

# 语言测试的反拨效应理论与实证研究\*

广东外语外贸大学 亓鲁霞

**提 要:**过去二十来年,语言测试反拨效应研究成果可观。理论研究涉及测试影响的属性、范围、层面和运行机制,还涉及其强度、长度、具体性、意图与性质,为实证研究提供了操作框架。实证研究反过来验证理论,主要发现有:反拨效应体现在教学的各个环节和层面,如教学内容、方法、教材编写等,正面负面影响兼而有之。相比之下,测试对学习效果的影响研究不多,两者之间的关系有待探明。此外,反拨效应的一些理论模型仍有欠缺,尚需完善,以增强解释力。

**关键词:**测试;反拨效应;涉考者

语言测试的反拨效应研究旨在探讨语言测试与教学之间的关系。对此关系,人们有截然不同的看法。有一种观点认为,如果考试实施得当,以考促教是既经济又奏效的好途径(Popham, 1987),许多考试的创设就是基于这个认识,如我国大学英语四六级考试的初衷主要是为了改善英语教学,提高教学质量(杨惠中 & Weir, 1998)。还有一种观点认为,考试会干扰和阻碍教学,不宜过分宣扬和滥用,应试教学害处多(王初明, 2011)。两种观点基于同一前提,均认同测试对教学有影响,即产生反拨效应。研究此效应是过去二十来年外语测试与教学界的热门课题,成果可观。本文从相关理论和实证研究两方面介绍、讨论和总结一些主要发现,以期获得启示与借鉴,促进外语测试、外语教学以及今后的同类研究。

## 一、反拨效应的理论探讨

反拨效应理论研究主要从三个角度展开:一是定义反拨效应的属性,二是界定它的各个维度,三是提出假设,构建模型,勾勒反拨效应范围、层面和运行机制。

对反拨效应属性的研究,是澄清研究的性质和目标。对属性的主要认识有,反拨效应是考试构念效度不可分割的一部份(Messick, 1989, 1996),或归属于超考试效度(李筱菊, 1997)。把考试所产生的影响提升到效度的高度来认识,意味着考试人员对反拨效应负有不可推卸的责任。在设计一个考试时,必需考虑到考试可能产生的影响,采取一切措施保证积极作用,减少负面影

响。美国著名语言测试专家 Bachman & Palmer (2010) 便持此观点。他们认为,设计一个考试首要考虑的问题,是考试结果的使用对涉考者(stakeholders)会产生怎样的影响和后果,怎样才能使他们从中受益。可见,反拨效应在测试领域受到高度重视。笔者完全赞同将反拨效应作为衡量语言测试工作最重要的指标,却不赞成将其纳入考试效度。理由是,用效度概括反拨效应易使问题简单化,误导考试使用者,以为考试对教学能起决定性作用,考试搞好了,教学也会跟着好转。实际情况是,测试对教学影响的大小和好坏,不全由考试决定,社会上和教学中的诸多因素与考试会发生交互作用,产生不同程度、不同性质的反拨效应,这点已被研究所证实。杨惠中、桂诗春(2007)认为,要想加强考试的正面反拨效应,减小负面影响,单靠考试人员是无法做到的,需要社会各方面的协同努力。

反拨效应维度的界定,是深入开展研究的前提。为此,Watanabe(2004)提出反拨效应的五个维度:具体性、强度、长度、意图、性质<sup>①</sup>。具体性指测试的某一方面或某种测试具体的反拨效应,比如一个考试如果包括听力测试,教学中听力就会受到重视。与此相对的是普遍性,指任何测试都具有的反拨效应,如促使学生加大学习投入。强度指测试对教学影响的程度有多大。长度指测试影响延续的时间。意图指考试的具体影响是设计者所期望的还是预料之外的。性质指反拨效应

\* 该论文得到 2010 年度国家社会科学基金项目 10BY031 的支持,特此致谢。

① 英文文献中有不同的词表达此概念,如 value (Watanabe, 2004) 和 direction (Green, 2006)。

是正面的还是负面的。在这五个维度中,性质最为重要,也受到最大关注,有关反拨效应的多数文章都会提及,而划分正、负面效应的标准却不甚明确(元鲁霞,2004)。研究一个考试的反拨效应,划分正面效应和负面影响,是一种价值判断,判断标准至关重要,但多数学者在讨论此问题时没有明确阐述所用标准。因此,反拨效应的性质,尚需更加广泛深入的讨论与研究,就判断标准达成共识。

