

基于国际标准的教育实证研究评价量表的构建和优化

杨洋^{1*}, 张宗芳², 欧倩儿³, 吴苡欣⁴

¹北京师范大学人文和社会科学高等研究院教育科技中心, 广东 珠海

²枣庄市市中区孟庄镇周村中心小学, 山东 枣庄

³北京师范大学(珠海校区)乐育书院, 广东 珠海

⁴北京师范大学(珠海校区)会同书院, 广东 珠海

收稿日期: 2026年3月9日; 录用日期: 2026年4月21日; 发布日期: 2026年4月30日

摘要

科学规范的教育实证研究评价标准, 是保障教育实证研究质量的重要前提。本研究以美国教育研究协会的实证社会科学研究报告标准、EASE英文科研论文写作和翻译指南等实证研究论文的规范和标准为基础, 构建了实证研究评价工具; 并通过专家的反馈和具体对实证研究文章进行评分的实践操作, 优化了评价标准。同时研究结果还发现在教育实证研究评价标准中, 研究结果和研究结果讨论的描述是至关重要的。

关键词

教育实证研究, 评价标准, 评价量表

Construction and Optimization of Evaluation Scales for Educational Empirical Research Based on International Standards

Yang Yang^{1*}, Zongfang Zhang², Qian'er Ou³, Yixin Wu⁴

¹Center for Educational Science and Technology, Institute of Advanced Studies in Humanities and Social Sciences, Beijing Normal University, Zhuhai Guangdong

²Zhoucun Central Primary School, Mengzhuang Town, Shizhong District, Zaozhuang City, Zaozhuang Shandong

³Leyu College, Beijing Normal University at Zhuhai, Zhuhai Guangdong

⁴Huitong College, Beijing Normal University at Zhuhai, Zhuhai Guangdong

Received: March 9, 2026; accepted: April 21, 2026; published: April 30, 2026

*通讯作者。

文章引用: 杨洋, 张宗芳, 欧倩儿, 吴苡欣. 基于国际标准的教育实证研究评价量表的构建和优化[J]. 创新教育研究, 2026, 14(4): 552-561. DOI: 10.12677/ces.2026.144303

Abstract

Scientific and standardized evaluation criteria for educational empirical research are the important prerequisites for improving the quality of educational empirical research. Based on the standards and standards of empirical research papers, such as the American Educational Research Association's Standards for Empirical Social Science Research Reports and the EASE Guidelines for Writing and Translating English Research Papers, this study constructs an empirical research evaluation tool. Through the feedback of experts and the practical operation of scoring empirical research articles, the evaluation criteria were optimized. At the same time, the results of the study also found that the description of the research results and the discussion of the research results is crucial in the evaluation criteria of educational empirical research.

Keywords

Educational Empirical Research, Evaluation Criteria, Evaluation Scales

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

改革开放以来，中国在教育领域见证了教育资源的极大扩充和教育体系的全面优化，标志着教育理论与实践的深刻革新。教育研究在这一背景下获得了前所未有的发展机遇，逐步走向成熟与国际化。2013年初，全国教育科研工作会提出了“没有教育科学，就没有科学的教育”这一重要论断，强调了教育科学的核心地位和基础性作用，并将实证研究作为教育研究的发展重心[1]。教育实证研究是制定基于证据的教育政策的基石，对全面提高教育质量和效率，实现教育现代化起着至关重要的作用。2015年10月，由华东师范大学教育学部、北京师范大学教育学部、全国教育科学规划办和光明日报社教育研究中心联合主办的“全国首届教育实证研究论坛”在华东师范大学举行，大大推动了实证研究在中国教育研究领域的发展[2]。2017年1月，《加强教育实证研究，促进研究范式转型的华东师大行动宣言》发布，进一步推动了基于事实和证据进行教育研究的范式和文化[1]。迄今，教育实证研究论坛已成功举办多届，持续推进新时代高水平教育实证研究。

实证研究强调基于事实的证据收集，通过提出理论假设并进行实地调查，以科学的方式获取数据，最终形成可靠的结论。这一过程不仅要求严谨的数据分析和逻辑推理，还需要研究结果能够被重复验证，以确保其科学性和客观性[3]。此外，实证研究更是一种科学思维和学术态度。它要求研究者保持客观中立，减少主观臆断，并且在整个研究过程中严格遵守科学规范和程序。这种学术风范和规范程序对于培养严谨的学术态度和科学的思维方法具有不可替代的作用。强调教育实证研究的同时，也不能忽略我国教育研究领域多年来以思辨为主的现状，理清并平衡实证与思辨研究的关系，是促进教育研究发展的重要条件[4]。目前学界普遍认为实证研究包含量化、质性和混合研究[5]-[7]。无论是量化研究、质性研究，还是混合研究都需要对“数据”进行收集和分析，来获得对事物或关系确切的认识。实证研究通过对前人研究在不同情况下的复现，获得类似的研究结果，这种可以通过不断重复被验证的知识即为科学[8]。

