概化理论在BCT口语测试中的应用研究

李海燕
[中国]北京大学

提要：本文采用概化理论的p×i×r的随机双面交叉设计，对商务汉语考试（BCT）的口语测试的信度和设计进行了探讨。结果表明，目前BCT口试采用试题数为2、评分员数为3的设计总体信度较高，概化系数$E\rho^2>0.8$，是经济可行的测试。若要进一步提高测量信度，使$E\rho^2>0.9$、$\Phi>0.6$，经济有效的设计方案为：评分员个数为3、试题数量为3，或者评分员数量为2、试题数量为4。BCT属于特殊目的测试，涉及商务专业内容，本文对口语试题的内容和难度控制也进行了讨论。

关键词：商务汉语考试（BCT）　概化理论　口语测试

1　引言

1.1　BCT口试的研发概况

商务汉语考试（BCT）作为国家汉办项目，自2003年开始研发，到2006年9月在新加坡首次开考，其口试的形式和内容不断在调整。

BCT口试考查的是考生在商务活动或与商务有关的生活社交活动中的汉语口头表达能力。2003年12月首次预测时，考试时间共15分钟。考试采用录音形式，共有三道题，难度分别为易、中和难。考试开始后，考生会先听到录音中考官问考生姓名、国籍、序号、试卷号码等问题，然后考官会给考生读一遍第一道考题和答题要求，考生准备，准备时间结束时播放一小段情景录音，考生开始回答。答题时间结束后，考官会读下一道题，过程与前一道题相同。考生三道题的答题时间分别为1分钟、2分钟和3分钟。

2003年预测后，部分专家和考生反映考题难度稍大，国家汉办也希望考试降低门槛，另外由于BCT由“听·读”和“说·写”两个分测验构成，很多考生要连续参加两个考试，考试时间不宜过长，因此研发办对预测情况进行分析后，将口试调整为两道试题，去掉了第三道较难的题。两道题在难度和任务类型上各有侧重：第一题较容易，要完成的交际任务主要是与商务有关的生活、社交类任务，大部分是日常生活中较熟悉的话题；第二题较难，多为商务业务类交际任务。口试程序不变。第一题准备1分30秒，回答1分钟，第二题准备2分30秒，回答2分钟（注：后因部分高水平考生受时间限制发挥不足，又将两道题的准备时间各减少30秒，而把答题时间各增加30秒）。这样把整个口试时间压缩到10分钟。

在评分阶段，评分员三人一组，根据评分标准独立评分，然后共同商定每道题的原始分。BCT研发办公室经过复审后根据原始分转成导出分数，确定考生的口语水平等级。

1.2　研究问题

BCT口试采用调整后的形式和内容，基本上得到了考生和用户认可，认为考题任务设计实

用、有针对性。在广泛听取各界反馈意见时，有部分企业用户认为目前两道题题量偏少，考试时间不足，担心不能全面反映考生的口语水平。

满足用户的要求是考试研发努力的方向。但在有限的考试时间内，到底选择两道题还是三道题或者更多的题呢？目前两道题的口试信度如何？在保证较高信度的前提下采取什么样的口试方案更加经济可行呢？本文运用概化理论对部分正式考试的口试数据进行分析，以期回答这些问题。

2 概化理论（GT）及其在衡量测试信度方面的方法

2.1 概化理论关于测试信度的估算

相对于CTT的真分数概念，GT提出了全域分数（Universe Score）的概念，即把被试的某种潜在特质水平定义在具体的测量条件全域（范围）上的分数。GT认为，被试的水平不能抽象地描述为真分数，而应根据决策的需要，把它置于指定的条件范围之中进行解释，因为每次测量工作所涉及的条件或者说影响测量结果的因素、侧面（facet）是不尽相同的，研究者对测验结果的用途，即推论或概括的程度也不尽相同，所以，在讨论被试水平时，同时应指出这种水平是在何种测量条件下取得的（杨志明，张雷，2003）。

