计算机与智能教育技术

2026年10月 第1卷第3期

人工智能教学技能评价研究进展

——聚焦教学语言、板书与教姿教态

姜彦非

江苏师范大学，江苏徐州，221116

摘　要 | 本文采用叙述性综述（按主题归纳和比较已有研究），梳理2022年至2026年10月人工智能辅助评价教学语言、板书和教姿教态的19篇研究，比较其样本、评价标准、验证方法及局限。纳入的研究中，语言评价涉及话语分类、表现评分和随机对照试验。板书研究主要识别书写行为、生成单字评语，对整幅板书结构及其与讲解配合的验证仍不足。教姿教态研究涉及手势识别、非言语表现评分和反馈试用。三项技能在评价对象、观察时长、所需课堂背景和人工评分标准上存在差异。这些差异可能解释研究为何侧重不同任务，但仍需检验。后续可借鉴训练与评估按教师分开、比较信息来源及设置对照试验的方法，并分别检验语言连贯性、板书内容关系和动作的情境适宜性。

关键词 | 人工智能；教学技能评价；教学语言；板书；教姿教态

Copyright © 2026 by the author(s).

This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

https://creativecommons.org/licenses/by-nc/4.0/

1 引言

教学语言、板书和教姿教态是课堂表达的重要部分，分别用于讲解与回应、呈现内容、指示与交流。评价既要看表达是否清楚，也要看是否有助于教学。板书内容少，可能是表达简洁；教师停顿，也可能是给学生留出思考时间。因此，仅凭次数、时长或动作幅度，难以判断技能质量。

人工智能可分析课堂录音、视频和书写图像，帮助教师回顾课堂表现。本文关注这些分析结果能否支持教学技能质量的判断。

下文按三项技能梳理研究，再比较评价对象、观察时长、课堂背景及评分标准，并区分研究结果与本文的方法建议。

2 研究范围与文献选取

2.1 三项技能的范围

教学语言指讲解、提问、追问和回应，涉及表达清晰度、语速、停顿及内容组织，不限于外语教学。板书指黑板、白板和电子板上的现场书写，关注字迹、内容结构及课堂运用。教姿教态包括站姿、手势、目光、表情和配合讲授的移动。本文不讨论学生语言测评、一般幻灯片制作，也不单独评价课堂秩序管理。

图1呈现从数据分析到技能质量判断的步骤。

作者简介：姜彦非，江苏师范大学研究生，研究方向为学科教学（物理）。

文章引用：姜彦非. 人工智能教学技能评价研究进展——聚焦教学语言、板书与教姿教态[J]. 计算机与智能教育技术, 2026, 1(3): 128-136.

https://doi.org/10.68070/2w1k5m19

识别话语、板面和动作特征后，还需结合课堂情境与评分规则解释结果。例如，教师指向板面时，需对照所指内容和同步讲解，判断指示是否清楚、时机是否恰当。综合评价三项技能，还需考察它们的配合。

图 1 三项教学技能从数据输入到质量判断的分析框架

Figure 1 Framework linking input data to quality judgments across three teaching skills

2.2 文献来源与分析方法

本文采用叙述性综述，按主题比较2022年1月1日至2026年10月公开的研究。开展两批公开网络补检，查找候选文献，再通过计算语言学论文库（ACL Anthology）、教育数据挖掘（EDM）会议论文库、机器学习研究论文集（PMLR）、期刊官网及机构论文存储平台核对全文。中文来源包括《内蒙古师范大学学报（自然科学版）》和《现代教育技术》等期刊网站。

中文检索式包括“教学语言评价”“人工智能”和“板书”“评价”“人工智能”“粉笔字”。英文检索式包括“automated feedback teacher classroom discourse evaluation”和“teaching gesture recognition automated body language assessment”。纳入期刊或会议发表的实证研究，要求至少涉及上述一项技能，或提供可用于比较的课堂评分方法。全文须能核对样本、评价标准和结果。排除学生能力测评、非教学场景行为识别、概念论述、新闻、课程页及一般书写教程。同一数据集上研究任务不同的论文可保留，但不重复计算样本。

