Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊

Volume 1 | Issue 1 Article 4

2020

去"构念"的测验效度验证

Stephen G. Sireci

Follow this and additional works at: https://www.ce-jeme.org/journal

Recommended Citation

Sireci, Stephen G. (2020) "去"构念"的测验效度验证," *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*: Vol. 1 : Iss. 1 , Article 4.
Available at: https://www.ce-jeme.org/journal/vol1/iss1/4

This Article is brought to you for free and open access by Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊. It has been accepted for inclusion in Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊 by an authorized editor of Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊.

去“构念”的测验效度验证

Stephen G. Sireci

University of Massachusetts Amherst

摘要

构念效度理论 (construct validity theory) 为教育和心理测验的“效度”提供了最全面的描述。“构念效度”一词最初由1954年的《心理测试与诊断技术的技术建议》引入 (American Psychological Association [APA], 1954)，随后由两名1954年委员会成员 Cronbach 和 Meehl (1955) 进行了阐释。构念效度理论对效度的理论描述产生了巨大影响，但没有在最近两版的《教育与心理测验标准》(American Educational Research Association [AERA] et al., 1999, 2014) 中得到明确支持。在本文中，我将回溯有关构念效度理论对测验效度检验的重要性的讨论历史，并且识别构念效度理论的基本要素——这些要素对于有着特定用途的测验的效度验证至关重要。同时，我将提出一个侧重于测验使用，而非测验构念的效度验证框架。这个解构（译：“去构念”）的方法包含四步：（1）明确测验目的，（2）确认测验使用可能造成的消极后果，（3）将测验目的、可能的误用与 AERA et al. (2014) 的《教育与心理测验标准》中的五个效度证据来源进行交叉比对，（4）优先考虑一些效度证据来源，以建立一个充足的、注重测验使用以及后果的效度论证。该去构念效度验证的目标是接受构念效度理论涉及的主要原则，利用这些原则发展一个心理测量学家、法院法官、决策者和一般公众都能理解的、连贯的、全面的效度论证，且与 AERA et al. (2014) 的《标准》保持一致。

关键词

构念效度；
测验标准；
效度

伴随着美国和欧洲标准化考试的兴起 (Sireci, 2009)，教育和心理测验的效度理论可追溯到 100 多年前 (比如， Thorndike, 1904)。在过去的一个世纪里，效度的概念引起了广泛的争论，研究者提出了多套理论和术语 (Newton & Shaw, 2013)。为了在效度观点上达成一致，美国心理学协会 (APA) 在 1952 年进行了第一次尝试，发表了《心理测试和诊断技术的技术建议：初步提案》(APA, 1952)。在APA、美国教育研究协会 (AERA) 以及全国教育测量委员会 (NCME[1]) 的共同努力下，这项关于测验开发、使用和评价的专业指南的提案促成了教育与心理测试标准的第一个正式版本的发布，其全称为《心理测试与诊断技术的技术建议》(APA, 1954)。《建议》开启了新的篇章，成为了测量研究者内部就效度是什么以及测验效度验证过程如何进行等达成共识的开端。而《建议》引入的一个新概念— 构念效度，则是这些共识观点的一部分。

编写 1954 年《建议》的委员会中的两名成员 Lee J. Cronbach 和 Paul Meehl 随后发表了一篇论文，更透彻地解释了“构念效度” (Cronbach & Meehl, 1955)。该文成为效度理论最具开创性的著作之一，代表了“教育和心理测验的一切效度都是构念效度”这一共识观点的诞生。然而，尽管构念效度理论的许多主要原则对于测验效度验证仍然很重要，但构念效度理论就像多数关于效度理论的共识观点一样，最终失去了它的显著地位。

在本文中，我回顾了一小段构念效度理论的历史，以解释为什么测验史上众多著名的效度理论家用以构念效度为中心的一元概念来描述效度。接下来，我依据 1952 年至 2014 年间的七版《教育和心理测验标准》(以下称为《标准》)，解释这个一元概念为什么没有在测验效度验证工作中得到重视。在第三节中，我将重点介绍在最近两版《标准》 (AERA et al., 1999, 2014) 中对效度的现代定义。我也将说明以 AERA et al. 的《标准》 (1999, 2014) 中的五种效度证据来源为框架的测验效度验证方法，是如何体现了构念效度理论的思想。本文的目的是回顾效度理论的历史文献，解释那些在 20 世纪的测验效度验证中一直存在并仍然很重要

[1]1961 年，教育中使用的测量国家理事会更名为教育测量国家理事会，该名称沿用至今(见NCME.org)。

通讯作者: Stephen G. Sireci. sireci@acad.umass.edu. University of Massachusetts Amherst. 813 North Pleasant St. Amherst, MA 01003.

