

心理治疗的忠实度测量：现状与改进

杨浩然[#] 孙焕翔[#] 刘晓明

(东北师范大学心理学院, 长春 130024)

摘要 心理治疗的忠实度(fidelity)指心理治疗按预定方案实施的程度,是保障临床实践质量,验证疗效可靠性的关键指标。为填补国内学界在心理治疗过程质量评估中忠实度研究的空白,本文系统梳理了其概念构成与测量必要性,比较了常用测量方法的优势与局限,并评述了现有改进策略的贡献与不足。综述现有研究发现:观察性编码、书面调查法和行为排练等方法虽各有特点,但存在耗时耗力、适用范围有限及评分一致性不足等问题;近年来,围绕测量框架建设、干预措施的精确描述和计算机辅助评估的探索,在一定程度上提升了忠实度测量的规范性与效率,但仍难以同时解决长期存在的“高成本”与“评估不一致”问题。未来研究需进一步完善忠实度测量框架,建立统一知识库,并持续推进自动化评估工具(如大语言模型)在心理治疗情境中的实证检验,从而推动心理治疗过程质量评估从理论倡议走向标准化实践。

关键词 忠实度, 心理治疗, 心理咨询, 测量方法, 评估框架

分类号 R395

1 引言

想象一下,你正饱受负面情绪的困扰,鼓起勇气走进了一间心理治疗室。治疗师看起来温和而专业,你们谈论了童年、人际关系和工作压力。几周过去了,你按时赴约,支付着不菲的费用,但内心深处的阴霾似乎并未散去,现实治疗效果与你的预期并不相符。你开始怀疑:治疗师真的运用了传说中有用的“心理治疗方法”吗?还是仅仅在进行一场昂贵的闲聊?那些治疗技术是否真的被付诸实践了?或者,治疗师是否不自觉地偏离了既定的方案?

上述情境不仅是来访者个体可能面临的困惑,更反映了心理治疗领域长期面临的重要挑战——如何确保治疗被准确、忠实且有质量地执行?大量研究证实,如果心理治疗得到规范实施,其临床疗效是显著的(Blanck et al., 2018; Kazdin, 2017; Öst & Ollendick, 2017)。但现实情况是,心理治疗的理论与实践间常存在鸿沟,预定的治疗方案与

实际的执行之间存在差异(Peterson et al., 1982; Peyser et al., 2021)。部分治疗师可能会偏离核心、循证的治疗重点,未能充分落实规定的治疗措施,或在干预中不加区分地混用其他理论取向的策略(Ahmadi et al., 2021);此外,同一治疗师的干预方式也可能随时间演变,对同一案例的早期和后期治疗在实施上也难以保持一致(Moncher & Prinz, 1991)。在国内,伴随着心理治疗队伍规模扩大的,是从业人员良莠不齐,专业水平参差不齐的乱象(陈祉妍等, 2016)。这些行业问题已通过临床实践影响到来访者。例如,一项针对心理咨询过程负面体验的质性研究发现,来访者认为一些咨询师存在能力与经验不足、专业态度欠佳,甚至违反伦理规范等问题(李牧阳等, 2025)。遗憾的是,当来访者产生负面体验时,通常缺乏直接证据来解释治疗效果不佳的原因,也难以判断问题根源在于疗法本身、治疗师的执行,还是来访者自身。从公众或监管等第三方视角看,心理治疗仿佛是一个神秘的“黑箱”,外界难以知晓其中具体发生了什么、哪些技术被使用、使用到了什么程度,以及督导或培训是否足以支持治疗师开展有效的治疗。这种不透明性既阻碍了治疗技术的传播,也可能削弱公众对心理治疗有效性的

收稿日期: 2025-12-30

[#]杨浩然和孙焕翔为本文的共同第一作者

通信作者: 刘晓明, E-mail: liuxm@nenu.edu.cn

信心,对行业的监管更无从谈起。当然,将这些问题完全归咎于治疗师并不公平,更根本的原因在于,当前普遍缺乏确保治疗被准确实施,并能被客观记录与评价的机制。已有论述指出,国内心理咨询和治疗领域存在对疗效的科学研究以及对专业工作的评估反馈机制重视不足等问题(陈祉妍等, 2016)。

在临床环境中,心理治疗的实施问题尚且能被察觉;而在研究中,这一问题则更为隐蔽。如果一个未被正确实施的新疗法被包装成有效的疗法进而推广,可能会带来更严重的后果。因此,无论是评估临床实施质量,还是验证新型或既有疗法的有效性,都离不开一个关键的前提——心理治疗是按照预定方案实施的,这被称为心理治疗的忠实度(fidelity, 亦译作保真度)。

国外学界对于忠实度的探索已久。20世纪70年代到80年代,“忠实度”的相关概念便已经在一些社会和行为研究中被讨论(如 Peterson et al., 1982; Quay, 1977)。此后,研究者逐步明确了忠实度的组成部分与评估重点(Moncher & Prinz, 1991; Waltz et al., 1993),推动了测量方法与工具(Bellg et al., 2004; Borrelli, 2011; Rodriguez-Quintana & Lewis, 2018)以及更系统的评估框架在心理治疗领域的发展与应用(Carroll et al., 2007; Des Jarlais et al., 2004; Hopewell et al., 2025)。与此同时,一些被广泛应用的心理治疗取向也完善了相应的忠实度评估工具,用于对治疗过程进行更标准化的质量控制与持续监测,涵盖认知行为疗法(Goldberg et al., 2020; Harstad et al., 2021)、动机式访谈(Schmidt et al., 2019)、人际关系疗法(Ravitz et al., 2025)、辩证行为疗法(Harned et al., 2021, 2023)、正念干预(Crane & Kuyken, 2019; Griffith et al., 2021; Sansom & Crane, 2025)等各类治疗方法。然而,时至今日,如何对忠实度进行系统性测量与报告,仍是未能妥善解决的难题(Beck et al., 2023),其在具体概念界定与测量方式等细节上仍存在模糊之处(Toomey et al., 2020),这限制了忠实度评估的推广。此外,在国内学界,仅有极少数文献提及心理治疗忠实度或其相关概念(艾小青等, 2012; 李宇等, 2019; 李晓苗, 李占江, 2014; 张静等, 2020),对其系统性的介绍尚属空白。

基于国内外的研究现状,本文将首先对心理治疗忠实度的概念进行辨析,进而对其测量的必

要性、常用方法及局限,以及现有的改进措施等问题进行系统的梳理总结,最后在此基础上从框架完善(6.1节)、知识库建设(6.2节)与大语言模型辅助测量(6.3节)三个方面,对未来研究方向进行展望。本文期望为国内心理治疗领域的研究与实践提供新的视角与思路,推动过程质量评估从理论倡议走向标准化实践。需说明,结合我国心理健康服务行业的现状及学界观点(陈祉妍等, 2016),本文对心理治疗和心理咨询两个术语不作严格区分。

2 心理治疗忠实度的概念辨析

心理治疗的忠实度,是指心理治疗在多大程度上按预定方案实施(Perepletchikova et al., 2007)。它是一个方法论上的概念,强调任何关于疗法有效性的判断、技术的传播及行业的监管,都必须建立在可检验的证据之上,即能够客观描述并量化治疗过程是否按预期发生(Moncher & Prinz, 1991; Perepletchikova et al., 2007; Rodriguez-Quintana & Lewis, 2018)。

首先,要准确把握忠实度这一综合性的概念,需对其构成要素进行厘清。忠实度的概念构成主要源于两篇奠基性文献(Waltz et al., 1993; Moncher & Prinz, 1991)。Moncher和Prinz(1991)认为,忠实度包括完整性(integrity)和区分性(differentiation)两个方面:前者指治疗师是否实施了特定的干预措施,后者指不同治疗方案在预期方式上的差异程度。随后,Waltz等人(1993)将忠实度进一步发展成依从性(adherence)和胜任力(competence)两个维度:依从性的含义与Moncher和Prinz(1991)所述的完整性基本一致,即遵循治疗手册的规定并避免使用禁止的干预措施;而胜任力则指治疗师在实施干预时所展现出的技能水平。在此基础上,Rodriguez-Quintana和Lewis(2018)对这些维度进行了整合,认为忠实度由依从性、区分性和胜任力三个部分组成。然而,亦有研究者认为,在明确了哪些要素不应被包含在治疗方案中后,对依从性进行测量就足以确立差异化水平,无需单独区分(Peyser et al., 2021; Waltz et al., 1993),故“区分性”可以不必作为一个独立维度进行评估。因此,评估的焦点主要集中在依从性和胜任力上。

其次,相较于容易理解的依从性维度,胜任

力的结构更加复杂。研究者通常将其细分为技术胜任力 (technical competence) 和基础胜任力 (common competence) (Brown et al., 2015; McLeod et al., 2013)。技术胜任力指治疗师在特定治疗模式或项目中实施特定治疗干预措施的能力, 其不同疗法中有不同要求。Waltz 等人(1993)将其解释为治疗师在实施干预时, 能在多大程度上结合治疗情境的相关因素(如来访者特征、治疗阶段、干预时机)并做出适当反应; 而基础胜任力则指一般的、适用于各种治疗模式的能力, 涵盖了工作同盟、个案概念化、元认知与自我监控, 以及临床判断与决策能力等多个方面。尽管已有研究尝试开发基础胜任力量表, 如针对特定群体的青少年心理社会治疗通用治疗师胜任力量表(Global Therapist Competence Scale for Youth Psychosocial Treatment, G-COMP) (Brown et al., 2018), 但能够提供详细维度且区分度高的基础胜任力评估工具依然稀少, 且研究表明技术胜任力和基础胜任力在当前测量中仍难以完全区分(Cecilione et al., 2021)。这种概念上的复杂性与操作性定义的不充分, 为胜任力的精确测量带来了挑战。

此外, 另有学者从来访者视角对忠实度概念进行了扩展。Lichstein 等人(1994)曾建议, 除了从治疗师角度评估实施外, 还应评估另外两个过程——治疗的接收(receipt, 来访者理解和展示治疗中学到的知识和能力的程度)和治疗的行动化(enactment, 来访者在生活中应用所学技能的程度)。然而, 一些学者对此提出质疑, 认为来访者的理解程度和运用技巧仅与干预效果相关, 而非与干预忠实度相关(Gearing et al., 2011)。

综合以上考量, 本文采纳 O'Shea 等人(2016)和 Toomey 等人(2020)的操作性定义, 将忠实度概念界定为依从性和胜任力两部分, 其中胜任力进一步划分为技术胜任力和基础胜任力。

3 心理治疗忠实度测量的必要性

从直觉上看, “确保心理治疗按照计划或预期实施”对部分读者而言可能显得陌生, 部分读者可能会质疑忠实度测量是一种僵化的做法。同时, 关于忠实度和心理治疗结果的实证研究亦呈现出不一致的结论。这些问题共同引发对忠实度测量必要性的疑问。因此, 在深入介绍忠实度测量方法之前, 有必要对这些争议进行澄清。

