文章编号:1006—9860(2021)11—0024—008

评他能力：人工智能时代学生必备的高阶思维能力*

张 生[1]，王 雪[1]，齐 媛[2①]

(1.北京师范大学 中国基础教育质量监测协同创新中心，北京 100875；2.中国教育科学研究院，北京 100088)

摘要：人工智能时代的评价改革的核心在学评融合，一线实践层面的核心在评他能力的发展。对他人作品进行评价的活动是一种重要的高阶思维训练活动，对学生关键能力和人格品质的发展具有重要的价值和意义，是落实评价改革的核心抓手。该研究研制信效度良好的评他能力测评工具，基于分层抽样设计，对全国东中西地区的18420名小学、初中和高中学生进行大规模调查。结果发现：我国中小学生评他能力的总体表现处于中等偏低水平，亟待整体提升和培养；内部结构发展不均衡，尤其是认知反馈维度相对较弱，学生对他人的作品提出问题、改进建议的能力训练不足；在城市特征、人口学特征上呈现低水平均衡发展，性别和学段的部分差异达到小效应量，需要进一步改革实验，研制有效提升策略；在影响因素方面，评价频率和回复频率对评他能力的解释力最高，说明当前评价的活动开展较少，在量的层面都有积极影响，应该加大评价活动的次数和频率。由此可见，我国中小学生的评他能力尚处于自由发展、并未有效的积极干预措施的阶段，还未发挥“互联网+”在评他方面的优势，需要在学与教方式变革过程中积极融入评他活动的设计，以进一步推进评价改革的落地。

关键词：学评融合；评他能力；评价改革；核心素养；人工智能

中图分类号：G434 **文献标识码：**A

随着《深化新时代教育评价改革总体方案》[1]的出台，呼吁新时代需要新的评价改革。当前对评价改革多数关注评价的诊断性作用，强化了互评在评价多元性、评价信效度方面的意义和作用，这对宏观层面的评价有非常大的价值；同时在一线层面，新的问题也随之会产生，随着多级评价的开展，学生发展的规律、学与教的现状、优点与缺点都能清晰地呈现出来，同时也会增加评价的负担，减弱诊断性评价的价值。随着时间的推移，在学生有限的学习时间里，如何提升质量是摆在大家面前的关键问题。近日，《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》[2]的印发，也体现对这一关键问题的高度关注，其背后的评价理念变革是落实“双减”政策的核心和探索方向。评价本身既具备诊断的作用，用在学习活动中，还是一种训练学生高阶思维发展的抓手，即评价具备诊断和学习双重价值，如何实现从诊断到改进过渡到从评价的学习性到诊断这一路径的变革，依赖于“互联网+”、大数据、测量与评价技术的不断发展，更依赖于强调评价学习性，即评他能力的改革落地，也需要学评融合的评价新理念的进一步阐明。现阶段同伴互评和自我评价作为融合世界中的重要评价方式，可行性与有效性等诊断性目的仍是关注焦点[3][4]，难以适用于人工智能时代学生评他能力的培育要求。在这一背景下，基于学评融合的评价新理念[5]，评他能力作为学评融合的重要一方面，亟需进一步理清其内涵与结构，掌握当前我国中小学生评他能力的现状和影响因素，挖掘数据背后的规律，形成可行的实践方法，将对我国评价改革的落地具有独特的意义。

一、问题提出

(一)核心概念界定

评他能力的内涵建构始于其上位概念，即评价能力。随着教育评价改革和育人目标的不断深化，评

* 本文系国家自然科学基金联合基金重点项目“基于‘天河二号’超级计算机的教育系统化监控评估、智能决策仿真和应用研究”(项目编号：U1911201)研究成果。

① 齐媛为本文通讯作者。

价能力的内涵应由知识的掌握性向能力的发展性、由对结果的诊断性向学习的过程性、由对作品的评价向人的评价不断发展变化。Tai等人[6]对评价能力的定义是：“相对于预先确定但不一定明确的标准，批判性地评估表现的能力。这需要一个复杂的反思过程，包括内部的自我评价，和外部决策他人‘工作’质量的同伴评价”。李泽文[7]认为评价能力是能够根据明确标准对有关事物做出一种高度自觉的价值判断能力。Tai等人[8]在此基础上提出了更简单的定义，认为评价能力是对自我和他人的工作质量做出决定的能力。至此，评价能力是一种价值判断力，作用对象是他人或自己的“工作”质量，衡量标准是与明确标准的一致程度，这一界定也与当前迫于升学和考试的压力，评价更多从诊断性评价入手有关[9]。例如，在同伴互评领域，大部分学者都关注学生评分的客观性或与教师评分的一致性上[10]，以此来衡量评价能力的高低具有一定的合理性，但已不符合当下的育人理念、促进学生学习的作用有限，因此，评他能力背后的评价理念革新是关键，面向学习性的学评融合新理念提供了重要路径。

