英汉版和英语版词汇量测试异同及效度比较

付玉萍，刘振前

（海南热带海洋学院 外国语与国际文化交流学院 海南 三亚 572000；山东大学 外国语学院 山东 济南 250000）

摘 要：我们以 228 名英语专业学生为被试，对 Nation 的词汇量测试卷英汉版和英语版进行了比较研究，并考察了两者的信度和效度。结果显示，英汉版成绩显著高于英语版；两版本均具有较高的内部一致性信度、结构效度和效标效度，但英汉版信度高于英语版（.855 vs .816），前者各等级得分与总分的相关性皆高于后者（介于.333 ~ .795 之间，5K 和 8K 除外），因子分析结果似乎也支持两版本的单维性（即只测量一个潜在变量——词汇知识），三组不同水平被试两版本词汇测试成绩均在.05 的水平上差异显著，且随着语言水平的提高呈递增趋势。随着词频等级的增加，测试成绩大致呈阶梯状下降趋势，英汉版各难度等级成绩均显著高于英语版。两版本（特别是英汉版）词汇量测试卷提供给广大英语教师和研究者一种全新的测量工具，以考量学生书面接受性词汇知识。

关键词：词汇量；词汇测试；英汉版；英语版；效度检验

中图分类号：H319 **文献标识码**：A **文章编号**：1673-9876（2018）01-0064-06

Abstract: The primary purpose of this paper is to provide further evidence to previous findings that the bilingual version of Nation's vocabulary size test (VST) can be more efficient than the monolingual one. It conducts a comparison between English-Chinese (E-C) and English-English (E-E) VST based on data from 228 English majors. Statistic results show that scores of E-C version are significantly higher than those of E-E one, which reflects a higher level of internal consistency reliability, structural and criterion-related validity. The correlation between scores of each frequency level and total scores of the former is also higher than those of the latter. The results of factor analyses reveal that single construct, presumably word knowledge, is underlying the two version tests. A one-way between-subject ANOVA also indicates that the two version tests can effectively distinguish English learners of different proficiency levels (significantly different at .05 level). As word frequency increases, participants' scores roughly drop in a stair-step pattern though clusters of 1,000-word frequency levels provide a relatively meaningful difficulty order, which is less susceptible to the idiosyncrasies at each 1,000 level. The two versions, especially E-C version, offers English teachers and researchers a new valid instrument to measure written receptive vocabulary size, if with some minor revision of the items in question.

Key words: vocabulary size; vocabulary test; English-Chinese version; English-English version; test validation

DOI:10.16362/j.cnki.cn61-1457/h.2018.01.013

1. 引言

词汇量测试历来受语言学家关注，因其结果不仅可用来判断学生词汇习得状况，而且可用来指导教学、课程设计与教材编写，同时还是进行词汇习得研究的重要工具。国内词汇测试的相关研究虽然已有很多（桂诗春 1985；陈晓扣 2000；马广惠 2001 等），但运用现代化统计和测量手段进行量化分析的实证研究，特别是效度研究并不多见（徐柳明、刘振前 2013：79-85）。国外最有代表性的标准词汇量测试工具当属 Nation 的词汇水平测试（vocabulary levels test，简称 VLT，Meara 1996：38），这可能是 20 年来词汇习得研究的最大进步（Meara & Alcoy 2010：222）。该测试以词汇概貌的形式粗略估计学习者接受性词汇量，因其操作简单、容易理解而在世界各地得以广泛使用，并成为标准的词汇量测试工具。然而，Nation（2001：21-22）指出：VLT 是一种诊断性测试，能帮助教师快速了解学生需要集中学习的是高频词还是低频词，但不是综合评估词汇量的工具。为提供一个准确、可靠、全面的测试，测量学习者从首个 1,000 词族起至第 14 个 1,000 词族的词汇量，Nation & Beglar（2007：9-13）合作研发出词汇量测试卷（vocabulary size test，简称 VST）。我们尝试对其英汉版和英语版词汇量差异及信度、效度进行比较，以验证其是否适合中国英语学习者。

