

中国学生英语语用能力的测试

广东外语外贸大学 刘建达

提要:本研究用定量和定性等实证的方法研究了书面话语填充、选择性话语填充和自我评估这三种测试中国英语学习者语用能力的方法。结果表明,三种方法都具有较高的信度和较好的效度,不同的测试方法对考生的成绩及考试中的思维过程有不同的影响。另外,参加此次研究的学生的语用能力并没有随其语言能力的提高而提高。这暗示着老师在课堂教学中应更多地教授学生英语语用知识,给学生更多的语用知识输入,以达到提高学生英语语用能力的目的。

关键词:语用能力、语言测试、话语填充、自我评估

[中图分类号] H319

[文献标识码] A

[文章编号] 1000-0429(2006)04-0259-7

1. 引言

随着外语教育中引入交际能力这一概念,人们越来越注重培养外语学习者的外语交际能力,交际教学法等意念大纲教学法也应运而生(Cohen & Olshtain 1981)。虽然交际教学法已普遍用于外语教学中,但作为检验教学效果的测试方法还大都停留在原来的语法—翻译法和结构测量法,与交际教学法配套的交际测试方法还处于探索阶段。语用能力是交际能力的一部分(Bachman 1990; Bachman & Palmer 1996),目前有关如何在课堂上教授语用知识以及外语学习者是如何学习语用知识的研究有所增加(见 Rose & Kasper 2001),然而,如何才能有效测试外语学习者语用能力的研究却非常少。已有的一些研究都集中于验证一些测试语用能力的方法。Farhady (1980)和 Shimazu (1989)研究了运用选择性话语填充法(Multiple-choice Discourse Completion Test, MDCT)测试语用能力的信度和效度。Hudson et al. (1992; 1995)探索了测试语用能力的其它一些方法,包括书面话语填充(Written Discourse Completion Test, WDCT)、听说话语填充(Listening Oral Discourse Completion Tasks, ODCT)、角色扮演(Discourse role-play Tasks, DRPT)、自我评估(Discourse Self-assessment Tasks, DSAT)以及角色扮演自我评估

(Role-play Self-assessment, RPSA)等方法。但他们没有进一步对所设计的测试方法进行验证。后来, Yamashita (1996, 1997)以及 Enoch & Yoshitake-Strain (1999)用定量方法对 Hudson 等设计的测试方法进行了研究,而 Yoshitake-Strain (1997)则用定性方法对这几种测试方法进行了研究。Roever (2001)探讨了网络化语用能力测试的信度和效度。这些研究的主要发现可归纳成以下几点。第一, ODCT, DRPT, DSAT 和 RPSA 都具有较高的信度和可接受的效度。第二,虽然 Farhady 和 Shimazu 的研究显示 MDCT 具有较好的效度和信度,但 Yamashita 和 Yoshitake-Strain 的研究表明 MDCT 和 WDCT 存在很多问题。第三,考生接触目标语的程度对考生的表现有显著影响。第四,考生的母语文化对考生的表现有很大影响。

这些研究的参加者都是美国日语学习者和日本英语学习者。由于文化在语用能力测试中起着很重要的作用(Brown 2001; Hudson 2001),适合日本文化或美国文化的语用能力测试题不一定适合其它文化。与此同时,虽然外语学习者的外语交际能力(包括语用能力)已经写进我国外语教学的教学大纲之中,但现有的一些考试却未能将学习者的语用能力涵盖在内,其原因之一可能是目前还没有一种得到验证的测试学习者语用能力的方法。因此,很有必要就测试中国外语学习者语用能力的方法进行研究。

2. 研究方法

2.1 研究问题

本研究首先设计三套试卷,分别采用三种不同的测试方法(WDCT, MDCT, 和 DSAT)来测试学生道歉这一言语行为能力。然后依据 Messick (1989) 的效度理论对这三种测试方法进行效度验证。具体回答以下几个问题:

- 1) 三种测试方法的信度和效度如何?
- 2) 测试方法对考试结果是否有影响?
- 3) 考试结果是否受考生的外语水平的影响?

