形成性评估的效度验证方法

顾永琦 李加义
新西兰惠灵顿维多利亚大学

© 2020 外语教育研究前沿（原名《中国外语教育》）(4)，57–63 页

提 要：有效的形成性评估对促进教学具有深远影响。然而，面向学习的形成性评估并不会自动促成学生的进步，形成性评估同样需要效度验证。本文简要介绍形成性评估的效度与效度验证，并着重阐述如何使用基于论证的效度验证框架以验证形成性评估的效度，为形成性评估的研究人员以及一线教师的评估实践提供切实可行的效度验证方法。

关键词：形成性评估；效度；基于论证的效度验证；外语教育

[中图分类号] H09 [文献标识码] A [文章编号] 2096-6105(2020)04-0057-07

1. 引言

所谓形成性评估的效度，就是对评估结果的解释与后续行动能否站得住脚、有无证据和理论支持。我们在《外语教育研究前沿》第3期上刊发的《形成性评估的效度》详细阐明了形成性评估效度的内涵（顾永琦、李加义 2020）。本文将重点探讨形成性评估的效度验证方法。

在实践中，抽象、复杂的效度验证框架在指导教师的评估实践过程中缺乏可操作性，这也对灵活动态的课堂评估的效度验证提出严峻的挑战。本文着重探讨可用于验证形成性评估效度的方法，旨在探索一套系统的、可操作的验证方法以检验形成性评估的有效性，以期对教师和研究人员在形成性评估的效度验证研究和实践上提供启示。

许多学者把形成性评估的效度集中在评估促进学生能力提高的程度上（consequential validity）（Kane & Wools 2019；Stobart 2012），认为只有促进学习进步的评估才是高效度的形成性评估。我们承认，评估效果固然是形成性评估效度的最终指标，但不应是唯一指标。在达到学习效果之前，形成性评估必须具备解释与使用效度。道理很简单：应该评估的都进行了评估，教师对评估信息的解释是准确无误的，也作了适当的反馈，并为学生提供了改进机会，这样的评估才能达成形成性的效果。

2. 效度验证

效度验证是确保测评效度的整体过程。这意味着评估前要明确评估目的和手段，以确保评估任务具有相关性和代表性，评估结果得到合理的解释和使用。在评估结束后，评估的效度验证也指收集证据证明评估实现了其目的并得到正确的解释和使用的过程。无论效度概念怎样发展，只有转化为可操作的效度验证模式，才能实现其真正的意义和价值（O'Sullivan & Weir 2011）。

2.1 传统的效度验证方法

对于课堂评估，首先需要解释的是课程忠实度（curriculum fidelity），即教师应确保评估内容与课程教学吻合，实现教、学、评在课堂中的融合统一。此外，评估应涵盖所有主要课程目标。核对清单（checklist）是衡量评估内容相关性和代表性的重要工具，把课程要求列在清单之上，需要评估的内容就一目了然了。课堂教学和评估具有灵活性，若教师发现学生理解目标内容存在问题，可以设计课堂任务来解决这一问题并再次进行评估，以判断是否需要进一步的讲解、练习，

或者可以进入下一个阶段。

除课程忠实度外，基于课堂的形成性评估还应该仔细研究其理论关联度。如果课程忠实度侧重对照课程目标检查评估活动的相关性和代表性，那么理论关联度则对照教学内容的理论构念来检查评估活动的相关性和代表性。例如，如果使用Bachman（1990）的语言能力框架，教师可能会问这样的问题：我似乎大部分时间都在评估“组构能力”，我评估过“语用能力”吗？怎么去评估呢？如果我相信交际理论，但我日常关注最多的是学生的语法是否正确，那么评估内容的代表性就出现了问题。

课堂评估作为教学和学习的组成部分，并不一定意味着它会自动有效。教师需要确保通过相关的和有代表性的评估活动获得的学习证据能得到适当的解释，并适用于教学和学习目的。课堂中的现场判断和决策对教师的专业知识及其对学生的了解提出了很高的要求。教师不断发展的专业能力可以确保他们对学生学习的解释不偏离课程要求，并且避免从学生的测试结果中得出笼统的结论。确保对评估结果作出适当解释的另一种方法是多次从多个来源获得证据以证明某一解释是正确的。课堂判断和决策大多涉及对学习的诊断。犹如人们多次拜访不同的医生，运用多种诊断手段，以寻求对健康状况最合理的解释，同样的原则也适用于诊断学生的学习。错误的诊断会导致不适当的教学行为。

