人工智能时代的教育测评通用理论框架与实践进路*

□ 杨华利 耿 晶 胡盛泽 黄 涛 徐晨曦

【摘 要】

人工智能技术的发展和深度应用为教育测评的智能化提供了技术支撑。本文结合国内外相关研究探讨教育测评理论的内涵、分类和应用场景，发现测评目标从宏观能力向微观知识转变、测评场景从单次静态向持续动态转变、测评方法从传统概率统计向深度学习转变。智能时代的教育测评需要以传统教育测量理论为指导，实现面向认知过程的时序动态建模分析。因此，本文提出人工智能时代教育测评模型通用性理论框架，包括教育数据分析、教育测评建模、模型参数估计、测评模型评估和创新教学应用等核心步骤，并通过经典模型的对比剖析智能化教育测评模型的内部运转规律。未来智能化测评研究的开展需要探索知识建构、认知发展和综合能力的理论价值，构建基于多维知识空间的认知诊断模型和面向时序数据的认知过程跟踪模型，并从智能化数据采集、智能化建模分析、教师数据素养等方面保证智能化教育测评的实施。

【关键词】 教育信息化；教育评价；教育新基建；教育测评；项目反应；认知诊断；知识追踪；智能化测评

【中图分类号】 G40-057 【文献标识码】 A 【文章编号】 1009-458x(2022)12-0068-10

DOI:10.13541/j.cnki.chinade.2022.12.007

引言

教育信息化促进了教育测评理念的变革，人工智能时代的教育更加关注以测评数据驱动的学习者个性化诉求。教育测评是有效教学的内存价值（Wiliam, 2011），已成为全球教育系统最突出的先行政策。2016年，*Science*杂志报道了美国国家科学基金会未来发展的六大科研前沿，大数据支持下的高级个性化学习支持（Mervis，2016）是前沿之一。期刊*Studies in Educational Evaluation*于2021年6月推出专题“从数据驱动转向基于数据决策（DBDM）”（Mandinach & Schildkamp, 2021），研究者们从多个维度探究通过教育测评数据分析辅助教育工作者开展科学决策，推动教育实践的快速发展。中共中央、国务院于2020年印发《深化新时代教育评价改革总体方案》，强调充分利用人工智能、大数据等信息技术提高教育评价的科学性、专业性、客观性，促进人工智能与教育测评的融合发展。2021年7月，教育部等六部门印发《关于推进教育新型基础设施建设构建高质量教育支撑体系的指导意见》，其中重点强调创新信息化评价工具，客观分析学生能力，支撑全过程纵向评价和全要素横向评价，鼓励探索尝试规模化在线考试，促进教育高质量发展。教育测评作为推动智慧教学生态的重要环节，在人工智能技术的影响下已发生巨大变化。智能化技术不仅丰富了数据采集的维度和数量，还记录了学生的过程性成长轨迹，通过全样本、海量测量数据分析可促进教师精准教学和学生个性化

* 本文系国家自然科学基金面上项目“面向时空融合的学习者认知诊断理论及关键技术研究”（项目编号：61977033）的研究成果。

学习的高效开展。

一、人工智能时代教育测评理论发展背景

（一）教育测评的概念界定

教育测评（Educational Assessment）是以现代心理学和教育学为基础，通过科学方法对学习者的认知过程和发展潜力等特征进行客观定量刻画，并对教育现象进行科学价值判断的过程。教育测评由教育测量（Measurement）与评价（Evaluation）组成，评价是价值判断的最终目标，测量是实现定量描述的手段（范涌峰, 等, 2019）。牛顿（Newton, 2007）从教育测评的目标出发定义教育测评的内涵，包括判断目标、决策目标和影响目标三个层次。从流程上看，教育测评分为定量、定性与价值判断三个步骤。桑德格勒等（Sondergeld, Stone, & Kruse, 2020）将教育测评分为常模参照测评和标准参照测评，前者包括轨迹、过渡表、学生增长百分位数和投影四个流行的增长模型，后者包括传统的Angoff高风险测试模型和现代更高级些的基于项目反应理论模型。布鲁姆等（Bloom, Hastings, & Madaus, 1971）将教育测评分为学习完成后进行的总结性评价和以学习过程为中心的形成性评价，前者侧重于宏观能力的评价，后者侧重于微观知识的精熟度评价。人工智能时代的教育测评亟须探索教育现象的本质和规律，将教育测评过程中抽象、潜在的属性精准量化，为价值判断与分析提供全面、有效的客观数据。

（二）教育测评理论分类

国内外有很多教育测评理论的分类方式。例如，有学者将其分为标准测验理论和新一代测量理论（Mislevy, 1993），前者包括经典教育测评理论、概化理论和项目反应理论。也有学者从传统教育测量（经典教育测评）和现代教育测量理论（项目反应理论）的视角来对比测评项目特征（Subali, 2021）。同时，以认知诊断测评为核心的新一代测量理论也已获得研究者们的关注（王立君, 等, 2020），该理论尝试挖掘学习者的内部信息加工过程规律。也有学者将项目反应理论归为认知诊断理论，即特殊的连续型认知诊断模型（Wang, et al., 2020）。近年来，伴随智能导学系统（Grossman & Salas, 2011）的兴起，加之新冠肺炎疫情的爆发，在线学习方式在国内外迅速推广，海量多次持续测评的过程性数据得以获取，由此推动了以“知识追踪”（Corbett & Anderson, 1994）为代表的智能化测评模型的发展。综上所述，我们发现教育测评可按照“宏观”（分数、能力）和“微观”（知识、技能）来分类，按照此规律我们将教育测评理论分为传统经典教育测量理论、现代测量理论（以项目反应理论为代表）、新一代测量理论（以认知诊断为代表）和人工智能时代的测评理论（以知识追踪为代表）四大类，体现了测评目标从宏观向微观的转变，从单次静态测评到多次动态测评的转变。