2.2 受试对象

在试卷设计阶段,总共有 108 位中国大学生(以下简称 NNS)和 22 位以英语为本族语的参加者(以下简称 NS)参与了本研究。在最后资料收集阶段,共有 99 名 NNS 参加了此研究,其中 37 名大学英语专业三年级学生,62 名大学英语专业一年级学生。三年级学生被分在高水平组,一年级学生在低水平组。为了检验两组学生的英语水平是否有显著差异,两组学生都参加了一次 TOEFL 考试(试卷获美国 ETS 授权使用)。结果显示两个水平组在听力($t = 6.344, p < .01$)、结构($t = 4.652, p < .01$)、阅读($t = 6.328, p < .01$)三个部分和总分($t = 8.642, p < .01$)上都有显著意义的区别,表明两个水平组的英语水平有显著区别。

2.3 试验材料

本研究所采用的试卷包括 12 个有关道歉的情景。试卷的设计经过了情景选择、情景可能性筛选、元语用调查、WDCT 试测和 MDCT 选项设计等几个步骤。

第一步是情景选择(Groves 1996; Rose & Kwarfan 2001)。笔者首先设计了一份收集情景的问卷,要求参加者写出最近发生在他们身上或者他们所目睹的发生在别人身上的十个有关请求的情景。总共有 30 位参加者交回了问卷。笔者从收集到的总共约 300 个情景中选出 46 个情景,然后对这 46 个情景在不改变原来意思的原则下进行了改写。

第二步是情景可能性调查。笔者用选出的 46 个情景设计了一份问卷。问卷采用 5 点 Likert 度量表,5 表示很可能发生,1 表示不可能发生,要求参加者在 1 至 5 中选一个点来表示每个情景发生的可能性的。总共 15 位参加者寄回了问卷。笔者从中选出平均分最高(即参加者认为最可能发生)的 30 个情景。

第三步为元语用调查。为了确定这 30 个情景每个情景的社会语用变量,笔者同样使用 5 点 Likert 度量表设计了另外一份问卷。问卷首先详细介绍了三个

社会语用变量:地位、熟悉程度和严重性。每个情景有三个问题,分别调查这三个变量。问卷包括中文和英文两个版本。为了使两个版本保持一致,笔者使用回译(back translation)(Fowler 1993; Kasper & Rose 2002)的办法来设计这两个版本的问卷。然后,中文版问卷发给 30 个 NNS,英文版问卷寄给了 30 个 NS。所有 NNS 都返回了问卷,但只有 15 个 NS 寄回问卷。根据所收集的资料,笔者确定一个筛选原则:对于每个情景的每个问题,首先 NNS 之间必须有 70% 达成一致意见,其次 NNS 和 NS 之间也必须有 70% 持相同意见。如果任何一个情景的任何一个问题不能达到这样的要求,这个情景就被放弃不用。最后,只有 15 个情景达到要求。

第四步是试测。笔者用这 15 个情景再设计了一份问卷,要求参加者根据情景描述,将他们在这些情景中可能会说的话写下来。总共有 34 个 NNS 和 6 个 NS 返回了问卷。然后,笔者邀请两位 NS 对收集的 34 份 NNS 的答卷进行评分。在评卷之前,笔者对两位评分者进行过培训。

最后是 MDCT 选项设计。笔者将收集的资料全部输入电脑,用定性分析软件 Winmax Pro (Kuchartz 1998)进行分析。首先,把所有 NS 的回答标记为“答案”,而把所有两位 NS 评分者认为不妥当的 NNS 的回答标记为“干扰项”。然后,笔者设计了一份问卷,每个情景有 4 到 8 个选项。15 位 NS 返回了此问卷。根据问卷结果,每个情景选出三个 NS 意见最一致的选项,其中来自 NS 的那个为答案,其余两个来自 NNS 的选项为干扰项。这样一份新的试卷(共 15 个情景)合成了,每个情景包括三个选项。接着,为了验证答案的准确性,笔者邀请 5 个 NS 选择每个情景的答案,结果他们在三个题目的答案上有不同的意见,笔者遂把那三个情景删除,这样整份试卷只剩下 12 个情景。最后,笔者对这份包括 12 个情景的试卷进行了试测。并且根据项目反应理论(IRT),利用 IRT 分析软件 Winsteps (Linacre 2003)对试测的结果进行了分析。最后,依据分析结果,笔者对试题的选项进行了一些必要的修改和验证。