3.1 忠实度与心理治疗效果的关系

对于心理治疗忠实度水平与心理治疗效果之间的关系这一问题, 不同研究的结果相互矛盾(Perepletchikova, 2011)。大量研究报告揭示了高忠实度与更优治疗效果之间的正向关联。例如, Power 等人(2022)对 62 个成人心理治疗研究的元分析表明, 依从性和胜任力与临床效果密切相关, 其程度可与更广泛的常见因素相媲美; Esposito 等人(2024)对团体治疗的综述也指出, 大多数研究(82%)的忠实度与治疗效果存在正向相关。尤其是将治疗方案推广至临床机构时, 现有文献表明, 忠实度似乎是治疗结果的强有力预测因素(McHugh et al., 2009)。然而, 亦有研究未能发现忠实度与治疗效果间的明确关联。例如, Husabo 等人(2022)发现, 基于学校青少年焦虑干预措施的忠实度并不能预测临床结果; 类似地, Dimitropoulos 等人(2020)针对神经性厌食症青少年家庭治疗的大规模多中心研究也表明, 治疗师的依从性与患者达到理想的体重目标无关。

既然忠实度与心理治疗效果之间的关系尚不明确, 那么, 是否意味着忠实度测量并无必要? 首先, 从概念层面分析, 忠实度体现的是治疗干预实施的精确程度, 其作用类似于心理测量的“信度”。这样看来, 忠实度这一变量便不能脱离干预而独立存在。一个干预若未被实施, 则谈不上其忠实度。若干预未达到某个阈限(即“最低剂量”), 从研究角度看, 治疗师实际上并未执行有效的干预(即“不存在”干预)。在此情况下, 探究忠实度与结果的关系便失去了意义。这种绕过干预这一核心自变量, 直接探讨一个类似于“信度”的过程性指标(忠实度)与因变量(治疗效果)之间关系的研究思路, 可能存在“循环论证”的嫌疑。此外, 对于实证结果的矛盾, Perepletchikova (2011)认为, 忠实度在概念上与治疗结果无关, 其他变量(如治疗特征、治疗环境、咨访关系等)可能在其中起到调节或中介的作用, 混淆了对二者关系的解释。

其次, 从测量方法的角度讲, 忠实度与心理治疗效果关系的不明确, 可能源于忠实度测量本身不够严谨。例如, Brown 等人(2015)在测量忠实度胜任力维度时发现, 评分者间的一致性信度较低, 并由此推断, 以往研究的矛盾结果可能由对胜任力的测量不充分所致。Power 等人(2022)也指出, 早期研究所使用的评估工具可能未经验证或

针对性不足,实施测量时也缺乏随机抽样、双评等提高可靠性的程序。在这种情况下,受限于所纳入实证研究的忠实度评估质量,既有元分析得出不一致的结论便不足为奇。Peterson 等人(1982)将这种对因变量的观测极其严谨,而对自变量的实施情况却缺乏同等程度系统性观察的现象,称为“奇怪的双重标准”。这一植根于心理学研究的方法学局限,在面对心理治疗这类复杂过程时,其负面影响尤为凸显。

最后,心理治疗效果研究的主要目标是明确治疗方案并评估其可行性和有效性。对治疗效果好坏的任何解释,都必须以确保治疗方案按设计实施为前提(Pereplechikova et al., 2007),因此心理治疗忠实度便成为疗效研究在方法论层面的核心考察指标(McHugh et al., 2009)。高忠实度本身并不必然增强疗效,低忠实度的治疗也可能取得效果,但这仅意味着该治疗的实施方式与最初的预期不同(Pereplechikova, 2011)。在现实临床环境中是否需要绝对的忠实度是另一个议题,但只要存在评估治疗方案效果这一需求,都应以高忠实度为前提,而这离不开对忠实度的持续测量。

3.2 心理治疗“要忠实,还是要灵活”?

反对治疗忠实度评估的另一个常见角度是,对干预忠实度的僵化追求会限制治疗师的灵活性(或自主性),从而损害干预效果(Greenhalgh & Papoutsis, 2018)。必须承认,在现实环境中,一位高水平的治疗师有时确实需要采用一些不严格遵循既有治疗模式的方法,这甚至会体现出治疗师的胜任力水平(Dobson & Singer, 2005)。然而,“治疗师偏离方案”与“是否知晓并记录治疗师对方案的偏离”是两个截然不同的事情。假设一位治疗师在实践中发现“偏离治疗方案”对特定来访者有益,那么从推动心理治疗科学发展的角度看,合理的做法是忠实地记录下这次“偏离”及相应治疗情境。对该现象的进一步分析,将有可能促进心理治疗理论的发展。而“忠实地记录”这一过程,恰恰可以通过测量心理治疗忠实度来实现。回到忠实度与灵活性的争论,在研究中,应尽可能确保治疗的高忠实度,这是保证研究结论科学性的前提。在临床实践中,Kendall 和 Frank (2018)提出的“忠实度中的灵活性(flexibility within fidelity)”很好地回应了这个争论。他们指出,任何干预措施都包括不可被更改的核心成分(core components),

也包括可以根据实施环境和来访者特征进行调整的可变外围部分(adaptable periphery)。参考现有证据科学地实施治疗方案,并非机械地照搬手册,而是在保留核心成分的前提下,根据来访者的具体情况进行个性化调整(Kendall & Frank, 2018)。

我们应承认心理治疗是一个高度个性化的过程,“整合”各种可行手段是必然的发展趋势。但正如“忠实”不应被刻板地执行,“整合”也不应被盲目地实施。在何种条件下,运用何种手段,对来访者的帮助最有效?对于这一问题,专家或许可以凭借丰富经验快速地给出解答;但对于大多数治疗师而言,更高效、科学的做法是遵循现有的最佳“证据”来将干预措施应用到临床环境,因为实证研究表明这样做是有效的。进一步地,如何判断治疗师是否遵循了现有证据?这就需要测量治疗的忠实度。由此可见,忠实度的监测应贯穿从疗法开发到临床评估的全过程。

除了理论层面的辨析,初步的实证研究也表明,在忠实度得以维持的前提下,适度的灵活性并不会减弱治疗结果,这同样说明二者并非相互对立(McHugh et al., 2009)。无论如何,在研究和临床中,针对“改善心理治疗”这一共同目标,一个共同的需求便是能够明确治疗师在治疗中的具体行为。无论是遵守治疗方案,还是对方案进行灵活修改,都需要通过严谨的忠实度评估将整个治疗过程进行记录和测量。

3.3 忠实度在研究和临床中的作用

在澄清了与忠实度相关的争议后,可以发现,忠实度测量的必要性并未因这些争论而减弱。作为评估心理治疗实际实施情况的关键指标,忠实度对提高实证研究的内部效度和外部效度,以及促进临床实践中的质量评估,均具有重要意义(Brown et al., 2015)。

在保证内部效度方面,忠实度测量旨在排除从“实施治疗”到“产生结果”这一路径中其他混杂因素的影响(Pereplechikova et al., 2009)。在未测量忠实度的情况下,即使治疗结果显著,也无法确定该结果到底是由干预措施本身,还是由其他因素所驱动;当治疗结果不显著时,则无法判断是治疗方案本身无效,或治疗师实施不当,还是其他因素干扰所致(Moncher & Prinz, 1991)。只有以心理治疗忠实度被测量为前提,才能回答“该心理治疗方案本身是否有效”这一问题。更进一步,

可以采用定量方法评估治疗结果在多大程度上由特定干预技术引起,又在多大程度上源于其他非特异性的治疗实施变量,从而更好地区分干预措施的效果和治疗师—来访者间互动的效果(Bellg et al., 2004)。

在提高外部效度方面,忠实度评估有助于将有效的心理干预措施从研究情境推广至其他情境,包括在研究中复现和在临床中实践。研究间的可重复性常因对治疗描述不充分或实施过程中的特殊变化未被报告而降低(Moncher & Prinz, 1991)。无论是新的干预方案还是被广泛验证的治疗方法,研究者或心理治疗师都需要了解干预措施的细节及有效性信息,这些内容将帮助他们判断是否有必要对这种治疗方案进行深入研究或迁移应用。只有当一项干预措施忠实度高且疗效良好时,其他研究者或治疗师才更有信心进行深入探究或应用(Bellg et al., 2004)。反之,若一项成功的干预试验未能监测忠实度,盲目地进行推广可能会导致实际疗效大幅减弱,甚至获得失败结果,后续纠错的成本将十分高昂(Borrelli et al., 2005; Moncher & Prinz, 1991)。

除了保障研究的科学性,忠实度也可以作为质量评估的客观依据,为督导师和治疗师提供参考。尽管忠实度本身不直接预测治疗结果,但评估忠实度的过程提供了治疗师实施干预的详细信息。督导师可据此评估治疗师的优势与不足、能够胜任的干预类型,以及是否需要额外培训(Borrelli, 2011; Waltz et al., 1993)。此外,从忠实度评估中获得的信息也可以为治疗手册的修订提供参考,例如判断哪些干预措施描述得不够清晰,哪些要素是核心的或非核心的,以及干预的适用人群等(Pereplechikova, 2011; Waltz et al., 1993)。通过评估忠实度,可以将“制定治疗方案、编写治疗手册、培训并考核治疗师”这三个过程连接起来,形成Waltz等人(1993)所称的“反馈循环(feedback loop)”。总之,忠实度的评估可提高心理治疗的应用价值和实施质量,使相关各方获益。

不同于一般的行为研究,心理治疗是一种复杂的干预措施。其持续时间长、混杂因素多、操作步骤难以详细描述等特点增加了详尽报告干预细节的难度。然而,如果对干预研究或临床实践继续采取“宽容”态度,不要求评估忠实度,那么干预产生的结果就难以令人信服。测量忠实度的

作用,就是对治疗中的各类事件进行全面、客观地“概括”。相较于单纯比较忠实度量表的得分,评估过程中产生的关于实施细节的记录以及评估执行质量的标准等信息可能更有价值,因为明确这些内容的前提是,研究者或治疗师有信心确保其主张的干预措施是明确、清晰、可检验的。从这个意义上讲,评估忠实度的本质是在不泄露来访者隐私的前提下,要求研究者或治疗师披露干预设计和实施的细节以供第三方检验,从而提高心理治疗的可信度和有效性。因此,测量心理治疗的忠实度,是讨论其他问题的基础。

4 心理治疗忠实度的现有测量方法及局限性

4.1 忠实度测量的“金标准”——观察性编码法

忠实度的直接测量方法属于观察法。用于现场观察的设备有单面镜、监视器等,用于事后观察的材料有录像带、录音带、心理治疗转录稿等(Moncher & Prinz, 1991; Pereplechikova et al., 2007; Schoenwald & Garland, 2013)。观察的详尽程度也有所不同。一种是最基础的回顾性评价,即评估者在观察会谈后,根据记忆、简单笔记和总体印象对忠实度进行简要评价;另一种则被称为观察性编码(observational coding),评估者需按照预设标准,对会谈中出现的每个关键事件进行标注,并以此为据得出忠实度结果(Brown et al., 2015; Waltz et al., 1993),其材料来源既可以是事后材料,也可以是现场观察(Rodriguez-Quintana & Lewis, 2018)。更全面的观察性编码法被公认为忠实度评估的“金标准”(gold-standard)(Borrelli, 2011; McLeod et al., 2013; Rodriguez-Quintana & Lewis, 2018)。

