·医学模拟教育·

提升医学模拟评价信度的12项实用建议

李力[1]　刘继海[2]　谢亚典[3]　史迪[2]　胡小芸[4]　陈志桥[5]　安海燕[6]
[1]武汉大学中南医院重症医学科/湖北省重症医学临床医学研究中心，武汉 430071；[2]中国医学科学院北京协和医院急诊科/疑难重症及罕见病全国重点实验室，北京 100730；[3]武汉大学中南医院教学办公室，武汉 430071；[4]中国医学科学院北京协和医院内科重症监护室，北京 100730；[5]武汉大学中南医院急救中心，武汉 430071；[6]北京大学人民医院麻醉科，北京 100044
通信作者：李力，Email：mrcatking@gmail.com

【摘要】 医学模拟评价已成为衡量医护人员临床能力的重要方法。这些评价为受评者提供了在还原真实医疗场景且受控的环境中展示临床技能的机会。然而，评价组织方必须采取特定的措施，才能保障这些评价能够产生有效且可靠的结果。模拟评价的信度是模拟评价质量保障的重要组成部分。本文从评价者、模拟环境和评价工具3个关键维度梳理了提升模拟评价信度的12项实用建议：①选择具备相关专业背景和丰富模拟教学与评价经验的评价者；②为评价者提供关于评价的规范化培训；③进行模拟评分和试点评价，以达成评价者之间的共识；④防止评价者超时工作或认知负荷超载；⑤为标准化人员提供充分的培训和再培训；⑥对标准化病人的角色演绎提供即时和长期的质量保障措施；⑦确保模拟器保持一致的运行状态；⑧确保模拟为受评者提供足够的功能仿真度；⑨明确定义评价工具中的要素和项目，并控制评价条目的数量；⑩如果使用多个评价工具，确保不同工具之间的一致性；⑪尽可能使用具有构念效度证据的评价工具；⑫建立确保数据（评分）采集一致性的系统，并确保数据安全。这些策略旨在帮助医学教育工作者构建结构化的、可信的模拟考核评价方案。

【关键词】 模拟医学；　评价；　信度；　质量控制
【中图分类号】 R33
基金项目：湖北省卫生健康委员会科研基金资助（HBJG-250007）
DOI: 10.3760/cma.j.cn116021-20230503-02099

Twelve practical recommendations for enhancing the reliability of medical simulation-based assessment

Li Li[1], Liu Jihai[2], Xie Yadian[3], Shi Di[2], Hu Xiaoyun[4], Chen Zhiqiao[5], An Haiyan[6]
[1]Department of Critical Care Medicine, Zhongnan Hospital of Wuhan University/Hubei Provincial Clinical Research Center for Critical Care Medicine, Wuhan 430071, China; [2]Department of Emergency Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences/State Key Laboratory of Complex Severe and Rare Diseases, Beijing 100730, China; [3]Teaching Office, Zhongnan Hospital of Wuhan University, Wuhan 430071, China; [4]Medical Intensive Care Unit, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing 100730, China; [5]Emergency Center, Zhongnan Hospital of Wuhan University, Wuhan 430071, China; [6]Department of Anesthesiology, Peking University People's Hospital, Beijing 100044, China
Corresponding author: Li Li, Email: mrcatking@gmail.com

【Abstract】 Medical simulation-based assessments have emerged as a vital method for evaluating the clinical competence of healthcare professionals. These assessments provide examinees with opportunities to practice clinical skills in a controlled environment that mirrors real-life medical scenarios. However, assessment organizers must implement specific measures to ensure that these assessments yield valid and reliable results. The reliability of simulation-based assessments is a crucial component of quality assurance. This article delineates twelve practical recommendations to enhance the reliability of medical simulation-based assessments across the three key dimensions of raters, simulation environment, and assessment tools.

These recommendations include ① selecting raters with relevant professional backgrounds and extensive experience in simulation-based teaching and assessment; ②providing raters with standardized training for the assessment; ③conducting mock ratings and piloting to reach consensus among raters; ④preventing excessive work hours or cognitively overload while rating; ⑤providing standardized personnel with adequate training and re-training; ⑥ implementing immediate and long-term quality assurance measures for standardized patient role performance; ⑦maintaining consistent operation of simulation equipment; ⑧guaranteeing the simulation provides adequate functional fidelity for the examinees; ⑨clearly defining each element and item in assessment tools while controlling their quantity; ⑩ensuring consistency when multiple assessment tools are used; ⑪using assessment tools with evidence of construct validity when possible; and ⑫establishing systems to ensure consistency in data (scoring) collection while safeguarding data security. These strategies aim to assist medical educators in developing structured and reliable simulation-based assessment schemes.

