面向儿童执行功能评估的低碳自适应导入设计：一种测量效度保持型人工智能交互框架

蓝炜[1]　任怡霖[2*]

（1.　庆应义塾大学 媒体设计研究科，日本，999081；

2.*通讯作者，暨南大学珠海校区，广东 珠海，519070）

摘　要：计算机化执行功能评估能够提升儿童认知测评的标准化、时间控制与评分自动化水平，但儿童在正式测验前的任务进入困难可能与目标执行功能混杂；与此同时，若自适应支持延伸至正式计分阶段，又可能改变任务结构并影响测量效度。面向儿童心理评估中的绿色数字化、教育公平与低资源部署需求，本文提出一种“测量保持型自适应导入”设计框架。

该框架将系统支持严格限定在任务导入与练习阶段，并在正式计分阶段锁定刺激材料、呈现时序、反应映射、反馈政策、评分规则和汇总指标。研究以维度变化卡片分类任务和倒背数字广度任务为应用场景，基于重复错误、长时间犹豫和输入映射困难等练习阶段遥测信号，设计简短提示、规则特异性动态示范和支架式恢复练习等轻量化、可撤回支持机制。

基于清洗后的 52 名儿童会话数据及 8 名观察者配对评分，本文比较了标准练习与自适应练习条件下的任务进入、支持触发、练习恢复、正式阶段表现和现场可部署性。结果显示，自适应导入能够在练习阶段形成可审计的支持与恢复过程证据，并改善任务进入流畅性、错误后恢复、协调员负担和流程标准化；正式计分阶段未呈现一致的自适应得分优势，尤其在被直接支持的 DCCS 任务中，两种条件下正式准确率基本一致。

研究表明，有界自适应的目标不是提高正式测验得分，而是在不改变正式测量核心结构的前提下，以低算力、可审计、边界清晰的人机交互机制降低儿童评估中的非目标交互负担与无效流程成本。由于本文尚未直接测量设备能耗、云端调用成本或碳排放变化，相关发现应被理解为具有潜在低碳部署价值的初步证据，而非已经证明碳减排效果。

关键词：儿童执行功能；计算机化评估；低碳人工智能；自适应导入；测量效度

Low-Carbon Adaptive Onboarding for Child Executive Function Assessment

Lan Wei[1]　Ren Yilin[2*]

（1. Embodied Media Laboratory, Graduate School of Media Design, Keio University, Hiyoshi Campus, Yokohama, Kanagawa, 223-8526, Japan；2.Jinan University Zhuhai Campus, 519070, Zhuhai, Guangdong Province, China）

Abstract: Computerized executive-function assessment improves standardization, timing control, and automated scoring in child cognitive assessment; however, task-entry difficulties before scored trials

may be confounded with the target executive-function construct, while adaptive support that extends into scored trials may threaten measurement validity. This study proposes a measurement-preserving adaptive onboarding framework for low-resource and potentially low-carbon child psychological assessment. The framework confines support to task entry and practice, while locking formal stimuli, timing, response mappings, feedback policy, scoring rules, and summary metrics during scored measurement. Using the Dimensional Change Card Sort and backward digit span as application scenarios, the system provides lightweight, withdrawable, and auditable support through brief prompts, rule-specific animated demonstrations, and scaffolded recovery practice triggered by repeated errors, prolonged hesitation, and input-mapping difficulty. Based on 52 cleaned child sessions and paired ratings from eight observers, the study compares standard and adaptive practice in terms of task entry, support triggers, recovery, formal-block performance, and field deployability. Results show that adaptive onboarding generated auditable practice-stage evidence and improved task-entry smoothness, post-error recovery, coordinator burden, and process standardization. Formal-block metrics did not show a uniform adaptive score advantage; in the directly supported DCCS task, formal accuracy was broadly comparable across conditions. These findings suggest that bounded adaptivity should be understood as a validity-preserving interface strategy rather than an intervention for raising test scores. Because energy use, cloud-call cost, session-level carbon emissions, and invalid-test reduction were not directly measured, the low-carbon claim is limited to potential deployment value rather than demonstrated carbon reduction.

Keywords: child executive function; computerized assessment; low-carbon AI; adaptive onboarding; measurement validity

1 引言

执行功能（executive function, EF）通常指个体在目标导向行为中用于调节注意、抑制冲动、更新信息与转换规则的一组高级控制过程。已有研究普遍将抑制控制、工作记忆和认知灵活性视为执行功能的核心成分，且这些能力在儿童期经历快速发展，并与学习准备、课堂适应和心理健康密切相关[1][2][3]。在儿童评估场景中，计算机化执行功能任务逐渐取代部分纸笔或完全主试主导的测验形式，原因在于其能够标准化指导语、精确控制呈现时序、自动记录反应时与正确率，并在施测结束后生成结构化评分结果。NIH Toolbox 等标准化测验电池已将维度变化卡片分类任务（Dimensional Change Card Sort, DCCS）和 Flanker 等任务改编为计算机化形式，为儿童执行功能和注意能力测量提供了可参照的

工具基础[4]。与此同时，关于儿童与青少年计算机化认知测验的综述也指出，工具的操作效率提升并不意味着心理测量证据自然充分，效度报告、可迁移性和不同施测情境下的稳定性仍存在较大差异[5]。

这一张力在儿童任务进入阶段表现得尤为明显。儿童在正式试次开始之前，往往需要理解规则标签、把刺激属性映射到反应选项、掌握何时作答以及如何在数字界面上完成输入。对于成年人而言，这些要求通常只是界面操作问题；但对 5—9 岁儿童来说，它们可能与目标执行功能要求混合在一起。以 DCCS 为例，儿童既要判断当前依据颜色还是形状分类，又要在规则转换后抑制先前维度并启用新的分类依据[6]。已有元分析显示，DCCS 表现会受到指导语、冲突强调和任务结构影响[7]。进一步的研究也发现，诸如“颜色”“形状”一类维度标签本身并不总能有效引导儿童注意相关维度，标签理解可能成为任务进入的关键瓶颈[8]。在倒背数字广度任务中，儿童若按听到的顺序输入数字，可能是没有理解倒序规则，也可能是工作记忆操作能力不足；若反复按错键，则又可能是输入映射困难。可见，低分并不必然只代表目标认知能力较低。

实践过程中，主试通常会通过重复解释、临时提示或手势指引来帮助儿童继续完成任务。这样的帮助在现场有其合理性，但也带来了流程差异和可记录性不足的问题。不同主试对“何时帮助、帮助多少、何时停止帮助”的判断并不完全一致，结果可能使原本标准化的计算机化测验重新受到人工施测风格的影响。相反，如果系统在正式计分阶段持续提供自适应线索，又会直接改变任务实际测量的心理结构。表现性执行功能测验本来就与日常执行行为评定存在有限一致性，二者并非完全测量同一构念[9]。因此，评估界面中任何改变反应复杂度、任务线索或反馈结构的设计，都不能被简单视为心理测量中性的技术优化。

