

破解教育评价难题 探索“融合评价”新方法¹

林梦泉 任超 陈燕 吕睿鑫

摘要：分析了教育评价面临的挑战以及评价改革与方法创新的紧迫性，重点剖析了计量评价的本质与“客观”性，同行评议的本质与“主观”性；以基于客观事实的同行评议为基础，提出“融合评价”的新方法，并对其理论框架和应用路径进行了深入研究；最后通过实例验证了融合评价的可行性和有效性，为丰富教育评价理论体系、创新实现路径、建立更加科学的评价体系提供理论支持，助推教育评价改革。

关键词：计量评价；同行评议；融合评价；结构性证据；评价改革

作者简介：林梦泉，教育部学位与研究生教育发展中心副主任，研究员，北京 100083；任超，教育部学位与研究生教育发展中心评估处副处长，副研究员，北京 100083；陈燕，北京工业大学高等教育研究院助理研究员，北京 100083；吕睿鑫，教育部学位与研究生教育发展中心评估处项目主管，北京 100083。

习近平总书记在全国教育大会上强调，要深化教育体制改革，健全立德树人落实机制，扭转不科学的教育评价导向，坚决克服唯分数、唯升学、唯文凭、唯论文、唯帽子的顽瘴痼疾，从根本上解决教育评价指挥棒问题。科学的评价方法论是扭转不科学的教育评价导向的理论基础。当前，在过于依赖定量评价的国内外大学排名备受诟病的背景下，形成了多种“唯”，造成了排名“迷信”化，存在“盲用”现象，极大影响我国教育发展与评价改革。

教育评价绕不过定量评价和定性评价，但孤立地采用定量评价和定性评价方法都存在一定缺陷，特别是过于依赖有限数据进行定量评价的大学学科排名，缺陷更加凸显。落实习近平总书记在全国教育大会上的讲话精神，需要深入分析评价理念、方法、背景和影响，遵循科学的评价方法论，是扭转不科学教育评价导向的理论基础。多年来，国内外关于定量评价和定性评价的持续争议，最根本的原因是孤立、排斥性地强调各自的方法优势，加上应用中的偏离、误读、误解、误用，引发定量评价过于被“客观”化，也成了“唯”的重要原因之一。而把种种错误导向归罪于评价本身，是一种简单化的认识。历史经验告诉我们，即使评价理念正确，价值导向正确，如果方法不当，评价结果不能反映事实，评价目标仍会落空。因此需要认真梳理当前教育评价的主要特征及面临的挑战，理性地分析定量评价和定性评价的本质特征，客观分析其优势和缺陷，在此基础上探索融合二者优势的新途径和新方法，破解评

¹ 基金项目：国家自然科学基金面上项目“基于大数据的一流学科水平评估与绩效评价研究”（编号：71673124）

价难题，为扭转不科学的教育评价导向提供理论和应用支持。

一、教育评价面临的挑战

教育是一种复杂的社会活动，教育评价在教育管理中地位、作用以及评价本身的价值观、质量观、方法论，大众对评价的认知等，各种因素形成了互促关系，系统性地影响着评价的导向和作用，从而影响教育的科学发展。教育评价方法论是在价值观、质量观的框架下形成和建立的，因此方法论研究是教育评价改革的重要基础。分析教育评价的基本功能和作用，梳理评价方法论在教育评价改革中的重要地位和面临的改革挑战，可为教育评价改革研究提供基础。

1.教育评价的基本功能和本质

教育评价是检验教育行为和成果的活动，从评价方法上看，大致可分为定量评价和定性评价两大类，对应的主要方法是计量评价和同行评议。定量评价和定性评价也被演绎成客观评价和主观评价，准确地说，这里的“客观”“主观”只是表示评价类型，并不意味着定量评价是客观的，定性评价是主观的。定量评价主要方法是计量分析，是基于清晰的定量信息进行计量分析，对相关教育现象进行判断，这个过程称为计量评价；定性评价主要通过人为评价，主要方法包括专家（同行专家或学术共同体、行业专家、管理专家）评议、学生评价、用人单位评价等。其中，最具代表性的是同行评议。