合计19篇，其中15篇作为三项技能的核心证据，4篇用于方法比较。逐篇提取样本类型与数量、人工评价标准、数据划分方法、验证结果及局限，按技能和评价任务比较，不合并不同指标的数值。

3 教学语言评价研究进展

3.1 话语分类与语言表现评分

Suresh等[1]的TalkMoves[1]数据集含567份中小学数学课堂文本、174186个教师话语单元。教师话语分为6类教学行为及“未出现上述行为”类。研究从公开子集中划分训练与测试数据，两部分教师不重叠；部分非公开数据用于验证。模型输入教师的一句话及其前一条学生话语，最佳宏平均F1为0.7632。该指标反映话语分类表现，不能换算为教学语言质量得分。

Ilagan等[2]评价教师“鼓励学生之间互动”的表

[1] TalkMoves 为课堂话语分析数据集。训练数据用于学习，验证数据用于调整设置，测试数据用于最终检验。F1同时考虑查准率（判为某类的结果中有多少正确）和查全率（该类样本中有多少被找出）。两者都高，F1才高。宏平均 F1是各类别 F1的平均值。准确率是全部样本中分类正确的比例。

https://doi.org/10.68070/2w1k5m19

现。79名师范生参加155场模拟科学课堂讨论，其中62人的121场用于训练，17人的34场用于测试。人工评分为1—3分整数，其中56场由两人评分，取均分；组内相关系数（ICC）[2]为0.635。融合内容与互动结构的模型评分，与人工评分的相关系数r为0.678。研究仅涉及模拟讨论中的一项表现。

仅使用结构特征或内容特征时，模型评分与人工评分的r分别为0.631和0.547[2]。两种模型的分数预测方法也不同。要判断差异是否来自特征类型，应保持预测方法和数据条件一致。

人工转写文本上的模型表现，不能直接代表自动处理录音的表现。多人同时说话、专业词语识别错误，以及教师和学生话语混淆，都可能影响后续判断。

除话语分类外，研究也对课堂片段的语言表现评分。Hou等[3]分析德国92节数学课的367个片段，评价课堂话语特征、提问及解释三个维度。每段由两人按1—4分评分，取均分。训练与测试按教师分开，模型设置仅在训练数据内部选择。文本模型的平均二次加权κ（QWK）为0.356，加入音频后为0.384，再加入视频后为0.364。人工评分者间平均QWK为0.326。前者比较模型分数与人工均分，后者比较两名评分者的分数。比较对象不同，不能据此认定模型优于人工。增加视频也未使QWK数值继续提高。

3.2 模型分析能力与教师学习效果

Wang与Chen[4]将北京一所高校的60名师范生随机分入BERT[3]组或Llama 3组，各30人。两组使用相应模型参加约3小时的对话教学工作坊。工具测评中，经任务数据训练的BERT分类准确率为0.869；获得少量示例及分类理由的Llama 3为0.573。两者使用条件不同，不能据此认定某类模型普遍更好。

考虑培训前知识成绩的影响后，两组培训后的知识成绩差异未达到统计显著[4]，p=0.374[4]。研究未设置不使用工具的对照组，知识测验也不能替代实际授课评价。模型分类是否准确、师范生是否学会分析对话、自身教学语言是否改善，是三个需要分别检验的问题。

3.3 自动反馈能否改变课堂语言行为

Jacobs等[5]组织21名数学教师试用TalkMoves约四个月，记录233节课，并结合问卷、访谈和前后比较。统一换算为55分钟课时后，后半程每课比前半程平均多6.5次话语行为，但六类行为的增幅均未达到统计显著。研究未设置独立对照组。因此，教师的积极评价和行为次数上升，尚不足以证明反馈带来了稳定改进。

Demszky等[6]在在线编程课程中研究教师回应学生想法的行为，筛选后有918名授课者。研究随机决定哪些教师收到反馈邮件提醒，所有教师均可查看反馈。收到提醒的教师比对照组每小时多回应约0.60次，增幅约7%。对因提醒而使用反馈的教师，研究估计其每小时多回应约1.13次，相当于对照组均值的13.2%。后一估计只适用于这类使用者，不是整个实验组的平均提升。