【**Key words**】 Medical simulation; Assessment; Reliability; Quality control

Fund program: Health Commission of Hubei Province Scientific Research Project (HBJG-250007)

DOI: 10.3760/cma.j.cn116021-20230503-02099

医学模拟评价(healthcare simulation-based assessments)通过高度仿真的临床情境,为医学教育提供了一个独特的临床能力考核平台,同时也为临床教师提供了一条有效评价学员能力的标准化途径[1-2]。然而,评价过程的信度(reliability),即可靠性,受到多重因素的影响,这些因素包括评价者的专业水平、模拟环境的真实性及评价工具的科学性[3]。信度不佳的评价可能会导致评价结果无法准确反映受评者的真实能力水平,进而影响教育决策的制定。目前,医学教育领域关于模拟评价信度的相关研究更多源自国际医学教育期刊,而我国在这方面的报道相对较少。本文通过全面回顾国际医学教育期刊中关于模拟评价的文献报道,结合最佳实践案例及模拟教师能力培训的经验总结,系统地提出了12项具有可操作性的建议,旨在为我国各级医疗教育机构和组织优化模拟评价信度提供参考。这些建议涵盖了评价前的准备、评价过程的控制及评价后的质量改进等关键环节。通过落实这些建议,教育人员可以最大限度地减少评价差异,提高评分一致性,从而确保评价结果能够真实反映受评者的临床能力。

1 评价者维度

评价者(rater)是实施模拟评价的人员,在不同的评价场景中通常承担着不同的具体职责并有着不同的名称。如在执业医师资格考试和住院医师规范化培训结业考核等高利害考试中,评价者通常被称为“考官”。

建议1:选择具备相关专业背景和丰富模拟教学与评价经验的评价者。评价者的选择对于保障模拟评价的信度至关重要。评价者作为内容专家,必须拥有准确评价表现所需的知识和经验[4]。这种专业素养使他们能够识别临床表现中的细微差别,而这些差别可能会被经验不足的评价者所忽略。临床经验丰富的评价者能够更好地识别各种临床情况中的关键要素,理解造成表现差异的影响因素[5]。他们通常对临床背景有更深入的理解,并能提供更有针对性的反馈。

在选择评价者时,必须考虑以下因素:①临床专业知识和技能。评价者应在所评价的临床领域具有丰富经验,这不仅包括对学科内容的深入了解,还应包括对与患者交互过程及其复杂性的全面理解。②既往评价经验。先前参与过评价工作,能为评价者提供进行有效评价所需的技能,而熟悉特定评分工具和评价标准的评价者更有可能一致性地应用它们。③模拟教学经验。承担模拟教学的临床教师通常知晓模拟情境的优势与局限性,因此具备识别并剔除模拟情境局限性对临床能力评价带来影响的能力。为保障信度,机构应进行结构化的评价者遴选流程,包括审查其教育背景、临床经验、既往承担临床教学评价的经验,并对候选人员进行面试和日常观察。

建议2:为评价者提供关于评价的规范化培训。选定评价者后,需要进行全面且规范化的培训,以统一他们对评价目标和所用具体工具的理解[6]。一方面,要说明评价的目的,以促进评价者对模拟评价的总体理解,这包括计划评价的临床能力,以及结果将如何影响教育或临床决策。另一方面,需要对评价过程中所使用的评价工具及其评价标准进行培训,以避免评价者因理解偏差或对工具不熟悉而影响评价的可靠性[7-8]。

对于高利害考试而言,不建议以避免泄题为由,在考试正式开始前几小时内进行考官一致性培训。因为考官培训和达成共识需要时间和反复的

磨合，考前时间紧迫，常无法就分歧较快达成共识，因此会影响高利害考试评分的一致性，进而影响考试评分结果的权威性。考官培训可通过提前对多项考试内容进行培训，或先进行相近内容的培训，随后将考试内容替换为实际考试内容的方式，以避免考题泄露；考官提前入围并接受培训也可防止考题泄露的情况发生。规范化的评价者培训，特别是针对高利害模拟评价的考官培训，通常需要通过考官错误培训（rater error training，RET）、表现领域培训（performance dimension training，PDT）、参考样式培训（frame-of-reference training，FOR），以及行为观察培训（behavioral observation training，BOT）4 个步骤实现[6,9]。其中，考官错误培训通常由命题人员和考试组织方就考官常见错误进行理论介绍；而表现领域培训则通过讲授、解释和示例，使考官明确所需要评价的内容。对于参考样式培训，则更多通过提供模拟情境下的考官评分，在展示了好、中、差等典型受评者（此时可称为“考生”）行为后，让考官进行打分和讨论，从而锚定打分行为。行为观察培训则聚焦于考官在发觉、感知和回忆受评者行为方面的执考能力培训，通常也需要通过模拟评分或试点评分来实现。