2. 英语版和双语版 VST

2.1 英语版 VST

VST 用于测量书面接受性词汇知识，选词来源基于英国国家语料库（BNC）确定的 14,000 词族表。该表按词频高低分为 14 级，每级 1,000 词族。VST 从每级中随

机抽取 10 个题目，组成包括 14 级共 140 题的测试卷，要求被试从每题 4 个简短释义中选取正确答案，释义用词选自通用词表（West 1953）中最常用的前 2,000 词。

VST 的测试词都置于简短的非定义式语境中，目的在于反映测试词的最常见意义和用法，提供测试词的词性信息，为识别任务创设更加有效的语言环境。下面是选自 3 级的例子：

3. scrub: He is **scrubbing** it.

a. cutting shallow lines into it

b. repairing it

c. rubbing it hard to clean it

d. drawing simple pictures of it

VST 采用多项选择题的形式，有很多优点，如利于控制题目难度、简化操作和评分程序、消除评分偏见、易于在线测试[①]等。VST 干扰项与正确答案属于同一词性，其评分方法是计算 14 个等级的所有正确答案数量，然后乘以 100。如果被试认识所有测试词，则得分为最高分 14,000，表明被试词汇量至少为 14,000 词族。Beglar（2010：101-118）的研究初步显示，VST 具有较高的信度和心理测量单维性，是测量书面接受性词汇的有效工具，可用来对词汇水平各异的被试进行大范围测试。

2.2 双语版 VST

VST 可以用来测试英语母语者和学习者的词汇量，测试结果可用来指导课程设计、阅读与词汇教学和教材开发等。然而，为选择有限的常用词进行恰当表达，有些选项难免过于复杂而冗长。下面是另一选自 8 级的例子：

8. octopus: They saw an **octopus.**

a. a large bird that hunts at night

b. a ship that can go under water

c. a machine that flies by means of turning blades

d. a sea creature with eight legs

该例选项释义涉及 that 关系从句和介词短语做后置定语的情况。因此，如果被试缺乏一定的语法知识和阅读技巧，即使了解单词的概念意义，也可能会误解某些选项释义。这有违设计者仅测量词汇知识而并非其他维度的初衷。为此，VST 开发了包括日语、汉语、越语和俄语在内的多个双语版本，这些都可在网上免费下载[②]。下面是上述英语版同一测试词 octopus 的英汉版举例：

8. octopus: They saw an **octopus.**

a. 猫头鹰

b. 潜水艇

c. 飞机

d. 章鱼

这种英汉版测试仍然是要求较高甚至更高的词汇知识测试（难以测量被试部分词汇知识的掌握情况），但避免了复杂的语法释义，并减少了对阅读技巧的要求，表达更简洁明了。被试理解选项的难度减小，焦虑感降低，从而更乐于尝试和配合。与英语版相比，双语版 VST 的非词汇因素和语言障碍减少了，应能更准确地测量被试的词汇量。因此，被试（特别是低水平被试）双语版测试成绩应高于英语版。

Nguyen & Nation（2011：86-99）以越南大学生为对象，对英语—越语版 VST 进行了研究。随后，Karami（2012：53-67）和 Elgort（2013：253-272）分别编制了英语—波斯语版和英语—俄语版 VST，并进行了效度检验。三项研究结果皆表明，双语版具有较高的信度和效度；作者均推荐编制并使用双语版进行词汇量调查。

我们将参考以往研究编制出英汉版 VST，并用来调查中等水平学习者英汉版和英语版 VST 的词汇量差异，同时检验两版本的信度和效度，重点回答以下 3 个问题：

1）两版本所测词汇量是否存在显著差异？

2）两版本是否皆有较高的信度和效度？

3）两版本的项目难度如何？

测试的信度和效度从内部一致性、结构效度和效标效度三方面探讨。

3. 研究方法

3.1 英汉版 VST 的翻译校正

首先，作者仔细研究并核对了网上下载的英汉版 VST 的汉语选项，发现很多地方表达并不准确。为确保测试信度和效度，作者和 3 位汉语母语者（都有英语语言学博士学位，且从事英语教学多年）就前人的翻译原则（Nguyen & Nation 2011；Karami 2012；Elgort 2013）进行了充分讨论和归纳总结，并制定了下列翻译原则：

