效度与信度的权衡

——交际性英语听力测试中题型问题的探讨

上海对外贸易学院 颜 薇 王 勇

提 要:本文根据语言测试理论对目前听力测试中经常采用的两类主要题型,即选择题型和建构题型的各自特点进行比较和分析。文章既从听力考试的信度和效度等主要衡量指标对听力题型的重要性展开讨论,同时又利用教学中的实验数据对建构题型在检验语言学习者的语言应用能力方面的重要意义加以论证。最后,作者提出逐步扩大建构题型比例是听力测试改革的一个趋势,在考试的效度和信度方面寻求最合适的平衡点是我们的主要研究目的。

1. 引言

在语言测试研究中,越来越多的语言研究者和语言教师开始注意到测试方式的重要性。如果说测试内容(test content)解决的是测试"什么"的问题,那么测试方法则是针对"如何"测试的问题(Bachman, 1990: 111; McNamara, 2000: 66)。题型是测试方法的一个重要方面,决定了考生答题的方式,影响其测试表现(test performance)(Bachman, 1990: 113; 1991: 678)。由于题型不受测试内容的限制,因此题型研究是语言测试学,乃至教育评估学中一个重要的课题。本文试图通过理论和实证研究,就交际性英语听力中选择题型和建构题型的效度和信度问题展开讨论,同时结合实际提出一些作者个人的看法。

2. 题型的分类

一般来说,题型可分为两种。第一种为选择题型(Clark, 1972: 27),又称选择回答(selected response)(Popham, 1978)或固定回答(fixed-response)(Bennett, 1993: 2),要求考生从几个选项中挑选答案。第二种为自由回答(free response)(Clark, 同上),又称建构回答(constructed response)(Popham, 同上; Bennett, 同上),要求考生自行给出口头或书面的回答。本文使用"选择题型"和"建构题型"的概念。

选择题型最早在一战时应用于筛选应征士兵,因其经济可信的特性而沿用至今(Mislevy, 1993: 76),是各大语言测试体系的主要题型之一。选择题型不仅是"美国当今标准化考试的支柱"(Bennett & Ward, 1993: ix),也是我国主要英语考试(如高考英语和大学英语四六级考试)的主要题型。

许多学者(Bennett, 1993: 2 - 5; Traub, 1993: 29 - 30; Bachman & Palmer, 1996: 54 等)认同建构题型包含一系列具体的题型。其中比较详细的阐述是Bennett(同上)将所有题型按开放性从低到高排列成一个连续体,即选择题—多选/辨认题—排序题—替代/改错题—填空题—回答题—演示/操作题。

3. 题型与效度-信度

题型不仅是表面形式,而且与测试的效度和信度有着紧密的联系。

3.1 题型与效度

Nuttall, C. 1996. *Teaching Reading Skills in a Foreign Language*. Oxford: Heinemann.

Pearson, P. D. & L. Fielding. 1991. "Comprehension Instruction". In R. Barr. et al. (eds.). *Handbook of Reading Research*. White Plains, NY: Longman.

Wassman, R. & L. Rinsky. 2000. *Effective Reading in a Changing World*. NJ: Prentice Hall.

Williams, M. & R. L. Burden. 1997. *Psychology for Language Teachers*. Cambridge: Cambridge University Press.

刘丹丹,2002,中国英语学习者的阅读策略研究,《外语界》第6期。

孟悦,2004,大学英语阅读策略训练的实验研究,《外语与外语教学》第2期。

王睎、黄慧娟、许明,2003,PISA:阅读素养的界定与测评,《上海教育科研》第9期。

(**通讯地址**:201102 上海市莲花南路588弄44号301室)

效度指根据测试结果所做出推断的合适性、有意义性和有用性(American Psychological Association, 1985:9)。效度又可分为内容效度、效标关联效度和构想效度等几个层面,其中与题型有密切关系的是构想效度。