（三）教育测评场景剖析

基于布鲁姆对教育测评的总结与过程分类，学者们将教育测评场景分为两大类别，如图1所示。单次规模化测评场景是一种终结性测评，隶属于判断范畴，即通过对某一时刻学习者群体的测评数据建模来判断学习者当前的认知水平，适用于国际大规模测评或者区域教学质量检测等教育情境，代表模型包括经典教育测量模型、项目反应理论和认知诊断模型等。多次持续性测评场景属于形成性测评（Wongwatkit, Srisawasdi, Hwang, & Panjaburee, 2017），隶属于决策范畴，在学习过程中通过动态连续跟踪采集学生测评数据来建模，实时了解学生的学习状态，通过“反馈–矫正”不断调整课堂进度与学习者学习的重点，其建模方式以知识追踪模型为代表，能够动态监测知识水平的演变和发展趋势。

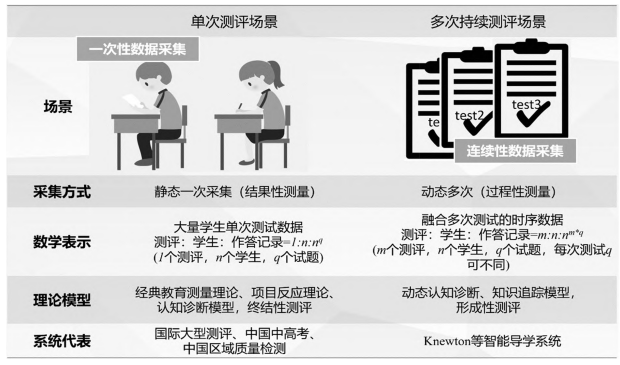


图1 教育测评场景对比

二、教育测评理论的跨越式变迁

伴随认知心理学和人工智能技术的迭代更新，教育测评理论经历了从传统数学统计向智能计算的跨越式发展。为探索其理论的发展与变迁，本文根据教育

测评的分类，选取项目反应理论、认知诊断模型和知识追踪模型三种理论来探讨。经研究发现，三种理论均经历了兴起阶段、发展阶段和智能化崛起阶段。

（一）项目反应理论：面向学习者反应的学习者潜在能力挖掘

针对基于随机抽样的早期教育测量理论“唯分数”的局限性，项目反应理论（Item Response Theory，IRT）采用项目特征函数描述单次测评场景或者自适应练习场景（Rasch, 1960）。具体来说，IRT是根据学习者在项目上的反应和其本身的潜在能力间的关系输出学习者能力值，实现由外在表现到潜在能力的转变（Embretson & Reise, 2013）。当前在教育领域广泛推广的Rasch模型，其本质即为参数IRT模型。针对项目反应理论的单维性假设的局限性，查默斯将其扩展至多维项目反应理论MIRT（Chalmers, 2012），利用多维隐藏能力刻画学生状态（Sympson, 1978），它对潜在能力的刻画更为精准。伴随人工智能技术的兴起，DIRT模型（Cheng, et al., 2019）和TC_MIRT模型（Su, et al., 2021），结合深度学习技术估计传统项目反应理论中的能力、区分度、困难度等参数，达到挖掘学习者潜在能力的目的，其对学习者能力的表示更为精准，同时兼顾了其模型的可解释性。

（二）认知诊断模型：基于Q矩阵的认知水平建模

认知诊断模型（Cognitive Diagnosis Model，CDM）以“微观认知属性”为测评目标（De La Torre, et al., 2009），在奥苏伯尔“有意义的接受学习”理论的指导下，通过融入心理学的认知特征，在教学过程中帮助师生挖掘学习中未掌握的技能，纳入已有认知结构，帮助学习者改进（Wiliam, 2011）。认知诊断适用于单次测评场景，支持终结性测评或诊断性测评，重在对学习者知识水平进行分析。认知诊断模型对学习者的认知结构进行建模，弥补了传统测量模型在内部认知结构上的不足和IRT中笼统的能力值无法判断被试微观层次认知的缺陷。目前，教育心理学中已有多种经典的认知诊断模型，主要包括补偿型与非补偿型两大类。非补偿型认知诊断模型，如采用Q矩阵理论的统一模型（DiBello, Stout, & Roussos, 1995），仅涉及“失误”“猜测”两个参数的确定性输入、噪声“与”门模型（DINA），其中DINA模型由于其参数的简洁性和易解释性受到研究者的青睐。补偿型认知诊断模型，典型的有通用型的G-DINA模型（De La Torre, 2011）。近几年来，学者们越来越重视利用数据挖掘方法来改善认知诊断效果。比如，通过“模型集理论”分别对主观题和客观题进行建模的模糊认知诊断模型FuzzyCDM（Liu, et al, 2018）；通过对项目反应精度和项目反应速度联合建模的JRT分层模型（Zhan, et al, 2017），拓展了多维认知诊断模型的输入信息；特别是NeuralCD神经认知诊断（Wang, et al., 2020）和RCD关系认知诊断（Gao, et al., 2021），通过神经网络自动训练与诊断函数结合，实现知识技能精准诊断的目标，代表着认知诊断实现智能化跃迁。