除测量目标外，GT把凡是会影响被试得分的条件因素都称之为测量侧面。对于BCT口试来说，测量目标是被试的口语能力，测量侧面有试题面和评分者面以及它们之间的交互作用等。

GT认为，测量误差包括两种：一是相对误差，即由所有随机误差引起的测量误差，所有与测量目标（被试）有关（被试与试题之间、被试与评分者之间，以及被试、试题和评分者之间）的交互效应（random effects variance components）构成测量的相对误差变异，记作δ，用相对误差估计出来的信度系数叫概化系数（Generalizability Coefficient）（$E\rho^2=\sigma^2(p)/[\sigma^2(p)+\sigma^2(\delta)]$）；二是绝对误差，包括了全部的系统误差和随机误差的变异。在概化全域（universe of generalization）上，除了被试主效应之外，所有侧面的主效应、侧面及测量目标之间及侧面之间的交互效应构成了测量的绝对误差变异，记作Δ，用绝对误差估计出来的信度系数叫可靠性系数（Dependability Coefficient）（$\Phi=\sigma^2(p)/[\sigma^2(p)+\sigma^2(\Delta)]$）。

由于概化全域可以有多个，因此，$\sigma^2(\delta)$和$E\rho^2$也会有多个，也就是说，GT可以针对测验结果概括程度的不同而估计出多个测量“信度”（杨志明，张雷，2003）。

2.2 概化理论研究方法

概化理论研究分为概化研究（G研究）和决策研究（D研究）两个分析步骤。

G研究首先明确测量目标、测量侧面和观测全域，以及测量设计和测量模式，然后收集样本数据做变异数分析，即使用方差分析（ANOVA）软件把各种因素（测量目标及各测量侧面）的效应及因素之间的交互效应一一估计出来，得到各个效应的变异分量（方差分量σ^2）。

D研究主要是在一个具体的概括全域上分别估计相对误差变异$\sigma^2(\delta)$和绝对误差变异$\sigma^2(\Delta)$，然后得到概化系数$E\rho^2$和可靠性指数Φ，并以此作为整个测验的精度指标。概化系数与可靠性指数是动态的，概括全域不同取值不同，因此可以根据对测量信度的不同要求，对测量侧面水平进行估计，以优化测量设计。

3 BCT口试的一元概化理论（UGT）研究

3.1 研究样本

为了比较全面地了解BCT口试的信度和稳定性，本文选取七次口试成绩作为研究样本。其中三次是在中国国内举行的，时间分别是2007年6月、10月和12月，考生人数分别为296、144和375；另外四次口试是在韩国举行的，时间分别是2007年9月、11月，2008年1月、3月，有效考生人数分别为219、244、147和74。[①]

3.2 研究设计

本研究为p×i×r的随机双面交叉设计测量模式。在这七次口试中，每位考生（person，p）都必须回答两道试题（item，i），被试和试题样本可以看做是分别从各自所对应的无限全域中随机抽取的。考试结束后每个被试样本均有三位经过培训的评分员（rater，r）分别独立评分。每个评分组的三位评分员及评判样本都是随机分配的。在这一双面交叉设计的随机测量模式中，被试的口语水平（p）是测量目标，观察全域由被试要回答的试题（i）和评分员（r）两个测量层面构成，这样就得到了被试、题目和评分者三种主效应（p，i，r）和四种交互效应（pi，pr，ri，pri），这七个方面的效应是误差的来源。

3.3 国内2007年6月BCT口试的G研究

我们使用SPSS13.0的一般线性模型（General Lineal Model）中的单元方差分析（Univariate），取被试得分为因变量，被试、试题和评分员为自变量对数据进行处理，得到离差平方和SS和均方MS，然后代入G研究方差分量估计公式（见杨志明、张雷，2003：94），得到各种效应的G研究变异分量估计值如下：