构想效度指"考试实际测得的东西与理论所假设的能力要素或心理特征相吻合的程度"(舒运祥,1999:54)。题型是引出考生语言能力的中介,"通过限制所测试的内容和过程而影响测试结果的意义"(Frederiksen, 1984),能够"影响考生的答题心理"(Snow,1993:46),有时能混淆欲测试的构想,如混淆答题技巧与掌握的知识、辨认与回想、事实知识与更高层次的思辨能力等(Bennett, 1993:6)。

3.2 题型与信度

信度指"考试结果的可靠性和稳定性"(舒运祥,1999:61),即假设在另一时间对相同考生再次进行测试,由不同评分者评分,所得结果与原结果的一致性(Hughes,1989:29)。Heaton(1988:162)列举了影响信度的五大主要因素:1)试题的覆盖面。试题覆盖面越大,题量越大,信度越高;2)测试的安排,如地点、条件等;3)试题要求,即试题要求简洁明了;4)考生的个人因素,如动机、健康状况等;5)评分,不同评分者或同一评分者在不同时间的评分结果应一致。

因此,题型对信度的影响表现为不同题型因所需答题时间不同,限制了测试中试题的数量;不同题型因评分方法不同而导致评分结果的一致性有异;不同题型对考生心理产生不同影响而形成考生的不同动机等等。

3.3 题型与效度和信度的对立

虽然效度和信度是考察测试质量的重要指标,但"在某些场合两者互相排斥"(Weir,1990:33),并形成一种对立。

在权衡效度和信度的关系时,许多学者将重要性放在效度上。Spolsky(1965:94)认为,没有效度,所有其他指标(包括信度)都失去意义。Davies(1990:32)也认为,一项测试最重要的是效度。Hawkey(1982:149)明确表示,为确保测试的效度,信度可以放在从属的地位。

在具体处理信度、效度的对立时,Moller(1981:67)认为部分试题可以追求较高的信度,而另一部分试题,尤其是测试交际能力的,必须注重其效度。Heaton(1988:165)提出,首先应该设计高效度的测试,其次再通过其他手段提高其信度,手段之一是分语言能力等级,设计详尽的评分标准。

不同的题型由于其内在性质在效度和信度上各有千秋,因此,选用哪种题型体现了试题设计者在效度—信度关系中的立场。

4. 选择题型和建构题型的特点

4.1 选择题型的特点

选择题型客观、经济,能增加测试的题量和考题的覆盖面,评分较客观,适合机器阅卷,因此具有较高的信度(Hughes, 1989:59-60)。但选择题型在效度上屡遭质疑。Heaton(1988:165)认为,语言能力决不会从几个选项中选出正确答案能表现出来的。选择题型的答题过程与交际语言的性质相去甚远(黄素华 1998:92)。Hughes(1989:60-62)列举了选择题型的下列缺陷:1)选择题型仅能体现考生辨认知识的能力,对产出性能力考查不够;2)考生很可能在答题过程中进行猜测,从而影响对测试结果的准确阐释;3)有些知识点不适合设计选择题型的干扰项,从而限制测试内容的广度;4)很难写出高质量的试题;5)其后效作用可能有害。在语言测试中大量运用选择题型,会导致大量运用选择题型来教学(Smith, 1991:10)。同时,选择题型会使语言学习者认为,任何语言问题都有唯一的答案,而答案就在教书或命题人脑中,学生的任务就是选出这个答案,即使用猜测的方法也无妨(Shepard, 1991);6)由于考生很容易通过非语言手段传送选择题答案,该题型利于考试作弊。

4.2 建构题型的特点

由于考生必须给出自己的推断、比较和概括等,建构题型能更好测试高层次的语言能力(Weir,1990:53)。由于没有选项,建构题型能大大减少猜测等考试技巧对考核考生真实语言能力的影响(Weir,1993:45,53)。建构题型更接近语言实际使用的特征,考生的答题过程也更类似实际语言使用任务。如在听力测试中,考生被要求记下口信或完成一次听课笔记,这种任务在日常生活和学术语境中很常见。同时,建构题型能促使教师将教学重点放在提高学生语言能力上,也能促进学生在考试前做广泛的准备,从而学得更多(Bennett,1993:18)。可以说建构题型在测试语言交际能力上具备较高的效度。但同时,Bennett(同上)也指出,"建构题型因其内在性质

