语言测试效度与公平性研究*

姜秀娟
(曲阜师范大学,曲阜 276826; 北京外国语大学,北京 100089)

提　要: 效度是评判一项测试质量的重要指标,而公平性又是效度的重要保证。本文结合测试效度观及其验证模式的发展变化,对近 50 年来语言测试公平性观念及其研究模式在分类、整体、论证 3 种效度观时期的演变进行梳理与思考,发现语言测试公平性研究采取的几乎是与效度研究一样的进路,学界对公平性研究的必要性存在争议。在以上分析的基础上,本文总结二者之间的关系,并指出未来测试公平性研究应继续努力的方向。

关键词: 语言测试; 效度; 公平性
中图分类号: H319　　文献标识码: A　　文章编号: 1000 - 0100(2018)01 - 0097 - 6
DOI 编码: 10.16263/j.cnki.23 - 1071/h.2018.01.015

Validity and Fairness in Language Assessment

Jiang Xiu-juan
(Qufu Normal University, Qufu 276826, China; Beijing Foreign Studies University, Beijing 100089, China)

Validity has been regarded as the key element of any assessment and fairness plays a very important role in achieving high validity. Based on the conceptual changes of validity and changes in the modes of validation, this paper deals with the evolution of fairness studies in language assessment over the past 50 years from the following three perspectives: categorized, unitary and argumentative validity concepts. It reveals that the methods and contents in the study of fairness are nearly the same as those in the study of validity and validation and the assessment experts have different views towards the necessity of the fairness exploration in language assessment. According to the analysis above, the paper summarizes the relationship between the two and points out what the further studies should explore thoroughly in the future.

Key words: language assessment; validity; fairness

1　引言

测试公平性研究始于 20 世纪 60 年代,80 至 90 年代被广泛关注(Zieky 2006: 360),是测试领域一个新兴的热点话题。长期以来,效度是评判一项测试质量的重要指标,而公平性又是效度的重要保证,二者交织在一起,不可分割(同上: 359)。80 年代中期以来,有关测试公平的观点、标准、文件不断涌现,专门探讨测试公平性问题的高层次学术会议也相继召开,测试公平性的重要性可见一斑。那么,公平性到底是什么,如何研究或检验一项测试的公平性? 语言测试效度观及其验证模式的变化对公平性观念及其研究模式产生怎样的影响? 语言测试公平性与效度有怎样的关系? 为了回答以上问题,本文结合测试效度观及其验证模式的发展变化,对语言测试公平性观念及其研究模式在分类、整体、论证 3 种效度观时期的演变进行梳理与思考,并指出语言测试公平性研究的未来趋势。

2　效度分类观与语言测试公平性研究模式

20 世纪 50 年代之前,教育与心理测量学普遍坚持“相关即有效”的效度观(韩宝成 罗凯洲 2013: 412)。但是,要想确定那个“相关”的东西绝非易事,因为一项测试可以与很多种事物相关。

* 本文系山东省社科规划基金项目“高考英语、托福和雅思听力测试的测量目标及任务设置比较研究”(16CZWJ29)的阶段性成果。

于是，不同类型的效度应运而生。1954 年，美国心理学会（APA）在《关于心理测验和诊断的技术建议》（*Technical Recommendations for Psychological Tests and Diagnostic Techniques*）中，将效度分为 4 种：预测效度（predictive validity）、共时效度（concurrent validity）、内容效度（content validity）和构念效度（construct validity）。1966 年，《教育与心理测验的标准与指南》（*Standards for Educational and Psychological Tests and Manuals*）（AERA et al.）把预测和共时合并为校标关联效度（criterion-related validity）。