全面推动教育实证研究，表明教育研究领域正在走向一条更为科学和系统的发展轨道。这一改变将

增强教育研究的质量及其对教育的影响力,有助于更好地制定和实施教育政策,推动教育公正和优质发展。实证教育研究的普及和应用将会成为推动中国教育改革深入进行的重要组成部分。

随着教育实证研究的发展,我国教育实证研究的发文量也呈逐年增长态势。实证研究发文量占比从2011年前的约20% [9],增长至2016年的24% [10],以及2019年的38.6% [11]。尽管如此,实证研究的真实占比可能低于先前的研究发现。作者通过对我国2012~2021年教育研究文献的摘要和部分原文的分析,发现这10年实证研究的实际发文量占比仅为15.7%,并且这些实证研究的质量参差不齐,评价标准也不一致 [4]。因此建立学界广泛认可且实践操作性强的实证研究评价标准是提升实证研究质量的重要前提。

布卢姆(1987)指出,“评价是一种获取和处理用以确定学术水平和教育有效性的证据的方法 [12]”。评价标准则是在正确的评价理念和原则指导下,根据目标、按照规定的程序,运用科学的方法,对实证研究的质量进行价值判断,同时通过评价过程与结果的反馈,促进研究更大的价值 [13]。汪晓赞认为评价标准是整个评价活动中最关键的部分,它是评价工作实际操作的标尺 [14]。因此,有效地进行评价从而实现评价目标,首先取决于评价标准是否科学合理。科学规范的实证研究评价标准可以确保研究的科学性和客观性,减少研究人员审稿专家的偏见和误差,从而保证研究结果的可靠程度。评价标准还能确保研究设计和执行的透明度,使得研究可以被独立地复制和验证。与此同时,将实证研究质量的影响因素进行整理并纳入评价标准,也能够更全面地对实证研究质量进行评判。

然而,现有的实证研究评价标准存在许多亟待解决的问题,胡中锋和禹薇指出,教育实证研究中出现的各种问题,表明现有评价标准的成效并不令人满意 [15]。具体而言,其一,现有的国际评价标准内容翔实、条目繁杂,侧重深度学术审查,无法满足期刊快速审稿、新手研究者自评、一线教师开展小型实证研究评价等实际需求。其二,标准缺乏本土化适配性与学界共识,部分国内评价标准直接照搬国际规范,未结合我国教育实证研究以思辨研究为传统、新手研究者占比高、中小规模研究为主的现状,同时不同机构、期刊采用的评价指标差异显著,导致评价结果缺乏可比性和公信力。其三,评价标准的动态优化机制缺失,现有标准大多为静态设定,较少通过专家共识论证、实际试评验证等方式进行迭代完善,也未关注指标的权重分配,难以精准反映教育实证研究的核心质量要求。正是基于上述问题,本研究以国际标准为基础,结合我国教育实证研究的实际需求,旨在构建一套指标科学、操作简便、适配性强的教育实证研究评价量表,通过专家咨询和实际试评对量表进行优化,为提升我国教育实证研究质量提供实用工具和参考依据。

2. 教育实证研究评价标准综述

教育实证研究论文的评价标准来源包括美国教育研究协会的实证社会科学研究报告标准 [16]、EASE英文科研究论文写作和翻译指南 [17]、美国心理学协会心理学研究报告标准 [18]、定性和定量研究设计 [19]、如何撰写和发表科学论文 [20]、教育科学研究 [21]、教育研究的设计和评估 [22]、国际期刊论文评价标准 [23]、评价实验教育研究报告的工具 [24]、定性研究检查表 [25]、评估心理学研究报告的标准 [26] 等。

通过对上述国际标准与研究成果的系统梳理和深度分析,提炼出实证研究论文质量评价的9项核心组成部分,构成评价的基础框架:1) 论文标题与摘要;2) 引言;3) 研究问题;4) 文献综述;5) 研究设计与方法(包括样本、工具、数据分析);6) 研究结果;7) 结果讨论;8) 理论与实践意义;9) 参考文献。同时,各标准还提出了多项特色评价要求,为评价维度的完善提供补充:Nunn与Adamson [23]还提到了论文的整体框架、语言风格与图标呈现;AERA(2006)强调了报告中的学术道德与伦理 [16];EASE(2018)提及关键词列表、缩略语列表和致谢 [17];Shavelson和Towne讨论了实证研究中的推理链 [21];Fraenkel等人提出应有术语的定义和对未来研究的建议 [22]。