后续在实体数学与科学课堂开展的试验[7]，将523名工具用户随机分组。按预先登记的条件筛选后，主要分析样本为224名教师、1509份录音。两组均获TeachFX课堂语音分析工具的常规反馈，实验组额外获得关于引导学生解释思路的提问反馈。实验组此类提问比对照组每小时多约4.61次，增幅约20%；学生发言时间、推理频次及教师回应等指标的组间差异不显著。两项试验的情境和反馈目标不同，不能把7%与20%的差别解释为工具效果逐年提高。

Wang与Demszky[8]从18个课堂片段中抽取36条GPT-3.5建议，由2名数学教师分工评价。约82%的建议涉及教师已采用的做法。两人共同评价的建议过少，未计算评分一致性指标。反馈评价还应检查建议是否符合课堂实况、切合教学目标并提供新信息。

随机对照试验已观察到部分提问或回应行为的增加[6,7]，但实体课堂试验未发现学生推理频次显著增加[7]。这些结果尚未说明语音是否更清晰、停顿是否更恰当，或讲解结构是否得到改善。

4 板书评价研究进展

4.1 书写动作识别提供过程线索

Ma等[9]使用105个课堂场景中的6660幅图像，识别6类教师行为，其中包含书写和朝向信息。5328幅用于训练，1332幅用于验证，以人工标注的类别

[2] ICC 即组内相关系数，用于评价评分者的一致程度，具体含义取决于计算方式；r 表示两组分数的相关程度。Fleiss κ衡量多名评分者的一致性，并扣除偶然一致的影响。二次加权κ（QWK）还考虑等级误差的大小。这些指标不能互相替代。

[3] BERT 用于提取文本特征；Llama 2、Llama 3、GPT-3.5和GPT-4可按提示生成文本。GPT-4o 还可处理图像等输入。本文比较的是各研究实际使用的模型版本和使用方式。

[4] p 值是在零假设（如无真实差异）及检验模型成立时，出现当前或更极端结果的概率。它不是“没有差异的概率”，也不表示效果大小。差异不显著，不等于两组效果相同。

和位置为依据。改进YOLOv7[5]后，整体mAP@0.50为0.962。这是所有类别的综合检测指标，不能当作板书识别准确率或板书质量分数。

Bueno等[10]分析195节小学数学与语言艺术课，含约164小时视频，其中68节有人工转写文本。研究逐秒标注24类活动，按课次划分数据。X-CLIP[6]较大版本的整体宏平均F1为0.577，“教师书写”的F1为0.501，见会议官网补充材料表7。但该类别还包括在纸张、板面及实物投影设备上书写或擦除，不能直接作为板书质量评价结果。

两项研究识别书写行为，未评价板面内容是否正确。文献[10]按课次划分训练与测试，两部分仍可能含有同一教师。分析讲解与板书的配合，还需将板面书写与其他书写行为区分开。

4.2 粉笔字规范性分析与评语生成

张伟等[11]从粉笔字图像中提取整字重心、倾斜、匀称性和笔画走势等特征，按特征名称与对应取值整理数据。人工编写的评价文本共9000条，其中7000条用于训练，测试与验证各1000条。模型根据字形特征数据生成评语。研究未报告书写者人数。

研究用BiGRU[7]结合多头自注意力生成评语，并通过方法比较和消融实验检验模型。BLEU-4为0.68、ROUGE为0.75、METEOR为0.67[11]。这些指标反映生成评语与参考文本的接近程度，尚不能验证从板面图像到教学质量评分的完整流程。

该研究未报告人工评语的一致性，也未明确同一书写者是否出现在训练和测试中。后续应对照原字形，由独立评价者检查评语是否准确，再考察新书写任务中的改进。

4.3 整幅板书评价仍需补充证据

单字书写规范，不等于整幅板书具有较高的教学价值。本文纳入的研究主要涉及书写识别和粉笔字评语生成，对板书结构及其与讲解配合的验证仍不足。

评价板书与讲解的配合，需要保留书写过程。可将录音、板面变化和教师动作按时间对应，考察内容是否正确、组织是否合理、呈现是否及时。仅看最终的板面照片，难以了解擦除、补写和逐步推导的作用。这类评价方法仍需验证。