1961 年，Lado 在现代语言测试的奠基之作《语言测试》（*Language Testing*）中首次将教育与心理测量学领域的效度概念引入语言测试领域，指出“效度本质上是一种关联。一项测试是否测量到它要测量的东西。如果答案是肯定的，那么它就是有效的”。之后，语言测试领域纷纷效仿 Lado 的观点定义效度（如 Valette 1967；Harris 1969；Heaton 1975；Finocchiaro，Sako 1983）。Heaton（1975：153）将语言测试效度分为表面效度、内容效度、构念效度和实证效度。这一时期的语言测试效度验证模式主要采取 Lado 提出的方法，如选择、设计与内容相关、与学习问题相关的题目；修改因非语言因素引起难度增加的试题；使用一项有效的测试和自己开发的测试，对一组有代表性的学生样本进行测试，计算两次测试成绩间的相关系数，从而确定测试效度（Lado 1961：328－329）。分析测试内容、计算校标关联系数是这一时期进行语言测试效度研究的主要方法（韩宝成 罗凯洲 2013：413）。

那么，如何分析测试内容，如何保证测试题目与测试构念相关，如何确定测试题目中没有包含与测试构念无关（construct-irrelevant）的因素？这些问题是该时期语言测试效度验证过程中必须解决的，对这些问题的回答也使测试专家学者开始关注测试公平性问题。早期的语言测试文献只是将测试公平性等同于测试中的题目对不同的考生群体不存在偏颇（bias）（AERA et al. 1985）。测试偏颇（test bias）指具有相同能力的不同群体的考生在相同题目上的得分不同。换句话说，测试偏颇就是与测试构念无关的考生特征（如性别、种族、社会经济地位等）对考生的考试成绩产生系统性的影响（McNamara，Roever 2006：82）。测试偏颇一般采用项目功能差异（Differential Item Functioning，DIF）研究。如果研究显示测试题目存在 DIF，就要确定 DIF 存在的原因是否与测试构念无关因素有关，如果有关，则说明试题存在偏颇，从而影响测试的公平性，必须去除或修改导致偏颇的题目。美国教育测试服务中心（Educational Testing Service，ETS）1986 年规定，在测试开发的过程中，为保证测试较高的效度和公平性，除了对编制的题目进行常规的项目分析外，还必须进行项目功能差异研究。受这一时期效度验证模式的影响，偏颇研究只是从技术的角度，对试题的心理测量学属性进行统计分析，控制与测试构念无关的因素，从而为效度验证提供数据和技术支持。20 世纪 80 年代末，随着效度分类观向效度整体观的转变，测试领域对公平性的认识也发生变化，公平性研究模式也随之发生改变。

3　效度整体观与语言测试公平性研究模式

20 世纪 80 年代，随着效度研究的不断深入，教育测量界发现基于分类方法进行测试的效度验证所得结果太零散，也没有考虑考试成绩的价值含义及考试成绩使用的社会后果。基于此，Messick（1988，1989）提出整体效度概念（unitary concept of validity），认为效度只有一个，即构念效度，而证明效度的证据可来自多方面，并用分层效度框架（又称效度渐进矩阵（progressive matrix））进行说明（参见表$_1$）。

表$_1$　分层效度框架

	测试解释	测试使用
证据基础	构念效度	构念效度＋相关性/实用性
后果基础	价值含义	社会后果

（改自 Messick 1989：20）

分层效度框架由测试解释、测试使用、证据基础和后果基础 4 个维度构成。Messick 的“一元多维”效度整体观更新人们的测试效度验证观念，自此，效度验证不仅仅是对测试本身及分数的评价，还包括对测试结果解释和使用的评价。但是，Messick 的“一元多维”效度理论太抽象，不能有效地指导测试效度验证。为解决操作性问题，Bachman 和 Palmer（1996）提出测试的有用性框架（test usefulness framework），通俗易懂地诠释 Messick 的效度理论。测试有用性框架包括信度（reliability）、构念效度、真实性（authenticity）、交互性（interactiveness）、影响力（impact）和可行性（practicality）6 个要素。信度指一项考试结果的稳定性；构念效度指对考试分数解释在多大程度上是有意义的、适切的；真实性指考试任务特征与目标语言使用任务特征的一致性程度；交互性指考生

