

# 学生评教研究的最佳样本量估算 ——基于拉格朗日乘数法的应用

刘颖 张敏强<sup>\*\*</sup> 甄锋泉

(华南师范大学心理学院, 广州, 510631)

**摘要** 样本量是影响测量有效性的因素之一, 受多个条件限制。文章阐述了拉格朗日乘数法用于推导预算限制下概化研究侧面最佳水平数的流程化操作思路; 通过学生评教的实证研究, 比较预算限制下概化研究不同设计各测量侧面的最优水平数, 说明拉格朗日乘数法的广泛适用性。结果表明: (1) 拉格朗日乘数法在预算限制下求解概化研究侧面最佳水平数时表现出稳健性; (2) 结合测量研究设计需要及实际情况, 可得概化研究中的最优设计。

**关键词** 预算限制 概化理论 学生评教 拉格朗日乘数法

## 1 引言

### 1.1 学生评教的现状

随着高等教育大众化, 教学质量折射着学校的教学水平, 成为社会关注的热点话题, 是各高校无法回避的问题, 而学生评教是教师教学质量评价的主要途径之一。学生评教 (students' evaluations of teaching, SET) 是指学生对教师的课堂教学质量和教学效果进行评价的活动 (赵平, 2018)。教育学家 Williams 和 Blackstone 强调应定期进行学生对教师教学状况的评价, 著名教育学者 Marsh 和 Overall 也将学生评教的重要性摆在突出位置 (谢博文, 史蒂, 2012)。可见, 学生评教的重要性已在学术界得到充分的认可。组织学生对教师教学状况进行评价可获得以下效果: ①促进教学质量的提升; ②关注学生的课程体验; ③为高校管理者提供决策依据。

作为目前高校最普遍的教师评价手段, 学生评教的有效性问题在学术界一直存在争议。“肯定派”认为学生评教虽在实践中存在质疑, 但由于学生作为教学主体的优势, 其实效性是可信的 (江利, 2017)。然而, “质疑派”认为目前多数学生对“评教”持无所谓的态度以及评教工具的科学性等问题影响了评教的有效性及可靠性。从测量学的角度看,

评价工具的质量、抽测群体的样本量、施测组织过程等都可能影响研究结果的有效性。

有学者认为, 学校在大规模实施教师教学水平评估之前应该对评估的可靠性作调查分析, 以保证评估的质量 (黎光明, 张敏强, 2009)。马秀麟、袁克定和刘立超 (2014) 的研究发现目前学生评教主要应用经典测量理论 (classical test theory, CTT) 进行描述性统计分析, 以均值、频数等统计量论证数据间的逻辑关系, 少有研究从测量学的角度对学生评教的有效性作进一步的探讨。然而, 应用经典测量理论进行数据分析只能从宏观上观测评教结果的有效性, 无法探查不同的测量情境、不同误差来源对研究结果的影响, 具有一定的局限性。

### 1.2 概化理论

概化理论 (generalizability theory, GT), 是测量行为可靠性的统计理论 (Shavelson & Webb, 1991), 广泛应用于心理与教育测量实践中 (杨志明, 2003)。概化理论引入因素实验设计和方差分析 (Analysis of Variance, ANOVA) 的思想, 通过概化研究 (generalizability study, G 研究), 将总变异按照来源分解到各个侧面, 再根据概化研究所得的各个侧面及其交互作用的方差分量, 进行决策研究 (decision study, D 研究)。概化理论中, 测量情境

\* 通讯作者: 张敏强。E-mail: 2640726401@qq.com

DOI:10.16719/j.cnki.1671-6981.20200413

关系中的重要组成部分是每个测量侧面的水平数，即一个侧面内的个别情况（陈社育，余嘉元，2001），研究者可在 D 研究中调整各测量侧面的水平数以改进测量设计。概化理论以概化系数（generalizability coefficient,  $E_p^2$ ）和可靠性指数（dependability index,  $\phi$ ）作为指标衡量测验的效用，分别用于常模参照测验和标准参照测验中表示信度水平，两个指标的值越高，代表测验信度越高。Cronbach、Gleser、Nanda 和 Rajaratnam (1972) 指出在概化理论中，随着一个侧面水平数量的增加，概化系数一般会随之提高，直至这个潜在的增长最终达到设想的数值。

### 1.3 预算限制问题

在决策研究中各测量侧面水平数的确定受测量成本的限制，提高测验有效性与控制测验成本是一个两难问题。因此，预算与成本是研究者在进行测量研究设计时不可忽略的问题。进行学生评教时，问卷的编制、参与评教的学生数量等问题均可能造成成本的变动。由此看来，在预算限制下，明确概化研究中各个侧面水平数以保证研究的效用是一项重要工作。

