语言测试的能力结构与因子分析法

（An exploratory study of language ability construct and factor analysis）

朱正才

（Zhu Zhengcai）

（上海交通大学 外国语学院 上海 200230）

摘要：本文首先探讨了语言测试效度研究中的一个关键问题——语言能力是什么？可分不可分、如何分？接着还探讨了 3 种因子分析方法在分析一份语言测试试卷的能力结构时的运用，提出：用语言能力聚合度，辨别度和拟合度来共同描述一份试卷的语言能力结构特征。实证研究部分表明，这 3 个指标确实能很好地刻画大学英语四级考卷的语言能力结构特征。

关键词：语言能力； 构念； 因子分析

Abstract： This paper starts with a discussion about a key problem in language test validation study: how to define language ability? Whether language ability consists of different components? If so, how to define them? As a research background for the present study, this paper explores the three models of factor analysis and their implications in the defining of language ability. This paper has put forward that degree of convergence, degree of discrimination and degree of fitness can be used to characterize the construct measured by a language test paper. Findings of empirical study show that the three parameters based on the three different factor analysis models very well describe the ability construct of CET4.

Key words: Language ability; Construct; Factor analysis

前言

任何一个语言测试项目的开发都是以某种目标语言能力假设为前提的，现在，大家都习惯了把这些假设统称为构念（Construct），可以说，构念的最后形成和语言学理论的新进展、语言教学现状以及社会需求驱动都有关系。考试效度研究就是要用考试的实测数据来验证这些构念有没有被测量到。这种复杂而漫长的效度验证工作光靠人的主观判断显然不行，于是，大量的统计分析方法就被引入到了语言测试的效度研究当中，其中，因子分析法可谓是最重要的方法之一——至今，它仍是分析一份试卷的能力结构模型的最成功的手段。本文就是想对这一问题做一番更深入的探讨。

1. 语言能力结构和 3 种因子分析法

1.1 语言能力可分不可分

何为语言能力呢？它和人类智力是什么关系呢？它是一个不可分割的综合体还是一个可解析的心理结构呢？

20 世纪初，心理学家们在使用因子分析法（Exploratory Factor Analysis，EFA）研究人类智力问题时，英国的 Spearman（1904）最先主张存在一种“一般能力”（即智力）。他认为，每一种人类活动的顺利完成都受到这种一般能力的影响。随后，有很多人对此提出异议，认为所谓智力，包含多种相互独立的能力因素。到 1927 年，Spearman 也承认了多因素的作用。于是，有关智力的 7 因素说、9 因素说，以至于 120 因素说等纷纷出笼。大量研究现已证实，每一种职业擅长的人确实都有与其相应的多因素能力剖析图，“多元智力观”得以确立（Gardner 2006）。

语言能力一直被认为是多元智力的组成部分之一。脑科学家和心理学家也一直都在探索语言能力的心理和生理机制。与人类智力研究类似，语言测试界围绕有没有单一的语言能力，语言能力是否可以进一步细分下去，早就展开过激烈的争论，而且这些争论与因子分析法有着密切联系。

最早提出单一的语言能力说的是 Oller（1983），他也是在使用统计学的主成分分析法（Principal Component Analysis，PCA）分析语言测试中各个组成部分的结构时，得到了一

个“主要因子”，并称之为 g 因子。后来大家也对此提出异议，说若使用可以旋转因子的因子分析法，对用主成分分析法处理过的数据重新分析，就会得出了好几个因子。Vollmer（1983）就指出：主成分分析法作为语言能力的分析方法特别不适合。……因为它倾向于过高的估计第一个因子的作用。 Oller 自己也做过重新分析，最后坦承，“关于语言能力中有一个无所不包的普遍因子的想法是错误的。”由此，“语言能力的可分立说”渐占上风。

笔者特别注意到，对于这种争执，英国统计学家兼语言学家 Woods（1986）给出的解释是：主成分分析法是一种描写性技术，它没有什么假设，而仅对数据提供另一种视角。而因子分析法却要对数据提出一个模型作为假设，即假设任何一个测量变量都是共性和个性的统一，人们因此可以对变量的总方差进行“公共方差”和“独特方差”的分解。Woods 还以 70 名香港考生考英国剑桥水平考试的数据为样本进行了实证研究，并且也把第一主成分解释为“一般英语语言能力”，并猜测：它可能与一个人的语言学习潜能和智力因素有关。