表1 2007年6月国内BCT口试p×i×r设计G研究变异分量估计

变异来源	自由度df	离差平方和SS	均方MS	方差分量的估计值σ^2	占总变异的百分比
Person	295	4173.36	14.147	2.0573	27.39%
Item	1	4089.541	4089.541	4.6018	61.27%
Rater	2	8.45	4.225	0.0036	0.05%
Person * Item	295	467.626	1.585	0.4257	5.67%
Person * Rater	590	310.216	0.526	0.1090	1.45%
Rater * Item	2	3.748	1.874	0.0053	0.07%
Person * Rater * Item	590	181.586	0.308	0.3080	4.10%

从表1可以看出，来自被试（p）的方差分量σ^2（p）为2.0573，它代表被试之间水平的差异，解释了分数总变异的27.39%，是主要的变异来源之一。被试水平之间的差异属于测量目标引起的有效变异，这个值越大越是我们希望看到的，这说明考试的测量目标比较明确。但相对于试题面（i）来说，σ^2（p）所占的比例偏少，来自试题面（i）的变异分量σ^2（i）占总变异的比例达到61.27%，试题面的变异分量代表试题之间的同质性，这说明试题之间的难度彼此差

① 在韩国考场有一小部分考生录音无法听辨，无法给出成绩，所以未计入本文统计的考生数。

异很大，不同难度的试题给分数变异带来了相当大的系统误差，不同的试题对考生的能力测试会造成较大的差异。这样的结果可能与BCT口试的设计相关。考试设计的两道试题一道偏易，一道偏难，难度有明显差别，难度差别过大会直接影响被试的分数，导致对被试水平过低或过高的估计而产生较大的系统误差，但试题的难易程度对一次测试中的所有考生来说是一视同仁的，因为所有被试都要回答相同的试题，所以试题主效应σ^2（i）不会改变被试的口语水平高低的排位顺序。

在表1中，被试与试题的交互效应的方差分量σ^2（pi）所占比例仅为5.67%，说明考生在不同试题上的表现是比较稳定的。表1中评分员的变异分量σ^2（r）只占到总变异的0.05%，在各效应的方差成分中所占比例最小，说明评分员评分的差异对分数变异的影响非常小，评分员之间评分标准的一致性即评分员的信度较高。被试与评分员的交互效应（pr）以及评分员与试题的交互效应（ir）的方差分量占总变异的百分比分别为1.45%和0.07%，所占比例很小，说明评分员对不同考生和不同试题的评分的稳定性也很好。

3.4 七次口试的G研究结果比较

按照上面同样的方法对另外6次的口试数据进行分析，得到这七次口试各种效应的G研究变异分量估计值如表2所示。

表2 七次BCT口试p×i×r设计G研究变异分量估计值及占总变异的百分比

变异 \ 来源		Person	Item	Rater	pi	pr	ri	pri
2007.6 国内	**df**	295	1	2	295	590	2	590
	σ^2	2.057 3	4.601 8	0.003 6	0.425 7	0.109 0	0.005 3	0.308 0
	百分比	**27.39%**	**61.27%**	**0.05%**	**5.67%**	**1.45%**	**0.07%**	**4.10%**
2007.9 韩国	**df**	218	1	2	218	436	2	436
	σ^2	3.694 3	3.681 5	0.001 24	1.043 75	0.083 99	−5.934 7*	0.235 2
	百分比	**42.27%**	**42.12%**	**0.014%**	**11.94%**	**0.96%**	**0**	**2.69%**
2007.10 国内	**df**	143	1	2	143	286	2	286
	σ^2	1.839 3	4.391 5	0.001 997	1.687 3	0.014 5	−0.001 82	0.284
	百分比	**22.38%**	**53.43%**	**0.024%**	**20.53%**	**0.18%**	**0**	**3.46%**
2007.11 韩国	**df**	243	1	2	243	486	2	486
	σ^2	6.37	1.345	0.003 4	1.786	0.029 5	−0.000 62	0.218
	百分比	**65.32%**	**13.8%**	**0.035%**	**18.31%**	**0.30%**	**0**	**2.24%**
2007.12 国内	**df**	374	1	2	374	748	2	748
	σ^2	1.253	1.816	0.008	0.591	−0.12	−0.00064	0.24
	百分比	**32.06%**	**46.46%**	**0.21%**	**15.12%**	**0**	**0**	**6.14%**
2008.1 韩国	**df**	146	1	2	146	292	2	292
	σ^2	5.093	1.39	−0.008 8	1.49	0.437	0.029	0.724
	百分比	**55.55%**	**15.21%**	**0**	**16.26%**	**4.77%**	**0.316%**	**7.896%**