针对实证研究包含的量化、质性两大类型，国际标准也提出了差异化评价要求：量化研究中的调查与实验形式，除遵循通用标准外，Creswell 等人指出实验研究需重点考虑结果影响因素，因此实验环境的详细描述成为关键评价点[19]；Schostak 提出将论点和反思纳入定性研究检查表[25]；Creswell 等人还讨论了定性研究的其他重要组成部分，例如研究者的角色对研究解释的影响[19]。上述标准来源对每一部分内容的书写规范与标准均给出了详细的要求与解释，但由于内容翔实，通常不会用于实证研究论文的快速审核与判断。为此，本研究以这些国际标准为核心依据，通过筛选核心评价维度、整合重叠要求、细化差异化标准、适配教育研究特征的路径，完成从国际标准到初步量表指标的映射与转化，依托这一转化过程构建高效、简易、操作性强的教育实证研究评价工具，并进一步验证该工具的有效性和可靠性。

3. 评价标准与评价工具

研究基于美国教育研究协会的实证社会科学研究报告标准(AERA, 2006) [16]、美国心理学协会心理学研究报告标准(APA, 2019) [18]等上述实证研究论文的规范与标准初步设计了教育实证研究论文的评价标准，该评价标准分成了 4 个大部分，分别是引言和文献综述、研究方法、研究结果和研究结果讨论。在这四个部分之下包含了研究问题/目的陈述、概念/理论框架、文献覆盖范围、文献总结/合成、研究方法与设计、研究样本/参与者、研究工具/测量、数据分析、研究结果/发现、结果讨论、研究局限、研究意义这 12 个一级指标，每个一级指标下面又包含着不同层级的二级指标，共 30 个二级指标(见表 1)。评分者可以在阅读文献的时候使用此表，通过阅读文献的具体内容，结合指标去分析文献中的内容是否达到了每一个指标要求呈现的程度，程度评价共有 3 个层次，分别是未呈现(0 分)，呈现弱(1 分)，呈现强(2 分)，通过评分，评分者能够明确一篇文献的质量，同样对这篇文章的行文结构有着更深层次的理解。

4. 研究方法

研究采用德尔菲法收集教育实证研究领域的专家意见，对评价工具进行修订；并收集 5 位研究新手使用评价工具对实证研究论文进行评价的结果和使用者反馈来验证工具的效果和一致程度。德尔菲法是一种利用专家的专长和经验进行直观预测的研究方法，通过一系列集中的专家调查问卷，并辅以有控制的观点反馈，从而得到一组专家最大程度的共识的过程[27]。专家人数的多少，可根据研究项目的大小和涉及面的宽窄而定，一般在 8~20 人左右为宜[28]。本研究选择了 9 位教育实证研究领域的专家进行咨询，每位专家都发表过多篇教育实证研究的论文。在咨询过程中，专家对评价标准中的一二级指标的重要程度进行评分，并在指标后填写评审意见。

验证过程中，两名本科生两名研究生和一位一线教师共同参与了论文的评价。考虑到实证研究的不同类型，研究选取了一篇量化研究和一篇质性研究文章作为分析与评价对象，评分者通过阅读所选文章，使用研究中开发的评价工具对文章进行评分。

5. 德尔菲专家咨询结果

本研究采用德尔菲法，选取 9 位教育实证研究领域发表过多篇相关论文的专家，对初步设计的 12 个一级指标、30 个二级指标的重要程度进行评分(评分越高代表重要性越强)，并收集专家对指标设计的修改意见，旨在通过专家共识优化评价量表指标体系。

5.1. 指标重要性评分结果

一级指标重要性排序：重要性程度最高的为研究结果和发现(得分 1.50)，其次为结果讨论(1.40)和数据分析(1.32)；重要性程度最低的为研究局限(0.72)，其次是文献覆盖范围(0.84)和研究样本(0.93)。

二级指标重要性排序：重要性程度最高的为研究结果具备适当数量的证据(质性/量化)的支持(1.36)，其次为通过对研究结果的讨论，明确回答了研究问题(1.34)和比较文献综述与研究发现，明确一致的与不一致的研究发现(1.28)；重要性程度最低的为清楚陈述了(本研究)的局限与不足(0.72)，其次为论证了研究样本或者参与者的选择方式与过程的合理性(0.73)，清楚陈述了将来研究的方向和文献综述中引述的文献在时间范围上是以近 5~10 年为主并列第三(均为 0.79)。