鉴于人类判断的内在主观性，课堂评估在教师对学生学习的判断一致性和稳定性方面提出了实质的挑战。以下提高测试信度的方法可以迁移至课堂评估中。1）提高教师对评估目标和成功标准的共同理解。当同一所学校同一年级的所有教师对教学目标都有相近的理解时，就不太可能出现判断不一致的情况。2）运用各个层次的学生表现的范例。系统地观察和收集学生的语言范例是记录和掌握学生语言进步的重要工具（Gardner & Rea-Dickins 2002；Rea-Dickins 2008）。教师可以收集和讨论课堂范例，并保存起来以备将来使用。教师在评估这些范例时越有经验，他们在评估未来课程中的类似任务时一致性和稳定性就越高。3）创建评估量规、核对清单和工作表。随着教师在使用这些评估量规、核对清单和工作表进行评估方面的经验愈发丰富，他们在偶然的课堂判断中就变得越来越一致。4）建立教师发展共同体。在同伴的持续支持下，教师能够增长学生学习诊断一致性方面的专业知识。

以上效度验证方法即是收集不同方面的效度证据的传统思路（Messick 1989）。课程忠实度主要是内容效度或实质效度；观察学生解决问题的过程基本是实质效度；理论关联度主要看的是结构效度和实质效度；评估的一致性则是类推效度；如果发现教师判断与学生考试分数有关系，那就是外部效度的证据；而判断问题是否得到解决则为后效效度的证据。

2.2 基于论证的效度验证框架

在过去20年中，一个允许所有证据作为一个连贯的整体呈现的验证框架，即基于论证的效度验证框架（argument-based validation），越来越为教育评估领域所接受。通俗地说，在这个框架下，效度验证就是一个论证的过程。早在20世纪80年代，Cronbach（1988）就开始将测试的效度验证视为收集证据，以支撑测试设计、解释和使用的论证。多年来，Kane（1992，2001，2006）和Mislevy *et al.*（2003）完善了基于论证的效度验证方法，将测试的效度验证发展为一个连贯和实用的论证框架。Bachman（2005）、Bachman & Palmer（2010）采用并修改了这一框架；托福网考也使用基于论证的模式指导其效度验证（Chapelle *et al.* 2008）。Chapelle（2020）最近用整本书的篇幅介绍了基于论证的效度验证。该模式清楚地展示了收集证据的先后顺序以及各证据间的内部联系，使效度验证不再是证据的简单收集和罗列，在体现效度整体观核心原则的同时展开推理论证，使效度验证更具可操作性。李清华、孔文（2015）通过分析和比较现有的效度评估框架，提出了外语形成性评估的效度验证框架，认为形成性评估系统包括社会文化环境下的教学阶段和评估阶段。但该框架较为复杂，缺乏可操作性，作者并未依据该框架分析形成性评估的效度。Hopster-den Otter *et al.*

(2019)采用Kane的基于论证的效度验证框架，将其开发的测试平台中的小学算术题目用作形成性评估并检验其效度，但未就形成性评估给出清楚的定义，也缺乏对实际课堂评估情境的指导。

基于论证的效度验证框架分为两个论证步骤：1)解释和使用论证(interpretation and use argument，IUA)（Kane 2006）；2)效度论证(validity argument)（Kane 2013）。第一步，搭建一个环环相扣的推理链，通过分析测试表现与决策形成的推理链及其依据的假设来阐明分数解释和使用论证。简言之，明确概述根据评估结果作出的主要推论和主张(见图1)。在解释和使用论证中，通过观察学生的表现评出观察分(即原始分)，通过概化推理得到全域分，通过外推推理借由全域分预测目标域分，使用目标域分进行决策。第二步，系统地论证每一项主张或推论的可靠性。效度论证使用了Toulmin(2003)的论证模型(见图2)。