续表

变异 \ 来源		Person	Item	Rater	pi	pr	ri	pri
2008.3 韩国	df	73	1	2	73	146	2	146
	σ^2	3.349	4.407	0.001 14	1.146 9	0.023 6	0.027	0.209
	百分比	**36.54%**	**48.09%**	**0.0125%**	**12.51%**	**0.258%**	**0.297%**	**2.28%**

* 当方差分量为负数时，极有可能是抽样误差造成的（Shavelson & Webb，1991，转引自杨志明，2003）。按照Cronbaeh等人（1972）的主张，通常把该值设定为0再进行D研究。

从表2可知，在中国国内和韩国的另外六次口试各方面效应的变异分量情况基本与中国国内2007年6月的一致，主要表现在评分员评分的差异对分数变异的影响非常小，在总变异中所占的百分比均不超过0.21%，并且由评分员与被试、评分员与试题之间的交互效应引起的分数变异的比重也很小，这说明评分员自身及评分员之间的信度很高很稳定。另外，被试能力和试题面仍是变异的主要来源。在韩国2007年9月、11月和2008年1月的测试中，被试面的方差分量占总变异的比例最大，最高达到65.32%，这说明大部分的差异来自被试的能力差别，是测试目标期望的分数变异。

通过比较这七次测试的情况，还可以看到，中国国内的三次测试与韩国的四次测试之间有一些不同，国内的测试中，来自试题面的变异分量所占比例明显高于来自被试面的变异分量，而在韩国的测试中，尤其是2007年11月的和2008年1月的测试，来自被试能力（P）的变异分量占总变异的比例最大，分别为65.32%和55.55%，而来自试题（i）的变异分量占总变异的比例分别只有13.8%和15.21%。分析这两次韩国口试的试题，造成这种不同的原因可能在于这两次口试中两道试题的难度没有其他几次口试题之间差别那么大。

在另外六次测试中，被试与试题的交互效应的变异分量所占的比例有所提高，特别是在国内2007年10月的测试中达到20.53%，这部分的变异是相对误差的来源之一，说明考生在两道试题上的表现不太一致，某些能力不强的被试因擅长某些试题在较难的试题上的表现反而更好。

3.5 国内2007年6月BCT口试的D研究

本文主要研究目的是要知道BCT口试的两道试题的样本容量能不能保证测验信度，这就需要进行概化理论的D研究。D研究采用与G研究同样的测量结构和模式，通过对两个测量侧面（题目与评分员）的样本容量（水平取值）的调整，来观察概化系数的变化，考察概括全域的变化对测量信度的影响。我们首先分析国内2007年6月的口试数据，得到表3、表4和表5，这三个表中的结果表示在评分员数分别为3、2、1的情况下，不同的试题数（从1道题到8道题）所对应的相对误差、绝对误差及概化系数、可靠性指数。

表3 国内2007年6月BCT口试P×I×R设计D研究的概化系数和可靠性指数（n′r=3）

效应（α）	n′i=1	2	3	4	5	6	7	8
相对误差 $\sigma^2(\delta)$	0.564 7	0.300 5	0.212 5	0.168 4	0.142	0.124 4	0.111 8	0.102 4
绝对误差 $\sigma^2(\Delta)$	5.169 5	2.603 5	1.748 2	1.320 5	1.063 9	0.892 9	0.770 7	0.679