第一，尽可能用汉语对应词、短语或同义词，而非释义进行翻译；

第二，选项都和测试词的词性一致；

第三，如果测试词是借词，就不用汉语翻译对应词（即音译词）而采用简短释义翻译。例如，测试词 caffeine 的汉语翻译对应词是咖啡因，这是 caffeine 的音译词，因此要译成一种让人兴奋的物质而非咖啡因；

第四，为避免被试基于选项长度和/或句法复杂性进行猜测，当选项中的汉语对应词或简短释义数量不对称时，将一个对应词译成释义。

总之，在选项的翻译校正方面优先考虑把所有选项都译成汉语对应词（见例 1），或汉语简短释义（见例 2），或两项汉语对应词，两项汉语简短释义（见例 3）；

其次，作者严格遵照上述原则，结合权威的英汉词典，对最初的英汉版选项汉语表达进行了修改或重译。之后，交由 3 位博士校对，负责检查汉译质量和翻译原则的执行情况；

最后，作者和 3 位博士对一些翻译不一致的选项进行了充分讨论，消除了所有歧义并达成了共识。

例(1)

6. **cranny**: We found it in the **cranny**!	6. **cranny**: We found it in the **cranny**!
a. sale of unwanted objects	a. 积压货
b. narrow opening	b. 裂缝
c. space for storing things under the roof of a house	c. 阁楼
d. large wooden box	d. 大木箱

例(2)

4. **augur**: It **augured** well.	4. **augur**: It **augured** well.
a. promised good things for the future	a. 预示未来的好事情
b. agreed well with what was expected	b. 和预料的很吻合
c. had a colour that looked good with something else	c. 很容易配色的东西
d. rang with a clear, beautiful sound	d. 发出悦耳的声音

例(3)

4. **counterclaim**: They made a **counterclaim**.	4. **counterclaim**: They made a **counterclaim**.
a. a demand made by one side in a law case	a. 反诉
b. a request for a shop to take back things with faults	b. 要求商店收回有瑕疵的东西
c. An agreement between two companies to exchange work	c. 两个公司之间交换工作的合同
d. a top cover for a bed	d. 被子

3.2 试测

为了解翻译校正后的英汉版是否还有歧义,作者于2014年11月以山东某市公立中学高三、同一城市综合性大学英语专业大一和大三共计236名学生为被试进行了试测。之后对参与试测的部分被试进行了非正式访谈,修改了个别翻译不到位或存在歧义的题目。

3.3 研究对象

海南省某综合性大学英语专业大二和大三9个自然班的271名学生参与了正式测试,剔除缺失数据后,实际人数为228人。其中男生25人,女生203人,被试平均年龄为21.19岁,学习英语的最短时间为6年,最长为15年,平均10.12年。根据高校英语专业教学大纲判断,这些被试应属于中等水平学习者。

3.4 测试材料

我们运用两个测试材料获取数据。一是以Nation & Beglar的VST为词汇量测试工具。为保证测试有效性,英汉/英语版测试卷均由前10级的题目构成,因为试测结果发现,被试面对大量不熟悉单词时极易猜测,得分较低;且英语专业四级考试大纲规定的词汇量不过6000单词。为避免反拨效应,英汉版题目顺序进行如下调整:每个等级的前5个题目与后5个题目对调,但等级顺序不变。二是以全国高校英语专业四级考试(TEM-4)为语言水平测量工具对被试进行水平分组。

3.5 数据收集与分析 K

2016年3月所有被试完成了英汉版和英语版VST测试。测试开始前,作者向任课教师讲明测试目的和注意事项,并和任课教师一起参与了整个过程,确保测试在没有查阅词典、参考资料和讨论的环境下进行。测试过程中提醒学生不认识的单词不要盲目猜测,并由任课教师告知,测试成绩将作为期末成绩的一部分,以保证其积极参与。英语版测试用时(包括汉语说明)大概45分钟,英汉版大概35分钟,两次测试前后相隔1周。所有任课教师和学生都非常配合,测试进展顺利。测试卷由作者人工评阅,结果输入EXCEL 2007进行初步整理。