（三）知识追踪模型：基于领域知识的时间序列建模

伴随ITS在线学习的兴起，知识追踪（Knowledge Tracing，KT）受到空前的关注。KT模型不再局限于对单次测评场景的研究，而是适用于学习者多次持续测评场景，实现对学习者学习历程的分析。贝叶斯知识追踪（Bayesian knowledge tracing，BKT）是基于概率的典型代表，将知识状态抽象为一组二元变量，以学生的实时交互作为输入，通过隐马尔科夫模型来模拟学习者学习过程中对知识掌握情况的变化（Yudelson, Koedinger, & Gordon, 2013）。BKT模型有许多变体，如融合猜测和失误因素、时间因素（Goldberg & Wang, 1991）和问题难度估计（Gan, et al., 2019）等，此类模型具备较好的可解释性。然而，随着深度学习技术的进步，学者们更加关注深度知识追踪（Deep knowledge tracing，DKT）。DKT将循环神经网络引入知识追踪，其表现效果明显优于概率模型（Piech, et al., 2015）。近几年来，国际上知识追踪的论文呈现爆发式增长趋势，其改进主要集中在技术改进与多特征融合改进两个方面：在技术改进方面，如基于记忆增强神经网络的动态键值存储网络（DKVMN）模型（Zhang, Shi, King, & Yeung, 2017）、基于卷积神经网络的CKT模型（Yang, Zhu, Hou, & Lu, 2020）和基于图神经网络的GKT模型（Nakagawa, Iwasawa, & Matsuo, 2019）等；在多特征融合改进方面，体现在遗忘特征、项目内容特征和学生能力特征等，如融入三种遗忘特征的DKT+F模型（Nagatani, et al., 2019）、考虑项目文本与知识并解决了冷启动问题的EKT模型（Liu, et al., 2019）和CF-DKD引入学习与遗忘特征

建模（Huang, et al., 2021）等。深度知识追踪系列方法的表现效果明显优于传统概率模型，但其训练过程类似黑盒，可解释性较差；基于教育心理学特征的模型则改善了深度学习模型的可解释性，但结果仍存在一定的波动与不稳定性。

如图2所示：在兴起阶段，基于概率与统计学方法对测评数据建模，从一定程度上解决了教育测评诊断学习者能力水平、知识水平与认知水平（Embretson, et al., 2013; De La Torre, 2009; Yudelson, et al., 2013）的问题；在发展阶段，利用数据挖掘与传统统计学方法联合建模，使其诊断精准性得到一定的提升（Chalmers, 2012; De La Torre, 2011; Zhan, et al., 2017）；在崛起阶段，将深度学习等智能化技术引入教育数据挖掘与分析领域，不论是传统概率统计模型，还是基于深度学习模型，其效果预测表现效果得到显著提升（Piech, et al., 2015），体现了人工智能技术对教育测评的深刻影响。智能时代的教育测评实现了从“经典分数论”向“能力论”转变，从“结果性测评”向“过程性测量”转变，从“静态单维诊断”向“动态多维追踪”的转变，以及从“经验主义”“数据驱动”向人工智能时代“数据决策”的精准测量的转变。

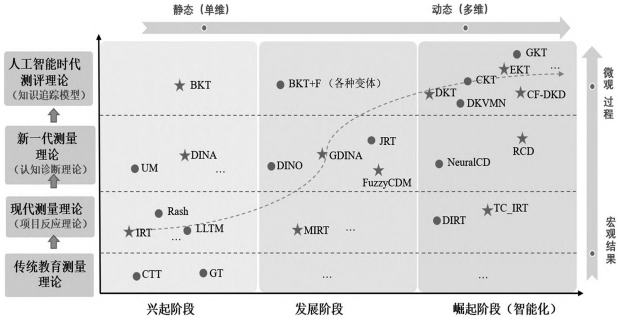


图2　教育测评理论跨越式变迁

三、人工智能时代教育测评框架和实现路径分析

教育测评模型的本质是通过学习者对项目的作答反应构建教育测评模型，并对学习者的认知状态进行追踪与测评。具体来说，模型通过分析可观测变量与不可观测变量的关系建立目标函数或挖掘潜在关联，对隐性变量进行参数估计，最终输出测评结果并应用于教育场景。智能化教育测评旨在应用人工智能技术，在提高测评精确度的同时更好地模拟不可观测变量的运行机制。结合教育数据挖掘流程、学习分析技术在教育中的应用框架，本文在剖析多种教育测评模型的实施路径基础上提出人工智能时代教育测评模型通用性理论框架，如图3所示。通用性理论框架包括教育数据分析、教育测评建模、模型参数估计、测评模型评估和创新教学应用等环节，通过对教育测评数据进行建模精准诊断学习者当前认知水平，挖掘认知变化内在的规律，通过应用实践助力于教育教学模式的创新。

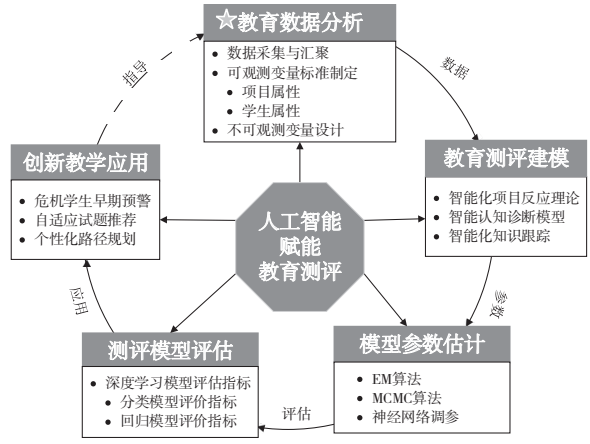


图3　教育测评模型通用性理论框架

基于以上教育测评理论框架，我们在每个阶段选取1~2个具有代表性的教育测评模型进行对比分析，研究各大经典测评模型的内涵与动作机理，剖析其模型函数的实施流程，探索其模型在人工智能时代的发展、应用以及存在的问题。

