Educational Measurement
https://jem.atu.ac.ir/

ALLAMEH TABATABA'I UNIVERSITY PRESS

17(65), 2026, 105-128
eISSN: 2470-6240; ISSN: 2252-004x
Article Type: Research Article
Received June 02, 2025..... Received in revised form September 13, 2025
Accepted November 16, 2025 ..Published Online September 28, 2026

Alignment of Educational Sciences Theses with Kirkpatrick's Model Standards: A Meta-Evaluation Study

Ali Mahdavikhah [1], and **Morteza Karami** [2]✉

1. Department of Educational Sciences, Faculty of Psychology and Education, Ferdowsi University of Mashhad, Mashhad, Iran.. E-mail: alimahdavikhah@gmail.com
2. *Corresponding Author*, Department of Educational Sciences, Faculty of Psychology and Education, Ferdowsi University of Mashhad, Mashhad, Iran. E-mail: m.karami@um.ac.ir

ABSTRACT

Educational evaluation is pivotal in enhancing the quality of educational programs. This study aimed to analyze the alignment of master's theses in educational sciences with the standards of Kirkpatrick's four-level model (reaction, learning, behavior, results). The statistical population comprised all theses from four top Iranian universities (Tehran, Allameh Tabataba'i, Shahid Beheshti, and Ferdowsi Mashhad) between 2005 and 2025, from which 37 theses were identified and 26 were selected through census sampling based on inclusion criteria. A researcher-made checklist with 21 indicators across the four levels was used, with validity confirmed by three experts (CVR = 0.78) and reliability established via Cohen's kappa (κ = 0.92). Findings showed strong alignment at the reaction level (95.4%) , declining to 54.6% for learning, 34.8% for behavior, and 13.2% for results. Key weaknesses included limited use of advanced assessment methods (15%) , insufficient long-term follow-up (8%) , and neglect of organizational indicators (77%). These results underscore the urgent need for revising evaluation research methodologies, developing localized tools for higher-level assessments, and prioritizing longitudinal evaluations to enhance educational evaluation research and program effectiveness.

Keywords: educational evaluation, compliance analysis, meta-evaluation, Kirkpatrick model

Cite this Article: Mahdavikhah, A., & Karami, M. (2026). Alignment of Educational Sciences Theses with Kirkpatrick's Model Standards: A Meta-Evaluation Study. *Educational Measurement*, *17*(65), 105-128. https://doi.org/10.22054/jem.2025.86666.3632

© **2026** by The author(s) retain the copyright
Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2025.86666.3632

EXTENDED ABSTRACT

Introduction

Educational evaluation is a cornerstone for improving instructional quality, with Kirkpatrick's four-level model (reaction, learning, behavior, results) recognized as the global gold standard (Kirkpatrick & Kirkpatrick, 2016). This meta-evaluation examines 26 master's theses in educational sciences from Iran's top universities (2005–2025) to assess their adherence to Kirkpatrick's standards. Globally, 65% of organizations employ this framework (ATD, 2023), yet studies reveal implementation gaps, particularly in developing contexts where Level 4 (results) evaluations are addressed in only 18% of cases (Alsalamah & Callinan, 2022). To address these gaps, this study uses a 21-indicator checklist, validated through expert review (CVR=0.78) and inter-rater reliability testing (κ=0.92). The checklist is grounded in Kirkpatrick's levels and aligned with the Joint Committee on Standards for Educational Evaluation (JCSEE) criteria—utility, accuracy, feasibility, and propriety—to support the meta-evaluation (Yarbrough et al., 2011). For instance, reaction aligns with utility (usefulness for stakeholders), learning and behavior with accuracy (data validity) , and results with feasibility and propriety (ethical and practical implementation). Findings reveal declining alignment across Kirkpatrick's levels, with alignment dropping from 95.4% (Level 1) to 13.2% (Level 4), highlighting the need for curricular reform and longitudinal assessment protocols in graduate research.

