

我国外语教学研究中的统计分析方法使用调查^{*}

鲍 贵

提要: 本研究纵向调查了 2000–2009 年间我国 8 种外语类主要期刊外语教学论文中统计分析方法的使用特点。研究的主要发现如下: (1) 使用的统计分析方法以简单推理统计为主, 纯描述性统计使用率总体呈下降之势, 多元统计使用率有微弱增长的迹象; (2) 较之于前阶段(2000–2004 年), 后阶段(2005–2009 年) 纯描述性统计报告的完整性和图表使用率有一定程度提高, 但是数据挖掘不够彻底; (3) 简单推理统计和多元统计的使用在前、后阶段均存在问题, 主要包括误解、误用和简化统计检验形式等。

关键词: 统计分析方法; 纯描述性统计; 简单推理统计; 多元统计

Abstract: This study is a longitudinal survey of the statistical methods used in the papers on EFL teaching published in 8 key Chinese EFL journals from 2000 through 2009. It reveals the following major findings. (1) Simple inferential statistics is most frequently used. The use of purely descriptive statistics presents a roughly downward tendency, whereas the use of multivariate statistics increases slightly. (2) With respect to purely descriptive statistics, the completeness of statistical reports and use of tables/graphs improve, but the data mining in the latter period (2005–2009) is not as thorough as in the former (2000–2004). (3) Problems arise in the use of simple inferential and multivariate statistics in both the former and latter periods, including misconceptions, misuses and simplifications of statistical tests.

Key words: statistical method; purely descriptive statistics; simple inferential statistics; multivariate statistics

中图分类号: H319 文献标识码: B 文章编号: 1004–5112(2012) 01–0044–09

1. 引言

外语教学研究离不开定量分析。通过定量分析, 我们既可以检验和发展外语教学理论, 也可以了解外语学习者语言发展的整体特点, 还可以探索出制约学习者语言发展的多重因素。一篇好的定量型研究论文有助于我们更新外语教学观念, 为外语教学实践提供有益的指导; 一篇差的论文则不但没有什么教学价值, 甚至还会误导我们。因此, 如何正确评价定量型研究论文的质量近年来引起了我国学者的注意。吴旭东和张文忠(2002) 从实验研究的多个环节着手, 系统评价了国内外语教学实验研究的质量。也有研究者(王立非 2002; 张少林 2004; 张少林, 夏荣良 2005; 赵蔚彬 2002) 聚焦于外语教学论文中统计分析存在的问题或问卷调查统计概念及方法的运用情况。还有研究者(如葛诗利 2010) 则探讨某个统计方法使用的恰当性问题。毫无疑问, 上述研究对我们提高定量型研究报告的质量具有重要借鉴意义。

总体而言, 外语界对定量型研究论文质量的探讨还不够深入。我们对外语教学研究中统计分析方法使用的纵向特点还不甚清楚, 对研究中不同统计分析方法使用问题的原因还缺乏全面的认识。鉴于此, 本研究拟纵向调查 2000–2009 年间我国 8 种外语类主要期刊外语教学论文中统计分析方法的使用特点, 指出存在的主要问题, 提供解决办法。

2. 研究设计

2.1 研究问题

本研究主要探讨两方面的问题:

(1) 我国外语教学研究使用的统计分析方法呈现怎样的发展特点? 具体而言, 统计分析方法的使用呈

^{*} 本文为教育部人文社会科学研究基金项目(编号: 09YJAZH044)、江苏省社会科学基金项目(编号: 09YYB010) 和江苏省高校“青蓝工程”(编号: 苏教师(2010) 27 号) 的部分成果。

现怎样的总体变化和阶段性变化?

(2) 我国外语教学研究中统计分析方法使用的正确性如何? 具体而言, 统计分析方法的使用存在怎样的问题以及出现这些问题的主要原因是什么?

2.2 期刊论文收集

本研究收集的外语教学论文为 2000-2009 年发表在我国外语类主要期刊上的定量型研究论文。期刊和论文的选择标准是: (1) 期刊为 CSSCI 来源期刊, 且较多地刊登实证性外语教学研究论文; (2) 既考虑期刊的排名顺序(以 2008-2009 年 CSSCI 来源期刊排名为依据), 又考虑期刊的地区覆盖面; (3) 论文为每种期刊每期第一篇实证性研究论文, 论文的研究对象是以英语为外语的中国英语学习者; (4) 论文检索源为“中国期刊全文数据库”、“中文科技期刊数据库(维普)”和“万方数据资源系统”。根据以上标准, 筛选出 8 种期刊, 即《外语教学与研究》、《现代外语》、《外语界》、《解放军外国语学院学报》、《外语学刊》、《外语教学》、《外语电化教学》以及《外语研究》。10 年中上述 8 种期刊每期刊登的首篇外语教学实证研究论文的总数依次为: 54、36、51、48、15、44、37 和 36, 合计 321 篇; 所有期刊各年首篇论文总数依次为: 17、23、26、30、32、36、37、40、38 和 42。