评价能力的内隐性多维结构是研究评他能力结构与测评的关键。学术界对评价能力的结构研究还未形成统一结论，现有研究主要从评价对象和评价过程两个方面进行探讨。从评价对象方面的分析多对应到互评和自评两种评价方式下的评价能力[11][12]；从评价过程方面进行的分析聚焦到一般评价过程中的评价能力结构，例如，苏倩[13]将评价能力的内部结构定义为三方面：一是将外在评价标准内化为自己的内在标准，二是要有意识、独立地进行判断和选择，三是要反思自我，不断地促进自我的发展。Dominique等人[14]在建构的三级维度评价能力模型中包括：(1)明确评价标准：思考哪些方面是评价作品时需要关注的；(2)判断作品表现：识别同伴作品的优缺点；为下一步的学习提供反馈：对同伴的作品给予建设性的反馈。至此，评价过程方面的研究可以概念化为四个主要组成部分：标准意识、判断选择、认知反馈和反思提升。因此，上述两种评价方式和四个主要过程是建构评他能力结构的基础和关键。

数字世界对物理世界的拓展和自身的时空泛在性赋能评他能力的常态化、精准性和科学性培育[15]。基于上述研究，评价目的、评价方式和评价过程是构建评他能力内涵与结构的基石。如图1所示，评他能力的内涵取决于y轴的评价目的这一本质性取向，与面向作品的诊断性评价相比，面向人的学习性评价是首要本质，他人或自己的“工作”质量只是衡量学生评他能力的载体，关键是在学生作为评价者参与评价活动的过程中，学生自身学习的发生与促进。评他能力的结构依附于x轴的两种评价方式和z轴的四个主要评价过程建构。以往研究表明，自我评价和同伴互评两种评价方式是形成性评价过程的必要环节[16]，二者相结合能够促进评价方法的有效性[17]。因此，评他能力同时关联自我评价和同伴互评两种方式，而无论哪一种方式，其内部蕴含的评价过程，尤其是学习性评价过程才是建构评他能力结构的关键。在上述四个主要评价过程中，按发生顺序在z轴上由低到高排列，标准意识是评价的基础过程，在两种评价方式和评价目的中均有体现。判断选择是对他人或自身作品进行量化打分，涵盖在两种评价方式中，但因其作品主导性，仍以诊断性为主要目的，故未被纳入评他能力结构中。认知反馈和反思提升以其认知加工和思维深度参与而体现为学习性评价，分别主要发生在学生互评和自我评价中。至此，评他能力结构体现在关联两种评价方式下的标准意识、认知反馈和反思提升一系列学习性评价过程中。

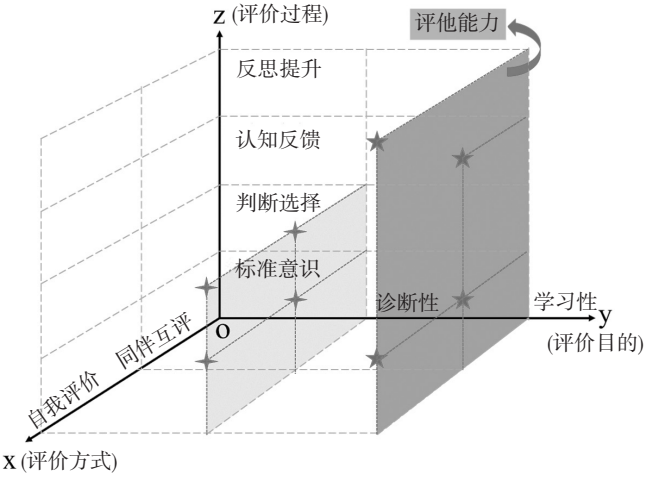

图1 评他能力

综上，评他能力是学评融合新理念的核心内容之一，强调在物理世界与数字世界融合育人过程中，发挥评价的高阶思维的学习特性，优化设计学生的评他活动，形成人人创造、分享作品，人人评价他人的作品，在评价他人(含自身)的作品中，不仅关注作品本身，更加关注作品的创作者，还关注创作过程，是基于创作者出发的一种系统的评价新观念。首先，它强调的是融合世界中，在学习时空得以拓展的情况下的评价活动；是关注人在评价他人过程中评他活动对学生自身成长的价值；也是关联自评和互评两种评价方式，强调评价过程中学生主体的标准意识、认知反馈和反思提升关键思维过

程；是指向学生直接发展的评价活动，是先于评价的诊断性活动，是基于评价活动数据开展自动化诊断评价的基础，是落实过程性、增值性、综合性评价的重要基础；最后，它还是减轻各级诊断性评价带来的评价负担的重要路径和抓手。

评他能力的培育既能支持原有自我评价和同伴互评活动的高质量开展，也超越其支持学生更高层级的思维能力发展。这主要体现在对学习性评价过程的关注，即不再关注自评和互评中的客观性与准确性，而是体现在根据一定标准，评价他人作品过程中自身的认知性反馈和对自我作品的反思提升上。其关键特征：一是理解质量标准并具备标准意识的能力；二是应用标准对他人作品表现进行认知性反馈的能力；三是评价他人过程中反思和修改，以提升自己作品的能力。因此，上述界定的评他能力结构将不止于价值判断，能将判断的结果外显为认知性反馈行动，如“指出问题”“提出建议”等，并在大量阅读、评价他人作品的过程中逐步认识到事物的多面性，对自我作品进行反思，并进一步提升自身作品、标准认知、学习方法等，才是完成一次有积极意义的完整的评他过程。