Homepage: https://jem.atu.ac.ir/

Literature Review

The Kirkpatrick model remains a cornerstone for evaluating educational program effectiveness, encompassing reaction, learning, behavior, and results (Kirkpatrick Partners, 2023). Recent adaptations include a fifth level (ROI) to enhance its scope (Aguinis & Kraiger, 2009). Meta-analyses show moderate inter-level correlations (e.g., learning-behavior: r=0.42; Alliger et al., 1997. The present study identified three problems: (1) limited use of diverse assessment methods (72% of cases), (2) follow-up periods shorter than six months (92% of studies), and (3) organizational resistance to data sharing (78% of cases). In the present study, reaction assessments achieved 100% compliance in questionnaire design, whereas assessments at higher levels showed weaker alignment with evaluation standards. Studies like Sharafinezhad and Sahraei (2024) advocate for mixed-methods approaches, while Kim and Patel (2021) emphasize stakeholder-engaged designs to

address these gaps. The integration of JCSEE standards (Yarbrough et al., 2011) with Kirkpatrick's levels, as applied in this study, supports a broader meta-evaluation framework, aligning with global benchmarks (Alsalamah & Callinan, 2022; Nawaz et al., 2022).

Methodology

This study employed quantitative content analysis to evaluate 26 theses from 37 identified works through census sampling. A researcher-developed checklist operationalized Kirkpatrick's levels into 21 indicators: reaction (7 items: questionnaire design, ethical compliance), learning (5 items: assessment diversity, control groups), behavior (5 items: longitudinal tracking, multiple sources of assessment), and results (4 items: KPI alignment, ROI calculation). Each indicator was scored on a 5-point Likert scale (1 = weak to 5 = excellent), converted to percentage-based tiers (excellent: 90–100%; weak: <30%). The checklist was reviewed by three experts (CVR = 0.78), and a second coder examined 20% of the sample (87% agreement). Data analysis used descriptive statistics, including frequencies, percentages, means, standard deviations, and coefficients of variation.

Conclusion

Three key findings emerged: First, a steep quality gradient exists across levels—reaction assessments scored 4.6/5 (95.4% alignment) versus 1.2/5 (13.2%) for results, indicating a substantial decline across levels. Second, methodological limitations were widespread: 85% of studies used only quantitative methods, and 89% did not address contextual or confounding variables. Practical recommendations include: (1) integrating advanced evaluation designs into curricula, (2) developing localized behavioral metrics, and (3) encouraging longer follow-up periods in impact studies. The 21-indicator checklist, aligned with JCSEE standards, provides a framework for evaluating the quality of educational evaluation research. Future research could examine other disciplines and institutions and undertake longitudinal comparisons.

Homepage://jem.atu.ac.ir/

Ethical Considerations

Adherence to Research Ethics Principles

The authors have observed all ethical principles in conducting and publishing this scientific research, and this is confirmed by all of them.

Author Contributions

Both authors contributed to all sections of the article.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We would like to express our gratitude for the cooperation of the respected librarians of the University of Tehran, Allameh Tabataba'i University, Shahid Beheshti University, and Ferdowsi University of Mashhad in providing access to the full texts of the dissertations.

Homepage: https://jem.atu.ac.ir/

اندازه‌گیری تربیتی
https://jem.atu.ac.ir/

انتشارات دانشگاه علامه طباطبائی

دوره ۱۷، شماره ۶۵، ۱۴۰۵، ۳۷-۶۲
eISSN: 2470-6240; ISSN: 2252-004x

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۴۰۴/۰۴/۱۲؛ تاریخ بازنگری: ۱۴۰۴/۰۶/۲۲
تاریخ پذیرش: ۱۴۰۴/۰۸/۲۵؛ تاریخ انتشار: ۱۴۰۵/۰۷/۰۶

انطباق پایان‌نامه‌های علوم تربیتی با استانداردهای مدل کرک‌پاتریک: مطالعه فراارزیابی

علی مهدوی خواه [۱]، مرتضی کرمی [۲]✉

۱. گروه مطالعات برنامه درسی و آموزش، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه فردوسی مشهد، مشهد، ایران.
رایانامه: alimahdavikhah@gmail.com

۲. *نویسنده مسئول* گروه مطالعات برنامه درسی و آموزش، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه فردوسی مشهد، مشهد، ایران. رایانامه: m.karami@um.ac.ir