当代著名的语言测试专家 Bachman（2005）在他的学术专著 *Statistical Analyses for Language Assessment* 中报告了一个类似的研究，他也保留语言测试能力结构中有一个“一般语言能力”因子的构想。

看来，在这一问题上，语言测试专家看法上的分歧仍然存在，有一个“一般语言能力”的假设仍没有被完全抛弃。

1.2 语言能力结构与因子分析法

如前所述，从智力和语言能力的研究历史看，一般智力和一般语言能力概念的提出都和一种叫主成分分析的统计方法有关，而多维智力与语言能力的可分立学说则和另一种统计方法——因子分析法相关。那么这两种统计方法到底有什么不同呢？竟然让基于同一批测试数据的研究得出完全不同的结论。

先来考察一下主成分分析法。

统计学上的主成分分析是研究用少数几个线性组合来解释原有全部变量的绝大部分信息（即分数方差）的一种多元统计方法。主成分分析假设原始变量之间有适度相关性，而求解出的主成分之间则相互独立。

假设研究对象有 p 个测量指标，即，有一组可测变量 $X=(X_1, X_2, \cdots, X_p)$，然后对 X 进行线性组合：

$$Y_i=u_{i1}X_1+u_{i2}X_2+\cdots+u_{ip}X_p$$

Y_i（$i=1, 2, \cdots, p$）就是所谓主成分。其中 Y_1 叫第一主成分，它最大化地保留了全部原始变量的信息。公式中每个变量前的系数 u 叫这个变量的权，权的大小反映了这个变量在组合中的重要性。一般地，人们把某一个主成分的方差占总方差的比例叫这个主成分的方差贡献率。据观察，许多中国人编制的英语试卷，其第一主成分方差贡献率会在 30%~50%之间（杨惠中 1998）。在笔者见过的诸多语言测试分析文章中，第一主成分所解释的分数方差也总是比其它主成分要大得多。因此，笔者认为：如果研究目的是要考查试卷中的题目变量是否向某个维度聚集或者收敛，第一主成分的方差贡献率就是一个很好的统计指标。第一主成分方差贡献率越大，说明它从全部题目变量中提取的信息量越大，或者说，全部题目得分变量向这个第一主成分聚集的程度就越高。

正如 Oller 所碰到的情况一样，笔者在用主成分法对 CET 试卷做结构分析时也发现：第一主成分的方差贡献率相对于其它主成分来说，总是特别大，而且大部分题目得分变量在第一主成分上都显得很重要（朱正才 2002）。笔者曾给出的解释是：试卷主要是在测量一般英语语言能力，它和大部分题目都有关。但这种解释的科学性和认知机制，笔者一直也没有看到相关的研究报告。令人困扰的还有：几乎所有的这类研究，结果都如此相似！难道这些题目、结构都不相同的英语试卷真的会一致性地在测量某种一般英语语言能力吗？

再来看因子分析法。

因子分析法（指 R 型探索因子分析）是根据一组变量 X_i（$i=1, 2, \cdots, p$）之间相关性的大小，把变量分成若干组，组内的变量相关性较高，不同组的变量之间相关性低。然后，每组变量都用一个不可观察的公因子来表示，而且假设每个观察变量的方差都可以分解成两个部分：一部分由公因子解释，另一部分与公因子无关——由变量的独特个性解释。

因子分析的数学模式如下：

$X_i=a_{i1}F_1+a_{i2}F_2+\cdots+a_{im}F_m+e_i$ （$i=1, 2, \cdots, p$；$m<p$）

F_m就是所谓的公因子，e_i就是变量 i 的独特因子，a_{im} 叫因子载荷矩阵，表示公因子 i 与某个变量之间的相关性。显然，因子载荷越大，表示该变量在公因子中越重要。

因子分析中求解公因子的数学方法很多，主成分法（Principal Component Analysis）和主轴因子法（Principal axis factoring）在语言测试的研究中比较常用。不管用何种数学方法求解因子载荷矩阵，结果都不惟一。任何一组因子载荷都可以由另一组因子载荷的线性组合来代替，这叫因子旋转（注意主成分分析没有这一步），旋转后的因子载荷矩阵将有利于因子意义的解释。

在做语言能力结构分析时，主成分分析法适用于分析全部变量向第一主成分聚集的程度，对每个主成分的语言学意义解释并不清晰；而因子分析则主要是为了识别一组变量的内部结构，重点关注因子之间的差异性，因子的意义比较清晰，尤其是因子旋转对因子的解释和命名意义重大。

