第1章
绪　论

“效度”这一概念最早于19世纪20年代由测量学家引入教育测量和心理测量之中（Newton & Shaw，2014），随后在测量界逐渐传开并得以接受，其定义也随着效度理论的发展而发生变化。关于效度，最早但如今仍普遍使用的一个定义是一种测试[1]在多大程度上真正测试到了其宣称要测试的内容（Garrett，1937），强调除了关注一种测试有什么作用，还要注意测了什么（Sireci & Sukin，2013）。但真正得到广泛认可的定义源自由美国教育协会、美国心理协会和美国国家教育评价委员会这三个协会制定的《教育与心理测试标准》（*Standards for Educational and Psychological Testing*）（以下简称《标准》）。《标准》起初来源于由美国心理协会于1954年发布的《心理测试和诊断技术建议》（*Technical Recommendations for Psychological Tests and Diagnostic Techniques*），1966年由前面提到的三个协会正式以《标准》的名称发布，之后分别于1974年、1985年、1999年对《标准》进行过多次修订。最新的《标准》于2014年对外发布。目前普遍接受的效度定义是1999年版《标准》中的界定，即“所收集到的证据和理论依据在多大程度上支撑对测试分数使用所作出的解释”（American Educational Research Association [AERA] et al.，1999：9，2014：11）。

人们谈及效度时一般会联想到两个方面的问题：第一，如何定义效度，包括效度的概念是什么，谁会使用效度，效度指什么，效度的界限在哪里，效度与其他相关概念有什么关系等。第二，如何体现效度，包括有哪些相关证据，需要多少证据，如何归类证据等。（Newton，2012）

1　测试（testing）和考试（test）在语言测试中常不加以区分，本书统一使用测试。

1.1 效度的定义

1.1.1 效度定义的发展

Newton（2012）从四个方面来谈论效度的定义：第一，把效度看作测试的特征不一定明智；第二，把效度当作解释的一个特征是个好的做法；第三，把效度看作一个整体概念是正确的；第四，把构念效度看作所有效度的整体也是不错的。他还从历史的角度阐释了这四个方面。

第一方面，测试是有效的（或无效的）。他指出，“效度”这一概念何时进入教育与心理测评界不得而知，但在 20 世纪 20 年代的有关文章中，“效度”这一概念经常以不同的形式出现在一些名家的文章和著作中（Newton，2012）。早在 1921 年，美国教育研究国家标准委员会（Standardization Committee of the National Association of Directors of Educational Research）就提出，测试测了什么是一个效度问题，而每次测量是否一致是个信度问题（Buckingham et al.，1921）。到了 1924 年，人们普遍接受的“经典定义”得到广泛传播，即效度就是测试在多大程度上测量到了要测量的（Ruch，1924）。如果测试测量到了要测量的，那就意味着测试具有效度。当然，这个简单的“经典定义”很快就受到了挑战，因为影响测试效度的因素太多——测试的程序、考务的安排、试卷的性质、试卷的质量、考生的特点、测试结果的使用等诸多方面都会对测试的效度产生不同的影响。所以，简单地从测试是否有效来定义效度显然是不够的。

第二方面，解释是有效的（或无效的）。接下来提出的新效度定义是几十年后的事了。Cronbach & Meehl（1955：297）就考虑到关于测试的解释问题，他们指出，“只询问‘测试有效吗？’很不成熟。关于测试人们可以做不同的推断，有些推断是有效的，有些则是无效的”，测评的效度就是对测试结果解释的验证，“人们验证的不是测试本身，而是对从中产生的数据的解释”（Cronbach，1971：447）。

第三方面，效度是个整体概念。人们普遍意识到效度根据测试目的不同而不同，同一个测试只要用于不同的目的就要单独验证其效度。《标准》第一版区分了四种不同的测试目的，因此也就区分了四种

不同种类的效度：内容效度（content validity）、预测效度（predictive validity）、同期效度（concurrent validity）和构念效度（construct validity）（American Psychological Association [APA] et al.，1954）。后续两版的《标准》把效度类型减少到三类：内容效度、效标关联效度（criterion-related validity）和构念效度，即后来的三一概念观（trinitarian conception）（AERA，1966，1974）。当然，到了后来的版本，效度被认为是个整体概念，整体概念观（unitarian conception）的效度开始流行（AERA et al.，1999，2014）。