完成测试任务时，参与其中的个人特质类型和程度；影响力指考试对个人、教育制度以及整个社会产生的影响；可行性指设计、开发和使用一项测试所需资源与可用资源间的关系。随后的十几年中，该框架是语言测试效度验证的权威模式（Weigle 2002），在指导语言测试的开发和使用方面发挥重要作用。

测试效度观念及其验证模式的改变，使人们意识到偏颇研究只是属于 Messick（1989）分层效度框架中的证据基础维度，公平性应该包括更广阔的研究内容，比如测试的社会价值与影响。而且，1999 年版的《教育与心理测量标准》（以下简称《标准》）专设一个部分讨论测试公平性，将公平性定义为无偏颇、考试过程公平、基于考试结果的决策公平以及学习机会均等。具体来讲，无偏颇就是控制构念代表性不足（construct under-representation）及与构念无关的因素（construct-irrelevant variance），消除影响构念效度的偏颇。比如，要保证内容样本的覆盖面、所有考生都熟悉答题形式等。考试过程公平指在施考过程中平等对待所有考生，考生要有相同的机会展示自己的能力。基于考试结果的决策公平指不同考生群体的考试结果具有可比性，能力相同的考生应享有同等的选拔机会。学习机会均等主要指在标准参照考试中，考生有相同的机会学习考试内容和接触复习资料，尤其是考试成绩用于决定是否留级或颁发证书时，学习机会均等更显重要。因此，测试专家学者开始构建更为全面的公平性研究框架。

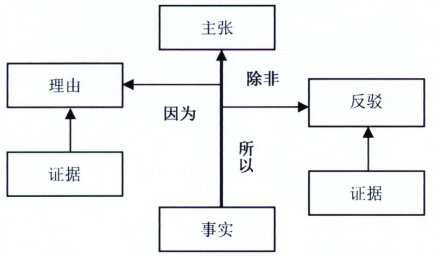
2000 年，Kunnan 在 Messick 整体效度观的指导下，以社会正义理论（Jensen 1980）和《教育公平测试行为准则》（JCTP 1988）为基础，参考 1999 年版的《标准》中关于测试使用、考生权利和责任、考生语言多样化以及残疾考生等涉及公平性话题的论述，进一步扩展传统的测试公平性研究范围，提出新的公平性研究框架。该框架包括效度、机会均等和公正性 3 个组成部分。其中，效度关注构念效度、考试内容与形式的偏颇、试题的差异效应、考试材料中语言使用的恰当性以及哪些考生群体处于不利地位；机会均等关注考试费用、考场选址、考试设备和条件是否有利于所有考生，考生受教育机会是否均等则关注对残疾考生是否有特殊待遇；公正性关注社会公正及法律挑战。可以看出，Kunnan 的测试公平性研究框架不再局限于心理测量学属性，已经扩展至社会、道德、法律和哲学层面（Kunnan 2000：5）。2004 年，Kunnan 对其 2000 年的公平性研究框架进行修改和完善，增加施考条件和社会后果两个部分。至此，测试公平性研究框架更加全面、更加深入，由原来的 3 个组成部分扩展到 5 个，形成由效度、机会均等、公正性、施考条件和社会后果构成的新框架，完全契合整体效度观的精神及其效验模式。该框架成为近年来语言测试公平性研究的主要依据。2009 年，Kunnan 又提出测试环境框架（the Test Context Framework），该框架试图从政治、教育、文化、社会、经济、法律和历史等诸多方面审视一项测试，同年，Kunnan 用美国公民入籍考试（the Naturalization Test）为例从 3 个方面对测试的公平性进行探讨：（1）测试的要求和目的：该考试的要求和目的是否有意义；（2）测试的理论基础、内容和操作：该考试是否能够测出英语语言能力以及关于美国历史与政府的知识；（3）测试后果：该考试是否能够带来民族主义或社会融合。通过分析以上 3 个方面，Kunnan 发现，此项美国公民入籍考试是 20 世纪 50 年代美国特定历史时期的产物，已经不符合时代要求，也不符合美国法律规定，因此，该考试的实施和分数的使用无意义。另外，该考试也测不出考生是否具有“民族主义”或“社会融合”能力，也就是说，该考试的内容和理论基础与预测构念不相关。可见，该考试对考生而言不公平。