本文将上述问题界定为“任务进入支持与正式测量效度保护之间的边界问题”。在教育健康数字化场景中，低碳人工智能不宜被简单理解为“更大模型”或“更复杂推理”，而应包括减少无效会话、降低重复解释与重复测试成本、减少对高资源主试协调的依赖，并提高系统在普通学校、社区或基层健康服务场景中的可部署性。基于这一理解，本文提出面向儿童执行功能评估的低碳自适应导入设计：系统不依赖大模型或高算力推理，而是使用练习阶段遥测信号和规则化触发机制，在导入与练习阶段提供有限、可撤除、可审计的支持；正式计分阶段则锁定刺激、时序、反应映射、反馈、评分和汇总指标。换言之，本文的贡献不是提出新的 AI 技术突破，也不是证明自适应支持能够提升儿童执行功能得分，而是提出并初步验证一种“面向儿童计算机化评估的低资源友好型交互设计与测量效度保护框架”。

2 相关研究

2.1 儿童执行功能与计算机化评估

儿童执行功能研究长期关注抑制、工作记忆和转换能力在学前期与学龄初期的发展。综合性框架指出，学前儿童在执行功能任务中的表现并非由单一能力决定，而是由注意保持、规则表征、语言理

解、动机状态和动作反应等多种过程共同构成[10]。这对于计算机化评估具有直接启示：系统记录到的正确率或反应时，既可能反映目标执行功能，也可能反映儿童对界面、指导语和作答方式的熟悉程度。

计算机化测评的优势在于可重复、可追踪和可规模化。与纸笔形式相比，计算机系统能够保持试次呈现的一致性，减少人工计时和人工评分误差，并保存更细粒度的反应过程数据。[4]然而，自动化并不等于效度自动成立。相关综述显示，儿童和青少年计算机化认知测验工具在心理测量报告方面并不均衡，不同工具的信度、效度、常模和跨情境适用性存在差异[5]。这意味着，面向儿童的计算机化测评系统在追求效率的同时，也必须把界面设计纳入效度论证。

DCCS 和倒背数字广度代表了两类典型任务进入困难。DCCS 要求儿童根据当前规则在颜色和形状之间转换，规则标签、刺激冲突和维度映射构成了进入任务的前提条件[6][7][8]。倒背数字广度则要求儿童在保持序列的基础上进行倒序转换，并通过键盘或触控界面完成输入。二者都存在一个共同特征：练习阶段的错误未必直接说明目标能力不足，却会影响儿童能否以可解释的方式进入正式测量。

2.2 自适应评估与测量效度风险

技术化评估长期关注新题型、交付系统和自动化评分可能带来的测量机会与风险[11]。证据中心设计进一步强调，评估并不是简单收集行为数据，而是要建立从任务表现到心理构念推断的证据链[12]。在这一视角下，界面支持并非中性背景，而是影响证据链是否成立的重要条件。若支持改变了作答线索、反应负担或反馈结构，研究者就难以判断儿童正式得分究竟来自目标能力还是来自系统帮助。

自适应界面在一般人机交互研究中通常被视为提升可用性、效率和任务表现的手段[13]。个性化界面也常通过适配用户能力、设备特征和任务需求来提高操作成功率[14]。但评估界面与普通交互任务有所不同：普通系统可以把“表现更好”作为直接目标，评估系统却必须谨慎区分“帮助进入任务”和“帮助完成被计分任务”。这一差别也是本文所谓测量保持型自适应的出发点。

人机交互领域关于 AI 交互的指南强调，系统应向用户清晰呈现其能力边界，提供情境相关信息，并支持用户从错误或不确定状态中恢复[15]。这些原则同样适用于儿童评估，但需要被重新约束：提示可以帮助儿童理解练习规则，却不应在正式试次中充当持续线索；恢复机制可以降低练习阶段挫败，却不能替代正式阶段的独立作答。基于这一分析，可推测，自适应评估真正需要解决的并不是“是否提供帮助”，而是“帮助被允许发生在何处、以何种形式发生，以及如何被记录”。

2.3 绿色 AI 与低碳数字服务

绿色 AI 研究提醒我们，人工智能系统不应只以性能指标衡量，还应关注计算成本、能源消耗和资源可及性[16]。可持续 AI 进一步将问题扩展到 AI 系统的整个生命周期，包括构想、训练、调试、部署、治理与社会影响[17]。在大模型快速扩张的背景下，模型规模与资源消耗之间的关系也受到越来越多讨论[18]。这些研究共同提示，在教育健康评估这种需要高频、低门槛、可规模化部署的应用中，低

碳 AI 不宜被理解为“以更大模型实现更复杂判断”，而应更多体现为轻量化、可解释和可运维的系统设计。

本文所讨论的低碳自适应导入并不表明已经直接测量或降低碳排放，而是提出的是一种具有潜在低碳部署价值的过程基础：第一，系统采用规则触发和任务特异性示范，不依赖高算力推理或云端大模型调用；第二，自适应只在练习阶段运行，避免全程持续监控和复杂生成；第三，遥测记录让无效解释、重复尝试、会话中断和人工介入变得可见，为后续优化流程成本、测试时长和设备能耗提供数据基础。类似 Eco2AI 的工具已经为机器学习模型能耗与碳排追踪提供了方法参照[19]，而本文则把低碳视角推进到交互流程层面：减少不必要的计算，也减少不必要的现场消耗。需要强调的是，本文尚未直接记录设备功耗、系统能耗、云端调用成本、会话时长减少、无效测试减少或碳排放变化，因此“低碳”结论应限定为设计层面的潜在部署价值。

2.4 ESG 场景下的儿童教育健康数字化

儿童心理评估同时具有教育公平、健康促进和公共服务属性。若计算机化评估只能在高资源研究场景中稳定运行，其社会价值会受到限制；若系统过度依赖专业主试实时解释，也难以在学校、社区或基层机构中推广。由此看，儿童执行功能评估中的低碳 AI 设计位于 ESG 中“社会责任”与“环境责任”的交叉处：一方面，它关注儿童是否能公平、顺畅地进入任务；另一方面，它要求数字服务以更低的算力、流程和人力成本实现稳定部署。