(二)研究问题

以往的评价能力相关研究中，对教师和学校层面的研究相对较多，对学生的评价能力研究较少[18]。并且，从与时俱进的发展角度审视，尚存不足：评价能力的内涵与结构界定中仍存在以诊断性的质量评价为主，尚未有研究基于促进学习评价的新取向对我国中小学生的评价能力特点进行深入分析。其中有一个引起许多兴趣的问题，学生的不同评价背景，即评价数量或频率，对其互评质量是否会有影响？具体而言，引入交互层数的概念，这种影响是源自评他的频率还是回复的频率？以往研究并未形成统一结论，一方面，交互数量与交互层数似乎都带来更深度和高质量的互评效果[19]。另一方面，交互数量多，深度互动的内容反而少[20][21]。那么当定位到本研究提出的评他能力时，评他频率和回复频率的影响效果如何还有待分析。

因此，本研究基于学评融合的评价新理念，提出评他能力这一面向新时代育人要求，统整当前自评和互评中实践困境，关注评价过程中学习性作用的高阶思维能力。研制信效度良好的学生评他能力测评工具，基于分层抽样设计，对全国东中西地区的小学、初中和高中学生进行大规模调查。确定以下4个研究问题：我国中小学生的评他能力现状如何？不同城市特征和人口学特征的群体存在哪些差异？家庭背景因素和评价背景因素对评他能力的影响如何？以评他能力测评作为一项循证指标，为推动高阶思维能力培养和核心素养发展提出对策建议。

二、研究设计

(一)研究对象

项目组对全国中小学生评他能力现状进行调查，共回收290352份问卷，有效问卷278571份(有效率95.94%)。为便于统计分析，使用分层抽样进行二次抽样，共抽取18420份问卷。先按地域分布和经济发展水平抽取省份，根据2020年的人均GDP数据由高到低排序，分为高中低三级，结合国家统计局2019年抽样调查公布的东西部学校的总体数量。最终确定东部4个、中部3个、西部3个，高人均GDP3个、中3个、低4个，共10个省、市、自治区①。依据2019年全国抽样调查数据中各学段的在校学生比例，在控制小、初、高的抽取比例大致为1:2:4的基础上，每个省、市、自治区抽取的问卷总数占所有抽取省份问卷总数的比例与我国2019年抽样调查数据中该省份在校学生总数的相对比例一致。这说明，本研究抽取的问卷数据对总体的代表性较好，基本上能反映全国中小学生评他能力的基本情况。样本的分布情况如表1所示。

表1 样本分布情况

变量	性别		学段			地域			人均GDP水平			父母最高受教育程度			
	男	女	小学	初中	高中	东部	中部	西部	高	中	低	初中及以下	中专或高中	专科或本科	硕士研究生及以上
人数	8920	9500	10342	5376	2702	8566	5784	4070	4729	7456	6235	8939	5796	3526	159
百分比%	48.4	51.6	56.1	29.2	14.7	46.5	31.4	22.1	25.7	40.5	33.8	48.5	31.5	19.1	0.9

(二)研究工具

① 这10个省市按照人均GDP水平由高到低排序，依次是北京市(东部、高GDP)、福建省(东部、高GDP)、山东省(东部、高GDP)、内蒙古自治区(西部、中GDP)、四川省(西部、中GDP)、河南省(西部、中GDP)、青海省(西部、低GDP)、河北省(东部、低GDP)、山西省(中部、低GDP)、黑龙江省(中部、低GDP)。

1.评他能力问卷结构及其信效度

采用自编《中小学生评他能力问卷》，经专家审校和两次预测试，最终形成14道题目的问卷供正式测评使用，采用李克特量表5级记分。分为认知反馈、标准意识与反思提升三个子维度，总量表和子量表的内部一致性信度在0.829–0.898之间，如表2所示。验证性因子分析显示，模型拟合良好。

表2 评他能力的维度及信度

维度	题数	子维度Cronbach' α系数	总Cronbach' α系数	拟合指标
认知反馈	4	0.829	0.898	χ^2= 2378.663***, df=71, RMSEA=0.042, CFI=0.984, TLI = 0.979
标准意识	4	0.862		
反思提升	6	0.898		

认知反馈是指学生在掌握基础知识及技能的基础上，通过对同伴作品所体现的知识技能的整合对比，进而做出认知性反馈的能力。包含4道题目，分别从“指出问题”“提出建议”“定位问题”“给出解决办法”4个方面出发，例如“我经常在评价中直接指出作品中的问题或错误”“我经常在评价中对作品提出具体的修改建议”。

标准意识是指学生在理解质量标准的基础上，能对自己的评价行为对应到相应的质量标准，将外在事物有意识地内化为内在标准的能力。包含4道题目，例如“我给出的评价符合一定的标准”“我给出的评价经过了仔细的思考”等。

反思提升是指学生在大量阅读、评价他人作品的过程中逐步形成对自我作品的反思，并进一步调节提升自身作品、标准认知、学习方法等的能力。包含6道题目，例如“评价他人让我能更好地完成我的作品”“我学会从评价者的视角看待自己的作品”等。