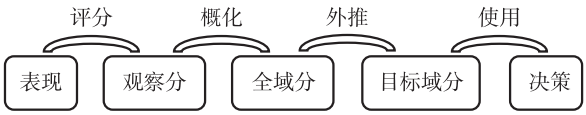

图1　解释和使用论证模式（Kane 2006）

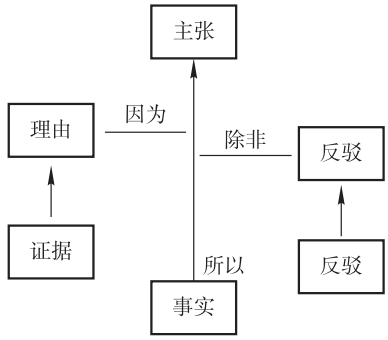

图2　论证模型（Toulmin 2003）

图2使用Toulmin(2003)的论证模型展示了一个简单的论证，主要包括：1)主张(claim)，即对学生能力的解释；2)事实(data)，即评估过程的特征及学生在评估过程中的表现；3)理由(warrant)，即在事实基础上提出主张的合理性；4)证据(backing)，即用以支撑理由的理论、研究结论或已有证据；5)反驳(rebuttal)，即与预期结论相反的解释。论证过程是："因为"有充分的证据支撑，"所以"主张成立，"除非"有证据证明反驳成立。图2的模型中因为反驳没有推翻主张(结论)，所以主张成立。

2.2.1 基于论证的形成性评估效度验证

效度验证的最大挑战是将效度理论付诸实践(Chapelle 1999)，基于论证的效度验证框架在一定程度上实现了效度整体观，结构清晰，论证严谨，但其抽象性和较强的逻辑性增加了效度验证的实际操作难度。为解决这一问题，笔者提出适用于形成性评估效度论证的方法，具体到各个推理的假设和需要搜集的证据等，并在此基础上结合课堂实例向教师阐明如何将其运用到形成性评估实践中。

第一步：解释和使用论证。图3展示了课堂形成性评估中的论证链，推理过程分为：1)评价(evaluation)：教师观察学生在课堂活动中表现出的能力并作出评价；2)概化(generalisation)：教师以此推断出学生在类似情况下完成类似任务的能力；3)解释/外推(explanation/extrapolation)：在对类似任务的成功表现进行一系列观察之后，教师进行两种类型的外推推理(解释和外推)，以此推断学生的语言能力是否已经达到课程标准(外推)，或者学生是否具备理论语言构念中所应具有的能力(解释)；4)使用(utilisation)：教师使用这些信息作出决策(如学生可以进入更高的层次，或者需要付出更多的努力进行改进)。

表1阐述了课堂形成性评估的四种主要主张。这四种主张及其相关推理构成了解释和使用论证。

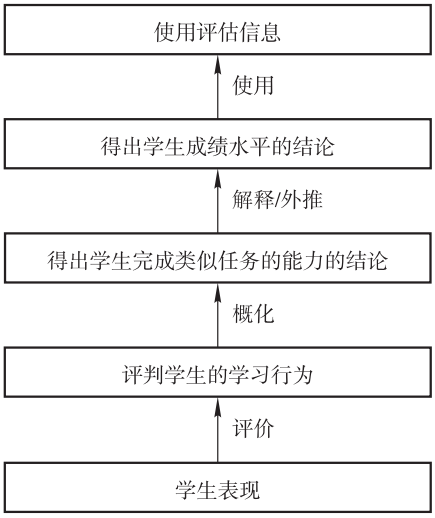

图3　课堂形成性评估的论证链

表1　课堂形成性评估的主张及其相关推理

形成性评估主张	关联推理
主张1：形成性评估的判断是正确的。	评价：将表现与判断联系起来
主张2：形成性评估对学生成绩的判断是可信的。	概化：将一次观察与对所有类似表现的概况判断联系起来
主张3：形成性评估反映学生真实的语言成绩。	解释：将判断与理论构念的解释联系起来
	外推：将判断与课程和教学目标联系起来
主张4：形成性评估用来改善学习结果。	使用：将解释与使用联系起来

第二步：效度论证。在阐明解释和使用论证后，下一步是论证所有的主张和推论都是正确的。论证过程遵循了Toulmin（2003）论证模型“事实→主张”的机制，也就是上一推理过程的结论（即主张）经论证后即为下一推理过程的事实。表2列出了形成性评估效度论证的根据及其支撑证据。