至此,本研究所使用的三份试卷的设计完成了。12 个包含道歉这一言语行为的情景构成 WDCT 试卷,要求考生根据每个情景的描述把他们在该情景下会说的话写下来。DSAT 试卷使用同样 12 个情景,考生需判断自己在此情景中有多大把握能做出恰当的回答,然后在 1 至 5 中选出一个点来表示自我评估的结果。MDCT 试卷也使用同样的 12 个情景,不同的是每个情

景配有上述描述的三个选项,要求考生为每个情景选出一个最恰当的选项。

2.4 评分规则

对于 DSAT 部分,考生自我评估的分数就是最后的结果;对于 MDCT 部分,每个正确的选择得 5 分,错误的得 0 分;对于 WDCT 部分,笔者依据以前的研究(如 Hudson et al 1995)并结合此次研究的特点设计了一套评分规则。此规则采取整体评分法(holistic scoring),分 1 至 5 五个级别。每个级别在言语行为、措辞的准确性、信息量以及表达的得体性四个方面做了具体规定。规则制定后,笔者邀请两位 NS 利用此评分规则试评三份答卷。结果显示两位评卷人之间的相关系数达到 .901,说明此规则基本可行。

2.5 评卷

对于 WDCT 部分,为避免因书写问题而影响评分,收集的资料全部不作任何改动地输入电脑。然后把电脑打印出来的材料给两位 NS 进行评卷。在评卷之前,笔者对这两位评卷人进行了培训。首先笔者向他们详细说明了评分的标准和注意事项,并且邀请他们进行试评。在确信他们完全掌握了评分标准后,让他们在一份专门设计的评卷单上采用 1 至 5 五个级别对每份答卷进行评分。每位考生 WDCT 部分最后的得分是两位评卷人所给分数的平均值。

3. 结果与讨论

所有统计都是用 SPSS11.0 进行的。表 1 列出了全部试卷的描述统计数据。从中可以看出,考生的自我评估成绩最高,需要老师评卷的 WDCT 分数最低。

表 1. 全部试卷的描述统计数据

标准差	人数	总分*	最小分	最大分	平均分	标准差
TOEFL	99	677	413.33	570.00	490.67	33.47
DSAT	99	60	28	57	41.41	6.11
MDCT	99	60	5	55	33.94	11.74
WDCT	99	60	12	42	26.65	6.00

(* 总分 = 试卷的满分)

3.1 试卷的验证

本研究依据 Messick (1989) 的效度理论来验证这三份试卷。Messick 的理论包括六个方面的分析:试卷内容分析、相关分析、过程分析、组间分析、考试条件分析以及考试后果分析。虽然 Messick 把这六个方面看

作一个整体,但验证一份试卷并不一定总要从全部六个方面来进行,根据具体考试的特点,可以集中就某几个方面进行验证(Messick 1996)。本研究拟从内容分析、相关分析和过程分析三方面来验证本研究所设计的试卷。

表 2 为各部分的内部一致性信度()、标准误差和评分者间信度。从表中可看出,DSAT、MDCT 和 WDCT 的信度均在 .88 以上,表明三份试卷具有可以接受的信度。这个结果与已有的一些研究结果(Enochs & Yoshitake-Strain 1999; Roever 2001; Yamashita 1996)相似。但值得注意的是,此研究中的 MDCT 的信度达到 .88,大大高于已有几个研究的结果。这很可能得益于本研究中试卷的设计步骤。本研究使用的情景不是照搬别人的情景,而是通过情景选择、可能性调查、原语用调查等多个步骤设计出来的。所有的情景都出自学生生活,干扰项也是来自学生的答卷。此外,评分者间信度达到 .903,说明两位评分者具有较高的评分一致性。

表 2. 试卷信度和评分者间信度

	人数	试卷信度()	标准误差	评分者间信度
DSAT	99	.9181	2.44	
MDCT	99	.8812	5.56	
WDCT	99	.9195	3.08	.903

