人工智能医疗器械创新合作平台
ARTIFICIAL INTELLIGENT MEDICAL DEVICE INNOVATION AND COOPERATION PLATFORM

数字疗法产品人机交互性能指标评估方法

人工智能医疗器械创新合作平台数字疗法工作组

2025 年 10 月

目录

一、 研究评测理论：奠基与指导思想 1
二、 建立评测模型：构建评估框架 3
三、 梳理评测指标：定义可测量的维度 3
四、 提出评测方法：如何实施评估 5
五、 形成可量化、可实现的感知质量评估体系 7

本文件为北京市科技计划项目《精神疾病数字疗法产品测试验证技术服务平台构建及应用》的研究成果，可作为产品设计开发和成品检验的技术参考，不具有强制性。

引言

数字疗法（Digital Therapeutics, 数字疗法）是一种由软件驱动、基于循证医学证据的干预方案，用于预防、管理或治疗疾病。其临床有效性不仅取决于干预方案本身科学性与算法模型的精准度，也在很大程度上受到用户与产品交互过程的质量与体验的影响。在多种交互模态中，音频与视频作为传递信息、建立联系、提供指导的核心渠道，其流畅性、清晰度、自然度及所引发的情感共鸣，直接影响到用户的参与意愿、理解程度、使用依从性，并最终对疗效产生关键影响。因此，建立一套科学、系统、可量化的人机交互性能评估体系，有助于提升数字疗法产品的质量，并为证明其价值提供支持。

本体系基于这一需求，从理论、模型、指标、方法四个维度进行系统化构建，形成一个层次分明、多维量化且理论与实践紧密结合的评测框架。该框架旨在为数字疗法产品开发团队客观测量与持续优化产品人机交互性能提供支持，并为进一步探索交互体验与临床疗效的关联性提供方法参考。

面向数字疗法产品的人机交互评测体系

数字疗法指由软件驱动，基于循证医学证据的干预方案，用于预防、治疗或管理疾病，改善疾病预后。其有效性不仅取决于底层算法的科学性，更极大地依赖于用户与产品交互的体验质量。音频、视频作为最核心的人机交互模态，其流畅性、自然度和情感共鸣直接影响用户的参与度、依从性乃至最终疗效。

一、 研究评测理论：奠基与指导思想

一个稳健的评测体系通常需要建立在坚实的理论基础上，以确保其科学性与有效性。针对数字疗法产品中人机音视频交互的特性，本评估方法融合了多个跨学科理论作为支撑。以用户为中心的设计理论强调，评估活动应紧密围绕目标用户展开，尤其是处于特定疾病状态下的患者。因此，评测设计需充分考虑其可能面临的认知负荷、情绪波动及社会功能挑战，力求使评估场景更贴合真实用户需求。

媒介丰富度理论为评估音视频作为信息传递载体的效能提供了框架，重点关注其是否被优化用于高效传达医疗信息、传递共情与支持，而非引发用户信息过载或误解。社会代理理论指出，用户往往会将交互界面（尤其是虚拟角色）视为具有社会属性的实体，因此评测需考察虚拟代理是否成功塑造出专业、可靠、可信赖的形象，这对于建立治疗同盟具有重要意义。

从动机心理学的角度看，自我决定论指出，数字疗法的长期有效

性依赖于用户对自主感、胜任感与归属感的基本心理需求。音视频交互的评估需关注其在多大程度上支持这些需求的满足。最后，整个评测体系的方法论遵循循证医学与循证设计原则，强调所选用的评测指标与工具应具备科学依据和实证支持，而非依赖主观判断，从而提升评估结果的可信度与实际应用价值。

具体而言，针对数字疗法的音视频交互，我们融合以下理论:

1.以用户为中心的设计（UCD）理论：核心思想是设计和评测必须紧紧围绕目标用户（即患者）的需求、能力和局限性展开。在数字疗法中，需特别考虑用户可能存在的认知负荷、情绪状态和社会功能挑战。

2.媒介丰富度理论（Media Richness Theory）：视频和语音是“丰富”的媒介，能传递大量语言和非语言（如表情、语调）信息。评测需关注这种丰富性是否被用于有效传递共情、支持和医疗信息，而非造成信息过载。

3.社会代理理论（Social Agency Theory）：即使交互对象是虚拟角色或算法，用户也会将其视为社会实体。评测需关注音视频代理（如虚拟治疗师）是否成功建立了信任感、可靠性和专业性。

4.自我决定论（Self-Determination Theory, SDT）：数字疗法的成功依赖于用户的长期参与和内在动机。音视频交互应支持用户的自主感（感觉是自愿的）、胜任感（感觉能够完成）和归属感（感觉被理解和连接）。评测需衡量这些心理需求是否被满足。

5.循证医学（EBM）与循证设计（EBD）：评测指标和方法的选

择不应是主观的，而应基于已有的临床证据和设计最佳实践，确保评测本身的科学性和有效性。

二、 建立评测模型：构建评估框架

基于上述理论指导，本文构建了一个名为“APC-S”的四层分层评测模型。该模型旨在系统化地解构用户体验，从基础的技术性能逐步上升到主观感知和潜在疗效关联，形成一个层次分明且相互关联的评估框架。该模型尝试将用户体验从基础到高级，从客观到主观进行拆解:

A - 音频质量层（Audio Quality）：音视频交互的技术基础。评估声音信号的清晰度、连贯性和真实性。

P - 视频质量层（Video Quality）：评估视频信号的技术质量和视觉呈现的舒适度。

C - 交互通信层（Communication & Interaction）：评估音视频作为“沟通媒介”的有效性，包括信息传递和交互流程。

S - 感知与疗效层(Subjective Perception & Therapeutic Efficacy): 评估用户最终的主观感受和心理影响，这是与疗效直接相关的最高层。

