目　录

第一章　效度验证模式[1]

罗凯洲　北京外国语大学

1.1 引言

效度[2]是教育与心理测量领域最引人注目的议题，其重要性在语言测试学科中也不例外。不同学者对效度的定义不尽相同，所采用的效度验证模式也各有差异，但人们普遍认可效度是一种表示程度的属性（AERA et al.，1985，1999，2014；Anastasi，1986；Cronbach，1971；Cureton，1951；Kane，2006，2016；Messick，1989；Newton，2012）。问题在于效度究竟是谁的程度属性？怎样才能确定程度的高低？其实，第一个问题触及效度的本质，第二个问题牵涉到效度如何验证。

效度真正引起教育与心理测量领域关注是在20世纪20年代后，从业人员秉持了一种朴素观念，笃信相关即有效（孙晓敏、张厚粲，2004）。50年代中期后的20年逐步形成标准效度（criterion validity）、内容效度（content validity）与构念效度（construct validity）三足鼎立的局面。70年代中后期，整体效度思想初现端倪。Messick（1980，1981，1988，1989）明确主张用构念效度指代效度，原先的效度类型均被“降级”为效度验证时的证据，整体效度思想一直影响至今。然而，无论效度理论（概念）如何演变，只有被转化为有实施步骤的效度验证模式后才算有真正价值（O’Sullivan & Weir，2011：26）。

本章称20世纪20年代至70年代初期为“分类效度观”时期，70年代中

1　本章为国家社科基金一般项目“复合型国际化高端外语人才沟通能力测评研究”（项目批准号：20BYY110）的阶段性成果。

2　限于篇幅，本章重点讨论效度（validity）及其验证模式（validation approach/model），未过多涉及效度验证的具体方法（validation method）（即获取效度证据的量化或质性方法）。遵循教育与心理测量领域的表达习惯，本章没有刻意区分approach、model与framework等词的用法，均指效度验证的模式。

后期至今为“整体效度观”时期。[1] 两大时期对效度本质的探讨和效度实践（效度验证）的摸索不尽相同。分类效度观将效度当作测试自身质量的程度属性，效度验证好似证据罗列；而整体效度观则把效度当作测试成绩解释与使用合理性的程度属性，效度验证倾向推理论证（罗凯洲，2019）。

总体上看，语言测试学科的效度研究追随了上述历程。然而，新概念和新模式应用到某一个新学科终究需要时间，语言测试学科对待效度及其验证模式就存在“迟滞”，甚至“误解”。弄清教育与心理测量领域的效度观与验证模式的发展，才能理顺语言测试效度研究。本章将在第二小节探讨分类效度观时期效度及其验证模式的演变；在第三小节讨论整体效度观时期的效度概念，以及在语言测试学科常见的四种效度验证模式；最后，讨论效度及其验证模式的困惑以及未来的发展趋势。

1.2 历史观点

1.2.1 分类效度观及其验证模式

20 世纪 20 至 50 年代初期，教育与心理测量界普遍认为效度是一项测试原本要测量东西的程度（Bingham，1937：3；Kelly，1927：3；Ruch，1924：13）。效度验证的主要模式是对测试与其他同类“标准”之间的相关性进行统计分析。Cureton（1951：623）在《教育测量》[2] 第一版中直接把效度定义为“测试实得分数”与“真实标准分数”之间的相关程度。按此逻辑，对一项测试进

1　也有学者（如 Brennan，2006：2；Chapelle，2013；Chapelle et al.，2010；Newton & Shaw，2014：16；韩宝成、罗凯洲，2013；李清华，2006）对效度及其验证模式的发展历程做了更详细的划分或采用不同的命名方式。例如，把效度发展初期称为单一效度观时期，把 Kane 提出的效度理论及框架称为“效度论证观”或“新理论”等。本章认为上述划分与命名略显繁复，依照对效度本质的看法（到底是谁的程度属性），本章只分为两个时期，即“分类效度观”时期和“整体效度观”时期。

2　教育与心理测量领域公认对效度、验证模式以及证据来源进行权威阐释的两本文献是《教育测量》（*Educational Measurement*）与《教育与心理测量标准》（*Standards for Educational and Psychological Testing*）。《教育测量》已经出版四版（1951，1971，1989，2006），《教育与心理测量标准》已经出版六版（1954，1966，1974，1985，1999，2014）。

