我国英语语言学博士生实验研究类论文质量评价

鲍贵
(南京工业大学 外国语言文学学院,江苏 南京 211816)

摘 要:本文首次利用实验研究效度指标体系实证评价我国英语语言学博士生实验研究类论文的质量。通过对 2005—2014 年间 104 篇博士论文在设计特征、操作程序、统计分析和报告实践等方面的调查发现,在 32 项效度指标中,14 项指标上的效度在至少 3/4 的博士论文中未能体现。在内部效度方面,博士生普遍缺乏效度威胁意识和设计局限意识。构念操作定义不充分以及没有使用双盲是构念效度的主要威胁。统计结论效度威胁主要包括严重忽略测量信度、效应量、统计效力以及统计假设检验。在外部效度方面,总体效度和子群体推广普遍被忽视。博士论文的研究效度没有呈现出随时间持续增加的趋势。

关键词:博士论文;实验研究;效度;评价
中图分类号:H313 **文献标志码**:A **文章编号**:1674-6414(2020)01-0098-09

0 引言

近年来,应用语言学领域开始重视研究质量评价。研究质量以研究报告的质量为前提。报告质量以充分性和透明度为原则,反映研究要素或事实陈述的清晰度与完整性。研究质量,即研究本身的质量,是依据报告事实或证据做出的价值判断,体现研究设计的合理性、变量测量的准确性、统计分析的恰当性和结论的可推广性。报告得当会增加研究结论的可信度,为研究价值判断提供依据;报告不当则会给研究质量带来不确定因素,削弱研究的价值。

论文报告指导原则或标准的建议(Larson-Hall et al. ,2015;Norris et al. , 2015)主要依据《美国心理协会出版手册》(2010)。Norris et al. (2015)就语言学习研究论文中方法论和结果部分的报告提出了一些基本原则。在这些原则中,有些适用于不同类研究,如实验和调查研究中实施测量,有些则具有研究特质性,如实验研究中使用随机分配。在结果报告标准方面,Larson-Hall et al. (2015)与 Norris et al. (2015)提出了大致相同的建议。不过,Larson-Hall et al. (2015)强调元分析思维模式的重要性,对结果报告建议的论述更充分。

论文质量评价性研究为数不多,主要集中于期刊论文实验研究(Plonsky et al. ,2011;Plonsky,2013/2014;Plonsky et al. ,2016;吴旭东 等, 2002)。在观察性研究评价领域,只有个别研究剖析调查类研究期刊论文存在的问题,如郑新民等(2014)。另外,也有一些研究(Lindstromberg,2016;潘珣祎 等,2008;何家宁 等, 2009;鲍贵, 2012)调查期刊论文数据收集或统计分析问题,一定程度上反映出期刊论文存在的质量问题。从整体上看,实验研究论文的评价尚需系统化。

坎贝尔(Campbell)及其同事开创的效度框架(validity framework)为系统化评价实验研究方法论的质量提供了理论依据(Campbell et al. , 1966;Cook et al. ,1979;Shadish et al. , 2002)。效度框架以效度分类和效度威胁清单为特色。根据 Shadish et al. (2002),效度分为四类:内部效度(internal validity)、构念效

收稿日期:2019-05-15
基金项目:江苏省社会科学基金项目“二语习得实验研究方法论评价研究”(18YYB013)的阶段性成果
作者简介:鲍贵,男,南京工业大学外国语言文学学院教授,博士,主要从事应用语言学及应用统计学研究。

度(construct validity)、统计结论效度(statistical conclusion validity)和外部效度(external validity)。关于效度框架的详细介绍与评论,参见鲍贵(2015)。迄今为止,尚没有应用语言学评价性研究完整地利用这一效度框架。本研究尝试采用这一效度框架较为系统地评价我国博士生学位论文报告的实验研究。

选择我国博士生实验研究类论文作为评价对象的主要理由在于学位论文方法论质量评价研究匮乏。郑新民(2009)发现,国内博士学位论文在这一方面存在更为严重的问题,因而有必要展开深入研究。

1　研究设计

1.1 研究问题

本文主要回答以下两个问题:

(1)博士生实验研究类论文在各类效度整体上呈现怎样的阶段性特点?

(2)博士生实验研究类论文在各类效度指标上总的特点和阶段性特点是什么?