最后来讨论一下实证因子分析法（Confirmatory Factor Analysis, CFA）。

实证因子分析是在探索性因子分析基础上发展起来的。在分析语言试卷的能力结构时，它要求先提出关于试卷语言能力结构的假想模型（即构念），然后再用大样本数据（一般要求 80 人以上）来拟合这一模型。如果数据和模型拟合良好，就接受关于这个模型的假设，否则就不接受或者修改模型假设。

可见，探索因子分析是从样本数据出发，用数学方法去探索因子结构——这是一个数据驱动的发现过程；而实证因子分析则反过来，先有因子结构的构念，再用样本数据去验证设想——是一个理论驱动的拟合检验过程。

2. 语言能力结构的聚合度、辨别度和拟合度

2.1 聚合度

笔者曾经多次用同一份试卷的同一批考试数据分别进行主成分分析和因子分析，研究结论很难统一解释；前者说存在一般英语语言能力因子，而后者说只有分立的听，说，读，写等因子（朱正才，2002；杨惠中，1998）。如果用实证因子分析来检验，两种模型又都能得到数据的支持——尽管模型的拟合程度上有些差异。那么，这样的研究结果是否就相互矛盾呢？

现在，笔者终于能比较好的解释这种现象了。其实，用主成分分析法分析试卷能力结构时，关注的焦点在第一主成分所能解释的全部题目分数方差大小上。第一主成分最大化地综合了全部题目变量信息，其方差贡献率的大小反映的是所有变量向某个共同因子集中的程度。因此，笔者认为：可以考虑把第一主成分的方差贡献率定义为试卷语言能力的聚合度。第一主成分就可以解释为人类语言能力的核心部分，叫一般语言能力也未尝不可。如果从认知语言学的角度看，这种一般语言能力就是人类赖以生存的基本认知能力（包括记忆，推理，辨音，空间形象的感知等）和基本语言知识（音位、词汇语法、句子和篇章以及社会语言学知识等）的复合体，也可以包含乔姆斯基的“人类一般语言能力”（一种天生的遗传的本能）。笔者认为，在大脑认知心理和认知神经机制上，人类语言能力都不是在大脑的某个区域独立完成的，语言能力是人在社会交往中运用其基本认知能力所创造的伟大成果。

石毓智（2007）也曾在反驳乔姆斯基的语言能力理论基础上提出了自己的“语言能力合成说”，认为语言能力不是天生的，也不是独立于其他认知能力的，它是由 7 种更基本的认知能力协同合作的结果。这 7 种基本认知能力是：符号表征能力，对量的认知能力，概括分类能力，记忆预见能力，联想推理能力，声音形状的辨别能力，空间时间的辨别能力。显然，这基本就是智力测量的内容。笔者比较赞同石氏的这一语言能力认知观。

乔姆斯基的抽象性的语言能力观和石毓智的语言能力合成说，都为主成分分析发现的“一般语言能力”提供了一种理论解释。这些能力也可能指示了人的一种英语学习潜能（Aptitude），这种理解与 Woods 的观点不谋而合。

2.2 辨别度

用因子分析法分析语言测试试卷的能力结构时，首先要确定一个合适的公因子数目，然

后再求旋转后的因子矩阵，并给各因子以适当的解释。显然，如果先能给定一个可接受的总分数方差解释百分比，是可以据此确定公因子数目的，这时，如果发现的公因子很多，说明试卷的语言能力结构分散，或者说试卷对不同的语言能力成分的辨别力很强。笔者认为，这个因子数目，可以作为这份试卷的语言能力辨别度指标。而对识别出的因子的解释则体现一定的语言能力观。

2.3 拟合度

用实证因子分析法验证试卷能力结构（其实，就是一种结构方程建模技术）就是要度量假设的语言能力模型和考试数据的拟合程度。AMOS 分析软件提供的统计指标：卡方似然比（x^2/df），通俗易懂而且很常用。在参考了一些这方面的专家意见后（温忠麟等，2004），笔者认为，可以把卡方似然比作为评价模型和数据的拟合良好性的首选指标。