2016年4月所有被试参加了TEM-4,我们以该考试成绩为参照,把被试分为低(79人)、中(76人)、高(73人)三个语言水平组。

运用EXCEL 2007和ISM SPSS 20.0软件对数据进行如下定量分析:1)运用描述性统计分析被试总体词汇量,并运用t-检验比较两版本所测成绩;2)运用同质性检验分析测试的内部一致性信度;3)运用相关分析、因子分析和单因素方差分析考察测试的结构效度和效标效度;4)运用项目难度分析找出影响信度和效度的题目。

4. 结果和分析

4.1 描述性统计和 t 检验

表1是228名被试词汇测试成绩的描述性统计结果。数据显示,被试100项测试(英汉版和英语版)的均值分别为56.49和42.10,标准差和分数跨度较大。这主要是因为被试来自全国各地,中学英语教学要求差别较大,专四成绩也存在很大差别,导致测试得分差异明显。原始分数均值乘以100,得到英汉版平均词汇量为5649词族,英语版平均词汇量为4210词族,相差1439。配对样本 t 检验结果显示,英汉版成绩显著高于英语版:$t(227)=23.535$, $p<.001$。

表1. 词汇测试成绩统计

测试版本	*Min*	*Max*	*M*	*SD*	平均词汇量	*t* 值	*Sig.*(双侧)
英汉版	26	91	56.49	12.471	5649	21.947	.000
英语版	16	71	42.10	11.018	4210		

有研究认为,识别单词的二语(英语)释义比一语释义需要更强词汇能力和更高语言水平(Laufer & Goldstein 2004: 423)。心理测量学也指出,测试题目的表达应采用被试最擅长的语言(Kaplan & Saccuzzo 2007: 188),以减少语言障碍,降低被试焦虑水平。因此,与英语版相比,英汉版题目更适合中等水平学习者,测试成绩当然更高。

4.2 测试信度分析

判断测试优劣的首要条件是其结果稳定性,即测试信度。我们运用经典测试理论的同质性检验方法

(Cronbach α 信度系数)分析其内部一致性,即各等级间是否保持大致的稳定性。228 名被试均完成了包括 100 个题目的英汉版和英语版测试,两份试卷的信度分别检验。结果显示(见表 2),两版本内部一致性信度均较高,其中英汉版 α 值介于.728 ~ .855 之间,英语版 α 值介于.715 ~ .816 之间。无论是各级系数还是总系数,英汉版皆略高于英语版(10K 除外),表明前者内部一致性优于后者。

表 2. 各词频等级信度分析(α 值)

版本	总系数	1K	2K	3K	4K	5K	6K	7K	8K	9K	10K
英汉版	.855	.767	.756	.742	.744	.738	.729	.728	.732	.739	.730
英语版	.816	.751	.744	.737	.731	.720	.727	.719	.715	.733	.735

4.3 测试效度分析

4.3.1 结构效度

结构效度分析可考察测试内容的单维性,即测试各部分只测量单一维度的内容。我们采用相关分析和因子分析检验测试的单维性。Pearson 相关分析结果显示(见表 3),两版本各等级得分与总分都在.05 的水平上显著相关。除英汉版 1 级得分与总分低度相关外($r = .161$),其他各级 r 值均在.333 ~ .795 之间,呈中高度相关,说明两版本各级线性相关,内部一致性较高,只有一个潜在构念。英汉版各级 r 值皆高于英语版(5K 和 8K 除外),表明前者内部结构一致性高于后者。

表 3. 各等级得分与总分的相关分析(r 值)

版本	1K	2K	3K	4K	5K	6K	7K	8K	9K	10K
英汉版	.161*	.471**	.667**	.654**	.693**	.795**	.781**	.744**	.713**	.716**
英语版	.333**	.451**	.556**	.598**	.712**	.690**	.754**	.759**	.618**	.556**