可能造成考分的信度有所缺陷"。的确,考生在写出答案时,其测试表现会受到单词拼写或写作能力的影响。另外,建构题型的评分主观性很大,进而影响到信度,也很难运用高效的机器阅卷手段。

5. 实验

虽然理论研究表明两者相比,建构题型的效度较高,选择题型的信度较高,但遗憾的是,从目前的文献看,只有为数不多的实证研究能证明建构题型的效度优势或选择题型的效度劣势。Bennett(1993:8)在研究了相关实验后,查找了几大原因。一是大多数实验所用试题不能完全显示两种题型的效度差别,例如演示题(presentation tasks)未被列入研究范围。许多实验者只是将选择题型直接转换成建构题型,由此不能区分两种题型实际考查的构想(Frederiksen,1984)。二是实验方法不当。要研究效度,尤其是构想效度,大多数研究所用的定量研究方法不合适。吴一安(2001)的实验也表明,内省法这样的定性研究方法能揭示考生回答某一题型时的思维过程,从而评价某一题型的构想效度。三是评分系统设计不当,许多研究从选择题型照搬过来的评分体系并不适合建构题型。

有鉴于此,本实验做到:1)为实验单独设计建构题,而不是从选择题转换而来;2)采用口头报告的方法让考生讲述答题思路;3)设计适合建构题型的评分标准。本实验的测试范围限定为交际性英语听力测试,旨在考查选择题型和建构题型的效度和信度差别。

5.1 实验程序

作者随机挑选30名大学二年级正准备参加大学英语四级考试的学生,这些实验对象根据前两学期的期末总评成绩和本轮听力测试的成绩分成3组,平均成绩高于80分的设为A组,70-79分的为B组,60-69分的为C组。每组分别由5名男生和5名女生组成。他们参加两次听力测试,测试一全部采用选择题型,测试二全部采用建构题型,目的是了解他们在做两种题型时的表现。测试一的试卷随机选用2002年1月大学英语四级考试听力部分及参考答案。测试二包含有填空、回答等不同开放度的建构题型。考虑到实验对象首次参加建构题型的测试,所有回答不超过3个单词或数字。评分标准经过仔细考虑,并列出所有可接受的答案。同时,为模仿实际听力场景,录音语调自然,有停顿和重复,在各题内容开始前说明场景和说话者之间的关系,录音内容有一定的信息冗余度。录音长度为30分钟,每部分留有20秒浏览问题的时间和30秒的检查时间。之后,这30名实验对象参加口头报告。

5.2 实验结果分析和讨论

由于篇幅有限,作者这里分别选取测试一的前5题和测试二的第一部分(共计12个空格)对30名实验对象做口头报告分析。根据杨惠中和Weir(1998:94)对听力技能的分类,实验用题所考核的听力技能归类见表1。

微技能	测试一	测试二
听懂重要的或特定的细节	无	第1-7题
理解中心思想	第3、5题	第9、10题
进行推论	第1、2题	第11、12题
判断话语的交际功能	第4题	第8题

表1 测试一、二中微技能列表

在测试一中,实验对象的测试过程表现可以归纳为以下几种:

表现	是否听懂录音内容	是否理解选项	答题手段	答题结果
1	是	是	基于理解*	正确
2	否	是	基于理解或猜测	错误
3	是	是	排除	正确
4	否	是	猜测	正确
5	是	否	基于理解	错误

表2 测试一中实验对象测试过程表现归类

*注:指实验对象根据自己的理解进行答题,包括正确和错误的理解,下同。

题号	表现1	表现2	表现3	表现4	表现5
1	13*	2	7	8	0
2	15	4	5	5	1
3	11	7	8	2	2
4	7	3	11	0	9
5	9	8	10	3	0
总计	55 (36.7%)**	24 (16.0%)	41 (27.3%)	18 (12.0%)	12 (8.0%)