20 世纪 70 年代起，国外学者关注如何在预算限制下寻找最优的测量设计，并提出了一系列解决方法。Woodward 和 Joe(1973) 将约束优化方法用于预算限制下的侧面最佳水平数的问题解决中，但该方法在三侧面及以上的设计中难以推广。随后，离散优化方法 (Sanders, Theunissen, & Bass, 1989)、柯西 - 希瓦兹不等式方法 (Sanders, 1992)，拉格朗日乘数法 (Goldstein & Marcoulides, 1991; Marcoulides & Goldstein, 1990, 1992; Marcoulides, 1993, 1994, 1995, 1997; Meyer, Liu, & Mashburn, 2014) 等方法也被用于解决预算限制下多侧面设计的侧面最佳水平数问题，其中，拉格朗日乘数法具有较好的灵活性。然而，关于拉格朗日乘数法的应用多在简单的交叉设计中，未在复杂的混合设计中推广。

本文基于前人的研究，阐述拉格朗日乘数法用于估算预算限制下概化研究侧面最佳水平数的流程化操作思路；通过一个学生评教的实证研究，比较不同概化设计在预算限制下不同测量侧面的最优水平数，说明拉格朗日乘数法的广泛适用性。

### 1.4 操作思路

拉格朗日乘数法是在数学最优问题中，一种寻找变量受一个或多个条件所限制的多元函数极值的方法；通过引入拉格朗日乘子（新的标量未知数），

解决等式约束的优化问题。拉格朗日函数可表示为：

$$L(x, y, \lambda) = f(x, y) - \lambda g(x, y) \quad (1)$$

公式 (1) 可以解释为引入新的标量未知数  $\lambda$ ，求解函数  $f(x, y)$  在函数  $g(x, y)$  的限制下的极值。以公式 (1) 为例，使用拉格朗日乘数法求解极值时，分别对未知数  $x, y, \lambda$  求导，得到偏导函数，并列出方程，如下所示：

$$\frac{\partial L}{\partial x} = 0. \quad (2)$$

$$\frac{\partial L}{\partial y} = 0. \quad (3)$$

$$\frac{\partial L}{\partial \lambda} = 0. \quad (4)$$

联合公式 (2)、(3)、(4) 可求得  $x, y, \lambda$  的值，代入目标函数  $f(x, y)$  可求得极值。

参考国外相关研究 (Goldstein & Marcoulides, 1991; Marcoulides & Goldstein, 1990, 1992; Marcoulides, 1993, 1994, 1995, 1997; Meyer et al., 2014)，使用拉格朗日乘数法求解预算限制下概化研究侧面最佳水平数时，包含以下 4 个步骤：

(1) 分析误差来源，进行概化设计。根据研究目的，确定测量目标；运用因素实验设计的思想，分析影响测验结果的误差来源，如被试水平、题目难度、施测次数等；根据侧面间的关系进行概化设计，包括交叉设计、嵌套设计和混合设计。

(2) 估计方差分量，求解平均误差方差和相对误差方差。对收集好的数据进行方差分析，估计各测量侧面的方差分量；同时，求解不同概化设计的平均误差方差和相对误差方差。估计方差分量时，常用的软件有：GENOVA、mGENOVA、urGENOVA 等。

(3) 根据实际研究的限制，量化限制条件。如在预算限制下，可预先估计单价，设定总预算，以确保总的测验成本不高于总预算。

(4) 引入拉格朗日乘子，求解预算限制下概化研究中测量侧面的最佳水平数。推导拉格朗日函数，获得预算限制下概化研究侧面最佳水平数方程，代入对应方差分量，求解侧面最佳水平数。

## 2 方法

### 2.1 数据收集

采用《高校教师教学水平评价量表(学生用)》（下称“评教问卷”）分别对广东省三所高校共 530 名学生进行施测，要求参测学生分别对其授课老师（共

19个不同科目)进行评价,每名学生均只评价一位授课老师,并在初测后3个月进行复测。

量表共有五个维度,分别为教学方法、教学内容、教学态度、教学组织、教学效果,每个维度含5个项目,共25个项目。量表五个维度的克隆巴赫系数依次为.86、.81、.87、.82、.88,量表的克隆巴赫系数为.95。