内容效度主要体现在两方面:内容的相关性和内容的覆盖范围(Messick 1996)。本研究所使用试卷是经过前述多个步骤设计出来的,其中邀请了专家、NS 以及 NNS 参与调查。专家普遍认为,试卷测试的是考生的语用能力。在试测中,笔者还邀请了一些 NS 和 NNS 利用有声思维等方法对试卷进行了定性研究,分析结果也显示,试卷测试的是考生的语用能力。这些都确保了试卷内容的相关性。此外,本研究还依据 Hudson 等(1995)的模式来选择情景。首先笔者按照交际双方的地位(power)、熟悉程度(distance)和道歉情景的严重性(severity)这三个变量排列出了 12 种情况(见表 3)。与 Hudson 等(1995)不同的是,对于地位这个变量,Hudson 他们只用了两个特征(高和低),而本研究还加入了地位平等这一特征。笔者希望能从收集的近 300 个情景中选出适合表 3 所列每种情况的情景,但最后经过多次试测和验证之后,未能得到 3 种情况的情景。

表 3. 情景和变量分布表

		分布												
情景		1	2	3	4	5	6	7	8	9	10	11	12	
	地位 P	+	+	+	+	=	=	=	=	-	-	-	-	
变量	熟悉程度 D	+	-	+	-	+	-	+	-	+	-	+	-	
	严重性 R	+	-	-	+	+	-	-	+	+	-	-	+	
有效项目数		0	1	1	2	2	2	1	1	1	1	0	0	12

(+ :听者比说话者地位高,听者和说话者彼此熟悉,道歉情景严重性高;- :听者比说话者地位低,听者和说话者彼此不熟悉,道歉情景严重性低;= :听者比说话者地位平等。)

表 4 为三种测试方法和水平考试之间的相关系数和决定系数。决定系数可以帮助解释两者之间有多大的相似性(Brown 1996)。从表中可以看出,TOEFL 与 WDCT 和 DSAT 没有显著意义的相关,决定系数也只有 .002 和 .001,说明 DSAT 与 WDCT 和 TOEFL 之间只有 0.2% 和 0.1% 的相似性。虽然 TOEFL 和 MDCT 有显著意义相关,但是决定系数只有 .05,两者之间只有 5% 的相似性。这说明 TOEFL 和三种语用能力测试方法可能测试了不同的能力。三种语用能力测试方法之间都有着显著意义的相关($p < .01$),三者之间的决定系数介于 .267 至 .469 之间,表示各种方法之间有 27% 至 47% 的相似性。虽然相似性不是很高,但可以在一定程度上说明此研究中的三种语用能力测试方法测试了相似的能力,即学生的语用能力。

表 4. 三种测试方法和水平考试之间的相关系数和决定系数

	TOEFL	WDCT	DSAT
TOEFL	.		
WDCT	.048(.002)	.	
DSAT	.033(.001)	.685**(.469)	.
MDCT	.226**(.050)	.675**(.456)	.517**(.267)

(** : 相关有显著意义, $p < .01$; 括号里为决定系数。)

因素分析也可以用来研究试题的结构效度。旨在测试同一能力的试题应该属于同一因素,而测试不同能力的试题应属于不同的因素(Brown 2001)。表 5 为各种试卷的因素分析结果。经过最大变异旋转后,两个特征值大于 1 的因素被提取出来。可以看到,测试语用能力的方法(WDCT, DSAT 和 MDCT)在因素 1 有较高的负荷值,而 TOEFL 水平考试在因素 2 上有很高的负荷值。共性方差也表明 TOEFL 考试已解释的方差达到 97.1%, WDCT, DSAT 和 MDCT 也分别达到

84.2%, 74% 和 74.4%。全部已解释的方差为 82.4%。以此,可以很清楚看到第一个因素就是语用能力因素,第二个因素为语言水平因素。这表明,WDCT, DSAT 和 MDCT 测试的是相似的能力,而 TOEFL 测试的是另外一种能力。

表 5. 因素分析结果

	成 分		共性方差
	1	2	
TOEFL	.041	.985	.971
WDCT	.918	-.004	.842
DSAT	.857	-.068	.740
MDCT	.815	.281	.744
方差比例	.561	.263	总共: .824