在科技、人才全球化背景下，科技竞争在于人才的竞争，人才竞争在于教育的竞争。在此背景下，教育呈现双重地位和作用，既具有战略性也具有战术性。人才培养及其科研支撑的社会经济发展具有战略性；而在WTO框架下，遵循的教育服务贸易特征，教育的学术流动和学生流动具有现实性和战术性。教育水平或排名评价逐渐成为这种流动简明易用的助推器。当前，主要基于有限数据定量的简易排名性评价，从学校外部直接影响高校的声誉、招生等，逐步成为各种教育、人才竞争的展示“平台”。但这种主要依据“头衔”“论文数”和获奖的简易性排名，可比指标少，反映大学学科发展情况具有很大局限性，加上其发布又缺乏明确解读，被当成大学学科整体实力而产生了越来越大的影响，客观上助长了“SCI、ESI至上”倾向，造成了计量评价的“客观性”“影响力”被扩大化。

同行评议作为定性评价的一种重要方式，也存在许多争议。主要认为其具有一定的主观性和不确定性，并不能完全替代定量评价或完全克服定量评价的缺陷。所谓的“学术共同体”评价，实际上就是“同行评议”。由于争议的客观存在，同行评议也并不等同于“精准学术性”和“高可信度”。因此，如何正确认识、科学改进及合理应用计量评价和同行评议，是当前教育评价面临的巨大挑战，是扭转不科学教育评价方法的重要课题。

2. 教育评价方法创新是破解评价改革难题的重要途径

作为管理、诊断和监督手段的教育评价活动，在评价理念和目标确定后，构建评价体系是实现评价目标的核心步骤，但单纯从指标设计出发构建体系往往受到限制。将评价理念、体系建构、方法创新“三个维度”系统谋划，才能更好地实现评价的总体目标，破解教育评价的关键问题，而这个顶层和系统性谋划往往被评价组织者忽略。

方法创新是建立中国特色、世界影响的教育评价体系的重要内容。在国内外一些基于少量公共数据进行大学学科排名的冲击下，要深入贯彻落实全国教育大会精神，坚持党的领导、坚持立德树人、坚持破“五唯”，坚持遵循高等教育发展规律，解决中国问题，贡献中国智慧，更需要研究能够凸显立德树人成效、社会服务贡献的评价方式^[1]，检验教育成效。由以上分析可见，过于依赖数据的定量评价，容易造成不良导向。因此，研究如何改进和应用计量评价，特别是同行评议，提升评价质量，成为新时代的新课题，是评价组织者和教育管理者面临的新挑战，对于破解不科学评价体系具有重要意义。

二、计量评价与同行评议的优势与不足

国际上，已形成了计量评价和同行评议相对独立的评价研究群体，并建立了相对独立的评价模式，各自强调其优势。为探索新的评价方法和模式，首先对计量评价和同行评议的优势与不足进行深入分析。

1. 计量评价的优势与不足

计量，是利用技术和法制手段实现单位统一和量值准确可靠的测量^[2]。在计量过程中，认为所使用量具是标准的，用它们来校准、检定受检量具，以衡量和保证使用受检量具进行测量时所获得测量结果的可靠性。计量涉及计量单位的定义和转换、量值的传递和保证量值统一所必须采取的措施、规程和法制等。

教育计量，是利用技术和法制手段实现对教育数据进行标准统一和量值准确可靠的测量。教育计量涉及基于计量标准的定义、量值的一致性，以及计量的举措和规范。

教育计量评价，是在计量的基础上，进行统计分析，形成结果数据，并对统计结果表征的教育状态、成效进行定义和研判。教育计量评价一般包含四个步骤：计量标准的制定、数据的获取和传递的真实性、规范的统计模型、统计结果体现教育成效（状态、质量、水平）的表达。根据计量评价的特征，其主要优势有：分辨率高，易于统计分析，可更具客观性；主要不足有：表征的教育现象单一，作为评价依据具有孤立、偏向性和不全面性，数据本身通常存在可靠性问题。