普适计算与 HCI 测量系统为这种思路提供了启发。相关研究已经开始利用凝视、亲子互动、多模态信号和课堂行为等过程数据来支持儿童或教育情境中的行为评估[20][21][22]。生态瞬时评估和情境敏感干预研究也显示，良好的界面设计能够减少报告负担、提高数据密度，并在不完全替代专业判断的前提下提供过程证据[23][24]。本文在此基础上进一步强调，儿童评估中的自适应支持应当被边界化和审计化，使界面成为效度保护的一部分，而不是只追求更强的自动化干预。因此，本文的 ESG 贡献应被限定为儿童教育健康数字化中的社会可及性与潜在环境效率：一方面，通过降低任务进入障碍和人工协调差异，提高儿童在学校、社区和基层场景中获得标准化评估服务的可能性；另一方面，通过本地化、规则化和少干预的系统机制，为后续减少无效会话、重复测试和高算力依赖提供流程基础。

3 理论框架：测量保持型低碳自适应导入

本文提出的“测量保持型低碳自适应导入”由四项原则构成。第一是边界原则。自适应支持只发生在导入与练习阶段，不进入正式计分阶段。该原则并不否定儿童需要帮助，而是要求系统把帮助限定在儿童学习任务入口规则的阶段。一旦进入正式试次，系统停止提示、示范与支架恢复，儿童必须回到独立作答状态。

第二是测量保持原则。正式阶段的刺激材料、呈现时序、反应映射、反馈政策、评分规则与汇总指标均保持不变。这样设计的目的，是把练习阶段的“进入任务支持”与正式阶段的“目标能力测量”区分开来。若正式阶段仍改变刺激、延长时限或降低反应映射难度，即使得分提高，也难以说明儿童的执行功能表现真正改善。

第三是轻量化原则。系统使用练习阶段遥测信号、阈值规则和任务特异性示范来触发支持，而不依赖大模型、云端推理或复杂个性化生成。支持机制包括简短提示、动态示范和支架式恢复练习，均可以在本地或低算力设备上运行。这里的“智能”并不来自模型规模或新的 AI 技术突破，而来自对任务瓶颈的细化建模、对支持边界的明确限定，以及对正式测量条件的严格控制。

第四是可审计原则。每次支持触发的原因、支持等级、支持持续时间、错误类型、恢复结果和是否发生人工介入都被记录下来。这样做不仅有助于研究者解释练习阶段发生了什么，也能帮助施测团队判断哪些任务环节最容易造成非目标负担。换言之，遥测并不是附属日志，而是低碳自适应导入得以承担效度责任的关键机制。

表 1 测量保持型低碳自适应导入的设计原则

原则	操作化方式	效度与低碳意义
边界原则	自适应仅在导入和练习阶段运行，正式计分阶段关闭支持。	防止支持延续到被计分试次，避免污染目标构念。
测量保持原则	正式阶段的刺激、时序、反应映射、反馈、评分和汇总指标保持不变。	保障标准练习与自适应练习后的正式指标具有可比性。
轻量化原则	使用遥测触发、规则阈值和任务特异性示范，不调用高算力大模型。	降低计算依赖，并提高基层场景部署可行性。
可审计原则	记录触发原因、支持等级、恢复结果和人工介入等过程数据。	将任务进入困难转化为可追踪、可解释的交互证据。

4 系统设计

4.1 任务场景与系统流程

系统以 DCCS 和倒背数字广度作为核心应用场景。选择这两个任务，并不是因为它们覆盖了全部执行功能成分，而是因为二者都存在清晰的练习/正式边界，也暴露出典型的任务进入瓶颈。DCCS 中的主要困难来自规则标签、维度选择和目标类别映射；倒背数字广度中的主要困难则来自倒序转换、反应启动时机和数字键盘输入。系统流程如图 1 所示：儿童在练习阶段产生行为表现，监控器根据遥测信号判断是否需要支持，支持发生后儿童回到恢复练习，达到标准后进入正式测量锁定状态。

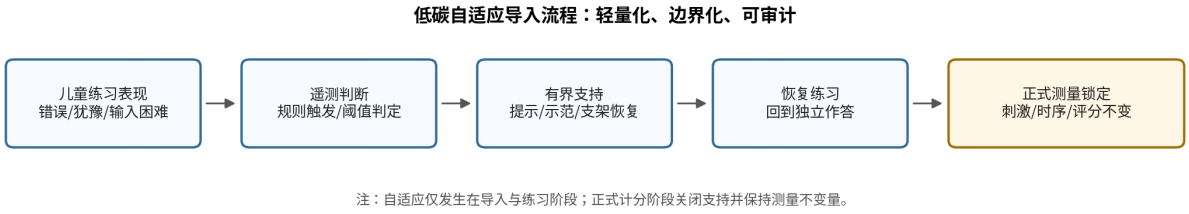

图 1　低碳自适应导入的系统流程图

图 2 实验示意图

4.2 遥测信号与触发规则

系统重点监测三类练习阶段遥测信号。第一类是重复错误，即儿童在同一任务瓶颈上连续出现相似错误；在当前系统协议中，重复错误以“同一练习瓶颈连续出现同类错误或达到协议版本预设次数”触发依据，而不是根据一次错误推断儿童能力水平。例如，在 DCCS 颜色规则下选择形状匹配目标，或在形状规则下选择颜色匹配目标，均被记录为规则/维度混淆事件。第二类是长时间犹豫，即儿童在练习反应窗口内超过预设上限仍未启动有效反应；该信号提示儿童可能尚未确定作答规则或输入方式，但不被解释为目标执行功能低下。第三类是输入映射困难，主要表现为无效按键、不完整输入、频繁删除、重复重输、输入顺序与界面要求不一致，或在倒背数字广度中将听到顺序直接作为输入顺序。

触发规则被设计得相对保守。系统不根据一次练习错误推断儿童潜在能力，也不尝试给儿童贴上能力水平标签。相反，它把这些信号视为界面事件：儿童可能需要任务进入支持。当前样本量不足以支持年龄分层阈值或个体化阈值拟合，因此 5-9 岁儿童使用同一套由协议版本锁定的保守触发规则；后续大样本研究可进一步检验不同年龄段是否需要差异化阈值。