## 2.2 研究设计

在学生评教课题研究中,评教学生数量、评教问卷的编制(包括维度划分、项目数量)以及评教次数均可能成为影响测验有效性的因素。本研究应用拉格朗日乘数法,探讨不同概化设计在同一预算限制下不同测量侧面的最优水平数,采用层层递进的思路,根据目前高校学生评教的常见场景,以教师 $t$ 为测量目标,依次纳入不同侧面进行概化设计,分别考虑 $t \times i$ 设计、 $(s:t) \times i$ 设计、 $(s:t) \times (i:v)$ 设计,以及 $(s:t) \times (i:v) \times o$ 设计共四个设计,其中, $t$ 代表教师, $i$ 代表量表项目, $s$ 代表参评学生, $v$ 代表量表维度, $o$ 代表评教次数。

## 2.3 分析工具

urGENOVA是一个用来估算含有侧面交互作用的概化设计的随机效应方差分量的ANSI C计算机程序。本文使用urGENOVA软件,估计不同概化设计中各个侧面及各侧面交互作用的方差分量。

## 2.4 预算限制

在测量研究中,预算与成本均是可变的,研究者可以根据实际研究设计,设定预算,估算成本。本文用 $B$ 代表完成一次评教的预算,用 $c$ 代表完成评教问卷中一个项目评估所需的成本(即单位成本)。根据实际研究情况,本文设定完成一次评教的预算 $B=2500$ 元。综合考虑问卷的设计、打印、施测、数据分析等费用支出,本研究设定完成一份评教问卷(以问卷为单位)评估的成本为10元,采用的问卷共25个项目,即单位成本(以项目为单位)为 $c=.4$ 元。

以 $(s:t) \times (i:v) \times o$ 设计为例,预算限制可表示为 $cn_sn_in_o \leq B$ ,其中, $n_s$ 代表评教学生的数量(即收取的评教问卷的总份数), $n_i$ 代表评教项目的数量, $n_v$ 代表评教问卷包含的维度, $n_o$ 代表评教次数(下同)。

## 2.5 比较指标

以概化系数作为比较指标。概化系数 $E^2$ 适用

于常模参照性测验和相对决策,被定义为全域分数方差 $\sigma^2(\tau)$ 与其和观察分数方差 $\sigma^2(\delta)$ 两者之和的比率,可表示为:

$$E^2_\rho = \sigma^2(\tau)/[\sigma^2(\tau)+\sigma^2(\delta)]. \quad (5)$$

实际计算中,概化系数即为测量目标方差与其和相对误差方差之和的比率,在本研究中,可表示为:

$$E^2_\rho = \sigma_t^2/(\sigma_t^2 + \sigma_\delta^2). \quad (6)$$

## 3 结果与分析

### 3.1 $t \times i$ 设计的侧面最佳水平数估计

测验由不同的项目构成,用来测量特定的心理特质。当测验项目较少时,测验的内容代表性可能不足,测验的有效性降低;但当测验项目较多时测验成本也将随之增加。在预算限制下合理地确定施测问卷的项目数量是研究者进行测量设计时不可忽视的问题。 $t \times i$ 设计是教师和评教项目的随机交叉设计,为评教中最简单且较常见的设计,视测量目标教师 $t$ 为整体,由 $n_s$ 位学生使用同一份含 $n_i$ 个评教项目的评教问卷评价教师 $t$ 。

$t \times i$ 设计的随机效应方差分量如表1所示:

$t \times i$ 设计的平均误差方差和相对误差方差分别为:

$$\sigma_X^2 = \frac{\sigma_t^2}{n_t} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{ti}^2}{n_t n_i} \quad (7)$$

$$\sigma_\delta^2 = \frac{\sigma_{ti}^2}{n_t n_i}. \quad (8)$$

在 $t \times i$ 设计中,数据的获得方式是由学生 $s$ 使用评教问卷对教师 $t$ 进行评价, $n_i$ 的实际意义是收取的问卷数量,也就是进行评教的学生数量 $n_s$ ,仅用相对误差方差代入拉格朗日函数无法推导求得最佳评教学生数和最佳评教项目数。参考Marcoulides(1993)的研究,在单侧面交叉设计中,用平均误差方差 $\sigma_X^2$ 代替相对误差方差 $\sigma_\delta^2$ 代入拉格朗日函数计算,即 $F(n_t, n_i, \lambda) = \sigma_X^2 - \lambda(cn_t n_i - B)$ 。在 $t \times i$ 设计中, $n_i = n_s$ , $t \times i$ 设计的最佳评教学生数 $n_s$ 和最佳评教项目数 $n_i$ 分别为:

$$n_s = n_t = \sqrt{\frac{\sigma_t^2 B}{\sigma_i^2 c}} \quad (9)$$

表1  $t \times i$ 设计方差分量

	$\sigma_i^2$	$\sigma_i^2$	$\sigma_n^2$
单侧面设计方差分量	.28548	.09267	.74400

$$n_i = \sqrt{\frac{\sigma_t^2 B}{\sigma_{ti}^2 c}}. \quad (10)$$

由公式(9)、(10)及表1计算可得,当B=2500, c=.4时,  $n_t=138.75808$ ,  $n_i=45.04242$ 。将 $n_t$ 和 $n_i$ 的计算结果四舍五入,取值为139和45,此时由公式(7)可得,平均误差方差 $\sigma_{\bar{X}}^2=.00423$ ,由公式(8)可得,相对误差方差 $\sigma_{\delta}^2=.00012$ ,由公式(6)可得,最优概化系数 $E_p^2=.99958$ 。

因此,在 $t \times i$ 设计中,当预算限制为2500元,评教学生数量为139人,评教项目为45题时,测量可靠性最大。

### 3.2 (s:t) $\times$ i设计的侧面最佳水平数估计

在概化理论中,将不同被试评价不同个体的测验关系定义为嵌套关系,可进行嵌套设计。(s:t)  $\times$  i设计考虑评教学生s嵌套于被评教师t中,与评教项目i交叉设计,即由 $n_s$ 位学生使用同一份含 $n_i$ 个评教项目的评教问卷评价不同的教师t。

(s:t)  $\times$  i设计的随机效应方差分量如表2所示。

(s:t)  $\times$  i设计的平均误差方差及相对误差方差分别为:

$$\sigma_{\bar{X}}^2 = \frac{\sigma_t^2}{n_t} + \frac{\sigma_i^2}{n_i} + \frac{\sigma_{s:t}^2}{n_s} + \frac{\sigma_{ti}^2}{n_i} + \frac{\sigma_{s:t:i}^2}{n_s n_i} \quad (11)$$

$$\sigma_{\delta}^2 = \frac{\sigma_{s:t}^2}{n_s} + \frac{\sigma_{ti}^2}{n_i} + \frac{\sigma_{s:t:i}^2}{n_s n_i}. \quad (12)$$

(s:t)  $\times$  i设计的拉格朗日函数为 $F(n_s, n_i, \lambda) = \sigma_{\delta}^2 - \lambda(c n_s n_i - B)$ ,求导可得最佳评教学生数 $n_s$ 及最佳评教项目数 $n_i$ 分别为:

$$n_s = \sqrt{\frac{\sigma_{s:t}^2 B}{\sigma_{ti}^2 c}} \quad (13)$$

$$n_i = \sqrt{\frac{\sigma_{ti}^2 B}{\sigma_{s:t}^2 c}}. \quad (14)$$

由公式(13)、(14)和表2计算可得,当B=2500, c=.4时,  $n_s=262.97048$ ,  $n_i=23.76693$ 。将结果四舍五入,则该设计中的最佳评教学生数和最佳评教项目数分别为263和24。此时,由公式(12)可得,相对误差方差为 $\sigma_{\delta}^2=.00212$ ,由公式(6)可得,

表2 (s:t)  $\times$  i设计方差分量

	$\sigma_t^2$	$\sigma_{s:t}^2$	$\sigma_i^2$	$\sigma_{ti}^2$	$\sigma_{s:t:i}^2$
双侧面设计方差分量	.11059	.26909	.02702	.02432	.44349

概化系数 $E_p^2=.98121$ 。

因此,在(s:t)  $\times$  i设计中,当预算限制为2500元,抽取263位评教学生,即每位教师大约由14位学生进行评价(注:本研究共19位被评教师),评教项目为24个时,测量可靠性最大。

### 3.3 (s:t) $\times$ (i:v)设计的侧面最佳水平数估计

进行测验设计时,不同项目分属不同的维度,不同的维度下需包含一定数量的项目。(s:t)  $\times$  (i:v)设计不仅考虑评教学生s嵌套于被评教师t中,同时考虑评教问卷的不同维度,评教项目i嵌套于维度v,两者交叉设计,即由 $n_s$ 位学生使用同一份含 $n_v$ 个维度、每个维度含 $n_i$ 个评教项目的评教问卷评价教师t。此时,最佳评教项目数 $n_i$ 为不同维度下应包含的项目数量。

