汪顺玉 著

语言测试构念效度研究

Validity is an evolving concept. Initially, it refers mainly to predictive validity or concurrent validity. The next stage of development is characterized by so called Holy Trilogy (content validity, criterionrelated validity, construct validity). At present, Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. Put it simply, validity refers the degree that test score interpretation confirm to the construct or test purpose. As a unitary concept embracing multiple facets, validity ...

Construct Validation in Language Testing

$$Rc = \frac{\sum_{1}^{n} \alpha_1 \beta_1}{\sqrt{\sum_{1}^{n} \alpha_1^2 \sum_{1}^{n} \beta_1^2}}$$

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} x_{ijk} + N\sum_{j=1}^{n} x_{ijk} - \sum_{j=1}^{n} x_{ijk}}{2(N-1)n}$$

$$\text{ID P-DII} = \frac{\sum_{k=1}^{k} [Wk(Pfk - \text{Pr}\,k)]}{\sum_{k=1}^{k} Wk}$$

四川大学出版社

汪顺玉 著

语言测试构念效度研究

Validity is an evolving concept. Initially, it refers mainly to predictive validity or concurrent validity. The next stage of development is characterized by so called Holy Trilogy (content validity, criterionrelated validity, construct validity). At present, Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. Put it simply, validity refers the degree that test score interpretation confirm to the con

concept embraced.

Construct Validation in Language Testing

四川大学出版社

责任编辑：黄新路
责任校对：夏　宇
封面设计：米茄设计工作室
责任印制：李　平

图书在版编目(CIP)数据

语言测试构念效度研究 / 汪顺玉著. —成都：四川大学出版社，2009.9
ISBN 978-7-5614-4563-1

Ⅰ.语… Ⅱ.汪… Ⅲ.英语-测试-研究 Ⅳ.H319.3

中国版本图书馆 CIP 数据核字（2009）第 160812 号

书名　语言测试构念效度研究

著　　者　汪顺玉
出　　版　四川大学出版社
地　　址　成都市一环路南一段 24 号 (610065)
发　　行　四川大学出版社
书　　号　ISBN 978-7-5614-4563-1
印　　刷　郫县犀浦印刷厂
成品尺寸　140 mm×202 mm
印　　张　8.75
字　　数　219 千字
版　　次　2009 年 9 月第 1 版
印　　次　2009 年 9 月第 1 次印刷
定　　价　26.00 元

◆读者邮购本书，请与本社发行科联系。电 话：85408408/85401670/85408023　邮政编码：610065
◆本社图书如有印装质量问题，请寄回出版社调换。
◆网址：www.scupress.com.cn

致　谢

在上海外国语大学攻读博士学位是一个艰苦而愉悦的历程，而博士论文的完成是我人生中迄今为止最丰厚的收获。在我的求学和研究道路上，我得到了许多人的指引、支持和帮助。没有他们的关心和帮助，要完成本研究是不可能的。

首先，我要衷心地感谢我的导师、上海外国语大学的邹申教授。是她的学识、教诲和关怀把我引入语言测试和评价的殿堂。在三年的求学过程中，她在学术上的指导使我学业不断进步；她在精神上的鼓励与支持使我在学习和研究上勇往直前；她无偿提供的资料和数据是我能顺利完成学业的保证。对我而言，她是一位值得尊敬的良师，也是一位和蔼可亲的益友。

我非常感谢上海外国语大学研究生部的全体老师，感谢他们给我提供了一个良好的学习和研究环境。感谢上海外国语大学戴炜栋、何兆雄、梅德明、许余龙、王彤福等教授。他们所开设的课程和讲座开阔了我在语言学领域的视野

我也感谢全国英语专业四、八级考试中心和美国“教育考试服务中心”（ETS），他们给我提供了宝贵的研究数据和资料。

在一起求学的博士生同学中，我的学长席仲恩、刘宝全、党争胜、张艳莉、刘琴以及同年级同学唐雄英、纪小凌、陈洁倩对我的研究都提出了许多宝贵意见；我前后一起的两位室友李华东和邵军航经常与我讨论语言学领域的问题，分享学习和研究心得，他们对我的研究启发很大；与周大军合作进行科研的经历使我感受了他治学严谨的风范；陈科芳、黄惠、齐快鸽、杨仙菊、刘春艳、张艳、赵得全、詹全旺、陈广兴等同学对我的学习和生活方面帮助很多，在此一并感谢。

我还要感谢我所在的重庆邮电大学的校长和书记们，以及外语学院张爱琳院长、杨祖鼎（前）院长、温平川（前）书记、吴世银书记和戴晓莉副院长，是他们的坦荡宽广的胸襟和无微不至的关怀，为我学业的顺利完成和博士论文的撰写提供了保证。

