大学英语网考研究专稿

大学英语四级网考效度初探
——影响考生评价和考试成绩的因素分析

金 艳,吴 江
(上海交通大学 外国语学院,上海 200030)

摘 要: 自 20 世纪 90 年代后期,信息技术开始被引入语言测试领域。与此同时,语言测试者越来越关注技术的使用对语言测试效度所产生的影响,以确保考试的公平公正性。作为大学英语四级网考效度研究的一部分,本文分析了影响考生对四级网考评价和考试成绩的因素。研究表明,语言水平和计算机熟悉程度对考生评价具有统计意义上的显著影响,地区经济和社会发展程度因素对考生评价的影响较小;计算机熟悉程度对网考成绩影响较大,其影响力既有统计意义上的显著性,又具有实际意义,地区经济和社会发展程度因素对网考成绩没有产生具有实际意义的影响。

关键词: 四级网考;考生评价;考试成绩;影响因素;影响力
中图分类号: H319.3 **文献标识码:** A **文章编号:** 1001-5795(2010)03-0003-0008

20 世纪 90 年代后期,美国教育考试服务中心(ETS)开始研究和推广计算机化的托福考试(TOEFL CBT),本世纪初 ETS 率先在全球范围内实施基于互联网的托福考试(TOEFL iBT),加快了语言测试的信息化进程。同时,语言测试界也开始关注信息技术的运用是否会影响语言测试的效度,基于计算机或网络的考试是否会对技术应用能力差的考生不公平。为了提高考试效度,改进考试后效,有效预防高科技手段的作弊,我国教育部高等教育司也于 2007 年 5 月启动基于计算机和网络的大学英语四、六级考试项目❶(以下简称四、六级网考;参见金艳,吴江,2009)。作为四、六级网考的设计者,在题型设计和考试方式上充分利用先进信息技术的同时,也必须关注考试的效度以及影响考试效度的因素。本文以四级网考试点考试的考生为样本,通过对问卷调查数据和考试成绩的分析,探讨影响考生对考试的评价及其考试成绩的因素,以期初步论证四级网考的效度。

1 研究问题

1.1 研究背景

基于计算机的考试不仅考核语言能力,同时还涉及考生的计算机应用能力和视听感知能力。因此,影响计算机化考试效度的因素主要产生于测试任务的呈现方式和考生答题方式,而且这些因素可能对不同的考生会产生不同的影响,包括考生对考试的态度和评价、考生的心理过程、认知策略、答题策略以及考试成绩等。

对基于计算机的语言测试的效度研究大多是机考与纸笔考试的等效性研究,研究对象是同样的测试任务在计算机上呈现或用键盘输入方式答题是否会对考试过程、考试结果和评分产生影响,即考生的计算机熟悉程度对答题过程和考试成绩的影响。Kirsch 等(1998)对 1996 年 4 月和 5 月全球托福考生的计算机熟悉程度的调查发现,计算机熟悉程度与考生的性别、母语、出生地以及考点所在地有关,与考生成绩也相关。Breland 等(2004)分析了 1998 年至 2000 年托福机考的 83 个作文题的成绩,考生分别通过手写和在计算机上打字完成写作。结果发现,对于语言水平相当

作者简介:金 艳:女,博士,教授,博士生导师。研究方向:语言测试理论与实践。
吴 江:男,教授,硕士生导师。研究方向:语言测试理论与实践。

收稿日期:2009-10-18

❶ 本文不区分基于计算机/网络的考试、计算机/网络化考试等术语。

的考生群体，手写作文的得分均高于计算机上完成的作文，这说明答题方式对写作成绩产生了影响。Weir等(2007)分析了262名雅思考生的纸笔和计算机写作测试的认知过程和成绩差异。与Breland等人的研究结果不同，Weir等人发现，纸笔和计算机上完成的作文得分没有显著差异，计算机熟悉程度和考试焦虑因素对考试成绩总体上没有产生显著影响。Wolfe & Manalo(2005)研究了写作答题方式对作文评分的影响。通过对ETS授权评分员所评的133,906篇托福作文成绩的分析证明，计算机上完成的作文评分信度略高于手写作文，机考作文与托福客观题部分的得分相关性也高于手写作文；客观题得分低的考生手写作文得分高于机考作文，而客观题得分高的考生两者得分差异不显著。