表 2 遥测信号、触发规则与自适应支持退出机制

遥测信号/场景	触发条件	支持等级与内容	升级条件	退出条件	记录字段与人工介入规则
重复错误	同一练习瓶颈连续出现同类错误，或达到协议版本预设次数；DCCS 中主要表现为按非当前维度分类。	一级简短提示；若为规则/维度混淆，则进入二级规则特异性动态示范。	支持后独立练习仍出现同类错误。	独立练习达到当前协议通过标准，或达到最大练习次数并转入人工记录。	记录任务名、错误类型、支持等级、支持时间、恢复结果和人工介入标记。
长时间犹豫	练习反应窗口内超过预设上限仍未启动有效反应；不将其直接解释为能力不足。	一级简短提示，重申当前规则、作答时机或输入要求。	提示后仍未形成有效反应，或再次出现超时。	出现有效独立反应并完成恢复练习。	记录反应时、超时标记、提示类型、恢复试次和退出状态。
输入映射困难	无效按键、不完整输入、频繁删除/重输、顺序错误，或倒背数字广度中按听到顺序直接输入。	一级输入提示；持续困难时进入二级示范或三级支架式恢复练习。	支持后仍出现无效输入或顺序错误。	完成一次符合协议要求的独立输入，或达到最大尝试次数。	记录按键序列、删除/重输次数、顺序错误类型、支持等级和恢复结果。
人工介入	儿童疲劳、拒绝继续、设备异常或达到协议最大尝试次数。	暂停自适应流程，由研究人员按标准化脚本处理。	不再自动升级；仅记录事件。	恢复任务、终止任务或排除相应任务记录。	记录介入原因、介入时间、处理方式和任务特异性无效标志。

4.3 支持机制：提示、示范与恢复

自适应支持被组织为三个层级。一级为简短提示，通常用于首次轻度困难、轻微犹豫或超时，内容只提醒当前规则或作答要求，不直接替儿童完成反应。二级为规则特异性动态示范，用于重复同类错误、持续规则混淆或输入映射仍不稳定的情形。动态示范强调任务结构，例如在 DCCS 中先呈现当前维度，再呈现相关特征如何映射到目标类别；在倒背数字广度中则展示从听到顺序到倒序输入的转换过程。三级为支架式恢复练习，用于示范后仍失败或困难持续存在的情形，系统提供更明确的高亮与有限引导，但随后仍要求儿童回到独立作答。

支持阶梯不是单向增加帮助的过程。每次支持后，控制器都会安排独立练习机会；如果儿童成功，支持状态即淡出，系统继续练习或进入标准门槛；如果儿童再次出现同类错误，系统才可能升级支持。退出条件包括一次或若干次独立练习达到当前任务协议规定的通过标准、超过练习最大尝试次数后转入人工记录、或出现儿童疲劳/拒绝继续等现场安全条件。所有触发、升级、退出和人工介入均被写入日志。这种“支持—独立—再判断”的循环，体现了支架理论中临时支持和责任回交的基本逻辑[25][26]。同时，worked examples、表征性动画、信号化提示和反馈机制分别对应不同形式的认知支持[27][28][29][30]。不过，与学习系统不同，本文的支持目标不是训练儿童获得更高正式得分，而是帮助其以更标准化的方式理解任务入口。

4.4 低碳与可部署设计

从实现角度看，本系统的低碳部署价值首先体现为轻量化。练习监控、触发判断和支持呈现均可由本地规则完成，不需要在线调用大模型，也不需要根据每个儿童生成复杂内容。其次，低碳部署价值体现在边界化。由于自适应只在练习阶段运行，正式阶段关闭支持，系统减少了不必要的持续推理和全程适配。再次，低碳部署价值还体现在流程节约。若系统能把常见任务进入困难转化为标准化提示和可记录的恢复练习，就有可能减少主试反复解释、会话中断和无效测试。需要强调的是，本文没有直接测量能耗或碳排，也没有比较不同设备的功耗差异，因此相关结果只能被解释为“具有潜在低碳部署价值”，不能被表述为已经证明碳排放降低。

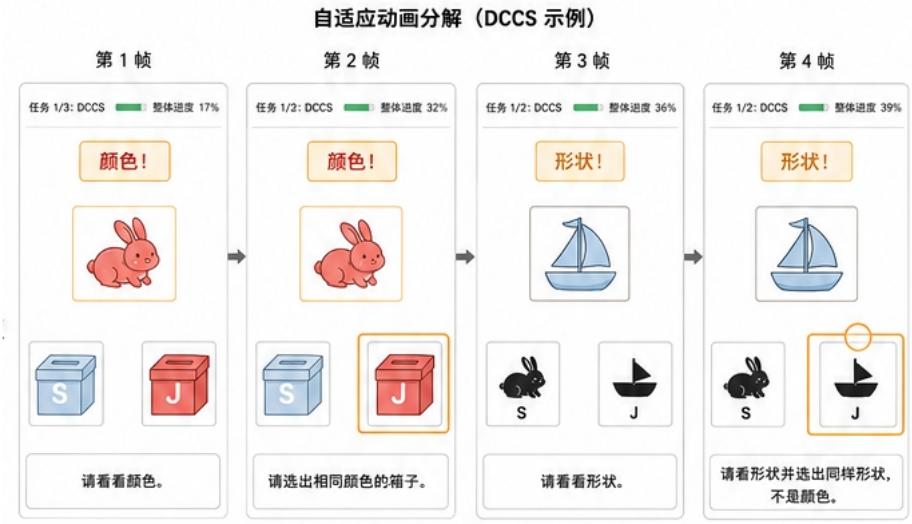

图 3 有界自适应导入 DCCS 实验动画示例

图 4 有界自适应导入倒背数字广度实验动画示例

5 研究方法

5.1 数据来源与研究条件

本研究使用清洗后的 52 名儿童会话数据，其中标准练习条件 26 例，自适应练习条件 26 例。儿童年龄为 5—9 岁（M=6.86，SD=1.00），其中男孩 35 名，女孩 17 名。每次会话由训练研究人员管理，系统记录实际交付的练习层、协议版本和遥测版本。需要说明的是，本研究以元数据中的实际练习层作为条件比较依据，属于初步设计验证和准实验性过程比较，并非严格随机分配实验。两组在年龄和性别构成上较为接近：标准练习组年龄 M=6.74（SD=0.99），自适应练习组年龄 M=6.99（SD=1.03）；标准组男孩 17 名、女孩 9 名，自适应组男孩 18 名、女孩 8 名。

标准练习条件提供固定指导语、固定练习项目和固定反馈。自适应练习条件在相同任务进入结构的基础上增加练习监控器；当监控器检测到重复错误、犹豫或输入映射困难时，系统在练习阶段提供相应支持。两种条件进入正式阶段后使用相同的正式任务模块，刺激、时序、反应映射、反馈规则、评分规则与汇总指标均保持一致。为提高可复现性，系统材料均为预先生成并嵌入本地程序的文本、图形和动画片段，不进行在线生成；输入方式主要包括鼠标/触控选择和数字键盘输入，系统记录每次练习与正式试次的时间戳、反应、正确性、反应时、错误类型、支持等级、恢复结果、协议版本、遥测版本和人工介入标记。屏幕尺寸、输入设备型号与具体功耗未被纳入当前统计分析，后续研究应将其作为部署环境变量系统记录。