4.1.1 观察性编码法的测量工具

观察性编码法主要使用的工具是检查清单(checklists)或评定量表(rating scales)。检查清单通常由研究者自行编制,使用最简单的“发生/非发生”二分指标。评估者对照清单勾选预期的治疗成分是否出现,最终以发生事件数占总事件数的百分比作为忠实度结果(Borrelli, 2011)。评定量表(忠实度量表)则提供更精细的评估。部分心理疗法拥有经过验证的忠实度评估工具,如认知行为疗法的忠实度评估量表(Cognitive Therapy Scale - Revised, CTS-R)(Blackburn et al., 2001)、动机式

访谈疗法的忠实度评估量表 (Motivational Interviewing Treatment Integrity, MITI) (Moyers et al., 2016)等。此外,也存在一些并不局限于特定疗法的通用忠实度量表,如青少年心理社会治疗通用治疗师胜任力量表(G-COMP) (Brown et al., 2018)。若研究涉及新疗法或某种疗法尚无现成工具,则需要研究人员自行编制量表(如 McLeod et al., 2025)。忠实度量表通常基于特定干预措施(如互动方式、谈话主题、具体行为等)出现的频率或广度,以及治疗师的胜任力水平高低进行评分,计分方式多样,如 3 点、5 点、7 点计分。研究者实际使用的指标各不相同,并且很难在研究报告中辨别出来(Schoenwald & Garland, 2013; Waltz et al., 1993)。在 Walton 等人(2020)的工作之前,学界长期缺乏规范编制忠实度量表的实用指南。Peyser 等人(2021)在一项针对治疗进食障碍的研究中发现,多数研究者倾向于为其特定研究开发未经检验的临时测量工具,即便少数研究使用了经过验证的量表,其选择也并不统一。在缺乏相对完善的工具时,研究者不得不重复自制细分领域的评估工具。这种做法不仅效率低下,也不利于研究间的横向比较。

4.1.2 观察性编码法的测量过程

综合较明确地描述忠实度观察性编码测量过程的文献(Behn et al., 2023; Brown et al., 2015; Devane et al., 2025; Hill et al., 1992; Jelsma et al., 2015; McLeod et al., 2025)及相应方法学标准(Heilemann et al., 2014),可将观察性编码的评估过程总结如下(见表 1)。

4.1.3 观察性编码法的优势与局限性

作为忠实度评估的“金标准”,观察性编码的优势有: (1)直接性。该方法的评估依据十分明确,能够捕捉治疗过程中变量的动态变化,为“过程—结果”研究提供数据(Waltz et al., 1993); (2)客观性。由独立评分者对治疗师的实际行为进行评估,排除治疗师或来访者主观因素的影响(Borrelli, 2011)。然而,该方法对人工评估的依赖也带来了两个核心问题。

首先是成本高昂。观察性编码需要聘请或培训熟练的评分者,这使其成为一项劳动密集且成本极高的工作(Ahmadi et al., 2021; Becker-Haimes et al., 2022; Peyser et al., 2021)。从设计忠实度评估流程、培训评估者,到最终实施评估,均需投入大量的人力物力(Pereplechikova, 2011)。在使用两名评分者的条件下,累计训练时间可达数十小时(Brueck et al., 2009; Schmidt et al., 2019; The Project SAFER Intervention Team et al., 2013)。在正式评估阶段,由于需要评估的会谈数量远多于训练阶段,所需时间将进一步增加。编码一份 1 小时的会谈材料需耗时 2.5 小时,即使是经验丰富的评分者也需要 30 分钟(Anis et al., 2021)。Hill 等人(1992)报告其团队总计花费了 50 个小时进行手册学习和试评,而正式评估周期更是长达 18 周。高昂的人力成本使得在研究中普及忠实度评估面临现实障碍,很难强求所有研究都提供忠实度的测量数据(Pereplechikova, 2011),更遑论将其推广至常规临床环境。

其次是缺乏统一的评估标准及由此导致的评

表 1 观察性编码法的测量过程

测量阶段	测量过程
I. 数据取样阶段	从所有待检查会谈中随机选取 10%~20%的样本作为测量对象。
II. 培训阶段	1. 确保评分者理解忠实度概念,熟悉评分表所有项目,明确评分标准。 2. 评分者对会谈进行试评,这些样本取自研究,但不用于最终统计。之后集体讨论以澄清分歧,直至对评分标准和项目理解达成共识。 3. 研究者根据讨论结果,进一步修订和完善评分标准。 4. 重复上述过程数次,每次均使用新的会谈材料,直至在各项目上达到令人满意的评分者间一致性(即评分者信度, inter-rater reliability)。
III. 忠实度评估阶段	不同评分者独立对正式样本进行评估。定期(如每周)组织培训和独立的编码练习,以维持评分者之间的编码一致性。
IV. 结果可靠性评估阶段	评估评分者间一致性。通常要求 Intraclass Correlation Coefficient (ICC)系数达到 0.75 以上,或 Cohen's Kappa 系数达到 0.70 以上。 评估评分者内一致性(intra-rater reliability, 即重测信度)。

估不一致性。第一,在测量实施层面标准化不足,如缺乏对抽样程序、治疗过程分段方式等的统一规定(Waltz et al., 1993),或培训评分者的最佳实践指南或模型(Rodriguez-Quintana & Lewis, 2018)。在缺乏统一标准的情况下,研究人员往往受限于实际资源而采用不同的测量方法,或忽视已有工具进行重复的开发(Waltz et al., 1993)。第二,在依从性评估层面,难以对复杂的治疗干预进行精确的操作性定义(Moncher & Prinz, 1991)。理想情况下,应制定一份基于理论的有效治疗要素清单,并确定评估这些要素实施程度的方法(Borrelli, 2011)。但实际情况是,一些干预措施(如建立脱敏层级、设定认知疗法议程)相对容易描述,而另一些干预措施(如在来访者存在阻抗时建立合作关系)则难以规范描述(Waltz et al., 1993)。为了在评分者之间达成对特定干预措施的一致理解,需要反复对其进行培训(Borrelli et al., 2005)。第三,在胜任力评估层面,普遍缺乏明确的定义和衡量标准(McLeod et al., 2013)。评估治疗师是否实施了某种技术(依从性)相对客观,但评估其执行质量(胜任力)则需要更多的经验参与(Violante et al., 2024)。事实上,目前尚无公认的判断治疗师最低胜任力的标准(Borrelli, 2011)。不同评分者因其隐含的判断标准和价值观不同,在胜任力评分时很难达成一致(Violante et al., 2024)。研究也证实,胜任力评估的评分者间一致性水平显著低于依从性评估(McLeod et al., 2025; Peyser et al., 2021)。第四,在忠实度结果解释上,亦缺乏区分不同水平的常模或标准(Miller & Rollnick, 2014)。即便得出了量表分数(Likert 等级),其绝对值也往往缺乏比较意义。标准的缺失引发了不一致问题;为了达成一致又需要更多的培训,从而增加成本;高昂的成本使得研究者难以投入足够资源评估忠实度,进而完善评估标准。由此可见,在缺乏标准化指南和依赖人工评估的双重作用下,成本高昂与评估不一致之间形成了恶性循环。这直接导致研究者缺乏评估心理治疗忠实度的动力(Perepletchikova et al., 2009),研究间的比较更无从实现,最终阻碍了心理治疗效果研究的进展。需要采取措施,分别针对标准缺失和依赖人工评估这两个弊端进行改进,以打破当前僵局。

4.2 忠实度的替代测量方法及局限性

尽管使用观察性编码法测量忠实度可以提供

精确的分析结果,但成本与不一致问题使其难以在资源紧张的研究和临床环境中开展。鉴于观察性编码法的局限性,研究者也在探索更简便的替代方法,致力于在常规的临床环境下,以较低成本获得可接受的忠实度评估结果(Hogue, 2022)。常见替代方法包括书面调查法(written survey)、行为排练法(behavioral rehearsal)等。

4.2.1 书面调查法

书面调查法以书面材料的形式进行。材料来源包括治疗师自我报告、治疗记录审查,以及来访者报告等(Perepletchikova et al., 2007; Schoenwald & Garland, 2013)。

治疗师自我报告,即要求治疗师通过问卷或清单报告其在会谈中实施的技术、频率、时间分配、是否遵循手册步骤等。鉴于其便捷性,治疗师自我报告可能是一种更可持续、负担更轻的评估方法。然而,治疗师可能因缺乏忠实度编码训练而难以量化自身行为(McLeod et al., 2023),或受社会赞许效应影响而高估干预措施的使用情况。此外,当治疗师同时管理很多案例时,若未在治疗后立即报告,记忆偏差也可能影响报告的准确性(Beidas et al., 2016)。

治疗记录审查通过对治疗师撰写的过程记录、电子病历或治疗计划进行编码,来判断计划的干预要素的实施情况。其本质是利用书面记录间接推断治疗的忠实度,一般使用结构化的编码模板进行评估(Wiltsey Stirman et al., 2022)。研究表明,治疗记录往往不能准确反映治疗技术的实施频率或强度(Stirman et al., 2021)。即便记录了治疗师对治疗过程中发生的事情的看法(提供依从性指标),也难以准确反映治疗师对各项要素的胜任力水平(Stirman et al., 2021)。

来访者报告,即通过来访者对会谈中所感知到的技术、互动模式与治疗体验的主观评价来推断治疗实施的程度。在这种方法中,来访者可通过问卷、访谈或治疗后的即时反馈报告其感受(McLeod et al., 2023)。通过审查来访者报告,能够得知其是否接受了所需的治疗成分、是否接受了不当的干预措施,还可以评估如治疗关系、是否感到被倾听或被催促、是否感到被尊重或被批评等非特异性成分。但与治疗师自我报告一样,来访者通常缺乏专业知识来准确描述具体的治疗内容(Borrelli, 2011)。同时,来访者也可能因不愿给

治疗师负面评价而提供不完全准确的信息(McLeod et al., 2023)。

4.2.2 行为排练法

行为排练法, 或称角色扮演(role-play)法/标准化病人(standard patient)法, 要求治疗师与受训的“演员”在标准化情境下进行模拟会谈。评估者通过直接观察治疗师在模拟情境中的表现来评估忠实度(Beidas et al., 2016)。该方法的优势在于情境高度标准化, 便于严格控制变量; 且评估者可对治疗师的行为(而非间接的书面材料)进行直接观察(Ottman et al., 2020)。然而, 由于角色扮演的语境是人为设计的, 这种方法的主要局限是无法反映治疗师在真实临床案例中的实际表现。因此, 或许更为恰当的看法是, 行为排练法更适合评估治疗师的“胜任力潜力”(实施干预的胜任力), 而非其在真实治疗中的“实施忠实度”(McLeod et al., 2023)。此外, 该方法同样面临与观察性编码类似的成本和评估不一致性问题(Ottman et al., 2020)。