5 教姿教态评价研究进展

5.1 手势识别及其教学含义

Chen等[12]用400条教师骨架动作序列[8]，识别随意性（无明显意图）、指示性（指向内容）、描述性（配合描述）和操作性（如书写）4类手势。序列随机分组，进行五折交叉验证，准确率为93.5%。教师人数及训练与测试是否含同一教师未充分说明，对新教师的适用性仍待检验。

Xiong等[13]使用含2908个片段、8类手势的TBU-G[9]数据集。研究用GPT-4o生成手部姿态描述，再结合视频特征分类。基础模型准确率为86.4%；加入文字辅助方案后为92.8%，再加入时间变化信息后为93.7%。文字辅助方案同时改变特征融合与分类方式，增幅不能只归因于文字描述。

两项研究的数据、类别数和验证条件不同，不能仅凭93.5%与93.7%判断方法优劣。Xiong等使用的是动作描述，而非课堂讲解的转写文本；该研究也未检验手势与讲解配合的教学效果。

5.2 非言语表现的评分标准与验证

Lazarides等[14]研究非言语亲近性[10]，视频涉及46名教师、135节课。模型开发最初使用403个30秒片段，训练324个、验证79个。其中3个片段因画面质量不足，未纳入整体人工评分，仅用于评价局部表现的子模型。剔除人工评分一致性不足的片段后，非言语亲近性模型实际使用316个训练片段。训练与验证按教师分开。

3名评分者用连续滑动量表评分，人工评分者间ICC约0.68。在验证集上，模型与3名人工评分者的结果共同计算得到ICC为0.690[14]。它衡量评分的一致程度，与准确率及相关系数r不同。研究另加220个片段，共623个片段用于关联分析。非言语亲近性与学生感知的教学热情正相关，但考虑学生初始兴

[5] YOLOv7用于定位并识别图像对象。mAP 是各类别平均精度的平均值，用于概括检测表现。@0.50表示位置匹配阈值：预测框与人工框的重叠面积占合并面积的比例至少为0.50。正确检出还要求类别匹配。mAP 不是单类准确率。

[6] X-CLIP 结合视频与文字信息识别活动。“语言艺术课”主要涉及阅读、写作等内容。

[7] BiGRU 即双向门控循环单元，从前后两个方向分析序列。多头自注意力让模型从不同角度分析输入信息之间的关系。BLEU-4、ROUGE 和 METEOR 比较生成文本与参考文本是否接近。消融实验通过去掉模型中的某个组件，判断该组件是否有效。

[8] 骨架动作序列记录人体关节位置随时间的变化。五折交叉验证将数据分成五份，每轮用四份训练、一份测试。按教师划分时，同一教师不能同时出现在训练与测试中。嵌套验证仅在每轮训练数据内部选择模型设置。

[9] TBU-G 是教师行为理解数据集的手势扩展版本。该研究的文本用于描述手部姿态和手势类别，不是课堂讲解的转写文本。

[10] 非言语亲近性指教师通过动作、目光和表情等拉近与学生距离的表现。连续滑动量表是在一条连续刻度上移动标记给出分数。

https://doi.org/10.68070/2w1k5m19

趣后，对六周后兴趣的直接预测未达到统计显著。评分接近人工，并不能证明学生兴趣有所提高。

研究因人工评分一致性较低，排除了“姿态放松程度”指标[14]。目光、表情和身体距离的评分规则需结合具体情境制定。某种姿态或微笑频率不宜作为通用的优质标准。

研究优先选取教师正面讲授、身体完整可见且镜头移动较少的片段[14]，结果主要适用于这些条件。遮挡、转身或走出镜头不必然是教学缺陷；证据不足时，系统应暂缓判断。

5.3 自动反馈后的变化如何解释

Dimitriadou与Lanitis[15]让9名教育工作者完成两次短讲授。两次之间，参与者查看系统反馈，并听取研究人员的解释。系统按预设界限对表情、手势、姿态、身体活动和语音评分。其中，系统用头部朝向估计目光接触，只能近似反映教师的注视方向，难以确定其注视对象和表达意图。

