国家社会科学规划基金资助项目（13CYY032）

# 标准化语言测试的标准制订与效度研究

## The Development and Validation of Standards in Language Testing

范劲松 著

復旦大學出版社
Fudan University Press

国家社会科学规划基金资助项目（13CYY032）

# 标准化语言测试的标准制订与效度研究

## The Development and Validation of Standards in Language Testing

范劲松 著

复旦大学出版社
Fudan University Press

**图书在版编目(CIP)数据**

标准化语言测试的标准制订与效度研究 = The Development and Validation of Standards in Language Testing:英文/范劲松著. —上海:复旦大学出版社,2018.8
ISBN 978-7-309-13704-0

Ⅰ. 标… Ⅱ. 范… Ⅲ. 语言-测试-研究-英文 Ⅳ. H09

中国版本图书馆 CIP 数据核字(2018)第 105717 号

**标准化语言测试的标准制订与效度研究**
范劲松 著
责任编辑/唐 敏

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143 出版部电话:86-21-65642845
常熟市华顺印刷有限公司

开本 890×1240 1/32 印张 8.75 字数 223 千
2018 年 8 月第 1 版第 1 次印刷

ISBN 978-7-309-13704-0/H·2824
定价:35.00 元

# 中 文 摘 要

随着语言考试在社会生活中发挥的作用越来越重要，语言考试在开发、实施和使用过程中的专业化问题成为近些年国际语言测试界关注的热点话题之一。实现语言考试专业化的重要手段之一是制订和实施语言测试标准。自上个世纪八十年代以来，我国的语言测试理论与实践发展迅猛。经过近三十年的发展历程，我国已经成为名副其实的语言测试大国。以英语为例。目前我国各种各样英语考试层出不穷，每年有数以千万计的英语学习者参加英语考试。我国的语言考试具有规模大、风险高、对语言教学反拨作用显著等特点。因此，这些考试在开发、实施和使用的过程中是否严格遵循相关的语言测试质量标准，考试是否能够有效地反映考生的语言能力水平，对考试结果的使用是否合理，这些问题不仅是教育和考试管理部门关注的重点，也是考生、教师乃至社会公众讨论的热点话题。

本研究的主要目的是对目前国际上已经制订和实施的语言测试标准进行详细的回顾和分析，并且对我国语言测试的现状进行深入、细致的调查和研究，为我国教育和考试管理部门在制订和实施语言测试质量标准的时候提供重要参考。本研究共分为三个阶段。在第一阶段，我们对语言测试开发、实施和使用的理论和实践进行了详细的回顾，以此确定指导本研究的理论框架。在众多现有的理论框架中，Bachman & Palmer（2010）提出的测试使用论据（Assessment Use Argument，AUA）集中体现了语言测试研究领域的最新成果，对语言考试的开发和评估具有重要的指导作用。因此，我们在本研究中采

用了 AUA 作为理论框架。在第一阶段，我们还收集、整理并分析了国际语言测试研究领域已经制订和实施的语言测试标准。我们以 AUA 框架为基础，对其中的五项标准进行了详细分析，包括国际语言测试协会（ILTA）制订的《ILTA 道德规范》和《ILTA 行为准则》、《教育和心理测试标准》、《ETS 质量和公平标准》以及《EALTA 语言测试和评估良好行为准则》。对于这些标准的系统分析能够为我国制订语言测试标准提供重要参考。

在第二阶段，我们在 AUA 框架的指导下对我国的英语测试现状进行了详细的调查研究。首先，我们采用问卷和一对一访谈相结合的方法研究了我国大学入学分级考试在开发、实施和使用各个步骤的具体做法，以此为例探讨我国语言测试的现状以及在考试过程中所面临的困难和挑战。全国共有 63 所高校参与了本部分的研究，并有 5 所院校的分级考试负责人参与了一对一访谈。研究结果表明，大学英语入学分级考试普遍缺乏标准意识，很多做法存在问题。例如，很少有高校开发分级考试细则，也很少有高校对命制的试题进行试测。我们在研究中发现，分级考试在部分学校与学生的学分以及课程选修计划挂钩，因此具有一定风险性。但是，考试开发过程中缺乏严格的质量控制措施可能对考试的质量产生影响，进而影响到基于考试结果所做出的决定的效度和合理性。因此，本部分的研究结果进一步说明在我国制订和实施语言测试标准的必要性和紧迫性。

