บทบาทของทฤษฎีการทดสอบทางภาษาในการเรียนการสอนภาษาจีน ในฐานะภาษาต่างประเทศ

หนิวปิน

วิทยาลัยนานาชาติจีน มหาวิทยาลัยครุศาสตร์หนานจิง หนานจิง สาธารณรัฐประชาชนจีน 210046

อีเมล: niubinnk@126.com

รับบทความ: 4 มิถุนายน 2561 แก้ไขบทความ: 7 พฤศจิกายน 2561 ตอบรับบทความ: 22 ธันวาคม 2561

บทคัดย่อ: การทดสอบทางภาษา ถือเป็นส่วนสำคัญในการเรียนการสอนภาษาต่างประเทศที่สอง การวิเคราะห์ตามแนวคิดวิทยาศาสตร์จะช่วยในการประเมินผลการสอนได้ดียิ่งขึ้น บทความฉบับนี้ใช้ข้อสอบปลายภาคของรายวิชาภาษาจีนระดับกลางจากมหาวิทยาลัยที่มีชื่อเสียงมาเป็นแบบทดสอบ โดยใช้โปรแกรม SPSS 15 เป็นเครื่องมือในการทดสอบ จากการประเมินความถูกต้องและประเมินความน่าเชื่อถือของตัวเลขที่ได้จากการทดสอบ พบว่าข้อสอบดังกล่าวมีความน่าเชื่อถือและมีความแม่นยำสูง แต่อย่างไรก็ดียังคงพบปัญหาอยู่บ้าง บทความฉบับนี้ต้องการแสดงให้เห็นว่าในการเรียนการสอนภาษานั้นต้องให้ความสำคัญกับการทดสอบทางภาษาและยิ่งไปกว่านั้นคือต้องทำให้แบบทดสอบแสดงบทบาทที่สมบูรณ์ด้านการวัดผลสัมฤทธิ์ในเรียนการสอนภาษาในมากที่สุด

คำสำคัญ: การทดสอบทางภาษา; ความน่าเชื่อถือ ; ความถูกต้อง

The Role of Language Testing Theory in Teaching Chinese as a Foreign Language

Niu Bin

International College for Chinese Studies, Nanjing Normal University, Nanjing, 210046, P.R. China

Email: niubinnk@126.com

Received: 4th June 2018 Revised: 7th November 2018 Accepted: 22nd December 2018

Abstract: Language testing is an important part of second language teaching. Scientific analysis of tests helps us to better evaluate the teaching effect. This paper uses a mid-level Chinese final examination of a top university in China as an analysis sample, and uses SPSS 15 software to evaluate the reliability and validity based on data. The results show that this examination has a high degree of reliability and validity, but also has some problems. This paper believes that Chinese teaching should pay attention to language testing and make language testing play an active role.

Keywords: language testing; reliability; validity

语言测试理论在对外汉语教学中的作用

——以某份中级汉语期末考试试题为例

牛彬
南京师范大学国际文化教育学院
电子邮箱：niubinnk@126.com

收稿日期：2018-06-04　修回日期：2018-11-07　接受日期：2018-12-22

摘　要：语言测试是第二语言教学的重要组成部分，对试卷进行科学的分析有助于我们更好地评估教学效果。本文以某中国重点大学的一份中级汉语期末试卷为分析样本，以 SPSS 15 软件为分析软件，根据数据对试卷的信度和效度进行评估。检验结果表明这份试卷具有比较高的信度和效度，但也存在一定的问题。本文认为对外汉语教学应重视语言测试，发挥语言测试的积极作用。

关键词：语言测试；信度；效度

0 引言

语言测试（language testing）是第二语言教学的重要组成部分。现代语言测试与传统的语言考试并不相同[1]。传统的考试观认为学生成绩不好，仅是由于学生单方面的原因造成。但现代语言测试理论认为，学生成绩不好有多种原因，除了学生自己的因素，还涉及到教材、教学法、教师的教学等多个相关因素。因此，要想正确地测量学习者语言能力，必须要有科学的测量工具，这就要求教师应该设计出具有较高信度（reliability）和效度（validity）的试题，并且能根据测试情况，对学生学习情况、教学内容和教学法等做出评估。