过程分析可以帮助了解考生考试中的思维过程(Messick 1989)。通过对考生思维过程的分析,可以了解考生做题的思维是否与试题所设计的相似,从而进一步验证试题的构念效度(Messick 1996)。本研究通过有声思维方法收集了部分考生考试中的思维过程。分析结果表明,考生在考试中的思维是围绕道歉这言语行为进行的,因而试题测试了学生的言语行为能力。

以上分析表明,本研究设计的试卷具有较好的内容和构念效度。信度分析也表明试卷具有较好的信度。一般认为(李筱菊 2001),选择填空试题具有较好的信度。但此研究及其它一些研究(如 Enochs & Yoshitake-Strain 1999; Yamashita 1996)表明,在语用能力测试中,MDCT 的信度没有 WDCT 和 DSAT 等方法高。影响试卷信度的因素之一是试题量(Brown 1996),本研究的试卷只包括 12 个题目,因此,增加题量也许能更好地提高试题的信度。

3.2 测试方法的影响

Levene 检验表明三种方法之间的方差符合方差分

析的要求。MDCT、WDCT、和 DSAT 三种方法之间的方差分析显示,三种方法存在显著差异($F = 114.803$, $p < .01$)。随后的 Scheffé 检验分析(表 6)也表明,三者相互之间也有显著意义的区别。同时,Scheffé 检验分析还显示出三个同质组(homogenous group),分别对应三种测试方法(表 7)。这些都表明测试方法对考生的成绩有显著影响。

表 6. Scheffé 检验结果

方法组(I)	方法组(J)	均差(I - J)	标准误差	p 值
WDCT	DSA	- 14.2500 *	.94109	.000
	MDCT	- 6.6596 *	.94109	.000
DSA	WDCT	14.2500 *	.94109	.000
	MDCT	7.5904 *	.94109	.000

(*) 均值差异有显著意义, $p < .01$ ()

表 7. 同质组检验

方法组	人数	Subset for alpha = .01		
		1	2	3
WDCT	99	29.2979		
MDCT	99		35.9574	
DSA	99			43.5479
p 值		1.000	1.000	1.000

此外,定性研究也表明,考生在三种测试方法中的思维过程也不同。在 WDCT 中,考生在阅读完情景之后,集中分析情景的社会语用变量。然而,整个过程却主要集中在生成答案上,包括计划答案、生成最初答案、修改答案、生成最后答案几个步骤。重读(re-reading)在整个思维过程都有发生。在 DSAT 中,分析情景的思维过程与 WDCT 大致相同。两者的不同之处主要在于生成答案阶段。由于试题的特性,在 DSAT 中,考生要评估自己的能力,而在 WDCT 中考生主要是生成答案。当然,笔者也发现,在 DSAT 中考生有时也会试着给出一个答案,但是主要还是集中分析情景,而重读也主要发生在分析情景阶段,在最后评估阶段很少发生。考生对自己能力不太肯定时,大都采取猜测的策略,而且最后自我评估的分数往往较低。MDCT 的思维过程则与 WDCT 和 DSAT 大不相同,考生在读完情景之后会马上阅读所有三个选项或者只读第一个选项。此外,与 WDCT 和 DSAT 不同,在 MDCT 中考生主要集中分析和权衡选项,而不是情景。社会语用变量分析一般是在分析第一个选项时进行。即使考生认为第一个选项正确,他们都会继续考虑第二个选项。

但是如果考生判定第一和第二个选项都是错误的,则有些不会认真去考虑第三个选项,而是直接判定第三个选项就是答案。如果考生对三个选项都没把握,他们往往会利用一些考试技巧,如排除法。还有,与 WDCT 和 DSAT 不同,在 MDCT 中,重读特别多,而且集中在选项上,考生很少重读情景。在本研究总共发现的 174 次重读中,MDCT 有 138 次(79%),而 DSAT 和 WDCT 则分别只有 12 次(7%)和 24 次(14%)。笔者还发现,考生在 MDCT 中常用“也许”、“大概”之类表示不确定的词,而在 WDCT 和 DSAT 中却很少见这些词。