5.2. 专家反馈修改意见与指标修订方式

指标整体评分特征：30 个二级指标中，得分均数在 1 分及以上的共 13 个，占比 43.3%，主要集中在研究结果和发现“结果讨论”模块；得分均数在 1 分以下的共 17 个，占比 56.7%，主要集中在研究局限“研究样本”“文献覆盖范围”模块。

专家反馈修改意见包括：1) “文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容”这一二级指标，建议 a) 增加文献综述评述的文献包括(本研究)概念最初提出的背景；b) 把文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容这一指标和文献综述中引述的文献与研究问题是相关的这一指标合并为一个指标；c) 在文献覆盖范围上减少一个指标。2) “文献综述中引述的文献在时间范围上是以近 5~10 年为主”这一二级指标，专家建议：不一定是 5~10 年，涵盖研究领域最重要的文献和最新进展。3) “提供了满足研究目的的研究样本或参与者的描述性统计信息”这一二级指标，专家建议：把提供了满足研究目的的研究样本或参与者的描述性统计信息这一二级指标和论证了研究样本数量对于研究问题的合理性这一二级指标进行合并。4) “论证了研究样本数量对于研究问题的合理性”这一二级指标，专家的建议：很多研究都不报告，也不计算样本量。5) “论证了研究样本或者参与者的选择方式与过程的合理性”这一二级指标，专家的建议如下：这一指标与论证了研究样本数量对于研究问题的合理性有重叠部分。6) “论证了研究工具或方法的合理性”这一二级指标，专家的建议如下：这一指标与解释了研究工具或方法的可靠性(信度)与有效性(效度)有重叠部分。7) “清楚描述了数据分析方法/技术与研究问题的联系与适配性”这一二级指标，专家的建议如下：这一指标与清楚描述了数据分析方法/技术与研究设计的一致性有重叠部分。

修改方式如下：1) 将“文献综述中引述的文献与研究问题是相关的”和“文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容”这两个二级指标进行合并，调整为“文献综述可以覆盖全部研究问题”；2) 将“文献综述中引述的文献在时间范围上是以近 5~10 年为主”这一指标调整为“文献综述中引述的文献能够涵盖研究领域最重要的文献和最新进展”；3) 删去“论证了研究样本数量对于研究问题的合理性”这一指标；4) 将“论证了研究工具或方法的合理性”和“解释了研究工具或方法的可靠性(信度)与有效性(效度)”这一指标进行合并，调整为：“论证了研究工具的合理性(信效度)”。5) 将“清楚描述了数据分析方法/技术与研究问题的联系与适配性”这一二级指标与“清楚描述了数据分析方法/技术与研究设计的一致性”这一二级指标进行合并，改为“清楚描述了数据分析方法与研究问题的一致性与适配性”。

6. 量表试评结果

6.1. 质性研究论文试评结果

在使用质性研究方法的实证研究例子中，评分者初次评分的存在一致性指标有 18 个，占比 60%，评分存在差异的共有 12 个指标，占比 40%。其中评分标准差处在 0.5~1 之间的指标有 6 个，分别是研究样本的指标 1 和指标 2，即提供了满足研究目的的研究样本或参与者的描述性统计信息，和论证了研究样本数量对于研究问题的合理性；以及数据分析中的四个指标，分别是 1) 清楚描述了数据分析方法/技术

与流程；2) 清楚描述了数据分析方法/技术与研究问题的联系与适配性；3) 清楚描述了数据分析方法/技术与研究设计的一致性；4) 讨论了数据分析的可靠性、有效性和潜在的影响因素。评分标准差处在 0~0.5 之间的指标有 6 个，包括：概念/理论框架中的清楚描述了(本研究)的概念界定或理论框架；文献覆盖范围中的文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容，和文献综述中引述的文献在时间范围上是以近 5~10 年为主；研究工具和测量中的描述了研究中收集数据所用到的工具或方法和解释了研究工具或方法的可靠性(信度)与有效性(效度)，以及研究意义中的清楚陈述了将来研究的方向。经过短暂的讨论后，评分差异较小的指标能够迅速达到一致，因此评分者数量在三个或三个以上，且评分差异不大时，可以根据平均得分来确定该指标的最终得分。得分差异较大的指标为样本描述与数据分析描述，结合德尔菲法的结果，研究样本中的论证了研究样本数量对于研究问题的合理性这一指标权重得分为 0.84；数据分析中的清楚描述了数据分析方法/技术与研究问题的联系与适配性与清楚描述了数据分析方法/技术与研究设计的一致性这一指标的权重得分分别为 1.01 和 0.89，均为专家评价重要性较低的标准，专家评价与实际应用结果的一致性较强，因此遵循前面的建议对标准进行修订。