四川外国语学院廖七一教授、重庆师范大学吴念教授、重庆邮电大学张爱琳教授评审了本博士论文，并对论文提出了诸多建设性意见，我深表感谢。

我也要感谢我妻子石竹屏，是她的爱和牺牲为我的求学提供了一个安全而舒适的港湾，她对我求学路上始终如一的鼓励是我不断追求进步的永恒动力。我也感谢我的儿子汪宏见，在孩子的成长过程中即孩子最需要父爱的时候，我有6年多的时间在求学，很少有时间和他在一起，对此我深感内疚。孩子对我回家的期待以及对我研究内容的天真的提问，使我的研究减少了些苦恼、增加了些乐趣。

在博士论文出版之际，我还要对关心和支持本书出版的机构和人士表示真诚的谢意：

我的导师邹申教授在百忙中为本书稿做序；上海外语音像出版社社长、上海外国语大学陈坚林教授对本书进行了介绍，

西南大学李力教授聘我为西南大学外国语学院硕士生导师，并为研究生开设语言测试课程，使我有机会在更多的人群中和更高的平台上讲授语言测试理论和方法。不少研究生下载了本博士论文，并与我讨论其中的诸多理论和应用问题，也提出了不少有启发意义的意见。

重庆邮电大学社会科学处为本书提供了出版基金。重庆邮电大学席仲恩博士在本博士论文出版前通读了书稿，就格式、观点和措辞等方面提出了宝贵的修改意见。陈昌川老师为书稿的文字以及图表编辑付出了大量心血。

四川大学出版社的黄新路先生以及其他编辑同志为本书的顺利出版做了大量工作，他们不仅在本书的规范和编辑上倾注了不少心血，而且对图书的版式和封面进行了精心的设计。对于他们的工作和贡献，我表示衷心的感谢。

由于笔者学识有限，谬误或疏漏之处在所难免，谨请识者指正。

汪顺玉

2009年5月于重庆邮电大学

序

汪顺玉的专著《语言测试构念效度研究》是在他的博士论文的基础上撰写而成的。该书正式出版了，作为导师，我颇感欣慰。

早在上世纪中期，语言测试界已开始认识到效度是考试评价中的最重要因素。20世纪八九十年代起人们对效度的认识有了新飞跃，效度的概念从多类效度发展到统一构念效度。效度的各个方面不再是互不关联的环节；构念效度上升到统领各个效度的地位；效度验证不仅包括对分数的解释，还包含对分数的使用及结果。这一系列认识上的变化都与Messick的经典论述有直接关联。与效度概念发展相联系，效度验证的范式和方法也产生了变化。因此，对于语言测试研究者而言，如何在实践中全面准确地诠释效度、探索和掌握效度验证的科学方法，是一个既有理论意义又有应用价值的课题。本书作者汪顺玉在这个领域作了有意义的探索。

作者从本体论、认识论、方法论的视角对统一效度概念进行了全面科学评述，在此基础上提出了一个英语专业八级考试效度验证的理论和方法框架。然后在该理论和方法框架下，对八级考试的客观试题从实证角度进行效度验证。效度验证旨在解答四个研究问题：（1）八级考试客观试题的测量学属性如何？（2）客观题目实际测量的维度与考试设计的理念是否一致？（3）客观试题分数的意义在不同的群体中是否具有类似的解释？（4）新增加的人文知识分测验是否存在考试偏差？

特别值得一提的是，作者在效度验证过程中采用不同的研究方法，为后来者提供了可借鉴的范式。作者对构念效度的理论和方法讨论采用文献综述方式，对八级考试客观题目的效度验证则运用定量方式，如：相关方法分析题目同质性、题目区分度、聚合和区别效度、构念一致性检验；因子和谐系数用于检验不同群体因子负荷之间的相似程度；单因素方差分析进行跨群体均值比较；因子分析用于因子维度探索和验证；标准难度

方法进行项目差异功能分析等。

本书的特点是理论阐述详尽，实例描述具体，具有很强的可操作性和借鉴价值，是理论与实践相结合的典范。我深信本书定会使语言教师和研究生读者受益匪浅。

是为序。

邹申

2009 年 6 月 23 日

于上海外国语大学

摘 要

2005年的英语专业八级考试是根据2004年新的八级考试大纲设计和施测的第一次考试，考试的性质、构念领域、任务要求、分数权重等都发生了较大变化。作为全国唯一的测量英语专业学生高年级英语水平的大规模考试，这些变化对个人、团体和社会将产生重大影响。测试界认为，越是高风险考试，越要对考试的技术和应用方面进行评价，对考试的效度验证要求越高。因此，运用先进的效度理念、分析技术和行业规范对我国的八级考试进行研究，不仅有理论价值，也具有现实意义。在对测验的评价中，效度是最重要的考虑因素。然而，在过去近一个世纪以来，效度的概念从多类效度发展到统一构念效度。与效度概念发展相联系，效度验证的范式和方法也产生了变化。因此，全面和准确地理解效度，具有十分重要的理论意义。而掌握效度验证的科学方法具有运用价值。