注: $^{**}p < .01$; $^{*}p < .05$

表 4. 测试成绩总体解释方差

	英汉版			英语版		
因子	特征值	方差百分比	累计百分比	特征值	方差百分比	累计百分比
1	4.429	44.289	44.289	3.811	38.114	38.114
2	1.17	11.701	55.989	1.729	17.293	55.407

表 5. 旋转后的因子矩阵

英汉版			英语版		
词频等级	因子一	因子二	词频等级	因子一	因子二
7K	.814		9K	.842	
10K	.803		10K	.805	
8K	.797		8K	.780	.295
9K	.792		7K	.667	.422
6K	.705	.383	2K		.727
2K	.203	.695	5K	.291	.722
4K	.441	.622	4K	.198	.626
5K	.515	.546	1K	-.154	.623
3K	.493	.537	3K	.210	.570
1K		.401	6K	.493	.513

为进一步了解测试的内部结构,我们进行了探索性因子分析。KMO 测度值为:英汉版.886,英语版.837,两版本 Bartlett 球体检验显著水平 $p < .001$,表明数据符合因子分析的要求。全部数据进行主成分分析,并经四次最大正交旋转,结果见表 4 和表 5。特征值大于 1 的因子英汉/英语版各 2 个,累计方差贡献率分别为 55.99% 和 55.41%。碎石图也显示,前面 2 个点的高度明显陡峭,形成碎石坡,坡后各点形成的曲线较平缓,最后形成一条直线(见图 1)。

2 个因子有如下两种解释:第一,因子 1 是词汇知识因素,因子 2 是区分不同词频等级的难度因素;第二,因子 1 代表低频词知识,因子 2 代表高频词知识(Schmitt et. al. 2001: 70)。两种解释似乎都支持英汉/英语版的单维性,即主要测量一个潜在变量——词汇知识。

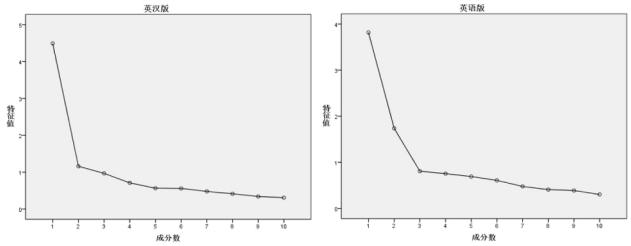

图 1. 英汉/英语版碎石图

4.3.2 效标效度

Bachman 指出,收集测试效标效度的常用方法是比较不同语言水平的学习者数据(1990: 248)。词汇能力是语言能力必不可少的部分,能否区分不同语言水平的学习者是检验词汇测试效度的有效手段,高水平学习者认识更多单词(Laufer & Nation 1999: 38)。我们以 TEM-4 成绩为参照,把被试分为低、中、高三个水平组,探讨其英汉/英语版测试能否区分不同语言水平的学习者。

表 6. 三组被试成绩描述统计

测试版本	组别	N	Min	Max	M	SD	平均词汇量
VST 英汉版	低水平	79	26	82	49.11	10.247	4911
	中等水平	76	36	86	56.66	10.973	5666
	高水平	73	35	91	64.30	11.383	6430
VST 英语版	低水平	79	16	61	35.33	9.141	3533
	中等水平	76	20	58	41.20	8.076	4120
	高水平	73	27	71	50.37	10.185	5037

描述性统计结果显示(见表 6):高水平被试英汉/英语版成绩明显超过中/低水平被试,中等水平被试成绩超过低水平被试。原假设成立,即英汉/英语版词汇测试能够区分不同语言水平的学习者。单因素方差分析结果显示(见表 7),三组被试英汉/英语版成绩均在.05 的水平上差异显著,且随着语言水平的提高呈递增趋势。Scheffe 事后检验还发现,高水平被试英汉/英语版成绩显著高于中/低水平被试,中等水平被试成绩显著高于低水平被试。