4.2.3 替代测量方法的整体局限

总体而言, 书面调查法虽然成本更低, 但容易受到社会赞许效应、自我表征扭曲、对过去的认知改变以及记忆偏差等主观因素的影响(Borrelli, 2011; Perepletchikova et al., 2007)。研究表明, 治疗师自我报告、审查治疗记录或来访者报告与观察性编码所得结果的相关性较低(McLeod et al., 2023), 即便是经过简要培训的治疗师, 其自我报告的依从性水平依然存在显著高估(Becker-Haimes et al., 2022), 可靠性存疑。行为排练法虽然在方法上更为严谨, 但其评估的并非真实治疗的忠实度, 且未能解决成本与一致性难题。因此, 虽然这些替代方法为忠实度评估提供了更全面的视角(Borrelli, 2011; Perepletchikova

et al., 2007), 但目前更适合作为观察性编码法的良好补充, 而非作为独立的评估手段。

4.3 现有忠实度测量方法的优缺点总结

综上所述, 将忠实度传统测量方法的优势和局限总结如下(见表 2):

三种方法中, 书面调查法作为间接方法, 并不能提供完整的治疗信息; 行为排练法与观察性编码法相比并未展现出更明显的优势。因此, 改进忠实度评估的核心在于如何解决“金标准”——观察性编码法所面临的成本和不一致两大难题。当前的研究进展也主要围绕这两个方面展开。

5 心理治疗忠实度测量的现有改进措施及其局限性

为改进现有测量方法的问题, 研究者们主要从两个方向寻求突破: 一是以建立统一的评估框架与干预描述体系为代表的方法学创新, 旨在解决“评估不一致”问题; 二是以引入计算机技术辅助评估为代表的技术创新, 旨在应对“高成本”问题, 部分解决“不一致”问题。

5.1 从概念到实施, 建立忠实度的测量框架

心理治疗忠实度框架是一套用于组织、界定并系统化评估心理治疗实施情况的结构化概念体系, 其作用在于: (1)组织整合忠实度的各个组成部分, 确保术语统一; (2)提供忠实度测量的基本原则; (3)建立从治疗理论模型到可观察行为的映射规则。构建框架的目的是尽可能统一忠实度的测量, 解决不一致性问题。现有框架大致可分为专用于心理健康领域的概念框架、源于公共卫生领域的通用框架, 及更具针对性的专业实施框架三类。

5.1.1 概念框架

Carroll 等人(2007)提出了一个干预实施忠实

表 2 心理治疗忠实度现有测量方法的优势与局限

测量方法	优势	局限
观察性编码 (观察法)	1. 评估依据直接、清晰、明确。 2. 公认的评估“金标准”。	1. 依赖人工评估, 成本高昂。 2. 缺乏统一标准, 评估一致性难以保证。 3. 成本与不一致问题相互加剧, 形成恶性循环。
书面调查法	1. 评估成本低。 2. 可提供多方视角。	1. 间接评估, 易受自我表征扭曲、记忆偏差、社会赞许效应等主观因素影响。 2. 可靠性低, 仅适合作为补充依据。
行为排练法	1. 可严格控制评估情境。 2. 采用直接观察。	1. 无法评估真实治疗的忠实度。 2. 同样存在成本高和评估不一致问题。

度的概念框架。该框架将“依从性”视为干预与临床结果之间的核心中介变量,并将其细化为四个可操作的构成维度,即:(1)干预的“内容细节”是否被完整传递;(2)“覆盖率”(即剂量)是否充分;(3)“频率”是否符合预期;以及(4)“时长”是否达到要求。框架进一步指出,依从性的水平受到一系列调节因素的影响,包括治疗师对干预策略的理解、其提高干预实施质量的策略、干预的交付质量,以及来访者的反应性。关于忠实度与治疗结果的关系,该框架倾向于认为二者在概念上相互独立,但同时承认实施忠实度的水平会调节干预与疗效之间的关系。高忠实度的干预产生预期效果,而对治疗效果的评估和对关键组成部分的识别,则能持续改进干预措施本身,由此构成了一个忠实度的完整概念框架。

Carroll 等人的框架明确了忠实度在干预和疗效之间的逻辑关系,细化了其维度,并提出了相关的调节因素,为忠实度的测量提供了理论参考。但该框架亦存在以下问题:(1)未明确相关因素如何测量,导致在应用中存在操作化困难;(2)尽管通过“治疗师的干预交付质量”这一方面触及了“胜任力”维度,但未能充分阐明依从性和胜任力二者间的关系。

5.1.2 通用实施框架

公共卫生领域为规范干预试验报告而提出的通用框架,经适当调整后,可为心理治疗忠实度测量提供参考。其中,应用较广的有非随机试验报告规范(The Transparent Reporting of Evaluations with Nonrandomized Designs, TREND)框架(Des Jarlais et al., 2004)和临床试验报告统一标准(CONsolidated Standards Of Reporting Trials, CONSORT)框架(Hopewell et al., 2025)。二者均旨在提高研究的透明度和试验报告质量,内容上亦有相似之处,其主要区别在于前者适用于非随机对照试验,而后者适用于随机对照试验。

TREND 框架于 2004 年 3 月由美国疾控中心(CDC)艾滋病病毒预防研究综合(PRS)项目组提出,包括一个 22 项的检查清单,旨在指导非随机对照试验报告的标准化。在 TREND 框架中,与忠实度相关的是“干预”一项,要求报告试验中干预措施的 8 个细节:干预内容、干预方式、干预提供的单元、干预提供者、干预环境、暴露量和持续时间(会谈次数或事件次数)、时间跨度、以及为提高

被试依从性而设计的活动(Des Jarlais et al., 2004)。

CONSORT 框架最初于 1996 年提出(Begg, 1996),经过 2001 年(Moher et al., 2001)、2010 年(Schulz et al., 2010)两次修订,最新版本于 2025 年发表(Hopewell et al., 2025),包括一个 30 项的检查清单。早期版本的 CONSORT 仅笼统地要求“提供详细的干预措施”,并未明确包含与忠实度相关的条目(Nielsen et al., 2024; Schulz et al., 2010)。为弥补此不足,描述干预措施的清单和报告规范(Template for Intervention Description and Replication, TIDieR)于 2014 年被提出,用以扩展 CONSORT 声明,提供干预措施描述细节(Hoffmann et al., 2014)。最新的 2025 版 CONSORT 框架已明确纳入与忠实度相关的“干预和对照”项,并在其扩展检查单中解释了需报告的 8 个方面:干预和对照措施的组成部分、实施方法、实施时间和实施持续时间、个性化定制程序、所用材料、允许或禁止的附属干预、指导干预修改的标准,以及提高实验者或参与者依从性的措施(Hopewell et al., 2025)。

TREND 框架和 CONSORT 框架(及 TIDieR 清单)为干预措施的报告提供了详细模板,有助于提高报告完整性和可重复性。借助这些框架,可以为心理治疗忠实度应报告哪些维度提供参考。在与心理治疗忠实度测量的相关性上,CONSORT 因推出过专门针对社会与心理干预试验报告的指南 CONSORT-SPI 2018 (Montgomery et al., 2018)而更具优势。然而,与 Carroll 框架一样,这些通用框架同样没有就干预措施的操作化给出详细建议。TIDieR 清单虽然给出了详细示例,但主要针对医疗领域。此外,从维度上看,Lopez-Alcalde 等人(2022)指出,通用框架未能涵盖所有与心理干预忠实度相关的关键成分。针对这一问题,他们开发了专门针对心理干预的面对面心理干预完整性报告(Reporting Integrity of Psychological Interventions delivered face-to-face, RIPI-f)清单,声称其比 TIDieR 等清单更全面、更聚焦于心理治疗,但该工具的有效性尚待广泛验证。

5.1.3 专用实施框架

Bellg 等人(2004)扩展了 Lichstein 等人(1994)的模型(见本文第 2 节),为美国国立卫生研究院(NIH)与美国心脏协会(AHA)共同成立的行为改

变联盟(The Behaviour Change Collaborative, BCC)的治疗忠实度工作组提供了一个专用框架(以下简称为 BCC 框架),用于评估行为改变联盟从1999年开始赞助的研究项目的忠实度。BCC 框架强调在整个干预生命周期中通过结构化策略维持干预的一致性与质量,认为忠实度评估应涵盖5个过程:研究设计、培训提供、治疗实施、治疗接收和治疗行动化,并为每个环节提出了详细建议。Borrelli 等人(2005)随后利用该框架成功开发了疗效研究忠实度报告的评价标准,并在342项研究中验证了其有效性。BCC 框架在2011年得到了一次更新,增补了治疗提供者的选择、其他混杂因素、理论测试和多元文化考量等项目内容(Borrelli, 2011)。

Borrelli 等人(2005)认为,相较于 CONSORT 框架(Moher et al., 2001)和 TREND 框架(Des Jarlais et al., 2004),BCC 框架在忠实度评估这一专门领域上更为完善、更具针对性,可作为通用框架的实用补充。然而,亦有研究者指出其不足。Anis 等人(2021)认为,BCC 框架在忠实度的具体评估流程、评估方法选择,以及如何界定“依从”和“偏离”原有治疗方案等方面,仍缺乏明确指导。此外,从适用范围看,心理治疗并不完全等同于行为改变,前者包含治疗关系等复杂的非特异性因素(这些因素应在基础胜任力维度中体现);从概念构成上看,BCC 框架更侧重于依从性,对胜任力维度的阐述不足。因此,尽管 BCC 框架具有很大参考价值,但不宜将其直接迁移至心理治疗领域。

5.1.4 忠实度测量框架的局限性

虽然现有框架对忠实度的各个组成部分进行了明确规定,但它们可能没有充分考虑如何解决干预忠实度测量的实际操作问题(Toomey et al., 2020),即“只呼吁,不指导”。过于笼统的建议,例如“应提供治疗手册”,对于改善具体研究和实践的测量困境帮助有限。此外,现有框架也大多未能指导研究者如何确定干预的活性成分(治疗方案中真正起作用的成分),或如何将编码结果映射到最终的忠实度结果。从观察性编码方法的局限性上看,现有框架的精细化程度尚不足以解决评分者之间理解不一致的根本问题。在这种情况下,框架提出的建议越多,反而可能因操作上的模糊性而加剧成本问题(Toomey et al., 2020)。这

些局限给研究者和从业人员尝试应用框架带来了困难。

5.2 从分类到体系,对干预措施进行精确描述

在忠实度测量框架发展的同时,研究者也意识到,需开发一种严谨、具体的方法,来精确地描述干预活性成分。这促进了对行为改变技术(Behavior Change Techniques, BCTs)的研究。BCTs 被定义为干预措施中可观察、可复制且不可简化的组成部分,主要用于改变或重塑调节行为的因果机制(Michie et al., 2013)。每个 BCT 可被认为是一个“有效成分”,既可单独使用,也可组合使用。该方法旨在为干预措施提供“字典式”的精确、统一描述,以应对测量忠实度时因理解的不一致而产生的问题。