(s:t)  $\times$  (i:v)设计的随机效应方差分量如表3所示。

(s:t)  $\times$  (i:v)设计的平均误差方差和相对误差方差分别为:

$$\sigma_{\bar{X}}^2 = \frac{\sigma_t^2}{n_t} + \frac{\sigma_v^2}{n_v} + \frac{\sigma_{t:v}^2}{n_t n_v} + \frac{\sigma_{s:t}^2}{n_s} + \frac{\sigma_{sv:t}^2}{n_s n_v} + \frac{\sigma_{tv}^2}{n_v} + \frac{\sigma_{tit:v}^2}{n_t n_v} + \frac{\sigma_{sits:v}^2}{n_s n_t n_v} \quad (15)$$

$$\sigma_{\delta}^2 = \frac{\sigma_{s:t}^2}{n_s} + \frac{\sigma_{sv:t}^2}{n_s n_v} + \frac{\sigma_{tv}^2}{n_v} + \frac{\sigma_{tit:v}^2}{n_t n_v} + \frac{\sigma_{sits:v}^2}{n_s n_t n_v}. \quad (16)$$

(s:t)  $\times$  (i:v)设计的拉格朗日函数为 $F(n_s, n_i, n_v, \lambda) = \sigma_{\delta}^2 - \lambda(c n_s n_i n_v - B)$ ,本研究采用的量表共5个维度,即在本研究中,量表的侧面水平数固定为5,求导可得(s:t)  $\times$  (i:v)设计的最佳评教学生数 $n_s$ 和最佳评教项目数 $n_i$ 分别为:

$$n_s = \sqrt{\frac{(n_v \sigma_{s:t}^2 + \sigma_{sv:t}^2) B}{n_v \sigma_{tit:v}^2 c}} \quad (17)$$

$$n_i = \sqrt{\frac{\sigma_{tit:v}^2 B}{n_v (n_v \sigma_{s:t}^2 + \sigma_{sv:t}^2) c}}. \quad (18)$$

由公式(17)、(18)和表3计算可得,当B=2500, c=.4,  $n_v=5$ 时,  $n_s=329.83942$ ,  $n_i=3.78972$ ,将结果四舍五入得 $n_s=330$ ,  $n_i=4$ 。此时,由公式(16)可得,相对误差方差为 $\sigma_{\delta}^2=.00380$ ,由公式(6)可得,概化系数 $E_p^2=.96630$ 。

表3 (s:t)  $\times$  (i:v)设计方差分量

	$\sigma_t^2$	$\sigma_{s:t}^2$	$\sigma_v^2$	$\sigma_{i:v}^2$	$\sigma_{tv}^2$	$\sigma_{tit:v}^2$	$\sigma_{sv:t}^2$	$\sigma_{sits:v}^2$
三侧面设计方差分量	.10884	.25595	.02005	.01032	.01045	.01561	.07887	.37777

表4  $(s:t) \times (i:v) \times o$  设计方差分量

	$\sigma_o^2$	$\sigma_t^2$	$\sigma_{s:t}^2$	$\sigma_v^2$	$\sigma_{t:v}^2$	$\sigma_{ot}^2$	$\sigma_{os:t}^2$	$\sigma_{ov}^2$	$\sigma_{oi:v}^2$
四侧面设计方差分量	.00147	.08300	.00288	.02021	.00850	.02704	.28058	-.00018	.00160
	$\sigma_v^2$	$\sigma_{ti:v}^2$	$\sigma_{sv:t}^2$	$\sigma_{st:v}^2$	$\sigma_{otv}^2$	$\sigma_{ot:v}^2$	$\sigma_{osv:t}^2$	$\sigma_{osi:v}^2$	
四侧面设计方差分量	.00692	.01110	-.00224	-.00366	.00059	.00602	.09078	.40095	

因此，在 $(s:t) \times (i:v)$ 设计中，当预算限制为2500元，评教问卷共分5个维度时，评教学生抽取数量为330人，即每位教师大约由17位学生进行评价（注：本研究共19位被评教师）；评教问卷每个维度含4个评教项目，共20个项目（注：本研究采用的量表共5个维度），测量可靠性最大。

### 3.4 $(s:t) \times (i:v) \times o$ 设计的侧面最佳水平数估计

在经典测量理论（CTT）中，重测信度反映测验跨越时间的稳定性和一致性，是测验结果一致性的重要指标之一。在概化理论中，测验的次数经常被作为测量侧面在统计分析中考虑。 $(s:t) \times (i:v) \times o$ 设计在 $(s:t) \times (i:v)$ 设计的基础上考虑评教次数 $o$ ，评教学生 $s$ 嵌套于被评教师 $t$ ，评教项目 $i$ 嵌套于评教问卷维度 $v$ ，两个嵌套与评教次数 $o$ 交叉设计，即由 $n_s$ 位学生使用同一份分 $n_v$ 个维度、每个维度含 $n_i$ 个评教项目的评教问卷先后 $n_o$ 次评价教师 $t$ 。