本文研究的目的有二：一是对统一效度概念从本体论、认识论、方法论视角进行较全面的评述，旨在为八级考试效度验证提供一个理论和方法框架；二是在统一效度概念下，对八级考试的客观试题从实证的角度进行效度验证。验证的问题包括四个：(1)八级考试客观试题的测量学属性如何？(2)客观题目实际测量的维度与考试设计的理念是否一致？(3)客观试题分数的意义在不同的群体中是否具有类似的解释？(4)新增加的人文知识分测验是否存在考试偏差？

针对两个目的，采用两种研究方法。对构念效度的理论和方法讨论采用文献综述方式，针对八级考试客观题目的效度验证是用定量的方式提供解释依据。

效度是一个不断演进的术语。最初的概念是指效标关联效度（预测或共时效度），发展到由内容效度、效标关联效度和构念效度构成的神圣三部曲阶段，现在的效度是一个包括多个方

面的统一概念，指的是测量分数解释理论和测量目的的程度。统一效度概念意味着只有一种效度，即构念效度，不再存在传统意义上的多种效度。蕴涵在构念效度的分数意义存在于所有基于分数的推断之中。效度的关键是分数的可解释性、关联性、实用性、作为行动基础的分数价值、分数使用所导致的社会后果的功能价值（对于效度的威胁主要存在于构念代表不良和构念无关因素两种形式中）。前者是指测量不能代表拟测构念所包含的某些种类的内容、所负载的某些心理过程，或者是排除了应有的某些方式的反应，因而使测验的意义被狭隘和局限化。后者是指测验包含与拟测构念无关的额外附加并且稳定的变异。它们都会导致对构念的不准确测量。

在理念上强调统一性和在技术上区分不同的方面是并行不悖的。效度可以从内容、心理实质、分数结构、分数概化、外在方面以及社会后果六大方面进行研究。内容方面包括内容关联性、内容代表性和技术质量；心理实质方面，指的是测验任务和被试反应所代表的理论原理和心理过程；结构方面研究通过测验题目体现的观察变量与测验分数的关系；概化方面试图回答的问题是，在什么程度上，测验分数的属性和解释适合于不同的群体、不同的情景和不同的任务；外在方面，指的是测验与其他测验或非测验行为之间的关系所反映出来的期望的相关关系，这种关系蕴涵在所测构念的理论中；社会后果研究用来评价分数解释以及长期及短期考试使用所导致的本意的和非本意的社会后果。

效度验证为测验分数所应有的解释和测验成绩及所建议的用途相关性提供佐证。效度验证在逻辑上应该以明确陈述测验分数解释方式以及测验使用的原理为开端。效度验证是一个无止境的过程，而不是一个程序。它需要多方面的依据：定量的、定性的、证实的、证伪的。这些依据常常交织在一起。效度验

证与具体目的相关联，考试开发者和使用有共同的责任从不同的方面提供效度依据。验证工作贯穿于测验开发、施测和评价的始终。

构念一致性问题是本研究中关于分数解释的切入点，解决的是测验所测量的构念对不同背景的考生的意义是否具有可比性的问题。当一个测验在一个组别（群体）中所测量的假设特质（或者心理构念）与另一组别相同，或者当一个测验在测量相同的特质时测量的准确程度相似时，构念具有可比性。构念一致属于分数结构和效度概化问题。它与考试偏差分析一起，为分数的可解释性和考试公正提供依据。

本研究的对象是 446 所大学参加 2005 年 TEM 8 考试的 96696 名考生。分析的数据是由上海外国语大学四、八级考试中心提供的全体考生在客观题目上的原始反应数据以及已经对反应进行判断过的数据。针对不同的研究问题，使用了不同的统计分析手段：描述统计用于基本数据探索；相关方法分析题目同质性、题目区分度、聚合和区别效度、构念一致性检验；因子和谐系数用于检验不同群体因子负荷之间的相似程度；单因素方差分析进行跨群体均值比较；因子分析用于因子维度探索和验证；标准难度方法进行项目差异功能分析。