Wilcoxon符号秩检验[11]显示，综合评分前后比较的p=0.011；语音和身体活动的p分别为0.317和0.069，均未达到统计显著[15]。研究样本小、无独立对照，前后均由同一系统评分。评分变化可能包含练习或熟悉规则的影响，不能据此认定各项教姿教态均有改善。

教姿教态反馈需结合动作发生的情境。例如，面向板面可能是为了讲解推导。表1汇总核心研究。

表 1 三项技能核心研究的样本、标准、验证与局限

Table 1 Samples, criteria, validation, and limitations of core studies

研究	样本与情境	评价标准或对象	验证方法与主要结果	主要局限
Suresh 等[1]	全语料567份课堂文本；模型实验使用公开子集	7类教师话语标签	训练与测试教师不重叠；输入教师话语及前一条学生话语；宏平均F1=0.7632	行为分类，不是语言质量评分
Ilagan 等[2]	79名师范生；155场模拟讨论	鼓励学生互动；1—3分，双人评分取平均	训练62人、测试17人；56场双人评分 ICC=0.635；融合模型 r=0.678	单一维度；模型间分数预测方法不同
Hou 等[3]	92节数学课；367个片段	话语特征、提问与解释；1—4分取双人均值	按教师进行嵌套五折验证；文本、加音频、再加视频的 QWK 为0.356、0.384、0.364	单一国家与数学主题；未验证反馈干预
Wang 与 Chen[4]	60名师范生；BERT 与 Llama 3组各30人	教师话语分类与对话教学知识	BERT 准确率0.869；Llama 3为0.573；控制培训前成绩后，培训后知识成绩组间 p=0.374	模型使用条件不同；缺少不用工具的对照组
Jacobs 等[5]	21名数学教师；233节课；约四个月	6类话语行为；问卷与访谈	前后比较；后半程每55分钟比前半程多6.5次；六类增幅均不显著	无独立对照；行为趋势不等于稳定改进
Demszky 等[6]	在线编程课程；筛选后918名授课者	回应学生想法的行为频次	随机邮件提醒；相对对照组每小时多约0.60次，约7%	提醒与实际使用不同；不代表全部语言技能
Demszky 等[7]	随机分配523名用户；主要分析224人、1509份录音	引导学生解释思路的提问	随机对照试验；主要样本相对对照组每小时多约4.61次，约20%	分析样本经过筛选；未据此证明学生成绩提高
Wang 与 Demszky[8]	18个课堂片段；36条建议	建议是否符合实况、切合教学目标、可实施并提供新信息	2名数学教师分工评价；约82%涉及已采用做法	两人共同评价的建议过少，未计算评分一致性；仅使用早期模型
Ma 等[9]	105个课堂场景；6660幅图像	6类教师行为，包含书写	5328幅训练、1332幅验证；整体mAP@0.50=0.962	不是板书单类指标；未明确按教师分开
Bueno 等[10]	195节课；约164小时视频	24类活动；书写不限于板书	按课次划分；整体宏平均F1=0.577；书写 F1=0.501（补充材料表7）	未明确按教师分开；未评板书质量
张伟等[11]	9000条粉笔字评价文本	字形特征数据转为评语	7000训练、1000测试、1000验证；BLEU-4=0.68、ROUGE=0.75、METEOR=0.67	未明确书写者人数；无技能改进检验
Chen 等[12]	400条教师骨架动作序列	4类教学手势	将动作序列随机分组，进行五折交叉验证；准确率93.5%	教师人数及训练与测试是否包含同一教师未说明

[11] Wilcoxon 符号秩检验用于比较同一组对象的前后测量。它先按变化幅度排序，再结合变化方向进行检验。

续表1

研究	样本与情境	评价标准或对象	验证方法与主要结果	主要局限
Xiong 等[13]	2908个视频片段	8类教学手势	方法比较与消融实验；完整模型准确率93.7%；文字辅助方案同时改变特征融合与分类方式	文本是图像描述；训练与测试是否包含同一教师未说明；未评价教学质量
Lazarides 等[14]	46名教师、135节课的视频；初始403片段	非言语亲近性；连续滑动量表	训练与验证按教师分开；316个训练片段；验证集上模型与3名评分者的共同 ICC=0.690	ICC 不是准确率；控制学生初始兴趣后，对六周后兴趣的直接预测不显著
Dimitriadou 与 Lanitis[15]	9名教育工作者；两次短讲授	按预设界限评价身体语言与语音	比较同组对象的前后评分；综合评分 p=0.011；部分分项差异不显著	小样本、无独立对照；前后由同一系统评分