6.2. 量化研究论文试评结果

在使用量化研究方法的实证研究例子中，评分者初次评分存在一致性指标有 13 个，占比 43.3%，评分存在差异的共有 17 个指标，占比 56.7%。其中评分标准差处在 0.5~1 之间的有 12 个指标：包括研究问题/目的陈述中的提出了清晰明确的研究问题/目的；文献覆盖范围中的文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容；文献总结/合成中对(本研究)相关领域的主要研究观点、理论、方法和发现进行了总结，和对(本研究)相关领域已经做了什么研究和需要做什么研究进行了区分以及将 1~3 合成的文献综述与(本研究)要探究的研究问题联系起来；研究样本/参与者中的论证了研究样本或者参与者的选择方式与过程的合理性；数据分析中的清楚描述了数据分析方法/技术与研究问题的联系与适配性，和清楚描述了数据分析方法/技术与研究设计的一致性，以及讨论了数据分析的可靠性、有效性和潜在的影响因素；结果讨论中的比较文献综述与研究发现，明确一致的与不一致的研究发现和联系文献综述或新的文献，对(一致/不一致)的研究发现进行解释；以及研究意义中的清楚陈述了将来研究的方向。其中评分标准差处在 0~0.5 之间的有 5 个指标，包括：研究方法与设计中的清楚描述了研究方式(质性/量化)、策略和设计(实验……)和清晰描述了研究方法与研究问题之间的联系与适配性；研究工具和测量中的论证了研究工具或方法的合理性；数据分析中的清楚描述了数据分析方法/技术与流程；以及研究局限中的清楚陈述了(本研究)的局限与不足。该结果与前面质性研究的结果具有一致性，也反映了专家评价与实际应用结果的一致性较强，因此遵循前面的建议对标准进行修订。

7. 研究结论与意义

7.1. 研究核心结论

通过德尔菲法，我们邀请了 9 位教育实证研究领域的专家对教育实证研究的评价标准的 12 个一级指标和 30 个二级指标的重要程度进行评分。结果显示，在一级指标中，研究结果和发现、结果讨论和数据分析指标最为重要。在以往的研究中也有研究者提出了研究结果和讨论等指标的重要性，例如，刘钦普认为研究结果是实验得出的数据，它是论文的核心，也是推理和论证的重要依据，是论文的生命之所在[29]。唐伟认为讨论是对实验结果的分析和论证，是一篇文章的灵魂，是对结果的必然性、偶然性、可靠性和重要性进行的全面、系统的论述，并上升到一定的理论高度[30]。同样的，这两个一级指标下所对应的二级指标在德尔菲权重中所对应的得分也是最高的。其中研究结果具备适当数量的证据(质性/量化)的支持这一二级指标，得分最高；其次是：通过对研究结果的讨论，明确回答了研究问题这一指标，和比

较文献综述与研究发现,明确一致的与不一致的研究发现这一指标。再次证明了研究结果和讨论在一篇实证研究文章中的重要地位。

从评分者使用评分标准对实证研究进行评分的结果可以发现,评分者对两篇研究的评分与德尔菲法得出的权重具有较高的一致性。仅在“样本描述”和“数据分析方法”这两个维度上,评分者之间具有较大评分差异,这与德尔菲法的权重占比结果相一致,因而对相关指标进行了进一步调整,从而提高评分表的适用性。最终教育实证研究文章的评价标准由12个一级指标,30个二级指标变为12个一级指标,26个二级指标。

7.2. 研究贡献、量表适用范围与研究局限性

1) 研究核心贡献

本研究以美国教育研究协会、EASE等国际权威实证研究标准为基础,结合德尔菲专家咨询与实际试评实践,完成了教育实证研究评价量表的构建与优化,核心贡献体现在理论与实践双维度。理论层面,明确了我国教育实证研究评价的核心维度与各指标的相对重要性,呈现了国际标准向本土化评价工具的转化路径,为后续教育实证研究评价标准的制定与完善提供了科学依据和可复制的研究思路;实践层面,形成了一套高效、简易的评价工具,能够有效解决现有国际标准内容繁杂、实操性弱,难以适配快速审核场景的问题。

2) 量表适用范围

本研究优化后的教育实证研究评价量表为通用性快速评价工具,其适用范围具有明确边界,核心适用于基础教育与高等教育领域的量化、质性教育实证研究论文,涵盖期刊论文、学位论文、一线教师小型研究报告等类型;适用场景以快速质量评价为主,包括期刊审稿人的初评、高校教师对学生实证研究作业/论文的评价、新手研究者的自我评估、教育实证研究培训中的质量评判等;评价主体为具备基本教育实证研究方法知识的研究者、审稿人、高校师生及一线教育工作者。本量表暂不适用于混合研究的深度评价,也无法满足极致细节化的学术审查、高端科研成果评审等专业场景。