第四方面，构念效度是效度的全部。Loevinger（1957：636）针对传统的效度定义指出，“预测效度、同期效度和内容效度基本都是特意设计的，从科学角度上说，构念效度才是效度的全部。”这个观点后来得到了 Messick（1989）的认可，也体现在当时出版的《标准》中（AERA et al.，1999）。《标准》首先指出，效度就是使用测试分数时对分数的解释，效度与分数的使用联系在一起。表 1–1 展示了《标准》各个版本对效度的分类。

表 1–1 《标准》各个版本对效度的分类

版本	效度分类
1952 年《心理测试和诊断技术建议》第一版	预测效度、身份效度（status validity）、内容效度、相容效度（congruent validity）
1954 年《心理测试和诊断技术建议》第二版	构念效度、同期效度、预测效度、内容效度
1966 年《标准》第一版	效标关联效度、构念相关效度、内容相关效度
1974 年《标准》第二版	效标关联效度、构念相关效度、内容相关效度
1985 年《标准》第三版	效标关联效度、构念相关效度、内容相关效度
1999 年《标准》第四版	证据类型：内容、反应过程、内部结构、与其他变量的关系、测试的后效
2014 年《标准》第五版	证据类型：内容、反应过程、内部结构、与其他变量的关系、测试的后效

总之，最早的效度理论学者从测试与标准之间的相关关系来定义效度。他们认为，相关关系可以显示测试是否测试到要测试的。但随着人们对测试后效越来越关心，仅考虑相关关系无法满足人们对测试效度的

要求。后来，人们更趋向于收集与测试效度有关的证据，包括理论的一致性和目的，对个人、群体甚至整个社会的积极的后效等（Sireci & Sukin，2013）。最新的一版《标准》提出五方面的相关证据：测试内容、与其他变量的关系、内部结构、反应过程和测试的后效（AERA et al.，2014）。

1.1.2 有关效度定义的争议

一直以来，效度是语言测试研究的一大热点。语言测试界的效度概念来源于心理测量学，最早由 Lado 于 1961 年引入（D'Este，2012；Lado，1961）。虽然《标准》对“效度”有明确的定义，也得到普遍的认可，但有关“效度”的定义仍存在着不同的诠释，且种类繁多，如测量工具的效度、测试分数的效度、测试分数解释的效度、决策过程的效度等（Newton，2012；Newton & Shaw，2014）。Newton & Shaw（2014）认为有必要就测量这一根本术语的用法达成一致意见，并在 2014 年美国国家教育测量委员会（NCME）年会上就“如何最佳使用效度一词”这个话题组织了一次协调会议，试图解决这些争议。效度专家踊跃撰文，这些文章集中发表在 2016 年《教育测量：原则、政策及实践》（*Assessment in Education: Principles, Policy & Practice*）第 2 期上。根据他们的效度观点，Newton & Shaw（2016b）把这些专家划分为四个不同的阵营：新自由主义者、传统主义者、新保守主义者和超级保守主义者（刘建达、贺满足，2020）。

新自由主义者主张大范围拓展效度概念，强调测试后果在效度和效度验证中的重要性，既关注预期测试使用，也关心非预期使用（Gafni，2016；Moss，2016；Sireci，2016b）以及非预期的测试后果（Haertel，2013；Kane，2016a）。Hubley & Zumbo（2013）和 Zumbo & Hubley（2016）甚至提出效度应包括社会及个人层面的后果和消极影响（为了与积极的后果进行区分，特意使用不同的词汇表示）。传统主义者则认为效度包括测量（测试分数解释）及预测（测试使用），需要效度验证的主要是测试分数的使用（Shepard，2016；Sireci，2016a，2016b）；

定义效度时忽略测试使用等同于定义“无用”测试的效度（Sireci，2016b）。新保守主义者强调效度的测量属性，建议回归经典定义，即效度指的是“测试在多大程度上测量了它想要测量的内容”。他们认为测试分数解释和使用是两个不相容的问题（Cizek，2012，2016a，2016b；Geisinger，2016），且效度包含信度、单维性、偏颇性（Cizek，2016a）。超级保守主义者以 Borsboom 及其同事为代表，提倡回归经典定义（Borsboom & Markus，2013；Borsboom & Wijsen，2016）。但他们的效度观更狭隘：效度是测量最基本的属性，如果测量的特质引起了测量结果的变化，那么这个测量在测量它想要测量的内容方面就是有效的。这种效度观范围狭小，不包括其他评价性概念，诸如信度 / 准确性、维度和偏颇性。