$(s:t) \times (i:v) \times o$ 设计的随机效应方差分量如表4所示。

$(s:t) \times (i:v) \times o$ 设计的平均误差方差和相对误差方差分别为：

$$\begin{aligned}\sigma_x^2 &= \frac{\sigma_o^2}{n_o} + \frac{\sigma_t^2}{n_t} + \frac{\sigma_v^2}{n_v} + \frac{\sigma_{s:t}^2}{n_i n_v} + \frac{\sigma_{ov}^2}{n_v n_o} + \frac{\sigma_{ot:v}^2}{n_i n_v n_o} + \frac{\sigma_{st:v}^2}{n_s} \\ &+ \frac{\sigma_{ot}^2}{n_o} + \frac{\sigma_{os:t}^2}{n_s n_o} + \frac{\sigma_{tv}^2}{n_v} + \frac{\sigma_{ti:v}^2}{n_i n_v} + \frac{\sigma_{sv:t}^2}{n_s n_v} + \frac{\sigma_{st:v}^2}{n_s n_i n_v} \\ &+ \frac{\sigma_{otv}^2}{n_v n_o} + \frac{\sigma_{oti:v}^2}{n_i n_v n_o} + \frac{\sigma_{osv:t}^2}{n_s n_v n_o} + \frac{\sigma_{osi:v}^2}{n_s n_i n_v} \quad (19)\end{aligned}$$

$$\begin{aligned}\sigma_\delta^2 &= \frac{\sigma_{s:t}^2}{n_s} + \frac{\sigma_{ot}^2}{n_o} + \frac{\sigma_{os:t}^2}{n_s n_o} + \frac{\sigma_{tv}^2}{n_v} + \frac{\sigma_{ti:v}^2}{n_i n_v} + \frac{\sigma_{sv:t}^2}{n_s n_v} + \frac{\sigma_{st:v}^2}{n_s n_i n_v} \\ &+ \frac{\sigma_{otv}^2}{n_v n_o} + \frac{\sigma_{oti:v}^2}{n_i n_v n_o} + \frac{\sigma_{osv:t}^2}{n_s n_v n_o} + \frac{\sigma_{osi:v}^2}{n_s n_i n_v} \quad (20)\end{aligned}$$

$(s:t) \times (i:v) \times o$ 设计的拉格朗日函数为 $F(n_s, n_i, n_v, n_o, \lambda) = \sigma_\delta^2 - \lambda(cn_s n_i n_v n_o - B)$ ，同样地，固定量表维度的侧面水平数为5，同时，本研究只探讨进行一次重测的情况，即评教次数的侧面水平数固定为2，求导可得 $(s:t) \times (i:v) \times o$ 设计的最佳评教学生数 $n_s$ 和最佳评教项目数 $n_i$ 分别为：

$$n_s = \sqrt{\frac{(n_v n_o \sigma_{s:t}^2 + n_v \sigma_{os:t}^2 + n_o \sigma_{sv:t}^2 + \sigma_{osv:t}^2) B}{n_v n_o (\sigma_{ti:v}^2 + \sigma_{oti:v}^2)}} \quad (21)$$

$$n_i = \sqrt{\frac{n_o \sigma_{ti:v}^2 + \sigma_{oti:v}^2}{n_v n_o (n_v n_o \sigma_{s:t}^2 + n_v \sigma_{os:t}^2 + n_o \sigma_{sv:t}^2 + \sigma_{osv:t}^2) c}} B. \quad (22)$$

由公式(21)、(22)和表4计算可得，当 $B=2500$ ， $c=.4$ ， $n_v=5$ ， $n_o=2$ 时， $n_s=183.35696$ ， $n_i=3.40865$ ，四舍五入可得 $n_s=183$ ， $n_i=3$ 。此时，由公式(20)可得，相对误差方差 $\sigma_\delta^2=0.01662$ ，由公式(6)可得，概化系数 $E_p^2=.83313$ 。

因此，在 $(s:t) \times (i:v) \times o$ 设计中，当单次评教的预算为2500元，进行两次评教的预算限制为5000元时，评教问卷分5个维度，且只进行一次重测时，评教学生数量为183人，即每位教师大约由10位学生进行评价（注：本研究共19位被评教师）；评教问卷每个维度含3个评教项目，共15个项目（注：本研究采用的量表共5个维度），测量可靠性最大。