表2　形成性评估效度论证的根据及其支撑证据

推理	假设（根据）	证据（支撑）
评价	·明确评估目标和成功标准 ·恰当地选择和使用评估工具 ·遵守形成性评估的关键程序（收集证据、解读证据、提供反馈、后续行动）	·对教师和学生进行访谈，了解他们对评估目标和成功标准的理解 ·进行课堂话语分析，了解评估类型及其如何进行 ·分析课堂录像的内容，了解如何进行评测和解释，提供了什么反馈，反馈后采取了什么行动
概化	语言任务的课堂表现在类似活动、评估者、评估形式和评估场景中是一致的	·多个证据来源 ·多次观察 ·观察活动样本在内容域活动中的代表性 ·观察条件样本在内容域条件中的代表性
解释	课堂评估任务涉及的能力和过程与语言能力理论构念中适用于教学环境的能力和过程相同	校验构念相关性和代表性： ·访谈 ·观察评估过程 ·语篇/会话分析 ·评估活动的分析
外推	评估任务和材料代表课程相关层面（内容域）的目标知识、技能和能力	·评估活动覆盖内容域的判断证据； ·评估活动内容的分析
使用	·向使用者提供的信息是有用和充分的； ·评估信息用于调整学习和教学	·分析反馈（类型、信息性）； ·分析学习与教学调整； ·必要的调整周期； ·提高考试成绩

2.2.2 课堂形成性评估实例

在实际课堂场景中，基于论证的效度验证方法可以直接应用于形成性评估的效度验证。笔者以一位英语教师的课堂形成性评估反思日志为例，展示如何使用基于论证的方法进行效度验证。教师写道："我们今天课上进行了一项阅读分享活动。我在班里巡视的时候观察了三组学生的表现，发现他们在理解方面存在很多问题。我意识到很多学生不能理解此类文章。这一定是因为缺乏词汇量，所以我让学生从现在起每周背50个生词。"

根据Gu（2020）对形成性评估的定义，以上评估事件包含了形成性评估实践的四个步骤：收集证据、解释证据、提供反馈、后续行动，构成了形成性评估实践完整的循环，因此可以认定为形成性评估。但此次形成性评估的质量如何呢？接下来我们使用基于论证的方式对此次评估进行效度验证。

第一步：解释和使用论证。图4展示了形成性评估实例的论证链，概括出形成性评估的主张及推理过程。

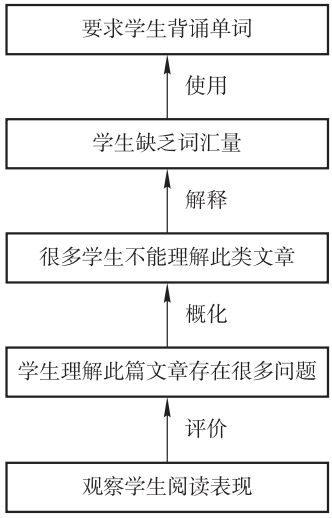

图4　课堂实例的解释和使用论证链

第二步：效度论证，即为上述每一项主张提供有效性论证，如果任何一项主张被推翻，则证明形成性评估论证链无效。由于篇幅有限，笔者不能一一展示每项主张的论证过程，仅以"解释推理"的论证过程为例（见图5），反证教师提出的主张不成立。换句话说，教师对评估结果的解释是错误的。在此情况下，无论后续行动多么有用，都无助于解决真正的学习问题，因而无法达到形成性评估的效果。