提出假设、构建模型,是对反拨效应理论研究的不断推进。学者们在这方面提出不少很有见地的理论模式。Hughes (1993, in Bailey, 1996) 用三分法来界定反拨效应的各个层面,认为考试的影响涉及参与者(考生、教师、教育管理人员、教材编写者、教科书出版商等)、过程(大纲制定、教材编写、教学方法、学习策略等)和结果(学到的知识与技能)。Alderson & Wall (1993) 针对反拨效应提出了 15 个假设,从教和学的内容、方法、顺序、进度、程度、深度等方面阐释考试对教学的影响,包括影响的大小与范围。根据实证研究结果, Alderson & Hamp-Lyons (1996) 后来新增了一个假设:反拨效应的类型和强弱程度对不同的教师和学生有所不同。

基于三分法和 15 个假设, Bailey (1996) 构建了考试反拨效应模型,其基础框架取自 Hughes 的三分法,内容包含 Alderson & Wall 的假设涉及到的各个层面。Bailey 的主要贡献在于整合了各种概念与假设,揭示了反拨效应运行机制的复杂性。在她看来,测试影响并非总是从考试直接到教育系统的各个环节与层面,例如,考试通过教师对教学产生直接影响,同时也通过教材对教学产生间接影响。因此,考试对教学系统各个环节的影响是双向的,甚至是多向的。在此过程中,其他因素(如教材编写者的教学理念)也起作用,很难把它们与考试本身产生的影响区分开来。

Bachman & Palmer (1996: 30) 则用一个简单的图形,展示反拨效应的各个层面与测试和社会的关系。他们认为,测试的影响始于测试本身及结果的使用,根源却在社会和教育系统,因为如何使用测试结果,由社会或教育系统决定并赋予其价值和意义,测试又反过来对社会、教育系统(宏观层面)以及个人(微观层面)产生影响。个人指与考试相关的所有人,包括考生、考生的亲属、老师等。这个图形既表明考试影响的范围与层面,

又揭示考试与社会、教育系统、以及个人之间的多重复杂关系。

以上理论模式各具特点。Bachmann & Palmer 的图形可视为宏观的理论框架,重点揭示考试反拨效应与社会的关系。Hughes 的三分法和 Bailey 的模型对反拨效应进行了较微观而更细致的描述,着眼点是受考试影响的整个教育系统各层面。Alderson & Wall (1993) 以及 Alderson & Hamp-Lyons (1996) 共同提出的 16 个假设则专门探讨考试对课堂教学各个方面的影响,深入全面,反映了考试反拨效应的复杂性(元鲁霞,2001)。但是,这些模型主要关注反拨效应的范围、层面以及测试与其他因素的交互关系,却忽略了前面提及的反拨效应维度。Green (2006) 提出的反拨效应模型,弥补了此缺陷(见图 1),拓展了早期的理论。

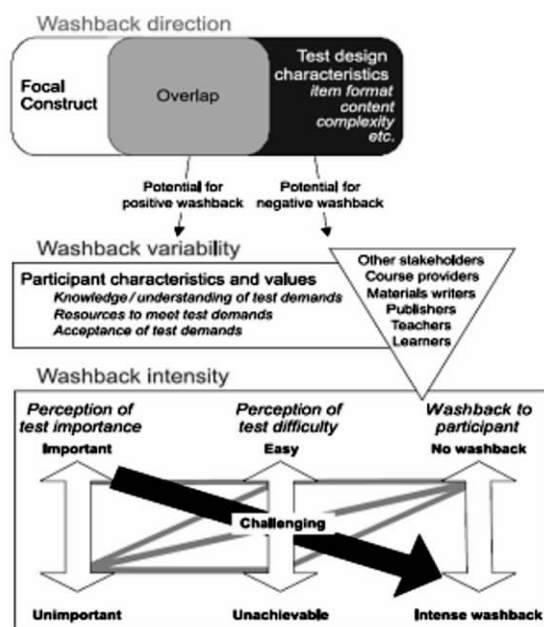


图 1. 反拨效应性质、可变性及强度

Green 的模型涵盖两个主要部分:一是考试的特征,包括考试整体设计和各部分的内容、题型、复杂程度等;二是涉考者特征,涉考者主要指考生,但也指教师、教材编写者等,其特征包括涉考者对考试要求的解读、接受程度、达到此要求所具备的资源、对考试重要性和难度的评估等。该模型的创新之处在于,将反拨效应的性质(direction)、可变性(variability)与强度(intensity)纳入其中,认为一个考试的构念效度越高(即反映拟测能力的全面性),正面反拨效应就越强。此外,反拨效应的强度,还受制于涉考者对考试重要性