表 3 清洗后的会话层分析集与人口学特征

特征	标准练习	自适应练习	总计
实验试次（N）	26	26	52
年龄 M（SD）	6.74（0.99）	6.99（1.03）	6.86（1.00）
年龄范围	5.00—8.50	5.00—9.00	5.00—9.00
男孩（n）	17	18	35
女孩（n）	9	8	17

5.2 测量指标

实验测量指标分为三类。第一类为过程指标，包括总提示次数、失败进入事件、DCCS 支持触发次数、DCCS 恢复成功次数、数字广度支持触发次数和数字广度恢复成功次数。这些指标用于判断自适应导入是否确实把练习阶段困难转化为可观察、可恢复的过程证据。

第二类为正式测量指标，包括 DCCS 转换准确率、DCCS 重复准确率、DCCS 转换成本反应时、倒背数字广度最大广度与总正确数，以及 Flanker 和变化检测等扩展任务指标。正式指标不是用来证明自适应支持改善执行功能能力，而是用于检查是否出现与测量边界不一致的一致性自适应得分优势。

第三类为部署指标，来自 8 名观察者的配对评分。观察者在审阅标准版本与自适应版本任务进入体验后，从任务进入流畅性、儿童参与、错误后恢复、协调员负担、流程标准化、可用性和可部署性等方面进行 1-7 点评分。分值越高表示评价越积极。

5.3 数据处理与统计分析

数据在会话层面进行去重和清洗。若某一任务记录存在任务特异性无效标志，则仅在涉及该任务的分析中排除该记录，同一会话中其他任务的有效数据仍予保留。练习阶段过程指标报告两条件均值、自适应减标准差异和 bootstrap 95%置信区间。Bootstrap 采用 2000 次重抽样并固定随机种子，在各条件内按会话重抽样。计数变量报告每次会话均值，并在必要时报告自适应条件下出现非零事件的会话数。

6 研究结果

6.1 自适应导入产生了可审计的支持与恢复过程证据

练习阶段遥测显示，自适应导入按设计产生了可审计的支持与恢复记录。标准练习条件不产生自适应提示和支持事件；自适应练习条件平均每次会话产生 0.81 次 HCI 提示，自适应减标准差异的 bootstrap 95%置信区间为[0.39, 1.35]。DCCS 支持在 26 个自适应会话中的 7 个被触发，平均每次会话 0.54 次，置信区间为[0.15, 0.96]。DCCS 恢复成功平均为 0.38 次，置信区间为[0.12, 0.73]；数字广度支持触发和恢复成功均为 0.27 次，置信区间分别为[0.08, 0.50]和[0.08, 0.46]。

表 4 练习阶段遥测指标

指标	标准均值	自适应均值	差值及 95% CI	自适应非零
总 HCI 提示	0.00	0.81	+0.81 [0.39, 1.35]	10/26
失败进入事件	0.46	0.54	+0.08 [-0.39, 0.54]	10/26
DCCS 支持触发	0.00	0.54	+0.54 [0.15, 0.96]	7/26
DCCS 恢复成功	0.00	0.38	+0.38 [0.12, 0.73]	6/26
数字广度支持触发	0.00	0.27	+0.27 [0.08, 0.50]	6/26
数字广度恢复成功	0.00	0.27	+0.27 [0.08, 0.46]	6/26

6.2 正式计分阶段未出现一致的自适应得分优势

正式阶段指标没有呈现一致的自适应得分优势。Flanker 不一致试次准确率在自适应条件下较高，标准均值为 0.935，自适应均值为 0.988，差值为+0.054，置信区间为[0.004, 0.121]；但数字广度和变化检测的部分指标则在自适应条件下略低。例如，数字广度最大倒背广度在标准条件下为 4.04，在自适应条件下为 3.69，差值为-0.35，置信区间为[-1.23, 0.59]；变化检测 Kmax 在标准条件下为 2.46，自适应条件下为 2.18，差值为-0.27，置信区间为[-0.64, 0.11]。

表 5 正式阶段主要执行功能指标

指标	标准均值	自适应均值	差值及 95% CI
Flanker 不一致准确率	0.935	0.988	+0.054 [0.004, 0.121]
Flanker 干扰反应时（ms）	106.03	82.45	-23.58 [-95.41, 53.05]
DCCS 转换准确率	0.931	0.928	-0.002 [-0.059, 0.050]
DCCS 重复准确率	0.933	0.932	-0.001 [-0.065, 0.061]
DCCS 转换成本反应时（ms）	232.98	174.83	-58.15 [-162.53, 43.20]
数字广度最大倒背广度	4.04	3.69	-0.35 [-1.23, 0.59]
数字广度总正确数	5.04	4.62	-0.42 [-1.84, 1.02]
变化检测准确率	0.849	0.827	-0.022 [-0.082, 0.038]
变化检测 Kmax	2.46	2.18	-0.27 [-0.64, 0.11]

6.3 DCCS 中被直接支持的任务正式准确率基本一致

DCCS 是本系统直接支持最充分的任务，因此其正式阶段结果对测量边界尤其关键。结果显示，DCCS 转换准确率在标准条件下为 0.931，在自适应条件下为 0.928，差值为-0.002，置信区间为[-0.059, 0.050]；重复准确率在标准条件下为 0.933，在自适应条件下为 0.932，差值为-0.001，置信区间为[-0.065, 0.061]。DCCS 转换成本反应时在自适应条件下数值较低，但置信区间跨越零。

这一结果具有双重含义。一方面，自适应导入确实在练习阶段对 DCCS 规则理解和维度映射提供了支持；另一方面，这种支持并未在正式准确率上表现为明显优势。对于测量保持型设计而言，这比单纯的成绩提升更重要。若自适应支持直接带来正式阶段准确率提高，研究者反而需要进一步担心支持是否改变了任务测量结构。

6.4 观察者评分显示任务进入和现场流程得到改善

观察者评分与遥测结果相互印证。自适应导入在协调员负担和流程标准化维度上的优势最大，标准条件均值为 4.58，自适应条件均值为 5.80，差值为+1.23，置信区间为[0.83, 1.68]。任务进入流畅性也明显倾向自适应导入，标准均值为 5.00，自适应均值为 5.83，差值为+0.83，置信区间为[0.53, 1.13]。错误后恢复与独立性、儿童参与以及可用性/可部署性维度也均呈正向差异。

需要注意的是，观察者并非盲法临床评定者，其评分反映的是对施测流程和现场可用性的判断，而不是对儿童执行功能能力变化的判断。因此，观察者评分最适合用于支持一个部署层面的结论：自适应导入有助于让儿童更顺畅地进入任务，并减少协调员在现场进行非标准化解释的压力。