该领域较为典型的成果是 Michie 等人(2013)开发的行为改变技术分类法第一版(The Behavior Change Technique Taxonomy v1, BCTTv1)。该研究通过德尔菲法,将93项 BCTs 归为16个类别,构建了一个具有国际共识的干预措施分类系统。除建立分类“字典”外,该研究的另一重要贡献是提供了描述一项干预措施的方法。在文章随附的手册中,作者为 BCT 提供了操作性定义:(1)包含动词(例如:提供、建议、安排、提示),以指代实施者(治疗师或来访者)所采取的行动;(2)包含术语“行为”,指代单个行动或一系列行动,包括执行期望行为和/或抑制不期望的行为;(3)相关的替代或附加编码说明;(4)与特定理论框架关联的技术术语。例如,“行为反馈”的定义是:监测并提供关于行为表现的信息性或评价性反馈(例如,形式、频率、持续时间、强度)。BCTTv1 因其明确的定义方法和较高的扩展性而具备参考价值。

BCTs 分类法只是第一步,它在最小层面(相关研究常称之为“分子层面”)解决了干预措施的描述问题。而如何将一个完整的疗法“映射”到一系列 BCTs 上,则可以尝试用干预映射(Intervention Mapping)等方法解决(Kok et al., 2016)。该研究强调,一个完整的行为改变干预应遵循“确定行为目标→寻找决定因素→匹配理论方法→应用干预措施”的路径。其中,决定因素指的是促进改变的心理变量或调节机制,包括态度、自我效能、社会规范、意图、风险知觉、技能、自我调节等方面。将行为目标与决定因素组合为改变目标矩阵,依据“参数”(理论方法的使用条件)匹配合适的干

预方法，从而构成一个完整的疗法描述体系。

BCTs 的相关研究为精确描述干预措施提供了方法论参考，对解决心理治疗忠实度测量的不一致问题具有重要意义。然而，与 BCC 框架存在的问题相同(前述 5.1.3 节)，行为改变并不完全等同于心理治疗，不能简单地将 BCTs 的描述方案照搬到心理治疗。此外，该体系主要关注“实施了什么”(依从性)，并未涉及“实施得如何”(胜任力)。也就是说，它为干预措施存在与否的判断提供了解决方案，但未能回答如何评估干预措施的执行质量。

5.3 从人工到自动，使用计算机技术辅助评估

上述方法学创新尝试用更完善的前置知识来改善评分者之间的不一致(评分者信度)问题，但只要依赖传统的人工编码，高成本问题就无法得到根本解决。将目光转向计算机领域，或许能为这一难题提供思路。事实上，利用计算机算法对心理学材料进行编码的研究并不少见，如通过语言探索和字词统计(Linguistic Inquiry and Word Count, LIWC)系统进行情感、认知等心理特征分析(Pennebaker et al., 2001)。聚焦于心理治疗忠实度评估领域，Imel 等人(2019)使用机器学习方法编码动机式访谈文本；近期如 Chen 等人(2022)使用 BERT 语言模型，对认知行为疗法文本进行自动化编码；以及，Zainal 等人(2025)使用自然语言处理(Natural Language Processing, NLP)算法评估指导式数字认知行为疗法(GdCBT)等。心理治疗忠实度测量使用的主要材料——文本，恰恰属于机器学习擅长处理的对象。Ahmadi 等人(2021)对使用机器学习方法评估心理治疗的综述表明，自动编码方法的表现优于随机水平(50%)，部分方法

已展现出接近人类评分者的性能。

自动编码方法拥有突出的成本效益，尤其是在资源紧张的临床环境中优势明显。此外，对于高频、定义明确的行为，训练完善的自动编码模型可以绕过繁琐的人类评估员培训过程，在评估的一致性上表现更佳。但与此同时，传统的机器学习方法也存在局限。第一，模型的训练依赖庞大、高质量的标注数据。对于心理治疗来说，伦理等限制使得获取大量真实的治疗转录稿极为困难，如果没有足够数量的训练材料，很难保证机器学习模型的有效性。除了数量，模型的效果还依赖于训练集的质量(即人工标注的质量)。然而，高质量的人工标注十分昂贵，导致“冷启动”阶段的成本高昂。第二，泛化能力差。为某些特定疗法(如动机式访谈、认知行为疗法)训练的模型，通常难以应用于其他流派的疗法。因此，尽管自动编码算法在成本问题和不一致问题上具有潜力，但仍需更强大的算法和更精细的调整来提升性能。

综上所述，将忠实度现有改进措施的优势和局限总结见表 3。

6 研究展望

综上所述，心理治疗忠实度的测量既需要理论框架的指引，也需要前沿技术的支撑。现有测量体系在改进措施方面的不足，凸显了理论与技术革新的迫切需求。鉴于此，本节将从框架构建、标准制定与技术应用三个层面展望未来方向，并据此探讨忠实度测量进一步走向规范化的可能路径。

6.1 持续完善忠实度测量框架

一个理想的忠实度测量框架应保持适当的更新频率，以纳入该领域的最新进展。然而，部分被

表 3 现有忠实度评估改进措施的优势/局限分析

改进措施	改进成本问题	改进不一致问题
忠实度评估框架	——	优势：明确忠实度的测量内容以改善不一致问题。 局限：未提供可操作的测量指南。
干预措施分类法	——	优势：为描述心理治疗干预措施提供分类和描述方法，以改善不一致问题。 局限：仅涉及干预措施(依从性)，未涉及胜任力(包括技术胜任力和基础胜任力)。
计算机辅助评估	优势：摆脱人工评估的局限性，拥有巨大的成本优势。 局限：传统模型训练仍需耗费大量人力物力，且泛化能力差。	优势：通过模型自身的稳定特性改善人工评估的不一致问题。 局限：如果模型质量不佳，输出结果仍不可用。

广泛使用的框架已多年未有更新,如 TREND 框架自 2004 年被提出后至今未有修订;BCC 框架的上一次更新距今也已十余年。相比之下,用于规范干预措施报告的 CONSORT 框架的最新版本(Hopewell et al., 2025)较上一版本(Schulz et al., 2010)在结构和内容上均有优化,并提供了完整的检查清单和详细的解释文件,值得借鉴。

此外,框架内容的深化同样关键。完整的框架不仅应全面覆盖忠实度的核心维度,还应提供各维度的规范化描述方法及忠实度水平判定的区分规则。尽管现有主流框架(如 TREND/CONSORT、BCC 等)对依从性维度的划分相对完善(Borrelli, 2011; Des Jarlais et al., 2004; Hopewell et al., 2025),基础胜任力维度亦有 G-COMP 等量表可供参照(Brown et al., 2018)。但在“如何规范描述心理治疗干预措施”以及“如何确定测量胜任力行为锚点的方法”方面仍然不足。针对这一问题,未来的框架设计可借鉴 BCTTv1 分类法对治疗成分的定义逻辑(Michie et al., 2013),以及 Smith 等人(2014)在设计编码手册时对行为锚点的描述方法。此外,判断治疗师的胜任力水平时,可引入德雷福斯胜任力习得模型(Dreyfus & Dreyfus, 1980)。该模型将胜任力发展分解为新手、高级新手、胜任者、精通者及专家五个阶段,并对各阶段的典型特征进行了界定,这一思路已被成功应用于认知行为疗法忠实度评估的胜任力分级中(Blackburn et al., 2001),体现出良好的适用性。

最后,心理治疗忠实度框架的有效解释和应用,有赖于循证实践(evidence-based practice, EBP)理念的指导。循证实践理念强调,需在充分考虑来访者的特征、价值观与偏好的前提下,将当前最佳的研究证据与治疗师的专业技能进行整合(American Psychological Association, 2021; American Psychological Association & Presidential Task Force on Evidence-Based Practice, 2006; Melchert et al., 2024; 杨文登, 2012; 杨文登 等, 2017),这与忠实度的评估逻辑一致,二者均需以方案实施过程中的“证据”为核心。一方面,忠实度测量是循证心理治疗得以成功实施、维持和扩展的前提(Bond & Drake, 2020; Cook et al., 2017; Marriott et al., 2026)。另一方面,“什么内容值得被忠实实施”本身需要由循证证据来界定。没有循证研究提供依据,忠实度评估将失去参照,容易沦为对治疗过

程表面形式的记录。McLeod 等人(2013)进一步阐释了二者之间的关系:“依从性”评估了治疗师是否选择并实施了最佳的研究证据;“技术胜任力”考察了治疗师的临床专业技能;而“基础胜任力”则可以反映出治疗师是否真正地结合了来访者的特征、价值观与偏好。由此可见,在循证实践理念的指导下持续完善忠实度测量框架,对于提升临床实践质量,推动心理健康服务的持续改进具有重要意义(Goldman et al., 2001)。

6.2 建立忠实度统一测量知识库

忠实度框架主要解决概念(测什么)和方法(如何定义)的问题,但要解决测量中的“不一致性”难题,仅有框架是不够的,必须将其转换为包含具体治疗技术和实践方式的统一知识库。

首先,需从各类常用的心理治疗技术中确定有价值的治疗成分。Leijten 等人(2021)的一篇综述对治疗成分的提取策略进行了总结,包括专家观点、共享成分分析、治疗成分与治疗效果间的相关分析、对治疗成分的忠实度与治疗效果间的相关分析、微试验、加成试验与拆解试验,以及因子实验设计共 7 种策略。一个完整的忠实度测量知识库应能综合利用不同策略的优势,从各类治疗流派中灵活地提取治疗成分,为忠实度评估提供知识基础。

其次,需为心理治疗领域引入一套能将实证证据与临床经验有机结合、并在专家间达成广泛共识的机制。在现实实践中,对于“何种操作是适当的”这一判断常依赖于督导师的个人经验。但不同督导师的主观判断往往存在差异,对同一案例的看法可能各有不同(Rodriguez-Quintana & Lewis, 2018)。为解决此问题,可借鉴医学指南制定中的“证据到决策”(Evidence to Decision, EtD)框架的思路(Alonso-Coello et al., 2016)。EtD 提供了一种系统性方法,帮助专家审视客观证据,从而结合临床经验形成共识性的医疗决策建议。这些建议包含具体的行为指导(回答“如何做是合适的”)、也包含提出建议的详细理由及适用范围等,既可为医生提供指导,也可用于检验医生是否按照合适的标准执行医疗操作。若将 EtD 引入心理治疗领域,可在确定评估框架后,由行业组织牵头,邀请领域内专家定期审查证据,形成干预实施的指导性建议。这些建议能够将忠实度测量中客观的编码标注转化为对忠实度水平高低的评判,

最终将分散、不稳定的“专家经验”与客观证据相结合,形成统一的测量知识库。

当然,在提供详细指导的同时,也应避免标准的教条化,防止因过度追求忠实度而牺牲临床应用价值(Leventhal & Friedman, 2004)。具体操作步骤应由教科书或治疗手册详述,而知识库则聚焦于提炼干预措施的实施“要点”。将这些共识标准进一步整合为通用的评估规范与知识资源,是改进忠实度评估的重要方向。归根结底,完善此标准体系并彻底解决不一致性问题的真正瓶颈不仅在于技术,更在于“人”——即从业者的共识与协作。需要行业各方久久为功,从零开始逐步建立起一套稳固的体系。