与难度的评估,考试越重要(或风险度越高),反拨效应也越强,反之亦然。而难度与反拨效应的强度不成正比,过难过易的考题或考试都不会导致高强度反拨效应,惟有那些考生认为具有挑战性且能通过训练提高成绩的考试,才会引发高强度的反拨效应。可变性主要指同一个考试对不同的涉考者所产生的影响有所不同。反拨效应性质、可变性和强度三个重要变量的引入,增强了Green模型的解释力。用此模型指导研究,不仅需要了解反拨效应的范围与层面,还要根据测试的构念效度和涉考者特征,甄别积极影响和负面作用,检测反拨效应的强度与变化,这有助于更加全面深刻地剖析反拨效应的本质。

黄大勇(2011: 299)综合分析前人对反拨效应的理论探讨,提出测试效应发生模式,探究其产生机制。他认为,测试效应是在一定的社会教育环境中产生的,根源在“效应引入方”即测试结果的使用者(包括误用),测试本身是“效应媒介”,效应通过测试产生并接受其影响,受到测试影响的考生、教师、学校等构成“效应接受方”。此模式的优点是强调反拨效应的环境(这点与Bachman & Palman 1996的观点一致),并把效应的起源与测试本身分开考虑,揭示了效应的始作俑者不限于测试,有可能来自测试之外的因素,符合测试研究学者们的共识。但是此模式与前面提到的几个早期理论框架一样,未将反拨效应的几个维度纳入其中综合考虑,解释力仍然有限。

二十来年的探索丰富了反拨效应理论研究,为实证研究提供了操作框架,而实证研究也反过来验证理论假设与模型的有效性,深化了我们对反拨效应的认识。下面回顾主要实证研究和相关发现。

## 二、反拨效应实证研究

1993年Alderson & Wall发表论文,分析当时对反拨效应研究的状况,提出前面提及的15个假设,并呼吁开展实证研究加以验证。该文被视作反拨效应研究的里程碑,由此引发了一系列相关实证研究,先后在国际上有三次集中报道。语言测试国际权威期刊“语言测试”(Language Testing, 1996)和“教育评估:原理、政策与实践”(Assessment in Education: Principles, Policy & Practice, 2007)均出版了反拨效应专刊;程李颖等人主编了反拨效应专题论文集(Cheng et al 2004),其他国际国内学术期刊也陆续发表了大量的反拨效

应论文。限于篇幅,下面仅讨论部分具有代表性的研究,多数涉及大规模的高风险语言测试。

首先介绍三项研究,其共同之处是所研究的考试都带有促进教学改革的目标,带有促进英语教学向注重运用、注重交际方向转变的目的,带有考试设计者的具体期望,这可视为期望反拨效应研究。Wall & Alderson(1993)通过访谈和课堂观察,研究了斯里兰卡的新英语中级证书考试的反拨效应。虽然考试设计者期望通过新考试去促进教学改革,研究结果却表明收效不大。考试对教学内容产生了较大影响,对教学方法基本没有触动,所观察到的影响是正负面兼而有之。研究还发现,教师们所说的和所做的有一定出入,彰显了多种方法收集数据的重要性。Wall & Alderson(1993)指出,对反拨效应的研究,应该调查老师对考试的态度和理解,以便解释为什么某些期望反拨效应得以实现,而另一些却毫无反响。

Cheng(2005)研究了香港教育证书英语科考试改革对中学教育系统产生的影响。研究方法包括访谈、课堂观察和问卷调查。调查对象是考试人员、教材编写人员、教师、学生等。研究者对比分析教学材料,还分析师生对考试的了解与态度、教学内容、教学方法等。结果表明,考试改革对教科书出版商、教材编写人员等影响颇大,对教师和学生也有影响,但主要体现在教学内容的改变上,教学方法却未随之而改变。虽然考试改革的一个重要目的是推动香港中学的英语教改,使学校更加重视语言的使用及交际功能,但此目的在很大程度上未能实现。

笔者(2004)研究了我国高考英语(NMET)的反拨效应。通过比较中学的英语教学现状和NMET设计者及命题人员的意图,揭示考试是否达到预期的效果。研究方法包括访谈、课堂观察和问卷调查,对象是NMET考试人员、英语教研员、教师和学生。高考英语的设计者们曾一度利用高考的巨大影响力,试图改变中学英语教学重知识、轻能力的倾向,通过重点考查语言运用能力,促使教学转向培养学生的英语运用能力。然而,这个良好的愿望却受制于考试目的和其他因素,未能在教学中得到充分体现。研究还探索了阻碍期望反拨效应实现的各种因素,包括考试的双重功能、考试结果的误用等。