分析结果发现：测量属性方面，对题目的测量学属性而言，题目的难度分布较宽，平均难度也较适中（0.63），难度值在不同群体的分布相似。题目的区分度整体上偏低，在 0.11–0.42 之间，平均区分度为 0.29。区分度在各个群体中的分布较相似；人文知识分测验题目的区分度整体上在技术规定的范围内，但是有两个题目在难度和区分度上都属于极值。各分测验内题目之间的相关系数也偏低，各个分测验的内部信度也偏低。测量的维度方面，听力分测验有两个维度，分别代表面试听力和新闻听力；人文知识分测验有两个维度，分别代表精神类知识和

历史地理知识；阅读分测验也是两个维度，所代表的意义有待进一步明确；全部客观题目有6个维度，其构成与分测验维度相同。对这些维度的上一级因子进行探索，发现3个因子比较清晰地代表了听力、阅读和人文知识三方面的能力和知识。这些发现，总体上说明，除了人文知识分测验的两个维度与考试设计的3个维度有出入外，考试所测量的东西与拟测量的构念是相符的。构念一致性方面，听力分测验和人文知识分测验总体上一致的。阅读理解分测验维度和全部客观题目在不同的群体中不一致。对全部题目的跨群体维度探索结果表明，主轴法比主成分方法能更加有解释力，而因子和谐系数往往比相关系数更倾向于得出因子相似的结论。在6因子方案中，因子1、因子5、因子6在不同群体之间相似程度低，因子2、因子3、因子4在不同群体之间相似程度高。在3因子方案中，因子1在不同群体中相似程度高，因子2和因子3在除外语院校外的群体中相似程度也高。人文知识分测验在外语和非外语院校之间以及外语本科生与其他专业类别学生之间没有明显的考试偏差存在。

本论文由五章构成。

第一章是引言，主要对所研究问题的社会背景和行业背景进行介绍。这部分介绍的基本逻辑是这样的：考试是教育改革的突破口，对社会和个人产生重大影响，需要研究考试，提高考试的质量，保证考试分数的合理使用；效度是考试评估的关键，新的构念效度理论与方法是效度评估的依据和途径；作为变化中的TEM 8需要及时地运用这些依据和方法进行构念效度验证。本章还提出了本研究要解决的几个问题。

第二章是文献综述部分，目的是为本研究提供理论基础和方法依据。本章对效度、构念、构念效度、考试偏差、项目功能差异、构念一致等概念进行讨论；对构念效度有关理论进行

较为广泛的评述，重点探讨构念效度的发展脉迹、效度欠缺的理据、构念效度验证的方面、构念效度验证的程序、跨群体构念效度的比较原理和建立构念效度的统计方法。

第三章介绍了研究的设计和采用的分析方法介绍，具体包括研究的具体问题、样本、研究的工具和统计方法。后者包括三个方面的统计：一是试题所测构念维度的探索方法；二是跨群体构念效度比较的方法；三是项目功能差异检验的方法。

第四章报告了本研究的核心发现。它们包括对研究具体几个问题的结果：考试的构念维度数、试题难度的跨群体比较、试题题目区分度的跨群体比较、分测试信度的跨群体比较、分测试间相关系数的跨群体比较、因子结构的跨群体比较、人文知识题的项目功能差异检验。

第五章是讨论和结论部分。总结了本研究所得到的发现及获得的结论，对使用的研究方法和过程进行了评述，在针对 TEM 8（2005）的研究结果的基础上，对 TEM 8 的设计者提出了建议；另外，本部分还指出了本研究对语言评价的贡献和不足之处，提出了此研究框架下今后研究的方向和思路。

关键词： TEM 8；构念效度；效度；效度验证；构念一致；项目；差异功能

Abstract

The Test for English Majors Band 8 (shortened as TEM 8) is a large scale nationwide criterion-referenced test for English majors in mainland China. Since its inception, Tem 8 underwent several modifications. The latest modification answering for the new requirements of The 2000 Curriculum for English teaching was made in the TEM 8 Syllabus in 2004. The 2005 version of TEM 8 is the first of its kind designed in line with the new TEM 8 test syllabus. This version substantiated the all major modifications in construct domain, testing time, test content, test formats, relative weight of subtest scores for listening, reading and writing. In short, a new subtest was added and testing time shorted and cognitive load increased. The newly added subtest is the general knowledge, which covers the fields of general linguistics, English and American literature and knowledge of major English speaking countries. Whether such modifications will result in psychometrically acceptable features and whether the test does test what it purports to measure remains unanswered. Moreover, it is generally acknowledged that the more a test is of high stakes in nature, the more demanding the validation should be for the test. TEM 8 is by nature a high stakes test. Therefore, a timely validation endeavor on TEM 8 in line with the state-of-art theoretical validity framework is of significance both theoretically and practically.