## 4 讨论

### 4.1 总预算舍入的影响

在预算限制下应用拉格朗日乘数法求得的概化研究侧面最佳水平数多为非整数，本研究通过四舍五入取得结果的整数解。然而，结果的舍入可能会导致实际测验成本低于或超出总预算。在本文的 $(s:t) \times (i:v)$ 设计中，舍入取整后，完成一次测验所需成本为2640元，比实际设定的预算高出140元（超预算5.6%），此超额尚可接受。如果实际的测验成本低于总预算，那么该问题可忽略。然而，在某些预算较大的研究中，舍入取整可能导致实际成本远远超出预算，此时研究者需进一步斟酌结果(Meyer et al., 2014)，可以列出舍入后所有可能的侧面水平数，进行排列，选择最佳的侧面水平数组合。

此外，本研究的预算条件基于数据的实际收集成本，并未考虑其他预算情况下的结果，无法对不同预算条件下的结果进行比较，后续研究可将预算作为影响因素进行考虑，以扩大研究结果的可推广性。

### 4.2 方差分量数值为负

本研究采用urGENOVA软件估计方差分量，其是用样本平均数来估计总体均值，容易受抽样的影

响,产生方差分量的负估计。在本文的 $(s:t) \times (i:v) \times o$ 设计中出现3个方差分量的负估计,但在本文的实际计算中没有对负值的方差分量作进一步处理。Brennan(2001)提到,如果出现方差分量负估计,可将其直接处理为零。在 $(s:t) \times (i:v) \times o$ 设计中,如果将负值的方差分量处理为零,得到舍入后的侧面最佳水平数为 $n_s=184$ , $n_i=3$ ,此时,概化系数 $E_\rho^2=.83262$ ,该结果与直接使用负值方差分量代入计算差异不大。因此,当负值方差分量极接近于零时,可不处理负值方差分量或将负值方差分量处理为零。当方差分量的负估计对研究结果影响较大时,研究者须考虑数据收集的有效性问题,或考虑使用不同的估计方法,如:马尔可夫链蒙特卡洛(MCMC)方法、Traditional方法、Bootstrap方法、Jackknife方法等(黎光明,张敏强,黄宪,王旭,2013)。

### 4.3 最优概化设计

本文将拉格朗日乘数法应用于国内学生评教的测量研究中,纳入了交叉设计及多侧面混合设计,说明拉格朗日乘数法在寻找概化设计侧面最佳水平数的广泛适应性。本文四个设计的概化研究结果如表5所示:

从概化系数的角度看, $t \times i$ 设计、 $(s:t) \times i$ 设计、 $(s:t) \times (i:v)$ 设计三个概化设计的概化系数均较大,均大于.95,但 $(s:t) \times (i:v) \times o$ 设计的概化系数最小。

从研究设计的角度考虑, $(s:t) \times (i:v)$ 设计既考虑了评教学生数量和评教项目数量,又划分了评教问卷的维度,该设计探测的误差来源更多样,对数据的分析解读更全面,其结果更具说服力。

从侧面最佳水平数的角度看,比较 $n'_s$ 和 $n'_i$ , $(s:t) \times i$ 设计和 $(s:t) \times (i:v)$ 设计的侧面最佳水平数与实际情况较为相符,其结果更为合理。

综合考虑概化系数、概化设计的完整性以及侧面最佳水平数的合理性,本文认为 $(s:t) \times (i:v)$ 设计属最优概化设计。

在评教实践中,单次测量的结果似乎欠缺说服力,但出于成本考虑,很少有学校会对同一任课教师进行重复测量,而该结果也从侧面说明了评教实

践无需进行重复测量的操作便可得到可靠性较高的测量结果。

## 5 结论

(1) 使用拉格朗日乘数法及不同概化设计下概化研究的方差分量,可推导出预算限制下不同概化设计的侧面最佳水平数。

(2) 通过学生评教的实证研究,比较了四个不同概化设计各测量侧面的最佳水平数,说明了拉格朗日乘数法能够有效地适用于预算限制下概化理论不同研究设计,且表现出较强的稳健性。

(3) 比较不同概化设计的研究结果,可得预算限制下的最佳测量设计,在本文四个概化设计中, $(s:t) \times (i:v)$ 设计为学生评教研究中的最优概化设计。

## 参考文献

- 陈社育,余嘉元.(2001).经典真分数理论与概化理论信度观评析.心理科学进展,9(3),258-263.
- 江利.(2017).我国高校学生评教研究综述.黑龙江教育(高教研究与评估),10,57-60.
- 黎光明,张敏强.(2009).用概化理论分析高校教师教学水平评估.高教发展与评估,25(2),68-73.
- 黎光明,张敏强,黄宪,王旭.(2013).概化理论方差分量变异量估计方法.心理学探新,33(3),239-245.
- 马秀麟,袁克定,刘立超.(2014).从大数据挖掘的视角分析学生评教的有效性.中国电化教育,10,78-84.
- 谢博文,史蒂.(2012).国外高校学生评教文献研究成果综述.大学教育,1(2),76.
- 杨志明.(2003).测评的概化理论及其应用.北京:教育科学出版社.
- 赵平.(2018).国内高校学生评教制度研究综述.教育观察,7(7),31-32.
- Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Goldstein, Z., & 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.