在本阶段的研究中，我们还采用定性和定量相结合的研究方法探讨考试利益相关群体对我国英语测试现状的看法，以此进一步勾勒出我国语言测试的环境特征。在定性研究阶段，我们采用开放式问卷（$n=248$）、一对一访谈（$n=14$）和核心组讨论（$n=12$）等方法收集学生数据，并采用一对一访谈（$n=12$）收集教师数据。所有的访谈和核心组讨论数据收集完成后进行转写，总计达 10 万多字。然后，我们采用定性分析软件 NVivo 对数据进行详细的编码和分析。分析结果表明，学生和教师对于语言测试的基本看法相似，均认为考试的设计和考试对教学的反拨效应是区分好的语言考试与不好的语言考

试的重要标志。两个群体整体上对于我国英语考试的现状评价比较积极。两个群体均认为口语能力在目前的语言考试设计中没有得到充分体现,考试成绩报告不够细化,无法为学生和教师提供丰富的反馈信息等。这些研究结果为我国的语言考试机构进一步提高考试质量提供了重要参考。基于定性研究结果,我们设计了用于收集定量数据的问卷。全国共有 381 名学生和 100 名教师参与了本部分研究。问卷数据收集完成以后,我们采用多种定量研究方法对其进行分析,包括探索新因子分析、信度分析、描述统计分析、配对样本 $t$ 检验等,以进一步深入探讨利益相关群体对语言测试现状的看法。该部分的研究结果与定性部分基本吻合,表明学生和教师对我国语言测试的现状总体看法比较积极,但是教师的看法比学生更为积极。在 EFA 所抽取的五个因子中,两个群体对考试实施和管理的评价最为积极,对考试设计的评价最不积极。本阶段的研究结果比较完整地展示了我国语言测试标准制订和实施的环境特征。

基于前两个阶段的研究结果,在第三阶段,我们详细探讨了我国制订语言测试标准的目的、方法、内容和基本原则,并尝试确定语言测试标准的基本框架。提高语言考试的信度、效度和公平性不是考试开发机构单方面的责任,而是考试利益相关群体的共同责任,这是制订语言测试标准需遵循的重要原则。也就是说,语言测试标准不仅可以为语言考试机构在考试开发和实施过程中提供重要指导,也能够方便教育和考试管理部门有效监管考试开发机构。同时,语言测试标准有助于提高利益相关群体的语言评估素养,使他们有效地参与到考试的过程中,提升考试的透明度。最后,语言测试标准能够帮助考试的使用者做出正确的决策。我们认为,传统的自上而下的标准制订方法不利于确保标准的效度。采用双向互动的方法既能够考虑到语言测试的共性,也能够有效体现我国语言测试的环境特征,因此是更为合理的标准制订方法。基于以上考虑,我们确定了标准的框架并尝试制订了语言测试开发和实施的标准。在本阶段的研究中,我们还尝试将该部分的标准应用到一项校本英语考试(FET,复

旦大学英语水平考试)的开发与实施中。为了探讨标准在实践中的应用效果,我们首先对考试的数据($n$=3 988)进行结构方程建模分析;其次,我们采用定量和定性相结合的方法探讨学生对考试的评价($n$=157)。研究结果表明,FET 的效度比较理想,考生对 FET 的评价也比较积极,认为考试有效地测量了他们的英语水平。语言测试标准在提高 FET 的质量方面发挥了重要作用。

本研究是制订和实施语言测试标准的尝试,相关研究结果对我国的语言考试机构进一步提高语言考试的质量和专业化水平具有重要的参考价值,对于教育和考试管理部门在将来制订和实施语言测试标准也具有借鉴意义。我们认为,制订和实施语言测试标准是提高语言考试信度、效度和公平性的重要举措,必将对我国的外语教育、人才培养和社会进步产生重要而深远的影响。