基于计算机的考试研究近年来也引起我国学者的关注。李清华(2006)综述了国外纸笔和基于计算机考试的等效研究，结论是，两种模式的阅读理解测试基本等效；在阅读速度测试方面，计算机化考试不比纸笔考试优越；计算机熟悉程度没有对测试行为和分数产生显著影响。王勇旗(2008)在探讨语言测试的信度和效度是否会向技术妥协时指出，虽然以往的实证研究基本证实计算机化测试与纸笔测试的结果具有对等性，但是这一结论是基于考生达到了一定水平的计算机熟悉程度。陈慧麟(2009)对试题呈现方式的研究表明，纯文字测试和多媒体形式呈现的计算机化考试即使考查范围相同，也会在效度上有较为显著的差异，过多的多媒体材料可能会降低对知识点的考查效度。李清华、孔文(2009)分析了基于计算机的语言测试效度研究的研究内容和方法，并指出基于计算机的语言测试的开发和推广应在教育和心理测量等相关理论指导下稳步进行，而效度验证是该领域发展的最迫切任务。

1.2　研究问题

与其他基于计算机的语言测试一样，四级网考首先关注的是考试效度，即四级网考是否有效考核了“学生英语综合应用能力，特别是听说能力”(教育部高等教育司,2007:1)，网考的设计和实施是否会给考试结果带来“与所测构念不相关的差异”(construct-irrelevant variance)(Messick 1989,1996)。差异可能来自多方面。例如，由于我国教育资源的分布不均衡，各地经济和社会发展程度也不同，大学生所在环境的计算机化和网络化程度不同。而且，在日常学习和生活中，大学生使用计算机的习惯不同，对计算机操作的适应程度也不同。因此，大学生的计算机使用背景和熟悉程度可能会影响他们对网考的接受程度，从而影响他们对网考的评价和网考成绩。具体来说，本文的研究问题是以下四个：

(1) 四级网考的效标关联效度如何？本问题将采用已经比较成熟的四级纸笔考试为效标，通过对网考和纸笔考试的总分及单项分之间的相关性研究，检验四级网考的同时效度。

(2) 考生对四级网考的评价如何？这个问题将分析考生对网考的有效性、难易度、试卷结构、答题时间、适应程度和紧张程度等方面的评价，以检验网考的表面效度。

(3) 影响考生对四级网考评价的因素是什么？这个问题将分析不同地区、语言水平和计算机熟悉程度的考生对网考的评价，以论证影响考生评价的因素。

(4) 影响四级网考成绩的因素是什么？这个问题将研究考生所在地区的经济和社会发展程度因素和考生的计算机熟悉程度因素对其网考成绩所产生的影响。

2　实验设计

在本研究的实验中，来自53所高校的4643名考生于2008年12月20日参加了全国统一实施的四级纸笔考试，并于第二天参加了四级网考试点考试。四级网考结束后全体考生参加了问卷调查。因此，考生的纸笔考试成绩、网考成绩和问卷数据可以对应❶。本节介绍实验工具和样本。

2.1　实验工具

2.1.1　*四级纸笔考试和网考*

四级纸笔考试由四个部分组成：听力、阅读、完型填空以及写作和翻译。听力部分包括听力理解题以及听写，听力理解题型为多项选择题；阅读部分包括快速阅读理解和仔细阅读理解，题型有多项选择、句子填空和集库式的选词填空题；完型填空部分是多项选择题型的篇章完型填空；写作和翻译部分是一篇120词以上的命题作文和五个单句的汉译英。2008年12月四级网考试点考试由三个部分组成：听力、阅读以及综合。听力部分为多项选择题型的听力理解，听力材料有短篇新闻、中篇报道或访谈以及长篇视频；阅读部分包括快速阅读理解和仔细阅读理解；综合部分有听写、

❶ 为保证考生成绩和问卷数据的私密性，本研究只分析群体数据，没有对任何个体进行单独研究和分析。

表1　四级纸笔考试和2008年12月四级网考的试卷构成和题型

四级纸笔	单项	听力测试	阅读测试	完型填空	写作和翻译
	比例	35%	35%	10%	20%
	题型	多项选择;听写	多项选择;句子填空;选词填空	多项选项完型填空	写作;单句汉译英
四级网考	单项	听力测试	阅读测试	综合测试	
	比例	25%	30%	45%	
	题型	多项选择	多项选择;句子填空	听写;跟读;多项选择(语法结构);写作	

注:四级网考的试卷构成和题型将根据试点情况进行调整。

表2　不同地区、层次和类型的院校数量和考生人数(N_1=53;N_2=3984)