2.3 期刊论文分析

本研究将期刊论文中使用的统计分析方法分为 3 大类: 纯描述性统计、简单推理统计和多元统计。纯描述性统计指限于样本本身数据特征的概括性描述, 包括平均数、标准差和频数等。这类统计的评价包括 3 个方面: 数据报告的完整性、图表的使用和数据挖掘(data mining) 的彻底性。数据报告的完整性指描述性统计的报告应尽量全面反映样本数据的整体特点。尽管研究者对“完整性”没有统一的意见, 但是至少对报告的统计量应该有个底线要求。如果研究变量是连续性变量, 至少应该报告平均数和标准差。如果研究变量是类别变量, 应该报告频数和百分比(或相对频数)。本研究根据以上底线要求将纯描述性统计一分为二: 完整与不完整。统计时, 如果一项研究既涉及连续性变量又涉及类别变量, 则将它们分别计为频数 1。图表可以使作者有效地报告大量的信息, 增强数据的可理解性。它们不仅有利于交流, 而且还增加了论文的视觉吸引力 (American Psychological Association 2010: 125-126)。本研究将是否使用图表视作描述性统计报告质量评估的一个方面。统计时, 只要一篇论文使用了图、表或兼而用之, 不管数量多少都计为频数 1。数据分析的彻底性指期刊论文对数据的挖掘是否彻底, 即某篇论文是否只限于纯描述性统计。本研究按照以下宽泛的标准确定一篇论文可以限于纯描述性统计: (1) 研究一个或多个组别, 但是小组样本太小, 比如小于 10, 或者针对不同组别的调查问题不同, 组间可比性不强; (2) 针对单个组的问卷调查; (3) 基于语料库的频数研究(这类研究可用推理统计, 但是不能用研究者普遍熟悉的卡方检验)。可以限于纯描述性统计的论文视作数据挖掘彻底, 否则视作数据挖掘不彻底。

简单推理统计指限于一个自变量和一个因变量或只有两个变量的推理统计分析, 分为 3 种统计检验类型。前两类是参数和非参数检验, 两个类型又包括具体的检验形式(详见下文表 3)。第 3 类为统计方法不明类, 指某研究只报告 p 值和(或) p 值的统计显著性程度, 或者在报告相关系数时没有指明也没有用系数符号(如 r 或 r_s) 暗示该检验是皮尔逊还是斯皮尔曼等相关分析。

多元统计比简单推理统计更为复杂, 涉及多个自变量(含协变量)和(或)多个因变量, 包括不同的检验形式(详见下文表 4)。由于多元统计除个别统计形式为非参数检验(如 Logistic 回归)之外, 其余均为参数检验, 所以多元统计不再按照参数和非参数检验进行细分。统计时, 逐年对期刊论文使用的具体检验形式进行计数, 对使用错误的检验形式也进行标记。若同一篇论文多次使用同一种统计检验形式, 则只按 1 次计数。若同一篇论文使用不同的统计检验形式, 则累积计数。同一项研究多处研究方法不明, 不重复计数, 只计为频数 1。计数过程中对各篇论文错误使用统计检验形式的原因进行定性分析。

最后, 根据研究问题对各个统计数据与分析结果进行汇总与归纳。为了便于清晰地把握期刊论文统计分析方法的使用特点, 在概括统计分析方法总体特点之外, 我们还按 2000-2004 年和 2005-2009 年两个阶段对论文中使用的不同类别统计分析方法展开详细的对比分析。

3. 结果与讨论

3.1 不同类型统计方法的纵向变化趋势

2000-2009 年 8 种期刊论文 3 类统计方法使用频数及比例见表 1。

表 1 2000-2009 年期刊论文统计方法使用频数及比例

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
纯描述性统计	7	11	8	8	9	13	11	7	10	6
	33%	37%	22%	17%	18%	28%	23%	11%	16%	8%
简单推理统计	13	18	25	30	32	29	30	45	44	56
	62%	60%	68%	65%	65%	63%	63%	73%	69%	75%
多元统计	1	1	4	8	8	4	7	10	10	13
	5%	3%	11%	17%	16%	9%	15%	16%	16%	17%
总计	21	30	37	46	49	46	48	62	64	75