建议 3：进行模拟评分和试点评价，以达成评价者之间的共识。模拟评分和试点评价是促进评价者达成共识的有效途径。这些途径通常通过评价者在模拟场景（通常是提前录制好的操作视频或由标准化考生现场演绎的一场模拟操作）中练习使用评价工具来实现。这种训练使评价者能够熟悉评分过程，并将自己的打分与同行进行校准[6,9-11]。它为评价者提供了一个在低风险环境中练习临床能力评价的机会。

在模拟评分过程中，需要注意实现解释的一致性，允许评价者就同一评分场景进行讨论和比较，以统一对评分标准的理解。过程中应充分识别评判差异，评价者之间的评分差异是影响模拟考试效度的重要因素[10-11]。通过讨论和澄清评价标准，可以有效解决这些问题。同时，通过模拟评分练习，评价者可以建立起实施临床能力评价的信心，从而在实际评价中获得更可靠的评分[6]。这些培训活动需要精心规划组织，应注意选择合适的情境、构建反馈流程，并为讨论和建立共识留出时间。此外，评价者应参加持续的培训和交流活动，以保持其知识和技能的更新，并通过定期的校准复训来减少评分中的差异，确保评价者能够一致地应用评价标准。

建议 4：防止评价者超时工作或认知负荷超载。认知负荷超载会严重损害评价者有效进行评价的能力。为最大限度地减少认知疲劳并提高评价的可靠性，在组织评价活动时，应控制评分工作量，以避免评价者在一轮评价过程中负担过重[12]。任务负荷重的评价往往出现在高利害的模拟考试中。在时间安排上，考试组织方应为每次评价留出充足的时间，包括反思和讨论的时间。同时，模拟考试中应规划适当的休息时间，定期休息有助于保持考官的注意力和专注力，可以帮助他们减轻认知负荷，恢复认知储备，从而保持评价质量[13]。另一方面，考试组织方应监测考官的疲劳迹象，如注意力下降和决策能力受影响。除此之外，提供一个积极且宽容的评分环境，让考官能够自主表达需要休息的诉求，这一点也十分重要。需要认识到，管理评价者的工作量和认知需求有助于维持评分过程的可靠性，确保评价者拥有可管理的工作量不仅能提高他们的表现，还能促进其对受评者进行更准确、更公平的评价。

2　模拟环境维度

模拟环境（simulation environment）指为了保障评价顺利进行而构建的模拟情境，包括情境中的标准化模拟人员、模拟器及其他所需的用品和物资，也包括情境中剧情的发展变化。

建议 5：为标准化人员提供充分的培训和再培训。标准化人员（standardized person，SP）在模拟评价中扮演着关键角色，不仅限于扮演患者（或其家属）角色的模拟/标准化病人（simulated/standardized patient），也包括协助受评者完成某一临床行为或提供所需信息的辅助医疗人员[14]。他们能否始终如一地演绎特定角色，包括准确提供病史和情绪状态，或以适当的程度完成协助任务，将直接影响评价的真实性[15]。为了提高评价的信度，可以通过以下方式为 SP 提供充分的培训和再培训[16]。①初始培训课程：初始培训应涵盖 SP 将要演绎场景的具体内容，包括角色背景、预期表现及在与受评者互动过程中的适当反应。②定期再培训：随着时间的推移，SP 的表现可能会偏离既定标准，因此定期再培训课程有助于确保 SP 在演绎中保持准确性和一致性。③表现指标：向 SP 提供结构化的反馈有助于长期维持其表现，如借助标准化的核查表来评价 SP 对指定角色的遵循程度，并据此向其提供反馈。④考前培训：类似于考官的考前培训，尽管许多 SP 拥有丰富的扮演经历并接受了规范化的培训，但情境考核案例可能有所差别。因此，针对 SP 的考前培训需要考核设计者提供专门的脚本来辅助 SP 理解模拟情境中的人物角色和表演尺度。当 SP 在不同评价

中持续、准确地扮演其角色，为受评者提供稳定且真实的互动时，可以极大程度地减少向受评者漏给、错给信息，或在错误时间点提供相关信息的情况，确保情境模拟考核案例能够按照计划顺利推进，从而保障模拟评价的信度。