随着中国的综合国力不断增强，越来越多的留学生开始学习汉语。中国的各所高校每年都会接收很多来自不同国家的留学生。每个学期末，这些高校都会对留学生进行期末测试，以此对一个学期的汉语教学进行总结和评价。但是根据我们的观察，很少有中国大学的汉语教学机构用语言测试理论对留学生汉语期末试卷的信度和效度进行分析。

语言测试理论认为，教育测量和评估可以保证语言测试的质量（Davys 1998）。本文将根据某中国重点大学 2014-2015 学年中级综合汉语期末测试中的考生样本资料，对该试卷的各题目做出信度和效度的分析和评估，并借此对汉语教学提出一些建议。

“本研究得到南京师范大学国际文化教育学院重点项目“来华留学预科教育研究”（ZD20161201）的资助。”

[1]在教育测量与评估领域，一般对“考试”（examination）和“测试”（test）进行严格区分根据 Davys（1998）的观点，考试是测试的特殊形式。测试的影响较大，并要求有详尽的计划和管理，而考试则无需如此精心准备。本文采用这种观点，下文除了为保持试卷原貌而保留“考试”这一术语外，一律用“测试”这个术语。

1 语言测试的相关概念

信度和效度是衡量语言测试是否有效的重要指标（Bachman 1990）。信度是测验分数的稳定性和一致性程度的指标。它可以反映测试结果受到随机误差的影响程度，是评价测试质量的最基本的指标。高信度是高效度的前提，没有信度，效度也谈不上。一个可靠的测试结果应该不受诸如测试取样不好、测试过程受到干扰、考生情绪不稳、身体不适等偶然因素的影响。如果测验的信度不够理想，该测验将是无意义的测验。效度是指一项测试的有效性。效度是针对测试目的而言，即最大程度地测试出需要测量的语言结果。从测试角度看，测试结果应达到出题人预期的目的，即测量出考生的语言知识或能力。如果测试结果能达到出题人的预期，该测试就有效度。测试结果与出题人的预期越符合，该测试的效度就越高。

成绩测试（achievement test）是语言测试的一种。这种测试方法是考察学生对所学知识的掌握程度，其目的是要“衡量对所讲授的内容的掌握程度，或者对教学大纲、教科书及其他资料的掌握程度”（Davys 1998：8）。平时测试和期末测试都属于成绩测试。考生成绩分析对于教师充分了解学生对学习内容的掌握情况以及班级的均衡和差异情况非常重要，因此对试卷进行信度和效度分析对于命题教师改进命题工作、提高教学质量意义重大。

2 留学生情况和测试题型

本研究以参加某中国重点大学 2014-2015 学年第二学期中级综合汉语期末测试的留学生为分析对象。他们是 2015 年春季班的留学生，一共 12 人，其中韩国学生 8 人，日本学生 2 人，俄罗斯学生 1 人，泰国学生 1 人。测试分为七个题型：选择、选词填空、组句、改写句子、完成句子、阅读理解、

作文。其中选择、选词填空和阅读中的判断题为客观题，其他题为主观题。测试时间为 100 分钟。题型的题量和分数分布如表 1 所示：

表 1　中级综合汉语期末考试题型、题量和分数

题型	选择	选词填空	组句	改写句子	完成句子	阅读	作文
题量	15	4	4	4	5	6	1
分数	30	8	4	4	10	8	36

留学生在各题的得分情况如表 2 所示：

表 2 留学生在各题中的得分情况

编号	总分	一（30 分）	二（8 分）	三（4 分）	四（4 分）	五（10 分）	六（8 分）	七（36 分）
	（100 分）	选择	选词填空	组句	改写句子	完成句子	阅读	作文
1	72	22	6	1	4	6	3	30
2	85.5	24	8	4	4	7.5	6	32
3	83.5	22	8	2	3.5	9.5	5.5	33
4	83.5	24	8	3	3.5	9	4	32
5	85	27	8	2.5	4	7.5	4	32
6	61	14	6	1.5	3.5	6	3	27
7	78	24	8	2	4	9	3	28
8	88	24	8	4	3.5	9.5	7	32
9	80.5	25	6	2.5	4	10	4	29
10	77.5	25	8	0.5	4	8	3	29
11	72	19	5	0.5	1.5	8.5	7.5	30
12	73.5	23	8	1.5	3	6	4	28
平均分	78.33	22.75	7.25	2.08	3.54	8.04	4.5	30.17