译者: 张佳慧. zhangjiahui_cicabeq@bnu.edu.cn. 北京师范大学. 北京市海淀区新街口外大街19号, 100875.

的关键概念。

“传统”效度术语简介

在重点讨论构念效度的历史之前，我们需要承认一个不幸的事实：有关不同效度种类的提法非常广泛，似乎渗透到了众多心理学和教育领域的入门教科书中。我在表 1 中列出了在 AERA、APA 和 NCME 的七版《标准》中使用的不同效度术语。如表 1 所示，效度术语从效度的“分类”演变为效度的“类型”，再到效度的“方面”，之后又回到“分类”，最后演变为“效度证据的来源” —— 最近 20 多年都保持了该说法。有趣的是，“内容”是整个七十年的历史中唯一持续存在的术语。但许多人认同的效度三种“传统形式”是：构念效度、内容效度和效标关联效度。我将在下一节通过追溯历史来描述构念效度。在这里，我只简单地介绍了内容和效标关联效度的传统提法 (更全面的描述见 Kane, 2006, 2013)。

内容效度指测试内容在多大程度上代表了测验所针对的能力，以及该内容与测试目的的一致程度。显然，内容的有效性是教育和认证考试的一个先决条件，因为这类考试需要显示出与目标课程或工作领域的“校准” (Crocker, 2003; Martone & Sireci, 2009; Sireci, 1998; Sireci & Faulkner-Bond, 2014)。内容有效性的评价通常由内容领域专家对测验项目与所测领域的相关性和代表性进行判断。

效标关联效度指为了对测验分数的解释进行评价，对测验分数与其他重要变量 (即效标) 之间的关系进行统计评价。效标关联效度的研究通常分为对预测效度和同时效度的研究。在预测效度研究中，对测验分数的评价是取决于测验需要预测的效标 (比如, 大学招生考试分数能够预测大学平均绩点的程度)。在同时效度中，对测验分数的评价根据它们与同一时间点收集的其他变量之间的关系 (比如, 5 年级数学考试分数与学生在 5 年级数学课上的成绩)来决定。在最近两版《标准》(AERA et al., 1999, 2014) 中，效标关联效度的理念在“基于与其他变量关系的效度证据”中得到了体现。

在下文中，我将回顾这些不同的效度证据来源的现代概念及其对效度验证的重要性。之后我将对构念效度理论进行简要的介绍。

1 构念效度简史

要理解构念效度，我们必须先理解“构念”的哲学概念。Cronbach 和Meehl (1955) 将构念定义为，

> 某些人为假定的人的属性，假设能够在测验表现中得到反映 ... 我们期望一个人在任何时候都拥有或不拥有一个定性的属性 ... 或者在某种程度上拥有一个定量的属性 ... (p. 283)。

他们对构念效度的描述是

> 每当测验被解释为某种没有“操作定义”的属性或质量的度量时，就会涉及构念效度验证。研究人员面临的问题是：“什么构念导致了测验表现的差异？” (p. 282)

这些定义中有一些同义反复问题，因为个体对测试项目的作答反映了测验理应测量的构念。但是很明显，构念的概念必须出现在测验被创造出来之前。

虽然 Cronbach 和 Meehl (1955) 强调了构念效度的重要性，但很重要的一点是他们并未认定构念效度适用于所有类型的测验。相反，他们引入了这一概念，是为了解决更偏心理的测验（比如投射测验和人格测验）的效度验证问题。诚然，从内容或效标关联的角度来评价这些测验是比较困难的。正如 Cronbach 和 Meehl 所描述的那样，“引入构念效度是为了在常规效度验证观点不能适用时，厘清测验开发所需的研究类型” (p. 299)。在不存在特定课程目标和内容细则的非教育测验环境中，我们需要这个被测量的“某物”的概念，例如潜在的特质或构念。为此，Cronbach 和 Meehl 声称“... 在没有公认的标准或内容领域足以定义要测量的量时，必须探讨”构念效度 (p. 282)。

Loevinger (1957) 认为，当人们被测试时，构念总是隐含的；任何操作定义、标准或内容领域都不足以取代对测验分数的构念解释。在用这种方法描述构念效度时，Loevinger 巩固了所有效度都是构念效度这一概念。她声称，“由于预测、同时和内容效度本质上都是特例，所以从科学的角度来看，构念效度就是效度的全部” (p. 636)。这一观点获得了认可，并逐渐成为共识 (cf. Ebel, 1961)。毕竟，没有计量心理学家想被指责为不科学。然而，达成这一共识还是花了一段时间。例如，题为《教育与心理测验和手册的标准》(AERA et al., 1966) 的下一版《建议》朝着效度一元论迈了一小步—— 它将 1954 年的《建议》中提到的四种效度“类型” (内容、同时、预测和构念)改为效度的三个“方面” (内容、效标关联和构念)。从“类型”到“方面”的变化是细微的，但该变化明确地反映出越来越多的人认为不存在独立的效度类型，而是存在同等的效度类型。