（一）教育测评数据分析

构建智能化教育测评模型需要对测评场景数据进行量化分析，从数据分析中挖掘出隐藏的显性变量与隐性变量。表1为经典测评模型数据采集特征与变量对比。

1. 教育测评数据集。基于两种教育测评场景，我们将数据集分为两类：①针对单次测评场景的固定项目的数据集，包括考试数据、国内外大规模测评数据和模拟数据等，常用于基于概率的项目反应与认知诊断模型，如国际知名的五项目SAT12数据集（Chalmers, 2012）和PISA2012数据集（Zhan, Jiao, & Liao, 2018）等；②针对多次连续测评时序数据集，呈现“项目数不固定、时间不固定”规律，适用于智能阶段测评模型，如Assistments/EDnet/KDD等公开

表1 代表模型的变量分析

理论	代表模型	数据集	可观测变量					不可观测变量					
			项目属性			学生属性		项目参数				学生参数	
			项目	知识点	其他	学生反应 R	做题行为	区分度 a	难度 b	猜测 c or g	失误	隐藏单元	能力
项目反应	IRT	单次	√	×	–	√	–	√	√	√	–	–	√
	MIRT	单次	√	×	–	连续	–	√	√	√	–	–	√
	TC_MIRT	多次	√	×	–	√	–	√	√	√	–	–	√
认知诊断	DINA	单次	√	Q	–	√	–	–	–	√	√	–	×
	JRT	单次		Q	–	√	反应时间	–		–	–	–	×
	FuzzyCDM	单次	√	Q	主观题	连续	–	√	√	√	√	–	√
	RCD	多次	√	多知识	关联	连续	–	–	√			√	√
知识追踪	BKT	时序	–	√	–	√	–	–	–	√	√		×
	DKT	多次	–	√	–	√	–	–	–	–	–	√	×
	DKVMN	多次	√	记忆	–	√	–	–	–	–	–	√	×
	EKT	多次	√	多知识	内容	√	–	–	–	–	–	√	×
	CF-DKD	多次	√	记忆		√	时间间隔	–	–	–	–	√	×

数据集、我国智学网数据集（Liu, et al., 2019）以及EAnalyst数据集（Huang, et al., 2021）等。

2. 可观测变量标准制定。根据测试主客体，可观测变量按学生与项目划分（Huang, et al., 2020）。在项目属性中，除了项目基础信息以外，项目所考核的知识点也是关注的焦点。总体来说，认知诊断模型采用Q矩阵表示关联的多个知识点，知识追踪要么假设项目只有一个知识点，要么采用“记忆模块”或神经网络来表示多个隐性知识点。学生属性包括学生与项目交互的学生反应特征（学生在参与测评过程中正确或错误回答项目）和学生行为特征（学生答题的反应时间和时间间隔等），学生行为特征对学习者的反应速度或记忆有影响。

3. 不可观测隐性变量设计。测量模型中涉及的不可观测的隐性参数有项目难度 b、区分度 a、能力 θ、猜测 c or g 和失误 s 等。前四者常见于项目反应理论，后两者常见于认知诊断模型。知识追踪模型采用神经网络的隐藏单元来表示隐性知识点状态，通过神经网络反向传播来调整参数。

（二）智能化测评模型分析

根据现有测评数据特征，如何利用智能化技术来构建智能化测评模型以实现教育过程性大数据的精准化诊断是智能化测评模型的核心。

1. 技术催生智能化项目反应理论

智能化项目反应理论遵循传统项目反应理论原理，以“学生的反应矩阵”为模型输入，以学习者能力为中心建模，以此预测学习者在未来项目上的表现。TC_MIRT（Su, et al., 2021）是智能化项目反应模型的代表，其原理仍是基于项目反应理论及其扩展的多维项目反应理论。其中，项目反应理论主函数如公式所示：

$$P(\theta)=c+\frac{(1-c)}{1+e^{-Da(\theta-b)}} \quad (1)$$

多维项目反应是基于被试多种能力之间相互作用的关系输出多维连续的能力值，将学习者的能力多维化，让学习者能力描述更加细化与精准，但由于多维能力的引入其计算参数按倍数增加，影响其参数估计的速度。TC_MIRT则通过长短时记忆网络与卷积神经网络等深度学习方法估计多维项目反应函数所需的参数，通过预测学习者反应的正确性来反向调节神经网络参数。相比于原始项目反应理论，其能力评价更多维化，计算速度更快；相比于多维项目反应模型，突破了其参数量大、估计速度慢的困境。

2. 技术驱动智能化认知诊断

人工智能时代的认知诊断，除了注重基于教育现象的深层教育规律分析以外，还利用先进的深度学习等智能化技术对不同教育测评中的变量建模，利用智能技术优越的表现性来提高认知诊断的精准性。关系认知诊断RCD是智能技术与认知诊断相结合的典型代表，其原理遵循传统认知诊断模型，以微观知识技能为核心，通过认知诊断函数获取学习者当前认知状态。传统经典认知诊断模型DINA主函数公式如下：

$$P_j(\alpha_i)=P\left(Y_{ij}=1|\alpha_i\right)=g_j^{1-\eta_{ij}}\left(1-s_j\right)^{\eta_{ij}} \quad (2)$$

DINA模型将学习者知识状态假设为掌握与未掌握二元向量，将学习者正确答题情况分为两种：学习者掌握了知识但答错（失误 s）和学习者未掌握知识而答对（猜测 g），输出学习者掌握模式 α_i。发展阶段FuzzyCDM融合项目反应理论与认知诊断理论，