1.2 实验研究效度评价指标体系

本次实验研究质量评价以研究效度为依据,效度评价指标体系的构建主要参照 Shadish et al.(2002)、《美国心理协会出版手册》(2010)以及鲍贵(2019),包括四类效度:内部效度、构念效度、统计结论效度和外部效度,涵盖 32 项效度指标,如图 1 所示。

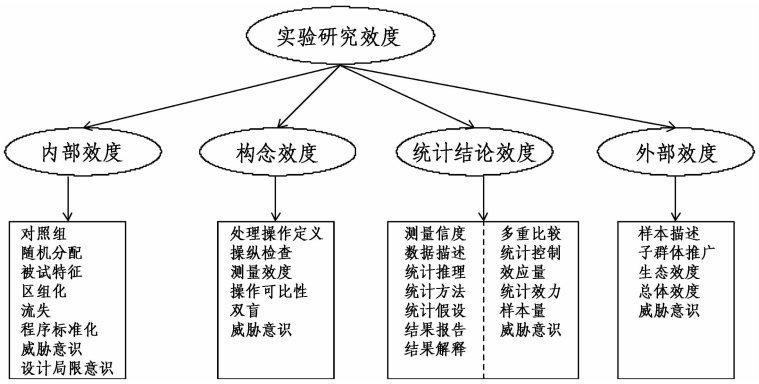

图 1　实验研究效度评价指标体系

图 1 中,内部效度评价指标有八项;构念效度评价指标有六项;统计结论效度评价指标有 13 项;外部效度评价指标有五项。每类效度评价指标体系均包括威胁意识指标。这是因为威胁意识能够体现研究者开展实验研究的能力。由于内部效度与研究设计紧密联系,因而在内部效度指标的选择上尽可能考虑实验设计的特点。譬如,在内部效度指标中设计“对照组”“随机分配”和“设计局限意识”等指标。使用对照组是确定变量之间因果关系的重要控制手段。使用随机分配是为了减少外扰变量对研究变量之间因果关系的干扰。使用“设计局限意识”指标的目的是考察研究者是否能够意识到某个研究设计在内部效度方面的局限性。本次评价使用的“被试特征”指标与 Shadish et al.(2002)列出的内部效度威胁清单中的“被试选择偏差”一致。这一指标用于考察研究者是否在准实验设计中使实验组或发现实验组在一个或多个前测或其他被试特征变量测量上相似,减少选择偏差。为了不使评价指标过于繁琐,本次评价将 Shadish et al.(2002)列出的内部效度威胁清单中的“历史”“成熟”“回归”和“测试”等威胁归入“威胁意识”指标。“程序标准化”指标与 Shadish et al.(2002)提出的“工具变化”威胁一致。Shadish et al.(2002)提出的内部效度威胁框架中的“流失”威胁在本次评价中得以保留,但是“模糊的时序性”威胁未予考虑,因为所有的实验研究都能排除这一威胁。在实际操作中,流失率低于 20% 视作被试流失不严重,否则视作被试流失严重(Bausell,2015)。

在构念效度方面,本研究使用的“构念操作定义”指标和“操作可比性”指标与 Shadish et al.(2002)列出的构念效度威胁清单中的“构念论述不充分”和“构念混淆”分别一致。设计“操纵检查”指标是为了考察研究者是否使用操纵检查或使用类似的方法检验并确保实验处理实施的忠实度(fidelity)。“测量效

度”指标考察研究者是否提供主要因变量测量的效度证据。如果一项研究能够提供效度证据,很大程度上就能够排除 Shadish et al.(2002)提出的“单一方法偏差”威胁。本研究主要评价定量型实验研究,Shadish et al.(2002)提出的“单一操作偏差”威胁未予考虑。将 Shadish et al.(2002)列出的构念效度威胁清单中的“对实验情境的反应性”和“实验者期望”两个威胁归入“双盲”指标。如果一项实验采用双盲技术,这两个威胁基本可以被排除。Shadish et al.(2002)列出的其他构念效度威胁归入“效度威胁意识”指标。

在统计结论效度评价方面,“测量信度”指标反映 Shadish et al.(2002)列出的统计结论效度威胁清单中的“测量无信度”威胁。本研究增加“数据描述”指标(至少包括样本量、平均数、标准差或频数和比率)是为了考察研究者是否较充分地报告描述性统计量。本研究还增加了“统计推理”和“统计方法”两项指标。统计推理是定量研究统计决策的必要手段。“统计方法”指标的重要性是不言而喻的。譬如,如果研究者对两个实验组在二项类别变量数据上分布差异的比较采用独立样本 t 检验,统计结果就没有意义,因为 t 检验使用的平均数不适用于类别变量数据。本研究使用的“统计假设”“多重比较”“效应量”和“统计效力”等指标分别对应于 Shadish et al.(2002)提出的“违背统计检验假设”“捕捉和错误率问题”“不精确的效应量估计”以及“统计效力低”等威胁。Shadish et al.(2002)列出的统计结论效度威胁清单中的“范围限制”“实验场景中的额外方差”和“研究单位的异质性”等威胁是导致统计效力低的主要原因,本研究将这些威胁归入“效度威胁意识”指标。样本量的大小也与统计效力密切相关,因而本研究将“样本量”列为一个效度指标。样本量多大才合适依具体的研究性质而定。为了不使问题复杂化,本次评价依据 Gersten et al.(2000),将每个实验条件下的被试数不少于 20 人作为质量评判的大致标准。此外,本研究统计结论效度评价体系还包括“结果报告”“结果解释”和“统计控制”指标。“结果报告”指标(指结果报告的完整性,如 t 检验报告中至少包括 t 值、正确的自由度和 p 值)同“数据描述”指标一样是应《美国心理协会出版手册》(2010)对研究结果报告的要求。“结果解释”指标体现研究者对重要统计概念正确理解和应用的能力。如果结果解释错了,统计结论就不可信。使用“统计控制”这一指标是为了与内部效度指标中的“被试特征”指标相一致。如果研究者在统计分析中包括了外扰变量,统计结论的信度就会提高。