下一版《标准》——《教育和心理测验标准》(AERA et al., 1974) —— 在推广效度是一个统一概念方面迈出了一大步。文中写道：

> 效度的种类取决于人们希望从测验分数中得出的推论的种类 ... 相互依赖的推理性解释通常被描述并大多用来总结测试用途：效标相关效度 ... 内容效度和构念效度 ... 效度的

表 1. 当前和以前版本的《标准》中使用的效度术语总结

发表物	效度术语
《心理测试和诊断技术的技术建议：初步提案》(APA, 1952)	效度的分类：预测，状态，内容，全等
《心理测试与诊断技术的技术建议》(APA, 1954)	效度的类型：构念，同时，预测，内容
《教育与心理测验和手册标准》(AERA et al., 1966)	效度的类型：效标相关，构念相关，内容相关
《教育与心理测验标准》(AERA et al., 1974)	效度的方面：效标相关，构念相关，内容相关
《教育与心理测验标准》(AERA et al., 1985)	效度的分类：效标相关，构念相关，内容相关
《教育与心理测验标准》(AERA et al., 1999)	效度证据的来源：测验内容，作答过程，内部结构，与其他变量的关系，测验结果
《教育与心理测验标准》(AERA et al., 2014)	效度证据的来源：测验内容，作答过程，内部结构，与其他变量的关系，测验结果

> 这些方面是可以被单独讨论的，但仅是为了方便。这些方面在操作上和逻辑上相互关联；只在少数特定情况下，仅有其中一种是重要的。对一个测验的深入研究通常涉及到关于所有类型的效度的信息 (AERA et al., 1974, pp. 25-26)。

《标准》提倡统一的效度观，同时保留了效度验证所涉及的不同方面 (Cronbach, 1971)。然而，他们并没有说这个一元概念化是以构念效度为中心的。下一版《标准》，现在叫做《教育与心理测验标准》(AERA et al., 1985)，更明确地提倡一元概念，但没有指出构念效度是构建这个一元概念的驱动力，如以下节选所示：

> 效度 ...是一个一元的概念。虽然证据可以通过多种方式积累，但效度一般指的是在多大程度上证据能支持从分数中得出的推论。被验证的是关于测验的特定用途的推论，而不是测验本身 (AERA et al., 1985, p. 9)。

因此，在 20 世纪 80 年代中期，人们一致认为效度是一个一元概念。同时期的一位颇具威信和影响力的效度理论家 Samuel Messick，借鉴了 Cronbach 和 Meehl (1955)、 Loevinger (1957) 和其他人的 (比如，Guion, 1977) 观点，提出这个一元概念本质上就是构念效度 (Messick, 1975, 1980, 1988, 1989)。在他收录于第三版《教育测量》的里程碑式的章节中 (Messick, 1989)，Messick 运用哲学、逻辑论证以及对教育和心理测验的文献和实践的全面综述，提出对测验分数的所有解释以及对测验使用情况的评价都必须与测验想要测量的构念相关联的观点。

Messick 关于构念效度作为统一力量的观点类似于 Loevinger (1957)的观点；然而，在讨论 Loevinger 关于效标和内容效度是构念效度的“特例”的原文时，Messick (1989) 评论道：

> 这一章节更进一步 ...这里认为 ...仅依靠效标效度或内容覆盖率是不够的。测量工具的意义及其构念效度要一直被探讨——这不仅是为了支持测验的解释，同时也是为了证明测验的使用是合理的 (p. 17)。

对 Messick 来说，测验过程意味着对构念的测量，因此所有效度的内在都是构念效度。正如他所说，“...统一的效度观的实质是：基于分数的推理的适当性、意义性和有用性是不可分割的，而基于经验的对构念的解释是使多种效度类型的得以统一的驱动力” (Messick, 1989, p. 64)。显然，他的意图是消除不同类型效度的概念。对于许多效度理论家 (比如， Guion, 1980) 来说，他成功了。

2 超越构念效度理论：测验效度验证

Messick (1989) 关于效度的章节可能是有史以来对这一主题最全面的论述。该章节由两个部分构成——测验的来源/正当性证明，以及测验的功能/结果。Messick (1989) 从六个哲学取向 (逻辑实证主义、相对主义、理性主义、工具主义、现实主义、建构主义) 和五个探询系统 (莱布尼茨、洛克、康德、黑格尔和辛格)对效度进行了描述。透过所有这些“哲学比喻” (p. 21)，Messick 论证了效度的一元概念即构念效度。正如他所说，“如果构念效度被认为依赖于单一的哲学基础，例如逻辑实证主义，且该基础是有缺陷和错误的，那么构念效度可能会因其根本的缺陷而被排除出讨论范围” (p. 22)。