表 1 显示,在 3 类统计中,纯描述性统计的使用率在 2001 年最高(37%),到最后一年降到最低值(8%)。不论在哪一年,3 类统计中简单推理统计的使用率都最高,均在 60% 以上。这类统计的使用率在 2001 年最低(60%),最后一年最高(75%),与纯描述性统计的使用率呈相反之势。除 2007 和 2009 年以外,在 3 类统计中多元统计的使用率每年都最低。从纵向变化来看,多元统计使用率在 2001 年最低(3%),2003 和 2009 年最高(17%),中间经历了一些起伏变化。

图 1 更清晰地揭示了不同类型统计分析方法使用率变化的特点:(1) 简单推理统计变化曲线位于图上方,远远高于纯描述性统计和多元统计的变化曲线,说明简单推理统计在我国现阶段外语教学研究统计分析中占主导地位。(2) 在前阶段(2000-2004 年),除 2002 年之外其他各年简单推理统计使用率的变化幅度不大;在后阶段(2005-2009 年),2007 年之后这类统计的使用率在一个新的高度起伏。(3) 纯描述性统计的变化曲线呈 3 座山的形状,3 座山的峰顶依次降低,说明纯描述性统计的使用率大体呈下降之势。(4) 多元统计的变化走势与简单推理统计相近,说明随着简单推理统计使用率的提高,使用多元统计的可能性也在增加。(5) 多元统计的变化出现两个高峰。第一高峰在前阶段末,坡度较陡。第二高峰在后阶段末,坡度较缓。(6) 2007 年前,纯描述性统计变化曲线高于多元统计曲线,尽管在 2003 年两者的使用率相同。在 2007 年后,多元统计变化曲线与纯描述性统计变化曲线交叉,多元统计的使用率有超越纯描述性统计之势。

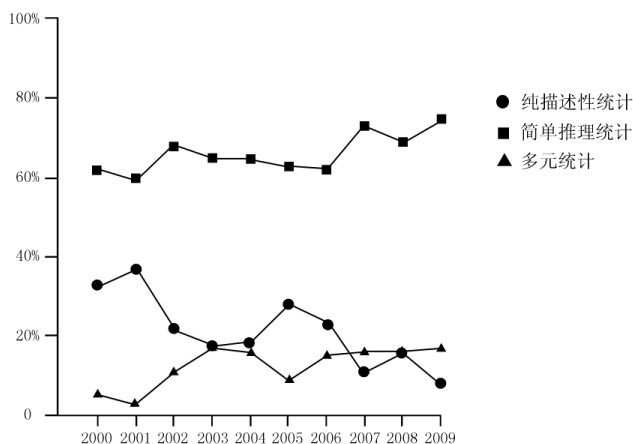


图 1 2000-2009 年期刊论文统计方法使用变化趋势

3.2 纯描述性统计

表2从3个方面比较了两个阶段期刊论文中的纯描述性统计使用情况。

表2 纯描述性统计两个阶段的使用比较

阶段(年)	完整性		图表		彻底性	
	完整	不完整	有	无	彻底	不彻底
2000-2004	23(45%)	28(55%)	35(81%)	8(19%)	33(77%)	10(23%)
2005-2009	38(72%)	15(28%)	46(98%)	1(2%)	28(60%)	19(40%)

从表2可见,两个阶段大多数研究报告结果时均使用了图表。数据报告完整的研究在前阶段没有过半,在后阶段则占到总数的近3/4。比较前、后阶段发现,在数据报告的完整性和图表使用两个指标上,后阶段均明显好于前阶段,但是数据挖掘的彻底性却差于前阶段。

3.3 简单推理统计

3.3.1 简单推理统计使用的阶段性特点

表3比较了两个阶段不同统计检验形式的使用情况。前、后阶段使用的简单推理统计总数分别为118例(使用参数检验88例,使用非参数检验14例,统计检验方法不明16例)和204例(使用参数检验145例,使用非参数检验33例,统计检验方法不明26例)。不论在哪个阶段,参数检验的数量远远高于非参数检验。对此可以做出以下解释:(1)参数检验的方法众多,研究者的选择面广;(2)研究数据更适合进行参数检验;(3)参数检验具有稳健性,研究者因此会不充分考虑数据分布的特点和参数检验的假设(assumption),过度使用参数检验;(4)研究者不知道参数检验的要求,一味使用参数检验。需要注意的是,参数检验稳健与否是有前提条件的,包括样本的大小、各单元格容量是否相等、正态分布和方差齐性假设满足情况或违反程度等。为了便于深入调查各种参数检验形式的使用情况,笔者假定各个检验条件都已满足,相关建议也以此为基础。