外部效度评价采用“样本描述”等五项指标。“样本描述”指标包括被试年龄、性别和外语水平。“威胁意识”指标涵盖 Shadish et al.(2002)列出的四种外部效度威胁,即“因果关系和场景的交互作用”“因果关系在处理变体上的交互作用”“因果关系和结果的交互作用”和“依赖于环境的中介作用”。但是,本研究将 Shadish et al.(2002)列出的外部效度威胁清单中的“因果关系和研究单位的交互作用”归入“子群体推广”指标。“生态效度”指实验场景、程序或处理方式等是否自然。“总体效度”指研究样本是否从被试总体中随机抽样得到。

1.3 数据收集

本研究使用的博士论文数据为 2005—2014 年间我国英语语言学博士生的学位论文,检索语料库为中国知网(CNKI)的“中国博士论文全文数据库”。选择检索的学科领域为:哲学与人文科学·外国语言文字·英语。检索词为“experiment”,检索年度为 2005—2014 年。符合初始检索条件的博士论文数为 353 篇。

文中有“实验”一词的博士论文未必就是实验研究,需要对初次检索的论文进行再次筛选。筛选的论文满足以下条件:(1)作者为英语语言学专业博士研究生;(2)以中国语境下的英语学习者为主要研究对象(被试);(3)作者在摘要或在研究方法论中采用术语“实验”“试验”“实验组”“控制组”或“对照组”等中、英文术语表明研究的实验性质,且为定量研究;(4)满足实验研究的基本特征:研究者有意地操纵一个或多个自变量,观察操纵水平的变化对结果变量(因变量)的影响;(5)如果作者在论文中声称开展了多项实验,则以第一个所谓的实验为评价对象。按照以上筛选标准,得到有效博士论文数 104 篇。博士论文的阶段性划分以每两年为一个时段,如 2005—2006 年为一个阶段,共五个阶段。每个阶段博士论文样本量依次为 12、20、25、31 和 16。

1.4 数据标注与统计分析方法

博士论文数据标注的范围是论文的研究方法、结果和结论章节。各类效度指标的标注采用二分法。凡某项指标在论文中得以显示,评价结果就为“是”,计数为“1”,表示在某项指标上有效度。凡某项指标在论文中缺失,评价结果就为“否”,计数为“0”,表示在某项指标上没有效度。譬如,若一项研究使用对照组,评价结果就为“是”,否则评价结果为“否”。一项研究没有被试流失现象,评价结果就为“是”。若流失率超过20%,评价结果则为“否”。

研究问题的回答采用描述性统计和推理统计相结合的方法。比较每类效度显示度的阶段性差异采用秩次型单因素稳健方差分析①。效度显示度定义为同类效度指标上的计数之和与指标题项数的比率。对各类效度指标变化总体特点的探索采用卡方拟合优度检验。每项指标上的效度显示度定义为各个阶段该指标上的计数之和与总样本量的比率。本研究的零假设为论文总体(population)中效度指标显示度可能有三种情形,即 $P_0 = 0.25$、$P_0 = 0.5$ 或 $P_0 = 0.75$。在零假设情况下,$P_0 = 0.25$ 指在总体中某个效度指标的显示度为0.25,缺失度为0.75,表示只有1/4的学位论文在该指标上体现了效度。$P_0 = 0.5$ 和 $P_0 = 0.75$ 的解释与之相似。0.25、0.5和0.75是三个有意义的比率,依次反映低、中、高效度。

各类效度指标阶段性变化特点的探索采用卡方列联表检验。由于分阶段统计中有些单元格观察频数较小,每项效度指标与阶段性关系的检验实际采用卡方置换检验(permutation test)。

