English, participated in this study. The subjects completed the Stroop color words task in Chinese and English respectively when fMRI was scanning. Results: Under the consistent condition of color words, the dorsal occipital lobe and the right dorsolateral dorsal prefrontal lobe were more strongly activated in Chinese task compared with English task. Compared to the reverse, the ventral occipital lobe was more strongly activated. Under the inconsistency condition of color words, the bilateral dorsolateral frontal lobe and right occipital inferior gyrus were more strongly activated in Chinese task compared with English task. Compared to the reverse, there was no significant activation of the brain region. In addition, compared with the English Stroop effect, the Chinese Stroop effect caused stronger activation of the left inferior frontal gyrus. Conclusion: Combined with previous studies, we concluded that for the dorsolateral prefrontal cortex which was important for attention control and the left inferior frontal gyrus which was associated with response inhibition, language proficiency affected the degree to which they involve in Stroop tasks, and further affected the performance of Stroop tasks.

Key words: stroop; language proficiency; fMRI; attention control; response inhibition