3 试卷分析

3.1 信度分析

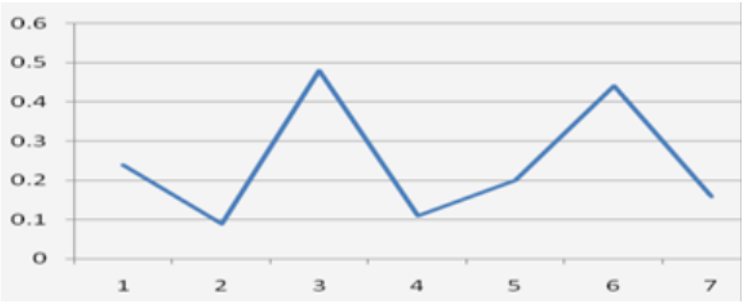
本研究通过 SPSS 15 得到各题和总分的信度和难度系数，具体见表 3。

表 3　试卷的各题与总分的信度

题号	均值	难度系数	标准差	信度
1、选择	22.75	0.24	3.387812381	0.750
2、选词填空	7.25	0.09	1.138180366	0.414
3、组句	2.08	0.48	1.183856052	0.452
4、改写句子	3.54	0.11	0.721687836	0.111
5、完成句子	8.04	0.2	1.453183548	0.377
6、阅读	4.5	0.44	1.60963011	0.273
7、作文	30.17	0.16	1.99240984	0.556

试卷的难度系数曲线如图 1 所示

图 1 试卷的难度系数曲线



从表 3 可以看出选择题的信度很高，其他题基本上接近或超过 0.4，但是在“改写句子”和“阅读”这两个题信度明显比其他题低。根据计算，整个试卷的信度是 0.707。语言测试理论认为信度系数在 0-1 之间波动。一份普通试卷的信度系数如果可以达到 0.7 就算合格的试卷（景美霞 2006）。本

套试卷不是标准化测试的试卷，其信度系数达到 0.707，已经可以说具有较高的信度。这表明本测试基本是可信而稳定的。

总体来看，试卷整体信度比较高。但是在个别题上信度不够，比如第四题“改写句子”和第六题“阅读”。第四题题目要求是“请用括号里的词语改写下面的句子”，括号中的词主要是副词，还有一个“你 V 你的”的固定搭配，主要考察学生对词语的理解和把握能力。这个题的难度是 0.11，是这份试卷中一道简单的题，所以学生普遍答得很好，平均分达到 3.54 分（满分 4 分）。这说明学生对这些词语已有了很好地掌握。但是从整个试卷而言，这个题没有体现出区分度，与试卷整体的一致性和稳定性也较低，所以这个题的信度不高。与第四题形成对比的是第二题选词填空，这个题目的难度系数是最低的，但区分度却不低。这一方面说明难度和区分度之间并不一定呈正相关的关系，另一方面也反映试卷题目本身没有问题，关键在题目内容的考察取舍。在第四题考察内容的取舍上，老师还应进一步斟酌。

第六题是“阅读理解”，里面既有客观题（如判断题），也有主观题，各占一半的分数，所阅读文章的题目是《品尝生活的橄榄》。这篇文章最大的特点是有较多的成语，例如“福祸相依”、“痛不欲生”、“不知生，焉知死”、“大刀阔斧”、“时运不济”、“天崩地裂”等等，文中还多处运用了比喻和拟人的修辞手段，如“痛苦像一把犁，它一面犁破了你的心，一面掘开了生命的新起源”等。6 道阅读题中只有两道题考查对全文的理解，其他题目都是考查对词语和修辞的理解掌握能力。阅读题在整个试卷中的难度系数达到 0.44，是比较难的题。从结果来看，学生对阅读题回答得不是很好，这可能是因为题目中很多带有文化色彩的成语超出中级汉语的学习范围。

这个题给我们的启示是老师在设置阅读理解题目时，应尽量考察学生对全文的理解，而不是对部分超纲的词语。

试卷的第一题是 30 分的选择题，主要是对课文中各个语法点的考察，例如趋向动词、量词、定语语序等，另外还考察了学生对重要词语的理解。本题的信度非常高，与其它题的相关度也比较高。这说明这个题在形式和内容上设计得都比较成功。在中级阶段语法教学非常重要，我们可以通过对语法点的测试准确了解学生的语言水平，而用选择题的形式可以快捷地测试出学生对语法点的掌握情况。

