外科技能评价工具在医学教育中的演进：从客观结构化临床技能评估到人工智能辅助评价

赵劲歌
四川大学华西医院，四川 成都 610041
DOI: 10.61369/ETR.2026050033

摘　　要： 外科技能培养长期依赖师承经验传授。随着住院医师规范化培训与能力本位教育的深化，仅凭主观印象已无法满足公平、可追溯与可改进的评价需求。客观结构化临床技能评估（OSATS）的出现，标志着外科技能评价从经验判断转向科学测量。其核心是将复杂操作分解为可观察的行为维度，为教学反馈提供共同语言。随后，程序特异性评价、工作场所评估、模拟训练及视频评价等方法不断发展，推动评价体系从单一量表扩展为多情境、多数据源的复合系统。近十年来，计算机视觉与深度学习等技术融入外科教育，形成人工智能（AI）辅助评价新范式。该范式旨在通过视频或运动学数据实现自动化分析，以提供更即时、一致的反馈。然而，其广泛应用仍面临金标准缺失、模型泛化能力不足、可解释性欠缺及数据治理模糊等挑战。本文围绕外科技能评估工具的演进脉络，梳理从 OSATS 到 AI 辅助评价的证据基础与现实瓶颈，并结合课程与质量管理需求提出实施建议，以期为外科教育评价体系的更新提供参考。

关 键 词： 外科教育；技能评价；客观结构化临床技能评估；视频评估；人工智能

Evolution of Surgical Skill Assessment Tools in Medical Education: From Objective Structured Assessment of Technical Skills to AI-Assisted Evaluation

Zhao Jin'ge
West China Hospital, Sichuan University, Chengdu, Sichuan 610041

Abstract: The cultivation of surgical skills has long relied on apprenticeship-based experiential teaching. With the advancement of standardized residency training and competency-based medical education, subjective impressions alone can no longer satisfy the demands for fair, traceable, and improvable assessment. The emergence of Objective Structured Assessment of Technical Skills (OSATS) marked a shift in surgical skill evaluation from experience-driven judgment to scientific measurement. Its core principle is to deconstruct complex procedures into observable behavioral dimensions, thereby providing a shared language for instructional feedback. Subsequently, procedure-specific rating scales, workplace-based assessments, simulation-based training, and video-based evaluation have continued to evolve, expanding assessment frameworks from single instruments into multi-context, multi-source composite systems. Over the past decade, computer vision and deep learning technologies have been increasingly integrated into surgical education, giving rise to a new paradigm of artificial intelligence (AI) - assisted assessment. This paradigm aims to enable automated analysis through video or kinematic data, providing more immediate and consistent feedback. However, its widespread implementation remains challenged by the lack of gold-standard labels, limited model generalizability, insufficient interpretability, and uncertainties in data governance. This article reviews the developmental trajectory of surgical skill assessment tools, synthesizing the evidence base and practical barriers from OSATS to AI-assisted evaluation. It further offers recommendations for implementation aligned with curricular and quality-management needs, with the goal of informing updates to surgical education assessment systems.

Keywords: surgical education; skill assessment; objective structured assessment of technical skills (OSATS); video-based evaluation; artificial intelligence

引言

外科技能的核心内涵，不仅在于能否完成手术操作，更在于能否在可控的风险范围内，稳定、高效地执行关键步骤，并能够合理解释每一步操作背后的临床决策逻辑。因此，一套完善的技能评价体系，必须超越对动作规范性与步骤完整性的简单核查，而应将操作者

的决策质量、应变能力、团队协作意识以及对患者安全的整体考量，纳入一个连贯的评价逻辑链条之中。对于外科培训体系而言，评价绝非一个附属或末端环节，而是串联课程设计、带教实施、学习进程与准入考核的核心枢纽。唯有当评价本身做到可度量、可比较、可反馈时，整个技能训练过程才能形成有效的教育闭环，驱动教学质量的持续提升。