“计量评价”不等于“客观评价”。要做到符合事实的客观评价，必须保证计量评价的“四个可靠”。一是标准可靠，数据度量的标准制定要符合办学等实际，可表达、可获取、可精准解读；二是数据可靠，数据获取渠道可靠，获取信息符合标准，对获取信息经过有效清洗；三是模型可靠，采用符合计量评价目标的统计模型，模型算法符合教育规律和科学规范，并具有清晰内涵；四是评价可靠，对分析结果的认知要符合教育内在规律、符合教育战略方向，即人们赋予统计分析结果用于对教育的评价所表达的教育发展状态和教育成效是正确的。可见，计量过程是技术（数据统计）的客观过程，但计量标准及计量结果表征教育状态和成效实际上是主观过程。对于计量结果而言，即使前三步都可信，若误读了分析结果，整个计量评价结论也是不真实的。因此，符合以上“四个可靠”的计量评价才是客观的评价。

从一个侧面看，“四个可靠”是计量评价破“唯”的一种理论基础。前三方面是技术层面，产生问题主要在第四方面，即对数据表征教育质量和成果的误读，或简单地认为计量评价就是客观评价，这是逐步形成各种“唯”、形成“SCI、ESI至上”的重要原因。

下面，我们以ESI计量数据为例来分析。统计者按照自己定义的学科口径，统计各学科所有发表论文被引用的总次数，并依据总次数统计该学科在全球学科中的百分位，这是计量分析过程。按“四个可靠”的理论，我们可以发现，“ESI体现我国学科学术发表水平”的认识不完全具有客观性。一是学科口径不一致和数据范围不科学。ESI定义的22个学科与我国学科目录差异较大，即使名称相同的学科，其内涵也不尽相同，计量标准存在较大差异，这对可比性造成很大挑战。数据范围不科学体现在数据统计范围主要以英文期刊为主，许多高水平中文等非英文期刊没有在统计范围内。二是质量导向不突出。ESI统计全部发文的总被引次数与学术人口有关，对体量大（发表论文数多）、质量并不很高（每篇论文引用量少）的学校和学科利好，“低水平”的积累可超过“高水平”论文引用数据。三是统计模型待优化。ESI统计论文的被引用次数时，未区分施引论文的层次，被大量低水平论文引用的论文容易冲进ESI论文榜单的前列，一些高质量论文由于理论超前，没有被广泛理解和接受，很难成为高被引论文。四是代表水平的认知需纠正。将ESI计量统计结果作为判定大学学科科研水平甚至是学科水平的认识不是客观过程，是人为判断。这些问题在人文社会学科体现得尤为突出。

可见，应用“四个可靠”分析，ESI计量评价不等同于客观评价。综上，ESI不能准确评价学科的学术发表水平，自然也不能准确地代表学科的科研水平，更不能用来体现学科建设水平，也不宜用来对不同学科的学术发表水平进行比较。认识和应用ESI时，应充分考虑以上因素和分析的事实，“盲用”ESI可能引起决策失误。

2.同行评议优势与不足

定性评价在这里主要是指评价对象非纯数据，需要通过人为参与研判而形成非定量或定

量结果的评价方式^[3]。人为参与评议主要有专家评议、学生评价、用人单位评价等。学术界提出的“学术共同体”评价，实际上也属于同行评议范畴。保证同行评议可靠性是个复杂的系统工程。

社会对同行评议的可信度存在一定质疑。深入分析其优势和不足，是改进同行评议的前提。其主要优势：一是学术性。通过专业性、学术性研判，实现对非定量信息的评价，形成定量或定性结果，基于其专业性，评价可体现教育规律；二是导向性。可设置刚性或柔性原则引导同行评议，实现定量评价无法实现的价值导向，克服定量评价将数据直接标定结论的刚性缺陷；三是综合性。可进行综合性、整体性和比较性研判。主要缺陷：一是不确定性。评价过程的人为评价可能存在不确定性、随意性；二是非学术影响。存在一定的人情等非学术性因素的可能性，影响评价的可信度；三是评议分级局限。一般情况下，同行评议存在评价分辨极限，根据对人为评议能力的专门研究，人为将评议对象分成9档以上，会产生较大误差，分档的精确度大大下降，因此对象数量较多时，评议结果将形成较多并列；同时，人为评议的对象数量也是有限的。基于以上两个原因，评议材料较多时不得不分组邀请不同的专家评议，分组评议必然形成不同组评价结果的不可比性问题，需要采取额外的提高跨组可比性举措。这是任何同行评议不可回避的问题。对于人为评价中无材料的声誉调查，则主要是权衡评价的分辨问题。