第三题“组句”、第五题“完成句子”和第七题作文都有一定的难度，“组句”的难度达到 0.48，是这套试卷最难的一道题。但是这些题都具有良好的区分度，信度也比较高。这部分题是主观题，要求学生根据所给的词填空、造句或者根据前文补充句子。题型虽然难度大，但检验效果也比一般的客观题好，能够较好地考出学生的语言水平和能力。这些题不是简单的对语言知识的考察，而是更侧重于语言的应用。这也说明主观性的、综合性的题型更便于考察语言的实际应用能力。

3.2 效度分析

效度主要观察一项测试能否实现预定的教学目标，以及试卷试题能否真实考查需要考查的知识和能力（曹昭 2015）。本文通过相关系数指标考察试卷的内容效度（Face Validity），即测量目标与内容之间是否吻合。本文采用单项题目与总分相关分析法获得评价结果，即计算每个题项得分与题项总分的相关系数，根据相关是否显著判断是否有效。本研究使用 SPSS 15（社会科学统计软件包）和 excel 软件对各题目进行了相关系数的分析。试卷内部各题与总分的相关系数见表 4。

表 4　试卷内部各题间及各题与总分的相关矩阵

	选择	选词填空	组句	改写句子	完成句子	阅读	作文	总分
选择	1.000(*)	.607(*)	.380	.469	.399	-.033	.438	.814(**)
选词填空	.607(*)	1.000	.489	.484	.158	-.099	.318	.643(*)
组句	.380	.489	1.000	.368	.368	.346	.552	.712(**)
改写句子	. 469	.484	.368	1.000	-.002	-.607(*)	.058	.316
完成句子	.399	.158	.368	-.002	1.000	.398	.421	.636(*)
阅读	-.033	-.099	.346	-.607(*)	.398	1.000	.553	.394
作文	.438	.381	.552	.058	.421	.553	1.000	.793(**)
总分	.814(**)	.643(*)	.712(*)	.316	.636(*)	.394	.793(**)	1.000

（*）表示 0.001 水平显著

表 4 显示的数据是试卷内部各题间以及各题与总分的相关矩阵（regression matrix）。相关矩阵会把一套试题各大题之间的相关系数算出并排成矩阵，是语言测试统计分析的一部分。每道题和自身的相关系数为 1.000，不同题目相关系数为 0-1 之间，如第三题和第一题的相关系数为 0.380。对于非标准化考试来讲，接近或达到 0.4 就有一定意义的相关。达到 0.5 以上，就可以说有相当显著的相关（景美霞 2006）。如表 4 所示，本试卷大部分试题能彼此区分，具有正相关关系，特别是在“总分”一栏中，相关度尤为高。但是我们也要看出，“改写句子”和“阅读”这两道试题的相关度较低，甚至出现了负相关。总体来看，本测试具有较高的内容效度。

效度也可以从测试形式和测试内容两方面进行逻辑分析，即分析题项和内容是否符合测试的目的和要求。测试形式涉及题型设计和主客观题的分布等方面。从题型设计上来看，本次测试题型类别多样，分布合理。这份试卷题型分为七大种，在个别题里还会出现其他题型，如排序题、判断题等。另外，试题设计也十分注意教学的目的性，课文中的语言点通过多种设计形式呈现出来。从主观题和客观题的分布上看，本套试卷主客观题搭配较为合理。

通过数据我们可以看出，这套试卷在客观题上设置得比较成功。出题人用客观题考察了学生对汉语学习词汇、语法点、句子整体的把握能力，课文中出现的语法点和重要词语，如量词、趋向动词、副词等，在测试中都有所体现。数据显示，这些题的信度比较高，说明出题人对这种题型的利用是比较充分的。试卷的主观性题型虽然有一定的难度，但能够较好的考出学生的语言水平和能力。试卷的主观题减少了对语言知识的考察，更侧重于语言运用。例如填空题“她唱得简直太好了，我们不禁____”， 给学生提供词语，根据情景答题，答案是半开放的，而且实用性也比较高。语言测试理论认为，主观性的、综合性的题型便于考察语言的实际应用能力，但是如果主观性题目太多，也可能会导致试卷效度的降低。本试卷的出题人注意到了这个问题，对主、客观题型的分布较为合理，命题质量普遍比较高，保证了试卷的效度。