表 7. 单因素组间方差分析摘要表

		平方和	自由度	均方	检验	事后比较
		SS	*df*	*MS*	*F*	Post Hoc (Scheffe)
VST 英汉版	组间	8754.533	2	4377.266	37.095*	高 > 中 > 低
	组内	26550.450	225	118.002		
	总数	35304.982	227			
VST 英语版	组间	8676.184	2	4338.092	51.703*	高 > 中 > 低
	组内	18878.496	225	83.904		
	总数	27554.680	227			

4.4 项目难度分析

词汇量测试开发中有个普遍假设——不同词频等级的题目难度会组成一个连续体(Nation 2006; Read 2000)。如该假设成立,被试成绩会逐级递减;另外,由于英汉版成绩显著高于英语版(见 4.1),前者各级成绩也应高于后者。表 8 和图 2 显示,被试均值大致呈逐级递减的趋势;两版本各等级的英汉版均值基本高于英语版,但也有个别例外。

表 8. 各测试卷各等级的均值统计

版本	1*K*	2*K*	3*K*	4*K*	5*K*	6*K*	7*K*	8*K*	9*K*	10*K*
英汉版	9.18	8.16	6.62	6.73	5.09	4.00	4.98	4.51	2.59	4.64
英语版	9.07	6.57	5.90	5.23	3.49	2.96	2.79	2.80	1.52	1.76

图 2 显示,各等级均值分布概貌并非完全逐级递减,这与前人测试结果相似(Nguyen & Nation 2011; Karami 2012),但与我们的假设——词频越低,题目越难,测试成绩越低——不完全相符。如英汉版 10K 均值比 9K 高近一半。显然,在决定难度顺序方面,除词频外还有其他因素也影响了难度顺序。为此,作者分别计算了英汉/英语版 100 题的难度系数,附录 2 列出了 5K 前难度小于.3 的 8 个较难题目和 5K 后难度大于.7 的 3 个较易题目。在较难题目中,5K 有 4 个,4K 有 2 个,3K 和 2K 各 1 个。在较易题目中,6K、7K 和 8K 共 3 个。

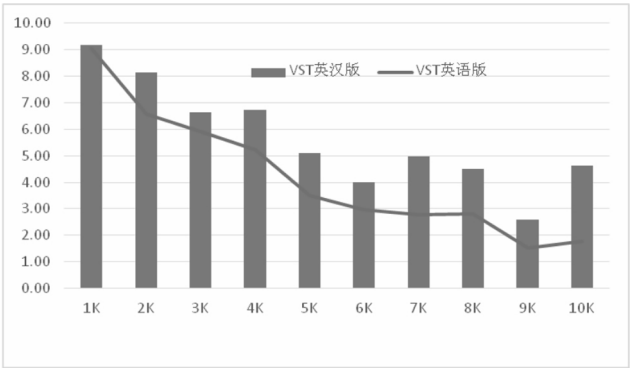

图 2. 两版本各等级的均值分布

为了解上述测试词为何较难/易,我们查阅了高中英语课程标准词表、大学英语课程教学要求词表和英语专业四、八级考试大纲词表,发现较易题目中 1 个测试词是高中词汇(kindergarten),2 个英语专业四级词汇(thesis 和 olive,其中 thesis 为英语专业学生论文写作必用的学术英语词汇,olive 是专四高频词)。至于较难题目,有 3 个测试词是专四词汇(allege, haunt 和 nun,均为专四低频词),2 个专八词汇(rove 和 fracture),1 个八级外词汇(compost)。只有 nil 属于大学英语较高要求词汇,但笔者查阅大学英语教材(大学体验英语)却没有发现该词。大纲规定,四级是英语专业基础阶段对学生的基本要求,八级是英语专业高级阶段的要求。228 名被试虽然都是英语专业大二和大三学生,但其专四成绩均值仅为 52 分,这些题目对多数被试来说较难,也就不足为奇了。另外也说明单纯从 BNC 词频表中选取测试词也许并不完全适合中国学生,建议今后的词汇量调查对该测试卷进行微调,把个别测试词替换为适合我国学生的单词。