6.3 使用大语言模型技术辅助忠实度测量

在测量实施层面,人类评分者因主观理解差异导致的的不一致问题,无法单纯依靠框架或知识库解决;高昂的人工评估成本也限制了忠实度测量的常规运转。因此,借助算法减轻人工编码负担、提升大规模评估效率,逐渐成为值得重点探索的改进思路。如前文所述(本文5.3节),传统机器学习方法存在高训练成本、泛化性差等弊端。而近些年迅速崛起的大语言模型(Large Language Model, LLM, 如 DeepSeek、ChatGPT、Gemini 等)为解决这一问题提供了契机,其在心理治疗领域中的应用也在近期的研究或观点中得到支持(Demszky et al., 2023; 黄峰 等, 2025; 朱廷劭, 2023)。LLM 是一类特殊的、基于 transformer 架构的语言模型,凭借在海量数据上的预训练及自身庞大的参数规模,它在许多自然语言处理任务中无需针对特定任务的训练即可取得卓越的性能(Kalyan, 2024)。

相较于机器学习方法,LLM 的突出优势在于其少样本(few-shot)学习能力。这意味着,仅需少量的示例和清晰的指令(提示词, prompts), LLM 便能执行各类复杂任务。在降低对特定领域大规模标注数据依赖(Kaplan et al., 2020)的同时,其效果仍能媲美甚至超越机器学习最先进的微调(fine-tune)范式(Brown et al., 2020)。更重要的是,心理治疗以语言为媒介,治疗师的技术实施与来访者的反应主要以言语形式呈现,忠实度测量的核心工作也因此集中在对言语内容与言语功能的识别、归类与量化之上(Gumz et al., 2015)。实证研究发现,LLM 可以从真实心理治疗逐字稿中有

效估计具有临床意义的心理构念(Abdou et al., 2025)。在此意义上,忠实度测量的本质属于 LLM 擅长处理的语言理解与语言分类任务(Brickman et al., 2025; Demszky et al., 2023)。因此,将 LLM 用于对心理治疗忠实度进行自动化量化评估,不仅在技术与方法上具备可行性,更有望摆脱传统机器学习方法对大规模领域数据和复杂模型训练的依赖。

同时也需承认,基于 transformer 架构的 LLM 在单次输出时本质上具有随机性(Atil et al., 2024)。然而,这种随机性并不会对 LLM 测量心理治疗忠实度的可靠性造成严重影响。第一,从原理上看,LLM 从海量语料中学习到的具有稳定中心趋势的语言统计规律,这使得模型倾向于输出训练语料中反复出现的稳定关联,保证了结果的平均一致性(Brown et al., 2020)。第二,从任务性质上看,心理治疗忠实度测量属于被严格限定的“评分任务”。除了待评估的心理治疗文本本身存在高变异性外,评估的输入格式可被规范化,输出结果的范围也被严格限定。因此,尽管 LLM 对提示词的输入形式高度敏感(Zhao et al., 2021),但可以通过严格的提示词工程(prompts engineering)对忠实度测量任务的边界、评分维度、输出格式与判定规则加以约束。在此,前述的框架和统一知识库发挥了关键作用。框架结构化的特性和知识库的准确性有助于向 LLM 提供高质量的提示词,从而提升 LLM 的稳定性。第三,从技术手段上看,已有多种技术手段能够提高 LLM 输出的一致性,如结构化输出(Structured Output)、集成(Aggregation)策略、事后校准(Post-hoc Calibration)等方法(Gu et al., 2026), LLM 自身的技术迭代也在持续改进其输出的稳定性。

在 LLM 出现以前,由于人工评估的成本问题长期无法解决,研究者只能尝试通过成本效益分析,对忠实度评估程序进行简化(Perepletchikova et al., 2007; Schoenwald et al., 2011),这是在尚无有效技术手段情况下的不得已妥协。而 LLM 技术的高效率和低成本,使得对心理治疗过程进行全面、细致的评估成为可能(Lim et al., 2025)。因此,应将 LLM 看作改进心理治疗忠实度测量的契机,而非盲目跟从潮流的“创新”。初步的实证研究已展现出利用 LLM 测量心理治疗忠实度的有效性(Wynia et al., 2025),但仍需更大规模的研究来进

一步验证其潜力与价值。

基于前述讨论可以发现,心理治疗忠实度的测量是一个以“规范性知识”为牵引、以“真实治疗文本”为对象、以“可检验的输出”为目标的结构化任务。即便在某些情境下忠实度的测量存在困难,但这不应成为忽视过程质量评估的理由。事实上,诸如心理动力学(Ogrodniczuk & Piper, 1999)、人本主义(Freire et al., 2014)、心智化治疗(Karterud et al., 2013)、家庭治疗(Robbins et al., 2011)等更强调关系性和体验性过程的心理治疗取向,也在积极开发自身的忠实度评估工具。这表明,尽管心理干预措施形式多样,治疗流派各具特色,但对忠实度的测量与监测很可能将逐步成为共同需求。未来研究应进一步聚焦在学校、社区等资源有限且问题性质复杂的环境,通过更多实证研究探索忠实度测量的适用性与有效性。

总的来说,本文为国内学界引入忠实度概念,旨在呼吁相关从业人员关注治疗的过程质量,从而提高各方对心理治疗服务有效性的信心。目前,我国政府机构和行业组织已发布一系列规范文件,如国家标准《GB/T 30446 心理咨询服务》(国家市场监督管理总局,中国国家标准化管理委员会,2018)、地区标准《DB31/T 1530-2024 心理咨询机构服务规范》(上海市市场监督管理局,2024),以及《心理咨询行业管理办法》(程万军等,2024)、《中国心理学会临床与咨询心理学工作伦理守则(第二版)》(中国心理学会,2018)、《中国心理学会学校心理工作伦理守则》(中国心理学会学校心理专业委员会,2026)等。如果能以这些扎实的工作为基础,汇聚多方力量持续完善忠实度测量体系,我国心理治疗过程质量评估的规范化水平有望得到进一步提升,进而保障心理健康服务领域的长远、有序发展。

参考文献

- 艾小青,曹玉萍,张亚林.(2012).内容分析方法在心理治疗研究中的应用.《中国心理卫生杂志》,26(5),373-376.
- 陈祉妍,刘正奎,祝卓宏,史占彪.(2016).我国心理咨询与心理治疗发展现状、问题与对策.《中国科学院院刊》,31(11),1198-1207.
- 程万军,贾晓明,钱铭怡,江光荣,王建平,肖旭,林秀彬.(2024).心理咨询行业管理办法.《中国临床心理学杂志》,32(5),1191-1194.
- 国家市场监督管理总局,中国国家标准化管理委员会.(2018).《心理咨询服务第1部分:基本术语》(GB/T 30446.1-2018)[国家标准].全国标准信息公共服务平台.
- 黄峰,丁慧敏,李思嘉,韩诺,狄雅政,刘晓倩,...朱廷劭.(2025).基于大语言模型的自助式AI心理咨询系统构建及其效果评估.《心理学报》,57(11),2022-2042.
- 李宇,万巧琴,于放.(2019).干预性研究中干预忠实度的提出与发展.《中华护理教育》,16(12),945-949.
- 李牧阳,雷霆,谢燮,国泽宇,刘可欣,陈雯青,...臧寅垠.(2025).心理咨询中来访者负面体验的质性探索.《心理学报》,57(7),1216-1230.
- 李晓苗,李占江.(2014).心理治疗疗效研究的方法学述评.《中国心理卫生杂志》,28(3),163-167.
- 上海市市场监督管理局.(2024).《心理咨询机构服务规范(DB31/T 1530-2024)》[地方标准].地方标准信息服务平台.
- 杨文登.(2012).《循证心理治疗》.北京:商务印书馆.
- 杨文登,李晓苗,张小远.(2017).心理治疗循证实践中“证据”的四个基本问题.《心理学报》,49(6),841-852.
- 张静,陶红霞,阮海慧,牛晓丹,迟俊婷,王艳红.(2020).动机式访谈法培训师的培养和实施保真度评估现状.《中华护理教育》,17(12),1134-1137.
- 中国心理学会.(2018).中国心理学会临床与咨询心理学工作伦理守则(第二版).《心理学报》,50(11),1314-1322.
- 中国心理学会学校心理专业委员会.(2026).中国心理学会学校心理工作伦理守则.《心理学报》,58(1),180-184.
- 朱廷劭.(2023).试析通用人工智能在心理学领域的应用.《人民论坛·学术前沿》,14),86-91.
- Abdou, M., Sahi, R. S., Hull, T. D., Nook, E. C., & Daw, N. D. (2025). Leveraging large language models to estimate clinically relevant psychological constructs in psychotherapy transcripts. *Computational Psychiatry*, 9(1), 187-209.
- Ahmadi, A., Noetel, M., Schellekens, M., Parker, P., Antczak, D., Beauchamp, M., ... Lonsdale, C. (2021). A systematic review of machine learning for assessment and feedback of treatment fidelity. *Psychosocial Intervention*, 30(3), 139-153.
- Alonso-Coello, P., Schünemann, H. J., Moher, J., Brignardello-Petersen, R., Akl, E. A., Davoli, M., ... GRADE Working Group. (2016). GRADE evidence to decision (EtD) frameworks: A systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*, 353, i2016.
- American Psychological Association. (2021). *Professional practice guidelines for evidence-based psychological practice in health care*. <https://www.apa.org/about/policy/psychological-practice-health-care.pdf>
- American Psychological Association, Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, 61(4), 271-285.
- Anis, L., Benzies, K. M., Ewashen, C., Hart, M. J., & Letourneau, N. (2021). Fidelity assessment checklist development for community nursing research in early childhood. *Frontiers in Public Health*, 9, 582950.
- Atil, B., Aykent, S., Chittams, A., Fu, L., Passonneau, R. J., Radcliffe, E., ... Baldwin, B. (2024). Non-determinism of "deterministic" LLM settings. *arXiv*. <https://doi.org/10.30446.1-2018>