表3 简单推理统计两个阶段的使用比较

参数检验									
	单样本 t 检验	独立样本 t 检验	配对样本 t 检验	t 检验 方法不明	Z 检验	单因素 方差分析	单因素重 复测量方 差分析	皮尔逊相 关分析	简单回 归分析
2000-2004	1	13	9	15	2	26	0	21	1
2005-2009	0	31	25	15	6	36	6	22	4
非参数检验									
	卡方检 验	斯皮尔 曼相关分 析	秩次相 关 G	Mann- Whitney U 检验	Kruskal- Wallis 检验	Wilcoxon 配对符号 秩次检验	/		
2000-2004	7	2	2	3	0	0	16		
2005-2009	25	0	0	6	1	1	26		

表3还显示,两个阶段使用的检验种类数相同。两个阶段使用频繁的参数检验形式依次为:t检验>方差分析>皮尔逊相关分析。非参数检验中,两个阶段使用率最高的检验形式为卡方检验。值得注意的是,后阶段使用卡方检验的比例为76%(25/33),明显高于前阶段50%(7/14)的比例。究其原因,随着语料库研究的

增多,使用卡方检验作为推理统计方式的研究随之增多。

3.3.2 简单推理统计使用的主要问题

在常用的检验形式中,t 检验、简单方差分析和卡方检验的使用问题最严重。下面分别对这 3 种检验形式进行讨论。

3.3.2.1 t 检验

表 3 表明,前阶段使用的 t 检验总量中,该检验使用方法不明的比例接近 40%(15/38)。后阶段的情况虽有好转,但是仍有 21%(15/71)的 t 检验方法不详。由于作者没有明确说明是使用独立还是配对样本 t 检验,加之统计结果的报告不完整,我们无法对这些 t 检验结果的可信度做出合理判断。

在前阶段使用的 3 类 t 检验中,单样本 t 检验有 1 例,且使用正确。前、后阶段的独立样本 t 检验分别有两例和 13 例存在问题,正确率分别为 85%(11/13)和 58%(18/31)。错误原因包括:(1) 错误理解独立样本 t 检验的概念:(1a)对同一批受试两个实验条件下的比较采用独立样本 t 检验,正确的应采用配对样本 t 检验;(1b)误将单样本 t 检验称作独立样本 t 检验。(2) 简化统计推理手段:(2a)对多个组别比较重复采用独立样本 t 检验,扩大第一类错误率,应采用单因素方差分析;(2b)将多因素设计应采用的因素方差分析简化为多次采用独立样本 t 检验;(2c)对混合设计(mixed design)采用独立样本 t 检验,以比较不同组组内变量在每个水平上的均值是否存在差异,正确的应采用混合方差分析;(2d)对混合设计并用独立和配对样本 t 检验;(2e)混合设计中发现交互作用,却使用独立样本 t 检验比较一个自变量两个水平在另一个自变量每个水平上的均值差异,正确的应采用交互作用对比(interaction contrast)。

前、后阶段配对样本 t 检验分别有 4 例和 10 例错误,正确率分别为 56%(5/9)和 60%(15/25)。错误原因包括:(1) 错误理解配对样本 t 检验的概念:(1a)对两个不同组受试的比较采用配对样本 t 检验,正确的应采用独立样本 t 检验;(1b)误将单样本 t 检验称作配对样本 t 检验。(2) 简化统计推理手段:(2a)多个(两个以上)重复测量水平上的比较采用配对样本 t 检验,扩大第一类错误率,应采用单因素重复测量方差分析;(2b)对混合设计并用配对样本 t 检验与其他检验形式。

从以上比较可以发现,前、后阶段配对样本 t 检验使用的正确率相近,但是后阶段独立样本 t 检验使用的正确率明显低于前阶段。出现后一个结果的原因在于:后阶段研究中过多使用了简化的推理统计手段。前阶段各有 1 例使用独立和配对样本 t 检验简化统计分析,错误率为 8%(1/13)和 11%(1/9),但是后阶段使用独立和配对样本 t 检验简化统计分析的案例分别达到 8 例和 7 例,错误率分别为 26%(8/31)和 28%(7/25)。这说明,随着外语教学研究的深入,研究设计也开始复杂起来,但是统计分析手段的使用却没有跟上研究设计的发展。