分类	地区						层次		类型	
	华北	东北	华东	华中	西南	西北	211	非211	综合	非综合
N_1	11	7	9	12	8	6	37	16	23	30
%	20.75	13.21	16.98	22.64	15.09	11.32	69.81	30.19	43.40	56.60
N_2	927	605	563	933	609	347	2 827	1 157	1 710	2 274
%	23.27	15.19	14.13	23.42	15.29	8.71	70.96	29.04	42.92	57.08

注:我国行政区域划分为华北3省2市、东北3省、华东6省1市、华中6省、西南4省1市、西北5省。

句子跟读、多项选择题型的语法结构以及基于听力材料的写作。考试时间均为两小时左右。成绩经过等值和常模调整后,报道单项分和总分。表1描述了四级纸笔考试和网考的试卷构成和题型。

与纸笔考试相比,四级网考的试卷设计和考试方式有较大变化,主要特点有① 考试任务更加综合。除听力理解外,听写、跟读、语法结构和写作均为基于听力的综合测试任务;② 增加了口语测试。尽管目前采用的单句跟读主要测试考生的语音和语调,但与纸笔考试相比,这是新增加的语言能力;③ 加强了对听的能力的测试。听力素材丰富,有短篇、中篇和长篇,有单人讲话、双人谈话或多人讨论,有视频和音频节目,且都是从广播、电视、网络等媒体采集的真实材料;④ 考生在计算机上完成考试。考生在计算机屏幕上阅读,在计算机上听录音、看录像,并在计算机上完成选择、录音、打字等答题操作。

2.1.2　问卷调查

问卷调查的目的是了解考生对网考的评价和计算机熟悉程度。其中,有关计算机熟悉程度的问题参考了Eignor等(1998)对托福考生计算机使用机会、目的、态度和经历等的调查问卷。问卷共分十个部分:考生的计算机使用背景(第1~6题),网考的听力测试过程(第7~16题),对网考有效性(第17~27题)、难易度(第28~41题)、试卷结构和分值比例(第42~49题)、答题时间(第50~57题)、在计算机上答题的适应程度(第58~65题)、考试过程中的紧张程度(第66~74题)、考试系统操作难易度(第75~80题)、答题操作难易度和实用性(第81~89题)的评价,以及对试题和考试系统的改进建议(第90题)。除第90题为开放式问题外,其余都是利克特五点量表或三至五个选项的单选题(限于篇幅,本文未附问卷全文)。

2.2　实验样本

数据经过整理,得到3984名考生的有效数据(各项考试成绩完整且问卷缺失答案少于5%)。考生所在的院校分布在华北、东北、华东、华中、西南和西北六大行政区域,院校的层次涵盖211院校和非211院校,院校类型有综合性和非综合性(各类院校数量和考生人数见表2)。

数据分析时,将考生分成三类对照组(见表3)。第一类为地区组,根据地区经济和社会发展程度划分为三个水平,华北和华东地区为发展程度相对较高的Ⅰ组,东北和华中地区为中等发展程度的Ⅱ组,西南和西北地区为发展水平相对较低的Ⅲ组,人数最少的Ⅲ组样本量为956;第二类为语言水平组,四级纸笔考试成绩被作为考生语言水平指标,根据四级纸笔考试总分划分为高、中、低三个水平,其中高分组和低分组人

表3　考生分组及各组人数和比例(N=3984)

分组	地区			语言水平[1]			计算机熟悉程度[2]		
	Ⅰ	Ⅱ	Ⅲ	高	中	低	高	中	低
N	1 490	1 538	956	931	2 083	970	607	2 738	639
%	37.40	38.60	24.00	23.37	52.28	24.35	15.24	68.72	16.04

注:1. 按四级纸笔考试总分划分,高分组551~710,中等组431~550,低分组220~430。2. 按问卷第1~3、5、58~65、75~86题总值划分,缺失值用均值取代,高组74~96,中组56~73,低组24~55。

数和比例基本均衡，且样本量均大于900；第三类为计算机熟悉程度组，根据考生对问卷中“计算机使用背景”、“网考答题适应程度”、“考试系统操作难易度”和“答题操作难易度”部分的回答，得出计算机熟悉程度指数，分成高、中、低三个水平，其中熟悉程度高和低两个组的人数和比例基本均衡，且样本量大于600。