方差分析和过程分析均表明,测试方法不仅对考生的成绩有显著影响,而且对考生考试中的思维过程也有不同的影响。

3.3 考生英语水平的影响

Levene 检验表明水平组之间的方差符合方差分析的要求。方差分析显示,高水平的学生比低水平学生在 TOEFL ($F = 74.767$, $p < .01$) 上表现明显要好,然而,在测试语用能力的三份试卷 MDCT ($F = 2.215$, $p = .108$)、WDCT ($F = .606$, $p = .437$) 和 DSAT ($F = .023$, $p = .878$) 上都没有显著差异。低水平学生在 DSAT 和 WDCT 上比高水平学生的平均分甚至还要高一些。这说明两个水平组的学生在英语水平上存在显著差异,但是他们的语用能力并没有随着他们水平的提高而显著提高。这个结果与以前一些研究(Hill 1997; Roever 2001; Yamashita 1996)的结果不尽相同。这些研究的结果显示高水平学生比低水平学生,接触目标语和目标语文化多的学生比接触少的学生在语用能力测试中表现要好。这些研究与本研究差异的原因可能有三。第一,这些研究的参加者都有一定的在英语国家生活的经历,而笔者对参加本次研究的学生的调查显示他们都没有去过英语国家。第二,这些研究的参加者大都曾经直接接触过英语国家的文化,而本研究的学生无此直接接触,他们都是通过课堂或者书本接触英语文化的。笔者的调查还显示,本研究中高水平 and 低水平学生在课堂上接受语用能力教育的机会也大致相同。两个水平组的考生大都报告说他们老师上课时只是偶尔谈及语用知识的问题。第三,这些研究的参加者来自社会不同领域和阶层,而本研究的学生全部是大学英语专业学生,教育背景相似。

从上可以看出,由于没有太多与目标语和目标语文化直接接触的环境,学生的语用能力并不会随他们语言能力的提高而提高。要提高学生的语用能力,课堂语用知识教学是个值得推荐的办法(Rose 1997a,

1997b; Rose & Kasper 2001; Rose & Kwarfan 2001)。Ellis(1994)就认为,要提高学生的语用能力,为学生提供足够和恰当输入是重要的。课堂上语用知识的输入主要来自老师的谈话和课本(Hill 1997)。因此,要提高学生的英语语用能力,教师在课堂上应该多给学生讲授一些语用知识,此外,还需在教材的编写和教师的培训等方面做出努力。

4. 结论

本研究用定量和定性等实证的方法研究了WDCT、MDCT和DSAT这三种测试中国英语学习者的语用能力的方法。结果表明,这三种方法都具有较高的信度和较好的效度。不同的测试方法对考生的成绩和考试中的思维过程都有着不同的影响。由于参加本次研究的学生没有与目标语和目标语文化直接接触的经历,他们的语用能力并没有随着他们语言能力的提高而提高。这就要求教师在课堂教学中更多地教授学生英语语用知识,以给学生更多的语用知识输入,从而达到提高学生英语语用能力的目的。

必须指出的是,外语学习者语用能力的测试,正如Hudson(2001)所说,目前只是处于研究阶段。虽然本研究的结果显示这三种测试方法均有较好的信度和效度,但在将它们用于大型考试之前还需做大量进一步的研究工作。本研究只涉及道歉这一种言语行为。语用能力还包括很多其它方面的能力,同样的结果是否会出现其它的语用能力中,还有待进一步研究。此外,本次研究只研究了三种测试语用能力的方法,事实上,Hudson等(1995)开发了六种测试语用能力的方法,为了验证更多的语用能力测试方法,有必要在中国英语学习环境下,对其它测试方法做进一步的验证研究。