效度具体应涉及哪些方面也是一个备受争议的问题。Borsboom & Wisjen（2016）认为应弄清效度是一个本体论问题、认识论问题还是伦理问题，不能将这三个问题混为一谈。Koretz（2016）认为不应把伦理问题纳入效度定义，而 Cizek（2016a）也认为这是一个概念上的争论，因此从逻辑上来讲不应把伦理方面的考量牵扯进来。但 Kane（2016a）却认为效度不应和伦理分开。Kane（2016b）认为关于效度的争议更多的是词汇层面的，即测试专家们对效度一词的多种使用使得其确切含义变得模糊。Newton & Shaw（2016a）则指出效度定义问题与其说是逻辑问题，还不如说是权衡不同视角的利弊问题。Twing（2016）认为效度最根本的问题不在于定义，而更多的是过程或程序问题，即如何收集测试分数解释及使用的各种证据。Sireci（2016a）认为当今效度面临的真正问题是如何提供连贯的信息以帮助测试使用者根据测试分数进行决策。

在有关效度定义的讨论中争论最大的仍是后果效度问题（Brennan，2006a）。有关后果效度在效度理论中的地位，Kane（2016a）提出三个可能的解释模型：唯解释模型（an interpretation-only model）、后果作为指标的模型（a consequences-as-indicators model）、解释及使用模型（an interpretation-and-use model）。在唯解释论的框架中（类似于传统主义者），效度验证只是分数解释。效度验证本身需进行补充，进行另外的分析以证明预期的使用有意义。但是在语言测试和教育与心理测量

中，意义和使用是密不可分的，测试开发之初就需明确分数的使用或用途（Sireci，2016b；Zumbo & Hubley，2016）。在后果作为指标的模型中，效度验证包括评价测试项目达到预期目标的程度以及潜在的非预期后果（包括积极和消极的后果）。这种模型主要集中在预期分数解释的可信度上，而消极后果本身并不会降低测试项目的效度。相反，消极后果可作为指标表明测试项目中可能存在的问题。解释和使用模型对分数解释和分数使用进行了区分，并对预期的分数解释和使用在多大程度上包含实际的分数解释和使用进行评价[1]。如果测试开发人员或用户声明该测试用于某些目的，通常需要评价分数使用的后果来决定主张是否得到支持（刘建达、贺满足，2020）。

当前效度讨论更多地围绕基于分数使用的评价是否应置于效度范畴中（Kane，2016a）。分数使用必然会产生一定的后果，Kane（2013a）认为测试项目评价涉及的后果主要有三种：（总体）预期后果、公平性（fairness）/ 偏颇性（bias）、系统性影响（而非个人影响）。Bachman & Palmer（2010）认为测试的使用及预期后果应是测试开发和评价的出发点。他们提出“评价使用论证”（assessment use argument，简称 AUA）框架并详述了推理过程，其中包括评价使用和对利益攸关者有益的决策所产生的后果。在该框架中，Bachman 和 Palmer 用 AUA 代替效度验证，以此强调对分数使用而非分数意义的合理性进行验证的必要性。对非预期的后果（特别是社会后果）的评价，各位专家观点不一，但测试的区别性影响（如不同组别考生中及格率或录取率不同）和其他非预期后果的评价大多在效度的框架下进行（Xi，2010）。

总之，虽然部分研究者认为社会后果不属于效度和效度验证范畴（Brennan，2006a），但是大多数专家都认同后果在效度理论中的重要性（Bachman & Palmer，2010；Chalhoub-Deville，2016；Chapelle，2012；Kane，2016a；Tsagari & Cheng，2017；Zumbo & Hubley，2016），并且认为如果将道德评价排除在效度之外，将存在无人对测试后果负责的风险。然而，一些专家支持狭隘的效度观，认为测试后果的评价虽是测试效度验证的重要部分，但不属于效度本身，因而不能将其与推论的质

1 “评价”和“评估”（assessment）在本书中不作区别，“测评”则指“测试”与“评价”（testing and assessment）。

量混为一谈（Koretz，2016）。这些专家建议在其他范畴内对测试后果进行研究，如“使用”（utilities）（Geisinger，2016）、“合理性”（Cizek，2012，2016a，2016b）、“政治”（Borsboom & Wijsen，2016）等方面。他们坚持认为把道德评价纳入效度范围内将使效度概念变得过于复杂而无法理解，并且对效度验证不具有任何有用的指导意义。