当前，外科技能训练与评估领域并不缺乏工具，但值得关注的是，真正经过严格效度验证、且能在不同机构间稳定复现的评价方案仍然有限。在拥有评估工具与拥有足够证据支持该工具能够产生积极教育后果之间，仍存在显著差距[1]。这一矛盾在我国住院医师规范化培训的复杂场景中尤为凸显。培训需要同时在多学科、多层级中并行推进，既要利用模拟环境为学员提供高频次、低风险的形成性反馈，又必须在真实的临床工作中，完成对学员在可控风险下独立胜任能力的最终确认。因此，系统梳理评估工具的演进逻辑，深入理解不同工具的适用场景与其证据支持的边界，并审慎思考人工智能技术应如何融入现有体系，对于提升我国外科培训的同质化水平、最终保障患者安全，具有重要的现实意义。

一、从经验判断到结构化评价：OSATS 的奠基意义

（一）分解技能：构建可观察与可评分的行为维度

OSATS 的历史性贡献，远不止于提供了一张评分表格，更在于它引入了一种结构化的教育测量思想。它通过两种核心组件——针对具体操作步骤的核查清单与评价整体操作质量的全局评分量表——将原本笼统的技能评判分解为一系列可观察、可定义的关键行为。这种方式使不同的评分者能够在一个相对一致的观察框架下进行判断，显著提升了评价结果的可比性与可追溯性[2]。时至今日，在外科专业的技术培训中，OSATS 所奠定的结构化评估思路，依然常被作为课程设计、评分者培训与教学效果验证的底层框架。它首次为外科教学提供了一种将隐性知识显性化的通用语言，使技能变得可见、可讨论。

（二）聚焦反馈：OSATS 在形成性评价中的核心价值

在外科技能训练中，评价的首要目的往往并非为了区分优劣或评定等级，而是为了精准识别学员的可改进之处，并据此规划下一阶段的训练任务。正是在形成性评价的语境下，OSATS 的优势得以充分发挥[2]。一方面，它将学员需要改进的方面，明确锚定到诸如组织处理轻柔度、器械操作精确性、流程效率与计划性等具体维度上，使反馈内容具体而微。另一方面，它为带教老师与学员搭建了沟通的桥梁，提供了指向明确的共同术语，有效减少了基于模糊感觉的无效交流，提升了教学互动的效率与深度。在国内住院医师规范化培训的实践中，已有探索将 OSATS 与其他评估工具结合，用于腹腔镜技能培训的过程性评估与阶段考核。这些实践表明，结构化量表有助于推动教学模式从传统的讲授与观摩，转向训练、评估、反馈、再训练的良性动态闭环，从而加速学习进程[3]。

（三）通用与专用的平衡：认识 OSATS 的应用边界

OSATS 的普适性既是其得以广泛传播的优势，也构成了其内在的局限性。作为通用框架，它难以完全覆盖某一特定手术术式中独有的关键风险点、核心技术要点以及那些绝对不容有失的操作环节。随着外科培训的目标从掌握普适性基本功进阶到具备独立完成特定手术的胜任力，单纯依靠 OSATS 已显不足。因此，当前的发展趋势是，越来越多的培训课程将 OSATS 作为评价通用技能的基线框架，同时在此基础上，叠加针对具体术式的、高度特异性的步骤核查清单、错误分类体系以及关键决策点评估。这种组合策略更适用于贴近真实临床风险的高阶准入评估。国内围绕模拟教学与分层培训的研究也印证了这一趋势：OSATS 常用于评估基础动作的规范性，而程序特异性指标则更适用于检验学员在模拟或真实临床情境中做出正确决策并安全完成核心操作的综合能力[4]。

二、工具谱系的扩展：从量表到多场景、多源数据的融合评价

（一）视频评价与程序特异性量表：手术过程的可视化与精细化分析

腹腔镜与机器人手术的普及，使得记录完整手术过程的高清视频成为天然且宝贵的教学与评价资源。与依赖瞬时记忆的当场观察相比，视频评价具有多重优势：支持反复回看、允许多名评分者独立异步评分、便于建立评分者一致性培训的案例库，同时也为培训资源分布不均的地区开展远程督导与同质化考核提供了可能。多项研究致力于将系统性的视频回顾与结构化量表相结合，例如在腹腔镜培训中采用 GOALS 量表。这些程序相关量表补充了 OSATS 对微创手术特有维度，如双手协调、深度感知与视野维持等能力的关注，从而使评价更贴合腔镜手术的实际能力要求[3]。