表 6 观察者评分结果

维度	标准均值	自适应均值	差值及 95% CI
任务进入流畅性	5.00	5.83	+0.83 [0.53, 1.13]
儿童参与	5.30	5.75	+0.45 [0.20, 0.73]
错误后恢复与独立性	5.88	6.38	+0.50 [0.25, 0.80]
协调员负担与流程标准化	4.58	5.80	+1.23 [0.83, 1.68]
可用性与可部署性	5.80	6.18	+0.38 [0.18, 0.63]

6.5 低碳部署意义：减少非目标交互、重复解释和人工协调负担

从低碳部署角度看，本研究的结果提供了过程层面的支持，而非直接碳排放证据。自适应导入在练习阶段记录了哪些儿童需要帮助、需要何种帮助以及帮助后是否恢复，这为减少重复解释、无效练习和现场中断提供了可分析的基础。观察者对协调员负担和流程标准化的评分提升，也说明这种轻量化自适应机制有可能降低施测对高强度人工协调的依赖。由于本文未记录设备功率、会话能耗、云端调用成本、会话时长减少或无效测试减少，低碳结论应被限定为“具有潜在低碳部署价值”，而非已经实证证明碳排放降低。

因此，本文所说的低碳 AI 并不是以大模型替代主试，也不是通过更复杂的算法追求更高预测能力；它更接近一种“少而准”的交互策略：在确实出现任务进入瓶颈时提供有限支持，在正式测量前撤回支持，并把整个过程保留下来供研究者审计。此种策略的潜在价值在于提高低资源场景下的可部署性，同时避免以牺牲效度为代价换取表面流畅。

7 讨论

本研究围绕儿童计算机化执行功能评估中的任务进入困难，提出并初步验证了一种测量效度保持型的低碳自适应导入框架。不同于以提高正式测验得分为目标的自适应训练或智能辅导系统，本文所讨论的自适应导入主要服务于正式测量前的任务进入过程。换言之，系统并不试图在正式计分阶段帮助儿童取得更高成绩，而是通过轻量化、可撤回、可审计的交互支持，帮助儿童在进入正式任务之前理解规则、熟悉反应映射并恢复独立作答状态。研究结果显示，自适应导入能够在练习阶段形成明确的支持触发与恢复记录，并改善观察者评价中的任务进入流畅性、错误后恢复、协调员负担和流程标准化；与此

同时，正式计分阶段并未出现一致性的自适应得分优势，尤其是在被直接支持的 DCCS 任务中，两种条件下正式准确率基本一致。上述结果为“有界自适应”作为儿童计算机化心理评估中的界面效度策略提供了初步证据，但由于样本量较小且条件分配并非严格随机，其结论应被限定在初步设计验证和过程可行性层面。

7.1 测量效度贡献：从“表现优化”转向“测量保持”

本研究的首要贡献在于将自适应支持从“提高表现”的逻辑中区分出来，并将其重新定位为正式测量前的任务进入支持。在一般人机交互或智能学习系统中，自适应机制通常被视为提升操作成功率、学习效率或用户体验的有效手段。然而，在心理测量场景中，表现提升并不必然意味着测量质量提升。如果系统在正式计分阶段持续提供提示、线索、额外反馈或更宽松的反应条件，那么儿童得分的提高可能来自界面帮助，而非目标执行功能能力本身。尤其对于儿童执行功能评估而言，任务表现本就受到语言理解、规则表征、反应映射、注意维持和动机状态等多重因素影响，任何改变正式任务结构的支持都可能影响构念解释。

研究结果与这一设计目标基本一致。自适应导入在练习阶段产生了支持触发和恢复记录，说明系统确实识别并回应了部分儿童的任务进入困难；但正式阶段并未出现统一的自适应得分优势，尤其 DCCS 转换准确率和重复准确率在两种条件下几乎一致。这一结果不能被解释为严格的心理测量等效性证明，但至少降低了一个重要担忧：即自适应练习并未简单地转化为正式计分阶段的系统性得分膨胀。从效度角度看，这一点比单纯观察到成绩提高更有价值。若自适应支持直接导致被支持任务在正式阶段显著得分上升，研究者反而需要进一步追问这种提升是否源自测量结构被改变。

由此可见，本文对儿童执行功能计算机化评估的启示在于：评估系统中的自适应不应无条件追求“更聪明”或“更强干预”，而应首先明确哪些环节可以适配、哪些环节必须保持不变。对于 DCCS、倒背数字广度等具有明确练习/正式边界的任务而言，将自适应支持限定在任务导入与练习阶段，可能是一种相对稳妥的效度保护路径。这一设计并不能替代严格的心理测量验证，也不能直接证明测验具有跨群体等值性。然而，它为儿童计算机化评估提供了一个重要中间环节：在正式测量之前，系统可以帮助儿童理解任务入口，并记录支持过程；正式测量开始之后，系统停止帮助并保持任务条件一致。这样的边界设计，有助于把“儿童不会做任务”与“儿童目标能力较弱”区分开来。

7.2 HCI 贡献：将任务进入困难转化为可审计的过程证据

本文的第二项贡献体现在人机交互层面，即把儿童任务进入困难从以往难以记录、难以比较的现场经验问题，转化为可观察、可编码、可审计的交互过程证据。在传统儿童测评现场，儿童如果出现犹豫、重复错误、规则混淆或输入困难，主试往往会根据经验进行解释和干预。这样的人工协助在实践中非常常见，也有助于保障测验继续进行，但其问题在于：不同主试的介入时机、提示方式和帮助程度并不完全一致，研究数据中通常也难以完整记录这些临场支持如何发生。自适应导入系统则提供了另一种

处理方式。系统并不简单地把练习阶段错误视为儿童能力不足，而是将其细分为重复错误、长时间犹豫、输入映射困难等可识别的交互事件，并进一步记录支持触发原因、支持等级、支持阶段、持续时间、恢复结果以及是否发生人工介入。这样一来，任务进入阶段不再是正式测量前的一段“黑箱”，而成为可以被分析和优化的测量前过程。

从 HCI 角度看，本研究的价值不只是设计了一个更友好的儿童界面，而是提出了一种评估系统中的过程责任机制。系统不仅要呈现任务和记录得分，还要能够说明：儿童在进入任务时遇到了什么困难，系统为什么提供支持，支持后儿童是否恢复独立作答，正式计分开始前支持是否已经撤回。这种过程证据对于儿童评估尤其重要。儿童不是缩小版成人，其界面理解、语言加工、规则转换和动作操作都具有发展特征[31]。如果系统忽视这些发展差异，就可能把可避免的交互障碍误记录为认知失败。相反，如果系统能以标准化方式支持任务进入，并保留支持轨迹，研究者就能更谨慎地解释正式数据，也能更具体地改进任务设计。