# ENGLISH ABSTRACT

As language tests are playing an increasingly important role in society, the pursuit for professionalism in language testing operations has in recent years become one of the foci of the international language testing community. The development and implementation of language testing standards provide a viable approach towards professionalism. In China, language testing theory and practice has been developing with great momentum since the 1980s, and as a result, China has grown into a huge testing country today. Take English language testing as an example. Every year, tens of millions of English learners take English tests in China. English language testing in China is defined by several salient features, including its colossal scale of operation, its high stakes, and the significant washback effects on English teaching and learning at all levels. Therefore, legitimate concerns have been raised about whether test developers subscribe to rigorous quality standards and whether these tests can accurately reflect students' language abilities.

The present study serves three purposes. First, it intends to conduct a systematic review and analysis of the language testing standards which have already been developed and implemented worldwide. Second, it aims to portray the contextual features of language testing in China through an in-depth empirical investigation into the *status quo* of language testing practices. Third, it represents a

preliminary attempt to develop and implement standards in language testing practices and examine their impact on the quality of language tests. By serving these three purposes, this study aims to provide insights for educational and examinations authorities when they set out to develop and implement standards in the future.

In view of the three purposes, the study consists of three phases. During Phase 1, we reviewed the theories and practice of language test development, administration, and use, to determine the theoretical framework for this study. Among the many frameworks that are currently available, the Assessment Use Argument (AUA), proposed by Bachman and Palmer (2010), was selected for this research as its guiding theoretical framework, thanks to its comprehensiveness and impact. During this phase, we also collected and analysed some leading professional standards in the field of language testing. Specifically, we conducted a systematic and detailed analysis of five leading standards, including *ILTA Code of Ethics*, *ILTA Guidelines for Practice*, *Standards for Educational and Psychological Testing*, *ETS Standards for Quality and Fairness*, and *EALTA Guidelines for Good Practice in Language Testing and Assessment*. The review paved the way for the discussion on developing and implementing language testing standards in China.

During Phase 2, we conducted an in-depth investigation into the *status quo* of language testing practices in China, under the guidance of the AUA. First, we examined the procedures that universities followed in the development of English tests for placement purposes. Research data were collected through questionnaires and one-on-one interviews. Sixty-three universities participated in the questionnaire survey; representatives from 5 universities accepted our requests for follow-up interviews. Results show that the test development procedures which

most universities followed were not consistent with relevant standards and did not suggest good awareness of quality and professionalism. For example, few universities developed test specifications; nor did they pilot their test items. It was worth noting that placement tests had reasonably high stakes in some universities because their scores were associated with students' credits and might largely determine the English courses from which students could select. Lack of quality control measures tended to affect the quality of tests, and worse still, the validity and appropriateness of the decisions made based on the scores. Results from this survey therefore further heightened the necessity and urgency of developing professional language testing standards in China.

We also surveyed stakeholders' views on language testing practices through a mixed-methods research design during Phase 2. The design featured the combination of a qualitative component and a follow-up quantitative component. In the qualitative component, we used open-ended questionnaires ($n=248$), one-on-one interviews ($n=14$), and focus group discussions ($n=12$) to collect data from students, and one-on-one interviews ($n=12$) to collect data from teachers. All data collected were transcribed verbatim, which amounted to around 100,000 Chinese characters. We then used NVivo, a qualitative data analysis software, to code and analyse the data. Results indicate that students and teachers had similar views on language testing. Both groups believed that test design and washback were the two essential features which distinguished a good language test from a bad one. Overall, both groups made positive comments on the *status quo* of language testing in China, regarding test development, administration, and use. Nevertheless, both groups believed that speaking was not well represented in test design, and test takers should be provided with more detailed feedback about their performance. These findings are

meaningful to test developers in their efforts to further improve testing practices. Based on the findings from the qualitative component, we developed a questionnaire to collect quantitative data from 381 students and 100 teachers from across the country who participated in the survey. The data collected were subsequently subjected to a variety of analyses, including exploratory factor analysis (EFA), reliability analysis, descriptive statistics, and independent-samples *t*-tests. Results generally aligned with the findings yielded from the qualitative component. Both groups were found to have generally positive views on language testing practices, although teachers' views were found to be more positive. Among the five factors extracted by EFA, the two groups had the most positive comments on test administration and the least positive ones on test design. These findings help to portray a comprehensive and more detailed picture of language testing practice in China.