- 48550/arXiv.2408.04667
- Beck, A. K., Baker, A. L., Britton, B., Lum, A., Pohlman, S., Forbes, E., ... Carter, G. (2023). Adapted motivational interviewing for brief healthcare consultations: A systematic review and meta-analysis of treatment fidelity in real-world evaluations of behaviour change counselling. *British Journal of Health Psychology*, 28(4), 972–999.
- Becker-Haimes, E. M., Marcus, S. C., Klein, M. R., Schoenwald, S. K., Fugo, P. B., McLeod, B. D., ... Beidas, R. S. (2022). A randomized trial to identify accurate measurement methods for adherence to cognitive-behavioral therapy. *Behavior Therapy*, 53(6), 1191–1204.
- Begg, C. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association*, 276(8), 637–639.
- Behn, N., Harrison, M., Brady, M. C., Breitenstein, C., Carragher, M., Fridriksson, J., ... Hilari, K. (2023). Developing, monitoring, and reporting of fidelity in aphasia trials: Core recommendations from the collaboration of aphasia trialists (CATs) trials for aphasia panel. *Aphasiology*, 37(11), 1733–1755.
- Beidas, R. S., Maclean, J. C., Fishman, J., Dorsey, S., Schoenwald, S. K., Mandell, D. S., ... Marcus, S. C. (2016). A randomized trial to identify accurate and cost-effective fidelity measurement methods for cognitive-behavioral therapy: Project FACTS study protocol. *BMC Psychiatry*, 16(1), 323.
- Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., ... Treatment Fidelity Workgroup of the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology*, 23(5), 443–451.
- Blackburn, I. M., James, I. A., Milne, D. L., Baker, C., Standart, S., Garland, A., & Reichelt, F. K. (2001). The revised cognitive therapy scale–revised (CTS-R). *Behavioural and Cognitive Psychotherapy*, 29(4), 431–446.
- Blanck, P., Perleth, S., Heidenreich, T., Kröger, P., Ditzén, B., Bents, H., & Mander, J. (2018). Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: Systematic review and meta-analysis. *Behaviour Research and Therapy*, 102, 25–35.
- Bond, G. R., & Drake, R. E. (2020). Assessing the fidelity of evidence-based practices: History and current status of a standardized measurement methodology. *Administration and Policy in Mental Health and Mental Health Services Research*, 47(6), 874–884.
- Borrelli, B. (2011). The assessment, monitoring, and enhancement of treatment fidelity in public health clinical trials. *Journal of Public Health Dentistry*, 71(s1), S52–S63.
- Borrelli, B., Sepinwall, D., Ernst, D., Bellg, A. J., Czajkowski, S., Breger, R., ... Orwig, D. (2005). A new tool to assess treatment fidelity and evaluation of treatment fidelity across 10 years of health behavior research. *Journal of Consulting and Clinical Psychology*, 73(5), 852–860.
- Brickman, J., Gupta, M., & Oltmanns, J. R. (2025). Large language models for psychological assessment: A comprehensive overview. *Advances in Methods and Practices in Psychological Science*, 8(3), 1–26.
- Brown, R., Iqbal, Z., Reynolds, L., Press, D. A., Shaker-Naeeni, H., Scrivener, L., ... Murphy, S. (2015). Inter-rater reliability of treatment fidelity and therapeutic alliance measures for psychological therapies for anxiety in young people with autism spectrum disorders. *International Journal of Developmental Disabilities*, 61(4), 190–199.
- Brown, R. C., Southam-Gerow, M. A., McLeod, B. D., Wheat, E. B., Tully, C. B., Reise, S. P., ... Weisz, J. R. (2018). The global therapist competence scale for youth psychosocial treatment: Development and initial validation. *Journal of Clinical Psychology*, 74(4), 649–664.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Brueck, R. K., Frick, K., Loessl, B., Kriston, L., Schondelmaier, S., Go, C., ... Berner, M. (2009). Psychometric properties of the German version of the Motivational Interviewing Treatment Integrity Code. *Journal of Substance Abuse Treatment*, 36(1), 44–48.
- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2(1), 40.
- Cecilione, J. L., McLeod, B. D., Southam-Gerow, M. A., Weisz, J. R., & Chorpita, B. F. (2021). Examining the relation between technical and global competence in two treatments for youth anxiety. *Behavior Therapy*, 52(6), 1395–1407.
- Chen, Z., Flemotomos, N., Singla, K., Creed, T. A., Atkins, D. C., & Narayanan, S. (2022). An automated quality evaluation framework of psychotherapy conversations with local quality estimates. *Computer Speech & Language*, 75, 101380.
- Cook, S. C., Schwartz, A. C., & Kaslow, N. J. (2017). Evidence-based psychotherapy: Advantages and challenges. *Neurotherapeutics*, 14(3), 537–545.
- Crane, R. S., & Kuyken, W. (2019). The mindfulness-based interventions: Teaching assessment criteria (MBI: TAC): reflections on implementation and development. *Current Opinion in Psychology*, 28, 6–10.
- Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S., ... Pennebaker, J. W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688–701.
- Des Jarlais, D. C., Lyles, C., Crepaz, N., & TREND Group.

- (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health*, 94(3), 361–366.
- Devane, N., Mazzoleni, S., Behn, N., Marshall, J., Wilson, S., & Hilari, K. (2025). Treatment fidelity in a feasibility trial of the aphasia intervention, virtual elaborated semantic feature analysis. *International Journal of Language & Communication Disorders*, 60(3), e70054.
- Dimitropoulos, G., Lock, J. D., Agras, W. S., Brandt, H., Halmi, K. A., Jo, B., ... Woodside, D. B. (2020). Therapist adherence to family based treatment for adolescents with anorexia nervosa: A multi-site exploratory study. *European Eating Disorders Review*, 28(1), 55–65.
- Dobson, K. S., & Singer, A. R. (2005). Definitional and practical issues in the assessment of treatment integrity. *Clinical Psychology: Science and Practice*, 12(4), 384–387.
- Dreyfus, S. E., & Dreyfus, H. L. (1980). *A five-stage model of the mental activities involved in directed skill acquisition* (Report No. ORC-80-2). University of California, Berkeley, Operations Research Center.
- Esposito, G., Di Maro, A., & Passeggia, R. (2024). The relationship between treatment integrity and outcome in group psychotherapy: A systematic review. *Clinical Psychology & Psychotherapy*, 31(1), e2952.
- Freire, E., Elliott, R., & Westwell, G. (2014). Person-centred and experiential psychotherapy scale: Development and reliability of an adherence/competence measure for person-centred and experiential psychotherapies. *Counselling and Psychotherapy Research: Linking Research with Practice*, 14(3), 220–226.
- Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J., & Ngeow, E. (2011). Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review*, 31(1), 79–88.
- Goldberg, S. B., Baldwin, S. A., Merced, K., Caperton, D. D., Imel, Z. E., Atkins, D. C., & Creed, T. (2020). The structure of competence: Evaluating the factor structure of the Cognitive Therapy Rating Scale. *Behavior Therapy*, 51(1), 113–122.
- Goldman, H. H., Ganju, V., Drake, R. E., Gorman, P., Hogan, M., Hyde, P. S., & Morgan, O. (2001). Policy implications for implementing evidence-based practices. *Psychiatric Services*, 52(12), 1591–1597.
- Greenhalgh, T., & Papoutsis, C. (2018). Studying complexity in health services research: Desperately seeking an overdue paradigm shift. *BMC Medicine*, 16(1), 95.
- Griffith, G. M., Crane, R. S., Baer, R., Fernandez, E., Giommi, F., Herbert, G., & Koerbel, L. (2021). Implementing the mindfulness-based interventions; teaching assessment criteria (MBI: TAC) in mindfulness-based teacher training. *Global Advances in Health and Medicine*, 10, 1–6.
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... Guo, J. (2026). A survey on LLM-as-a-Judge. *The Innovation*, 7, 101253.
- Gumz, A., Treese, B., Marx, C., Strauss, B., & Wendt, H. (2015). Measuring verbal psychotherapeutic techniques: A systematic review of intervention characteristics and measures. *Frontiers in Psychology*, 6, 1705.
- Harned, M. S., Korslund, K. E., Schmidt, S. C., & Gallop, R. J. (2021). The dialectical behavior therapy adherence coding scale (DBT ACS): Psychometric properties. *Psychological Assessment*, 33(6), 552–561.
- Harned, M. S., Schmidt, S. C., Korslund, K. E., & Gallop, R. J. (2023). Development and evaluation of a pragmatic measure of adherence to dialectical behavior therapy: The DBT adherence checklist for individual therapy. *Administration and Policy in Mental Health and Mental Health Services Research*, 50(5), 734–749.
- Harstad, S., Bjaastad, J. F., Hjemdal, O., Compton, S., Waaktaar, T., & Aalberg, M. (2021). Competence and Adherence Scale for Cognitive Behavioural Therapy (CAS-CBT) for anxiety disorders in youth: Reliability and factor structure. *Behavioural and Cognitive Psychotherapy*, 49(6), 745–757.
- Heilemann, C., Best, W., Johnson, F., Beckley, F., Edwards, S., Maxim, J., & Beeke, S. (2014). Investigating treatment fidelity in a conversation-based aphasia therapy. *Aphasie und verwandte Gebiete/Aphasie et domaines associés*, 37(2), 14–26.
- Hill, C. E., O'Grady, K. E., & Elkin, I. (1992). Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. *Journal of Consulting and Clinical Psychology*, 60(1), 73–79.
- Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ*, 348, g1687.
- Hogue, A. (2022). Behavioral intervention fidelity in routine practice: Pragmatism moves to head of the class. *School Mental Health*, 14(1), 103–109.
- Hopewell, S., Chan, A. -W., Collins, G. S., Hróbjartsson, A., Moher, D., Schulz, K. F., ... Boutron, I. (2025). CONSORT 2025 statement: Updated guideline for reporting randomised trials. *BMJ*, 389, e081123.
- Husabo, E., Haugland, B. S. M., McLeod, B. D., Baste, V., Haaland, A. T., Bjaastad, J. F., ... Wergeland, G. J. (2022). Treatment fidelity in brief versus standard-length school-based interventions for youth with anxiety. *School Mental Health*, 14(1), 49–62.
- Imel, Z. E., Pace, B. T., Soma, C. S., Tanana, M., Hirsch, T., Gibson, J., ... Atkins, D. C. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy*, 56(2), 318–328.