3.3.2.2 简单方差分析

前阶段有 8 例错误使用单因素方差分析,正确率为 69%(18/26);后阶段有 14 例错误使用该分析,正确率为 62%(23/37),1 篇论文中有 1 例正确、1 例错误,反而低于前阶段。错误原因包括:(1) 错误理解单因素方差分析的基本要求:(1a)对同一组受试多个实验条件下测量结果的比较是组内比较,研究者错误采用单因素方差分析,忽视了组内设计和组间设计的区别,应采用单因素重复测量方差分析;(1b)对不独立频数的组间比较错误采用单因素方差分析;(1c)虽然论文声称采用单因素方差分析,但是根据自由度判断单因素方差分析统计结果为误。(2) 简化统计推理手段:(2a)对多个组间变量各水平进行比较时,错误地对每个组间变量分别采用单因素方差分析,正确的应采用因素方差分析;(2b)研究设计为混合设计,但是对组间和组内变量分开进行统计检验,包括对组间因素采用单因素方差分析,该设计适合采用混合方差分析。

前阶段没有使用单因素重复测量方差分析,后阶段有两例错误使用该分析,正确率为 71%(5/7),1 篇论文中 1 例正确、1 例错误。错误原因包括:(1) 对一个组间变量多个水平上的比较错误采用单因素重复测量方差分析,应采用单因素方差分析。(2) 简化统计推理手段:研究设计为双因素重复测量设计,但是研究者对各个组内因素分开采用单因素重复测量方差分析和配对样本 t 检验(组内变量有两个水平),恰当的统计方法则是双因素重复测量方差分析。

3.3.2.3 卡方检验

前、后阶段均只有 1 例正确使用了卡方检验,正确率分别是 14% (1/7) 和 4% (1/26, 1 篇论文中 1 例正确、1 例错误)。由此可见,错误使用卡方检验的问题很严重。究其错误原因,有以下 4 种情形,有时多种情形并存于同一项研究中:(1) 研究数据不是频数,对连续性因变量组间多个水平上的比较错误使用卡方检验,可考虑使用单因素方差分析。(2) 频数数据不独立,解决方法可参考下文相关建议。(3) 计算事件发生的频数时却没有计算事件未发生的频数,恰当的做法是同时计算事件发生与未发生的频数。(4) 作者对配对样本频数差异错误采用卡方检验,正确的应采用 McNemar 检验。

卡方检验最重要的假设是独立性假设(assumption of independence)。该假设意味着,每个受试只能进入列联表中的一个单元格。当受试反应不止一次时,观测值便不独立。关于独立性概念的详细解释,可参见 Hatch 和 Lazaraton (1991: 408-409) 的论著。

当前,语料库研究最常用的统计方法是卡方检验,由于研究的频数在语料库中不具有独立性,因而使用卡方检验并不恰当。Rietveld 等(2004) 通过模拟实验证实了在频数数据不独立时,卡方检验大大提高了第一类错误率,并建议在处理 2×2 列联表时先对每个文本中非独立频数的比数进行对数转换,再使用独立样本 t 检验比较组间差异。

3.4 多元统计

3.4.1 多元统计使用的阶段性特点

表 4 报告了两个阶段各种多元统计形式的使用数量。前、后阶段使用多元统计的研究分别为 22 例和 44 例。从使用类别来看,后阶段较之于前阶段多使用了 4 种,但是频数很小。因此,从多元统计的使用数量和种类来看,后阶段相对于前阶段有一定程度的提高,呈现出发展的迹象。前、后阶段均使用较多的检验形式是高级方差分析和多元回归分析,且在多元回归分析中又以多元线性回归分析占绝大多数。虽然两个阶段结构方程模型的使用数量比因子分析少,但是使用出现的问题却较多。下一节重点讨论高级方差分析、多元线性回归分析和结构方程模型使用存在的问题。

表 4 多元统计两个阶段的使用比较

	单因素 重复测 量协方 差分析	因素方 差分析	多因素 重复测 量方差 分析	混合方 差分析	多元方 差分析	偏相关 分析	多元线 性回归 分析	Logistic 回归	分类 回归	因子 分析	结构方 程模型
2000-2004	0	5	1	3	1	0	5	0	0	5	2
2005-2009	1	8	6	6	1	2	9	1	1	5	4

3.4.2 多元统计使用的主要问题

3.4.2.1 高级方差分析

从因素方差分析来看,前、后阶段分别有 3 例和 8 例错误。错误原因包括:(1) 研究设计为多因素重复测量设计,却错误地使用因素方差分析,恰当的统计方法是多因素重复测量方差分析。(2) 研究设计为混合设计,却错误使用因素方差分析,恰当的统计方法是多因素混合方差分析。