上述三项研究表明,虽然考试在有些方面发挥了积极影响,但离设计者的期望反拨效应相距

甚远。这个发现带来的启示是,考试并非决定教学的唯一因素,包括教师因素在内的其他诸多因素与考试交互作用,导致考试的实际反拨效应。对此,一些学者进行了比较深入的探讨,下面介绍几项相关研究。

Alderson & Hamp-Lyons (1996) 研究托福考试对教学的影响。调查对象是两名既上普通英语课也上托福备考课的教师。通过访谈和课堂观察,发现考试对教学的内容和方法均有影响,但影响的类型与程度却因人而异,说明教师因素起了作用。Watanabe (1996) 研究的是日本各大学的入学考试对备考的影响。他首先分析了入学考试,然后在一所辅导学校对两位教师进行访谈和课堂观察,发现教师本身的因素(如学历、教学经验、教学理念)对教学的影响更大,超过考试的反拨效应。Green (2006) 针对雅思写作进行研究,通过课堂观察和访谈,对比雅思备考课和大学英语课的写作训练,发现两者有不少相似之处,反拨效应主要体现在备考课训练的内容与范围明显窄于大学英语课应该涵盖的内容;同时还发现教师因素在教学中起着重要作用,这与 Alderson & Hamp-Lyons (1996) 以及 Watanabe (1996) 的发现相同。

辜向东 (2007) 以反拨效应的性质(即负面和正面效应)为切入点,针对大学英语四六级考试开展研究。调查对象包括考试人员、教师和学生。研究工具有问卷调查、访谈、课堂观察等。结果发现,大学英语四六级考试对教学各方面均有影响,其程度因学校、年级、教师而异;考试的正面影响大于负面影响,主要体现在促进了教学大纲在教学中的贯彻落实,激发了教师和学生的积极性等;负面影响包括教学进度过快、课堂上使用测试模拟试题,教材完成情况不够理想等。金艳 (2000) 的研究侧重于大学英语四、六级口语考试的反拨效应,在口试举行第二次试点之后,用问卷在考生和主考教师中收集数据。结果表明,口试产生了良好的反拨效应,致使学生更加重视口语学习,甚至在课外参加各种各样的口语活动。

Shohamy 等人 (1996) 在以色列同时调查两个全国性的语言测试,一个是阿拉伯语作为第二语言的考试,另一个是英语作为外语的考试,当时两个考试均进行了改革。研究者通过访谈和问卷,调查了教研员、教师和学生。结果发现,对阿拉伯语考试的改动未在教学中引起任何反响,而

对英语考试的改革却导致了备考时间、教学活动、教材等多项变动。这表明,英语考试产生了很大的反拨效应。调查的结论是,语言考试的反拨效应取决于诸多因素,其中包括考试的用途和所考查的语言在社会中的地位。Shih (2010) 研究台湾英语水平考试(GEPT)对学校制定学生毕业政策的影响。调查对象是台湾两所私立高校的应用英语系,通过访谈和文件分析,证实 GEPT 对政策制定的影响,发现了其他与考试交互作用的因素,如政府规章、教育环境、学生英语水平、学校行政工作、教学资源等。在所有上述因素的共同作用下,一个系把 GEPT 的听读考试成绩作为学生毕业的条件之一,而另外一个系则无此要求。

以上综述表明,测试对教学内容、教材编写、教学活动的影响比较明显,对教学方法的影响只在某些研究中体现,但这些研究却未涉及测试对学习结果的影响。一个考试的反拨效应是正面还是负面,一个重要的衡量标准是考试对学习结果的影响,遗憾的是相关研究极少。下面仅就为数不多的研究开展一些讨论。