行效度验证的主要方法就是计算实得分数与标准分数之间的相关系数。所以，这个相关系数（也称为效度系数）也由此成为最重要的效度证据，结果越接近于 1，效度越高。Kane（2006：18）把此类效度验证模式称为“标准模式”（criterion model）。所谓“标准”，是用作参照的“可靠”依据，可以是专家对考生能力的排序，也可以是考生实际表现，但更重要的是考生参加被认可的同类测试后的数值结果。名为“标准”，实为“数值”。效度系数看似客观，但用于计算效度系数时对标准进行的选择往往又偏向主观。所以，“标准模式”的难点在于如何确定标准自身的质量，通常只适用于有现成标准可以参照的情况。

编制学业成就测试（achievement test）时所参照的标准往往是教学大纲，这类测试很难用“标准模式”验证效度。因此，有些教育测量界的学者倾向于采用逻辑分析的方法来确定成就测试的效度。所谓逻辑分析，就是请相关专家对测试内容与教学大纲内容之间的匹配程度进行判断，相似程度越高，效度越高。Kane（2006：19）把此类效度验证模式称为“内容模式”（content model）。“内容模式”的最大问题在于过度依赖主观判断。如果这些判断都是测试开发方主导，那很有可能得出对开发方有利的结果。此外，无论“标准模式”还是“内容模式”都无助于（基于理论的）人格测试（personality test）的效度验证。

20 世纪 50 年代后期至 70 年代初，人们对效度本质的看法并未发生根本性转变，始终认为效度是测试质量的程度属性。然而，在这一时期，效度验证模式却倒逼了效度类型的划分。1954 年美国出版的《关于心理测量和诊断技术的建议》（也就是《教育与心理测量标准》第一版，下文简称《标准》），直接将效度划分为预测效度（predictive validity）、共时效度（concurrent validity）、[1] 内容效度（content validity）。此外，这份重要文献还引入了一种新的效度类型，即构念效度（construct validity）。在这一时期，构念 [2] 指人通过测试表现出来的某种假定存在的特质，反映的是依照心理学理论（或假说）构建出的概念，无法直接观察或测量（Cronbach & Meehl，1955：283）。1966 年第二版《标准》

1　所谓预测效度是指一次测试结果与后来的“标准”测试结果（如真实的工作表现）进行的“比较”，由此来判定测试的预测力，而共时效度是指在同一时间段完成的两次测试的“比较”，由此判定测试在当下的质量。预测与共时效度其实是标准效度（criterion validity）的两种不同形式。

2　在整体效度观下，绝大多数学者或机构（如 AERA et al.，2014；Bachman，1990；Bachman & Palmer，1996，2010；Kane，2016；Messick，1989；Sireci & Sukin，2013）倾向于用 construct 指代任何被测量的（理论）概念或（可观测）特质（trait/attribute）。

把预测效度和共时效度合并，统称为标准相关效度（criterion-related validity）（也称标准效度）。标准效度主要为以能力倾向测试（aptitude test）为代表的选拔类测试服务，内容效度主要为学业成就测试服务，而构念效度则主要为以人格测试为代表的基于理论的测试服务，效度三分模式（trinitarian model of validity）从此确立（Guion，1980）。1966 年版《标准》还提到，当一项测试找不到合适标准时，再考虑采用构念效度（APA et al.，1966：13）。引入构念效度的初衷是为了解决基于理论编制的测试效度验证问题，对构念效度的验证主要借助标准和内容效度验证时所采用的方法与证据。所以，构念效度在 20 世纪 60 年代中期前仍处于边缘地位。

构念效度验证好似自然科学的理论（假说）检验；与自然科学不同的是，它主要依靠统计结果来检验测试成绩与所测构念之间的关系。如果结果支持，则说明测试成绩有解释力，从另一个角度来看，测量工具的构念效度也得到证实（Cronbach & Meehl，1955）。Kane（2006：19）把此类验证模式称为构念模式（construct model）。这一时期以经典测量理论为代表的各类心理统计方法也得到了极大发展，为构念效度验证提供了方法论支持。例如，因子分析方法（factor analysis）、多质多法（multitrait-multimethod）都在这一时期逐渐成熟，一度成为构念效度验证的专属方法。

1.2.2 Lado 等人的效度观及其验证模式

一般认为，1961 年是现代语言测试的开山之年。Lado 在 1961 年出版《语言测试》（*Language Testing*）被视为标志性事件之一，该书也是第一本系统论述语言测试的著作（Kunnan，2014：xiii）。Lado 从教育与心理测量领域引入了“效度”（如内容效度、标准效度）、“信度”等有关测试（评分）质量的概念。他认为：“效度本质上是一种关联（relevance）。一项测试测量到了它原本要测的东西了吗？如果答案是肯定的，那么它就是有效的”（Lado，1961：321）。他在书中还介绍了效度验证的方法。例如，建议考生参加新开发的测试和可作为标准的已有测试，根据两次测试的成绩计算相关系数来确立新测试的效度（Lado，1961：30）。Lado 对语言测试效度及其验证方式的阐述同教育与心理测量领域的观点一脉相承，对后人产生了深远影响。20 世纪六七十年代语