随着研究的不断深入，理论专家对测试后果的认识逐渐深化，意识到研究测试后果的重要性和必要性，争论的焦点也从效度是否包含测试后果转移到在哪个范畴内研究测试后果（Kane，2011，2013a）及后果的性质和范畴（Kane，2016a）。Kane（2011）指出，有关测试社会后果研究的三个真正问题是：应关注哪些社会后果？应如何评判这些社会后果？谁负责评判这些社会后果？

虽然语言测试界就效度定义并未完全达成共识，但作为教育与心理测量领域的纲领性文件之一，《标准》的效度定义被称为公认定义（AERA et al.，1999，2014）。语言测试界对效度的界定也主要是依据这个定义进行的。

效度是个概念，如前所述，指测试在多大程度上完成了既定的目标以及测试结果在多大程度上得到恰当的解释。与效度概念不同，效度验证则指为了评价测试结果是否得到恰当使用而收集和报告相关证据的过程（Sireci & Sukin，2013）。其实，效度与效度验证是一个硬币的两面，效度验证是对测试效度的调查，所以效度是要调查的对象；效度验证则是调查效度的过程（Newton & Shaw，2014）。《标准》提出的五类证据也是效度验证中需要重点去收集的。在测试结果的使用过程中，测试效度的各种证据的重要性会有所不同，其中有些证据可能会比其他证据更为重要。例如在课程终结性评价中，内容效度会较为重要。不过仅一类效度往往说服力不够，需要多种证据来支撑测试和评价的效度。在效度证据的收集以及研究过程中，理论学者也提出了不同的效度理论。效度理论为效度验证提供了概念框架。

1.2 20 世纪语言测评效度理论的主要发展

如上所述，效度的定义发展过程大致可分为三个主要阶段：前三一概念观（pre-trinitarian）、三一概念观（trinitarian）、整体概念观（unitarian）。与之相关联的效度理论发展大致可以分成五个主要阶段：（1）孕育期（gestation），即 19 世纪中叶至 1920 年；（2）形成期（crystallization），即 1921 年至 1951 年；（3）碎片化期（fragmentation），即 1952 年至 1974 年；（4）（重新）一体化期（reunification），即 1975 年至 1999 年；（5）解构期（deconstruction），即 2000 年至今。五个阶段没有明显的分界线，但反映出效度理论的发展过程。其实，几个转折和过渡点都与《标准》的各个版本有关（Newton & Shaw，2014）。

标准化测试最早可以追溯到中国的汉文帝前元 15 年（公元前 165 年），汉文帝为参加“贤良方正科”的士子出题测试（杨学为，1991）。现代的测试一般都会联想到 Binet 的工作，他致力于使巴黎学区的每个孩子不经过正式的测评就可以享受到教育（Binet & Simon，1905；Sireci & Sukin，2013）。之后随着测试的流行，测试效度的概念越来越受到重视，相应的效度理论也随之而生。

20 世纪的测试效度理论最早起源于 19 世纪（19 世纪中叶至 1920 年）。当时，欧洲和北美的学校主要采取结构性的测评作为进一步决策的基础。例如，美国采用了笔试，英国则为学校开发了地方性测试。到了 19 世纪末，人们越来越多使用结构性测试的结果来满足不同的选拔目的，例如公务员招聘和学校教学质量检测和追责等（Newton & Shaw，2014）。然而，随着测试结果的广泛使用，人们逐渐担心测试的效度和信度，一些结构性较强、主观性较弱的测试应运而生；主要的测试题型包括正误判断、多项选择、句子填充等，标准化测试也随之诞生。在那之后，人们进一步关注测试的公平性、测试技术的有效性和测试的效率。当时统计学的发展加速了这方面的发展和需求，人们开始使用各种统计方法来验证测试的信效度，并把测量学的概念引入语言测试。利用统计方法验证测试质量在 20 世纪初得到更广泛的应用，这也是语言测试效度理论验证的孕育期。