表3 测试一中各题答题情况归类

*注:单位为人次,下同。

**注:括号内的数值为该表现出现人次的总数除以参加1-5题测试的总人次的百分比。

从构想效度理论看,实验对象作出正确回答应缘于其对所听内容的完全理解,而实验对象作出不正确回答时,则应表明其对所听内容存在错误理解,那么表现1和2是符合效度要求的,而表现3、4、5则与效度要求不符,或者说表明该考题

效度的某种不足,且从表 3 看,这三者出现的频率并不低。现将测试一的 5 道题中表现 3、4、5 的分布情况列表如下。

	表现 3	表现 4	表现 5	总计
A 组	2	0	0	2
B 组	3	3	0	6
C 组	2	5	0	7
总计	7 (23.3%)*	8 (26.7%)	0 (0%)	15

表 4　测试一第 1 题中三种考试表现分布情况

*注:括号内的数值为该表现出现人次的总数除以参加该题测试人次的百分比,下同。

	表现 3	表现 4	表现 5	总计
A 组	1	0	0	1
B 组	1	2	0	3
C 组	3	3	1	7
总计	5 (16.7%)	5 (16.7%)	1 (3.4%)	11

表 5　测试一第 2 题中三种考试表现分布情况

	表现 3	表现 4	表现 5	总计
A 组	2	0	0	2
B 组	3	1	0	4
C 组	3	1	2	6
总计	8 (26.7%)	2 (6.7%)	2 (6.7%)	12

表 6　测试一第 3 题中三种考试表现分布情况

	表现 3	表现 4	表现 5	总计
A 组	1	0	3	4
B 组	5	0	2	7
C 组	5	0	4	9
总计	11 (36.7%)	0 (0%)	9 (30.0%)	20

表 7　测试一第 4 题中三种考试表现分布情况

	表现 3	表现 4	表现 5	总计
A 组	3	0	0	3
B 组	3	1	0	4
C 组	4	2	0	6
总计	10 (33.3%)	3 (10.0%)	0 (0%)	13

表 8　测试一第 5 题中三种考试表现分布情况

从表现 3 的数据可得,在听懂了录音内容的情况下,各组的部分实验对象仍然用排除法来回答选择题型,5 道题中分别有 23.3%、16.7%、26.7%、36.7%和 33.3%的实验对象使用了排除法,比重相当高。其中包含两种情况,部分实验对象虽理解所听内容,但面对选择题型仍习惯用排除法答题;部分实验对象产生不同于 4 个选项的答案,而用排除法选择较接近于其理解的答案。这样的答题表现有悖于真实的语言运用过程和听力的交际性原则,一定程度上有损于选择题型的效度。从表现 4 看,部分实验对象在没有理解所听内容的情况下,采用猜测手段做出正确的回答,最多时有 26.7%的实验对象进行了猜测。这种情况违背了构想效度的要求,不能真实反映这部分实验对象的语言水平。同时,猜测手段的运用与实验对象的语言水平有关,即通过猜测而非通过理解来做出正确回答的情况较多出现在语言水平较低的实验对象上。从表现 5 看,部分实验对象虽然理解所听内容,却因为误读选项而导致回答错误。而且从数据分布看,这种情况与实验对象的语言水平无关,A、B、C 三组的实验对象均有此现象发生。这说明在听力测试的时间限度内,选项的阅读不利于对听力测试的准确性。在这里,选项成为选择题型的效度问题的源头。假设在测试时没有选项的束缚,那么这些实验对象就完全可能做出正确的回答,显示真实的语言水平。

值得注意的是,从上述表中的数据可以看出,表现 4 的出现频率与所考的听力技能相关。实验对象在做第 1、2 题这样的推论题时,容易在未完全听懂语篇片段的情况下猜测出答案。另一方面,表现 5 的出现频率与题干用词有关。实验对象在做第 4 题时,由于时间压力,容易出现听懂录音内容,却因误解选项而造成回答错误的情况。由此可见,选择题型在考查推断能力时效度有所欠缺,且设计难度可见一斑。