APC-S 模型强调下层是上层的必要支撑，上层是下层的最终目标，例如，较差的音频质量会直接损害交互通信层的有效性，进而影响用户的主观感知和治疗依从性。该模型为系统性地定位问题、关联因果提供了结构化视角。

三、 梳理评测指标：定义可测量的维度

为“APC-S” 四层分层评测模型中的每一层模型定义具体、可量化的关键绩效指标（KPIs）。

表 1.“APC-S” 四层分层评测模型关键绩效指标（KPIs）

层级	评测维度	具体指标（示例）
A.音频质量层	清晰度	可懂度（%）、语音质量感知评估（PESQ 得分）、信噪比（SNR）
	连续性	卡顿次数、音频断线率、端到端延迟（毫秒）
	舒适度	音量稳定性、无背景噪音/回声、无尖锐刺响
P.视频质量层	清晰度	分辨率、帧率、视频质量感知评估（VMAF 得分）
	连续性	视频卡顿时长占比、花屏/马赛克发生率、端到端延迟（与音频同步）
	视觉舒适度	画面亮度/对比度适宜、无过度闪烁、虚拟形象表情自然度
C.交互通信层	信息传递有效	任务成功率(如: 正确理解指导语并执行)、信息回忆准确率、交互效率（完成特定任

	性	务所需时间/步骤）
	交互自然流畅性	响应延迟（用户结束说话到系统回应的时间）、中断率（用户被迫打断系统的次数）、对话非语言匹配度（如点头、微笑时机）
	系统代理表现	虚拟治疗师的语音情感表达丰富度（语调、节奏）、面部表情、共情表达准确性
S.感知与疗效层	主观体验	系统可用性量表（SUS）、接受度问卷（TFA）、信任度量表、共情商数
	情感与认知状态	积极/消极情绪量表（PANAS） 前后测、自我效能感量表、工作同盟量表（用于衡量用户与虚拟治疗师的关系质量）
	疗效相关行为指标	治疗任务完成率、用户粘性（周活/月活）、治疗脱落率、临床结局指标（如 PHQ-9、GAD-7 评分改善与音视频交互使用的相关性）

四、 提出评测方法：如何实施评估

采用混合方法研究，结合客观数据采集与主观用户反馈，在实验室和真实世界环境中进行。

1.客观性能测试（针对 A 层和 P 层）

工具：使用网络仿真工具（如 NS3、Appliance）模拟不同网络

条件（4G/5G/Wi-Fi， 弱网），利用音视频质量评测工具（如 ITU-T P.1203 标准工具）自动化采集 MOS 分、延迟、卡顿等数据。

实验室控制测试：在标准化的实验室环境下，控制变量，精确测量音视频基础性能指标。

2.可用性测试（针对 C 层）

任务型测试：招募目标用户群体（如特定病症患者），在观察室中完成一系列预设任务（如“跟随虚拟治疗师完成一次正念呼吸练习”）。记录任务成功率、耗时、错误点击，以及发声思维法下的用户口头反馈。

眼动追踪：通过眼动仪分析用户在交互过程中的视觉焦点，评估界面布局、虚拟代理注视是否自然，是否存在干扰因素。

3.主观评估（针对 S 层）

标准化问卷：在每次测试或特定使用周期后，要求用户填写上述 S 层提到的各类量表（SUS, TFA, PANAS 等），进行量化分析。

深度访谈与焦点小组：选取代表性用户进行定性研究，深入挖掘量化数据背后的“原因”，获取对交互体验、代理信任度、情感影响的深层洞察。

4.真实世界纵向研究（Correlating C & S with Efficacy）

A/B 测试：对不同版本的音视频交互设计（如不同的虚拟形象、语音语调、反馈节奏）进行分组测试，比较其在用户粘性、脱落率和临床结局指标上的差异。

后台数据分析：在产品上线后，匿名采集聚合级的 C 层行为数

据（如平均会话时长、互动次数）并与 S 层的疗效数据（如每次使用后的情绪自评）进行关联分析，建立交互行为－主观感知－临床效果的关联模型。

五、 形成可量化、可实现的感知质量评估体系

基于上述模型、指标和方法，本文整合形成一套可操作的流程供数字疗法产品开发团队参考:

1.设定基准：开发团队可根据产品定位和目标用户，为每个层次的关键指标设定合格线（Must-have）和优秀线（Nice-to-have）等参考数值。

2.构建评估矩阵：一个可行的方法是创建一个包含所有指标的评分矩阵，为每个指标赋予权重（例如，S 层权重 ＞C 层 ＞A/P 层）。权重可根据产品阶段调整（初期更关注技术性能，成熟期更关注疗效关联）。

3.综合评分算法：可考虑开发一个算法，将客观数据（如延迟毫秒数）通过模型映射为主观感知分（如 MOS 分），再结合问卷分数和行为指标，最终计算出一个整体感知质量指数（Perceptual Quality Index, PQI）。该指数可以是一个综合分数（如百分制）。

4.迭代反馈闭环：可将 PQI 和分项指标得分反馈给设计、开发和临床团队，以期用于指导产品迭代优化，形成一个“设计－评测－优化”的闭环。

5.标准化与报告：可考虑建立标准化的测试协议和报告模板，以

利于评测过程的一致性和结果的可比性(如版本间对比、竞品 分析)。