6 核心比较与后续研究方向

6.1 评价单位与时间跨度

评价单位（一次标注或评分对应的对象）不同于模型实际接收的信息范围。单句话语分类也可能使用相邻对话[1]。语言研究还对整段讨论或课堂片段评分[2,3]；板书研究使用图像、书写片段或字形特征数据[9,10,11]；教姿教态研究分析动作序列或短视频[12,13,14]。判断一句话是否属于追问，与判断一段讲解是否连贯，所需观察时长并不相同。

Wang等[16]从41名教师的561段视频中选取50段，每段抽取连续50句，共标注2500句。4名评分者的Fleiss κ为0.712。不加入前句信息时，模型分类准确率为0.7864；根据分类结果得到的教学支持评分，与人工评分的r仅为0.1197。这表明，在该研究采用的汇总方法下，逐句分类结果尚不能可靠地反映片段质量。

Whitehill与LoCasale-Crouch[17]在同一批41名教师的561段视频上，按教师进行五折验证。模型分别使用词频特征、Llama 2提取的话语行为特征及两者组合，预测教学支持评分。预测分数与人工评分的五折平均 r 依次为0.47、0.42、0.48。两项研究的方法不同，不能把结果差异仅归因于特征类型；数据来自同一批教师，也不构成独立样本上的重复验证。

6.2 情境依赖与人工评分标准

情境依赖指作出判断需要借助课堂背景的程度。本文建议分别保留学生在教师回应前的发言、板书中的概念关系与呈现顺序，以及手势所指对象和当前任务。人工评分标准是否完善，要看规则是否明确、评分结果是否一致，以及能否用于新情境。语言研究已有分级规则[2,3]；粉笔字评语研究未报告人工评语的一致性[11]；非言语评分研究因一致性不足排除了“姿态放松程度”[14]。

王梦珂等[18]分析10节小学数学教学比赛录像。3名专家采用五级评价，专家间Fleiss κ为0.573；系统与专家综合评分的r为0.778。前者衡量专家评分的一致程度，后者衡量系统分数与专家分数的相关程度。专家未获得系统各维度的具体指标和分级规则，因此仍需核对双方是否依据相同标准评分。整体互动评分也不能替代三项技能的分别评价。

话语类别较明确时，可以逐句标注；评价整幅板书，需要观察内容关系如何逐步形成；评价教姿教态，则需理解动作在当前教学中的作用。评价对象及所需背景信息的差异，可能解释三项技能研究为何侧重不同任务。这是本文提出的解释，尚需专门验证。后续可在同一任务中改变观察时长和背景信息，比较人工评分一致性与模型表现。

6.3 可借鉴的方法与分项验证

Hou等[19]以92节课的367个片段评价“鼓励与温暖”。多层感知机[12]、零样本方式下的GPT-4及两者加权组合，五折平均r依次为0.441、0.341、0.513；数据按课次划分。组合方法在该任务中的相关系数较高，而文献[3]加入视频后QWK数值下降。因此，后续应比较单一信息来源与组合方案，检验增加信息是否改善评分。文献[3,14,19]的数据来自相关项目，不能作为独立样本累加。

将训练数据与评估数据按教师分开[3,14]，可检验模型能否适用于新教师。反馈研究可借鉴语言研究的随机分组设计[6,7]，但评价目标须分别确定：语言关注解释与回应是否恰当，板书关注内容关系及呈现时机，教姿教态关注动作是否帮助表达。这些做法能否用于其他技能，仍需实际研究检验。表2并列比较三项技能。