多年来，教育评价领域对以上问题有诸多激烈争论，也进行了许多有益的研究和实践，取得了一些成果，提高了同行评议的可信度，但主要还是就同行评议论同行评议，问题的破解存在局限性。

在解决教育评价指挥棒问题的形势下，要破解“ESI、SCI至上”等只依据数据定量评价教育状态、水平和成效而产生的不利倾向，坚持立德树人成效导向，坚持服务贡献导向，需要加快对定性评价特别是同行评议模式的研究和实践，助推破解不科学的教育评价问题。

三、融合评价

在总结近年研究生教育评估以及中外合作办学评估经验基础上，吸取计量评价和同行评议的优势，提出“融合评价”新方法。融合评价突破了计量评价和同行评议各自的局限性，希望在破解上述难题、为教育评价改革方面作出有益贡献。

1. 基本理论框架

思路与目的。“融合评价”是吸收计量分析的客观性和同行评议的学术性、综合性的优点，充分应用数据、证据的客观事实特征，与人的学术和专业性智慧融合交互、相互支撑，进行更加科学的评判，提升同行评议的导向性、约束性和易控性，同时提高计量评价的广泛性和

可靠性。这种方式克服了一般同行评议的短板，提高客观性和可信度，提升评价精度，从而提升评议分辨率，突破了评议结果 9 档的局限。同时提高评议效率，评审的材料数量可以扩大。反之，其融合理念应用于定量评价，也将提升计量评价的可信度。融合评价为解决不科学的教育评价问题提供了新路径，丰富了教育评价理论体系。

方法含义。方法一：借助数据计量的客观事实，为同行评议提供支持，提升同行评议的可信度；方法二：借助同行专家的专业性，解决数据的不确定性问题，提升计量评价的准确性；方法三：是融合评价更加宏观的应用，超过微观的计量或同行评价融合，评价方法设计时可以融合多维评价方法、多维对象信息、多维信息类型。如对一个项目或指标的评价，除了同时可采用计量评价和同行评议、优势互补外，还可在指标信息类型和信息来源渠道上进行多维融合，全面提升评价的科学性，破解“五唯”等问题。

应用步骤。以同行评议为主线的融合评价共分为四个步骤：一是规范评议信息，提供客观事实。对评议信息的提供提出要求，包含统一要求的“结构性证据”，如关键数据、主要成效的证据、证据链、证据可鉴性和可寻路径等；二是就评议对象的公共数据进行写实性分析，提供给专家参考；三是根据评议价值导向，制定评议规则和评价标准。主要包含价值导向、评价重点、评价结果的分类规则等；四是组织评议。要求同行专家按照规则和导向，关注对象事实、参考数据及其分析结论，进行综合研判。为充分发挥融合评价的优势，可利用现代化评议平台，依据数据事实和分析结果有序呈现，协助同行专家评议，并对评议行为进行规范、提醒、约束，以及评议结果反馈验证等技术，提高评议客观性，克服人情因素，提升评价的整体可靠性。

以上关于“结构性证据”是融合评价新方法的重要内容。本文提出“结构性证据”是指从结构性考虑多维元素之间的内在关系，从而更科学、系统地呈现可靠证据，助力教育评价。具体方法是：从延续性、立体性、互鉴性的视角，形成结构性事实，使得教育评价的依据更加可靠、可信、准确，更加实事求是，更能呈现教育内涵发展规律。“结构性证据”包含了“证据链”和“立体证据”。“证据链”主要指延续关联的证据；“立体证据”主要指跨越链条式的、包括立体空间式的具有交互关系的证据。“结构性证据”的提出拟从多维、结构性评价理念出发，创新证据提供方式、要求和分类，为融合评价等各类评价方式的改进、评价可信度的提高提供有力支撑。

2.融合评价在计量评价中的应用

前面分析了独立的计量评价和同行评议的优势和缺陷，采用融合评价机制，是提升两种评价方式可信度的有效方法。下面，分析以计量评价为主线的融合评价方法。根据前面的讨论，任何计量评价都面临可靠性两个方面的问题，要保证可靠性，关键在于数据本身的真实性和对数据代表的教育现象的合理定义。从定量信息的类型角度，计量评价可大致分为两类：一是数据的计量标准清晰，数据体现的教育特征清晰，如具有学科信息的博士学位论文抽检