3　数据分析

3.1　考试成绩和相关性分析

从表4的描述统计数据来看，四级网考和四级纸笔考试的总分均值和标准差很接近。由于本研究采用的是经调整后的报道分[❶]，两个考试的分数并不直接可比。但是，从得分率数据看，网考的听力和阅读高于纸笔考试的相应部分；而网考综合部分得分率偏低，这部分是网考与纸笔考试差异最大的部分，包括基于听力材料的听写、写作、语法结构和句子跟读。

表4　2008年12月四级纸笔考试和网考单项及总分报道分的均值和标准差（N=3984）

	四级纸笔					四级网考			
	听力	阅读	完型	写译	总分	听力	阅读	综合	总分
比例	35%	35%	10%	20%	100%	25%	30%	45%	100%
均值	172.06	173.65	49.60	94.01	489.32	134.12	174.33	184.75	493.20
得分率(%)	69.24	69.88	69.86	66.20	68.92	75.56	81.85	57.82	69.46
标准差	30.64	29.67	10.76	13.35	73.76	22.30	28.06	35.25	71.80

从表5的相关数据来看，四级网考和纸笔考试的总分相关为0.80，所测能力相近的单项之间相关为：听力0.57、阅读0.58、纸笔完型与网考综合0.50、纸笔写作和翻译与网考综合0.59；其他单项相关中，纸笔听力与网考综合最高(0.72)；纸笔单项与网考总分的相关中，纸笔听力与网考总分相关最高(0.76)；网考单项与网考总分的相关中，网考综合与总分相关最高(0.89)。对比网考和纸笔考试内部相关数据发现，网考的内部相关低于纸笔考试。以听力和阅读为例，网考为0.48，纸笔考试为0.70。原因之一是纸笔听力和阅读比例(70%)高于网考听力和阅读(55%)，另一原因是两者在听力测试任务上的差异。两个考试的阅读部分基本一致，但听力测试差异较大，纸笔听力含有复合听写，而网考听力测试仅含听力理解，听写归入网考综合部分。此外，纸笔听力采用文字稿朗读的录音制作材料，其中听力短篇的文字比较书面语化，而网考听力采用广播电视的录音和录像节目，具有更多口语特征。从考试的设计原则来看，一个考试的单项之间相关在0.4至0.6之间更为合理，因为这些单项测试了语言能力的不同方面。

表5　2008年12月四级纸笔考试和网考单项及总分的相关

	纸笔听力	纸笔阅读	纸笔完型	纸笔写译	纸笔总分	网考听力	网考阅读	网考综合	网考总分
纸笔听力	1	.70	.56	.62	.91	.57	.59	.72	.76
纸笔阅读		1	.56	.59	.90	.50	.58	.61	.68
纸笔完型			1	.49	.70	.39	.49	.50	.56
纸笔写译				1	.77	.43	.51	.59	.62
纸笔总分					1	.59	.66	.74	.80
网考听力						1	.48	.59	.79
网考阅读							1	.56	.81
网考综合								1	.89
网考总分									1

注：所有表内的相关（双尾）都在0.01水平上显著。

3.2　网考评价数据分析

从表6全体考生对四级网考的评价数据来看，考生对网考整体有效性的评价为3.45，分值比例合理性为3.27，适应程度为2.96，紧张程度为3.00，答题时间适中的比例为73.3%。但是，考生对网考整体难易度的评价为偏难(2.42)。考生对网考单项测试的各项指标评价数据显示，均值最高的是仔细阅读、快速阅读、写作和语法结构，其次是听力理解；各项指标评价均值都相对较低的是听写和跟读，而且这两个单项在各项评价指标上的标准差均大于其他单项，说明考生

表6　全体考生对四级网考的评价[1]（斜杠后的值为标准差）

	有效性	难易度	分值比例[2]	适应程度	紧张程度	答题时间
网考整体	3.45/.81	2.42/.80	3.27/.82	2.96/.95	3.00/.88	73.34
听力理解	3.32/.97	2.45/.86	64.43	2.88/.95	2.75/.92；2.77/.92[3]	62.10
听写	3.10/.98	2.16/.90	59.79	2.49/1.01	2.47/.98	47.34
跟读	3.01/1.05	2.31/.91	54.34	2.47/1.04	2.46/1.00	58.48
写作	3.57/.85	2.90/.83	78.61	3.15/.93	3.16/.85	74.47
语法结构	3.39/.81	2.90/.79	74.85	3.09/.86	3.07/.81	76.41
快速阅读	3.74/.82	2.89/.83；2.93/.86[4]	83.61	3.27/.92	2.99/.89	69.85
仔细阅读	3.78/.79	2.94/.79	81.28	3.35/.88	3.12/.83	77.33