This research embodies two main aims: one is to review and evaluate the literature concerning unitary validity in the

perspectives of ontology, epistemology and methodology in the hope of providing relevant rationale for the validation argument. The second aim is, in line with unitary validity, to provide empirical (statistical) evidence to the score interpretation to the objectively scored items in TEM 8 (2005 version). Due to the many unclaimed testing error variance in subjective items, the study analyzed only the objective subtest data, namely, listening component in MC, reading comprehension in MC, and general knowledge in MC, taking up 40% of the total score. There are four questions in this regard. What are the psychometric properties of the objective items like? Are the tested traits (or dimensions) consistent with what is expected in the test specifications? Do the scores of objectively scored items mean the same across different population groups? Does the subtest of general knowledge demonstrate substantial test bias?

For the above two aims, two research methods are used. Literature review is conducted in reviewing and evaluating the theoretical issues of validity and validation. Statistical analysis is employed for interpreting the dimensions and meaning of scores for TEM 8 objective items.

Validity is an evolving concept. Initially, it refers mainly to predictive validity or concurrent validity. The next stage of development is characterized by so called Holy Trilogy (content validity, criterion–related validity, construct validity). At present, Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. Put

it simply, validity refers the degree that test score interpretation confirm to the construct or test purpose. As a unitary concept embracing multiple facets, validity means a comprehensive concept–construct validity, in stead of many types of validity in traditional sense. The meaning of scores for construct validity resides in score–based inference. The key issues of test validity are interpretability, relevance, utility of score, value implications of scores as a basis of action, and the functional worth of score in terms of social consequences resulted of their use. There are two kinds of threats to validity: construct underpresentation and irrelevant test variance. The former indicates that the test is too narrow and fails to include important dimensions or facets of the constructs. The later means that the test contains excess reliable variance that is irrelevant to the interpreted construct. There are two types of irrelevant test variance in language achievement testing: construct–irrelevant difficulty and construct–irrelevant easiness. Generally, construct–irrelevant difficulty leads to construct scores that are invalidly low for those individuals adversely affected.

The nature of unitarity by no means implies that validation can not be conducted from various facets operationally. Messick distinguishes 6 kinds of components for validity. They are content component, structural component, substantive component, external component, generalizability component and social consequence component. In effect, these six aspects function as general validity criterion or standards for all educational and psychological measurement, including performance measurement. The content aspect of construct

validity includes evidence of content relevance, representativeness and technical quality. The substantive aspect refers to theoretical rationale for the observed consistence in test responses along with empirical evidence that the theoretical processes are actually engaged by respondents in assessment tasks. The structure aspect appraises the fidelity of the scoring structure to the structure of the construct domain. The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings and tasks. The external aspect includes convergent and discriminant evidence from multitrait multimethods comparison as well as criterion relevance and applied utility. The consequential aspect appraises the value implications of score interpretation as the basis for action as well as the actual and potential consequences of test use, especially in regard of sources of invalidity related to issues of bias, fairness and distributive justice.

Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed interpretation. Validation logically begins with an explicit statement of the proposed interpretation of scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation refers to the construct or concepts the test is intended to measure. Validation in nature is an unending process rather than a procedure. Therefore, evidence from various sources is imperative: qualitative and quantitative

evidence, evidence for proving and refuting. Different sources of evidence are often interwoven. Moreover, validation is use–specific. When test score are used or interpreted in more than one way, each intended use or interpretation must be validated. Test developer and test user share joint responsibility in validation. The test developer is responsible for furnishing relevant evidence and rationale in support of the intended test use. The test user is ultimately responsible for evaluating the evidence in particular setting in which the test is to be used. Validation begins at the test constructing, not at the last stage of test development.

Construct invariance is the focus of this research. It involves the comparability of score interpretation across population groups. Operationally, construct across population groups is comparable when the test demonstrates the same constructs or hypothetical traits or dimensions across population groups, or when the construct is measured to similar extent of accuracy across population groups. Construct invariance, a type of evidence for score generalizabilty, function as evidence, along with test bias detection, for test fairness.

The study utilized a unitary approach to evaluate the construct validity of TEM 8 （2005version）. The sample consisted of 96696 examinees of five populations taking TEM8 in 2005. Of all examinees, 14977 are from comprehensive universities, 25728 from polytechnic universities, 35965 from normal universities, 11510 from universities of international studies and 8516 from else universities.

The study into TEM 8 is confined exclusively to empirical