但是，随着测试效度及其验证模式研究的深入，人们发现 Bachman 和 Palmer（1996）测试有用性框架的 6 大要素间缺少关联，效度验证只是证据的简单罗列，而且无从知晓证据收集从哪儿开始，到哪儿结束。对测试有用性框架“重操作性、轻连贯性”缺陷的认识，也使人们意识到 Kunnan（2004）测试公平性框架存在同样问题，该框架的 5 个组成部分没有形成一个连贯的令人信服的测试公平性论证（Bachman 2005）。Kunnan（2009）框架也没有解决这一问题，无法为测试公平性的评估和实证研究提供切实有效的指导（Xi 2010）。如何明确语言测试公平性各要素间的关系；如何整合各类证据，使它们成为一个连贯的相互联系的整体？人们期待新观点新模式的出现。

4　效度论证观与语言测试公平性研究模式

1999 年版的《标准》把效度定义为“证据及理论对测试分数解释与使用的支持程度”，指出效度验证就是对“分数的预期解释与使用的论证”（AERA et al. 1999：9）。但是，在效度验证中如何组织证据，该版《标准》没有给出一个可供参考的论证模式，效度验证基本上采取证据罗列模式。

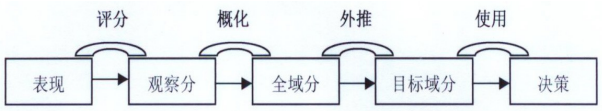
当然，教育测量界并没有停止探索效度验证中的证据组织方法（如 Kane 1992，2002，2004，2006；Kane et al. 1999；Mislevy et al. 2002，2003），最终将 Toulmin（2003）的实用推理模型（practical reasoning model）（参见图$_1$）用于效度验证，提出基于论证的验证模式（argument-based approach to validation）。该模式明确收集证据的类别与数量，效度证据的组织也不再是简单的罗列，而是形成一个环环相扣的证据链，使效度验证成为一个有始有终、逻辑严密的论证过程。



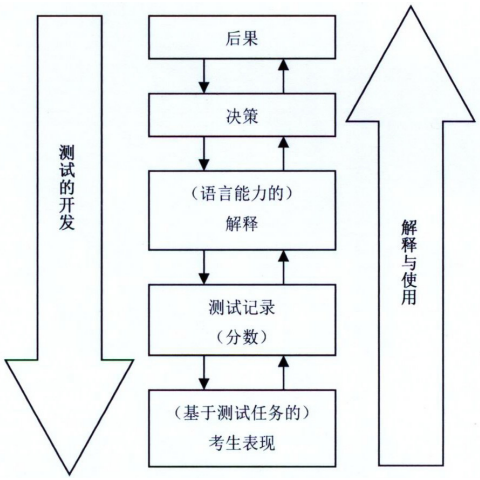
图$_1$　Toulmin 的实用推理模型
（改自 Toulmin 2003：97）

典型的基于论证的效度验证模式有两个，一个是 Kane（2006）的解释性论证（interpretive argument）与效度论证（validity argument）。该模式分两步：第一步，搭建理论框架（解释性论证）（参见图$_2$）；第二步，检验理论框架（效度论证）。另一个是 Bachman 和 Palmer（2010）的测试使用论证（Assessment Use Argument，简称 AUA）（参见图$_3$）。