虽然 Messick 的学术论证是令人信服的，但也有一些批评者认为他的理论体系限制性过强 (Ebel, 1977; Sireci, 1998; Yalow & Popham, 1983)，或过于晦涩，无法有效地促进对测验效度检验的实践 (Shepard, 1993)。这些批评很可能起源于他坚定地认为所有效度都是构念效度。例如，虽然 Messick (1989) 得出结论，所有效度都是构念效度，但他也指出，“效度是一个一元但有多个侧面的概念” (p. 14)，且各个侧面之间“不仅是相互关联的，而且有重叠” (p. 20)。像“统一的侧面”和“相互联系但重叠”这样的术语恰当地描述了 Messick 的概念体系，但这些术语在概念上是复杂的，也没有帮助实践者和非专业受众更多地了解效度。由于这个原因，Messick (1989) 的章节被批评为难以理解，同时缺乏对效度检验实践的指导。

例如，Shepard (1993) 在批评其 Messick 的章节时，赞扬了 Messick 的主要观点，但要求 Messick 提供一个“更简单的模型，从而将效度相关的问题进行排序，并且说明若要支持测验的使用，哪些效度问题必须得到回答，同时，哪些问题是为了改进学术理论而提出而并非是眼前的急需回答的问题” (p. 407)。Shepard 同意 Messick 关于构念效度的理论观点，但是认为这无助于指导效度验证实践。与 Messick 不同，Shepard 鼓励测验评估人员“ . . . 直接问‘这项测验实践主张做什么？’并围绕这个问题组织收集效度证据” (p. 408)。这种方法无论对测量实践者还是对非专业人士都很有吸引力。虽然 Shepard 的章节支持 Messick 的构念效度理论，但值得注意的是，她提出的的建议，即通过评估测验主张来开展效度验证调查，并不需要使用构念效度理论的深奥术语。

另一个用于测验效度验证的解构方法是由 Kane (1992, 2006, 2013) 发表的基于论证的方法。Kane (1992) 通过提供一个以大量解释证据来支持测验使用和解释的过程，从而回避了对效度的理论层面的系统表述。他借鉴了Cronbach (1971, 1988)，提议确定测验分数的预期用途和解释，并将其作为效度验证的框架。他的“基于论证的方法”涉及建立一个“解释性论证作为收集和呈现效度证据的框架” (p. 527)。解释性论证使用包括观察、概括、外推和理论在内的“证据类别”来建立论证，以支持为特定目的服务的测验使用或测验分数解释。这种基于证据的论证被称为“效度论证”，代表了一种对解释论证的合理性的评价。

Kane (2006) 将这种基于论证的效度验证方法分为两步。首先，通过解释性论证，明确想对测验分数继进行怎样的解释和其用途。然后，通过评价该解释性论证的元素来发展效度论证。该方法提供了一种实用的方式——它可用于评价一个测验对于某个特定目的的效用，并且无需援引不同类型或方面的效度。 Kane (2013) 将他的基于论证的方法扩展到包括“解释和使用论证”，并描述了他在提出这一方法时的逻辑：

> 这个基于论证的方法是为了避免强形式的构念效度所要求的充分发展的正式的理论；同时也避免弱形式的构念效度的开放性和模糊性 — 即任何涉及被测属性的关系的数据都可以被认为是有用的 (pp. 8-9)。

当提到“强”和“弱”形式的构念效度时，Kane指的是 Cronbach (1989) 的区分，即构念效度验证的理想的 (强的) 方法会由正式理论和一系列可能持续出现的假设检验驱动；而不足的 (弱的) 方法则侧重于已有的数据和“混杂的、仅仅边际相关的发现的集合” (Kane, 2013, p. 7)。因此，基于论证的效度验证方法承认，理想的效度验证通常是不可能的，但无论如何须提出充足的证据来支持为实现特定目的而使用该测验的做法。

Kane基于论证的方法基本上得到了当前版本的《教育和心理测验标准》(AERA et al., 2014) 的支持，文中指出，“效度是指证据和理论在多大程度上支持针对测验目的的测验分数解释” (p. 11)。关于效度验证，《标准》很明显采用了基于论证的方法。例如，他们指出：

> 一个强有力的效度论证将各种证据整合为一个连贯的整体，来说明现有的证据和理论在多大程度上支持为特定用途而对测验分数的解释 . . . 对某一特定用途的测验解释取决于一系列构成效度论证的命题，等效度证据积累到一定程度就可以对计划的测验解释做一个总结性的判断 (pp. 21-22)。