在多因素重复测量方差分析方面,前、后阶段各有 1 例和 3 例错误。错误原因包括:(1) 文中声称使用多因素重复测量方差分析,但是根据提供的数据和统计量可以判定没有正确使用该分析。(2) 简化统计推理手段:研究设计为混合设计,却错误地在每个组间水平上分别使用重复测量方差分析,恰当的统计方法是多因素混合方差分析。

前、后阶段各有 1 例和两例错误使用混合方差分析。错误原因包括:(1) 文中声称使用混合方差设计,但是根据提供的数据和统计量可以判定没有正确使用混合方差分析。(2) 简化统计推理手段:研究设计含组间、组内变量以及多个因变量,却错误使用多个不同的检验方法,其中包括混合方差分析,恰当的统计分析方法是混合多元方差分析。

前、后阶段各有 1 例使用多元方差分析,但是前阶段分析正确,后阶段分析错误,错误原因在于简化统计推理手段:研究设计含组间、组内变量以及多个因变量,却错误使用多个不同的检验方法,其中包括多元方差分析,恰当的统计分析方法是混合多元方差分析。

此外,后阶段有 1 例单因素重复测量协方差分析,但是研究设计包括两个组内因素和 1 个协变量,研究者分析时采用简化的方法在各个组内变量上分别进行单因素重复测量协方差分析,恰当的方法是多因素重复测量协方差分析。

概括起来,前阶段共错误使用 5 例高级方差分析,正确率为 50% (5/10);后阶段共有 15 例错误使用高级方差分析,正确率仅为 32% (7/22)。虽然后阶段高级方差分析的使用种类和数量略多于前阶段,但是使用正确率却低于前阶段。出现这一情况的主要原因是,外语教学研究设计趋于复杂,但是统计分析手段运用水平却没有相应提高。

3.4.2.2 多元线性回归分析

前、后阶段共有 14 例多元线性回归分析,除后阶段有 1 例误用多元线性回归分析(数据为频数)之外,其余 13 例(数据都为连续性数据)符合线性回归分析的要求。在所有 14 例研究中没有 1 例讨论多元线性回归分析的一些重要统计假设,如检查异常值和多元正态性、诊断多元共线性和进行残差分析等。我们需要把统计假设的检验当做一项常规的检查工作,以便提高结论的可信度。下面从两个主要方面讨论多元线性回归分析存在的问题。

一方面,忽略重要的统计假设会导致不可靠的结果。由于研究论文均忽略了重要的统计假设,我们只能将相关分析与回归分析对照起来检查研究中存在的问题。前阶段 5 例中有 4 例提供相关系数矩阵,但只报告了预测变量与因变量之间的相关系数,因而报告不完整,另 1 例则没有提供相关分析结果。后阶段 9 例中有 5 例完整地报告了各个变量之间的相关系数矩阵,4 例没有提供相关分析结果。通过检查相关系数矩阵,我们发现前阶段至少有 1 例存在多元共线性的可能性,后阶段有 1 例结果存在疑问。其中,有 1 例发现多个听力策略能够预测听力成绩变异的 1/4,但是每个听力策略对听力成绩均没有显著的独特贡献($p > .05$)。从文中提供的不完整的相关系数矩阵来看,每个听力策略与听力成绩均呈低、中度显著相关,这与回归分析中每个听力策略都没有预测力的发现相矛盾。造成这一结果的原因有可能是各个听力策略之间存在多元共线性,因为随着共线性的增加,回归系数的标准误差也相应增大,从而影响了回归系数的统计显著性。多元共线性有多种解决办法,包括删除或合并预测变量,也可以采用岭回归(Ridge Regression)。另一项研究用逐步回归的方法调查英语学习者自主学习、弥补策略对英语学业成就的预测力,发现两个预测变量均有独特的预测力,其中自主学习(第一个预测变量)的预测力为 98.5%,弥补策略(第二个预测变量)的预测力为 0.3%。但是,相关矩阵显示自主学习、弥补策略与英语学业成就的相关系数分别是 .203 ($p < .05$) 和 .034 ($p > .05$)。相关系数的平方值与回归预测力的极大差异令人怀疑研究结果,作者也未对此提供解释。一般情况下,如果采用逐步回归法,第一个预测变量的预测力应为 4% (0.203^2),第二个预测变量很有可能没有预测力。