7.3 低碳 AI 贡献：轻量化、边界化与少干预

本文的第三项贡献在于从低碳 AI 视角重新理解儿童心理评估中的“智能”。近年来，人工智能在教育和健康领域的应用常常与大模型、复杂预测和高算力推理联系在一起。但在儿童执行功能评估这样的具体场景中，真正需要的未必是更大的模型，而是更适度、更清晰、更能承担效度责任的交互机制。本文所提出的低碳自适应导入并不依赖大模型，也不依赖云端持续推理，而是通过练习阶段遥测、规则触发和任务特异性示范提供有限支持。

这种低碳性首先体现在轻量化。系统通过重复错误、犹豫和输入映射困难等可解释信号触发支持，不需要对儿童能力进行复杂建模，也不需要生成开放式个性化内容。支持形式主要包括简短提示、动态示范和支架式恢复练习，这些机制可以在本地设备上运行，具备较好的可维护性和可迁移性。其次，低碳性体现在边界化。由于自适应支持只在导入和练习阶段运行，正式计分阶段关闭支持，系统避免了全流程持续适配所带来的计算负担和效度风险。再次，低碳性还体现在流程层面。自适应导入通过标准化提示和恢复练习，有可能减少主试重复解释、会话中断、儿童反复失败和无效测试，从而为低资源学校、社区和基层健康服务场景中的部署提供基础。

本研究提供了一种具有潜在低碳部署价值的设计路径：以较低计算复杂度实现更稳定的任务进入流程，以可审计数据减少现场协调的不确定性，并在不改变正式测量核心的前提下提高系统可部署性。这样的低碳 AI 不是“用更大的模型替代人”，而是“用更少、更准、更有边界的支持减少无效流程”。所以，低碳 AI 与测量效度并不是相互冲突的目标。相反，轻量化和边界化可能有助于效度保护。系统越少在正式阶段进行复杂适配，正式测量结果越容易解释；系统越清楚记录支持过程，研究者越容易判断支持是否影响了测量核心。由于本文尚未直接测量能耗与碳排，上述讨论应被理解为设计理论与部署逻辑层面的贡献，而非碳减排效果的实证证明。

8 局限与未来研究

本研究存在若干局限。第一，样本量较小。52 个儿童会话能够支持初步设计验证和过程分析，但不足以开展更细致的年龄分层、个体差异建模或正式等效性检验，也不足以将结论推广到所有儿童执行功能测评场景。未来研究应扩大样本，并在不同年龄段、不同学校和不同施测环境中重复验证。第二，条件比较并非严格随机分配实验。虽然两组在年龄和性别构成上较为接近，但仍不能完全排除施测场景、主试、任务顺序、设备环境和样本清洗差异带来的影响。后续研究应采用随机对照或交叉设计，并预注册主要过程指标和正式测量指标。第三，本文没有直接测量系统能耗、设备能耗、云端调用成本、会话时长减少、无效测试减少或碳排放，也没有估计减少人工协调所对应的环境成本。因此，本文只能讨论潜在低碳部署价值，不能声称已经证明碳排放降低。后续研究可结合本地能耗监测、会话时长、重复测试率和人员投入记录，建立更完整的低碳评估指标。第四，观察者评分可能受到主观判断影响。虽然配对评分有助于比较标准练习与自适应练习的流程差异，但观察者并非盲法临床评价者，其结果更适合解释为部署可用性证据。第五，本研究主要验证了 DCCS 和倒背数字广度两个任务，不能直接推广至所有执行功能任务或其他儿童心理测验。第六，本文不能声称自适应导入提升了儿童执行功能测量能力，更不能解释为一种训练或干预效果。更谨慎的结论是，自适应导入改善了任务进入和流程标准化，并在现有数据中未显示出系统性正式得分膨胀。

9 结论

本文围绕儿童计算机化执行功能评估中的任务进入困难，提出并初步验证了一种测量保持型低碳自适应导入框架。该框架把自适应支持限定在导入与练习阶段，在正式计分阶段保持刺激、时序、反应映射、反馈、评分和汇总指标不变；同时通过遥测记录支持触发、错误类型、恢复结果与人工介入，使任务进入过程可追踪、可解释、可审计。基于 52 名儿童会话数据和 8 名观察者配对评分，研究发现自适应导入能够形成练习阶段支持与恢复证据，并改善任务进入流畅性、错误后恢复、协调员负担和流程标准化；正式阶段未呈现一致的自适应得分优势，DCCS 正式准确率在两条件下基本一致。需要强调的是，上述结果属于小样本、准实验性设计验证，不能被解释为强因果证据或正式等效性证明。

总体而言，有界自适应导入不是为了提高儿童正式得分，而是为了以低算力、可审计、效度保持的方式降低任务进入摩擦和现场部署负担。对于儿童心理评估中的绿色 AI 设计而言，这一路径的关键不在于追求更大模型，而在于用更轻、更边界清晰、更能承担效度责任的交互机制，支持儿童公平、稳定地进入正式测量。

附录：方法透明性说明

系统版本与数据处理说明：被评估系统为基于 Python/PsychoPy 的儿童执行功能电池，包含标准练习层与自适应练习层两个版本。会话元数据记录参与者编号、实际交付练习层、练习协议版本和遥测版本。自适应层仅在导入和练习阶段运行，正式任务阶段在两种条件下使用相同刺激、时序、反应映射、评分规则和汇总指标。系统材料包括预先制作的任务指导语、图形刺激和动画示范，不进行在线文本或图像生成；运行方式为本地程序调用，主要输入方式为选择反应和数字键盘输入。日志字段包括任务名、试次编号、练习/正式标记、刺激条件、反应内容、正确性、反应时、错误类型、重复错误标记、犹豫/超时标记、输入映射困难标记、支持等级、支持触发原因、恢复练习结果、退出状态、人工介入标记、协议版本和遥测版本。屏幕型号、设备功耗与具体能耗未在本研究中作为统计变量记录，后续低碳验证应补充设备与能耗层面的标准化采集。

统计说明：本文报告的置信区间均为 bootstrap 95%区间。儿童会话分析在条件内按会话重抽样，重复 2000 次并固定随机种子。除非另有说明，效应均为自适应练习减标准练习。计时变量因可能偏态而以描述性方式解释。观察者评分的 bootstrap 区间基于维度内配对观察者—条目差异计算。