续表

效应(α)	n′i=1	2	3	4	5	6	7	8
概化系数 $E\rho^2$	0.784 6	0.872 5	0.906 4	0.924 3	0.935 4	0.943	0.948 5	0.952 6
可靠性指数Φ	0.284 7	0.441 4	0.540 6	0.609 1	0.659 1	0.697 4	0.727 5	0.751 8

由表3可知，当依照目前的评分模式固定评分员个数为三个人（n′r=3）时，在试题数n′i=1的情况下，概化系数$E\rho^2$=0.784 6，这表示只用一道口试题目对考生的口头表达能力进行测试，并任选三个评分员进行评分时，这样得到的分数与学生的真实水平的相关可以达到0.88（两者的相关系数等于概化系数的开方）。而当不断增加试题面水平数时，相对误差和绝对误差不断降低，概化系数和可靠性指数也在不断提高。当试题数n′i=2时，概化系数达到0.872 5，说明目前两道试题的口试信度较高。如果希望测试达到0.9以上的信度指标，可以考虑测试三道试题。当试题数为4、5、6、7、8时，概化系数的提高幅度开始变得很小，所以三道试题就足够保证测试的较高信度，更多的试题没有更大的意义。

在同样的条件下，概化系数总是比可靠性指数的值大，原因在于可靠性指数是测量目标的分数变异与全体分数变异之比，其中包括了全部的系统误差和随机误差的变异。表3中测试的可靠性指数相对于概化系数来说明显低了很多，这是因为本次测试的试题面（i）的方差分量所占比例很大（61.27%，见表1），使绝对误差变异增大，可靠性指数降低。从表3来看，当只有两道试题时，概化系数达到0.872 5这一比较高的水平，说明测试作为常模参照测验的话具有很高的信度，但由于可靠性指数仅为0.4414，所以测试不太适合作为标准参照测验进行解释。要想使概化系数和可靠性指数都达到较理想的水平（0.924 3和0.609 1），试题数最好增加到4，但试题数继续增加时，概化系数和可靠性指数增加的幅度并不大。

表4 国内2007年6月BCT口试P×I×R设计D研究的概化系数和可靠性指数（n′r=2）

效应(α)	n′i=1	2	3	4	5	6	7	8
相对误差 $\sigma^2(\delta)$	0.634 2	0.344 4	0.247 7	0.199 4	0.170 4	0.151 1	0.137 3	0.127
绝对误差 $\sigma^2(\Delta)$	5.240 5	2.648 4	1.784 4	1.352 3	1.093 1	0.920 3	0.796 9	0.704 3
概化系数 $E\rho^2$	0.764 4	0.856 6	0.892 5	0.911 6	0.923 5	0.931 6	0.937 4	0.941 9
可靠性指数Φ	0.281 9	0.437 2	0.535 5	0.603 4	0.653	0.690 9	0.720 8	0.745

表5 国内2007年6月BCT口试P×I×R设计D研究的概化系数和可靠性指数（n′r=1）

效应(α)	n′i=1	2	3	4	5	6	7	8
相对误差 $\sigma^2(\delta)$	0.8427	0.4759	0.3536	0.2924	0.2557	0.2313	0.2138	0.2007
绝对误差 $\sigma^2(\Delta)$	5.4534	2.783	1.8929	1.4478	1.1808	1.0027	0.8756	0.7802

续表

效应（α）	n′i=1	2	3	4	5	6	7	8
概化系数 $E\rho^2$	0.709 4	0.812 2	0.853 3	0.875 5	0.889 4	0.898 9	0.905 9	0.911 1
可靠性指数Φ	0.273 9	0.425	0.5208	0.586 9	0.635 3	0.672 3	0.701 5	0.725