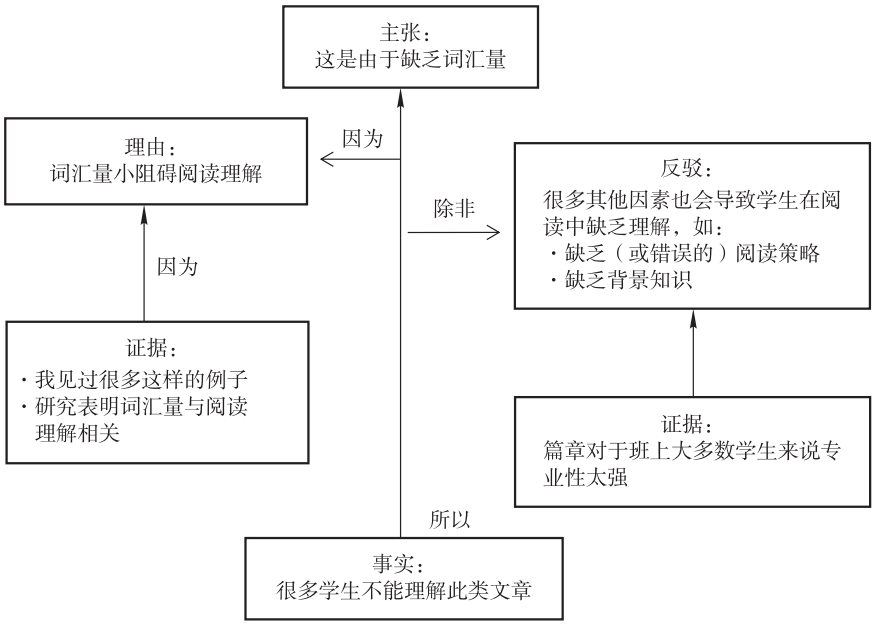

图5　效度论证课堂实例

为了清楚地阐释形成性评估的效度论证方法，图5几乎是一段论证过程的“慢动作”回放。在现实课堂实时的教学互动中，形成性评估发生在教与学的临时瞬间(moments of contingency)(Black & Wiliam 2009)，每时每刻都可能发生即时形成性评估。与课前计划好的形成性评估不同，课堂即时形成性评估具有“即时性”和“动态性”特征(杨华、文秋芳 2013)。教师要实时调整教学，以便有效地发挥形成性功能(Black 2009；Leahy *et al*. 2005；Wiliam 2006)。教师应非常快速地、非正式地进行效度验证，而不需要在纸上列出所有的推理或绘制论证图。

虽然本文的实例聚焦课堂即时形成性评估，但提出的效度验证框架适用于所有的形成性评估实践。无论是课前计划好的课堂形成性评估，还是跨越几节课或更长循环的形成性评估，评估的基本主张是类似的，不同的只是评估任务的规模而已。因此，既然本框架可以适用于转瞬即逝的即时形成性评估，那么对中长循环的计划性形成性评估的效度验证就更容易进行了。

3. 总结与建议

形成性评估不一定都会有利于教与学，有时也会存在质量问题，因此需要进行效度验证。本文通过介绍效度验证方法，即如何确保评估对其预期目的有效，详细地阐释了基于论证的形成性评估效度验证方法，以期为教师将其应用于形成性评估实践、检验课堂形成性评估效度、调整教学，进而提高学生的学习能力提供指导和借鉴。

笔者建议：1)教师应理解和掌握形成性评估的效度验证方法，尝试在形成性评估实践中加以运用。在任何情况下，计划性形成性评估和即时互动性形成性评估都应进行效度验证。非正式的形成性评估验证应随时在课堂教学发生时进行，正式的效度验证可以采取同伴互评和课堂观察的形式；2)教师应建立校内外形成性评估实践共同体，定期组织座谈会，以便同伴教师通过共享平台对形成性评估的效度验证进行互相帮助；3)研究人员应不时地加入评估实践共同体，加强各领域之间的合作，以便带来更多的专业知识和实践经验，并监督形成性评估的达成；4)教师可以录制自己的课堂视频，以便对教学设计和教学实践进行总结，分析计划性形成性评估和即时互动性形成性评估做法，进一步提炼和拓展包含形成性评估的教学实践。

参考文献

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing* [M]. Oxford: Oxford University Press.

Bachman, L. F. 2005. Building and supporting a case for test use [J]. *Language Assessment Quarterly* 2: 1-34.

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

Black, P. 2009. Formative assessment issues across the curriculum: The theory and the practice [J]. *TESOL Quarterly* 43: 519-524.

Black, P. & D. Wiliam. 2009. Developing the theory of formative assessment [J]. *Educational Assessment, Evaluation and Accountability* 21: 5-31.

Chapelle, C. A. 1999. Validity in language assessment [J]. *Annual Review of Applied Linguistics* 19: 254-272.