表 9. 两版本试卷的难度差异

难度等级	1 级		2 级		3 级		*F*	重复对比
英汉版	*M*	*SD*	*M*	*SD*	*M*	*SD*	(1.9, 424.1)	1 级 > 2 级 > 3 级
	23.96	2.772	15.82	4.740	12.53	5.317	837.94**	
英语版	M	SD	M	SD	M	SD	(1.9, 438.6)	1 级 > 2 级 > 3 级
	21.539	3.115	11.689	4.663	6.655	4.619	1244.865**	
配对样本 t 检验					1 级-1 级		2 级-2 级	3 级-3 级
英汉版-英语版			t(227)		13.014		15.083	18.085

注: ** $p < .001$

以上分析表明,被试的构词法知识、英语水平、BNC 词频表与我国英语教学大纲词表的不一致性等因素,掩盖了测试成绩逐级递减的趋势,即测试成绩会随词频等级的增加而逐步降低,但总体趋势应如此。为此,我们把 10 个词频等级合并为 3 个难度等级:1,000-3,000 词频为第 1 级,4,000-6,000 词频为第 2 级,7,000-10,000 词频为第 3 级。第 3 级有 4 个水平,为便于比较,我们重新换算了该等级分数。两版本试卷 3 个难度等级的统计结果显示(见表 9):测试成绩随难度等级的增加而逐步下降;单因素组内方差分析表明,难度等级对测试成绩有显著影响;多重比较分析说明,被试较低级均值皆显著高于较高等级:1 级 > 2 级 > 3 级。由此可得出结论:被试成绩随难度等级的增加而呈阶梯状下降趋势。配对样本 t-检验结果显示,各级成绩英汉版均显著高于英语版。

根据 Messick(1995: 745)的观点,结构效度的真实性既包括测试反应一致性的理论阐述,又包含考生答题情况的实证证据。数据分析显示,本测试假设的难度顺序与被试答题情况并非完全一致。难度分析表明,这种情况也许与被试的词汇知识、英语水平及学习背景等因素有关。10 个词频等级合并为 3 个难度等级后发现,被试成绩随难度等级的增加而明显下降,难度等级对测试成绩有显著影响,且各难度等级间皆有显著差异。显然,结构效度的真实性也获得支持。以难度等级而非词频等级为基础解释测试结果也许更明智,因为这种难度等级不易受词频等级特性的影响。

5. 结论

本文通过描述性统计分析、同质性检验分析、Pearson 相关分析、探索性因子分析、单因素方差分析和项目难度分析,比较了英汉版和英语版 VST 的词汇量差异及

其信度和效度。多层面的证据均表明英汉版的信度和效度高于英语版，因此前者具有一定的优势；中等水平学习者的英汉版词汇量可能比英语版大 10% 左右，英汉版测试能更准确地估计其词汇知识广度。作为测试学习者英语能力的工具，英汉版能更准确地测试单词的概念意义，可用来测评中国学习者的英语词汇量。课程标准设计者也可用该测试分析学生的词汇需求，因为如不清楚地了解学生的词汇量，将难以帮助他们在词汇学习中做出最佳选择。本研究强调了设计和使用双语词汇量测试需考虑的问题，重视这些问题并了解测试语言、被试水平和词汇频率等因素对词汇量的影响方式，将有益于教师、测试开发者和实施者选择适合其语言学习者群体的 VST 版本和构成。

本文在一定程度上弥补了国内词汇测量工具研究之不足，有望为词汇量相关研究提供重要参考，从而提高研究结果的可信度，让词汇测试更好地为教学服务。当然，本文也存在局限性。例如，VST 计算词汇量的原始方法（被试词汇量 = 100 题得分 × 100）是否可靠？计算结果是否真正代表被试绝对词汇量？是否需要不同方法计算词汇量并解释测试结果？这些问题本研究没有涉及。另外，由于被试样本不够大，代表性不够强，本研究结论的广泛性有待进一步检验。

注释：

① 在线互动版网址：http://my.vocabularysize.com/

② www.victoria.ac.nz/lals/about/staff/paul-nation

参考文献

[1] Bachman, L. F. *Fundamental Considerations in Language Testing* [M]. Oxford: Oxford University Press, 1990.