在测试二中,实验对象的实际测试过程表现可以归纳如下。

	对录音内容的理解情况	答题情况
1	理解	答案正确
2	未理解	答案错误
3	未理解	答案中出现拼写错误
4	未理解	未给出答案

表 9　测试二中实验对象测试过程表现归类

从结果看,这道题体现了建构题型良好的效度,实验对象的正确回答缘于对录音内容的正确理解,而错误的回答则大多缘于对录音内容的理解失误,未出现理解正确而答案错误,或理解错误而答案正确的情况。表 10 为实验对象在各题中的考试表现。

限于篇幅,作者将出现表现 3、4 的各题具体情况列表如表 11。

题号	表现1	表现2	表现3	表现4
1	28	2	0	0
2	30	0	0	0
3	22	0	8	0
4	30	0	0	0
5	30	0	0	0
6	25	5	0	0
7	24	6	0	0
8	19	3	2	6
9	20	3	2	5
10	20	1	4	5
11	18	8	0	4
12	18	6	2	4
总计	284 (78.9%)*	34 (9.4%)	18 (5.0%)	24 (6.7%)

表10 测试二中各题答题情况归类

*注:括号内的数值为该表现出现人次的总数除以参加1-12题测试的总人次的百分比。

从表11可见,各组实验对象均出现了拼写错误和答案空缺,且语言水平越低的实验对象越频繁地出现这两种情况。这一定程度上说明建构题型易受考生拼写能力的影响,也提醒命题人员在设计建构题型时,应充分考虑考生的书写负担,控制答案的长度和文字难度。另外,值得注意的是各组实验对象未给出答案,并承认没有理解相关内容的情况均有发生。由此可见,建构题型能有效减少考生答题时使用猜测的考试技巧,从而使测试结果更有效反映考生实际语言应用能力。从所考技能来看,考查人名等的细节题易导致考生拼写错误;而在考查推论、理解中心思想、理解交际功能等较高能力时,建构题型有助于防止考试进行猜测,从而反映考生的真实能力,保证考试的效度。

题号	表现3			表现4			总计
	A组	B组	C组	A组	B组	C组	
3	1	3	4	0	0	0	8
8	0	1	1	1	1	4	8
9	0	1	1	0	1	4	7
10	1	1	2	1	1	3	9
11	0	0	0	0	1	3	4
12	0	0	2	0	2	2	6
总计	2 (3.3%)*	6 (10.0%)	10 (16.7%)	2 (3.3%)	6 (10.0%)	16 (26.7%)	42

表11 测试二中表现3、4分布情况

*注:括号内的数值为该表现出现人次的总数除以参加该6道题测试的总人次的百分比。

对比测试一和测试二的口头报告结果,我们可以看出,实验对象面对选择题型时,较频繁地使用了猜测、排除的方法,语言水平较低的实验对象更是频繁地使用这些无关语言水平的答题方法,使得测试结果不能完全准确地反映他们的实际语言水平。当面对建构题型时,实验对象根据对录音内容的理解作答,甚至因未理解而放弃作答,可以说建构题型的答题结果能很好地反映实验对象的听力理解能力,从而享有更高的效度。

同时,由于测试二的答案长度控制在3个词或数字之内,建构题型的信度也有一定的保证。具体体现在:1)由于答案短小可以在考试中容纳较多试题,从而确保试题的覆盖面;2)考点较明确,可预见的答案数量得到有效控制,很大程度上确保评分环节的信度。

另一方面,实验表明,选择题型的选项可能被考生误读而影响答题结果,而建构题型也对考生的拼写造成压力。这两种情况需经试题设计者的努力而得到解决。

6. 对大学英语四、六级考试的启示

大学英语考试曾因效度和题型问题而受到学术界的批评(牛强,2001;韩宝成等,2004)。为配合2004年出台的《大学英语课程教学要求(试行)》,大学英语考试进行了改革,包括增加建构题型的比重,新四、六级考试采用的建构题型有填空、改错、简短回答、翻译和写作等。