چکیده

ارزشیابی آموزشی نقش محوری در بهبود کیفیت برنامه‌های تربیتی دارد. این پژوهش با هدف تحلیل میزان انطباق پایان‌نامه‌های کارشناسی ارشد علوم تربیتی با استانداردهای مدل چهارسطحی کرک‌پاتریک انجام شد. جامعه آماری شامل کلیه پایان‌نامه‌های چهار دانشگاه برتر کشور (تهران، علامه طباطبایی، شهید بهشتی و فردوسی مشهد) طی سال‌های ۱۳۸۴ تا ۱۴۰۴ بود که از میان ۳۷ پایان‌نامه شناسایی‌شده، ۲۶ پایان‌نامه به روش سرشماری و بر اساس معیارهای ورود انتخاب شدند. ابزار پژوهش چک‌لیست محقق‌ساخته‌ای متشکل از ۲۱ نشانگر در چهار سطح واکنش، یادگیری، رفتار و نتایج بود که روایی آن توسط سه متخصص تأیید و پایایی آن با ضریب کاپای کوهن ۰,۹۲ محاسبه شد. یافته‌ها نشان داد سطح واکنش با میانگین انطباق ۹۵,۴ درصد بهترین عملکرد را داشت، درحالی‌که این میزان در سطح یادگیری به ۵۴,۶ درصد، در سطح رفتار به ۳۴,۸ درصد و در سطح نتایج به ۱۳,۲ درصد کاهش یافت. مهم‌ترین نقاط ضعف شامل استفاده ناکافی از روش‌های پیشرفته سنجش (۱۵ درصد)، عدم پیگیری بلندمدت (۸ درصد) و غفلت از سنجش شاخص‌های سازمانی (۷۷ درصد) بود. نتایج نشان‌دهنده نیاز مبرم به بازنگری در روش‌شناسی پژوهش‌های ارزشیابی، توسعه ابزارهای بومی سنجش سطوح بالاتر و توجه به ارزشیابی‌های طولی‌مدت است. این یافته‌ها برای بهبود کیفیت پژوهش‌های ارزشیابی آموزشی و اثربخشی برنامه‌های تربیتی حائز اهمیت است.

کلیدواژه‌ها: ارزشیابی آموزشی، تحلیل انطباق، فراارزشیابی، مدل کرک‌پاتریک

استناد به این مقاله: مهدوی خواه، علی، و کرمی، مرتضی. (۱۴۰۵). انطباق پایان‌نامه‌های علوم تربیتی با استانداردهای مدل کرک‌پاتریک: مطالعه فراارزیابی. *اندازه‌گیری تربیتی*، *۱۷*(۶۵)، ۱-۲۴.
https://doi.org/10.22054/jem.2025.86666.3632

© ۲۰۲۶ نویسنده(گان)
ناشر: دانشگاه علامه طباطبائی

مقدمه

ارزشیابی آموزشی به‌عنوان یکی از ارکان اساسی نظام‌های یاددهی-یادگیری، در چند دهه اخیر تحولات چشمگیری را تجربه کرده است (Hattie, 2020; Patton, 2017). مطالعات نشان می دهد که نظام‌های آموزشی پیشرفته، سالانه حدود ۱۵-۲۰ درصد از بودجه خود را به ارزشیابی و سنجش اثربخشی برنامه‌های آموزشی اختصاص می‌دهند (UNESCO, 2022). این توجه ویژه ناشی از نقش تعیین‌کننده ارزشیابی در بهبود کیفیت آموزش و تخصیص بهینه منابع است (Kellaghan & Stufflebeam, 2020). در این راستا، پژوهش‌های متعددی به بررسی الگوهای مختلف ارزشیابی پرداخته‌اند که از میان آن‌ها می‌توان به مدل سیپ (Stufflebeam & Zhang, 2017)، مدل فلیپس (Phillips, 2022) و مدل کرک‌پاتریک (Kirkpatrick & Kirkpatrick, 2016) اشاره کرد. مدل چهارسطحی کرک‌پاتریک که نخستین بار در سال ۱۹۵۹ توسط دونالد کرک‌پاتریک معرفی شد، به‌عنوان پرکاربردترین چارچوب ارزشیابی آموزشی در جهان شناخته می شود (Garavan et al., 2019). بر اساس آمارهای منتشرشده توسط انجمن توسعه استعدادها[1] (2023) حدود ۶۵ درصد از سازمان‌های پیشرو در سطح جهانی از این مدل برای ارزشیابی برنامه‌های آموزشی خود استفاده می‌کنند. این مدل در سال‌های اخیر با افزودن سطح پنجم (بازگشت سرمایه) تکامل یافته و به‌عنوان چارچوبی جامع در ارزشیابی اثربخشی آموزشی مورداستفاده قرار می‌گیرد (Aguinis & Kraiger, 2021).