注：1. 评价量表是1-完全无效、很难、很不合理、很不适应、很紧张，5-很有效、很易、很合理、很适应、很放松；答题时间是选择“适中”的比例。2. 除“网考整体”为五点量表外，其余各题选项是“太多、较多、适中、较少、太少”，表中是选择“适中”的比例。3. 听力录音和录像的紧张程度。4. 快速阅读选择题和填空题的难易度。

❶ 本研究未采用原始分，因为网考采用题库组题，所以考生所做试题不同。

之间意见分歧较大。

从表7分组考生对网考整体的评价数据来看,地区组之间的差异较小,除难易度外,地区Ⅲ组对网考各项指标的评价均略高于其他两个地区组;语言水平组之间的差异较明显,除答题时间外,其余各项指标的评价趋势均为高分组比中等组好,中等组比低分组好;计算机熟悉程度组的差异最明显,除答题时间外,其余各项指标的评价趋势均为熟悉程度高组比中等组好,中等组比熟悉程度低组好。

表7　分组考生对四级网考整体的评价(斜杠后的值为标准差)

		有效性	难易度	分值比例	适应程度	紧张程度	答题时间
地区	Ⅰ	3.43/.83	2.45/.83	3.27/.85	2.95/.99	2.98/.93	70.67
	Ⅱ	3.42/.81	2.43/.79	3.25/.79	2.95/.94	2.99/.86	73.21
	Ⅲ	3.52/.78	2.36/.76	3.31/.81	3.01/.92	3.03/.84	77.72
语言水平	高	3.55/.78	2.61/.75	3.40/.82	3.09/.92	3.08/.84	75.73
	中	3.47/.80	2.42/.79	3.30/.79	2.99/.95	3.01/.87	72.54
	低	3.29/.84	2.24/.82	3.10/.85	2.79/.97	2.87/.93	72.78
计算机熟悉程度	高	4.03/.69	2.83/.90	3.83/.79	3.96/.69	3.34/.96	74.46
	中	3.47/.71	2.45/.72	3.28/.73	2.96/.80	3.00/.79	76.33
	低	2.78/.86	1.90/.74	2.72/.85	2.03/.77	2.65/1.03	59.47

注:评价量表见表6。

表8是ANOVA组间差异显著性检验的结果。可以看出,地区组之间仅有Ⅰ组和Ⅲ组、Ⅱ组和Ⅲ组对考试有效性和难易度的评价有显著差异,其余各项评价指标的组间差异均不显著;语言水平组之间仅有高分组和中等组对紧张程度的评价无显著差异,其余各项评价指标都有显著性的组间差异;计算机熟悉程度组对网考评价的所有组间差异都显著,且F值远大于语言水平组。

表8　对四级网考整体评价的组间差异显著性检验结果

		有效性	难易度	分值比例	适应程度	紧张程度
地区	Ⅰ-Ⅱ	.82	.46	.50	.92	.80
	Ⅰ-Ⅲ	.01	.00	.35	.15	.24
	Ⅱ-Ⅲ	.00	.03	.13	.17	.34
	组间 F/Sig.	5.21/.01	4.28/.01	1.16/.31	1.25/.29	.73/.48
语言水平	高-中	.01	.00	.00	.01	.06
	高-低	.00	.00	.00	.00	.00
	中-低	.00	.00	.00	.00	.00
	组间 F/Sig.	27.51/.00	49.87/.00	32.61/.00	23.74/.00	14.04/.00
计算机熟悉程度	高-中	.00	.00	.00	.00	.00
	高-低	.00	.00	.00	.00	.00
	中-低	.00	.00	.00	.00	.00
	组间 F/Sig.	466.70/.00	240.80/.00	329.44/.00	956.15/.00	101.06/.00

注:差异显著性水平为0.05。

3.3　影响网考成绩的因素分析

3.3.1　两因素方差分析

表9的描述性统计数据显示,地区组的网考成绩都是Ⅰ组高于Ⅱ组,Ⅱ组高于Ⅲ组;随着计算机熟悉程度的提高,考生的网考成绩也呈明显上升趋势。ANOVA差异显著性检验证明,网考总分和单项分的所有组间差异都在0.05水平上显著。但是,纸笔考试成绩除完型部分Ⅲ组成绩略高于Ⅱ组外,其余各组的成绩也呈现出与网考同样的趋势。因此,分析影响网考的因素时有必要对考生的语言水平(即纸笔考试成绩)加以控制。