2　研究结果

2.1 博士论文实验研究效度阶段性分析

各个阶段博士论文每类效度平均显示度的比较如图2所示。

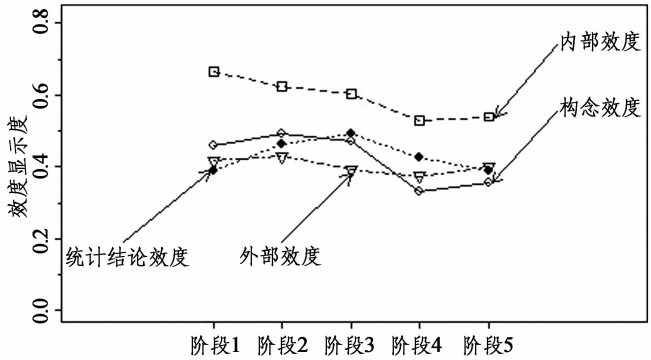

图2　每类效度平均显示度阶段性变化

图2显示,各阶段博士论文内部效度显示度在四类效度中最高,介于0.5－0.7之间,有随阶段缓慢下降的趋势,只是在近期两个阶段基本持平。构念效度显示度在前三个阶段保持较高的水平(介于0.45－0.5之间),后两个阶段处于较低的水平(介于0.3－0.4之间),下降趋势较明显。各个阶段统计结论效度显示度呈前升后降之势,大致介于0.4－0.5之间,最大值位于第三阶段。外部效度显示度阶段性变化不明显,大致维系在0.4的水平。总体上看,在四类效度中,只有内部效度显示度高于0.5的水平。各类效度均有不同程度的阶段性变化,阶段性变化最明显的是构念效度,变化最平缓的当属外部效度。

为进一步了解博士论文中每类效度显示度在不同阶段是否存在统计显著性差异,本研究采用秩次型单因素稳健方差分析,统计结果如表1所示。

表1显示,内部效度和构念效度前三个阶段的相对效应(relative effects)②较明显地大于后两个阶段

① 秩次型单因素稳健方差分析允许方差不齐和等值(tied values),详见 Wilcox(2017)。关于稳健统计的基本概念,见鲍贵(2017)。

② 合并 J 个独立组数据,对之分配秩次,得到各个组数值的秩次 R_{ij}(第 j 组第 i 个秩次)。各个组平均秩次 $\bar{R}_j$ 为:$\bar{R}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} R_{ij}$,其中 n_j 为第 j 组样本量。相对效应 $\hat{q} = \frac{1}{N}(\bar{R}_j - \frac{1}{2})$,其中 N 为各组合并样本量。相对效应反映各组平均秩次的相对大小。

的相对效应,但是只有构念效度显示度在阶段变化上有统计显著性($p<0.05$)。稳健多重比较发现,第二阶段构念效度显示度显著好于第四阶段($\widehat{P}(X<Y)=0.19$,$\widehat{P}(X>Y)=0.61$,$p=.006<0.01$)①。统计结论效度和外部效度测量上的相对效应接近,在阶段变化上没有统计显著性差异($p>0.05$)。

表 1　各阶段效度显示度稳健方差分析

效度	相对效应($\widehat{q}$)					F	分子、分母自由度	p
	阶段 1	阶段 2	阶段 3	阶段 4	阶段 5			
内部效度	0.64	0.56	0.53	0.42	0.43	1.88	3.15,46.72	0.143
构念效度	0.56	0.61	0.59	0.39	0.40	2.92	3.36,52.37	0.037*
统计结论效度	0.44	0.52	0.57	0.50	0.42	0.80	3.39,53.06	0.515
外部效度	0.53	0.54	0.49	0.46	0.51	0.27	3.86,77.56	0.890

2.2 博士论文各类效度指标推理统计分析

为了推断在博士论文总体中各项效度指标的变化模式,排除随机误差的干扰,本研究在 $P_0=0.25$、$P_0=0.5$ 和 $P_0=0.75$ 三种假设情形下,采用卡方拟合优度检验推导各项指标变化模式,统计分析结果如表 2 所示。