关于考试对学习结果的影响,有两种普遍看法:其一,备考有助于提高考试分数,其二,考试压力下的教学导致高分低能。正是由于第一种观点,所有高风险考试都会引发应试教学,但应试是否有助于提高分数与相应能力,是否导致高分低能,目前尚无定论。Andrews, Fullilove, Wong (2002) 针对香港高级程度会考的“英语运用”考试开展研究,发现该考试增加了口试之后,中学加大了口语训练,学生的口试分数得到提高,但是分数的改善不具显著意义。此外,对学生口试中的口语特征进行深入分析后发现:分数的提高主要得益于学生对考试更加熟悉,更好地掌握了应试策略,使用了一些固定短语,因而无法证明由于考试学生的口语能力得到提高。Muñoz & Álvarez (2010) 研究了基于课程的口语评价系统对教学的影响。该评价系统包括口语课程目标、系列评价任务、评分等级和标准、评分表格、评价结果反馈报告单等。此外,学生还需要根据标准开展口语自评。调查对象是 14 位英语教师和 110 名学生,分成两组,均使用同一个口语评价系统。一组教师接受培训和指导(实验组),另一组教师则根据自己的理解独立操作该系统(对照组)。结果表明,与对照组相比,口语评价系统在实验组产生了明显的积极反拨效应,学生参加数次小型口试

的平均分高于对照组,具有显著意义。这样的结果说明,如果有一个完善的考试或评价系统,并对教师进行培训指导,会对学习产生积极的反拨效应,提高相应能力。但值得注意的是,在该实验中起作用的不仅仅是完善的评价系统,还有培训与指导这个重要变量,调查结果不能视为单纯的考试影响。

与高分低能密切相关的一个问题是,考试培训或应试教学是否破坏了考试的效度和区分度。在提高分数的同时,如果相应能力也得到提高,回答便是否定的。如果只提分不提能,应试教学就有可能影响了考试的效度。Xie(2010)发现,通过停课备考,学生的分数得到一定的提高,但是对考试的效度未构成威胁。遗憾的是,该研究调查的备考期限较短,仅两个月。如果备考时间延长,结果是否不同? 元鲁霞(2007)调查了英语高考短文改错题的反拨效应,发现高三大量训练短文改错,而且接受调查的师生主观估计考前训练可以提高2至3分(总分10分),可是高考成绩分析却表明,整体考生的短文改错题得分在全卷最低,其区分度指标仍保持高位。由此可以推测,备考未破坏考试的区分功能。

从上述几项研究结果来看,考试本身似乎不能有效改善学习或备考结果,但如果考试质量高,并对教师进行培训指导,考试或许有助于促进学习,并提高分数及相应能力。不过由于此类研究太少,下结论还为时过早。

### 三、结语

上述实证研究结果有几个主要发现。第一,考试对教学的各个环节和层面均有影响(包括教学内容、方法、教材编写等),但对不同的涉考者(即不同的考生、教师、学校等)有所不同。第二,反拨效应虽然源于测试,但诸多其他因素(包括社会、教学环境、教师、学生等因素)交互作用能够促进反拨效应性质和强度的变化。第三,大规模高风险考试较难取得设计者所期望的反拨效应。第四,考试兼有正面与负面效应,取决于看问题的角度和研究者的判断标准。<sup>①</sup> 第五,针对学习结果开展的研究不足。在教学内容和方法层面形成的反拨效应是否进一步影响学习结果? 这种影响的性质是什么? 这些问题尚需深入研究。

以上发现为反拨效应理论提供了实证依据,

加深了我们对反拨效应的了解。但是,现有理论框架仍有欠缺。如前所述,早期的理论未将反拨效应的几个维度纳入其中,解释力受限。虽然Green的反拨效应模型弥补了这个缺陷,但该模型也有不足,主要是将影响反拨效应性质的因素限制于考试的特征,未明确阐释涉考者因素的作用。而涉考者因素似乎只影响反拨效应的强度与可变性,与反拨效应的性质无关。如果一个考试有较好的构念效度,能够反映拟考查的知识与能力,基于这样构念效度的模型可以预测,其反拨效应应该是正面的。可是有实证研究表明,反拨效应的性质不仅与考试设计有关,还与涉考者对考试的理解和备考策略相关,设计者的理念不一定被涉考者理解和接受。由此可见,目前的理论框架还可完善,可以进一步整合提炼反拨效应的几个维度、运行机制、范围层面等,以增强解释力。

### 参考文献

- Alderson, J. C. & D. Wall. 1993. "Does washback exit?". *Applied Linguistics* 14/2.
- Alderson, J. C. & L. Hamp-Lyons. 1996. "TOEFL preparation courses: a study of washback". *Language Testing* 13/3.
- Andrews, S., J. Fullilove & Y. Wong. 2002. "Targeting washback — A case study". *System* 30/2.
- Bachman, L. F. & A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.
- Bachman, L. & A. S. Palmer. 2010. *Language Assessment in the Real World: Developing Language Tests and Justifying Their Use*. Oxford: Oxford University Press.
- Bailey, K. M. 1996. "Working for washback: a review of the washback concept in language testing". *Language Testing* 13/3.
- Cheng, L. & Y. Watanabe with A. Curtis. (eds.). 2004. *Washback in Language Testing: Research Contexts and Methods*. New Jersey: Lawrence Erlbaum Associates.
- Cheng, L. 2005. "Changing language teaching through language testing: A washback study". *Studies in Language Testing (Volume 21)*, Cambridge: Cambridge University Press.
- Green, A. 2006. "Watching for washback: observing the influence of the international English language testing system academic writing test in the classroom". *Language*