2.评价背景因素

包括评他频率和回复频率，各有一道题目：“我评论别人作品的频率是”“别人评论了我的作品后，我对评论进行回复的频率是”，采用4点计分，从几乎没有到经常。

三、数据分析

(一)中小学生评他能力的总体表现

学生评他能力总分的均值为3.497(SD=0.647)。进一步采用重复测量方差分析，结果表明，中小学生在评他能力三个维度上的得分存在显著差异，F(2，36838)=8016.349，p<0.001。配对比较发现，三个维度上的得分两两之间均存在显著差异，得分由高到低依次为：反思提升、标准意识、认知反馈。

(二)中小学生评他能力的差异分析

以城市特征和人口学特征为分组变量，采用独立样本t检验和单因素方差分析，从地域、GDP水平、性别和学段背景信息上，对我国中小学生评他能力的差异进行了考察。t检验的效应量以d值估计，方差分析的效应量采用η^2估计，结果如表3所示。

表3 中小学生评他能力及各维度的方差分析和差异检验结果

维度	地域			F	p	η^2
	西部(n=4070)	中部(n=5784)	东部(n=8566)			
总分	3.518(0.647)	3.495(0.631)	3.487(0.658)	3.107	< 0.05	0.03%
认知反馈	3.020(0.927)	2.990(0.889)	2.986(0.919)	2.027	0.132	——
标准意识内化	3.600(0.772)	3.608(0.754)	3.591(0.775)	0.788	0.455	——
反思提升	3.796(0.757)	3.757(0.754)	3.752(0.774)	4.684	< 0.01	0.05%
维度	人均GDP水平			F	p	η^2
	低(n=6235)	中(n=7456)	高(n=4729)			
总分	3.462(0.630)	3.508(0.639)	3.524(0.679)	14.226	< 0.001	0.15%
认知反馈	2.967(0.901)	3.008(0.905)	3.011(0.936)	4.570	< 0.05	0.05%
标准意识	3.549(0.743)	3.609(0.761)	3.646(0.807)	22.727	< 0.001	0.25%
反思提升	3.735(0.750)	3.774(0.755)	3.785(0.795)	6.854	< 0.01	0.07%
维度	性别			t	p	d
	女(n=9500)	男(n=8920)				
总分	3.483(0.618)	3.511(0.676)		2.980	< 0.01	0.043
认知反馈	2.928(0.900)	3.066(0.919)		10.289	< 0.001	**0.152**
标准意识	3.595(0.733)	3.602(0.804)		0.571	0.571	——
反思提升	3.778(0.735)	3.748(0.794)		–2.635	< 0.01	0.039
维度	学段			F	p	η^2
	小学(n=10342)	初中(n=5376)	高中(n=2702)			
总分	3.564(0.635)	3.426(0.665)	3.380(0.622)	133.147	< 0.001	**1.425%**
认知反馈	3.055(0.912)	2.926(0.925)	2.904(0.866)	50.971	< 0.001	0.550%
标准意识	3.660(0.766)	3.532(0.782)	3.494(0.725)	78.800	< 0.001	0.848%
反思提升	3.839(0.744)	3.689(0.795)	3.622(0.746)	124.413	< 0.001	**1.333%**

注：黑体表示达到小效应量。

不同地域的中小学生，在评他能力总分和反思提升维度上存在显著性差异；不同GDP水平的中小学生，在评他能力总分及各维度上均有显著性差异。但均未达到小效应量。

不同性别的中小学生，在评他能力总分、认知反馈和反思提升维度上存在显著性差异，结合数据来看，男生在评他能力总分和认知反馈维度上的得分显著高于女生，达到小效应量；不同学段的中小学生，在评他能力总分及各维度上均有显著性差异。事后检验表明：在评他能力总分和反思提升维度上，小学生得分显著高于初、高中生，初中生得分显著高于高中生，达到小效应量。

(三)中小学生评他能力的影响因素分析

将评他能力总分及各维度得分与父母最高受教育程度、评他频率和回复频率做相关分析，结果如下页表4所示。所有变量两两间均存在显著性相关。

表4 评他能力总分及各维度得分与父母最高受教育程度、评他频率和回复频率的相关系数

变量	M	SD	父母最高受教育程度	评他频率	回复频率	总分	认知反馈	标准意识
父母最高受教育程度	1.723	0.797	1					
评他频率	2.680	0.933	0.053***	1				
回复频率	2.874	1.009	0.026***	0.578***	1			
总分	3.497	0.647	0.074***	0.268***	0.263***	1		
认知反馈	2.995	0.912	0.026**	0.201***	0.171***	0.671***	1	
标准意识	3.598	0.768	0.091***	0.216***	0.222***	0.847***	0.343***	1
反思提升	3.764	0.764	0.065***	0.225***	0.236***	0.874***	0.299***	0.730***

注：*表示<0.05，**表示p<0.01，***表示p<0.001，下同。

综合前述结果，采用多元分层回归分析，考察上述所有变量对评他能力总分和各个维度得分的影响，结果如表5所示。城市特征、人口学变量、父母最高受教育程度和评价背景分别解释了总分方差的0.2%、1.5%、0.3%和8.7%。模型2–4的调整R^2项显示，评他能力3个二级维度中，反思提升维度(模型4)受自变量影响程度较高，认知反馈维度(模型2)受自变量影响程度较低(讨论详后)。