[2] Beglar, D. A. A Rasch-based validation of the vocabulary size test [J]. *Language Testing*, 2010, 27(1): 101-118.

[3] Elgort, I. Effects of L1 definitions and cognate status of test items on the vocabulary size test [J]. *Language Testing*, 2013, 30(2): 253-272.

[4] Kaplan, R. M. & D. P. Saccuzzo. *Psychological Testing: Principles, Applications and Issues* [M]. Beijing: Beijing World Publishing Corporation, 2007.

[5] Karami, H. The development and validation of a bilingual version of the vocabulary size test [J]. *RELC Journal*, 2012, 43(1): 53-67.

[6] Laufer, B. & Z. Goldstein. Testing vocabulary knowledge: Size, strength, and computer adaptiveness [J]. *Language Learning*, 2004, 54(3): 399-436.

[7] Laufer, B. & I. S. P. Nation. A vocabulary-size test of controlled productive ability [J]. *Language Testing*, 1999, 16(1): 36-55.

[8] Meara, P. The dimensions of lexical competence [A]. In G. Brown, K. Malmkjaer & J. Williams (eds.). *Performance and Competence in Second Language Acquisition* [C]. Cambridge: Cambridge University Press, 1996: 35-53.

[9] Meara, P. & J. Alcoy. Words as species: An alternative approach to estimating productive vocabulary size [J]. *Reading in a Foreign Language*, 2010, 22(1): 222-236.

[10] Messick, S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning [J]. *American psychologist*, 1995, 50(9): 741-749.

[11] Nation, I. S. P. *Learning Vocabulary in Another Language* [M]. Cambridge: Cambridge University Press, 2001.

[12] Nation, I. S. P. How large a vocabulary is needed for reading and listening? [J]. *Canadian Modern Language Review*, 2006, 63(1): 59-81.

[13] Nation, I. S. P. & D. Beglar. A vocabulary size test [J]. *The Language Teacher*, 2007, 31(7): 9-13.

[14] Nguyen, L. T. C. & I. S. P. Nation. A bilingual vocabulary size test of English for Vietnamese learners [J]. *RELC Journal*, 2011, 42(1): 86-99.

[15] Read, J. *Assessing Vocabulary* [M]. Cambridge: Cambridge University Press, 2000.

[16] Schmitt, N., D. Schmitt & C. Clapham. Developing and exploring the behaviour of two new versions of the vocabulary levels test [J]. *Language Testing*, 2001, 18(1): 55-88.

[17] West, M. *A General Service List of English Words* [M]. London: Longman, Green, 1953.

[18] 陈晓扣，郑庆珠. 论英语词汇测试中的高低、一致原则 [J]. 解放军外国语学院学报，2000(6): 64-66.

[19] 桂诗春. 我国英语专业学生英语词汇量的调查和分析 [J]. 现代外语，1985(1): 1-6.

[20] 马广惠. 理工科大学生英语词汇水平研究 [J]. 外语教学，2001(2): 48-52.

[21] 徐柳明，刘振前. 英语词汇量测试卷的编制及其信度与效度检验 [J]. 外语教学理论与实践，2013(1): 79-85.

基金项目：本文系 2017 海南省哲学社会科学规划项目“南海问题中外新闻语篇批评话语分析研究”（项目编号：HNSK(YB)17-29）、2017 年海南省高校科研项目“社会心理视角下海南少数民族地区学生英语学习负动机与对策研究”（项目编号：Hnky2017-44）、2016 年海南省教育科学“十三五”规划项目“微信环境下大学英语移动教学模式探索与实践”（项目编号：QJY13516033）和 2016 年海南热带海洋学院教改项目（项目编号：RDJGb2016-54）的部分研究成果，同时也是海南省应用外语研究基地资助项目。

作者简介：付玉萍，海南热带海洋学院外国语与国际文化交流学院副教授，研究方向：语言测试、二语习得。

刘振前，山东大学外国语学院教授，博士生导师，研究方向：二语习得。

责任编校　郑　荣