尽管 AERA et al. (2014) 的《标准》提倡使用效度论证的概念，他们没有沿用 Kane (1992, 2006, 2013) 的说法。也就是说，他们没有要求建立一个解释性论证，也没有通过基于构念的视角描述效度。不过，他们要求对测验所测构念要有清晰的定义，以及对测验目的要进行明确声明。正如《标准》中对效度的定义，效度验证被描述为提供证据支持明确声明的测试目的的过程。因此，AERA et al. 的《标准》 (2014) 中描述的基于论证的方法从本质上讲是可行的。

为了指导效度论证的开发，即支持为了特定目的而使用某测验，AERA et al. (2014) 的《标准》规定了五种有效证据来源，“可用于评价某一特定用途的测验分数的拟议解释的效度” (p. 13)。这些效度证据来源是分别基于 (1) 测验内容、(2) 作答过程、(3) 内部结构、(4) 与其他变量的关系以及 (5) 测验结果。值得注意的

是，在建立该效度验证框架时，《标准》将效度验证的重点放在测验解释和使用上，并且避免了不同类型或方面的效度——包括构念效度。事实上，就像在 AERA et al. (1985) 的《标准》中一样，过去两版《标准》将效度定义为一个一元的概念，但未赋予构念效度权威地位。例如，在描述五种效度证据来源时，他们说道，

> 这些证据来源可能说明了效度的不同方面，但它们并不代表不同类型的效度。效度是一个一元概念。是针对所提出的测验使用，所有累积的证据支持测验分数预期解释的程度 (pp. 13-14)。

2.1 Mislevy的社会认知视角

在我们结束对效度文献的历史回顾之前，有必要承认 Mislevy 的工作的重要性 (比如， 2009, 2018)。Mislevy 以“社会认知”视角构建了效度验证框架，扩展了 Messick (1989) 的对测验使用的社会考量；同时，Mislevy 讨论了测验在心理测量模型上的差异是如何体现以及要求模型与从测验分数得出的推论之间的正式联系。正如 Mislevy (2009) 所述，“测验效度的一个基本要素是，在某种应用中使用给定的模型是否能如预期一样为组织观察和指导行动提供坚实的基础” (p. 83)。这种效度观点强调测验的使用， AERA et al. (1999, 2014) 的《标准》也是如此；和Messick一样，它也承认评价所处的社会环境的差异。

借用科学哲学中的术语，Mislevy 的社会认知视角支持 “…建构-现实主义的效度观” (Mislevy, 2009, p. 84)——Messick (1989) 也曾对此进行了讨论。这个观点是“现实主义的”，因为它假定被测量的客体真实存在，但同时也是“建构主义的”，因为它承认对于不同的测验开发人员、以及在不同测量条件和情境下，构念的概念化和测量存在很大差异。正如 Mislevy (2009) 所描述的那样，“建构主义-现实主义观点认为，虽然模型是人为构造的，但成功的模型能够识别并且体现某些模式，而这些模式描述了现实世界中更为复杂现象的多个方面” (p. 95)。因此，社会认知视角与《标准》的定义是一致的：效度指的是在多大程度上一个有特定目的的测验能够被证据和理论证明是合理的，因此它比从纯粹现实主义角度出发的相对狭义的概念化更有用 (比如，Borsboom et al., 2004)。

2.2 效度理论与标准进化的总结

在结束了对构念效度理论的历史以及对《教育与心理测验标准》的变化的简短回顾后，我们来到了一个注重实践的部分。由于效度指的是证据和理论在多大程度上支持为实现特定目的而使用某一测验，因此验证过程中收集和评价证据的重点是提供为达成该目的而使用测验分数的合理解释。因此，AERA et al. (2014) 的《标准》为我们提供了一个收集和组织效度证据的框架。《标准》指出对效度的评价必须根据测验分数的具体用途而进行，并强调单一证据不太可能构成充分的效度论证。以上论述都没有提到构念。既然这种解构的检验效度验证方法已经被证明是合理的了，那么我将说明如何在实践中应用它。

3 使用 AERA et al. (2014) 《标准》作为效度验证框架

在 Sireci (2013) 一文中，我提出了一个使用 AERA et al. (2014) 《标准》作为框架的三步效度验证过程。步骤包括：(1) 明确阐明测验目的；(2) 考虑测验可能的误用；(3) 将测验目的与潜在的误用与《标准》中效度证据的五个来源进行交叉比对。不过，我补充了第四步以承认不可能进行所有理论上可行的效度研究的事实。第四步是，(4) 对要做的效度研究划分优先顺序。最后一步是必要的，它可以确保效度论证是建立在证据的基础上的——这些证据聚焦于将测验用于其目的是否会带来更为积极而非消极的结果；也就是说，使用该测验利大于弊的证据。我在下文描述这个四步过程。