表 2　效度指标卡方拟合优度检验

		$\widehat{P}$	P			$\widehat{P}$	P
内部效度	对照组	0.95	$P>0.75$	统计结论效度	测量信度	0.32	$P=0.25$
	随机分配	0.27	$P=0.25$		数据描述	0.76	$P=0.75$
	被试特征	0.84	$P>0.75$		统计推理	0.87	$P>0.75$
	区组化	0.45	$P=0.5$		统计方法	0.73	$P=0.75$
	流失	0.93	$P>0.75$		统计假设	0.13	$P<0.25$
	程序标准化	0.83	$P=0.75$		结果报告	0.39	$0.25<P<0.5$
	威胁意识	0.29	$P=0.25$		结果解释	0.66	$0.5<P<0.75$
	设计局限意识	0.11	$P<0.25$		多重比较	0.54	$P=0.5$
构念效度	构念操作定义	0.30	$P=0.25$		统计控制	0.21	$P=0.25$
	操纵检查	0.86	$P>0.75$		效应量	0.05	$P<0.25$
	测量效度	0.08	$P<0.25$		统计效力	0.01	$P<0.25$
	操作可比性	0.72	$P=0.75$		样本量	0.77	$P=0.75$
	双盲	0	$P<0.25$		威胁意识	0.27	$P=0.25$
	威胁意识	0.54	$P=0.5$	外部效度	样本描述	0.39	$0.25<P<0.5$
					子群体推广	0.28	$P=0.25$
					生态效度	0.77	$P=0.75$
					总体效度	0.03	$P<0.25$
					威胁意识	0.52	$P=0.5$

$\widehat{P}$表示样本中指标显示度;P 表示总体中指标显示度

表 2 显示,内部效度指标变化有两极化趋势。总体中,随机分配、威胁意识和设计局限意识三项指标上的效度显示度很低($P\leqslant0.25$)。区组化指标显示度处于中间水平($P=0.5$)。其他四项内部效度指标上的效度显示度较高($P\geqslant0.75$)。

在构念效度方面,构念操作定义、测量效度和双盲三项指标上的效度显示度很低($P\leqslant0.25$)。操作可比性和操纵检查指标的显示度较好($P\geqslant0.75$),威胁意识显示度达到了中等水平($P=0.5$)。

① 稳健多重比较方法采用 Cliff(1996)介绍的方法。该方法允许方差不齐和等值,是对传统的 Mann-Whitney 检验的改进。本例采用 Hochberg 方法控制族第一类错误率。鉴于阶段性研究样本量较小,配对比较数较多,族错误率设定为 $\alpha=0.1$。$\widehat{P}(X<Y)$ 和 $\widehat{P}(X>Y)$ 分别表示第二阶段任一显示度小于和大于第四阶段显示度的概率。

统计结论效度指标上的效度显示度分布较为分散。有近一半的效度指标(六项指标)显示度很低($P \leqslant 0.25$)。这些指标包括测量信度、统计假设、统计控制、效应量、统计效力和威胁意识。结果报告指标上的显示度处于较低水平($0.25 < P < 0.5$)。结果解释和多重比较指标上的显示度处于中等水平($0.5 \leqslant P < 0.75$)。数据描述等其他四项指标上的显示度较高($P \geqslant 0.75$),说明至少有 3/4 的博士论文体现这些指标测量的效度。

在外部效度方面,只有生态效度指标上的显示度较高($P = 0.75$),威胁意识指标显示度次之($P = 0.5$),其他三项指标上的显示度处于较低或很低的水平($0.25 < P < 0.5$ 或 $P \leqslant 0.25$)。

2.3 博士论文各类效度指标与阶段性之间的关系

虽然 2.1 节只在构念效度上发现阶段性差异,但是这未必意味着构念效度的每项指标均有阶段性差异,也未必意味着其他效度的每项指标均没有阶段性差异。各类效度指标与阶段性关系的卡方置换检验结果如表 3 所示。

表 3　各类效度指标阶段性变化的卡方置换检验

		χ^2	p	$\hat{w}$			χ^2	p	$\hat{w}$
内部效度	对照组	6.90	0.130	0.26	统计结论效度	测量信度	10.23	0.034 *	0.31
	随机分配	4.14	0.397	0.2		数据描述	2.14	0.727	0.14
	被试特征	3.96	0.427	0.2		统计推理	4.18	0.392	0.2
	区组化	3.23	0.530	0.18		统计方法	2.41	0.677	0.15
	流失	5.81	0.190	0.24		统计假设	1.35	0.871	0.11
	程序标准化	9.44	0.047 *	0.3		结果报告	4.31	0.378	0.2
	威胁意识	4.68	0.328	0.21		结果解释	4.10	0.403	0.2
	设计局限意识	8.59	0.067	0.29		多重比较	5.86	0.212	0.24
构念效度	构念操作定义	6.54	0.163	0.25		统计控制	3.87	0.448	0.19
	操纵检查	12.04	0.015 *	0.34		效应量	5.14	0.267	0.22
	测量效度	4.56	0.332	0.21		统计效力	4.24	0.456	0.20
	操作可比性	7.87	0.094	0.28		样本量	10.38	0.032 *	0.32
	双盲	0	1.00	0		威胁意识	3.55	0.483	0.18
	威胁意识	12.18	0.015 *	0.34	外部效度	样本描述	4.24	0.385	0.2
						子群体推广	12.87	0.011 *	0.35
						生态效度	0.94	0.941	0.09
						总体效度	2.46	0.752	0.15
						威胁意识	3.66	0.460	0.19