显然，这三大指标：聚合度、辨别度和拟合度共同描述了一份试卷语言能力结构的主要特性，而且彼此不矛盾。

聚合度显示的是一份试卷的题目向一个核心语言能力收敛的程度，聚合度高说明试卷中有很多题目，彼此之间存在较好的同质性——当然这也并不一定是考试设计者所想要的结果。笔者甚至怀疑，语言测试中有时会出现这样一种极端情况，就是所有的题目在很大程度上都在测量人的智力水平，但表面上看似乎都只是在测量人的语言能力。

当一个试卷的语言能力辨别度很高时，说明考试考了多种不同的语言能力，不同公因子下的题目彼此之间差异明显。如果这时对公因子的解释又恰好包括了听说读写等诸多必要的语言能力成分，说明这个考试有一个好的设计，这时，向用户报道各分测验成绩就有了统计上的依据而且意义明确。

拟合度指标是越高越好，因为它反应的是观察数据支持考试达到预期目标的程度——而这正好是效度研究的经典范式。

3. 试卷语言能力结构实证研究

为了检验用以上 3 个统计指标来评价一份试卷语言能力结构特征的有效性，笔者用一份 2013 年 6 月的大学英语四级考试（CET4）试卷来做大样本的数据分析。

当我们用主成分分析法和因子分析法对同一份试卷的原始数据进行分析时，为了控制数学模型的规模和尽可能减少题目之间的多重共线性，一般要先对试卷全部题目（可能多达 70~150 题）做适当合并。这意味着：合并到一个变量内的小题目，彼此之间要非常相似，基本上没有能力目标上的差异（这个假设还需进一步验证）。

3.1 数据采集

研究数据来自 CET4 观察点学校的当届考生。随机抽取的样本量为 1364 人。考生原始题目得分被合并为 12 个变量——就是把基于同一个语篇的同种题型的题目分数加起来。

3.2 分析和讨论

对样本数据依次进行三种因子分析。SPSS 球形检验显示：KMO>0.85，说明样本数据满足了因子分析的统计要求。

3.2.1 主成分分析

分析题目变量相关矩阵，默认提取特征值大于 1 的主成分，这时第一主成分方差贡献率高达 47.5%。这说明 12 个题目变量同质性很高，有向一个核心的语言成分聚集的倾向，只要用一个主成分就可以提取全部题目变量大约一半的信息量。对第一主成分贡献信息量最大的题目变量前 4 名都是听力题目（载荷值依次为：0.839，0.828，0.761，0.753）。可见，这份试卷主要在测量与听力理解有关的语言能力。这个结果和许多英语教师的主观感受一致。如果教学和测试都在强调语言能力的综合运用，这也是一个近乎必然的令人满意结果。作为交际工具的语言，听不懂，还能指望什么呢？

3.2.2 因子分析

仍然分析变量相关矩阵。规定：累计解释方差不小于 50%（笔者认为：在语言测试领域，公因子即使是旋转后，其对全部变量方差解释比率也应达到 50%~60%以上），再强制提取这一要求下的最少数目公因子。经过反复尝试，当公因子为 4 时，解释的分数方差达可以达到

52%，符合要求。再用主轴因子法提取这 4 个公因子，并经过方差极大正交旋转得到因子载荷矩阵如下表 1（因子载荷大于 0.3 时才列出）。仔细分析这个载荷矩阵，发现：第 1 因子并不只是和听力有关，而是主要和有语言输出的题目变量有关，解释为“产出性语言能力”比较合适。第 2 因子很明显是听力理解因子。第 3 因子是快速阅读理解因子。第 4 因子是仔细阅读因子（其中，CLOZE 载荷 0.595 最高，有些出人意料，看来 CLOZE 主要是在测量阅读理解能力）。考虑到这份试卷在解释分数方差 52%的前提下，识别出了 4 个公因子，而且意义明确，可以认为这份试卷对不同语言能力成分有较强的辨别力。

表 1.　旋转后的因子矩阵

题目变量	Factor			
	1	2	3	4
LC_WORDS：听短文3遍（填单词）	.664	.426		
LC_MEAN：听短文3遍（填意思）	.614	.565		
TRANSLAT：翻译（开放题）	.591			.337
RC_CARE1：仔细阅读（配对题）	.506			.331
WRITING：作文（开放题）	.466			
LC_SCONV：听短对话（选择题）	.343	.673		
LC_LCONV：听长对话（选择题）		.572		
LC_3PASS：听短文1遍（选择题）	.394	.499		
RC_SCAN1：快读1（判断题）			.568	
RC_SCAN2：快读2（填空题）			.550	
CLOZE：综合填空（选择题）	.336			.595
RC_CARE2：仔细阅读（选择题）				.346