参考文献

- Bachman, L. F. 1990. *Fundamental Considerations in Language Testing* [M]. Oxford: OUP.
- Bachman, L. F. & A. S. Palmer. 1996. *Language Testing in Practice* [M]. Oxford: OUP.
- Brown, J. D. 1996. *Testing in Language Programs* [M]. Upper Saddle River, NJ: Prentice Hall Regents.
- Brown, J. D. 2001. Pragmatics tests: Different purposes, different tests [A]. In K. R. Rose & G. Kasper (eds.). *Pragmatics in Language Teaching* [C]. New York: CUP.
- Brown, J. D. 2001. *What is an eigenvalue?* Retrieved . 264 .
- August 3, 2004, from the World Wide Web: <http://www.jalt.org/test/bro-10.htm>
- Cohen, A. D. & E. Olshtain. 1981. Developing a measure of sociocultural competence: The case of apology [J]. *Language Learning* 31/1:113-134.
- Ellis, R. 1994. *The Study of Second Language Acquisition* [M]. Oxford: OUP.
- Enochs, K. & S. Yoshitake-Strain. 1999. Evaluating six measures of EFL learners' pragmatic competence [J]. *JALT Journal* 21/1:29-50.
- Farhady, H. 1980. *Justification, Development and Validation of Functional Language Testing* [D]. Unpublished doctoral dissertation. Los Angeles: University of California.
- Fowler, J. F., Jr. 1993. *Survey Research Methods* [M] (2nd edn.). Newbury Park, CA: Sage Publications.
- Groves, R. 1996. How do we know what we think they think is really what they think? [A]. In N. Schwarz & S. Sudman (eds.). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* [C]. San Francisco: Jossey-Bass.
- Hill, T. 1997. *The Development of Pragmatic Competence in an EFL Context* [D]. Unpublished doctoral dissertation. Tokyo: Temple University.
- Hudson, T. 2001. Indicators for pragmatic instruction: Some quantitative tools [A]. In K. R. Rose & G. Kasper (eds.). *Pragmatics in Language Teaching* [C]. New York: CUP.
- Hudson, T., E. Detmer & J. D. Brown. 1992. *A Framework for Testing Cross-cultural Pragmatics* [M]. Honolulu: Second Language Teaching & Curriculum Center University of Hawai'i at Manoa.
- Hudson, T., E. Detmer & J. D. Brown. 1995. *Developing Prototypic Measures of Cross-cultural Pragmatics* [M]. Honolulu: Second Language Teaching and Curriculum Center, University of Hawai'i at Manoa.
- Kasper, G. & K. R. Rose. 2002. *Pragmatic Development in a Second Language* [M]. Michigan: Blackwell.
- Kuchartz, U. 1998. *WinMAX: Scientific Text Analysis for the Social Sciences* [M]. Berlin: BBS.

- Linacre, J. M. 2003. *A User's Guide to WINSTEPS: Rasch-model Computer Programs* [M]. Chicago: MESA Press.
- Liu, J. 2006. Measuring Interlanguage Pragmatic Knowledge of EFL Learners [M]. Frankfurt am Main: Peter Lang.
- Messick, S. 1989. Validity [A]. In R. Linn (ed.). *Educational Measurement* [C]. New York: Macmillan.
- Messick, S. 1996. Validity and washback in language testing [J]. *Language Testing* 13/3: 241-256.
- Roever, C. 2001. *A Web-based Test of Interlanguage Pragmatic Knowledge* [D]. Unpublished doctoral dissertation. Hawai'i: University of Hawai'i.
- Rose, K. R. 1997a. Pragmatics in teacher education for nonnative-speaking teachers: A consciousness-raising approach [J]. *Language, Culture and Curriculum* 10/2: 125-138.
- Rose, K. R. 1997b. Pragmatics in the classroom: Theoretical concerns and practical possibilities [A]. In L. F. Bouton (ed.). *Pragmatics and Language Learning. Monograph Series* [C] (Vol. 8). Urbana-Champaign: University of Illinois.
- Rose, K. R. & G. Kasper (eds.). 2001. *Pragmatics in Language Teaching* [C]. New York: CUP.
- Rose, K. R. & C. Ng Kwai-fan. 2001. Inductive and deductive teaching of compliments and compliment responses [A]. In K. R. Rose & G. Kasper (eds.). *Pragmatics in Language Teaching* [C]. New York: CUP.
- Shimazu, Y. M. 1989. *Construction and Concurrent Validation of a Written Pragmatic Competence Test of English as a Second Language* [D]. Unpublished doctoral dissertation. San Francisco: University of San Francisco.
- Yamashita, S. O. 1996. *Six Measures of JSL Pragmatics* [M]. Honolulu: Second Language Teaching & Curriculum Center of University of Hawai'i at Manoa.
- Yamashita, S. O. 1997. Self-assessment and roleplay methods of measuring cross-cultural pragmatics [A]. In L. F. Bouton (ed.). *Pragmatics and Language Learning* [C] (Vol. 8). Urbana-Champaign: University of Illinois.
- Yoshitake-Strain, S. 1997. *Measuring Interlanguage Pragmatic Competence of Japanese Students of English as a Foreign Language: A Multi-test Framework Evaluation* [D]. Unpublished doctoral dissertation. Novata, CA: Columbia Pacific University.
- 李筱菊, 2001, 《语言测试科学与艺术》(第2版) [M]。长沙: 湖南教育出版社。