建议 6：对标准化病人的角色演绎提供即时和长期的质量保障措施。质量保障措施对于维持 SP 表现的高标准至关重要[17-18]。SP 在评价中的表现应接受定期的观察与评估。这些观察可以揭示需要改进的领域，并确认 SP 是否遵循既定的表演要求。在有条件的评价考核过程中，可以通过让 SP 佩戴耳麦，接受 SP 培训师或考试设计者的远程指导，在指定时机做出规定的动作或进行口头表达，从而推动剧情发展，同时也可以避免错误的行为误导受评者。

与评价者类似，同样需要考虑 SP 因长时间工作而可能出现的认知负荷过重或疲劳等情况。一方面，要避免单位时间内要求 SP 呈现过多的情境内容，以免导致他们的认知储备过早耗竭；另一方面，对于需要 SP 多次重复表演相同内容的组织安排而言，也应确保 SP 得到适当的休息，以保持其注意力与专注力。

评价结束后，组织者应与 SP 进行总结反馈，讨论其表现。可以收集 SP 和受评者的反馈进行分析，以发现其中的差异，从而确定未来可能需要改进的方面。利用结构化评价工具，如表现评分量表，可以对 SP 的演绎进行客观评价。这些工具应侧重于沟通清晰度、情绪可靠性和角色准确性等领域。

实施质量保障措施能确保 SP 的演绎保持准确和一致，从而维持模拟评价的完整性。

建议 7：确保模拟器保持一致的运行状态。模拟评价情境的构建在很大程度上依赖于模拟器的有效运行[19]。模型的损耗及模拟器因技术原因在运行过程中的稳定性，均可能影响模拟环境一致性的呈现。

单项技能考核常规用到的任务训练器主要以还原解剖为主要功能，如穿刺模型、心肺复苏模拟人等。这类模拟器成本相对低廉，总体不易磨损，但在反复穿刺、消毒、按压的过程中，可能会出现穿刺部位痕迹过于明显或模拟皮肤颜色改变的情况，进而影响操作者对于操作部位的选择与识别。这种情况可能导致受评者过于轻易地完成考核任务，或误导受评者无法正确完成相关操作。因此，考试设计者应选择不易磨损材质的模拟器，或定期更换模拟器的零部件，以避免影响受评者的表现，从而保证考核的信度。

对于专科技能考核，或是以情境变化与判断为主的情境模拟案例考核，通常会用到高科技模拟器。高科技模拟器通常会对受评者的行为表现作出反馈，并与其进行互动，从而影响受评者作出后续判断与决策，进而影响其基于思维活动的临床行为表现。此类模拟器往往要求操纵人员提前接受培训，方可具备一定的资质，要求能够正确设置、调整模拟器配置，并能熟练地进行故障排除。另一方面，此类模拟器需要定期维护和校准。评价组织方应制定常规维护计划，检查模拟器的功能和校准情况，包括验证软件和硬件的性能，以便在评价正式实施前解决任何潜在问题。

通过确保模拟器的一致运行，机构可以提高评价结果的可靠性。受评者在稳定可靠的模拟环境中，才能更有效地展示自己的技能[20]。

建议 8：确保模拟为受评者提供足够的功能仿真度。模拟中的功能仿真度（functional fidelity）是指模拟能够准确反映真实临床情况复杂性的程度。校准模拟环境的功能仿真度可以显著提高评价的相关性和可靠性。

情境案例的真实程度应确保受评者能够相对自然地完成所要求的临床行为，包括创建场景以促进受评者展现批判性思维，并以现实的方式应用知识[21]。为此，模拟应在一定程度上体现临床环境，包括医疗设置、设备、模拟人功能及医患互动，但应避免纳入过多元素干扰受评者的临床决策判断。例如，当考核目标是“正确实施气管插管步骤”时，若模拟场景中提供不同型号的气管导管，可能会引入额外的考核目标，即“选择合适型号的气管导管”。这种额外目标的引入会影响受评者的行为表现，且由于没有对应的考核项目，反而会降低该模拟评价的可靠性。

模拟评价案例的设计应让教师或临床专家充分参与，以确保模拟能够满足教育目标，并准确反映专业领域当前的最佳实践[22]。另一方面，模拟评价案例应先进行试运行，并收集参与该案例的人员的体验和反馈，从而获得哪些方面感觉真实或不真实的独到见解。而这些反馈可用于优化和改进考核情境案例。