表4和表5显示，当评分员为两个的情况下，两道试题的概化系数仍可以达到0.85以上，三道试题和四道试题的概化系数分别可以达到0.89和0.91以上，与三个评分员的结果差别并不是很大。当评分员只有一个时，那么至少需要三道试题才能使概化系数达到0.85以上，要想达到0.9左右的概化系数，就至少需要六到七道试题。

3.6 七次BCT口试的概化系数和可靠性指数比较

按照上面同样的方法根据另外六次口试的G研究方差分量计算其在固定评分者水平数为3的情况下的概化系数和可靠性指数，得到表6：

表6 七次BCT口试P×I×R设计D研究概化系数和可靠性指数（n′r=3）

		n′i=1	2	3	4	5	6	7	8
2007.6国内	$E\rho^2$	0.78	0.87	0.91	0.92	0.94	0.94	0.95	0.95
	Φ	0.28	0.44	0.54	0.61	0.66	0.697	0.73	0.75
2007.9韩国	$E\rho^2$	0.76	0.86	0.90	0.92	0.94	0.95	0.95	0.96
	Φ	0.43	0.60	0.69	0.75	0.79	0.82	0.84	0.85
2007.10国内	$E\rho^2$	0.51	0.67	0.75	0.80	0.84	0.86	0.88	0.89
	Φ	0.23	0.37	0.47	0.54	0.60	0.64	0.67	0.70
2007.11韩国	$E\rho^2$	0.77	0.87	0.91	0.93	0.94	0.95	0.96	0.96
	Φ	0.66	0.80	0.86	0.89	0.91	0.92	0.93	0.94
2007.12国内	$E\rho^2$	0.65	0.79	0.85	0.88	0.90	0.92	0.93	0.94
	Φ	0.33	0.50	0.60	0.67	0.71	0.75	0.78	0.80
2008.1韩国	$E\rho^2$	0.73	0.83	0.88	0.89	0.91	0.92	0.93	0.93
	Φ	0.61	0.75	0.81	0.85	0.87	0.88	0.90	0.90
2008.3韩国	$E\rho^2$	0.73	0.84	0.89	0.91	0.93	0.94	0.95	0.95
	Φ	0.37	0.54	0.64	0.70	0.75	0.78	0.80	0.82

从表6可以看出，当试题数为2时，在韩国举行的四次口试的概化系数与前面分析的2007年6月份的测试一样均达到了0.8以上，国内2007年12月份的概化系数也达到了0.79，因此可以说，目前BCT口试的总体信度很高，也非常稳定。

如果试题数增加到3的话，这七次口试的概化系数有三次可以达到0.9以上，有两次接近0.9（0.88，0.89），一次达到0.85，一次为0.75。同时有五次口试的可靠性指数可以达到0.6以上，

另外两次分别为0.54和0.47。如果试题数超过4的话，概化系数和可靠性指数提高的幅度越来越小。因此，如果BCT口试要进一步提高信度的话，在评分员数为3的情况下，只要增加到三道题就可以保证足够高的信度。即使是把BCT口试作为标准参照测验，主要以可靠性指数来衡量的话，最多也只需四道试题就可以保证相当高的信度。

在表6中我们也发现只有国内2007年10月的概化系数相对较低，为0.67。从前面的表2中可以看到，国内2007年10月口试的被试与试题之间的交互效应（pi）的方差分量比较大，占总变异的百分比达到了20.53%，它所导致的分数变异是相对误差变异的组成部分。当被试与试题之间的交互效应较大时，被试之间的排位顺序会因为试题的不同而出现差异，也就是说，会出现被试甲在第一题上的得分比被试乙高，却在第二题上的得分比被试乙低的情况。造成这种情况的原因可能是这次口试在两道试题的难度或者内容的控制上出现了问题，这需要进一步分析试题，在命题过程中尽量控制试题的难易度并且尽量避免某些考题的内容对考生造成不公平。