收稿日期: 2005—6—17;

本刊修订稿, 2006—5—11

通讯地址: 510420 广东外语外贸大学外国语言学及应用语言学研究中心

Abstracts of major papers in this issue

A study of Chinese university students' use of English articles, by *CAI Jinting & WU Yi'an* (English Dept., The PLA Univ. of Foreign Languages, Luoyang 471003, P. R. China), p. 243

By analyzing writings and cloze data, this paper investigated the relationship between Chinese university students' use of English articles and NP's three features, including referent specificity, hearer's knowledge and countability. Besides, task types and the absence of articles in Chinese were also taken into account. Three findings were obtained: (1) All of NP's three features had significant effects on the appearance and accuracy of English articles. (2) The subjects used the articles obviously better in writing than in cloze. (3) They often overgeneralized zero article, probably due to the joint influence of L1 and the Principle of Economy.

EFL learners' acquisition of English Adjective-Noun collocations — A quantitative study, by *ZHANG Wenzhong & CHEN Shuichi* (School of Foreign Languages, Nankai Univ., Tianjin 300071, P. R. China), p. 251

The present study attempts to provide a profile of Chinese EFL learners' acquisition of English adjective-noun collocations. Using the cross-sectional research design, the researchers designed a test to examine the reception and production of three different types of English adjective-noun collocations by three groups of subjects at different proficiency levels. The results show that (1) in general the subjects do not appear to have a good command of English adjective-noun collocations, (2) there is a gap between and within different groups of subjects in their receptive knowledge of the adjective-noun collocations, (3) differences exist between and within different groups of subjects in their productive use of English adjective-noun collocations, and (4) the transformation from receptive knowledge into productive skills appears to be a slow and gradual process.

Measuring interlanguage pragmatic knowledge of Chinese EFL learners, by *LIU Jianda* (National Key Research Center for Linguistics and Applied Linguistics, Guangdong Univ. of Foreign Studies, Guangzhou 510420, P. R. China), p. 259

This study attempts to validate three test methods — Written Discourse Completion Test (WDCT), Multiple-choice Discourse Completion Test (MDCT), and Discourse Self-assessment Test (DST) — used to test the interlanguage pragmatic knowledge of Chinese EFL learners in the context of apologies. The development of the test papers using these three methods underwent various stages, from exemplar generation, likelihood investigation, metapragmatic assessment to multiple-choice option generation and pilot study. Results showed that all the three methods investigated in this study have measured learners' interlanguage pragmatic knowledge of English in the Chinese EFL context, specifically in terms of the speech act of apology. Reliability studies revealed that WDCT and DST were highly reliable and MDCT was also reasonably reliable. Different quantitative analyses showed that the tests tapped the intended construct and the test methods measured a similar construct. Analyses of verbal reports yielded results which supported the quantitative analyses and revealed that the construct-relevant knowledge was involved in the test takers' cognitive activities. Participants of higher level of English proficiency seemed not to have correspondingly higher interlanguage pragmatic ability. Cognitive processes involved and strategies the participants adopted in answering questions with different test methods appeared to differ. It was suggested that teachers should pay more attention to the teaching of interlanguage pragmatic knowledge in class, so that learners could have more such input and thus improve their interlanguage pragmatic competence.

Investigating Chinese learners' developmental feature of noun colligation — A corpus based study, by *SUN Haiyan & CHEN Yongjie* (College of Foreign Languages, Henan Normal Univ., Xinxiang 453007, P. R. China), p. 272

This paper investigates the developmental feature of Chinese-speaking learners' use of noun colligation on the basis of