当模拟具有高水平的功能保真度时，受评者能更如实地展示其临床实践水平[23]，从而确保模拟评价的实施意义及其有效性。需要注意的是，尽管随着科技的发展，模拟设备能够在更大程度上具备功能保真度，但仍可能存在无法完整反映临床实际情况的问题。例如，即使使用 3D 打印技术，依然无法穷尽所有患者的解剖变异；或者由于模拟人内置音响数量的限制，可能导致原本在真实病人身上可以

进行听诊的部位无法进行呼吸音听诊。因此，在设计模拟评价时，应考虑技术局限性因素，并允许受评者知晓这些局限性，同时给予正确引导，以实现“搁置疑虑”（suspension of disbelief）的效果。

3　评价工具维度

在模拟评价中使用的评价工具（assessment tool）通常是提前制定好的评分表，以及对应的条目解释或评分标准。评价工具的设计与开发需遵循一定的方法，并符合测量学原理，以保证其信度和效度[24-25]。

建议 9：明确定义评价工具中的要素和项目，并控制评价条目的数量。在模拟评价中使用的评价工具应当表述明确，无歧义，各项内容具体且不模糊，以尽可能提高其可靠性和一致性[8, 26]。

为实现可靠的评价，明确每个评价条目及其标准至关重要。首先，有必要创建一个详细的评分表格，该表格应能概括所有社会或专业团队对受评者就评价项目的具体期望，并包括对应表现水平的清晰描述，以便帮助评价者做出准确的判断。常使用的两类评分表样式为核查表（checklist）和等级评分表（rating scale）[26]。其次，每个表现水平应提供相应的示例或“行为锚定点”，以便帮助评价者明确区分优秀、良好和不合格的表现，从而减少评分时的模糊性。此外，多方利益相关者共同参与开发评价工具，可以促进对标准的广泛共识，确保其既全面又符合教育目标。最后，应定期对评价工具进行审查和修订，通过评价者和受评者的反馈不断完善，从而保持其与当前教育标准的契合度。评价工具中应明确定义所有要素和项目，这样可以让评价者对评价中的期望有共同的理解，从而最大限度地减少评价者之间的差异，提高评分的可靠性。

关于评价条目的数量与相关条目的赋分权重，往往是评价工具开发者容易陷入的误区。他们希望条目或权重越细致越好，以期增加评价的准确性与有效性。然而，对于评价者而言，条目过多会增加评价过程中的认知负荷，反而可能降低评价的信度。对于高利害模拟评价而言，如果没有足够可靠的评价数据作为支撑，就无法讨论评价结果是否准确而有效。赋分权重的确定则是根据不同评价目的，以及考试组织方更为关注的方面，而作出相对灵活的调整。

建议 10：如果使用多个评价工具，确保不同工具之间的一致性。在使用多个评价工具的教育环境中，工具之间的一致性对于保持评价的完整性至关重要[27-28]。一方面，应确保评价工具中所使用的术语和标准的标准化，以防止混淆，并确保评价者对不同工具的标准有相似的理解。当评价工具涉及多个学科时，应在评价开始前组织校准会议，让不同学科的评价者讨论并统一他们的评分实践，以促进一致性。校准会议应在考官培训和考试开始前进行，以确保所有评价者都在一个共同的框架下工作。同时，评价组织方应定期比较不同评价工具的结果，以发现潜在的差异。这种做法能够使教育者确保评价产生可比较的结果，从而增强评价过程的可信度。除此之外，还应建立评价工具的周期性持续改进反馈机制。从评价者和受评者的角度获取每个工具的反馈，评价组织方可以确定需要调整和改进的领域。通过寻求意见并根据实际经验进行调整，可以提高工具自身的质量，以及提升不同评价工具之间的一致性。而一旦需要对此类评价工具进行调整或修订时，为了保证评价结果的一致性，工具中的评价项目修改程度越大，越需要再次通过专家共识达成一致。

通过在多个评价工具中保持一致的标准和方法，评价组织方能够增强评价的可靠性，确保无论评价背景如何，都能准确地反映受评者的能力[29]。

建议 11：尽可能使用具有构念效度证据的评价工具。选择具有既定构念效度（construct validity）的评价工具对于确保可靠评价至关重要。构念效度是指评价工具是否准确测量了它声称要测量的技能或能力[24-25, 30-33]。具有高度构念效度的评价工具能够为评价结果的解释提供信心，并有助于避免对受评者能力的错误分类。因此，在选择评价工具时，特别是进行高利害考核评价时，应优先考虑那些既往已经通过严格研究论证、具备既定构念效度的实证证据的评价工具[34-35]。而对于低利害关系的模拟评价，尽管也可以使用相同的评分表格进行评价，但对于评价工具是否达成专家共识的要求并不像高利害评价那样严格。