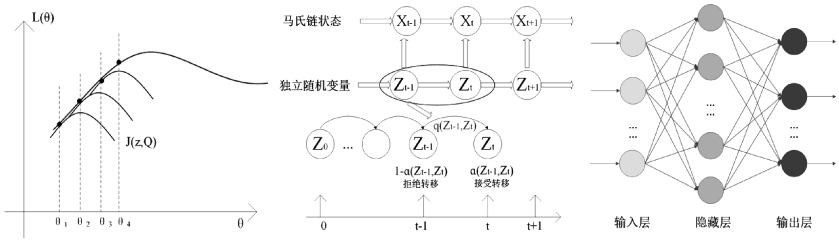
构建多层认知诊断框架，同时用“模糊集理论”建模联结型客观题和补偿型主观题，实现更精准的学习者知识状态表征。鉴于前两种模型存在依赖完备的Q矩阵问题，很难在真实的教育测评场景中实施与应用，因此智能时代的RCD（Gao, et al., 2021）继承了认知诊断思想与项目反应理论的优点，利用多级注意力网络实现学习者、试题与概念三层关系图的信息聚合，避开Q矩阵完备性难题，并采用认知诊断函数输出学习者认知状态，实现认知诊断对学习者认知状态的精准诊断，更适用于认知诊断模型在教育领域的实施与落地。

3. 技术赋能知识追踪方式变革

人工智能技术加速了深度知识追踪方式的变革，不论是采用新型深度学习技术，还是融入认知心理学特征，其本质仍以认知过程时序数据处理为核心，以预测学习者未来知识表现为目标，来探究学习者的认知演变机理。智能化的知识追踪遵循传统贝叶斯知识追踪基本原理，以学习者知识状态为隐藏状态，通过深度神经网络的隐藏层跟踪学习者的知识掌握状态的演化。基于学习与遗忘的动态认知诊断模型CF–DKD（Huang, et al., 2021），是基于我国多次连续测评场景，能够体现学习者认知过程的学习与遗忘规律的记忆神经网络，通过学习者的每次答题动态更新学习者的知识状态和认知过程，以此增强时序神经网络对知识的跟踪，实现对学习者未来表现更精准的预测。

（三）参数估计方法

兴起与发展阶段的测评模型均基于概率统计方法，其参数估计方法采用最大期望EM算法（Dempster, Laird, & Rubin, 1977）和马尔可夫链蒙特卡罗MCMC算法（Liu, et al., 2018）。智能化阶段通过深度学习模型的反向传播来更新神经网络参数，使目标损失函数达到最小值。三种参数估计方法如图4所示。



（a）EM参数估计方法　（b）MCMC参数估计方法　（c）神经网络参数估计方法

图4　三种主流参数估计方法对比

EM算法隶属迭代优化方法，适用于数据缺失情况，对初始值较为敏感（De La Torre, 2011）。MCMC算法是在概率空间中通过随机采样来估算参数的后验分布，具有良好的可扩展性，针对多参数的模型仍具有良好的效果，弥补了EM算法缺陷。然而，智能化测评模型中的参数估计采用反向传播计算，重点解决迭代过程中的梯度问题（Werbos, 1974），通过“前向传递输入”“反向传递误差”，误差由输出层反向传播，将误差分配给各层来修正各层权重，解决了隐层传播中权重值计算问题，在深度学习领域广泛使用。

（四）测评模型评估

鉴于项目反应与认知诊断理论中输出的学习者能力与知识掌握状态均为不可观测值，教育测评模型采用可观测的学习者表现来评估模型（Wang, et al., 2020; Gao, et al., 2021; Huang, et al., 2021）。从各模型实验结果的对比分析中，我们发现当前主流的智能化模型评估主要从“精确度”来评估模型，包括分类（ROC曲线下的面积AUC指标和准确性ACC）和回归（均方误差MSE、均方根误差RMSE）两个角度。发展阶段的测评模型在预测精确度方面优于兴起阶段的测评模型，而崛起阶段加入深度神经网络的智能化测评模型的预测精确度明显优于传统模型，证明了其预测结果具有较好的效度。基于教育和心理测量学理论，良好的测评模型需对其信效度进行系统检测（骆方, 等, 2021），然而，当前智能化测评模型对测评模型的区分度（将所测认知与其他区分开）、信度（预测结果的可靠性）和公平性（对不同群体的偏差）等方面的评估较为缺乏。

（五）测评分析结果的教育应用实践

根据教育测评模型输出的知识技能状态、能力等学习者认知水平信息，我们可将其应用在教育领域危机预警、资源推荐和路径规划等方面。第一，通过测评模型对学生的认知水平动态追踪，科学构建学习仪表盘，可实现危机学生早期预警。比如，阿吉拉尔开发的基于仪表盘的早期预警系统（Aguilar, Karabenick, Teasley, & Baek, 2021）在美国中西部某公立大学201名学生中展开实证研究，证明预警系统对学生学习动机和自我调节产生了重要影响。第二，通过对学习者认知状态的量化分析，

可实现向学习者自适应推荐试题，采用最少的试题达到测评认知水平的目标。比如，基于IRT实时计算学习者能力水平设计实现的试题提示的自适应推荐的脚手架系统（Ueno & Miyazawa, 2017），在某大学93名学生中开展实验，证明了系统的有效性与有用性，同时还探索了其最佳推荐的能力阈值。第三，根据认知水平的测评结果，可以为学习者自动规划学习路径（Shou, et al., 2020）。比如，通过认知诊断工具，利用K–means聚类分析方法来构建学习者的学习路径（Wu, Wu, Zhang, Arthur, & Chang, 2021），在TIMSS 2015数学测试中抽取726名学习者进行验证，有效构建了学习者的个性化学习进程，促进个性化学习。