3.1 步骤1：阐明测验目的

效度验证包括收集和分析为了某一特定目的而使用某一测验的正当性的证据。因此，效度验证从识别测验分数将被如何使用开始。AERA et al. (2014) 《标准》中描述道：“从逻辑上讲，效度验证的过程始于明晰测验分数将被如何解释以及说明该解释与测验用途之间的相关性” (p. 11)。在大多数情况下，这个初始步骤仅仅是重申一个特定测验的目的。事实上，《标准》要求测验机构明确地指出测验实施的目的。文中写道，“测验开发者应该清楚地说明其计划如何解释和使用测验分数” (p. 23)。《标准》还要求测验开发者清楚地描述与测验目的相关的构念 (p. 85)。定义被测构念以及测验目的奠定了效度验证的基础。

测验目的应该在技术报告中有清晰的定义 (AERA et al., 2014, p. 125)。然而，在许多情况下，测验的目的是复杂或不明确的。在这种情况下，测验目的则需要从测验机构的声明中推测出来。在明确了测验目的，包括测验分数的预期用途知后，研究人员可以提出用以评价该用途的效度研究。

3.2 步骤2：确定测试使用的潜在负面后果

尽管将效度验证的重点放在测验分数的预期用途上很重要，但考虑测验项目潜在的负面影响也十分关键 (Messick, 1989)。识别潜在负面影响的一种方法是

关注公众对测验项目的批评。例如，被用作高中毕业要求的成就测验常常被批评“窄化课程”以及给学生造成过大压力以至于辍学。另一种批评是“不利影响” (adverse impact)，即通过考试的考生的比率因性别和种族等人口因素的差异而不同，因此该影响也是一种需要考虑的潜在负面后果。潜在的负面后果代表了需要研究的假设，这些研究也应被包括在效度验证的框架中。

3.3 步骤3：将测验目的与潜在的误用与《标准》中效度证据的五个来源进行交叉比对

这一步涉及到明确纳入了 AERA et al. (2014) 《标准》的效度证据的五个来源——基于 (1) 测验内容、(2) 反应过程、(3) 内部结构、(4) 与其他变量的关系，以及 (5) 测验结果的效度证据。对每个证据来源的完整描述超出了本文的范围，因此笔者鼓励读者参考《标准》中更为完整的描述。这里只做简短描述。

基于测验内容的效度证据指的是评价某测验的内容充分反映了被测领域的程度，以及测验内容与测验目的一致性的研究 (Martone & Sireci, 2009; Sireci & Faulkner-Bond, 2014)。这类证据（比如, 资格考试中的工作分析）通常为内容领域专家或能定义该领域的人通过审查测验题目而做出的对题目与目标内容领域的关系的评判。

基于反应过程的效度证据指的是“关于构念和考生实际的表现或反应的细节之间是否匹配的证据”(AERA et al., 2014, p. 15)。这类证据包括对考生就试题反应进行访谈、对测验反应行为进行系统观察、对裁判在给表现任务打分时使用的标准的评估、对项目反应时间数据的分析，以及对考生在解决测验项目时使用的推理过程的评价(Embretson, 1983; Messick, 1989; Mislevy, 2009)。

基于内部结构的效度证据指对题目和子分数数据进行的统计分析，其目的为评价测验数据的维度是否与测验背后的理论以及用于评分的统计模型 (比如, 一个单维项目反应理论模型) 所假设的维度相一致。如 (探索性和验证性) 因素分析、多维量尺化或基于模型的残差分析等的统计方法可以用来评价假设的维度是否在考生对测验项目的反应中得到了体现。相关的统计分析还包括维度分析支持子分数的程度以及项目功能差异 (differential item functioning) 研究。

基于与其他变量关系的效度证据指传统意义上的效标相关的效度证据，如同时效度和预测效度研究，以及更为全面的对测验分数与其他变量之间的关系的研究，如多特质-多方法研究 (Campbell & Fiske, 1959)，和对不同学生群体之间的分数差异的研究，如学习不同课程的学生。这些外部变量可用于评价测验分数与其他学生成就指标之间的假设关系 (如考试分数和教师评分)，评价不同测验在多大程度上实际测量了不同的技能，以及测验分数在预测具体效标方面的效用。

最后，基于测验结果的效度证据指对与测验项目相关的预期和意外后果的研究。这类的例子之前在步骤2已经介绍过，更多的例子可以在 Messick (1989), Shepard (1993)，和 Lane (2014)中可以找到。从某种意义上说，所有的效度研究都可以被认为是对测验结果的评价，因为测验目的代表了预期的结果。然而，这类证据通常侧重于评价是否存在与测验相关的意外的负面影响。