通过因素分析或其他统计方法分析评价工具的内部结构（internal structure），可以帮助确认该评价工具是否有效地测量了预期的构念。一个结构良好的工具将在不同人群和环境中展示出一致的评分模式除了难度、区分度等经典评价理论所关注的内容外，考官间一致性（inter-rater reliability）与考官自身前后一致性（intra-rater reliability）也是评价工具重要的测量学监测指标。前者反映在同一模拟评价环境中，多个不同评价者对分配分数（或权重）的一致性程度；后者则反映同一评价者在不同情况下对相同表现评分的一致性。这些监测指标不仅反映了评价工具能否产生可信的评价结果，还

可以作为评估考官的指标。定期实施评价者评价可以揭示评分中的潜在差异，从而用于考官的质量控制与执考能力的提升。

通过使用具有既定效度证据的评价工具，组织方可以加强模拟评价的完整性和可靠性[34]，最终提升受评者的教育体验。上述与构念效度相关的测量学指标更多反映的是考试实施过程中的情况，属于横断面的质量控制范畴。评价组织方可以通过多次跟踪评价工具的有效性，进一步强化其构念效度和信度，这一过程即为纵向研究（longitudinal study）。随着时间的推移收集数据，实施评价的机构能够持续验证工具的有效性，并适应不断变化的教育标准。

建议 12：建立确保数据（评分）采集一致性的系统，并确保数据安全。评价数据的采集和管理方式同样会影响评价的可靠性和准确性。建立强大的数据管理系统能够提升评价的质量。

目前，许多单位均在探索相关的实践标准，包括利用信息技术实现数据采集自动化[36-37]，以降低与人工录入相关的人为错误和不一致性风险。电子系统能够简化流程，并确保不同评价中评分的一致处理。制定全面的数据收集、存储和分析协议，对于保持一致性至关重要。这些协议应当解决格式、数据录入流程及处理敏感信息等方面的问题。另一方面，在评价过程中实施实时监控以跟踪数据采集，可以提高评价的可靠性。通过持续监督，管理员能够及时识别并解决数据差异或技术问题，从而确保评价过程高效推进。为确保上述措施得到有效实施，数据管理人员应接受与评价者相关的培训，重点强调准确采集和规范管理数据的重要性，以及高质量数据对评价结果可靠性的关键影响。

借助强大的数据分析技术，甚至未来利用人工智能技术对评价结果进行深度分析，能够有效识别发展趋势和潜在的改进方向[38]。对评价数据的持续监测与分析可以揭示受评者表现和评价者一致性的模式，为持续改进建立证据基础。许多机构采用数据仪表盘来实现实时数据可视化和动态报告，提供关于评价者表现、受评者结果及整体评价有效性的即时反馈，从而在出现问题时能够及时干预。

通过建立系统化、标准化的数据采集和管理流程，机构可以提高模拟评价的可靠性。甚至可以通过信息技术手段（如录制考试过程并进行基于视频的评价）来减少考官的任务负荷，进而对受评者的能力做出更准确、有效的评价。然而，评价组织方需严格审视借助数字化平台实施的评价。一方面，技术故障或使用不当可能影响评价实施过程；另一方面，这些问题可能增加由技术实施所带来的外在认知负荷，甚至出现危及评价数据存储与处理完整性的情况，从而最终降低评价的信度。

对于高利害考试而言，对相关数据的保存和备查有着极高的要求。数据安全，包括考生的个人隐私信息保护、考试内容的保密需求，以及数据的存储与保存稳定性等方面。以上各方面均需满足指定的标准与规范后方可付诸实践。

4　结语

模拟评价作为医学教育的重要工具，为评价未来医疗人员的能力提供了有效手段。在快速变化的医疗保健环境中，确保模拟评价的完整性和有效性始终是卓越医学教育的核心焦点。未来，要提升模拟评价的信度，需要更多地聚焦于非技术性技能的模拟评价、跨专业教育中的模拟评价，并开展更多纵向研究，以积累关于这些评价方法有效性的重要证据。此外，还应积极探索新兴技术在提升模拟评价信度方面的应用价值。本文将提升模拟评价信度的 12 项建议归纳为“评价者–模拟环境–评价工具”三维度模型，或称“Rater-Simulation environment-assessment Tool”的“RST 模型”。