Based on the research in the first two phases, we went a step further to discuss the purpose, methods, content, and guiding principles in developing standards for language testing in China during Phase 3. We argued that maintaining test reliability, validity, and fairness is the shared responsibility of all stakeholder groups. This principle was then applied to determining the structure of the standards. In other words, examination boards can use standards to guide test development and administration; relevant supervisory authorities can use standards to facilitate their supervision of the examination boards; stakeholder groups like students and teachers can use standards to improve their language assessment literacy; and test users can apply the standards to use language tests in a more responsible way. It was also argued that the interactive approach, which is a combination of the top-down and bottom-up approach to standards development, could better

reflect both the commonalities in language testing practices and the salient features of a particular testing context. With these principles in mind, we set out to develop tentative standards in relation to test development and administration. During this phase, through a case study, we explored whether the implementation of the standards could yield any impact on the quality of language tests. The standards were piloted on a university-based English proficiency test, known as the Fudan English Test (FET). The standards were rigorously observed in the development and administration of the FET. To examine the quality of the FET, we first modelled the test data ($n = 3\ 988$), using structural equation modelling (SEM). Next, we employed a mixed-methods approach to investigate students' evaluation of the test. Results from the SEM analysis lent support to the construct validity of the FET; students expressed generally positive views on the quality of the FET. These findings suggest that standards might have played a vital role in enhancing the quality of language tests.

In conclusion, this study represents an attempt to develop and implement language testing standards based on empirical data collected from different perspectives and sources. Findings from this study have significant implications for educational and examination authorities in their endeavour to develop language testing standards in China in the future. The development and implementation of professional standards is undoubtedly a significant move towards reliability, validity, and fairness of language tests, which, in turn, will have a far-reaching impact on foreign language education, cultivation of talent, and social progress in China.

# 序

“标准”一词在我们的日常生活中司空见惯。例如，企业在生产产品时需参照产品质量标准，达不到质量标准的产品无法进入销售环节；单位在招聘新员工的时候有遴选标准，只有通过遴选标准的候选者才能进入到面试环节。在语言测试领域，标准既可以指语言水平或某一语言水平所代表的语言知识和技能，也可以指语言测试工具开发、实施和使用过程中所应遵循的质量标准或行业规范。范劲松老师的新著《标准化语言测试的标准制订与效度研究》中的“标准”属于后一种理解。该书基于作者主持的同名国家社科基金项目。

作者在书中指出，本书的目的并非是要制订一项语言测试标准。作为语言测试研究人员，作者的真正意图是通过大量的实证研究，深入了解目前我国语言测试在开发、实施和使用等环节的现状，从而为我国教育和考试主管部门将来制订语言测试标准的时候提供参考。

纵观本书，作者开展了大量有关语言测试标准的实证研究。首先，作者分析了世界各地已经颁布的多项语言测试行业标准。其次，作者采用定量、定性和混合研究设计，从多视角研究我国英语语言测试的现状和面临的挑战。作者采用问卷、核心组讨论、一对一访谈等方法收集了大量的研究数据。仅定性数据经转写后就达 10 多万汉字。作者不仅细致地调查了英语考试的开发和实施，也深入地研究了教师和学生这两个群体对我国语言考试现状的看法和态度。研究结果对于我国的语言考试机构、考试从业人员和考试主管部门等都有重要的启示。最后，作者探索性地制订了关于考试开发和实施的