四、人工智能时代教育测评发展趋势

（一）探索人工智能时代教育测评理论的价值内涵

基于传统教育测量与心理学理论，通过对经典教育测量理论、现代教育测量理论和新一代教育测量理论的深度剖析，结合学习者认知加工过程的客观规律，人工智能时代的教育测评应着重利用智能化技术挖掘学习者在“知识建构、认知发展和综合能力”方面的理论价值，如图5所示。基于美国能力本位理论（程新奎, 等, 2021），剖析“知识–认知–能力”内部认知机理，探索由知识“现象”到认知“本质”的学习者认知结构诊断，由“局部”知识到“整体”能力的学习者综合能力测评，以及由能力“结果”到认知“过程”的学习者认知过程剖析。以此来探索传统教育测评理论在人工智能时代的价值底蕴，实现“为学习而测评”的目标，为新时代的教育测评研究的开展提供理论指导。

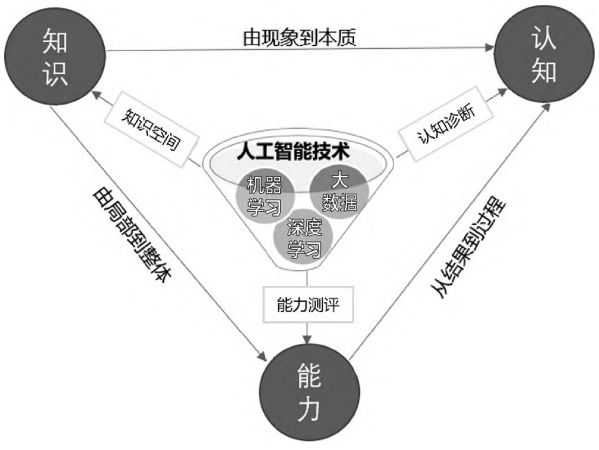


图5　教育测评认知理论框架

（二）基于多维知识空间的认知诊断

智能化教育测评模型，将依托学科领域知识图谱，结合记忆、语言、实践等丰富的内部认知刻画机理（Massa, et al., 2015），将知识图谱扩展到多维知识空间，融合项目、知识和学生等多维认知特征，强化知识、项目、学生两两之间的多对多复杂关联（Corbett, et al., 1994），全面剖析学习者认知结构本质属性。基于学科多维知识空间，学习者每次的答题均会生成一个独立的子空间，借助图神经网络算法来聚合当前知识空间中的邻居节点信息（图谱中先备、后继和关联关系）（黄涛, 等, 2015），实现当前知识点更全面的表征。针对单次规模化测评数据，可借助认知诊断函数获取学习者的当前知识状态（Gao, et al., 2021）；针对多次连续型测评数据，伴随时间的演变，借助时序神经网络来更新当前空间节点的状态（Nakagawa, Iwasawa, & Matsuo, 2019）。最终获取当前知识空间维度学习者的认知结构与认知状态，形成预警机制，帮助教师和学习者进行合理干预。因此，人工智能时代的认知诊断是在保留其可解释性优势的前提下，探索多种深度学习技术来改善其认知诊断的效果。

（三）面向时序数据的认知过程建模

当前主流教育测量模型大多基于单次测试结果数据对学习者知识水平进行静态评估，各阶段测评结果相对独立，仅实现对学习者当前知识水平的结果性评价，因此需重视学习者测评过程中认知结构随时间的变化过程（张生, 等, 2021）以及知识技能水平不断提升的规律。通过对以“时间序列分析”为核心的知识追踪模型发展史的深度对比分析，我们发现现有模型能够深度挖掘学习者知识状态在长周期时间序列上的演变规律。但由于学习者所掌握的知识会因遗忘和记忆因素而发生变化（Huang, et al., 2020），学生在学习过程中的知识内化与长期依赖的知识关联还需要深度挖掘。因此，人工智能时代的教育测评模型需充分利用时间序列上的动态知识诊断方法对学习者长周期测量数据进行关联建模，结合学习者对知识技能的潜在学习与遗忘规律，全方位持续监测随着时间演化的知识技能发展状况，客观刻画学习者深层次的知识建构和认知发展水平，提升时间序列模型在教育认知领域的可解释性。

五、结论与建议

智能化技术推动测评范式变革，教育测评经历了从传统概率统计向人工智能技术的跨越式变迁。智能时代的测评以“为了改进而测评”为核心理念，深度剖析学习者“知识–认知–能力”的内部认知机理，为长周期时序数据构建智能化测评模型，实现学习者全息认知刻画，促进教育高质量发展。目前，国外智能化测评的研究和实施主要集中在在线自适应学习场景，而国内基础教育由于测评数据获取困难、分析结果难以及时反馈等问题应用并不广泛。根据智能化测评的实施现状，本研究提出如下几点建议：

1. 探索智能化感知技术，支撑多场景测评数据的伴随式采集

基于数据挖掘与深度学习等技术的智能化测评模型需要依赖一定规模的数据集，以提高模型的精确性。然而，由于教育数据采集时存在增加教师负担、纸笔测试难以数字化、数字安全与隐私等问题，教育数据获取面临困难。各地方教育局应积极探索“产学研”合作，借助产业界先进的智能化感知技术，汇聚线上线下相结合的多空间、多场域的持续性测评数据，覆盖学习过程中的行为、心理与生理多模态数据，为形成性测评和诊断性测评等教育测评建模提供精准的特征表示（顾小清, 等, 2021）。通过多元测评数据的伴随式无感汇聚，在不增加师生负担的前提下提高数据获取的便利性，从根本上解决教育数据获取的困难，为教育测评提供海量数据支撑。