- Jelsma, J. G. M., Mertens, V. -C., Forsberg, L., & Forsberg, L. (2015). How to measure motivational interviewing fidelity in randomized controlled trials: Practical recommendations. *Contemporary Clinical Trials*, 43, 93–99.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling laws for neural language models. *arXiv*.
- Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6, 100048.
- Karterud, S., Pedersen, G., Engen, M., Johansen, M. S., Johansson, P. N., Schlüter, C., ... Bateman, A. W. (2013). The MBT adherence and competence scale (MBT-ACS): Development, structure and reliability. *Psychotherapy Research*, 23(6), 705–717.
- Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, 88, 7–18.
- Kendall, P. C., & Frank, H. E. (2018). Implementing evidence-based treatment protocols: Flexibility within fidelity. *Clinical Psychology: Science and Practice*, 25(4), e12271.
- Kok, G., Gottlieb, N. H., Peters, G. -J. Y., Mullen, P. D., Parcel, G. S., Ruiter, R. A. C., ... Bartholomew, L. K. (2016). A taxonomy of behaviour change methods: An Intervention Mapping approach. *Health Psychology Review*, 10(3), 297–312.
- Leijten, P., Weisz, J. R., & Gardner, F. (2021). Research strategies to discern active psychological therapy components: A scoping review. *Clinical Psychological Science*, 9(3), 307–322.
- Leventhal, H., & Friedman, M. A. (2004). Does establishing fidelity of treatment help in understanding treatment efficacy? Comment on Bellg et al. (2004). *Health Psychology*, 23(5), 452–456.
- Lichstein, K. L., Riedel, B. W., & Grieve, R. (1994). Fair tests of clinical trials: A treatment implementation model. *Advances in Behaviour Research and Therapy*, 16(1), 1–29.
- Lim, K., Jung, Y. C., & Kim, B. H. (2025). Evaluating motivational interview quality using large language models and hidden Markov models. *BMC Psychiatry*, 25(1), 908.
- Lopez-Alcalde, J., Yakoub, N., Wolf, M., Munder, T., Von Elm, E., Flückiger, C., ... Barth, J. (2022). The RIPI-f (Reporting Integrity of Psychological Interventions delivered face-to-face) checklist was developed to guide reporting of treatment integrity in face-to-face psychological interventions. *Journal of Clinical Epidemiology*, 151, 65–74.
- Marriott, B. R., Meyer, A. E., Feagans, A., Batch, B. L., & Adams, Z. W. (2026). Characterizing fidelity monitoring practices in community behavioral health care. *The Journal of Behavioral Health Services & Research*, 53(1), 45–64.
- McHugh, R. K., Murray, H. W., & Barlow, D. H. (2009). Balancing fidelity and adaptation in the dissemination of empirically-supported treatments: The promise of transdiagnostic interventions. *Behaviour Research and Therapy*, 47(11), 946–953.
- McLeod, B. D., Porter, N., Hogue, A., Becker-Haimes, E. M., & Jensen-Doss, A. (2023). What is the status of multi-informant treatment fidelity research? *Journal of Clinical Child & Adolescent Psychology*, 52(1), 74–94.
- McLeod, B. D., Southam-Gerow, M. A., Tully, C. B., Rodriguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice*, 20(1), 14–32.
- McLeod, B. D., Wood, J. J., Cecilione Herbst, J., Dunn, D., Kendall, P. C., Storch, E. A., Cho, A. C., & Rosenau, K. A. (2025). Treatment fidelity and outcome in CBT for youth with autism: The MEYA Fidelity Scale. *Journal of Clinical Child & Adolescent Psychology*, 54(1), 136–151.
- Melchert, T. P., Halfond, R. W., Hamdi, N. R., Bufka, L. F., Hollon, S. D., & Cuttler, M. J. (2024). Evidence-based practice in psychology: Context, guidelines, and action. *American Psychologist*, 79(6), 824–837.
- Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... Wood, C. E. (2013). The Behavior Change Technique Taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, 46(1), 81–95.
- Miller, W. R., & Rollnick, S. (2014). The effectiveness and ineffectiveness of complex behavioral interventions: Impact of treatment fidelity. *Contemporary Clinical Trials*, 37(2), 234–241.
- Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357(9263), 1191–1194.
- Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, 11(3), 247–266.
- Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). Reporting randomised trials of social and psychological interventions: The CONSORT-SPI 2018 Extension. *Trials*, 19(1), 407.
- Moyers, T. B., Rowell, L. N., Manuel, J. K., Ernst, D., & Houck, J. M. (2016). The motivational interviewing treatment integrity code (MITI 4): Rationale, Preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65, 36–42.
- Nielsen, A., Dyer, N. L., Lechuga, C., McKee, M. D., & Dusek, J. A. (2024). Fidelity to the acupuncture intervention protocol in the ACUpuncture in the emergency department for pain management (ACUITY) trial: Expanding the gold standard of STRICTA and

- CONSORT guidelines. *Integrative Medicine Research*, 13(2), 101048.
- Ogrodniczuk, J. S., & Piper, W. E. (1999). Measuring therapist technique in psychodynamic psychotherapies: Development and use of a new scale. *The Journal of Psychotherapy Practice and Research*, 8(2), 142–154.
- O'Shea, O., McCormick, R., Bradley, J. M., & O'Neill, B. (2016). Fidelity review: A scoping review of the methods used to evaluate treatment fidelity in behavioural change interventions. *Physical Therapy Reviews*, 21(3–6), 207–214.
- Öst, L. G., & Ollendick, T. H. (2017). Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: A systematic review and meta-analysis. *Behaviour Research and Therapy*, 97, 134–145.
- Ottman, K. E., Kohrt, B. A., Pedersen, G. A., & Schafer, A. (2020). Use of role plays to assess therapist competency and its association with client outcomes in psychological interventions: A scoping review and competency research agenda. *Behaviour Research and Therapy*, 130, 103531.
- Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC2001* [Computer software manual]. Mahwah, NJ: Lawrence Erlbaum Associates.
- Perepletchikova, F. (2011). On the topic of treatment integrity. *Clinical Psychology: Science and Practice*, 18(2), 148–153.
- Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: Survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology*, 77(2), 212–218.
- Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, 75(6), 829–841.
- Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, 15(4), 477–492.
- Peyser, D., Sysko, R., Webb, L., & Hildebrandt, T. (2021). Treatment fidelity in eating disorders and psychological research: Current status and future directions. *International Journal of Eating Disorders*, 54(12), 2121–2131.
- Power, N., Noble, L. A., Simmonds-Buckley, M., Kellett, S., Stockton, C., Firth, N., & Delgadillo, J. (2022). Associations between treatment adherence–competence–integrity (ACI) and adult psychotherapy outcomes: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, 90(5), 427–445.
- Quay, H. C. (1977). Measuring dimensions of deviant behavior: The behavior problem checklist. *Journal of Abnormal Child Psychology*, 5(3), 277–287.
- Ravitz, P., Heeney, N., Lawson, A., McAnanama, E., Pain, C., Kiss, A., ... Singla, D. R. (2025). Scaling up interpersonal psychotherapy training: A pilot randomized controlled trial of digital asynchronous self-directed vs. synchronous group workshop training. *Journal of Affective Disorders Reports*, 19, 100863.
- Robbins, M. S., Feaster, D. J., Horigian, V. E., Puccinelli, M. J., Henderson, C., & Szapocznik, J. (2011). Therapist adherence in brief strategic family therapy for adolescent drug abusers. *Journal of Consulting and Clinical Psychology*, 79(1), 43–53.
- Rodriguez-Quintana, N., & Lewis, C. C. (2018). Observational coding training methods for CBT treatment fidelity: A systematic review. *Cognitive Therapy and Research*, 42(4), 358–368.
- Sansom, S. A., & Crane, R. S. (2025). Mindfulness-based interventions: Teaching assessment criteria (MBI: TAC). In O. N. Medvedev, C. U. Krägeloh, R. J. Siegert, & N. N. Singh (Eds.), *Handbook of assessment in mindfulness research*. Springer.
- Schmidt, L. K., Andersen, K., Nielsen, A. S., & Moyers, T. B. (2019). Lessons learned from measuring fidelity with the Motivational Interviewing Treatment Integrity code (MITI 4). *Journal of Substance Abuse Treatment*, 97, 59–67.
- Schoenwald, S. K., & Garland, A. F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment*, 25(1), 146–156.
- Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research*, 38(1), 32–43.
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ*, 340, c332.
- Smith, J. E., Gianini, L. M., Garner, B. R., Malek, K. L., & Godley, S. H. (2014). A behaviorally-anchored rating system to monitor treatment integrity for community clinicians using the adolescent community reinforcement approach. *Journal of Child & Adolescent Substance Abuse*, 23(3), 185–199.
- Stirman, S. W., Gutner, C. A., Gamarra, J., Suvak, M. K., Vogt, D., Johnson, C., ... Resick, P. A. (2021). A novel approach to the assessment of fidelity to a cognitive behavioral therapy for PTSD using clinical worksheets: A proof of concept with cognitive processing therapy. *Behavior Therapy*, 52(3), 656–672.
- The Project SAFER Intervention Team, Seng, E. K., & Lovejoy, T. I. (2013). Reliability and validity of a treatment fidelity assessment for motivational interviewing targeting sexual risk behaviors in people living with HIV/AIDS. *Journal of Clinical Psychology in Medical Settings*, 20(4), 440–448.
- Toomey, E., Hardeman, W., Hankonen, N., Byrne, M., McSharry, J., Matvienko-Sikar, K., & Lorencatto, F. (2020). Focusing on fidelity: Narrative review and

- recommendations for improving intervention fidelity within trials of health behaviour change interventions. *Health Psychology and Behavioral Medicine*, 8(1), 132–151.
- Violante, S., McLeod, B. D., Southam-Gerow, M. A., Chorpita, B. F., & Weisz, J. R. (2024). Using adherence and competence measures based on practice elements to evaluate treatment fidelity for two CBT programs for youth anxiety. *Behavior Therapy*, 55(3), 605–620.
- Walton, H., Spector, A., Williamson, M., Tombor, I., & Michie, S. (2020). Developing quality fidelity and engagement measures for complex health interventions. *British Journal of Health Psychology*, 25(1), 39–60.
- Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology*, 61(4), 620–630.
- Wiltsey Stirman, S., La Bash, H., Nelson, D., Orazem, R., Klein, A., & Sayer, N. A. (2022). Assessment of modifications to evidence-based psychotherapies using administrative and chart note data from the US department of veterans affairs health care system. *Frontiers in Public Health*, 10, 984505.
- Wynia, Z., Delity, M., Morris, L., Costello, K., & Ashar, Y. (2025). Interrater agreement between a large language model and human raters for pain psychotherapy fidelity assessments. *The Journal of Pain*, 29, 105081.
- Zainal, N. H., Eckhardt, R., Rackoff, G. N., Fitzsimmons-Craft, E. E., Rojas-Ashe, E., Barr Taylor, C., ... Newman, M. G. (2025). Capitalizing on natural language processing (NLP) to automate the evaluation of coach implementation fidelity in guided digital cognitive-behavioral therapy (GdCBT). *Psychological Medicine*, 55, e106.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 12697–12706). PMLR.

Fidelity measurement in psychotherapy: Current status and improvements

YANG Haoran, SUN Huanxiang, LIU Xiaoming

(School of Psychology, Northeast Normal University, Changchun 130024, China)

Abstract: Fidelity in psychotherapy refers to the extent to which treatment is delivered as intended, serving as a key indicator of clinical quality and the reliability of therapeutic outcomes. Given the limited domestic research on fidelity in the evaluation of psychotherapy processes, this article reviews its conceptual structure and the rationale for its assessment, compares common assessment methods, and discusses current strategies for improvement. Existing studies indicate that observational coding, written surveys, and behavioral rehearsals each have distinct features and limitations. Across current approaches, major challenges include substantial demands on time and labor, limited applicability, and low scoring consistency. Recent efforts to develop assessment frameworks, specify intervention procedures, and introduce computer-assisted assessment have improved standardization and efficiency to some extent. However, the persistent issues of high costs and inconsistent assessment remain unresolved. Future research may further refine the assessment framework, establish a more systematic knowledge base, and continue exploring and validating automated tools such as large language models in psychotherapy settings, thereby advancing the quality assessment of psychotherapy processes from theoretical advocacy to standardized practice.

Keywords: fidelity, psychotherapy, psychological counseling, measurement methods, evaluation framework