然而，必须清醒认识到，视频评价本身并不等同于完全客观。其评价结果的可靠性仍然高度依赖于几个关键因素：评分者是否经过充分培训、视频截取策略是全程还是仅关键片段、评分标准的清晰度与可操作性，以及视频本身无法传递的信息，如组织触觉反馈、牵拉张力感知等所造成的信息缺失。因此，在许多机构的成熟实践中，视频评价常被定位为形成性反馈与周期性质量改进的核心工具；而在涉及晋级、认证或高风险操作准入的总结性考核中，则倾向于采用视频评价、现场直接观察、模拟器考核及并发症日志审查等多源证据相互印证的组合策略。

（二）融入临床：工作场所评估的现实与挑战

工作场所评估，如操作技能直接观察评估或迷你临床演练评

估，强调在真实的临床工作场景中，对学员的日常表现进行直接观察与即时反馈。其最大优势在于生态效度高，评价内容与真实的临床决策、患者管理及团队协作无缝衔接。然而，其实施难点也显而易见：观察机会因病例而异、不稳定；不同病例的复杂程度差异巨大，影响可比性；不同带教老师的评分标准可能不一致，影响信度。

国内已有研究探索将同伴互助学习模式与操作技能直接观察评估相结合，用于神经外科住院医师的培训。该实践提示，在引入工作场所评估时，对评价者进行系统培训、对评估流程进行标准化管理，例如规定最低评估频次、使用统一的记录与反馈模板、明确核心观察要点，对于维持教学效果的稳定性和有效性至关重要[5]。需要特别指出的是，工作场所评估能否发挥其应有价值，很大程度上取决于管理设计与教育文化。若仅将其视为一项必须完成的行政任务，则极易流于形式；唯有将其与清晰的能力里程碑体系、个性化的学习发展计划以及病例难度分级制度有机结合，它才能真正转变为促进学员在真实世界中持续学习和成长的有力工具。

（三）模拟训练的深化：从提供练习到实现精准达标

各类模拟训练平台，包括基础干箱、虚拟现实模拟器、动物实验等，为外科医师早期技能学习提供了一个安全、可重复、低风险的环境，同时也为实施高度标准化的技能考核创造了理想条件。国内研究在妇科腹腔镜模拟教学中，引入了基于 Miller 金字塔能力模型的分层教学模式，并运用 OSATS 作为核心考核工具之一。结果表明，这种将分层递进的训练路径与结构化的评价工具相结合的方式，显著增强了培训过程的可控性，并提升了学员的学习体验与效能感[4]。

更进一步，现代模拟训练平台正日益整合更精细的过程性指标数据与质量管理思维。例如，住院医师规范化培训中引入腹腔镜模拟训练系统并采用结构化评价，可显著提升学员的操作表现与达标效率，同时为个体化短板定位提供可追踪的过程数据[6]。在此基础上，若将训练设计为“设定标准—实施训练—评估反馈—针对不足再训练—复核达标”的闭环，其教学效果更稳定、可复制性更强；在腹腔镜外科基础规范化教学的住培实践中，规范化教学路径与评价结合能够改善学员技能考核表现 [7]。需要强调的是，模拟训练的价值不仅在于提供练习机会，更在于以标准化与可量化证据推动持续改进，从而更可靠地支撑患者安全。

三、AI 辅助评价：从自动化打分到可解释的智能反馈

（一）技术驱动的必然：AI 进入技能评价领域的逻辑

人工智能与数字化技术在外科技能评价领域受到高度关注，其现实动因在于传统评价模式存在难以克服的规模瓶颈：专家时间有限，面对大量学员难以实现高频次、同质化的观察与反馈。以手术机器人训练为例，训练系统可自动记录任务完成时间、路径与错误等过程指标，并据此生成初步评分或学习曲线，用于支持阶段性评估与个体化训练剂量调整[8]。需要强调的是，这类自动化工具的价值在于“增强”而非“替代”——教师仍需承担目标设定、情境判断、关键风险点纠偏与最终准入决策等核心职责。