4 结果与讨论

4.1 BCT口试的信度

现在我们可以回答本文引言部分提出的两个问题：一是目前BCT口试的信度怎么样？根据第三部分的统计数据和分析，目前BCT口试两道试题、三名评分员评分的测试信度总体上是比较高的，概化系数基本上可以达到0.8以上，测试的科学性较强，是一个可以信赖的测试。二是BCT口试只有两道试题是不是少了？我们的回答是目前两道题的样本容量已经能够达到一个较高的信度，两道题是可行的。如果要进一步提高测量信度，使概化系数达到0.9以上、可靠性指数达到0.6以上的话，比较经济有效的设计方案可以有：评分员个数为3、试题数量为3或者评分员数量为2、试题数量为4。

4.2 BCT口试试题的内容与难度

BCT作为商务汉语考试，是一种特殊目的的汉语水平测试。为了真正测出被试在商务环境中的语言交际能力，试题的语料和任务大多取自真实的商务环境。被试除了要具备一定的语言能力，还应该了解一定的商务专业知识，掌握一些常用的商务专业领域的词汇和表达方式，才能在BCT考试中取得好成绩。这些知识和能力也一个从事商务工作的人应该具备的。BCT口试的两道题正是在这种需求下兼顾了生活和业务两方面的测试内容，并且作为一个跨度考试，要通过一次考试把从初级到高级水平的考生的口语能力区分为五个等级。从本文的分析中也可以清楚地看到这种测试方式是比较经济有效的。

但凡事都有利有弊，试题要对不同水平的人有很高的区分度，题目的难度必然会有很大的差异，这正是造成试题主效应的方差分量比较大、测试的绝对误差增大、可靠性指数较低的主要原因。因此要改进BCT口试，首先要注意的是两道试题之间的难度不宜相差过大，要避免第一题过于容易和第二题过难。其次，国内2007年10月的测试概化系数相对较低，主要是因为来自被试与试题之间的交互效应的分数变异较大，这可能与试题涉及不同的商务业务层面，一些被试对这些业务内容的了解和掌握的程度不同有关。从这个角度来说，试题在内容上对商务领域的业务覆盖面要尽量大一些。要做到这一点，有两个办法：一是增加一道试题，可以加大试题对商务任务的覆盖面，增加考试信度；二是在选择交际任务时，既要选择有代表性的业务

内容，又要使这些内容为非专业的普通人所熟悉的，测试要在这两者之间找到最佳的平衡点，以避免试题内容对考生造成不公平的现象，减少来自被试与试题之间的交互效应带来的分数变异，提高测试信度。

参考文献

1 中国国家汉语国际推广领导小组办公室，北京大学商务汉语考试研发办公室．商务汉语考试大纲［S］，北京：北京大学出版社，2006：9–13.

2 刘远我，张厚粲．概化理论在作文评分中的应用研究［J］．心理学报，1998，30（2）．

3 王佶旻．概化理论在汉语初学者口语测验中的应用［J］．云南师范大学学报（对外汉语教学与研究版），2007，5（2）．

4 王占礼．试卷样本同质性对概化理论测评精度的影响［J］．青岛远洋船员学院学报，2004（4）．

5 王占礼，张红梅．SEPT口试方案设计［J］．外语电化教学，2005（102）．

6 杨志明．标准参照测验及其等级线信度的概化理论分析［J］．心理学探新，2003（3）

7 杨志明，张　雷．改进普通话测试的概化理论分析［J］．湖南师范大学教育科学学报，2003，2（1）．

8 杨志明，张　雷．测评的概化理论及其应用［M］．北京：教育科学出版社，2003.

9 Cronbach, Rajaratnam & Gleser. Theory of Generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963:16.

10 Cronbach, L.J., Gleser, G.C., Nanda, H.& Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.