【参考文献】

[1] Diamond A. Executive functions[J]. Annual Review of Psychology, 2013, 64: 135-168. DOI:10.1146/annurev-psych-113011-143750.
[2] Miyake A, Friedman N P, Emerson M J, et al. The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis[J]. Cognitive Psychology, 2000, 41(1): 49-100. DOI:10.1006/cogp.1999.0734.
[3] Best J R, Miller P H. A developmental perspective on executive function[J]. Child Development, 2010, 81(6): 1641-1660. DOI:10.1111/j.1467-8624.2010.01499.x.
[4] Weintraub S, Bauer P J, Zelazo P D, et al. NIH Toolbox Cognition Battery (CB): Introduction and pediatric data[J]. Monographs of the Society for Research in Child Development, 2013, 78(4): 1-15. DOI:10.1111/mono.12031.
[5] Tuerk C, Saha T, Bouchard M F, et al. Computerized cognitive test batteries for children and adolescents: A scoping review of tools for lab- and web-based settings from 2000 to 2021[J]. Archives of Clinical Neuropsychology, 2023, 38(8): 1683-1710. DOI:10.1093/arclin/acad039.
[6] Zelazo P D. The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children[J]. Nature Protocols, 2006, 1(1): 297-301. DOI:10.1038/nprot.2006.46.
[7] Doebel S, Zelazo P D. A meta-analysis of the Dimensional Change Card Sort: Implications for developmental theories and the measurement of executive function in children[J]. Developmental Review, 2015, 38: 241-268. DOI:10.1016/j.dr.2015.09.001.
[8] Buss A T, Nikam B. Not all labels develop equally: The role of labels in guiding attention to dimensions[J]. Cognitive Development, 2020, 53: 100843. DOI:10.1016/j.cogdev.2019.100843.
[9] Toplak M E, West R F, Stanovich K E. Practitioner review: Do performance-based measures and ratings of executive function assess the same construct?[J]. Journal of Child Psychology and Psychiatry, 2013, 54(2): 131-143. DOI:10.1111/jcpp.12001.
[10] Garon N, Bryson S E, Smith I M. Executive function in preschoolers: A review using an integrative framework[J]. Psychological Bulletin, 2008, 134(1): 31-60. DOI:10.1037/0033-2909.134.1.31.
[11] Bennett R E. The changing nature of educational assessment[J]. Review of Research in Education, 2015, 39(1): 370-407. DOI:10.3102/0091732X14554179.
[12] Mislevy R J, Steinberg L S, Almond R G. On the structure of educational assessments[J]. Measurement: Interdisciplinary Research and Perspectives, 2003, 1(1): 3-62. DOI:10.1207/S15366359MEA0101_02.

[13] Jameson A. Adaptive interfaces and agents[M]//Jacko J A, Sears A, eds. The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications. Mahwah: Lawrence Erlbaum Associates, 2003: 305-330.
[14] Gajos K Z, Weld D S, Wobbrock J O. Automatically generating personalized user interfaces with Supple[J]. Artificial Intelligence, 2010, 174(12-13): 910-950. DOI:10.1016/j.artint.2010.05.005.
[15] Amershi S, Weld D, Vorvoreanu M, et al. Guidelines for human-AI interaction[C]//Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York: ACM, 2019: 1-13. DOI:10.1145/3290605.3300233.
[16] Schwartz R, Dodge J, Smith N A, et al. Green AI[J]. Communications of the ACM, 2020, 63(12): 54-63. DOI:10.1145/3381831.
[17] van Wynsberghe A. Sustainable AI: AI for sustainability and the sustainability of AI[J]. AI and Ethics, 2021, 1(3): 213-218. DOI:10.1007/s43681-021-00043-6.
[18] Bender E M, Gebru T, McMillan-Major A, et al. On the dangers of stochastic parrots: Can language models be too big?[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2021: 610-623. DOI:10.1145/3442188.3445922.
[19] Budennyy S A, Lazarev V D, Zakharenkov I V, et al. eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI[J]. Doklady Mathematics, 2022, 106(Suppl 1): S118-S128. DOI:10.1134/S1064562422060230.
[20] Chong E, Chanda K, Ye Z, et al. Detecting gaze towards eyes in natural social interactions and its use in child assessment[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3): 43:1-43:20. DOI:10.1145/3131902.
[21] Kalanadhabhatta M, Rahman T, Grabell A S, et al. Tandem: At-home behavior assessment using multimodal signals from the parent-child dyad[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2025, 9(4): 182:1-182:25. DOI:10.1145/3770705.
[22] DiSalvo B, Bandaru D, Wang Q, et al. Reading the room: Automated, momentary assessment of student engagement in the classroom: Are we there yet?[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2022, 6(3): 112:1-112:26. DOI:10.1145/3550328.
[23] Ponnada A, Haynes C, Maniar D, et al. Microinteraction ecological momentary assessment response rates: Effect of microinteractions or the smartwatch?[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2017, 1(3): 92:1-92:16. DOI:10.1145/3130957.
[24] Paredes P E, Zhou Y, Hamdan N A, et al. Just Breathe: In-car interventions for guided slow breathing[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2018, 2(1): 28:1-28:23. DOI:10.1145/3191760.
[25] Wood D, Bruner J S, Ross G. The role of tutoring in problem solving[J]. Journal of Child Psychology and Psychiatry, 1976, 17(2): 89-100. DOI:10.1111/j.1469-7610.1976.tb00381.x.
[26] Reiser B J. Scaffolding complex learning: The mechanisms of structuring and problematizing student work[J]. Journal of the Learning Sciences, 2004, 13(3): 273-304. DOI:10.1207/s15327809jls1303_2.
[27] Atkinson R K, Derry S J, Renkl A, et al. Learning from examples: Instructional principles from the worked examples research[J]. Review of Educational Research, 2000, 70(2): 181-214. DOI:10.3102/00346543070002181.
[28] Höffler T N, Leutner D. Instructional animation versus static pictures: A meta-analysis[J]. Learning and Instruction, 2007, 17(6): 722-738. DOI:10.1016/j.learninstruc.2007.09.013.
[29] Schneider S, Beege M, Nebel S, et al. A meta-analysis of how signaling affects learning with media[J]. Educational Research Review, 2018, 23: 1-24. DOI:10.1016/j.edurev.2017.11.001.
[30] van der Kleij F M, Feskens R C W, Eggen T J H M. Effects of feedback in a computer-based learning environment on students’ learning outcomes: A meta-analysis[J]. Review of Educational Research, 2015, 85(4): 475-511. DOI:10.3102/0034654314564881.
[31] Druin A. Cooperative inquiry: Developing new technologies for children with children[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM, 1999: 592-599. DOI:10.1145/302979.303166.