3) 研究局限性

本研究虽完成了量表的构建与优化,但受研究设计、样本规模等因素限制,仍存在明显局限性,需在后续研究中进一步完善:其一,专家咨询样本量偏小且覆盖面有限,仅选取9位教育实证研究领域专家,专家的研究方向、地域分布相对单一,未涵盖职业教育、特殊教育等领域,所得意见的代表性与普适性有待提升;其二,量表试评样本具有单一性,仅选取1篇量化、1篇质性研究论文为评价对象,未覆盖不同质量等级、不同研究主题的实证成果,无法充分验证量表在各类实证研究中的适配性;其三,量表未设置具体评分权重,仅通过“未呈现、呈现弱、呈现强”进行等级评价,无法实现教育实证研究质量的量化评估,评价结果的精准性与可比性有待加强。

7.3. 反常发现的批判性讨论——专家对“研究局限”低评分的背后逻辑与学术生态反思

本研究在专家评分环节发现一处与国际实证研究规范相悖的反常结果:“研究局限”维度在所有一级指标中评分最低(0.72),其下二级指标“清楚陈述了(本研究)的局限与不足”也为全部二级指标中得分最低,而客观陈述研究局限是国际实证研究标准中保障研究科学性、可重复性的核心要求,这一结果背后折射出我国教育实证研究领域的深层发展特征与学术生态问题,值得深入批判性探讨。

从领域发展阶段来看,我国教育实证研究尚处于规范化建设的初级阶段,与国际成熟的实证研究体系相比,当前领域的核心需求是实现“研究基础规范化”——即确保研究问题清晰、研究方法与问题适配、研究结果证据充分、结果讨论回应研究问题,而“清楚陈述研究局限”属于研究质量提升后的高阶

反思性要求。专家对“研究局限”的低评分，本质是对我国教育实证研究发展阶段的客观回应：在多数研究尚未达到基础规范性要求的背景下，评价体系自然更聚焦于决定研究质量底线的核心维度，对高阶反思性维度暂时降低权重。

从学术生态与评价导向来看，这一结果更反映出我国教育研究领域长期存在的“重成果呈现、轻批判性反思”的固有问题。一方面，当前教育实证研究的评价导向存在偏差，期刊审稿、学位论文评审、科研成果认定等环节，多侧重研究的创新性、实用性与成果价值，对研究局限的陈述缺乏硬性要求，甚至部分研究者形成“陈述研究局限会降低研究认可度”的认知，导致研究局限的陈述成为论文的“形式化模块”，未得到真正重视；另一方面，我国教育研究长期以思辨研究为主导，学术文化中更强调观点的论证与表达，而非对研究过程的全面反思、自我审视与客观批判，这种文化惯性延伸至实证研究领域，使得研究者对研究局限的识别、分析与陈述能力普遍不足，也让专家在构建评价体系时，自然降低了该维度的重要性。

需要明确的是，对研究局限的客观识别与清晰陈述，是实证研究科学性、严谨性的重要体现，更是研究持续深化、领域不断发展的重要基础。专家对“研究局限”的低评分，并非意味着该维度不重要，而是我国教育实证研究发展阶段的阶段性特征。随着我国教育实证研究规范化水平的不断提升，未来的评价体系需逐步强化对“研究局限”维度的要求，推动研究者树立科学的反思意识，同时引导学术评价从“重成果呈现”向“重科学严谨、重批判性反思”转型，培育更健康、更成熟的教育实证研究学术生态。