3.2.3 实证因子分析

笔者使用 AMOS4.0 软件对样本数据反复进行了实证因子分析尝试。

模型 1（图 1）是根据 2007 年改革后的 CET4 考试大纲和样卷说明的语言能力目标设计出来的，是一个典型的理论驱动模型；模型 2（图 2）则是依据前文探索因子分析结果所提示的信息而设计出来的数据驱动模型。拟合效果显示：考试设计者预想中的能力结构并没有得到数据的支持，“卡方似然比”大于 2，其它拟合指标也不理想；但是，数据驱动的模型 2 则拟合较好（见表 2）。这个结果其实一点也不出乎意料，因为现在的 CET4 试卷，各部分多使用综合性试题，测量的语言能力目标自然也就是以综合性语言运用能力为主，这和要把它们清晰地区分为不同的语言能力成分的想法是矛盾的。笔者曾对 2007 年改革前的 CET4 试卷进行过类似研究，发现模型就可以显著地识别出听、读、写、语言知识等不同的语言能力成分来。其中的原因就在于老的 CET4 试卷各部分题目综合性不强，听力就是听力（选择题），不要求写，阅读也是如此。

模型 2 中的第一公因子是语言产出能力，是一个典型的综合性语言能力。阅读能力被分成了两块：快速阅读和仔细阅读。Cloze 不是语言的综合运用，而是最好的仔细阅读（在因子 4 里载荷最高）。写作被包含在语言产出因子里。笔者认为，这个结果除了不能完美支持目前的分数分项报道外，并没有什么不妥。这种试卷模式和语言能力结构可能正是社会所需，反映了英语教学和测试的进步。

表 2.　因子结构模型拟合结果

Model	模型说明	卡方似然比	GFI	AGFI	RMSEA	NFI/RFI	拟合效果
1	理论驱动	303.432/51=5.95	0.961	0.941	0.06	0.957/0.944	拟合不好
2	数据	145/44=3.3	0.982	0.969	0.041	0.979/0.969	拟合较好