* 表示在.05 概率水平上有显著关联。$w = 0.1$、0.3 和 0.5 分别表示小、中、大效应(Cohen,1988)

表 3 显示,在 32 项测量指标中,只有六项指标有阶段性变化。在内部效度指标中,只有程序标准化指标与阶段性有显著关联($p < 0.05$),效应量($\hat{w} = 0.3$)达到中等水平。程序标准化指标在第二、第三阶段显示度很高(均在 0.95 以上),其他阶段的显示度在 0.65 – 0.75 之间。在构念效度指标中,只有操纵检查和威胁意识指标与阶段性有显著关联($p < 0.05$),效应量($\hat{w} = 0.34, 0.34$)达到中等水平。操纵检查指标第四阶段的显示度相对较低(0.68),而其他阶段的显示度较高,均处于大约 0.9 的水平。威胁意识指标上的效度显示度在前三个阶段较高(0.6 – 0.8),第四和第五阶段较低(0.3-0.4)。在统计结论效度指标中,只有测量信度和样本量指标与阶段性有显著关联($p < 0.05$),效应量($\hat{w} = 0.31, 0.32$)达到中等水平。测量信度指标在第五阶段的显示度过低(0.06),其他阶段的显示度在 0.23 – 0.48 之间。第四阶段样本量指标上的显示度最高(0.90),第三阶段的显示度过低(0.56),其他阶段的显示度介于 0.75 – 0.88之间。在外部效度指标中,只有子群体推广指标与阶段性有显著关联($p < 0.05$),效应量($\hat{w} = 0.35$)达到中等水平。第一和第二阶段子群体推广指标上的显示度最高(均为 0.5),第四阶段显示度较低(0.25),第三和第五阶段的显示度过低(近似 0.10)。

概而言之,虽然有六个效度指标的显示度与阶段性有关联,但是并未体现随阶段性稳步提升的迹象,甚至出现“不进则退”的现象,譬如在构念效度威胁意识指标上,后期论文的效度显示度不及前期论文。

3 讨论与建议

本研究得出以下主要结论。其一,整体上,在博士论文体现的四类效度中,只有构念效度存在阶段性变化,不过这种变化没有显示阶段性或历时性稳步提升。其二,以三种效度显示度($P_0 = 0.25$、$P_0 = 0.5$ 和 $P_0 = 0.75$)为参照,14 项指标上的效度显示度在 0.25 以下,占指标总数(32 项)的 44%。即是说,这些效度指标上的效度在至少 3/4 的博士论文中没有得到体现。两项指标上的效度显示度在 0.25 – 0.5 之间,占指标总数的 6%。16 项指标上的效度显示度在 0.5 以上。这意味着 16 项指标在一半以上的博士论文中得到体现。其三,绝大部分效度指标(26 项,占指标总数的 81%)与阶段性没有关联。虽有少数效度指标(六项,占指标总数的 19%)与阶段性有关联,但是这些关联没有体现效度的历时性稳步提升。

这些结果表明,博士论文实验研究的质量不容乐观,特别是在经过近 10 年之后依然没有出现质量明显提高的迹象。针对我国英语语言学博士生实验研究论文中普遍存在的主要问题,建议研究生教学和论文指导以效度为抓手,重视实验设计、实验实施和统计分析的三位一体性。

3.1 提高研究生实验设计能力

此次评价发现,博士生对实验设计意识、内部效度威胁以及统计结论效度威胁意识相对薄弱。因此,研究方法论课程的教学应突出研究设计的重要性。

研究设计在整个研究过程中发挥着统领的作用。在实验设计阶段,博士生需明确研究的具体设计形式、研究中的自变量和因变量如何定义和测量、有哪些外扰变量需要通过设计本身、通过实施程序或通过统计程序加以控制。建议方法论课程的教学多开展实验设计案例分析,增强博士生的感性认识,明确一种实验设计形式可能面临哪些效度威胁以及如何排除或降低这些威胁。Shadish et al. (2002) 系统、深刻地论述了实验研究的原理、原则和方法,被尊奉为实验研究的“圣经”。Bausell (2015) 从实用的角度简明扼要地阐述了设计与开展实验的基本原则。推荐将这些著作作为方法论教材或研究生必读书目。

3.2 重视使用双盲技术,加强构念定义与操作之间的联系

本次评价的博士论文在构念操作中几乎没有使用双盲技术。这一方面是由于有些教学实验研究是由研究者本人实施的,或者被试知情,因而双盲技术很难实现。另一方面,很多博士生可能不了解双盲技术的重要性,未能在研究中应用这项技术。双盲技术能够避免实验者效应和被试对实验情境的反应性(如霍桑效应)。由研究助手或其他教师(非研究者本人)实施实验,可以避免实验者效应。如果被试不知情不会对他们造成伤害,则在被试不知情的情况下参与实验就会避免被试对实验情境的反应性威胁。