（二）研究脉络：从可行性验证到临床适用性探索

当前，AI 技能评价的研究主要循着两条技术路径展开：一是基于机器人手术平台或可穿戴传感器采集的运动学数据；二是基于常规手术视频的计算机视觉分析。基于机器人或传感器的路线，由于数据本身具有结构化程度高、噪声相对较小、易于量化的特点，在研究报告中常能取得较高的技能等级区分准确率。而基于视频的路线，虽然更贴近无附加设备的真实临床场景、具有更广泛的推广潜力，但其分析难度也更大，易受术中出血、烟雾、组织遮挡、镜头移动与切换等多种因素的干扰。

在国内机器人与腔镜外科实践中，AI 辅助评价的落地往往以“可解释的过程证据”为切入点。例如，有研究基于深度学习 YOLOv5 对机器人辅助单孔腹腔镜子宫切除术视频进行实时解剖标志指示，将关键解剖结构的识别从隐性经验转化为可视化、可追溯的过程线索，从而为培训中的过程反馈与一致性复核提供了新型证据来源[9]。与此同时，围绕手术机器人培训的标准化实施，国内亦有研究系统梳理培训流程、评价节点与质控要点，强调将自动化输出纳入“评分者培训—分级授权—持续质控”的治理框架之中，以提升工具在真实教学场景中的可用性与可信度[10]。

除纯视频分析外，另一类可行路径是将“任务表现”与“工作负荷 / 团队行为”等指标结合，构建更贴近真实手术情境的综合评价。例如，基于 NASA-TLX 量表的研究尝试量化团队心理负荷，并探讨其与机器人手术熟练度之间的关联，提示在技能评价中引入主观负荷与过程性指标有望增强对“胜任力状态”的解释力[11]。这类探索为未来在不额外增加过多硬件负担的前提下，将客观过程数据与可解释的人因指标共同用于教学反馈提供了可行思路。

总体而言，该领域的研究证据正在从早期验证技术是否可行的阶段，逐步转向关注其是否可靠、是否可被教育者理解、是否能切实改善教学效果等更具实践意义的维度。这也是为什么近期的系统分析更加强调：除了展示算法区分新手与专家的能力外，研究更需要提供与教育后果直接相关的证据，例如 AI 反馈能否有效指导学员改进特定缺点、能否预测或减少实际操作中的错误与并发症发生率。

（三）应用壁垒：数据、算法与信任的三角难题

AI 技能评价面临的最根本挑战之一在于金标准的界定。目前绝大多数研究仍将专家小组基于量表的评分作为训练算法的近似金标准标签。然而，专家评分本身并非绝对客观的真理，其内部信度与外部信度的差异是公认的难题。一项针对机器人手术技能评价客观工具与 AI 方法的系统分析明确指出，尽管新方法层出不穷，但按照严格的效度证据框架进行完整验证的研究仍相对稀缺，这正是制约这些工具从实验室走向广泛临床教学应用的重要原因之一[12]。

此外，AI 评价模型在实际部署前，还必须妥善应对至少四个层面的现实问题：1. 泛化能力与域迁移难题：在不同医院、使用不同品牌设备、不同画质参数、不同术者习惯及病例复杂谱系

下采集的数据，其分布可能存在显著差异。在一个中心数据上训练表现优异的模型，直接应用于新环境时，性能可能出现显著下降。2. 算法偏倚与公平性隐患：如果训练数据主要来源于少数顶尖专家或单一医疗中心，模型可能会学习到对某种特定手术风格或流程的偏好，从而对那些操作风格不同但同样安全有效的学员做出不公平的负面评价，甚至可能扼杀技术创新的多样性。3. 决策过程的可解释性欠缺：教育场景尤其需要知其然，更知其所以然。学员和教师不仅需要一个最终的分数或等级，更需要了解为何在此处扣分、具体的缺陷是什么以及应当如何改进。当前许多高性能的深度学习模型在这方面仍显不足，其黑箱特性影响了教学双方的信任与接受度。4. 数据治理与伦理责任边界：手术视频与运动学数据涉及患者隐私、数据安全、医院合规管理等诸多法律与伦理问题。必须从制度层面明确谁有权收集和使用这些数据、如何使用以及出现评价争议或不良后果时的责任归属。