2. 增强教育场景建模，加快智能化测评模型的落地实施

鉴于智能化测评评估以精准度为主，无法保证信度与公平（骆方, 等, 2021），故当前绝大部分测评模型只是应用于在线自主学习场景。而且在国内主流纸笔测评场景中，由于行为特征少、数据稀疏等问题，不满足智能化测评的实施条件，以致智能测评模型在我国真实的教育场景中仍难以实施。针对形成性与诊断性测评等单次纸笔测评场景，借助智能化技术从规模化学习者测评过程数据中挖掘学习者的认知结构缺陷，提高其实验的信效度。针对多次测评场景，在汇聚的海量过程性数据基础上，通过对学习者的学习过程与遗忘过程建模，实现学习历程跟踪与监测，发现学习者的波动趋势，形成可靠的预警机制。同时，设计适应于智能化技术的新型测验效度检测方法，在保证认知水平诊断准确性的同时提高实验的实用性与可靠性，以此加快智能化测评模型在教育领域的落地实施。

3. 提高教师数据素养，促进测评数据驱动的精准化教学

随着教育大数据学情分析系统的推广与应用，学科知识与能力的伴随式诊断已成为教育界的热点。然而面对已有的学情分析数据，可能因为教师能力或时间有限的问题，只有少部分教师将学情分析数据用于指导课堂精准化教学。教师应该坚持“数据驱动决策”的核心理念，提高自身信息化教学能力与数据素养（冯晓英, 等, 2021）。加强数据素养在学科教学中的应用培训可从四个方面来展开（张进良, 等, 2021）：首先，提高教师对数据的意识，积极探索数据间的关联；其次，加强对数据的定位，确保能迅速定位至关键的数据分析；再次，提高数据分析与解读能力，能从数据中获取有用的信息来指导教学；最后，通过基于数据的反思，进行科学的决策，指导教学活动的设计与实施。最终实现以学习者学情为中心的个性化教学，促进精准化教学的有效实施，提高教育教学质量。

[参考文献]

程新奎，张瑾. 2021. 美国能力本位教育的新发展及其对我国远程开放教育的启示[J]. 中国远程教育(12)：28–37.

范涌峰，宋乃庆. 2019. 大数据时代的教育测评模型及其范式构建[J]. 中国社会科学(12)：139–155，202–203.

冯晓英，郭婉瑢，黄洛颖. 2021. 智能时代的教师专业发展：挑战与路径[J]. 中国远程教育(11)：1–8，76.

顾小清，李世瑾，李睿. 2021. 人工智能创新应用的国际视野——美国NSF人工智能研究所的前瞻进展与未来教育展望[J]. 中国远程教育(12)：1–9，76.

黄涛，施枫，杨华利. 2015. 知识地图模型及其在教学资源导航中应用研究. 中国电化教育(7)：73–78.

骆方，田雪涛，屠焯然，等. 2021，教育评价新趋向：智能化测评研究综述[J]. 现代远程教育研究，33(05)：42–52.

王立君，唐芳，詹沛达. 2020. 基于认知诊断测评的个性化补救教学效果分析：以“一元一次方程”为例[J]. 心理科学，43(06)：1490–1497.

张进良，李保臻. 2015. 大数据背景下教师数据素养的内涵、价值与发展路径[J]. 电化教育研究，36(07)：14–19，34.

张生，王雪，齐媛. 2021. 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(02)：1–8，16，76.

Aguilar, S. J., Karabenick, S. A., Teasley, S. D., & Baek, C. (2021). Associations between learning analytics dashboard exposure and motiva-

tion and self-regulated learning. *Computers & Education*, 162, 104085.

Bloom, B. S., Hastings, J. T. & Madaus, G. F. (1971) *Handbook on formative and summative evaluation of student learning*. New York, McGraw-Hill.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software, 48*(1), 1-29.

Cheng, S., Liu, Q., Chen, E., Huang, Z., Huang, Z., et al. (2019, November). DIRT: Deep learning enhanced item response theory for cognitive diagnosis. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (pp. 2397-2400).

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User *Modeling and User-Adapted Interaction, 4*(4), 253-278.

De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115-130.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

Dibello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment*, 361389.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Gan, W., Sun, Y., Ye, S., Fan, Y., & Sun, Y. (2019). Field-Aware Knowledge Tracing Machine by Modelling Students' Dynamic Learning Procedure and Item Difficulty: IEEE, 1045-1046.

Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., & Wang, M., et al. (2021). RCD: Relation Map Driven Cognitive Diagnosis for Intelligent Education Systems, 501-510.

Goldberg, M., & Wang, L. (1991). Comparative performance of pyramid data structures for progressive image transmission. *Ieee Transactions On Communications, 39*(4), 540-548.

Grossman, R., & Salas, E. (2011). The transfer of training: what really matters. *International journal of training and development, 15*(2), 103-120.

Huang, T., Yang, H., Li, Z., Xie, H., Geng, J., & Zhang, H. (2021). A Dynamic Knowledge Diagnosis Approach Integrating Cognitive Features. Ieee Access, 9, 116814-116829.

Huang, Z., Liu, Q., Chen, Y., Wu, L., Xiao, K., & Chen, E., et al. (2020). Learning or forgetting? a dynamic approach for tracking the knowledge proficiency of students. *ACM Transactions on Information Systems (TOIS), 38*(2), 1-33.

Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., & Su, Y., et al. (2019). Ekt: Exercise-aware knowledge tracing for student performance prediction. *Ieee Transactions On Knowledge and Data Engineering, 33*(1), 100-115.

Liu, Q., Wu, R., Chen, E., Xu, G., Su, Y., & Chen, Z., et al. (2018). Fuzzy cognitive diagnosis for modelling examinee performance. *ACM Transactions on Intelligent Systems and Technology (TIST), 9*(4), 1-26.

Mandinach, E. B., & Schildkamp, K. (2021a). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in Educational Evaluation, 69*, 100842.

Massa, M. S., Wang, N., Bickerton, W. L., Demeyere, N., Riddoch, M. J., & Humphreys, G. W. (2015). On the importance of cognitive profiling: A graphical modelling analysis of domain-specific and domain-general deficits after stroke. *Cortex, 71*, 190-204.