大多数博士生对构念的操作及其与构念定义之间关系的重视程度明显不足。如果构念的操作不能体现构念的核心要素,或者构念的操作中引入了其他外扰变量,构念效度就会受到威胁。要提高构念效度,既要有明确、合理的操作程序,又要保证实施程序的严谨性。实验正式实施前的先导研究几乎是必不可少的。通过先导研究发现可能出现的外扰变量,并制定有效措施在正式实施中加以控制。建议研究生方法论教学中对构念定义与构念操作之间的联系给予足够的重视,通过案例来提高研究生批判性学术思维的能力。

3.3 提高研究生统计分析能力

本次评价发现,博士生普遍忽视测量信度、统计假设、效应量和统计效力报告。忽视测量信度和统计假设为统计结论的效度画上了问号。报告信度的博士论文数占论文总数的 32%,同吴旭东等(2002)在期刊论文调查中报告的 14% 相比有很大的进步。但是,在当今实证研究重视测量的大背景下,信度报告如此不足还是令人不安的。譬如,Plonsky et al. (2011) 发现,64% 的期刊论文报告了信度估计。当然,期刊论文对统计假设和统计效力的忽略程度也是相当严重的。譬如,在 Plonsky et al. (2011) 的调查中,只有 3% 的研究检验了统计假设,只有 2% 的论文开展了效力分析。在 Plonsky(2013) 的调查中,17% 的研

究检验了统计假设,只有1%的论文开展了效力分析。这说明很多博士论文中存在的问题在期刊论文中同样存在,是普遍性问题。大多数博士论文忽视效应量(报告效应量的论文比率为5%),使研究结论过度依赖统计显著性。相比之下,Plonsky(2014)通过对两个阶段期刊论文的调查发现,效应量报告的比率由前期的3%增至42%,说明效应量的报告越来越受到期刊作者的重视。样本量小是导致统计效力不足的主要原因之一。虽然有不少博士生意识到样本量的重要性,但是他们只将样本量问题与外部效度联系在一起,而没有意识到样本量不足会降低统计效力。

本次评估暴露出来的问题为我们的研究生教学敲响了警钟。长期以来,博士生课程教学不重视统计理论教学或者统计学教学过于强调统计分析的软件操作,未能使博士生真正掌握统计学的基本原理,未能认识到统计假设检验以及效应量等统计量报告的重要性。我们建议在研究生课程设置中增加应用统计学课程,或增加原有应用统计学课程的技术含量,切实提高博士生统计分析的能力。

3.4 研究生导师要重视过程性指导

学位论文写作是一个较长的过程。在这一过程中,除了博士生本人的努力之外,也需要导师的精心指导。

研究设计是实验成败的关键。研究生导师首先要把好设计这一关,最好能够结合研究实际列出问题清单逐一审查博士生论文的研究设计,内容包括研究问题、研究设计的具体形式、设计形式与研究问题的关联性、实验处理的核心要素、测量方法和被试招募等。其次,研究生导师要确保实验程序制定和执行的有效性,最好能列出问题清单,内容包括构念定义与操作的一致性、实验处理的忠实度、实验实施者的能力和测量的信度和效度等。最后,建议研究生导师规范统计分析流程,避免统计分析和认知误区。

4　结语

本文依据效度框架制定了效度指标体系,并借以评价2005—2014年间我国英语语言学方向博士生百篇实验性学位论文方法论的质量。我国博士研究生整体上初步具备开展实验研究的能力,但是也有不少"短板"。譬如,对实验研究设计的意识比较淡薄,对外扰变量的控制能力不强,统计分析与报告能力较弱。这些"短板"为博士生培养方案和课程设置的改革指明了方向。

本研究制定的效度指标体系具有普适性,为研究者开展实验研究评价或审查自身的实验研究问题提供了参考框架。在实际应用中,研究者可以结合具体的研究领域将指标体系进一步细化。另外,本研究没有对效度指标设定不同的权重。毋庸置疑,不同指标体现的难易度是不一样的。譬如,控制外扰变量比样本量报告要难得多,因为外扰变量控制与因果推论息息相关,不仅需要研究者有专业的知识和技能,还要有研究经验,而样本量报告只体现报告的完整性,技术含量低。能否或如何设定效度指标的权重或许是未来评估研究的一个难点。

参考文献:

American Psychological Association. 2010. *Publication Manual of the American Psychological Association* (6th ed.) [M]. Washington, DC: Author.