为了举例说明步骤 3，表 2 展示了测验目的和潜在误用与《标准》中效度证据五个来源的交叉比对。这个例子来自马萨诸塞州成人能力水平测验 (Massachusetts Adult Proficiency Tests, MAPT)，这是马萨诸塞州成人教育学生的数学和阅读考试。它的技术手册中明确记录了考试目的：“MAPT 的目的是测量（成人教育学生）在数学和阅读方面的知识和技能，以便评价他们在达到教育目标方面的进展情况...[它]旨在为了政府监督和问责而对学习者的教育水平进行测量”(Zenisky et al., 2018, p. 10)。根据这个目的声明, 研究者应该提出几类关于将 MAPT 用于这些目的合理性的效度问题。这些问题是，

1. MAPT 是否真正测量了成人教育学生数学和阅读的知识和技能？
2. 它是否按照马萨诸塞州成人教育课程框架中的定义对这些知识和技能进行测量？
3. MAPT 的分数是否准确地提供了关于学生数学和阅读能力的信息？
4. MAPT 分数对于评价学生为达到教育目标的进步是否有用？
5. MAPT 分数是否适合用来评价由联邦政府所定义的学生发展？
6. 由多个部分组成的 MAPT 分数对于评价 ABE 项目的有效性是否有用？

此外，MAPT 技术手册 (Zenisky et al., 2018) 明确了两种潜在的错误使用，并提出警告：即将测验用于诊断目的以及将学生分配到不同的教学项目中。这两种测验使用又引出了额外的与负面后果有关的效度问题。即，

7. 教师是否不当地使用 MAPT 分数来诊断学生的优势和弱势？

8. 成人教育项目是否使用 MAPT 用于分班目的？

9. MAPT 对成人教育的教学有什么影响？

最后一个问题并非来自于任何明确的测验目的或对不当使用的警告，而是出于一个隐含的、更无私的目的——教育测验应与教学相结合并改进教学。这9个效度问题分别对应表 2 中的一行，而效度证据的五个来源则分别对应了表 2 中的一列。表 2 提供了一个框架，以将《标准》的效度证据的五个来源与从目的声明与警告中自动产生的效度问题联系在一起。表中打勾的地方 (√) 表明该处需要相应的证据来回答效度问题。每个单元格中的勾代表的具体研究这里没有描述 (见 Sireci, 2012 的例子)，但大多可以很容易推断出来。例如，表格中的第一个单元格有关的研究的一个例子就是校准研究。

3.4 步骤4：优先考虑要进行的效度研究

表 2 中的单元展示了一系列研究，代表了一个全面的验证效度的研究清单。然而，由于时间和资源的有限性，进行这样一系列全面的研究通常是不可能的，因此需要设置一些优先次序。

需要注意的是，从 MAPT 的目的声明中收集到的所有效度问题都以某种形式在表 2 中得到了解决，并且都涉及至少一个效度证据来源。确定效度问题的优先次序必须考虑 MAPT 最重要的目的，以及它产生的主要原因——满足联邦问责规定的要求。根据联邦规定，马萨诸塞州必须有一个与其课程框架挂钩的评定，以及必须用该评定根据NRS的成就水平对学生的教育增益进行评估。这种评定必须是准确的，同时根据定义，该评定将被用于项目评估。因此，被优先考虑的效度研究必须提供能证明 MAPT 正在实现其预期目的所需的最少量的证据。表 2 中的星号 (*) 表示为支持一个充分的效度论证而被优先考虑的研究。从星号可以看出，着重点为确保评定内容的适当性以及满足联邦要求的效度研究是被优先考虑的。

3.5 总结四步效度验证过程

本文提出的指导效度验证过程的四步过程，就像基于论证的方法一样，体现了一种折衷：即穷尽所有理论上可以用来评估某一特定目的的测试使用的效度研究，与为证明某一特定目的测验使用的合理性所需的最低限度二者之间的折衷。这种方法有一定的局限性。其一，它要求负责的测验开发者和评估者清楚地阐明测试目的和预期用途，识别潜在的误用，并进行质量研究以提供所需信息。其二，它需要设定研究的优先次序，但可能很难就优先次序达成一致。然而，该过程的一个好处是它提供了一种基于 AERA et al. (2014) 的《标准》的标准化方法来处理效度验证。而《标准》是以近七十年的学术合作成就为基础的。它还强调测验的使用和其使用的效果。因此，它符合AERA等对效度的定义以及它的建议——将效度验证看作证明为某一特定目的的测验使用的合理性的努力。

4 讨论

50 多年来，构念的概念一直是教育和心理测量的核心。因此，许多测验专家接受构念效度理论作为对效度在哲学上最正确的描述也就不足为奇了。在这篇文章中，我认为我们可以使用构念效度理论的许多原则来设计和进行效度验证工作，而不是在术语中裹足不前。通过解构效度验证，我们使得这个工作变得不那么哲学化，而是更可实践。我们的重点不在于构念本身，而在于测验目的和测验分数的具体使用。体现该重点的一种方法是本文中提出的四步效度验证过程。