1921 年北美国家教育研究协会（The North American National

Association of Directors of Educational Research）发布了旨在为测量行为提供统一标准的文本，其中效度被定义为测试测量到了要测量的构念的程度，这为之后的效度理论发展奠定了基础。然而，该定义也带来了一个问题，即效度的程度如何进行验证，需要什么样的证据来证明测试的效度。早期最常用的方法包括两种：对测试内容进行逻辑分析以及对测试的实证数据进行相关分析（Newton & Shaw，2014）。对测试和要测量的能力之间进行相关分析以验证测试的效度成为一种被高度认可的做法。但是，要计算相关度，那测试结果应与哪些代表测试要测量的能力（即构念）的因素进行相关验证呢？也就是说，需要找到一定的标准来进行相关验证。然而，什么是标准？标准包括哪些？当时对于这些问题尚未有定论。所以，有些学者建议还是主要依靠分析测试内容是否符合要测量的能力来进行效度验证，而进行内容分析的主体则是内容相关领域的专家。如果专家一致认为测试的内容与教学大纲或要测量的能力一致，则测试具有较好的效度。在这一时期（1921—1951 年），效度理论得到了初步发展（Newton & Shaw，2014）。

到了 20 世纪 50 年代，美国心理学会（American Psychological Association，简称 APA）在 Cronbach 的带领下，制定了第一个专业标准来指导测试的效度验证工作，并于 1952 年发布了《标准》的第一稿（American Psychological Association，1952）。该版标准中有一章专门论述了效度，也开始区分不同种类的效度，从以前的内容效度和相关效度拓展到 4 种效度：内容效度、预测效度、身份效度（status validity）和相容效度（congruent validity）。1954 年正式出版时，身份效度更改为同期效度，相容效度更改为构念效度（APA et al.，1954）。非常重要的一点是，效度理论中引入了构念效度这一概念。此概念在 Cronbach & Meehl（1955）里程碑式的文章中得到完整的诠释。Cronbach & Meehl 认为，有些测试注重内容的相关性（内容效度），有些测试偏重标准测量（效标关联效度，包括预测效度和同期效度），但是对于一些没有这些衡量标准的测试来说，效度验证还得另辟蹊径。这方面主要考虑的就是构念的问题（Newton & Shaw，2014）。构念就是假设的特征（postulated attribute），可以在具体测试中体现出来。构念效度可以通过逻辑分析以及收集和分析实证证据这两个途径进行验证。

《心理测试和诊断技术建议》在 1966 年得到修订，以《标准》第一版的名称出版。效度的种类也由四种压缩到三种：内容相关效度、效标关联效度和构念相关效度。这种分类一直持续到 1985 的第三版《标准》。第一版《标准》指出，这三类效度彼此并不是完全分开的。然而，实际的测试验证却有分开的趋势，因为人们倾向于使用某种效度来验证某种类型的测试，进而呈现出某种碎片迹象。因此 1952 年至 1974 年这段时间也被称作效度碎片化时期（fragmentation of validity）（Newton & Shaw，2014）。

随后，1974 年至 1999 年这段时间被称为效度的“Messick 时期”（Newton & Shaw，2014），即前文所述的一体化期。Cronbach & Meehl（1955）认为，每当效标关联效度和内容效度不足以界定测量质量时，就应考虑构念效度。Messick 指出，效标关联效度和内容效度总是不充分的，所以效度基本上就是构念效度（Messick，1989），理由是对于分数的所有解释都隐含着某种构念。构念效度的验证就是“收集与测试分数解释有关的各种证据”（Messick，1989：17），所以，Messick 把所有效度证据都纳入构念效度范畴中（Messick，1989：20），把效度定义为“经验证据和理论依据在多大程度上支持基于测试分数或其他评价方式所做出的推断或采取的行动是充分和适当的？对这个问题的综合评价性判断就是效度”（Messick，1989：13；李清华，2006）。Messick（1989）用“效度框架层面”（facets of validity framework）表达自己的效度思想，构建了一种“渐进矩阵”（progressive matrix）（如表 1–2 所示）：构念效度在每个层面都出现，从左上角到右下角，每个层面增加一个新成分（李清华，2006）。Messick 重塑了效度验证，强调调查分数的意义，鼓励尽量多地收集和积累能够得到的任何证据和分析，警示效度验证不能仅依靠孤立的单次数据分析或实证调查。这种将效度理论重新一体化的观点是 Messick 的一大贡献（Newton & Shaw，2014）。Messick 的定义完全摒弃了分类效度观，具有突破性意义，成为 20 世纪晚期效度思想的“时代思潮”（Newton & Shaw，2014：99）。

李清华（2006）总结 Messick 对效度理论的贡献主要表现在：（1）明确了效度验证的对象是测试分数的解释和使用，而不是测试本身；（2）确立了构念的核心地位，加强了对构念效度作为效度整体概念