Chapelle, C. A. 2020. *Argument-based Validation in Testing and Assessment* [M]. Thousand Oaks: SAGE.

Chapelle, C. A., M. Enright & J. Jamieson (eds.). 2008. *Building a Validity Argument for the Test of English as a Foreign Language* [M]. London: Routledge.

Cronbach, L. J. 1988. Five perspectives on validity argument [A]. In H. Wainer & H. I. Braun (eds.). *Test Validity* [C]. Hillsdale: Routledge. 3-17.

Gardner, S. & P. Rea-Dickins. 2002. *Focus on Language Sampling: A Key Issue in EAL Assessment* [M]. London: National Association for Language Development in the Curriculum.

Gu, P. Y. (顾永琦). 2020. *Classroom-based Formative Assessment* [M]. Beijing: Foreign Language Teaching and Research Press.

Hopster-den Otter, D., S. Wools, T. J. H. M. Eggen & B. P. Veldkamp. 2019. A general framework for the validation of embedded formative assessment [J]. *Journal of Educational Measurement* 56: 715-732.

Kane, M. T. 1992. An argument-based approach to validity [J]. *Psychological Bulletin* 112: 527-535.

Kane, M. T. 2001. Current concerns in validity theory [J]. *Journal of Educational Measurement* 38: 319-342.

Kane, M. T. 2006. Validation [A]. In R. L. Brennan (ed.). *Educational Measurement (4th Ed.)* [C]. Washington, D. C.: Rowman & Littlefield. 17-64.

Kane, M. T. 2013. Validating the interpretations and uses of test scores [J]. *Journal of Educational Measurement* 50: 1-73.

Kane, M. T. & S. Wools. 2019. Perspectives on the validity of classroom assessments [A]. In S. M. Brookhart & J. H. McMillan (eds.). *Classroom Assessment and Educational Measurement* [C]. New York: Routledge. 11-26.

Leahy, S., C. Lyon, M. Thompson, D. William. 2005. Classroom assessment: Minute-by minute, day-by day [J]. *Education Leadership* 63: 18-24.

Messick, S. 1989. Validity [A]. In R. Linn (ed.). *Educational Measurement (3rd Ed.)* [C]. Washington, D. C.: American Council on Education. 13-103.

Mislevy, R. J., L. S. Steinberg & R. G. Almond. 2003. On the structure of educational assessments [J]. *Measurement: Interdisciplinary Research & Perspective* 1: 3-62.

O'Sullivan, B. & C. Weir. 2011. Language testing and validation [A]. In B. O'Sullivan (ed.). *Language Testing: Theory and Practice* [C]. Oxford: Palgrave. 13-32.

Rea-Dickins, P. 2008. Classroom-based language assessment [A]. In E. Shohamy & N. H. Hornberger (eds.). *Encyclopedia of Language and Education (2nd Ed.) Vol. 7: Language Testing and Assessment* [C]. Berlin: Springer. 257-271.

Stobart, G. 2012. Validity in formative assessment [A]. In J. Gardner (ed.). *Assessment and Learning* [C]. London: SAGE. 133-146.

Toulmin, S. E. 2003. *The Uses of Argument* (*Updated Ed.*) [M]. Cambridge: Cambridge University Press.

Wiliam, D. 2006. Formative assessment: Getting the focus right [J]. *Educational Assessment* 11: 283-289.

顾永琦、李加义，2020，形成性评估的效度[J]，《外语教育研究前沿》(3)：34-41。

李清华、孔文，2015，外语形成性评估的效度验证框架[J]，《外语教学理论与实践》(1)：24-31。

杨华、文秋芳，2013，课堂即时形成性评估研究述评：思考与建议[J]，《外语教学理论与实践》(3)：33-38。

作者简介

顾永琦，新西兰惠灵顿维多利亚大学语言学与应用语言研究学院副教授，博士生导师。主要研究领域：语言测评、学习策略、词汇教学、教师发展。电子邮箱：peter.gu@vuw.ac.nz

李加义，新西兰惠灵顿维多利亚大学语言学与应用语言研究学院博士研究生。主要研究领域：语言测评、外语教学。电子邮箱：jiayi.li@vuw.ac.nz

（审稿编辑：许宏晨）