另一方面,对多元回归分析中的交互作用问题重视不够。传统上,忽视多元回归分析中的交互作用问题是各个学科统计分析中普遍存在的问题。近年来,有学者(如 Jaccard *et al.* 1990)开始专门研究这一课题。值得一提的是,我们发现后阶段研究中有 1 例涉及了多元回归分析中的交互作用问题,研究了单词解码与英语语言理解对英语阅读理解影响的交互作用。交互作用问题凸显了回归分析中进行残差分析和检查散点图的重要性。我们在后阶段论文中发现 1 例存在交互作用。该例调查汉语阅读能力向英语阅读能力迁移的问题。研究检验高、低两个英语水平组汉语阅读能力和英语水平对英语阅读能力的预测作用,发现低水平组的两个预测变量均没有预测力,而高水平组的汉语阅读能力有独特预测力,英语水平则没有。此例说明,汉语阅读能力与英语水平有交互作用存在。但是,作者没有进行交互作用分析,而是将英语水平(连续性变量)分为二分类别变量分别进行回归分析,削弱了研究结果。

3.4.2.3 结构方程模型

结构方程建模是当前最复杂的统计技术,综合了多元回归分析、因子分析和多元方差分析的优点。在建模前,检查多元正态性、线性和异常值等是必不可少的步骤。由于前、后阶段 6 例研究中的结构方程模型分析

没有报告这些信息,对此我们不再讨论。下面从3个方面探析6例结构方程模型分析中存在的问题。

首先,建模所需的信息报告不全。在前面讨论多元回归分析时,我们提及了报告相关矩阵的重要性。同样,在结构方程建模中,我们也应该报告此类信息。美国心理协会(American Psychological Association 2001: 161)建议研究报告路径分析和线性结构关系分析使用的所有输入变量的平均数、标准差和相关矩阵,因为这些数据对读者验证分析或重新进行分析至关重要,对于归档也必不可少,比如对研究进行元分析(meta-analysis)就用得到。按照这一标准我们发现,后阶段各有1例研究报告了部分变量的平均数和标准差,1例报告了相关系数矩阵,其他4例均未提供这些基本信息。

其次,拟合指数种类报告不全。拟合指数有多种,每种又有多个具体的指数。报告多少指数才算报告完整,这没有定论。结构方程模型的拟合指数分3类,分别从不同层面评价模型的拟合情况(Kelloway 1998: 24-33)。第一类是绝对拟合指数(absolute fit index),检验观测的与模型再生的方差及协方差矩阵拟合的程度,包括卡方(χ^2)和拟合优度指数(goodness-of-fit index)等。第二类是比较拟合指数(comparative fit index),比较假设的模型优于独立模型(independence model)的程度,包括基准拟合指数(normed fit index)和比较拟合指数等。第三类是简约拟合指数(parsimonious fit index),检查达到某种程度拟合的必要参数量,惩罚参数过多的模型,包括Akaike信息准则(Akaike information criterion)和简约拟合指数等。按照这一标准,后阶段研究的拟合指数报告好于前阶段:前阶段1例研究报告第一类指数,1例则没有报告任何拟合指数;后阶段1例研究报告第一类指数,3例报告第一、二类指数。需要指出的是,这些研究没有进行分类报告,且对指数概念解释得不透彻。我们建议研究分类报告模型拟合指数,每类报告2-3个指数。

最后,对卡方统计量显著性和标准化系数大于或等于1的解释存在问题。前阶段有1例、后阶段有两例研究报告卡方值具有统计显著性($p < .001$),却错误地认为该结果表明模型拟合。卡方检验的零假设是指观测的与模型再生的方差及协方差矩阵相同。卡方值的统计显著性意味着拒绝零假设,模型再生的与观测的方差及协方差矩阵不同,模型不理想。因此,卡方统计量实际上是模型拟合劣度指数(badness-of-fit index)。当然,卡方统计量的显著性并不表示一定要放弃假定的模型。决策时需要考虑样本的大小和观测变量的多元正态性,并参照其他拟合指数值。关于标准化系数问题,通常情况下我们期望系数值小于1。但是,前、后阶段各有1例研究报告路径系数为1或大于1,研究者没有对此做出合理的解释。尽管路径系数等于或大于1未必意味着模型一定不可接受,但是研究者需要查找原因,比如各个观测变量之间是否存在共线性问题。

4. 结语

本研究采用定量统计和定性分析相结合的方法调查了2000-2009年我国8种外语类主要期刊教学研究论文使用统计分析方法的特点。总体上,统计分析方法的使用以简单推理统计为主,纯描述性统计使用率呈下降之势,多元统计使用率呈上升之势。纯描述性统计数据报告的完整性和图表的使用在后阶段有进步,但是数据挖掘不如前阶段彻底。简单推理统计和多元统计的使用在前、后阶段均存在问题,且问题的原因较为复杂。无论在哪个阶段,统计检验的使用均存在误解、误用和简化等现象。