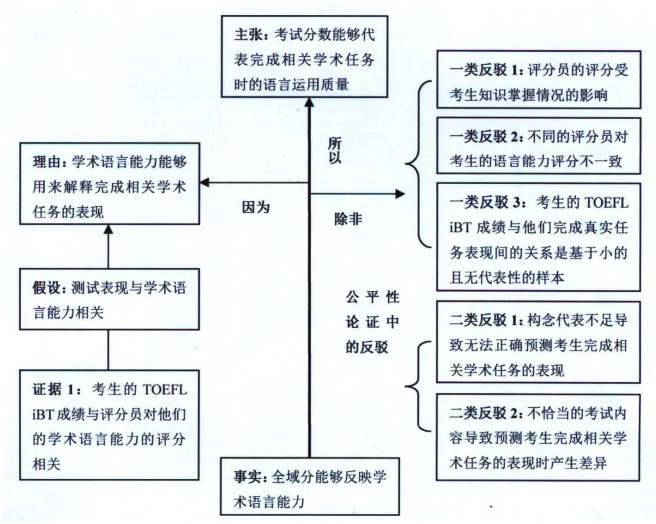
近年来，随着测试效度论证观的出现及其验证模式转变，测试学界也纷纷从论证的角度对语言测试公平性进行研究，提出基于论证的公平性研究模式，如 Xi（2010）的公平性论证框架（Fairness Argument Framework）。Xi 认为，测试公平性指测试所有环节对所有的相关考生群体具有相同的有效性，即对于所有相关考生群体而言，与构念无关因素、构念代表性不足、不一致的施测行为以及不恰当的决策程序或测试结果的使用，对考试分数及其解释以及基于分数所作的决定与后果不会产生系统性的影响（Xi 2010：154）。基于该定义，Xi 提出研究公平性的框架——公平性论证框架，该框架内嵌于效度论证框架，称作“效度论证中的公平性论证”，并用 TOEFL iBT 为例进行说明（同上：155）。Xi 的效度论证包含 6 个分论证（sub-argument）：（1）证据表明目标语言使用域能够提供对考生测试表现进行观察的有意义的基础；（2）证据表明观察分是考生目标语言使用的反映，而不是构念无关因素的反映；（3）证据表明观察分具有概推性，即考生在类似的其他考试中得分相同；（4）证据表明观察分的概推性是有理论基础的，即是基于构念的推论；（5）证据表明构念能够解释非测试环境下的目标语言使用；（6）证据表明基于考试结果对考生语言能力水平的判断具有相关性，对决策具有有用性与充足性（同上：156－157）。可见，Xi（2010）的效度论证框架经过目标域的界定（Domain definition）、评价（Evaluation）、概化（Generalization）、解释（Explanation）、外推（Extrapolation）与使用（Utilization）6 次推论，从考生的测试表现到基于测试结果对考生语言能力的判断与使用形成一个严密而连贯的推论链，从而明确证据收集的起点、终点、数量与种类，在此过程中也完成测试的公平性论证，每次效度论证和公平性论证都采用 Toulmin（2003）的实用推理模型，由事实、主张、理由、证据、假设以及反驳构成。其中，反驳有两类，一类是对所有考生来说，由于缺乏相应的反面证据（counter-evidence）而使结论的说服力减弱；另一类是指对特定考生群体而言，结论是无效的或是站不住脚的（Xi 2010：158－164）。Xi 就效度论证中外推环节的公平性论证以 TOEFL iBT 为例进行说明（参见图$_4$）（Xi 2010：165）。



图$_2$　解释性论证的推理链
（改自 Kane 2006，Bachman 2005）



图$_3$　AUA 框架（Bachman，Palmer 2010：91）



图$_4$　效度论证外推环节中的公平性论证举例(改自 Xi 2010: 165)