言测试著作纷纷效仿 Lado 的观点来阐释效度概念（如 Harris，1969；Heaton，1975；Valette，1967）。80 年代甚至 90 年代出版的语言测试教材和专著以及有关效度验证的研究，大都采用了分类效度观（如 Henning，1987；Hughes，1989；Wood，1993），这些著作几乎无一例外地把效度与信度看作测试质量属性的代名词。

1.3 当今视角

1.3.1 Messick 的整体效度观

20 世纪 70 年代后，不少学者（如 Guion，1977；Messick，1975；Tenopyr，1977）都提议采用整体视角看待效度概念，但效度验证的方式可以多样，仍可为不同测试类型服务。这一时期，构念效度验证方法愈发灵活，效度证据来源愈发多元，都使得构念效度地位不断攀升。Cronbach（1971）曾强调构念这一概念对各类教育或心理测量具有普遍意义，为效度从分类走向整合埋下了伏笔。十余年后，Cronbach（1984：126）更是明确提出“所有的效度验证其实都是构念效度验证”，这一理念在 1985 年版《标准》中也有所体现。构念一词的含义也悄然发生着变化，由原先专指依理论构建的概念，发展到可以指代任何被测量的东西（AERA et al.，1985，1999）。20 世纪 80 年代，构念效度已逐步统领了其他效度（标准效度、内容效度），甚至囊括信度概念。从某种意义上说，正是构念效度验证模式（包括各类验证方法与证据）促使构念效度成为整体效度概念的代名词（Kane，2006：21）。

Messick（1988，1989）正式提出了整体效度观（unified/unitarian view）。这一思想在他为第三版《教育测量》撰写的“效度”一章中得到了充分阐释：“效度是一个整体概念，是实证证据（empirical evidence）和理论依据（theoretical rationale）对测试成绩[1]解释与使用合理性的支持程度”（Messick，1989：13）。简言之，上述定义蕴含着对效度本质看法的转变；虽然效度仍是一个程度概

1　Messick 话语体系下的“成绩”是一个较笼统的概念，可以指用数值表示的分数或等级，也可以指其他能被观测并记录的测试行为表现（Messick，1989：13）。

念，但已不再是测试质量的程度，而是有关成绩解释与使用的充分性与合理性程度。所谓对成绩的解释与使用实际上都是人们做出的推断（inference），推断好似假设（hypothesis），因此需要检验。对推断的效度验证过程实际上就是对假设的检验过程，这一观点对后来出现的基于论证的效度验证模式产生了深远影响。Messick（1989：20）随后用多面效度模式（facets of validity）来阐述“一元多维”的整体效度思想（见表 1.1）。

表 1.1 多面效度模式（基于 Messick，1980，1989，1995）

	测试（成绩）解释	测试（成绩）使用
（解释与使用的）证据基础	构念效度	构念效度 + （成绩的）相关性 / 有用性
（解释与使用的）后果基础	构念效度 + 价值内涵	构念效度 + 价值内涵 + （成绩的）相关性 / 有用性 + 社会后果

多面效度模式好似一个“渐进矩阵”（progressive matrix），两个纵向维度分别是测试（成绩）解释和使用，各自包含不同要素，可看作成绩解释和使用的“操作定义”。评价这些要素的程度则需要不同类型的证据支撑，由横向维度表示，分为传统意义上的证据类型（如与内容和标准有关的证据类型）以及 Messick 个人特别强调的体现价值观和成绩使用后果的证据类型。所以，渐进矩阵从左上构念效度出发，向右下的社会后果演进，每次演进增加一个或多个新的要素。

整体效度观下的多面效度模式有如下特点。首先，构念效度永远处于核心地位，但其含义不同于分类效度观下的构念效度概念。传统的标准效度、内容效度、构念效度均被统一到构念效度概念之下，分类效度的各种验证方法和证据类别从此也只为构念效度验证服务。其次，一元多维的整体效度观，特别强调了测试成绩解释的价值内涵（value implications）以及成绩使用所带来的社会后果（social consequences），使测试开发者和使用者都肩负举证责任。由于强调测试成绩使用后果（也称影响力），语言测试学科自 20 世纪 90 年代中期到 21 世纪初炒红了“反拨作用”研究。

然而，也正是这些特点导致效度验证的实践中出现了新问题。首先，既然效度是一个整体，又何必单独强调构念效度的核心地位呢？在渐进矩阵中