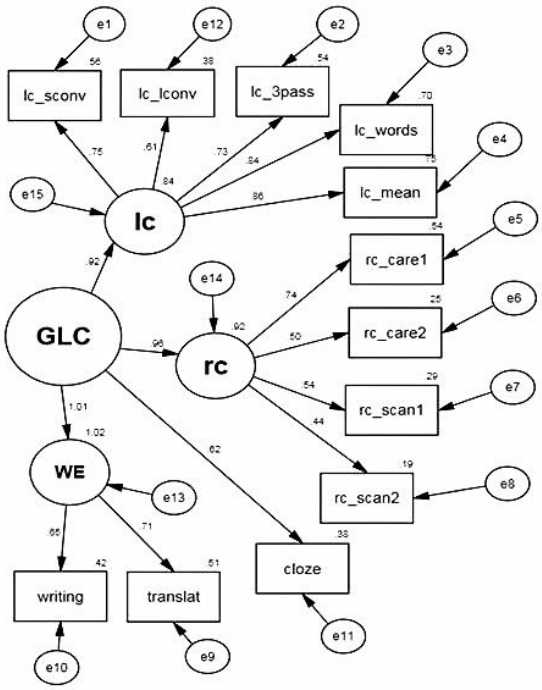

图 1. 模型 1

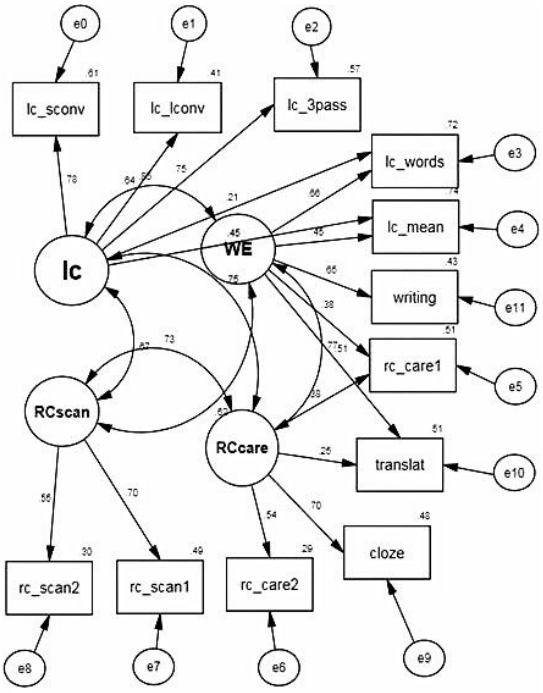

图 2. 模型 2

4. 小结

本文主要探讨了语言能力的心理结构问题。综合起来说，笔者主张把语言能力看作一个“多层次多侧面的可分的心理空间结构”。首先，语言能力可以根据人脑语言认知加工从简单到复杂划分出不同的层次：从原始的语音、符号等低级图式（Schema）加工到形象和意义的高级图式加工，再到社会交际语境中的综合运用（最复杂社会语言学图式）；其次，语言能力还可以在每个层次上都呈现出多个语言运用的侧面（如听说读写等）。可能在认知层次的低端，语言能力侧面彼此之间的辨析度不高，变化较少，个体之间差异小；而在认知的高端——社会交际层面，其侧面的变化就丰富多彩，听说读写的不同组合和强弱变化可以搭配出不同风格的交际能力来。有的人能说会道，有的人只能读不能说，有的人擅长阅读和翻译等等。这些不同特色不同风格的交际能力对应于语言测试不同类型题目测量目标。每个人在这些不同的测量目标量表上都有一个自己的位置。从语言学习的阶段性结果来说，每一个具体的人，不但在每个题目能力目标量表上有一个大致位置，而且在这个分层次多侧面的心理结构空间中也有一个大致位置。语言测试的任务就是给每个被试大致确定这个位置。如果从语言能力空间结构的几何意义上看，聚合度描述的是全部能力观察点向某个空间维度收拢的趋势，辨别度描述的是这些点能被彼此辨析出来的程度，而拟合度描述的则是这个观察到的能力空间结构形态多大程度上符合了设计者的构念。

显然，人的语言能力是不能无限的细分下去。从社会科学研究的角度讲，细分的程度要适当，不能像智力的 120 因素说那样分得太细，结果让人难以理解也难以准确把握，最终走向不可知论。对语言能力的细分应该以满足语言应用的实际需要为度。人们应该在模型精度和测量结果的可靠性之间保持某种平衡，这对一个只有 2 小时左右时间的英语考试来说尤为重要。

本文实证研究表明，2007 年改革后的 CET4 试卷主要是在测量被试的以听力理解为主的综合性英语语言能力，CET 考试大纲所描述的目标语言能力结构与试卷的实际情况稍有不符，尚需修订。

显然，这些结论还有待更深入研究，而本文的这种从多个视角来观察同一个问题的方法无疑为语言测试的效度研究提供又一种可能性。

参 考 文 献

[1]Bachman, L. *Statistical Analyses for Language Assessment* [M]. Cambridge University Press, 2005:281-282.

[2]Gardner, H. *Multiple Intelligences: The theory in Practice* [M]. New York, Basic Books，2006:1-25.

[3]Oller, J. *Issues in Language Testing Research* [M]. Rowley, MA: Newbury House, 1983:3-28.

[4]Spearman, C. “General intelligence” objectively determined and measured [J]. *American Journal of Psychology* 15,1904: 201-223.

[5]Vollmer, H. The structure of foreign language competence [A]. In A. Hughes & D. Porter (eds.), *Current Developments in Language Testing* [C]. UK: University of Reading, 1983:3-30.

[6]Woods, A. *Statistics in Language Studies* [M]. Cambridge University Press, 1986:283-285.

[7]石毓智. 认知能力与语言学理论[M]. 上海：学林出版社,2008：2-22.

[8]温忠麟, 侯杰泰, 马什赫伯特. 结构方程模型：拟合指数与卡方准则〔J〕.《心理学报》2004（36）：186.

[9]杨惠中, Weir, C. 大学英语四、六级考试效度研究[M]. 上海：上海外语教育出版社, 1998：55-62.

[10]朱正才.大学英语考试电脑自适应测验[M]. 上海：上海交通大学出版社, 2002：110-114.

基金项目：本研究得到了教育部人文社会科学研究规划基金项目“语言测试公平研究——科学与道德”（项目编号：13YJA740088）资助。

作者简介：朱正才，上海交通大学外国语学院教授，博士，博导，研究方向：语言测试、心理测量与统计、认知语言学等。