表5 各变量对评他能力总分及各维度得分的回归系数[B(SE)]和解释贡献率(ΔR^2)

自变量	模型1		模型2		模型3		模型4	
	B(SE)	ΔR^2	B(SE)	ΔR^2	B(SE)	ΔR^2	B(SE)	ΔR^2
城市特征		0.002***		0.000*		0.002***		0.002***
中部	–0.035(0.020)	0.000*	——	——	——	——	–0.050(0.020)*	0.001**
东部	–0.047(0.019)*		——		——		–0.057(0.019)**	
中GDP	0.071(0.017)***	0.002***	0.046(0.017)**	0.000*	0.079(0.017)***	0.002***	0.050(0.017)**	0.001**
高GDP	0.096(0.019)***		0.049(0.019)*		0.126(0.019)***		0.060(0.019)**	
基本人口学特征		0.015***		0.011***		0.008***		0.014***
男性	0.044(0.015)**	0.000**	0.151(0.015)***	0.006***	——	——	–0.039(0.015)**	0.000**
初中	–0.216(0.017)***	0.015***	–0.141(0.017)***	0.006***	–0.167(0.017)***	0.008***	–0.199(0.017)***	0.014***
高中	–0.287(0.022)***		–0.164(0.022)***		–0.212(0.022)***		–0.288(0.022)***	
父母最高受教育程度	0.053(0.007)***	0.003***	0.013(0.007)	0.000	0.076(0.007)***	0.006***	0.046(0.007)***	0.002***
评价背景		**0.087*** **		0.045***		**0.058*** **		**0.065*** **
评他频率	0.260(0.007)***	**0.067*** **	0.198(0.007)***	0.039***	0.208(0.007)***	0.043***	0.218(0.007)***	0.047***
回复频率	0.266(0.007)***	**0.070*** **	0.176(0.007)***	0.031***	0.222(0.007)***	0.049***	0.237(0.007)***	**0.056*** **
调整R^2	**0.107**		**0.056**		**0.075**		**0.083**	
F	220.166***		136.089***		212.957***		167.120***	

注：1.所有连续变量均经过标准化处理；2.城市特征和人口学特征在编码为哑变量时，分别将西部地区、低GDP、男、小学作为参照变量；3.模型1的因变量为评他能力总分，模型2–4的因变量分别为：认知反馈、标准意识和反思提升；4.黑体表示达到中等效应量；5.“——”标识根据差异检验和相关分析结果，该变量在因变量上的差异或相关不显著，不将其作为自变量纳入回归模型。

四、讨论与结论

(一)我国中小学生的评他能力整体情况处于中等偏低水平

本研究的目的之一，是希望在以核心素养指导的新评价理念下，衡量我国中小学生评他能力现状，并得到一个基本判断。在调查问卷的量尺上，我国中小学生评他能力总分的均值为3.497，高于3分这一理论中值，我国中小学生评他能力水平在本调查问卷量尺上，达到中等水平。若以百分制的思路，人为划定60分至70分为及格，则这个分数仍处于及格水平，未达到中等线(3.5分)。因此，结合理论中值和中等线两个参考标准，可认为我国中小学生评他能力处于一个中等偏低的水平。诚然，本研究界定的评他能力强调评价的学习性，对于评他诊断性上的能力体现还未有涉及，但本团队在以往研究中，已略述诊断性评他能力的内涵[22]，它不同于传统学生评分的准确性，而是基于数字世界各种人工智能技术支撑下的应用数据决策、助力学习的能力，将在未来的研究中进一步论述。

(二)我国中小学生的评他能力内部结构发展不均衡，认知反馈维度亟待提升

在评他能力三个维度上，学生的得分两两之间均存在显著差异，反思提升维度表现较好，这一结论与以往研究一致[23]，说明学生在评价他人作品的过程中能够反思、意识到其对自我作品的视野拓展和灵感激发，同时愿意将这种认知转化为具体的改进、提升自我作品的行为外部倾向。相反，认知反馈维度表现较差，均值仅为2.995。究其原因，认知反馈维度的界定是面向学生的高阶思维发展和核心素养育人要求，表现在“判断观点”“指出问题”“提出建议”“定位问题”“给出解决办法”五个方面，需要学生进行深度的思考和认知能力加工，而中小学生由于认知水平正处于发展阶段，且受制于学校、家庭对该方面能力的培养忽视[24]，故在该维度的表现不足，上述研究结论与已有研究相似[25]。该结果对评他能力的提升工程或有所启示：当前我国中小学生拥有较好的基础践行评他活动，能够认识到评他活动对自我的认知反思和改进提升具有重要意义，也呼应学评融合新理念中发挥评价学习性功能的最后一个阶段，即反思与改进；应将干预工作的重点放在学生认知反馈中的思维深度和认知加工等方面，及要彻底变革为以学生生成为核心的学与教方式。