5　语言测试效度与公平性的关系

通过以上分析可以看出，语言测试公平性及其研究模式随着语言测试效度及其验证模式的变化而变化，二者之间的关系较复杂，学界存在 3 种观点：二者是并列的、效度包含在公平性之中以及公平性包含在效度之中。

语言测试效度与公平性是并列的，即二者分别是一个独立的概念。首先，1999 版的《标准》对二者分别给出定义(见前文)。从两个定义来看，二者没有直接联系且各有侧重：前者偏重检验分数解释和使用是否有意义，后者着重衡量考生在考试的设计、开发和使用过程中是否享受平等待遇。再者，《教育公平测试实践规范》(*Code of Fair Testing Practices in Education* 2004)也明确规定测试开发者与使用者对整个测试过程进行独立的公平性研究，具体包括试卷的编制与题目的选择、考试的实施与评分、分数的报道与解释以及考试信息的反馈 4 个环节。

效度包含在公平性之中，即效度被看成是公平性的一部分。比如 Kunnan(2000)的公平性研究框架包括效度、机会均等和公正性 3 个组成部分，很明显，效度是衡量公平性的重要指标。Kunnan(2004)公平性研究框架由 3 个组成部分扩展到 5 个后，效度依然被认为是公平性的一部分。

公平性包含在效度之中，即公平性是测试效度的重要方面，甚至把公平性称作可比性效度(comparable validity)(Willingham，Cole 1997: 6－7)，是效度的一个种类。可比性效度指在一项公平的测试中，测量误差与基于测试结果对考生能力的推论对所有考生来说具有可比性。可比性效度贯穿测试的整个过程，涉及考试内容的选取、施考困难的避免、相同的评分过程等方面，无非是尽量避免与构念无关因素的影响与构念代表性不足，这两者也是效度研究的重要方面。

简单来讲，语言测试效度与公平性的关系问题其实就是如何看待二者重要性的问题。如果研究者把效度和公平性看成是测试同等重要的两个方面，就会把二者当做两个并列的独立的概念进行研究；如果认为效度更重要些，就会把公平性看成是效度的一部分；反之，亦然。

6　结束语

效度是评价一项测试质量的重要指标，一直是测试界的研究主题。近些年来，随着测试领域由重视技术向重视测试结果的使用及决策的社会影响的转变，公平性研究也成为测试界热议的话题。但是，学界在某些问题上还没有达成共识，比如，什么是公平性，如何处理效度与公平性之间的关系，公平性研究是否有必要，对最后一个问题的争论尤为激烈。2010 年，Davies 曾撰文回应“How do we go about investigating test fairness”(Xi 2010)一文，认为没有必要进行测试公平性研究，因为公平性研究与效度研究如出一辙，而且根本不可能有测试公平，测试公平只是一种幻想(Davies 2010: 173－175)。因此，今后的研究应多关注此类问题，深入探究测试公平性的性质、研究内容与方法，设计出令人信服的研究框架，从而摆脱与效度研究如出一辙的套路。

参考文献

韩宝成　罗凯洲. 语言测试效度及其验证模式的嬗变[J]. 外语教学与研究，2013(5).

AERA，APA，NCME. *Standards for Educational and Psychological Testing* [Z]. Washington: American Psychological Association，1985.

AERA，APA，NCME. *Standards for Educational and Psychological Testing* [Z]. Washington: American Psychological Association，1999.

APA. Technical Recommendations for Psychological Tests and Diagnostic Techniques [J]. *Psychological Bulletin Supplement*，1954(51).

APA，AERA，NCME. *Standards for Educational and Psychological Tests and Manuals* [Z]. Washington: American Psychological Association，1966.

Bachman, L. Building and Supporting a Case for Test Use [J]. *Language Assessment Quarterly*, 2005(2).

Bachman, L., Palmer, A. *Language Testing in Practice* [M]. Oxford: OUP, 1996.