基金项目

2023 年广东省教育科学规划课题(高等教育专项)“基于师生双向评价的高等教育课程教学质量评估工具开发与验证”(2023GXJK677)。

参考文献

- [1] 华东师范大学学报(教育科学版). 加强教育实证研究, 提高教育科研水平——“第二届全国教育实证研究专题论坛”及“全国教育实证研究联席会议”成果览要[J]. 华东师范大学学报(教育科学版), 2017, 35(3): 18-36.
- [2] 王春丽, 顾小清. 形成基于证据的教育研究文化——“全国首届教育实证研究论坛”综述[J]. 中国远程教育, 2015(12): 5-11.
- [3] 王艳玲. 教育研究的实证框架: 一个案例分析[J]. 全球教育展望, 2007(12): 3-6+93.
- [4] 陈雅川, 杨洋. 实证还是思辨——两种教育研究的路径探析[J]. 西北师大学报(社会科学版), 2024, 61(2): 82-90.
- [5] 姚计海. 教育实证研究方法的范式问题与反思[J]. 华东师范大学学报(教育科学版), 2017, 35(3): 64-71+169-170.
- [6] 尹弘飏. 教育实证研究的一般路径: 以教师情绪劳动研究为例[J]. 华东师范大学学报(教育科学版), 2017, 35(3): 47-56+168-169.
- [7] 王思遥. 教育实证研究的理论依据、争议与去向[J]. 大学教育科学, 2020(5): 12-17.
- [8] 袁振国. 实证研究是教育学走向科学的必要途径[J]. 华东师范大学学报(教育科学版), 2017, 35(3): 4-17+168.
- [9] 姚计海, 王喜雪. 近十年来我国教育研究方法的分析与反思[J]. 教育研究, 2013, 34(3): 20-24+73.
- [10] 王兴宇. 关于高等教育实证研究之研究: 基于期刊的分析[J]. 黑龙江高教研究, 2017(12): 5-9.
- [11] 朱军文, 马银琦. 教育实证研究这五年: 特征、趋势及展望[J]. 华东师范大学学报(教育科学版), 2020, 38(9): 16-35.
- [12] 布卢姆, 等. 教育评价[M]. 邱渊, 等, 译. 上海: 华东师范大学出版社, 1987.
- [13] 屈宏强. 学校体育均衡发展评价指标体系的构建与实证研究[D]: [博士学位论文]. 福州: 福建师范大学, 2012.
- [14] 汪晓赞. 我国中小学体育学习评价改革的研究[D]: [博士学位论文]. 上海: 华东师范大学, 2005.
- [15] 胡中锋, 禹薇. 教育实证研究之深度反思[J]. 华南师范大学学报(社会科学版), 2020(5): 138-149+191-192.

-
- [16] American Educational Research Association (2006) Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*, **35**, 33-40.
- [17] EASE (2018) EASE Guidelines for Authors and Translators of Scientific Articles 2018. *European Science Editing*, **44**, e1-e16. <https://doi.org/10.20316/ese.2018.44.e1>
- [18] American Psychological Association (2019) Publication Manual of the American Psychological Association. American Psychological Association.
- [19] Creswell, J.W. and Creswell, J.D. (2022) Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications.
- [20] Day, R.A. and Gastel, B. (2012) How to Write and Publish a Scientific Paper. Cambridge University Press.
- [21] Shavelson, R.J. and Towne, L. (2002) Scientific Research in Education. National Academy Press.
- [22] Fraenkel, J., Wallen, N. and Hyun, H. (2022) How to Design and Evaluate Research in Education. 10th Edition, McGraw-Hill Education.
- [23] Nunn, R. and Adamson, J. (2007) Towards the Development of Alternative Criteria for International Journal Paper Evaluation. *The Asian EFL Journal*, **9**, 206.
- [24] Suydam, M.N. (1968) An Instrument for Evaluating Experimental Educational Research Report. *The Journal of Educational Research*, **61**, 200-203. <https://doi.org/10.1080/00220671.1968.10883645>
- [25] Schostak, J. (2015) Research Assignment/Dissertation/Thesis Check-List. Open Course Ware, University of Palestine. <https://ocw.up.edu.ps>
- [26] Gottfried, G.M., Vosmik, J.R. and Johnson, K.E. (2008) A Rubric for Evaluating Apsychology Research Report. Office of Teaching Resources in Psychology, Society for the Teaching of Psychology.
- [27] Dalkey, N. and Helmer, O. (1963) An Experimental Application of the DELPHI Method to the Use of Experts. *Management Science*, **9**, 458-467. <https://doi.org/10.1287/mnsc.9.3.458>
- [28] 徐国祥. 统计预测和决策[M]. 上海: 上海财经大学出版社, 2005: 11.
- [29] 刘钦普. 论科技论文的写作技巧[J]. 中国校外教育, 2010(18): 76-77.
- [30] 唐伟. 科技论文中结果与讨论的写法分析[J]. 科技资讯, 2014, 12(23): 216-217.