要提高统计分析的质量,外语教学研究者需了解统计学基本理论,熟悉不同统计分析方法的使用条件和适用范围。对于外语类期刊来说,细致审查论文统计分析方法的正确性和合理性是提升办刊质量的关键环节之一。如果条件许可,审稿不妨采用双审制,即聘请外语教学研究专家和应用统计学专家同时评阅论文。如果审稿者既精通外语教学理论又熟悉应用统计学理论与实践,那就再好不过了。

参 考 书 目

- [1] American Psychological Association. *Publication Manual of the American Psychological Association* (5th Ed.) [Z]. Washington, DC: Author, 2001.
- [2] American Psychological Association. *Publication Manual of the American Psychological Association* (6th Ed.) [Z]. Washington, DC: Author, 2010.
- [3] Hatch E. & Lazaraton A. *The Research Manual: Design and Statistics for Applied Linguistics* [M]. Boston: Heinle & Heinle Publishers, 1991.

(下转第60页)

- Sequences: Acquisition, Processing and Use* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2004. 1-22.
- [31] Selinker L. Interlanguage [J]. *International Review of Applied Linguistics*, 1972, (10): 209-231.
- [32] Slabakova R. L2 fundamentals [J]. *Studies in Second Language Acquisition*, 2009, (31): 155-173.
- [33] Song H S & Schwartz B. Testing the fundamental difference hypothesis: L2 adult, L2 child, and L1 child comparisons in the acquisition of Korean Wh-constructions with negative polarity items [J]. *Studies in Second Language Acquisition*, 2009, (31): 323-361.
- [34] Urponen M I. *Ultimate Attainment in Post-puberty Second Language Acquisition* [D]. Boston: Boston University, 2004.
- [35] White L. *Second Language Acquisition and Universal Grammar* [M]. Cambridge: Cambridge University Press, 2003.
- [36] Wray A. Formulaic sequences in second language teaching: Principle and practice [J]. *Applied Linguistics*, 2000, (21): 463-489.
- [37] 戴曼纯. 对“根本性差异假说”的质疑[J]. 外语与外语教学, 2012(待刊).
- [38] 戴曼纯, 康悦. 二语习得研究语言学视角的理论思考——形式—形态分离与句法—形态映射[J]. 语言教学与研究, 2009, (4): 55-63.
- [39] 段士平. 国内二语语块教学研究述评[J]. 中国外语, 2008, (4): 63-67.
- [40] 骆雁雁. 基于语块理论的大学英语词汇教学模式研究[J]. 外语学刊, 2009, (6): 168-170.
- [41] 王立非, 陈香兰. 语言语块教学与研究在中国的进展——“首届全国语言语块教学与研究学术研讨会”综述[J]. 外国语, 2009, (6): 90-94.
- [42] 于秀莲. 语块教学法与提高英语应用能力的实验研究[J]. 外语界, 2008, (3): 54-61.

作者单位: 北京外国语大学中国外语教育研究中心 北京 100089

////////////////////////////////////
(上接第 51 页)

- [4] Jaccard J, Turrissi R & Wan C K. *Interaction Effects in Multiple Regression* [M]. London: Sage Publications, Inc., 1990.
- [5] Kelloway E K. *Using LISREL for Structural Equation Modeling: A Researcher's Guide* [M]. London: Sage Publications, Inc., 1998.
- [6] Rietveld T, Van Hout R & Ernestus M. Pitfalls in corpus research [J]. *Computers and the Humanities*, 2004, (38): 343-362.
- [7] 葛诗利. 语料库间词汇差异的统计方法研究[J]. 现代外语 2010 (3): 249-257.
- [8] 王立非. 应用语言学统计研究方法的实证调查——态度与实践[J]. 外语研究 2002 (1): 66-70.
- [9] 吴旭东, 张文忠. 我国外语教学实验研究质量调查[J]. 外语教学与研究 2002 (1): 35-44.
- [10] 张少林. 科学形式下的“不科学”——析外语科研论文中虚构统计推断结果现象[J]. 国外外语教学, 2004 (4): 48-54.
- [11] 张少林, 夏荣良. 外语定量实证研究统计失误: 分析与对策[J]. 湖南大学学报(社会科学版) 2005 (5): 97-101.
- [12] 赵蔚彬. 十种外语类核心期刊(1995-2000) 统计手段使用调查[J]. 解放军外国语学院学报 2002 (5): 62-66.

作者单位: 南京工业大学 江苏 南京 210009