以网考总分为因变量,地区和计算机熟悉程度为因素,以纸笔考试总分为协变量,进行两因素方差分析(各组均值和标准误及图示见表10和图1)。数据表明(显著性水平为0.05),两个因素对网考总分的影响不存在交互作用($F=0.39, p=.82$),但每个因素的主效应显著(地区 $F=3.91, p=.02$;计算机熟悉程度 $F=63.81, p=.00$)。所有的计算机熟悉程度组间均有统计意义上的显著差异($p=.00$);地区组仅有Ⅰ组和Ⅱ组之间有显著差异($p=.01$),Ⅰ组和Ⅲ组、Ⅱ组和Ⅲ组之间差异均不显著($p=.15$;$p=.31$)。

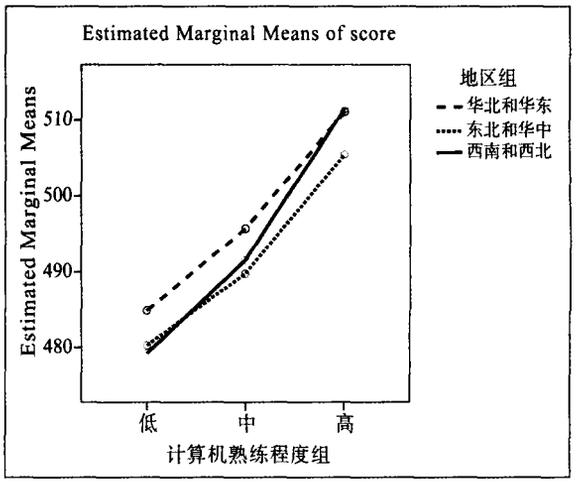

图1　地区和计算机熟悉程度组的四级网考总分

3.3.2　因素影响力分析

然而,统计意义上的差异显著性不一定具有实际意义,而且差异显著性检验无法分析因素影响力的大小。因此,本研究参考Taylor等(1998)托福计算机化考试的影响因素分析方法,运用Cohen(1988)系数 d 进行差异显著性检验,并用影响力(effect size)来表示组间差异的程度,影响力分析的另一好处是不受对比

表9　四级网考和纸笔考试总分和单项分的均值和标准差(括号内的值为标准差)

		网考总分	网考听力	网考阅读	网考综合	纸笔总分	纸笔听力	纸笔阅读	纸笔完型	纸笔写译
地区	Ⅰ	505.16(73.44)	136.59(22.68)	177.17(28.21)	191.39(36.29)	500.28(73.63)	177.16(30.12)	176.97(29.73)	50.81(10.65)	95.34(15.51)
	Ⅱ	488.07(70.46)	133.20(21.90)	173.21(27.83)	181.66(34.83)	486.43(74.17)	170.99(31.25)	173.11(29.57)	48.78(11.17)	93.55(15.32)
	Ⅲ	482.84(68.75)	131.76(21.99)	171.69(27.85)	179.39(32.63)	476.87(70.90)	165.85(29.12)	169.33(29.13)	49.03(10.09)	92.66(15.02)
计算机熟悉程度	高	526.50(66.23)	142.07(20.22)	184.31(26.11)	200.12(34.16)	511.88(68.63)	183.09(29.73)	179.75(27.43)	51.36(10.85)	97.67(14.15)
	中	493.16(69.85)	134.13(22.06)	174.32(27.50)	184.71(34.53)	490.44(73.12)	172.16(30.38)	174.29(29.51)	49.76(10.68)	94.24(15.15)
	低	461.75(70.97)	126.52(22.59)	164.89(28.99)	170.33(33.20)	463.05(73.28)	161.19(28.78)	165.10(30.55)	47.24(10.67)	89.52(16.25)

表10　地区和计算机熟悉程度组四级网考总分[a]

计算机熟悉程度	地区	平均分	标准误
低	Ⅰ	484.88[a]	2.74
	Ⅱ	480.29[a]	2.70
	Ⅲ	479.21[a]	3.34
中	Ⅰ	495.68[a]	1.35
	Ⅱ	489.76[a]	1.27
	Ⅲ	491.54[a]	1.64
高	Ⅰ	511.06[a]	2.54
	Ⅱ	505.46[a]	3.01
	Ⅲ	511.51[a]	3.62