这种趋势与本文的研究思路是一致的。纵观大学英语四、六级考试改革,作者认为,可以进一步加大建构题型的比重,特别是听力部分。改革后的大学英语四、六级考试听力部分的建构题型只涉及了填空题。作者认为,只要控制得当,听力部分及其他考题还应采用更加开放的题型,以增加该考试的效度,更准确地反映考生的真实语言水平和交际能力。当然,提倡建构题型,并不意味着立即、完全废除选择题型。鉴于其内在优点,选择题型可在大规模考试的题型中起补充作用,以期保证大规模考

试的信度。总之,题型是涉及考试效度和信度的重要因素,题型研究的最终目的是在考试的效度和信度之间找到最合适的平衡点。

参考文献

American Psychological Association. 1985. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

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

Bachman, L. F. 1991. "What does language testing have to offer?". *TESOL Quarterly* 25/4.

Bachman L. F. & A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Bennett, R. E. 1993. "On the meanings of constructed response". In R. E. Bennett & W. C. Ward (eds.). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, Inc., Publishers.

Bennett, R. E. & W. C. Ward (eds.). 1993. *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, Inc., Publishers.

Clark, J. L. D. 1972. *Foreign Language Testing: Theory and Practice*. Philadelphia, Pa.: Center for Curriculum Development, Inc.

Davies, A. 1990. *Principles of Language Testing*. Oxford: Blackwell.

Frederiksen, N. 1984. "The real test bias: Influences of testing on teaching and learning". *American Psychologist* 39.

Hawkey, R. 1982. *An Investigation of Inter-Relationships between Cognitive/Affective and Social Factors and Language Learning*. PhD thesis: Department of English for Speakers of Other languages, Institute of Education, London University.

Heaton, J. B. 1988. *Writing English Language Tests* (new edition). New York: Longman.

Hughes, A. 1989. *Testing for Language Teachers*. Cambridge: Cambridge University Press.

McNamara, T. 2000. *Language Testing*. Oxford: Oxford University Press.

Mislevy, R. J. 1993. "A framework for studying differences between multiple-choice and free-response test items". In R. E. Bennett & W. C. Ward (eds.). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, Inc., Publishers.

Moller, A. D. 1981. "Assessing proficiency in English for use in further study". In J. A. S. Read (ed.). *Directions in Language Testing*. Singapore: Singapore University Press.

Popham, W. J. 1978. *Criterion-Referenced Measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Shepard, L. 1991. "Interview on assessment issues with Lorrie Shepard". *Educational Researcher* 20/5.

Smith, M. L. 1991. "Put to the test: The effects of external testing on teachers". *Educational Researcher* 20/5.

Snow, R. E. 1993. "Construct validity and constructed-response tests". In R. E. Bennett & W. C. Ward (eds.). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, Inc., Publishers.

Spolsky, B. 1965. "Review of two tests". In O. Buros (ed.). *The Sixth Mental Measurements Yearbook*. Highland Park, N.J.: Prentice Hall.

Traub, R. S. 1993. "On the equivalence of the traits assessed by multiple-choice and constructed-response tests". In R. E. Bennett & W. C. Ward (eds.). *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, Inc., Publishers.

Weir, C. J. 1990. *Communicative Language Testing*. Hemel Hempstead: Prentice Hall International (UK) Ltd.

Weir, C. J. 1993. *Understanding and Developing Language Tests*. Englewood Cliffs: Prentice Hall.

韩宝成、戴曼纯、杨莉芳,2004,从一项调查看大学英语考试存在的问题,《外语与外语教学》第2期。

黄素华,1998,科学公正地测试学生的口语能力,邹申主编,《英语语言测试:理论与操作》,上海:上海外语教育出版社。

教育部高等教育司,2004,《大学英语课程教学要求(试行)》,上海:上海外语教育出版社。

牛强,2001,现行高校英语测试中的问题,《外语教学与研究》第2期。

舒运祥,1999,《外语测试的理论和方法》,上海:世界图书出版社。

吴一安,2001,题型与听力测试的有效性,《外语教学与研究》第2期。

杨惠中、C. Weir, 1998,《大学英语四、六级考试效度研究》,上海:上海外语教育出版社。

(通讯地址:201620 上海市上海对外贸易学院外语学院)