附录

论文部分	质性研究文章评分					量化研究文章评分														
	一级指标	二级指标	二级指标非权重	评分者1	评分者2	评分者3	评分者4	评分者5	平均分	标准差	评分者1	评分者2	评分者3	评分者4	评分者5	平均分	标准差			
引言和文献综述	研究问题/目的陈述	1.提出了清晰明确的研究问题/目的	1.00	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1.6	0.55	
	概念/理论框架	1.清楚描述了(本研究)的概念界定或理论框架	0.96	2	1	2	2	2	2	1.8	0.45	2	2	2	2	2	2	2	2	0.00
	文献覆盖范围	1.文献综述中引述的文献与研究问题是相关的	1.03	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00
		2.文献综述中引述的文献在内容范围上可以覆盖全部研究问题的内容	0.84	2	2	2	2	1	2	1.8	0.45	1	2	1	2	2	2	2	1.6	0.55
		3.文献综述中引述的文献在时间范围上是以近5-10年为主	0.79	2	2	2	2	2	1	1.8	0.45	2	2	2	2	2	2	2	2	0.00
	文献总结/合成	1.对(本研究)相关领域的主要研究观点、理论、方法和发现进行了总结	1.12	2	2	2	2	2	2	2	0.00	1	2	1	0	1	1	1	1	0.71
		2.对(本研究)相关领域的研究的优缺点进行了评述	0.94	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00
		3.对(本研究)相关领域已经做了什么研究和需要做什么研究进行了区分	0.98	2	2	2	2	2	2	2	0.00	2	2	0	2	2	2	2	1.6	0.89
		4.将1-3合成的文献综述与(本研究)要探究的研究问题联系起来	1.24	2	2	2	2	2	2	2	2	0.00	2	2	2	1	1	1.6	0.55	
	研究方法与设计	1.清楚描述了研究方式(质性/量化)、策略和设计(实验.....)	1.04	2	2	2	2	2	2	2	0.00	2	2	2	2	1	1.8	0.45		
2.清晰描述了研究方法与研究问题之间的联系与适配性		0.92	2	2	2	2	2	2	2	0.00	1	1	2	1	1.2	0.45				
研究样本/参与者	1.提供了满足研究目的的研究样本或参与者的描述性统计信息	0.87	1	2	0	2	2	2	1.4	0.89	2	2	2	2	2	2	2	2	0.00	
	2.论证了研究样本数量对于研究问题的合理性	0.84	2	1	2	0	2	2	1.4	0.89	2	2	2	2	2	2	2	2	0.00	
	3.描述了研究样本或者参与者的选择方式与过程(抽样方法)	0.84	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00	
	4.论证了研究样本或者参与者的选择方式与过程的合理性	0.73	2	2	2	2	2	2	2	0.00	2	2	0	2	2	1.6	0.89			
研究方法	1.描述了研究中收集数据所用到的工具或方法	1.08	1	2	2	2	2	2	1.8	0.45	2	2	2	2	2	2	2	2	0.00	
	2.论证了研究工具或方法的合理性	0.82	2	2	2	2	2	2	2	0.00	2	2	2	1	2	1.8	0.45			
	3.解释了研究工具或方法的可靠性(信度)与有效性(效度)	0.94	2	1	2	2	2	2	1.8	0.45	2	2	2	2	2	2	2	0.00		
数据分析	1.清楚描述了数据分析技术与流程	0.99	2	2	2	0	2	2	1.6	0.89	2	1	2	2	2	1.8	0.45			
	2.清楚描述了数据分析技术与研究问题的联系与适配性	1.01	2	2	2	0	2	2	1.6	0.89	1	2	0	1	2	1.2	0.84			
	3.清楚描述了数据分析技术与研究设计的一致性	0.89	2	2	2	0	2	2	1.6	0.89	2	0	2	1	1.2	0.84				
	4.讨论了数据分析的可靠性、有效性和潜在的影响因素	0.87	1	2	0	0	0	0	0.6	0.89	0	0	1	0	1	0.4	0.55			
研究结果	1.客观呈现与每个研究问题相关的研究结果	1.26	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00	
	2.研究结果具备适当数量的证据(质性/量化的)支持	1.36	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00	
研究结果讨论	1.比较文献综述与研究发现,明确一致的与不一致的研究发现	1.28	2	2	2	2	2	2	2	0.00	2	2	1	1	2	1.6	0.55			
	2.联系文献综述或新的文献,对(一致/不一致)的研究发现进行解释	1.24	2	2	2	2	2	2	2	0.00	2	2	0	1	2	1.4	0.89			
	3.通过对研究结果的讨论,明确回答了研究问题	1.34	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	0.00		
局限	1.清楚陈述了(本研究)的局限与不足	0.72	2	2	2	2	2	2	2	0.00	0	0	1	0	0	0.2	0.45			
意义	1.清楚陈述了(本研究)对理论和实践的意义	1.00	2	2	2	2	2	2	2	0.00	2	2	2	2	2	2	2	2	0.00	
	2.清楚陈述了将来研究的方向	0.79	0	1	0	0	0	0	0.2	0.45	0	0	0	0	0	0.4	0.89			
总分			55	56	54	47	55	53.4		49	49	44	46	52	48					