注:a 表中的平均分是在协变量均值为489.32分的水平上进行估计的。

组样本大小差异的干扰(关于 d 的计算和解释,参考 www. uccs. edu/ ~ faculty/lbecker/)。表11的数据表明,地区因素对Ⅰ组和Ⅲ组之间的成绩有较小程度的影响,对Ⅰ组和Ⅱ组之间的总分和综合部分有较小程度的影响,地区因素对其余组间差异没有产生影响;计算机熟悉程度因素对高组和低组之间的差异产生了中等以上程度的影响,其中对总分和综合部分的影响程度最大;计算机熟悉程度因素对其他组间的差异影响较小。

表11　影响四级网考成绩的因素影响力大小

		总分		听力		阅读		综合	
		d	ES	d	ES	d	ES	d	ES
地区	Ⅰ-Ⅱ	.24	S	.15	N	.14	N	.27	S
	Ⅱ-Ⅲ	.08	N	.07	N	.05	N	.07	N
	Ⅰ-Ⅲ	.31	S	.22	S	.20	S	.35	S
计算机熟悉程度	高-中	.49	S	.38	S	.37	S	.45	S
	中-低	.45	S	.34	S	.33	S	.42	S
	高-低	.94	L	.73	M	.70	M	.88	L

注:d 为 Cohen 影响力系数;ES:effect size,N: No effect ($d<0.2$); S: small ($0.2\leq d<0.5$); M: medium ($0.5\leq d<0.8$); L: large ($d\geq 0.8$)。

4　讨论

描述统计数据表明,四级网考的综合部分相对较难,考生可能不适应这部分所包含的跟读题型,也可能不习惯用键盘输入完成听写答题,还可能因为打字速度影响了其写作部分的表现。四级网考和四级纸笔考试之间的相关性分析表明,网考具有较好的同时效度,而且具有较理想的单项内部相关。相关性数据分析还表明,网考所测试的能力结构与纸笔考试有差异,各单项所测能力与纸笔考试不能完全对应。

问卷调查数据分析表明,考生对四级网考整体及各单项的有效性、试卷构成和答题时间的评价较好,网考适应程度和答题紧张程度也是比较理想的中等程度。考生最满意的是阅读、写作和语法结构部分,其次是听力理解部分。但是,考生认为四级网考整体偏难,其中听写和跟读部分的难度最大,考生对这两个单项的评价相对偏低,而且看法差异较大。对比组数据分析表明,西部地区考生对网考难度的评价显著大于其他地区考生,但是他们对网考有效性的评价却高于其他地区;地区组对于网考其它方面的评价没有显著差异。除了紧张程度外,不同的语言水平和计算机熟悉程度组对网考各方面评价均有统计意义上的显著差异。数据显示,语言能力或计算机熟悉程度越高的考生对网考有效性、试卷结构、适应程度等方面的评价越好,对网考难度评价越低,紧张程度评价也越低。

影响四级网考总分的两因素方差分析数据表明,地区与计算机熟悉程度对网考总分的影响不产生交互作用,但是两者分别对网考总分有统计意义上的显著影响。结合影响力数据分析表明,地区经济和社会发展程度因素对网考成绩的实际影响程度较小,计算机熟悉程度对网考成绩有一定程度的实际影响,特别是熟悉程度高组和低组之间的成绩差异影响较大。

5　结语

本研究初步验证了四级网考具有较理想的效标关联效度,考生对网考各方面总体评价较好,影响考生评价的因素包括语言水平和计算机熟悉程度,计算机熟

悉程度对考试成绩产生了较大的影响。从问卷第90题所收集的反馈信息来看,计算机熟悉程度对网考总分产生影响的一个原因是考生对计算机上答题的操作不熟悉,如不少考生认为听写和跟读的答题操作难度较大,也有考生反映不习惯在计算机上写作或阅读。因此,如果进一步完善考试系统,可降低计算机熟悉程度对考试成绩的影响。2008年12月试点考试后,全国大学英语四、六级考试委员会对网考的考生操作系统进行了改进,提高了界面的友好程度,如听写测试中增加了暂停键,跟读测试中新增了重录键,阅读测试中增加了标记笔功能,写作测试中增加了恢复、复制等操作键等。计算机熟悉程度因素对网考成绩产生较大影响的另一原因是,本研究的考生样本考前只参加过一次上机操练。目前网考的系统设计已趋完善,考生在考前可以进入四、六级网考的专用网站(www.ccets.org)模拟操练四级网考的各种题型。此外,随着基于计算机和课堂的新教学模式在大学英语教学中的进一步推广,大学生对网考的适应程度也一定会不断提高。