近期围绕 AI 技能评价的讨论普遍强调两点：其一是建立更可迁移的客观指标体系，系统性处理数据异质性与模型偏倚；其二是提升可解释性，使输出能够直接服务于教学决策与改进。在机器人外科领域，学习曲线研究以手术时间、关键步骤稳定性与并发事件等过程指标刻画技能获得与平台适应的轨迹，为“分级训练—达标准入—持续监测”的评价体系提供了可操作的量化依据[13]。同时，围绕团队沟通与工作方式的研究表明，诸如语音交流效率等团队行为指标也可能影响协作质量与操作流畅度，提示未来更可能走向“视频 / 运动学过程—任务表现—人因与团队行为”多维证据融合的联合评价框架[14]。

四、面向课程与考核的实施路径：从选择工具到构建生态

（一）目的先行：依据评价目标配置评估工具组合

在外科培训中，技能评价通常服务于三种不同的核心目的：第一，形成性反馈，旨在帮助学员识别不足、定向改进；第二，阶段性结业或操作准入考核，用于决定学员是否能进入下一培训阶段或获得特定手术操作的独立权限；第三，项目质量改进与学科管理，用于发现培训体系中的系统性薄弱环节，指导课程优化。这三种目的对评价工具的特性要求截然不同：形成性评价强调工具的敏感性、反馈的具体性与可重复性；准入考核则极度重视工具的信度、效度与标准化程度；质量改进则需要工具产生的数据易于汇总、分析与跨期比较。若不对评价目的进行清晰界定，很容易导致工具先进，但与实际教学需求错配的尴尬局面，造成资源浪费与师生抵触。

在实际的培训项目设计中，一个务实且有效的策略是采用分层组合评价方案。在培训早期，以高结构化的模拟器训练配合 OSATS 或基础技能核查清单为主，重点在于建立规范的动作模式与操作流程。在培训中期，逐步引入基于视频的回顾性评价与标准化的工作场所评估，重点考察学员在接近真实或真实临床情境中应用技能、稳定发挥的能力。在培训后期或高风险操作准入节点，则采用包含现场专家观察、关键步骤视频评审、模拟器复杂场景考核以及并发症回顾在内的多源证据综合评价，以最大程度保证决策的审慎与公正。国内对于住院医师技能培训体系的讨论也指出，强化形成性评价与贯穿培训全过程的质量控制，例如引入持续改进循环的管理思维，对于提升培训项目的可持续性与同质化水平具有重要的现实意义[1]。

（二）赋能评分者：将评价一致性作为教学质量的核心投资

只要评价过程中存在人的判断环节，无论是 OSATS、视频评分还是工作场所评估，评分者间的一致性就是保障评价质量的生命线。实践中，行之有效的做法包括：建立包含典型表现水平的锚定案例视频库；定期组织评分者开展校准会议，讨论评分分歧并统一认识；对常见错误与关键风险操作节点进行清晰定义与示例说明；设计结构化的反馈模板，强制要求包含本次表现最需改进的具体方面及下一次练习的明确任务。这些投入虽然短期内增加了教学管理的工作量，但长期来看，它们能换来更稳定可靠的评价结果、更令学员信服的反馈指导，从而根本上减少学员对评价的抵触情绪和仅为分数而练习的功利倾向，回归教育本质。

（三）人机协同：AI 技术落地的务实定位与渐进策略

在可预见的未来，AI 在外科技能评价中最具可行性和接受度的角色，并非完全取代人类专家，而是作为第二观察者或智能预筛与分析工具融入现有流程。具体而言，AI 可以用于自动识别并截取手术视频中的关键步骤或疑似问题片段；实时或事后提示可能存在的风险动作或流程偏差；基于历史数据生成初步的评分建议与学习要点提示；对海量训练数据进行趋势分析，发现共性问题。随后，再由带教老师对 AI 的输出进行复核、结合具体情境进行解读，并最终形成给予学员的反馈。这种人机协同的路径，既符合医学教育对安全、责任与伦理的严格要求，也有利于在合作中逐步建立教师与学员对 AI 工具的信任。有分析提出，如果 AI 辅助工具能在提供反馈的即时性、评价标准的一致性以及节省专家时间资源方面展现出明确优势，同时通过增强可解释性来管控偏倚风险，那么它将更有可能被现有的课程与考核体系逐步采纳和整合[10]。