Bachman, L., Palmer, A. *Language Assessment in Practice: Developing Language Assessment and Justifying Their Use in the Real World* [M]. Oxford: OUP, 2010.

Davies, A. Test Fairness: A Response [J]. *Language Testing*, 2010(27).

Finocchiaro, M., Sako, S. *Foreign Language Testing: A Practical Approach* [M]. New York: Regents Publishing Company, 1983.

Harris, D. *Testing English as a Second Language* [M]. New York: McGraw-Hill, 1969.

Heaton, J. *Writing English Language Test* [M]. London: Longman, 1975.

Jensen, H. R. *Bias in Mental Testing* [M]. New York: Free Press, 1980.

Joint Committee on Testing Practices(JCTP). *Code of Fair Testing Practices in Education* [Z]. Washington: American Psychological Association, 1988.

Joint Committee on Testing Practices(JCTP). *Code of Fair Testing Practices in Education* [Z]. Washington: American Psychological Association, 2004.

Kane, M. An Argument-based Approach to Validity [J]. *Psychological Bulletin*, 1992(112).

Kane, M. Validating High-stakes Testing Programs [J]. *Educational Measurement: Issues and Practice*, 2002(21).

Kane, M. Certification Testing as an Illustration Argument-based Validation [J]. *Measurement: Interdisciplinary Research and Perspectives*, 2004(2).

Kane, M. Validation [A]. In: Brennan, R.(Ed.). *Educational Measurement* [C]. Westport: Greenwood Publishing, 2006.

Kane, M., Crooks, T., Cohen, A. Validating Measures of Performance [J]. *Educational Measurement: Issues and Practice*, 1999(18).

Kunnan, A. J. Fairness and Justice for All [A]. In: Kunnan, A. J.(Ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing* 9 [C]. Cambridge: Cambridge University Press, 2000.

Kunnan, A. J. Test Fairness [A]. In: Milanovic, M., Weir, C.(Eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference* [C]. Cambridge: Cambridge University Press, 2004.

Kunnan, A. J. Testing for Citizenship: The U. S. Naturalization Test [J]. *Language Assessment Quarterly*, 2009(6).

Lado, R. *Language Testing* [M]. London: Longman, 1961.

McNamara, T. F., Roever, C. *Language Testing: The Social Dimension* [M]. Oxford: Blackwell, 2006.

Messick, S. The Once and Future Issues of Validity: Assessing the Meaning and Consequences of Measurement [A]. In: Wainer, H., Braun, H. I.(Eds.), *Test Validity* [C]. Hillsdale: Lawrence Erlbaum, 1988.

Messick, S. Validity [A]. In: Linn, R. L.(Ed.), *Educational Measurement* [C]. New York: American Council on Education and Macmillan, 1989.

Mislevy, R., Steinberg, L., Almond, R. Design and Analysis in Task-based Language Assessment [J]. *Language Testing*, 2002(19).

Mislevy, R., Steinberg, L., Almond, R. On the Structure of Educational Assessments [J]. *Measurement: Interdisciplinary Research and Perspectives*, 2003(1).

Toulmin, S. *The Use of Argument*(Updated Edition) [M]. Cambridge: CUP, 2003.

Valette, R. *Modern Language Testing* [M]. New York: Harcourt, Brace & World, 1967.

Weigle, S. *Assessing Writing* [M]. Cambridge: CUP, 2002.

Willingham, W. W., Cole, N. *Gender and Fair Assessment* [M]. Mahwah: Lawrence Erlbaum Associates, 1997.

Xi, X. How Do We Go about Investigating Test Fairness? [J]. *Language Testing*, 2010(27).

Zieky, M. Fairness Review in Assessment [A]. In: Downing, S., Haladyna, T.(Eds.), *Handbook of Test Development* [C]. Mahwah: Lawrence Erlbaum, 2006.

定稿日期: 2017 – 11 – 21　　**【责任编辑　孙　颖】**