就测试内容而言，汉语教学主要考查学习者运用汉语的准确性和交际能力，这也体现了第二代和第三代语言测量理论的理念。第二代语言测量理论是语言测试的心理测量—结构主义时期，该理论认为语言是由语音、词汇、语法组成的系统，这一系统是可以分解的，因此我们可以设计出具有客观性的离散的题目，但是这就忽视了这些要素在更大的交际环境中的互动。第三代语言测量理论强调语言是一种动态的、具有创造性的功能体系，强调学习者在真实（authentic）语境中运用语言达到交际目的的能力。这一理论认为掌握一种语言是指在真实语境中“能够使用所学的语言进行有效的交际，交流思想感情，达到相互沟通的目的”（Alderson, Clapham & Wall 2000：11）。我们在笔试中很难制造真实的语境，但是本试卷在主观题和客观题中都有对语境的强调，例如在选择题中“我的自行车被朋友___走了”，选项

“偷”和“借”都符合语法，但是“借”才符合语境中“朋友”这个词。从这个题的设置可以看出，出题人具有将两代语言测试理论有机结合的意识。

本套试卷不太成功的地方反映在“改写句子”和“阅读”这两个题目上。这两个题目的整体效度比较低。在相关度方面，“改写句子”和“完成句子”、“阅读”甚至是负相关。本文觉得问题不在题目类型的本身，而是在题目的内容上。“改写句子”这个题内容的取舍上有一些问题，区分度不明显。在阅读题中存在大量抽象性词汇。在汉语学习的中级阶段，抽象词汇的教学会大量增加，其中成语和固定短语比较多。留学生学习的内容既有现代汉语知识，又有古代汉语知识。在词汇量有限的情况下，留学生学习这些内容的困难是不少的。老师在教学时应重点对这些抽象性词汇进行教学，使学生对其有较好的掌握。从这份试卷来看，学生对汉语的抽象词汇的掌握不好，从而使学生成绩区分度低，效度也因此较低。这提示汉语教师应加强对此类词语的教学。

4 结论及其对汉语教学的启示

本文对某中国重点大学的一份中级汉语期末试卷进行信度和效度分析。总体上看，这套试卷具有比较高的信度和效度，基本上可以正确反映学生的学习水平。试卷的出题思路有不少可以借鉴的地方。例如选择题包含了多方面教学内容，主观题和客观题的题型分布较为合理，试题设计也有一定的针对性和实用性。另外，试卷考查的内容不仅涉及语言知识，对语境也比较重视。但是也要注意到该试卷有些题目的区分度和难度系数不够，影响到整个试卷的信度和效度。这暴露出来一些教学问题，需要进一步解决。

语言测试是二语教学过程的重要组成部分，其结果往往被作为评估教师水平、教学效果和教学质量的重要依据，也是用来评估学生表现、做出相关决定的重要依据。有学者甚至认为“语言测试在语言教学中处于中心地位”（Davys 1998:1）。我们对试卷进行评估，不仅仅是为了保证试卷可以正确反映学生水平，同时还应根据评估结果发现教学问题，为下一步教学提供一个正确的侧重点。因此，对外汉语教学应该重视语言测试，认真地对待每一次测试的出题，努力使语言测试的内容和方法与教学目标相一致，促进教学目标的实现，以最大限度地发挥语言测试对教学和学习的积极促进作用。

参考文献

曹昭（2015）河北省汉字应用水平测试信度和效度分析，《语言文字应用》第 3 期。

景美霞（2006）语言测试中试卷分析的方法，《伊犁教育学院学报》第 4 期.。

林连书（2001）《应用语言学实验研究方法》，广州：中山大学出版社。

孙成岗（2000）现代语言测试与试卷分析，《解放军外国语学院学报》第 4 期。

谢小庆主编（2005）《中国汉语水平测试（HSK）研究报告精选》，北京：北京语言大学出版社。

张凯（2005）《语言测验理论与实践》，北京：北京语言大学出版社。

张凯主编（2005）《语言测试及测量理论研究》，北京：北京语言大学出版社。

Bachman, Lyle F. (1990). *Fundamental Considerations In Language Testing*. Oxford: Oxford University Press.

Davys, A.著，任福昌等译（1998）《语言测试原理（*Principles of Language Testing*）》，北京：经济科学出版社。

J. Charles Alderson, Caroline Clapham & Dianne Wall. (2000). *Language test construction and evaluation*. 北京：外语教学与研究出版社.