(三)我国中小学生的评他能力在外部群体差异上呈现低水平均衡，性别和学段的部分差异达到小效应量

调查结果表明，地域分布、人均GDP水平、性别和学段这些常见的社会学分类指标，并不能大效应地区分不同组别人群的评他能力水平。可能的原因：一是我国中小学生评他能力发展比较均衡；二是中小学生的评他能力差异主要来自于个体之间的差异(组内差异)。但是结合上述评他能力整体情况处于中等偏低水平的现状判定，这种均衡是基础水平上的均衡，如何在全局尺度上提升我国中小学生评他能力，将是一个体量巨大的任务。

在上述分类指标中，仅有性别和学段的部分差异达到小效应量，地域分布和人均GDP水平未达到小效应量，与已有研究相似[26][27]。具体而言，就性别而言，男生在认知反馈维度上的得分显著高于女生。可能原因是由于性别的生理和心理差异，男生在提出问题和发表质疑时会更加果断自信，而女生在思考问题时表现得更为全面细腻。该结论与以往研究一致，男生在批判性思维开放思想、分析能力以及自信心上都要显著高于女生[28]。就学段而言，在评他能力总分和反思提升维度上，呈现出小学显著优于初、高中，初中显著优于高中的趋势。中学阶段的学生理应比小学阶段掌握了更多的知识技能，为什么评他能力不升反降？分析其原因可能有以下几点：首先，小学阶段相对中学阶段更注重学习策略的教授和运用，该结论与已有研究一致。例如，孙智昌等人[29]发现，学生的学习策略水平随学段升高而显著下降，Clear等人[30]和Eur等人[31]的研究结果也表明，随着年级的升高，学生自我调节学习策略的使用会下降，而无论是学习策略中的反思策略、认知策略还是自我调节学习策略，都是评他能力总分和反思提升维度的重要基石。其次，随着年级的升高，迫于考试和升学的压力，学生较少有机会以自我总结的方式巩固知识[32]，教师在教学中用于评他活动的时间和深度也不足，导致学生的评他能力总分和反思提升维度表现较差。这一猜想，在问卷中有具体体现：随着学段升高，评他频率的均值由小学2.72，初中2.63，到高中2.62，频率低且呈下降趋势，结合本研究中评他频率对评他能力的解释率最高的这一结论，也证明了这一可能原因。再次，学生的评他能力很大程度受到非智力因素和学生主体性人格的影响，而很多非智力因素都是随年龄升高而降低的[33]，学生的主体性人格也有随年龄升高而降低的趋势，例如申景玉等人发现小学五年级的学生在主体性人格的多个方面都高于初二学生，在初中阶段也出现低年级明显高于高年级的状况[34]。最后，马郑豫等人还发现，小学生自我感知的学习能力高于初、高中生[35]，即自我评价的评他能力更客观而表现为逐步下降的趋势。总之，这些对中学生评他能力的不利因素综合起来发挥的作用最终超过了随年龄增长认知能力的提高带来的有利因素，从而导致整体上评他能力总分和反思提升维度呈现出随学段升高而下降的趋势。

(四)在本文模型涉及的变量中，评价背景因素对评他能力的解释率最高。认知反馈维度更不易受影响

评价背景中的评他频率和回复频率在评他能力总分、标准意识和反思提升维度上的影响达到中等效应量。表征家庭背景的父母受教育程度，对评他能力的影响效果有限，在评他能力总分及各维度上，仅能最高解释0.6%的差异。因此，再进行细分比较意义不大，整体趋势表现为：父母最高受教育程度越高，学生的评他能力越高。较之区域、人均GDP水平、性别、学段等，评他频率和回复频率对评他能力的解释力高出数倍。具体而言，在评他能力总分、标准意识、反思提升维度上，回复频率的贡献程度相对更高，而在认知反馈维度上，评他频率的贡献程度相对更高。这一结论进一步印证了以往研究观点[36][37]。回复频率是发生在学生对他人评价进行再次回复的过程中，交互层数由一层变为二层，是在他人对自身作品评价内容上的再思考，这一过程必然会调动学生已有知识技能，对照标准进行整合比较，进而形成自我思想沉淀，完成内容回复，因此，有益于学生反思提升和标准意识维度的提高。在认知反馈维度上，由于需要学生以同伴作品为载体，对其所体现的知识技能进行整合对比，进而做出认知反馈，因而评他频率相比于回复频率，对该维度的解释率更高。同时，以往研究表明，互评准确性作为学生高阶思维评他能力的一种体现，训练对其有积极的提升作用[38]。综上，评价频率的提高是提升学生评他能力的有效切入点，鼓励多层次的评他活动可收事半功倍之效，也符合量变到质变的基本认知。

上述分析同时也指向另一结论：评他频率和回复频率对学生反思提升维度影响较大，对认知反馈维度影响较小。从前面表5中相应的ΔR^2项可以看出。这一结论也从侧面反映出，仅增加评价频率，难以达到认知反馈维度提升的理想效果。同时，比较模型整体的调整R^2项也可以发现，认知反馈维度相较于其他两个二级维度，更不易受城市特征、人口学特征和家庭背景因素的影响，这表明认知反馈维度的提升任务艰巨，其差异不在个体内的性别、学段等，仅考虑外部的家庭背景因素作用效果微弱，提升关键是个体内部因素的认知加工层级和思维深度。