[12] 多模态方法同时使用音频、视频、文本等信息。多层感知机通过多层计算，把输入特征转成预测结果。零样本方式指提示中没有提供带有答案的任务示例。

https://doi.org/10.68070/2w1k5m19

表 2 三项技能的评价对象、观察范围与评分依据比较

Table 2 Comparison of assessment units, observation scope, and rating criteria across three teaching skills

比较方面	教学语言	板书	教姿教态
评价单位	单句话语、讨论或课堂片段[1,2,3]	图像、书写片段、字形特征数据[9,10,11]	动作序列、手势片段及非言语表现片段[12,13,14]
观察时长	单句分类也使用相邻话语[1]；片段评分需观察更长过程[3]	静态图像或连续书写活动[9,10]	动作持续过程或30秒片段[12,14]
人工标准	类别标签及特定维度的分级规则[1,2,3]	行为类别与人工评语；评语一致性未报告[9,10,11]	手势类别及连续量表；部分指标因一致性不足被排除[12,13,14]
补充建议	保留学生此前的发言及教师随后的回应	保留概念关系、书写顺序及同步讲解	保留动作所指对象、教学任务及同步话语

注：前三个比较方面根据纳入的研究归纳；“补充建议”是本文提出的后续研究建议。

图2列出三类评价结果的验证要求。除核对指标，还应回看话语、板面或动作。对生成评语，应检查内容是否准确、有无新信息，以及重复分析同一材料时结果是否稳定。

图 2 三类评价结果的验证方法与反馈效果检验

Figure 2 Validation of three types of assessment results and evaluation of feedback effects

注：“人工一致性”指人工评分的一致程度；“独立评分”指由未参与系统开发、也未提供反馈的评价者评分；“延迟测量”指间隔一段时间后的复测。虚线框标出仍需加强的效果验证，箭头表示先后步骤，不表示因果关系。

检验反馈效果时，前后教学任务应具有可比性，并由不了解分组情况的评价者评分。研究应报告教师的初始水平、变化幅度、估计结果可能有多大误差，以及参与者中途退出的情况。间隔一段时间后再测，考察改进能否保持。板书与教姿教态仍需补充这类验证。

7 结语

纳入的研究显示，语言评价已涉及分类、评分和反馈的随机对照试验；板书研究主要识别书写行为、生成局部评语；教姿教态研究涉及动作识别、非言语评分及小规模反馈试用。评价对象、观察时长、课堂背景和人工评分标准的差异，可能解释这些研究为何侧重不同任务，但这一解释仍需验证。后续应分别制定评分规则、选择验证方法，并在新教学任务中检验反馈能否促进改进。

参考文献

[1] SURESH A, JACOBS J, HARTY C, et al. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student

discursive moves[C/OL]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille: European Language Resources Association, 2022: 4654-4662[2026-10-14]. https://aclanthology.org/2022.lrec-1.497/.

[2] ILAGAN M J, BEIGMAN KLEBANOV B, MIKESKA J N. Automated evaluation of teacher encouragement of student-to-student interactions in a simulated classroom discussion[C/OL]//Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications. Mexico City: Association for Computational Linguistics, 2024: 182-198[2026-10-14]. https://aclanthology.org/2024.bea-1.16/.

[3] HOU R, BÜHLER B, FÜTTERER T, et al. Multimodal assessment of classroom discourse quality: A text-centered attention-based multi-task learning approach[C]//Proceedings of the 18th International Conference on Educational Data Mining. Palermo: International Educational Data Mining Society, 2025: 241-253.

[4] WANG D, CHEN G. Evaluating the use of BERT and Llama to analyse classroom dialogue for teachers' learning of dialogic pedagogy[J]. British Journal of Educational Technology, 2025, 56(6): 2671-2704.

[5] JACOBS J, SCORNAVACCO K, HARTY C, et al. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change[J]. Teaching and Teacher Education, 2022, 112: 103631.

[6] DEMSZKY D, LIU J, HILL H C, et al. Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course[J]. Educational Evaluation and Policy Analysis, 2024, 46(3): 483-505.

[7] DEMSZKY D, LIU J, HILL H C, et al. Automated feedback improves teachers' questioning quality in brick-and-mortar classrooms: Opportunities for further enhancement[J]. Computers & Education, 2025, 227: 105183.