表5 四个概化设计研究结果

设计	相对误差 $\sigma_\delta^2$	概化系数 $E_\rho^2$	$n_i$	$n_s$	$n'_i$	$n'_s$
$t \times i$	.00012	.99958	40	139	8	7
$(s:t) \times i$	.00212	.98121	24	263	5	14
$(s:t) \times (i:v)$	.00380	.96630	20	330	4	17
$(s:t) \times (i:v) \times o$	.01662	.83313	15	183	3	10

注:  $n'_i$ 代表量表的每个维度包含的项目数量,  $n'_s$ 代表评价每一位教师的平均评教学生。

- 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. (1992). The optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement*, 52(2), 301–308.
- Meyer, J. P., 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.
- Sanders, P. F. (1992). Alternative solutions for optimization problems in generalizability theory. *Psychometrika*, 57(3), 351–356.
- Sanders, P. F., Theunissen, T. J. J. M., & Baas, S. M. (1989). Minimizing the number of observations: A generalization of the Spearman-Brown formula. *Psychometrika*, 54(4), 587–598.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Thousand Oaks: SAGE Publications.
- Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika*, 38(2), 173–181.

## Estimating the Best Sample Size for Students' Evolution of Teaching—Based on the Application of LaGrange Multiplier Method

Liu Ying, Zhang Minqiang, Zhen Fengquan

(School of Psychology, South China Normal University, Guangzhou, 510631)

**Abstract** The effectiveness of Students' Evolutions of Teaching (SET) is one of the problems that many colleges pay attention to. A great study of SET has the following influences: (1) Promoting the improvement of teaching quality; (2) Focusing on students' experience in class; (3) Providing decision-making ground for managers. The measurement study of SET is affected by the measurement design, the effectiveness of study tools, sample size, and so on. At present, Classical Test Theory (CTT) is most used in the analysis of measurement study of SET. The common problem is without consideration of the effectiveness of SET from the perspective of measurement.

Generalizability Theory (GT) is a statistical theory to evaluate the reliability of behavior, holding the views that the total variance can be decomposed into variance component representing the error of object of measurement and other variance components. Generalization coefficient is the standard of reliability in Generalizability Theory. Generally speaking, the larger the sample size, the higher the generalization coefficient value and the test reliability. However, the measurement cost will increase due to the increase of sample size, which are a dilemma. Therefore, researchers should consider the budget and cost when they explore a measurement procedure. How to get an optimal sample size under budget constraints is one of the problems that cannot be ignored.

LaGrange multiplier method is a method to solve the extremum in the mathematical field, which is most widely used for obtaining the optimal levels of different facets of generalized design under budget constraints.

According to a series of studies of Marcoulides and Goldstein, we summarized the steps of using LaGrange multiplier method to obtain the optimal facets of measurement in generalizability theory under budget constraints. The steps are as follows: (1) Discussing the sources of errors and carrying out generalizability design; (2) Estimating the variance components; (3) Quantizing the constraints on the basis of practical situation; (4) Imposing the LaGrange multiplier to structure the LaGrange function that is used to solve the optimal levels of different facets in generalized design under budget constraints.

This article used Teaching Quality Evaluation Scale for College Teachers (For Students) as a tool to collect data from three different colleges. A total of 530 students took the test and evaluate their teachers. We considered  $t \times i$  design,  $(s:t) \times i$  design,  $(s:t) \times (i:v)$  design, and  $(s:t) \times (i:v) \times o$  design respectively, where  $t$  represents teachers,  $i$  represents items,  $s$  represents students,  $v$  represents the dimensionality of the scale, and  $o$  represents the times of evaluation. We estimated the variance components through urGENOVA. What's more, the budget of one evolution was set at 2500 yuan. The generalization coefficient was used as the comparison index.

The results showed: (1) Using LaGrange multiplier method and variance components to derive the function of optimal levels of different facets is feasible; (2) LaGrange multiplier method can be effectively applied to different designs in generalizability theory under budget constraints and has stronger robustness. (3) Combined with analysis of actual situation, we know the  $(s:t) \times (i:v)$  design is the optimal generalized design in this SET study.

**Key words** budget constraints, generalizability theory, students' evolutions of teaching, LaGrange multiplier method