五、建议

评他能力面向的是一种自下而上的评价改革实践，评他活动是针对一线实践层面提升质量而提出的一种新型高阶思维能力活动。将评他能力测评作为核心素养教育发展的一项循证指标，以调查结果为依据，结合当前评价在教育实践活动中的应用，在学与教方式变革过程中积极融入评他活动，进而提升教育质量。本研究提出如下建议。

第一，相关政府部门和行业人士要对当前现状有清晰和足够的认识。要正确把握我国中小学生评他能力发展的现状，充分认识到评他能力提升任务的艰巨性。当前，有两个问题值得关注，一是我国中小学生评他能力整体水平中等偏低，二是评他能力内部结构发展不均衡，认知反馈维度亟待提升，但该维度不容易受个体内差异和家庭背景因素的影响而改变。前者意味着提升工程的任务巨大，后者意味着提升工程的难度巨大。目前，结合党和政府相关规划中，对评价改革和核心素养教育的要求目标，相关政府部门和行业人士对此挑战要有清晰和足够的认识。

第二，教育实践者应将评他能力的培养融于日常的教学活动中去。要注重发挥评价的学习性功能，考虑性别差异，着力提升初高中学生的评他能力。初中阶段作为学生自我调节学习发展较为迅速和逐渐成熟的重要时期，或是当前评他能力提升的重要阶段。同时，初高中学生在繁重的课业任务下，学生的自主时间相对较少，如果在此背景下，将评价与学习相剥离，评他能力的提升也必然以进一步加重学生学业负担而落幕。因此，只有将评价融于日常的教学活动，强化在线学习环境的支持，对其进行全空间、全时段的设计[39]，发挥评价的学习性功能才是提升初高中学生评他能力的本质选择。另外，鉴于性别表现出来的评他能力差异，还应关注男生的反思提升维度、女生的认知反馈维度培养。

第三，教师要基于在线学习环境，提升评他活动的频率和思维的深度。要有效建立提高评价频率与提升评他能力之间的良性循环。充分利用在线学习环境的优势，营造有利于评他活动开展的育人环境，设计有效的学生的评他活动，重视思维的互动[40]，提升评他活动的频率和思维的深度，提高学生的评他能力。基于在线学习环境，要注重设计激发多层次的评他活动，让学生在多次的评他迭代中，促进自身思维的螺旋式上升。另外，评价频率的效用发挥，要依赖于教师的良性设计与过程指导，教师的指导要尤其体现在学生评价中高水平认知加工层级上。

第四，研究者要深入探究评他能力、尤其是认知反馈能力的提升策略与机制。要深入研究评他能力中认知反馈维度提升的条件和机制。从上述研究发现来看，评价频率对认知反馈维度的促进作用相对较低，内部的变化特点还有待进一步分析，在某种程度上，可能存在增加评价频率也不一定导致认知反馈的积极变化。如何从本质上研究提升条件和机制，尤其是结合人工智能技术和学评融合的评价新理念是我国学者需要结合中国实际，深入探索的现实问题。

参考文献：

[1] 中共中央、国务院.中共中央 国务院印发《深化新时代教育评价改革总体方案》[EB/OL].http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202010/t20201013_494381.html,2020-10-13.

[2] 中共中央办公厅、国务院办公厅.中共中央办公厅 国务院办公厅印发《关于进一步减轻义务教育阶段学生作业负担和校外培训负担的意见》[EB/OL].http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202107/t20210724_546576.html,2021-07-24.

[3] 范逸洲,冯菲等.评价量规设计对慕课同伴互评有效性的影响研究[J].电化教育研究,2018,39(11):45-51.

[4] Admiraal W,Huisman B,et al.Self- and Peer Assessment in Massive Open Online Courses [J].International Journal of Higher Education,2014,3(3):119-128.

[5][22] 张生,王雪等.人工智能赋能教育评价：“学评融合”新理念及核心要素[J].中国远程教育,2021,(2):1-8+16+76.

[6] Tai, J.,Canny, B. J.,et al.The role of peer-assisted learning in building evaluative judgement:opportunities in clinical medical education [J].Advances in Health Sciences Education,2016,21(3):659-676.

[7] 李泽文.评价能力:不容忽视的素质[J].语文教学与研究,2002,(5):8-9.

[8] Tai H M,Ajjawi R,et al.Developing evaluative judgement:enabling students to make decisions about the quality of work [J].Higher Education,2018,6(3):467-481.

[9] Muchlis M,Ibnu S,et al.Students' Result of Learning at Chemistry Department through Assessment of,for,and as Learning Implementation [J].International Journal of Instruction,2020,13(2):165-178.

[10] Steendam E V,Rijlaarsdam G,et al.The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL [J].Learning & Instruction,2010,20(4):316-327.

[11][13] 苏倩.小学班级生活中学生评价能力养成研究[D].昆明:云南师范大学,2014.

[12][18][24] 彭杰.初中生评价能力培养策略研究[D].长春:东北师范大学,2013.

[14] Sluijsmans D,Brand-Gruwel S,et al.Training teachers in peer-assessment skills:effects on performance and perceptions [J].Innovations in Education & Teaching International,2004,41(1):59-78.