Mervis, J. 2016. NSF director unveils big ideas. American Association for the Advancement of Science.

Mislevy, R. J. (1993). Foundations of a new test theory: in N. Frederiksen, RJ Mislevy, and I. Bejar, eds., *Test Theory for a New Generation of Tests*. Hillsdale, NY: Erlbaum.

Nagatani, K., Zhang, Q., Sato, M., Chen, Y., Chen, F., & Ohkuma, T. (2019). Augmenting knowledge tracing by considering forgetting behavior, 3101-3107.

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: modeling student proficiency using graph neural network: IEEE, 156-163.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in education, 14*(2), 149-170.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., & Guibas, L. J., et al. (2015). Deep Knowledge Tracing. Advances in Neural Information *Processing Systems, 28*, 505-513.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Shou, Z., Lu, X., Wu, Z., Yuan, H., Zhang, H., & Lai, J. (2020). On learning path planning algorithm based on collaborative analysis of learning behavior. *Ieee Access, 8*, 119863-119879.

Sondergeld, T. A., Stone, G. E., & Kruse, L. M. (2020). Objective standard setting in educational assessment and decision making. *Educational Policy, 34*(5), 735-759.

Su, Y., Cheng, Z., Luo, P., Wu, J., Zhang, L., & Liu, Q., et al. (2021). Time-and-Concept Enhanced Deep Multidimensional Item Response Theory for interpretable Knowledge Tracing. *Knowledge-Based Systems, 218*, 106819.

Subali, B. (2021). The Comparison of Item Test Characteristics Viewed from Classic and Modern Test Theory. *International Journal of Instruction, 14*(1), 647-660.

Sympson, J. B. (1978). A model for testing with multidimensional items.

Ueno, M., & Miyazawa, Y. (2017). IRT-based adaptive hints to scaffold learning in programming. *Ieee Transactions On Learning Technologies, 11*(4), 415-428.

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., & Yin, Y., et al. (2020). *Neural cognitive diagnosis for intelligent education systems*, 6153–6161.

Werbos, P. (1974). *New tools for prediction and analysis in the behavioral sciences*. Ph. D. dissertation, Harvard University.

Wiliam, D. (2011). What is assessment for learning?. *Studies in Educational Evaluation, 37*(1), 3–14.

Wongwatkit, C., Srisawasdi, N., Hwang, G., & Panjaburee, P. (2017). Influence of an integrated learning diagnosis and formative assessment–based personalized web learning approach on students learning performances and perceptions. *Interactive Learning Environments, 25*(7), 889–903.

Wu, X., Wu, R., Zhang, Y., Arthur, D., & Chang, H. (2021). Research on construction method of learning paths and learning progressions based on cognitive diagnosis assessment. *Assessment in Education: Principles, Policy & Practice*, 1–19.

Yang, S., Zhu, M., Hou, J., & Lu, X. (2020). Deep knowledge tracing with convolutions. arXiv preprint arXiv:2008.01169.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models: Springer, 171–180.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262–286.

Zhang, J., Shi, X., King, I., & Yeung, D. (2017). Dynamic key–value memory networks for knowledge tracing, 765–774.

收稿日期：2021－09－17
定稿日期：2022－01－04

作者简介：杨华利，博士，讲师，武汉纺织大学计算机与人工智能学院（430200）。

耿晶，博士研究生，华中师范大学教育大数据应用技术国家工程研究中心（430079）。

胡盛泽，博士研究生，华中师范大学人工智能教育学部（430079）。

黄涛，博士，教授，博士生导师，本文通讯作者，华中师范大学教育大数据应用技术国家工程研究中心（430079）。

徐晨曦，本科生，华中师范大学人工智能教育学部（430079）。

责任编辑　郝　丹

（上接第67页）

cross–disciplinary perspective. Routledge.

Snow, E., Rutstein, D., Basu, S., Bienkowski, M., & Everson, H. T. (2019). Leveraging evidence–centered design to develop assessments of computational thinking practices. *International Journal of Testing, 19*(2), 103–127.

Vista, A., Awwal, N., & Care, E. (2016). Sequential actions as markers of behavioral and cognitive processes: Extracting empirical pathways from data streams of complex tasks. *Computers and Education, 92*, 15–36.

Voogt, J., & Roblin, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies, 44*(3), 299–321.

Wang, C. Y., Tsai, M. J., & Tsai, C. C. (2020). Predicting cognitive structures and information processing modes by eye–tracking when reading controversial reports about socio–scientific issues. *Computers in Human Behavior, 112*, 106471.

Watson, G., & Glaser, E. M. (1980). *Watson–Glaser critical thinking appraisal*. San Antonio, TX: Psychological Corporation.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Wolf, R., Zahner, D., Kostoris, F., & Benjamin, R. (2014). A case study of an international performance–based assessment of critical thinking skills. New York, NY: Council for Aid to Education.

Uto, M., Xie, Y., & Ueno, M. (2020). Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6077–6088), Barcelona, Spain(Online). International Committee on Computational Linguistics.

Zahner, D. (2014). *Reliability and validity of CLA+*. New York, NY: Council for Aid to Education.

Zlatkin - Troitschanskaia, O., Shavelson, R. J., Schmidt, S., & Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology, 89*(3), 468–484.

收稿日期：2022－03－18
定稿日期：2022－09－02

作者简介：姜力铭，博士研究生，北京师范大学心理学部（100875）；

刘玉杰，高级工程师，中国人民解放军海军招收飞行学员工作办公室（100071）；

骆方，博士，教授，博士生导师，本文通讯作者，北京师范大学心理学部（100875）。

责任编辑　单　玲