情况、某学科学生的成果等；二是数据的计量标准不清晰，如涉及交叉学科研究成果，比如国家奖励、国际奖项、高水平学术发表等情况，数据本身界定清晰，但数据的大学或学科归属不易划分。按融合评价的理念，可借助专家的专业性和学术性对数据信息的大学、学科等归属进行“审核”，为定量评价提供“矫正”机制，破解定量评价的数据归属难题，同时是解决交叉研究评价难题的良方，助力交叉研究^[4]通过专家制定某信息量表，作为计量分析的依据，也是融合评价在计量评价中的有效运用。

3.融合评价在同行评议中的应用

“基于客观事实的同行评议”是“融合评价”最典型的体现和应用。为有效应用新方法，以下按评议对象的信息分类分析其应用路径。

一是对象信息数据相对丰富，但又难以进行简单的计量评价，此类情况在教育评价中最常遇见。评议对象信息具有若干数据支撑表来表达一个目标，但又难以分别单独进行计量评价，可采用以同行评议为主线的融合评价方法。首先，可对数据进行适当分析，提供分析结果；其次，参考结果制定评议的约束限制，制定特色的同行评议标准和规则，由同行进行综合研判。这种方式的效果介于计量评价和同行评议之间，可发挥计量评价和同行评议的最大优势，属于最典型的融合评价。如学科师资水平评价、中外合作办学评估等，下文将具体论述；二是对象信息具有突出的非定量成果，反映建设成效。此类情况可通过增强证据和证据链，特别是提供规范的证据表述，既增强可评性和可比性，又不失开放性的特色贡献成效呈现；同样，也要制定与此相对应的同行评议标准、规则和评价导向，实现基于成效客观事实的同行评议，提升评议的质量和可信度。如学科评估中学科社会服务贡献代表性案例评价；三是对象信息介于以上两类之间，可采用一、二结合的评价方法。

四、融合评价实证分析

基于上述理论研究结论，将“融合评价”理念应用于学科评估^[5]和中外合作办学评估等教育评价实践中，来分析融合评价的可行性和有效性。

1.学科评估师资水平评价

过去各类评估在评价师资水平时，通常做法为统计各学科拥有特定“具有学术头衔”的教师数，如院士、“长江学者”等，进行计量评价。从数据标准的角度而言，“学术头衔”可以在一定程度上体现师资水平，但不能完全代表师资团队水平。同时，单纯基于这种计量评价的方式，也在某种程度上引发了重金引进“戴帽”人才，“孔雀东南飞”等人才无序流动现象。

学科评估在方法上的创新体现了融合评价新理念，在学科师资队伍水平评价上进行了改

革探索^[6]。基本目标是采用“整体结构质量”和“代表性骨干教师”相结合的方式，请专家参考客观数据进行评价。一是提供结构性证据。针对评价目标，要求单位提供简洁、结构化信息，包括师资队伍职称结构、学历结构、年龄结构、学缘结构等丰富结构数据，以及代表性骨干教师的学科方向、学术头衔、学术兼职、海外经历等支撑信息；二是为专家提供规范的客观数据，简洁、直观地为专家综合判断提供参考；三是提供评价基本规则，要求重点关注“代表性”教师总体情况，不强化“帽子”，要参考师资总体结构，同时也提出了约束性要求。该评价改革和结果受到了各方的普遍肯定。这种方法既避免了单纯使用计量信息的弊端，也保证了同行评议的可靠性和科学性。与传统的“数帽子”的计量评价相比，师资队伍水平的融合评价将评价重点从过去的学术头衔转向实际水平、学术影响及发展潜力，但仍然参考了计量信息。这种方式对构建良好的学科建设生态及师资队伍发展环境具有良好的导向作用。

2. 中外合作办学评估

中外合作办学评估是对本科及以上中外合作办学机构和项目的办学规范性以及教育质量进行的合格性评估。按“融合评价”理念建立专家评议机制：一是规范被评信息。要求《自评报告》内容结构与指标体系相对应，填报内容存在一定钩稽关系，围绕评价指标构成“结构性证据”，为专家提供事实依据。二是通过计量分析，为专家提供客观数据排序或对比供参考，帮助专家综合研判。如对“引进外方课程”及“外方教师授课比例”、学生满意度调查、社会声誉度调查数据进行分析统计后，提供省市、全国平均值，供专家评价参考。三是评价标准制定的培训。研讨制定评价规则和流程，并通过研讨培训，明确评价规则和标准，引导专家统一认识，把握评价的一致性，保证公平，提高同行评议的可信度。