[15] 齐媛,张生.学评融合：落实评价改革的重要路径[N].中国教育报,2020-10-31(03).

[16] 蔡旻君,王心怡等.在线学习者参与评价的理论探讨及实证研究[J].中国电化教育,2021,(3):15-23.

[17] Bouzidi L,Jaillet A.Can Online Peer Assessment Be Trusted? [J].Educational Technology & Society,2009,12(4):257-268.

[19][36] 严亚利,黎加厚.教师在线交流与深度互动的能力评估研究——以海盐教师博客群体的互动深度分析为例[J].远程教育杂

志,2010,28(2):68–71.
[20] 赵婴,何克抗.基于微信的跨文化网络交流互动深度研究[J].电化教育研究,2019,40(10):35–39+60.
[21][25] 李红霞,赵呈领等.促进学习的评价:在线开放课程中同伴互评投入度研究[J].电化教育研究,2021,42(4):37–44.
[23] 项纯.中小学生自我评价能力的现状、问题与对策[J].教育科学研究,2018,(11):56–61.
[26] 黄瑄,李秀菊.我国青少年科学态度现状、差异分析及对策建议——基于全国青少年科学素质调查的实证研究[J].中国电化教育,2020,(12):69–77.
[27][39] 张生,张平等.人工智能时代下的精准减负:提升减负政策效能的关键——基于小学生学习投入与主观课业负担类型的划分及特征分析[J].中国电化教育,2020,(1):114–121.
[28] 卢家楣,刘伟等.中国当代大学生情感素质的现状及其影响因素[J].心理学报,2017,49(1):1–16.
[29] 孙智昌,项纯等.我国中小学生学习动力与学习策略的现状与对策[J].课程·教材·教法,2016,36(3):78–85+77.
[30] Cleary,T.J.,et al.Self–regulation,motivation,and math achievement in middle school:variations across grade level and math context [J]. Journal of School Psychology,2019,47(5):291–314.
[31] Eur,J & Educ,et al.Grade level,study time,and grade retention and their effects on motivation,self–regulated learning strategies,and mathematics achievement:A structural equation model [J].European Journal of Psychology of Education,2013,28(4):1311–1331.
[33]《非智力因素及其培养》全国协作组.我国儿童青少年非智力因素发展的研究[J].心理发展与教育,1995,(4):1–6.
[34] 申景玉.中学生主体性人格及其与学业成就的相关研究[D].天津:天津师范大学,2006.
[35] 马郑豫,张家军.中小学学生学习策略的调查研究[J].教育研究,2015,36(6):85–95.
[32][37] 张生,陈丹等.中小学生自主学习能力对在线学习满意度的影响[J].中国特殊教育,2020,(6):89–96.
[38] 李菲茗,李晓菲等.训练对同伴互评评分准确性的影响——以“三维动画设计与建模”课程为例[J].中国远程教育,2018,(5):63–67+78.
[40] 黄蔚,曹榕等.人工智能时代批判性思维能力的提升策略——思维图示的应用对小学生批判性思维能力提升的实证研究[J].中国电化教育,2019,(10):102–108.

作者简介:

张生:副教授,博士,研究方向为教育大数据、智慧测评、信息技术与学科深度融合。

王雪:在读硕士,研究方向为教育测量、评价与统计。

齐媛:助理研究员,博士,研究方向为学习心理、教育技术。

Skills of Assessing Others: The Necessary Higher Order Thinking Skills of Students in the Era of AI

Zhang Sheng[1], Wang Xue[1], Qi Yuan[2]

(1.Beijing Normal University, Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing 100875; 2.National Institute of Education Sciences, Beijing 100088)

Abstract: The core of assessment reform in the era of Artificial Intelligence (AI) is the integration of learning and assessment, and the core of its front-line practice is the development of the skills of assessing others. The activity of assessing others' works is an important training process of higher order thinking skills, which is significant for the development of students' key abilities and personalities and is the core starting point for the assessment reform. The research developed a reliable and valid evaluation tool for evaluating the skills of assessing others. A large-scale survey was conducted among 18420 primary and middle school students in the eastern, central and western regions of China by stratified sampling. The results show that the overall performance of Chinese primary and middle school students' skills of assessing others is at a moderately low level, which needs an overall cultivation and improvement. Besides, the development of its internal structure is unbalanced, especially the cognitive feedback dimension is relatively weak, which means the training of asking questions and giving suggestions to improve others' works is insufficient. In terms of urban characteristics and demographic characteristics, it shows a low-level balanced development. Differences between gender and school section reach a small effect size. Therefore, it is necessary to further reform the practices and develop effective promotion strategies. In terms of influencing factors, the assessment frequency and response frequency have the highest explanatory power on the skills of assessing others, indicating the frequency of the assessment activities should be improved. The research showed the skills of assessing others of Chinese primary and middle school students is still in a stage of free development without effective intervention measures, and has not yet played the advantage of Internet Plus in assessing others. Thus, assessing others' works needs to be actively integrated into the design of learning and teaching reform, and then further implement the assessment reform.

Keywords: integration of learning and assessment; skills of assessing others; assessment reform; core competences; Artificial Intelligence (AI)

收稿日期:2021年8月9日

责任编辑:李雅瑄