<sup>①</sup> 前边已提及,判断反拨效应的性质(即正面还是负面)学界仍缺乏较为统一的标准,对此笔者在另一篇文章中有更详细的讨论(元鲁霞,待发表)。

- Assessment Quarterly* 3/4.
- Messick, S. 1989. "Validity". In R. Linn, (Ed.), *Educational Measurement*. New York: ACE/Macmillan.
- Messick, S. 1996. "Validity & washback in language testing". *Language Testing* 13/3.
- Muñoz, A & M. Álvarez. 2010. "Washback of an oral assessment system in the EFL classroom". *Language Testing* 27/1.
- Popham, W. J. 1987. "The merits of measurement - driven instruction". *Phi Delta Kappan*, 68/9.
- Shih, C. 2010. "The washback of the General English Proficiency Test on university policies: A Taiwan case study". *Language Assessment Quarterly* 7/3.
- Shohamy, E., S. Donitsa - schmidt & I. Ferman. 1996. "Test impact revisited: Washback effect over time". *Language Testing* 13/3.
- Wall, D. & J. C. Alderson, 1993. "Examining washback: The Sri Lankan impact study". *Language Testing* 10/1.
- Watanabe, Y. 1996. "Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research". *Language Testing* 13/3.
- Watanabe, Y. 2004. "Methodology in washback studies". In L. Cheng, & Y. Watanabe with A. Curtis (eds). *Washback in Language Testing: Research Contexts and Methods*. New Jersey: Lawrence Erlbaum Associates.
- Xie, Q. 2010. *Test Design and Use, Preparation, and Performance: A Structural Equation Modeling Study of Consequential Validity*. Unpublished PhD thesis. The University of Hong Kong.
- 辜向东 2007, 大学英语四 - 六级考试对中国大学英语教学的反拨效应实证研究,《重庆大学学报》(社会科学版)第4期。
- 黄大勇 2011 构建语言测试效应研究的理论基础,《现代外语》第3期。
- 金艳, 2000, 大学英语四、六级考试口语考试对教学的反拨作用,《外语界》第4期。
- 李筱菊, 1997,《语言测试科学与艺术》, 湖南教育出版社。
- 亓鲁霞 2001, 考试对教学的反拨作用: 推动还是阻碍? 载董燕萍、王初明编,《中国的语言学研究与应用》, 上海: 上海外语教育出版社。
- 亓鲁霞 2004,《意愿与现实: 中国高等院校统一招生英语考试的反拨作用研究》, 北京: 外语教学与研究出版社。
- 亓鲁霞 2007, 高考英语的期望后效与实际后效——基于短文改错题的调查,《课程·教材·教法》第10期。
- 王初明 2011, 外语教学三大情结与语言习得有效路径,《外语教学与研究》第4期。
- 杨惠中 桂诗春 2007, 语言测试的社会学思考,《现代外语》第4期。
- 杨惠中, C. Weir, 1998,《大学英语四、六级考试效度研究报告》, 上海: 上海外语教育出版社。
- (通信地址: 510420 广州市广东外语外贸大学外国语言学及应用语言研究中心)  
(文字编校: 严文庆)

## A review of washback studies

by QI Luxia

**Abstract:** In the past two decades or so, washback or the influence of testing on teaching and learning has been intensely researched. Hypotheses and models have been formulated to uncover its nature, scope, patterns and mechanisms. Its intensity, length, specificity, intentionality and value have also been explored. Theorizing has done much to provide frameworks for empirical studies, which in turn have afforded evidence for or against concepts and theories of washback. An overview of the washback studies has shown that language tests have imposed both positive and negative washback on all aspects of teaching and learning, including teaching content, methods, and materials. However, tests' influence on learning outcome remains largely nebulous due to scarcity of studies on this issue. Further, washback theories and models still leave considerable room for improvement before their explanatory power can be desirably enhanced.

**Key words:** testing; washback; stakeholders