Bausell, R. B. 2015. *The Design and Conduct of Meaningful Experiments Involving Human Participants: 25 Scientific Principles* [M]. New York: Oxford University Press.

Campbell, D. T. & J. C. Stanley. 1966. *Experimental and Quasi-Experimental Designs for Research* [M]. Chicago: Rand McNally.

Cliff, N. 1996. *Ordinal Methods for Behavioral Data Analysis* [*M*]. Mahwah, NJ: Erlbaum.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.) [M]. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T. D. & D. T. Campbell. 1979. *Quasi-Experimentation: Design & Analysis Issues for Field Settings* [M]. Boston: Houghton Mifflin Company.

Gersten, R., Baker, S. & J. W. Lloyd. 2000. Designing High-quality Research in Special Education: Group Experimental

Design [J]. *The Journal of Special Education*, 34(1): 2-18.
Larson-Hall, J. & L. Plonsky. 2015. Reporting and Interpreting Quantitative Research Findings: What Gets Reported and Recommendations for the Field [J]. *Language Learning*, 65(Suppl. 1): 127-159.
Lindstromberg, S. 2016. Inferential Statistics in Language Teaching Research: A Review and Ways Forward [J]. *Language Teaching Research*, 20(6): 741-768.
Norris, J. M., Plonsky, L., Ross, S. J. & R. Schoonen. 2015. Guidelines for Reporting Quantitative Methods and Results in Primary Research [J]. *Language Learning*, 65(2): 470-476.
Plonsky, L. & S. Gass. 2011. Quantitative Research Methods, Study Quality, and Outcomes: The Case of Interaction Research [J]. *Language Learning*, 61(2): 325-366.
Plonsky, L. & Y.-J. Kim. 2016. Task-based Learner Production: A Substantive and Methodological Review [J]. *Annual Review of Applied Linguistics*(36):73-97.
Plonsky, L. 2013. Study Quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research [J]. *Studies in Second Language Acquisition* (35): 655-687.
Plonsky, L. 2014. Study Quality in Quantitative L2 Research (1990—2010): A Methodological Synthesis and Call for Reform [J]. *Modern Language Journal*, 98(1): 450-470.
Shadish, W. R., Cook, T. D. & D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* [M]. Boston: Houghton Mifflin Company.
Wilcox, R. R. 2017. *Introduction to Robust Estimation and Hypothesis Testing* (4th ed.) [M]. San Diego, CA: Elsevier.
鲍贵.2012.我国外语教学研究中的统计分析方法使用调查 [J].外语界(1):44-51,60.
鲍贵.2015.坎贝尔实验研究效度框架在应用语言学中的应用 [J].外语研究(3):7-12.
鲍贵.2017.应用语言学研究中的图示与稳健统计方法 [J].外国语文(6):135-142.
鲍贵.2019.理解与评价应用语言学实验研究 [M].上海:上海交通大学出版社.
何家宁,张文忠. 2009. 中国英语学生词典使用定量实证研究数据收集与统计方法现状分析 [J].现代外语(1):94-101.
潘珣祎,简庆闽. 2008. 外语类学术期刊论文错失析评 [J]. 外语与外语教学(1):60-64.
吴旭东,张文忠.2002.我国外语教学实验研究质量调查 [J]. 外语教学与研究(1):35-44.
郑新民,王玉山. 2014. 如何在外语教育研究中科学地使用调查法——基于我国外语类 CSSCI 期刊文章(2008—2013 年度)的分析 [J]. 外语电化教学(4):8-13.
郑新民. 2009. 中外应用语言学学位论文写作对比研究 [J]. 外语电化教学(3):72-77.

An Assessment of Quality of Experimental Research in Chinese Doctoral Dissertations of English Linguistics

BAO Gui

Abstract: This article constitutes the first empirical assessment of experimental research quality in Chinese doctoral dissertations of English linguistics, using a system of validity indicators of experimental research. A total of 104 Chinese doctoral dissertations from 2005 to 2014 are surveyed on design features, procedures, statistical analyses and reporting practices. The validity evidence based on 14 out of 32 validity indicators fails to be demonstrated in at least three fourths of the dissertations. Regarding internal validity, there is a general lack of awareness to threats to validity and of limitations in research designs. Inadequate operationalization of constructs and a lack of double-blindness are typical threats to construct validity. Threats to statistical conclusion validity mainly involve serious omissions of reliability of measures, effect sizes, statistical power and tests of statistical assumptions. With respect to external validity, population validity and generalizing across subpopulations are generally neglected. There is no clear trend for the research validity of the dissertations to improve consistently over time.

Key words: doctoral dissertations; experimental research; validity; assessment

责任编校:蒋勇军