语言测试标准并将其应用到一项校本英语考试中。尽管这些标准还有待补充和完善,但这是语言测试领域标准制订和应用的一次有益尝试。

标准的制订和实施是提升语言测试专业化水平的重要举措。目前,我国已经制订并出版《中国英语语言能力等级量表》。相关的教育考试管理部门正在酝酿制订语言测试行业标准。因此,相信本书的出版将推动学界对语言测试标准问题的深入探讨,进而进一步提升我国语言测试的专业化水平。

金艳
上海交通大学外国语学院教授
全国大学英语四、六级考试委员会主任
2018 年 5 月

# 前　言

我国是考试的故乡，历史上绵延千年的科举考试制度对我国的政治、经济和社会发展发挥了重要的推动作用。但是，现代语言测试理论和实践是在近30年左右才开始迅猛发展起来。我国的语言考试具有规模大、风险高、对教学反拨效应明显等特点，在社会上具有广泛的影响力。因此，这些考试本身的质量如何？考试是否能够准确地反映学生的语言水平？这些问题不仅是教育和考试管理部门关注的重点，也是学生、教师乃至社会共同讨论的热点话题。制订语言测试标准是提升考试质量和实施问责制度的有效途径。

近些年，语言测试的专业化成为国际语言测试研究领域关注的重点和热点之一。语言测试专业化体现在多个方面，最突出的体现是制订和实施语言测试标准。但是，目前关于语言测试标准的实证研究数量稀少。本书源于我主持的国家社科基金项目《标准化语言测试的标准制订和效度研究》。项目组成员包括赵冠芳副教授（澳门大学）、闵尚超副教授（浙江大学）、何静副教授（复旦大学）和王丽副教授（西安外国语大学）。作为语言测试研究人员，我们承担该项目的目的并非要制订语言测试标准，而是通过大量的实证调查研究，了解我国目前英语语言测试的现状，从而为我国相关的教育和考试管理部门在将来制订语言测试标准的时候提供参考。在该项目开展的过程中，项目组成员齐心协力，共同推进项目的进展。我们收集了大量的定量、定性研究数据，深入探析我国语言测试的现状以及面临的挑战。仅定性的访谈数据经过转写后就达10多万汉字！

在本书的写作过程中，上海交通大学的金艳教授和上海外国语大学的邹申教授经常给我鼓励并提供了很多有益的建议。本书的大部分写作工作是我受国家留学基金委资助在墨尔本大学语言测试中心（Language Testing Research Centre，LTRC）访学期间完成的。我想在此感谢墨尔本大学的 Tim McNamara 教授、LTRC 中心主任 Ute Knoch 副教授和 Sally O'Hagan 博士、Kellie Frost 博士和 Annemiek Huisman 等在我访学期间给我提供的各种帮助和支持。同时，我也想借此感谢国际语言测试协会现任主席、墨尔本大学的 Cathie Elder 副教授和墨尔本皇家理工大学的 Amrohit Amputch 对我的帮助和支持。

在本书初稿完成以后，以下语言测试界的同事和朋友帮我阅读书稿并且提出修改意见：邹绍艳博士（青岛农业大学）、关晓仙博士（华东师范大学）、陈颖副教授（中国海洋大学）、王隽（上海交通大学）和张晓艺（上海交通大学）。本书的第八章的数据来自于复旦英语水平考试。我想借此感谢季佩英教授（复旦大学）和范烨教授（复旦大学）长期以来对我工作的支持。在本项目的开展过程中，我的两位硕士研究生郝原悦和雷志娟帮忙收集和整理语言测试标准，在此一并谢过。

国际语言测试研究领域的前辈 Lyle Bachman 教授在世纪之交借用爱因斯坦的名言发表了题为 *Modern language testing at the turn of the century: Assuring that what we count counts* 的重要论文，回顾语言测试领域的发展趋势并展望未来，指出制订和实施语言测试标准是未来语言测试研究的重点之一。制订符合我国国情和考试特点的语言测试标准是进一步提升我国语言考试信度、效度和公平性的重要举措，必将对我国的外语教育、人才培养和社会进步产生重要而深远的影响。

范劲松

2018 年 5 月于上海

# 目　录