3. 多维结构化论文评价体系（MSPA）

学术论文是大学学科学术成果的重要表征。统计“发表SCI、Ei等论文数”“影响因子”“ESI高被引论文数”等尽管具有一定的客观性，但只采用一种评价方法、一个维度、线性孤立的结构，无法全面反映论文质量，易导致学术研究急功近利、助推“唯”、催生学术不端、学术造假等问题，因而在学术界饱受诟病。

学科评估研究探索了第三类“融合评价”方式。针对“一个学科的学术论文质量”评价，单一融合评价难以从根本上做到科学有效，鉴此研究采用了宏观“融合评价”新理念新方法，称为“多维结构化论文评价体系”（Multi-dimensional Structured Paper Assessment System, MSPA）。MSPA体系的特征是：拥有多个维度，且维度的确定具有合理逻辑结构。这里采用“三结合”结构：计量评价和同行评议结合；国外与国内期刊结合；质量和数量结合。以期破解唯定量和“唯论文”数形成的不科学评价难题。具体做法是：除“中国高被引论文数”等计量评价外，重点关注“代表性论文”的同行评议；合理把握中、外期刊论文的综合评价，规定代表性论文须

包含一定比例的中国期刊论文,克服 ESI 等计量指标主要统计英文期刊的问题,确保学术论文评价覆盖范围合理;以质量作为各维度评价的基本前提,兼顾有质量保证的定量信息。同时,同行专家学术论文质量评价,关注了结构性证据,包括发表刊物名称、收录类型、期刊影响因子、引用次数、署名情况,以及论文摘要等内容,为结构性的综合评价提供支撑。

五、结语

论述了教育评价方法在教育评价改革中的重要地位,全面分析了计量评价和同行评议在数据本身可靠性、数据体现的教育属性、评议过程科学性、评价结果可信度等方面的优势和不足。在此基础上,对提出的融合评价的理念、内涵、程序和应用方法,进行了详细论述和深入研究。从教育评价方法论的视角看,融合评价方法吸收了计量评价和同行评议的优点,初步形成了理论框架和应用路径。融合评价方法在同行评议的应用中,突出了“基于客观事实的同行评价”新思路,对改进同行评议、提升评议质量方面具有明显的优势。通过新方法在学科评估和中外合作办学评估等方面的应用实例分析,进一步验证了融合评价的可行性和有效性,为融合评价模式的应用和推广提供了有效路径。通过新方法的理论研究和应用分析可见,“融合评价”对解决孤立的计量评价和同行评议的难题具有重要意义,对破解“五唯”,从方法、方向上提供了新的路径和解决方案,对改革不科学的教育评价体系和方法,对拓展教育评价理论和实践体系,具有一定的理论和现实意义。

参考文献

- [1] 全国教育大会讲话报道[EB/OL]. (2018-09-11) [2019-03-18]. <http://edu.people.com.cn/n1/2018/0911/c1053-30286253.html>.
- [2] 金华彰.“计量”一词溯源 回顾我国“计量”名词定义的制定[J]. 中国计量, 2010(3): 53-56.
- [3] 陈喜乐, 李腾达, 刘伟榕. 定性评价方法的极限与超越——基于对同行评议的研究综述[J]. 未来与发展, 2014(5): 16-20.
- [4] MOED H F. The future of research evaluation rests with on intelligent combination of advanced metros and transparent peer review[J]. Science and public policy, 2007, 34(8): 575-583.
- [5] 黄宝印, 林梦泉, 任超, 等. 努力构建中国特色国际影响的学科评估体系[J]. 中国高等教育, 2018(1): 13-18.
- [6] 林梦泉, 陈燕, 毛亮, 等. 以立德树人为核心的中国特色人才培养成效评价初探[J]. 学位与研究生教育, 2019(4): 1-7.

(选自《学位与研究生教育 2019 年第 12 期》)