这种效度验证方法的第一步是明确说明测验目的。明确的目的说明为效度验证工作奠定了基础，因为其中暗含了需要解决的关键效度问题。随后应对这些问题进行优先排序，以确定效度验证安排和时间表。在排序的过程中还应该考虑测验项目会收到的批评意见，以及收集效度证据时受到的限制。有些限制可能是由于财政和人力资源造成的，而另一些则可能是由于需要等待一段时间才能获得足够的测验成绩数据，或直到测验的影响显现出来。这些限制不应该成为过早停止效度验证工作的借口。相反，在关于如何以可用的资源充分系统地回答重要效度问题上，它们应该成为讨论的一部分。最后，随着测验项目逐渐成熟，或随着新的重要的研究方向的出现，效度验证相关的方案应被定期更新。

在制定效度验证方案时，我们可以从效度文献中得到有益的建议。我最喜欢的效度引言之一不出意外出自 Messick (1989)，他说，

> 测试是构念的不完美度量，因为它们要么遗漏了根据构念理论应该包含的东西，要么包含了应该忽略的东西，或者两者兼而有之 (p. 34)。

这句引言对于效度验证方案很重要，因为如果在效度验证中尽量确保测验没有遗漏任何东西 (比如, 充分代表了内容域)，并且不包含任何种类的偏差，就说明我们在支持测验用于特定目的方面取得很大进展。

当一个效度研究项目收集了充足的证据来支持测验的某一特定用途时，我们还有一个问题需要解决。

表 2. 将测验目的与潜在的误用与《标准》中效度证据的五个来源进行交叉比对的示例

效度问题	效度证据的来源				
	内容	内部结构	与外部变量的关系	作答过程	测验后果
测量正确的技能？*	√*	√*	√*	√	
和框架一致？	√*				
准确？		√*	√		
测量发展？	√*	√*	√*		
符合联邦要求？	√*	√			
对项目评价有用？	√	√			√
不合适的诊断用途？					√
不合适的分班？					√
对教学的影响？	√				√

*优先的研究.

在考虑要做多少效度验证才能让我们确信测验分数能实现其目的时，我之前建议使用法庭类比。在 Sireci (2009) 中，我提出了以下建议：

> 指导测验效度验证工作的最佳问题也许是："如果我为达成一个目的所使用的测验在法庭上受到挑战，我是否有足够的证据来说服法官或陪审团并赢得官司？"如果答案是肯定的，那么该证据将构成一个可靠的效度论证...如果没有，就需要更多的证据，否则测验的使用无法得到维护" (p. 31)。

5 结论

有关测验测量了设想的构念的证据在效度验证中的重要性毋庸置疑。不过，该证据可由 AERA et al. (2014) 的《标准》颁布的五个效度证据来源中的至少四个提出和整理。在这篇文章中，我提出，我们可以在不使用"构念效度"这个术语的情况下对效度进行正确和充分的调查。从科学哲学的角度看，构念效度性理论是优雅的，但即使是 Messick (1989) 也承认，"科学哲学更多地是哲学而不是科学" (p. 21)。希望本文中提出的四步效度验证过程所提供的实践指导纲领，能够促进充足的效度论证的开发，支持为特定目的服务的测验使用，促成对社会有益的测验项目，并将负面影响降至最低。

参考文献

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt.2), 1–38.

American Psychological Association. Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist*, *7*, 461–475.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice*, *22*(3), 5–11.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, New Jersey: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, *16*, 640–647.

Ebel, R. L. (1977). Comments on some problems of employment testing. *Personnel Psychology*, *30*, 55–63.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Guion, R. M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, *1*, 1–10.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*, 385–398.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Washington, DC: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*, *26*, 127–135. doi: 10.7334/psicothema2013.258

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694 (Monograph Supplement 9).

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research 4*, 1332–1361.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, New Jersey: Lawrence Erlbaum.

Messick, S. (1989). Valifity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, D.C.: American Council on Education.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing Inc.

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*, 301–319.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405–450.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 19–37). Charlotte, NC: Information Age Publishing Inc.

Sireci, S. G. (2012). *Smarter balanced assessment consortium: Comprehensive research agenda.* Available at http://www.smarterbalanced.org/assessments/development/additional-technical-documentation/.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement*, *50*, 99–104.

Sireci, S. G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*, 100–107. doi: 10.7334/psicothema2013.256

Thorndike, E. L. (1904). *An introduction to the theory*

of mental and social measurements. New York, NY: Teachers College Press.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher*, *12*, 10–14.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O' Donnell, F., Wells, C. S., ... Garcia, A. (2018). Massachusetts adult proficiency tests for college and career readiness: Technical manual. *Center for Educational Assessment research report No. 974*. Amherst, MA: Center for Educational Assessment.