需要注意的是，不同利害级别的模拟评价，其严谨程度有差别。利害级别越高的模拟评价，对信度的需求程度越高。基于三维度模型提出的 12 项建议，为教育工作者提供了一个系统性的指导框架，可以显著提高模拟评价的质量和信度。实施这些建议不仅能增强评价的可靠性，还将有助于实现培养有能力、自信和称职的医疗保健专业人员这一总体目标。

利益冲突　所有作者声明不存在利益冲突

作者贡献声明　李力、刘继海、陈志桥：提出论文构思、凝练主要观点、撰写论文；谢亚典、史迪、胡小芸、安海燕：总体把关、审订论文

参考文献

[1] McGaghie WC, Issenberg SB, Petrusa ER, et al. A critical review of simulation-based medical education research: 2003-2009 [J]. Med Educ, 2010, 44(1): 50-63. DOI: 10.1111/j.1365-2923.2009.03547.x.

[2] Holmboe ES, Sherbino J, Long DM, et al. The role of assessment in competency-based medical education [J]. Med Teach, 2010, 32(8): 676-682. DOI: 10.3109/0142159X.2010.500704.

[3] Hawkins RE, Margolis MJ, Durning SJ, et al. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research [J]. Acad Med, 2010, 85(9): 1453-1461. DOI: 10.1097/ACM.0b013e3181eac3e6.

[4] van der VC, Schuwirth LW. Assessing professional competence: a model integrating measures of health care outcomes, performance assessment, and the assessment of knowledge [J]. Adv Health Sci Educ Theory Pract, 2005, 10(4): 205-211. DOI: 10.1007/978-3-319-30064-1.

[5] Kogan JR, Conforti LN, Bernabeo E, et al. How faculty members experience workplace-based assessment rater training: a qualitative study [J]. Med Educ, 2015, 49(7): 692-708. DOI: 10.1111/medu.12733.

[6] Feldman M, Lazzara EH, Vanderbilt AA, et al. Rater training to support high-stakes simulation-based assessments [J]. J Contin Educ Health Prof, 2012, 32(4): 279-286. DOI: 10.1002/chp.21156.

[7] Cook DA, Dupras DM, Beckman TJ, et al. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial [J]. J Gen Intern Med, 2009, 24(1): 74-79. DOI: 10.1007/s11606-008-0842-3.

[8] McMullan M, Endacott R, Gray MA, et al. Portfolios and assessment of competence: a review of the literature [J]. J Adv Nurs, 2003, 41(3): 283-294. DOI: 10.1046/j.1365-2648.2003.02528.x.

[9] Newman LR, Brodsky D, Jones RN, et al. Frame-of-reference training: establishing reliable assessment of teaching effectiveness [J]. J Contin Educ Health Prof, 2016, 36(3): 206-210. DOI: 10.1097/CEH.0000000000000086.

[10] Davis DA, Mazmanian PE, Fordis M, et al. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review [J]. JAMA, 2006, 296(9): 1094-1102. DOI: 10.1001/jama.296.9.1094.

[11] Kogan JR, Conforti L, Bernabeo E, et al. Opening the black box of clinical skills assessment *via* observation: a conceptual model [J]. Med Educ, 2011, 45(10): 1048-1060. DOI: 10.1111/j.1365-2923.2011.04025.x.

[12] Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments [J]. Adv Health Sci Educ Theory Pract, 2013, 18(2): 291-303. DOI: 10.1007/s10459-012-9370-3.

[13] Paravattil B, Wilby KJ. Optimizing assessors' mental workload in rater-based assessment: a critical narrative review [J]. Perspect Med Educ, 2019, 8(6): 339-345. DOI: 10.1007/s40037-019-00535-6.

[14] Cleland JA, Abe K, Rethans JJ. The use of simulated patients in medical education: AMEE Guide No_{42}[J]. Med Teach, 2009, 31(6): 477-486. DOI: 10.1080/01421590903002821.

[15] Mirza N, Cinel J, Noyes H, et al. Simulated patient scenario development: a methodological review of validity and reliability reporting [J]. Nurse Educ Today, 2020, 85: 104222. DOI: 10.1016/j.nedt.2019.104222.

[16] MacLean S, Geddes F, Kelly M, et al. Simulated patient training: Using inter-rater reliability to evaluate simulated patient consistency in nursing education [J]. Nurse Educ Today, 2018, 62: 85-90. DOI: 10.1016/j.nedt.2017.12.024.

[17] Shirazi M, Labaf A, Monjazebi F, et al. Assessing medical students' communication skills by the use of standardized patients: emphasizing standardized patients' quality assurance [J]. Acad Psychiatry, 2014, 38(3): 354-360. DOI: 10.1007/s40596-014-0066-2.