[8] WANG R, DEMSZKY D. Is ChatGPT a good teacher coach? Measuring zero-shot performance for scoring and providing actionable insights on classroom instruction[C]//Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications. Toronto: Association for Computational Linguistics, 2023: 626-667.

[9] MA L, ZHOU T, YU B, et al. Improving YOLOv7 for large target classroom behavior recognition of teachers in smart classroom scenarios[J]. Electronics, 2024, 13(18): 3726.

[10] BUENO I, HOU R, BÜHLER B, et al. Exploring automated recognition of instructional activity and discourse from multimodal classroom data[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l.]: IEEE, 2026: 6806-6817.

[11] 张伟, 李成城, 邬友, 等. 一种融合多头自注意力机制的粉笔字书写规范性评价方法[J]. 内蒙古师范大学学报（自然科学版）, 2024, 53(3): 321-330.

[12] CHEN Z, HUANG W, LIU H, et al. ST-TGR: Spatio-temporal representation learning for skeleton-based teaching gesture recognition[J]. Sensors, 2024, 24(8): 2589.

[13] XIONG Y, CHEN S, CAI T, et al. Natural language-driven teacher gesture recognition[C]//Proceedings of the 18th International Conference on Educational Data Mining. Palermo: International Educational Data Mining Society, 2025: 482-488.

[14] LAZARIDES R, FRENKEL J, PETKOVIĆ U, et al. "No words" — Machine-learning classified nonverbal immediacy and its role in connecting teacher self-efficacy with perceived teaching and student interest[J]. British Journal of Educational Psychology, 2025, 95(S1): S15-S31.

[15] DIMITRIADOU E, LANITIS A. Evaluating the impact of an automated body language assessment system[J]. Education and Information Technologies, 2025, 30(3): 3509-3539.

[16] WANG J, HANKOUR K, ZHANG Y, et al. Classroom observation evaluation with large language models[C/OL]//Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop (Proceedings of Machine Learning Research, Vol. 273). [S.l.]: PMLR, 2025: 83-93[2026-10-17]. https://proceedings.mlr.press/v273/wang25a.html.

[17] WHITEHILL J, LOCASALE-CROUCH J. Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback[J]. Journal of Educational Data Mining, 2024, 16(1): 34-60.

[18] 王梦珂, 陈增照, 石雅文, 等. 智能技术支持的多模态互动教学评价框架设计与应用效果[J]. 现代

教育技术, 2024, 34(9): 91-101.

[19] HOU R, FÜTTERER T, BÜHLER B, et al. Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and ChatGPT[C]//OLNEY A M, CHOUNTA I A, LIU Z, et al. Artificial Intelligence in Education (Lecture Notes in Computer Science, Vol. 14829). Cham: Springer, 2024: 60-74.

Research Progress in AI-Based Assessment of Teaching Skills: Instructional Language, Boardwork, and Nonverbal Teaching Behaviors

Jiang Yanfei

Jiangsu Normal University, Xuzhou 221116

Abstract: This narrative review examines 19 studies on artificial intelligence (AI)-assisted assessment of instructional language, boardwork, and nonverbal teaching behaviors published from 2022 to October 2026. It compares their samples, assessment criteria, validation methods, and limitations. Among the included studies, instructional language assessment covers discourse classification, performance ratings, and randomized controlled trials of feedback. Boardwork studies mainly recognize writing behavior and generate comments on individual handwritten characters. Evidence remains limited for assessing whole-board organization and its coordination with spoken explanations. Research on nonverbal teaching behaviors includes gesture recognition, performance ratings, and feedback pilots. The three skills differ in assessment targets, observation periods, required classroom context, and human-rating criteria. These differences may help explain why studies focus on different tasks, although this explanation requires testing. Future research can draw on teacher-independent training and evaluation data, comparisons of information sources, and controlled feedback trials. It should separately examine discourse coherence, conceptual relationships in boardwork, and the contextual appropriateness of nonverbal actions.

Keywords: artificial intelligence; teaching skill assessment; instructional language; boardwork; nonverbal teaching behaviors