当然,四级网考的效度还需进一步论证,如网考所测的能力与纸笔考试有何差异,网考是否有助于提高测试的准确度,主观题的评分效率和信度如何,对教学的反拨作用如何等。此外,大学英语四、六级网考的长远发展方向是计算机自适应性考试,在这方面我国学者虽有所探索和研究(如何莲珍,2004;曾用强,2002;朱正才,2002),但尚未有付诸实践的大规模考试项目。总之,效度研究是语言测试研究中一个永恒的主题,是一个持续不断地收集各类效度证据的过程,是从不同侧面论证考试成绩是否准确反映了我们所要测试的能力的过程(Weir, 2005)。因此,四、六级网考的效度研究也将随着项目的推进而不断深入。 □

参 考 文 献

[1] Breland, H., Lee, Y. & Muraki, E. Comparability of TOEFL CBT writing prompts: Response mode analyses (TOEFL Research Report[R]. No. RR-75). Princeton, NJ: ETS. 2004.

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

[3] Eignor, D., Taylor, C., Kirsch, I. & Jamieson, J. Development of a scale for assessing the level of computer familiarity of TOEFL examinees(TOEFL Research Report. No. RR-60)[R]. Princeton, NJ: ETS. 1998.

[4] Kirsch, I., Jamieson, J., Taylor, C. & Eignor, D. Computer familiarity among TOEFL examinees (TOEFL Research Report. No. RR-59)[R]. Princeton, NJ: ETS. 1998.

[5] Messick, S. Validity [A]. In Linn, R. L. (ed.), *Educational Measurement*(3rd ed.). NY: Macmillan, 1989:13 - 103.

[6] Messick, S. Validity and washback in language testing[J]. *Language Testing*, 1996, 13:241 - 256.

[7] Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. The relationship between computer familiarity and performance on computer-based TOEFL test tasks(TOEFL Research Report. No. RR-61)[R]. Princeton, NJ: ETS, 1998.

[8] Weir, C. J. *Language Testing and Validation: An Evidence-based Approach*[M]. NY: Palgrave Macmillan, 2005.

[9] Weir, C. J., O'Sullivan, B., Jin, Y. & Bax, S. Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS writing component: effects and impact[R]. IELTS Research Reports. IELTS Australia and British Council, 2007, 7:311 - 347.

[10] Wolfe, E. W. & Manalo, J. R. An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study(TOEFL Research Report. No. RR-72)[R]. Princeton, NJ: ETS. 2005.

[11] 陈慧麟. 基于纸笔的语言测试和基于计算机的语言测试之间对等性验证模式初探[J]. 外语界, 2009/3:73 - 80.

[12] 何莲珍. 计算机化的认知适应性测试[M]. 杭州:浙江大学出版社, 2004.

[13] 教育部高等教育司. 大学英语课程教学要求[M]. 上海:上海外语教育出版社, 2007.

[14] 金艳, 吴江. 大学英语四、六级网考的设计原则[J]. 外语界, 2009/4: 61 - 68.

[15] 李清华. 基于纸笔的语言测试与基于计算机的语言测试的等效研究综述[J]. 外语界, 2006/4:73 - 78.

[16] 李清华, 孔文. 基于计算机的语言测试及其效度验证[J]. 外语界, 2009/3:66 - 72, 96.

[17] 王勇旗. 计算机化语言—新测试形式带来的思考[J]. 吉林省教育学院学报, 2008/5:28 - 29.

[18] 曾用强. 个性化自适应性测试探索[J]. 外语教学与研究, 2002/4:278 - 282.

[19] 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.

A Preliminary Study of the Validity of the Internet-Based CET-4
——Factors Affecting Test-takers' Perception of and Performance on the Test

JIN Yan, *WU Jiang*

(School of Foreign Languages, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Information technology was introduced into the field of language testing in the late 1990s. Since then, language testers have been concerned that the use of technology may bring construct-irrelevant variance to the measurement of test-takers' language proficiency, resulting in test bias and unfairness. As part of the validation study of the Internet-Based CET-4, this paper undertakes to analyze the factors affecting test-takers' perception of and performance on the test. The analysis of the questionnaire data indicates that language proficiency and computer familiarity significantly affected test-takers' perception of the test; however, no statistically significant differences were identified among groups of test-takers from regions at different levels of economic and social development. The analysis of the test scores evidences that the effect of computer familiarity on test performance is of both statistical and practical significance. The level of regional economic and social development did not have a practically significant effect on test-takers' performance on the test.

Key words: IB CET-4; Test-taker evaluation; Test performance; Factors; Effect size