[18] Bagnasco A, Rossi S, Dasso N, et al. Basing patient safety education on real student experience: development of a multinational simulation scenario [J]. J Prev Med Hyg, 2025, 65(4): E547-E554. DOI: 10.15167/2421-4248/jpmh2024.65.4.3110.

[19] Elendu C, Amaechi DC, Okatta AU, et al. The impact of simulation-based training in medical education: a review [J]. Medicine (Baltimore), 2024, 103(27): e38813. DOI: 10.1097/MD.0000000000038813.

[20] Cook DA, Hatala R, Brydges R, et al. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis [J]. JAMA, 2011, 306(9): 978-988. DOI: 10.1001/jama.2011.1234.

[21] Palaganas J, Raemer DB. Defining the role of simulation in health professions education: a qualitative study [J]. Simul Healthc, 2017, 12(2): 90-97. DOI: 10.2196/62803.

[22] Lavoie P, Michaud C, Bélisle M, et al. Learning theories and tools for the assessment of core nursing competencies in simulation: a theoretical review [J]. J Adv Nurs, 2018, 74(2): 239-250. DOI: 10.1111/jan.13416.

[23] Mitchell S, Blanchard E, Curran V, et al. Effects of simulation fidelity on health care providers on team training-a systematic review [J]. Simul Healthc, 2024, 19(S1): S50-S56. DOI: 10.1097/SIH.0000000000000762.

[24] Downing SM. Reliability: on the reproducibility of assessment data [J]. Med Educ, 2004, 38(9): 1006-1012. DOI: 10.1111/j.1365-2929.2004.01932.x.

[25] Downing SM. Validity: on meaningful interpretation of assessment data [J]. Med Educ, 2003, 37(9): 830-837. DOI: 10.1046/j.1365-2923.2003.01594.x.

[26] Ilgen JS, Ma IWY, Hatala R, et al. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment [J]. Med Educ, 2015, 49(2): 161-173. DOI: 10.1111/medu.12621.

[27] Furman GE, Smee S, Wilson C. Quality assurance best practices for simulation-based examinations [J]. Simul Healthc, 2010, 5(4): 226-231. DOI: 10.1097/SIH.0b013e3181da5c93.

[28] van der Vleuten CM, Schuwirth LT, Driessen EW, et al. A model for programmatic assessment fit for purpose [J]. Med Teach, 2012, 34(3): 205-214. DOI: 10.3109/0142159X.2012.652239.

[29] Norcini J. On purpose: the case for alignment in assessment [J]. Acad Med, 2023, 98(11): 1240-1242. DOI: 10.1097/ACM.0000000000005430.

[30] Borsboom D. Educational measurement (4th Ed.) [J]. Struct Equ Model A Multidiscip J, 2009, 16(4): 702-711. DOI: 10.1080/10705510903206097.

[31] McGaghie WC, Hinkamp MM, Harden RM. International best practices for evaluation in the health professions [M]. London: CRC Press, 2022.

[32] 王佳玉, 史霆, 徐玲玲. 基于Delphi-AHP法构建医学情境模拟课程实施质量评价工具的探索[J]. 中华医学教育杂志, 2025, 45(3): 235-240. DOI: 10.3760/cma.j.cn115259-20240116-00058.

[33] Downing SM. Twelve steps for effective test development [M]// Handbook of Test Development. London: Routledge, 2015.

[34] Johnston L, Sawyer T, Nishisaki A, et al. Neonatal intubation competency assessment tool: development and validation [J]. Acad Pediatr, 2019, 19(2): 157-164. DOI: 10.1016/j.acap.2018.07.008.

[35] Stufflebeam DL. Guidelines for developing evaluation checklists: the checklists development checklist (cdc)[EB/OL]. [2025-03-10]. https://files.wmich.edu/s3fs-public/attachments/u350/2014/guidelines_cdc.pdf.

[36] McQueen S, McKinnon V, VanderBeek L, et al. Video-based assessment in surgical education: a scoping review [J]. J Surg Educ, 2019, 76(6): 1645-1654. DOI: 10.1016/j.jsurg.2019.05.013.

[37] Fu Y, Zhang W, Zhang S, et al. Applying a video recording, video-based rating method in OSCEs [J]. Med Educ Online, 2023, 28(1): 2187949. DOI: 10.1080/10872981.2023.2187949.

[38] Boal MWE, Anastasiou D, Tesfai F, et al. Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review [J]. Br J Surg, 2024, 111(1): znad331. DOI: 10.1093/bjs/znad331.

（收稿日期：2023-05-03）
（本文编辑：邬小梅）