五、总结

从 OSATS 的创立到 AI 辅助评价的兴起，外科技能评价演进的主线并非简单地追求工具的新颖性，而是致力于实现证据基础更扎实、反馈指导更可执行、教育体系更可持续的深层目标。OSATS 所代表的结构化评价思想，首次使外科技能从依赖经验的模糊判断走向了基于行为的科学测量；手术视频化与高仿真模拟训练的普及，则将评价的时空范围从即时的现场观察扩展到了可随时回放、反复审视的证据域；而人工智能技术的引入，则为在更大规模上实现即时、一致且个性化的技能反馈开辟了新的可能性。

然而，无论技术工具如何迭代升级，外科教育评价的最终归宿仍需锚定两个核心原则：第一，评价是否真正服务于学员的学习与发展，成为促进成长的动力，而非增加负担的阻力；第二，评价是否能够通过提升培训质量，最终切实服务于患者安全与手

术结局的改善。将评估工具置于包含课程系统设计、师资队伍建设、数据治理规范以及教育文化塑造在内的整体生态中进行考量与部署，远比孤立地引入某一项尖端算法或工具更为重要。未来，构建一个融合结构化量表、多源数据与智能辅助技术，且与培训目标紧密咬合的评价生态系统，将是推动外科教育迈向更高水平同质化与专业化的关键所在。

参考文献

[1] 梁馨予，刘晓岚，杨超，栾岚 . 临床医学专硕住培技能培训体系的探讨 . 中国继续医学教育 2022；14(05)：161-5.

[2]Martin JA，Regehr G，Reznick R，et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997；84(2)：273-8.

[3] 梁耿祺，关礼贤，赵振华，廖俊发，徐勋 . 视频教学结合双评分系统对提高住院医师规范化培训腹腔镜手术的研究 . 中国当代医药 2021；28(35)：199-202.

[4] 张丹丹，李佳，张明杰，宋子璇，王晓雪 . 基于 "Miller 金字塔 " 原理的分层教学模式在妇科腹腔镜模拟教学中的应用 . 中国医学教育技术 2022；36(01)：70-4+101.

[5] 温志锋，吴安华 . PAL 结合 DOPS 在神经外科住培学员批判性思维能力培养中的实践 . 浙江医学教育 2021；20(05)：28-30+23.

[6] 张喆，胡晨浩，时飞宇，张磊，孙学军，佘军军 . 腹腔镜模拟训练系统在住院医师规范化培训中的有效性评价 . 医学教育研究与实践 2021；29(06)：912-5.

[7] 王李，徐琰，黄彬，et al. 腹腔镜外科基础规范化教学在住培医师教学中的应用 . 中国继续医学教育 2022；14(16)：137-40.

[8] 蒋凌箫，文志勇，陈高杰，杨琨，王行环 . 不同阶段医学生手术机器人训练效果对比 . 武汉大学学报 (医学版) 2024；45(02)：169-74.

[9] 马周，易跃雄，陈雨柔，et al. 基于深度学习 YOLOv5 网络的机器人辅助单孔腹腔镜子宫切除术实时解剖标志指示系统 . 武汉大学学报 (医学版) 2024；45(02)：152-8.

[10] 陈紫嫣，文志勇，杨琨，王行环 . 手术机器人培训标准化过程中的关键问题分析及应对 . 武汉大学学报 (医学版) 2024；45(02)：165-8+95.

[11] 向梦，张棣，张依云，刘黎明，赵国艳 . 基于 NASA-TLX 量表的团队心理评估与机器人手术熟练度的相关性 . 武汉大学学报 (医学版) 2024；45(02)：159-64.

[12]Boal MWE，Anastasiou D，Tesfai F，et al. Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment：a systematic review. Br J Surg 2024；111(1).

[13] 程尼涛，蒋鹏飞，王现国，黄静宇，刘俊，胡卫东 . 机器人辅助微创食管癌切除术的学习曲线 . 武汉大学学报 (医学版) 2024；45(02)：127-31+37.

[14] 陈高杰，李露，郑航，杨琨，王行环 . 蓝牙耳机对机器人手术团队语言交流的影响 . 武汉大学学报 (医学版) 2024；45(02)：175-9.