مطالعات متعددی به بررسی کاربرد این مدل در حوزه‌های مختلف پرداخته‌اند. در آموزش پزشکی، تحقیقات Yardley and Dornan (2011) نشان داده‌اند که این مدل در ۷۲ درصد از برنامه‌های آموزش مداوم پزشکی مورداستفاده قرار می‌گیرد. در حوزه آموزش مدیریت نیز پژوهش‌های Desimone and Pak (2016) حاکی از آن است که مدل کرک‌پاتریک در ۶۸ درصد از مطالعات ارزشیابی به کار رفته است. بااین‌حال، مطالعات Alsalamah and Callinan (2022) نشان می‌دهند که تنها ۱۸ درصد از پژوهش‌ها به سطح چهارم مدل (نتایج سازمانی) پرداخته‌اند. در زمینه فراارزشیابی پژوهش‌های آموزشی، یافته‌های Forss و همکاران (2011)

1. Association for Talent Development (ATD)

Homepage: https://jem.atu.ac.ir/

حاکی از آن است که کیفیت ارزشیابی‌های آموزشی در محیط‌های دانشگاهی نیازمند بازنگری اساسی است. در ایران، پژوهش‌های متعددی به بررسی کیفیت ارزشیابی‌های آموزشی پرداخته‌اند. مسئله اصلی این پژوهش، مسئله کیفیت پژوهش‌های مبتنی بر مدل کرک‌پاتریک است. با در نظر گرفتن این نکته که کیفیت هر ارزشیابی در گروه رعایت استانداردهای آن است، با وجود کاربرد گسترده این مدل، مطالعات معدودی به بررسی نظام‌مند کیفیت اجرای آن پرداخته‌اند (Okebukola & Uvalić-Trumbić 2023). این در حالی است که پژوهش‌های Mertens and Wilson (2022) نشان می‌دهند که کیفیت ارزشیابی‌های آموزشی تأثیر مستقیمی بر اثربخشی برنامه‌های آموزشی دارد. مطالعه حاضر با هدف تحلیل میزان انطباق پایان‌نامه‌های کارشناسی ارشد دانشگاه‌های برتر ایران رشته‌های علوم تربیتی که به ارزشیابی آموزشی مبتنی بر مدل کرک‌پاتریک پرداخته‌اند، با استانداردهای این مدل طراحی شده است. این پژوهش از سه منظر حائز اهمیت است: نخست از دیدگاه نظری، به‌عنوان اولین مطالعه نظام‌مند در زمینه فراارزشیابی پایان‌نامه‌های ایرانی بر اساس مدل کرک‌پاتریک؛ دوم از منظر روش‌شناسی، با ارائه چارچوبی بومی برای ارزیابی کیفیت پژوهش‌های ارزشیابی؛ و سوم از جنبه کاربردی، با ارائه راهکارهای عملی برای بهبود استانداردهای ارزشیابی آموزشی.

مدل چهارسطحی کرک‌پاتریک به‌عنوان چارچوبی جامع برای ارزشیابی اثربخشی برنامه‌های آموزشی شناخته می‌شود که شامل سطوح واکنش، یادگیری، رفتار و نتایج است (Kirkpatrick Partners, 2023). با وجود کاربرد گسترده این مدل در سطح جهانی، پژوهش‌های معدودی به بررسی میزان انطباق پژوهش‌های بومی با استانداردهای این مدل پرداخته‌اند (Ahn, Ames, & Myers, 2012). نوری (۱۳۹۴) در مطالعه‌ای با عنوان «فراارزشیابی در آموزش‌های فنی و حرفه‌ای» به فراارزشیابی عنصر سنجش در آموزش‌های فنی و حرفه‌ای پرداخت. این پژوهش که با استفاده از چارچوب استانداردهای ارزشیابی انجام شد، نشان داد نظام مذکور در معیارهای اخلاقی بودن (میانگین = ۴/۶ از ۵) و عملی بودن (میانگین = ۴/۴) در سطح خیلی خوب، و در معیارهای سودمندی (میانگین = ۳/۸) و درستی فنی (میانگین = ۳/۷) در سطح خوب قرار دارد. تحلیل داده‌ها حاکی از آن بود که اگرچه نظام ارزشیابی به استانداردهای مطلوب نزدیک است، اما برای ارتقای کامل نیازمند بهبود در ابعاد سودمندی و دقت فنی است. قائمی پور و همکاران (۱۳۹۴) در پژوهشی با

Homepage: https://jem.atu.ac.ir/

عنوان «میزان اشاعه استانداردهای ارزشیابی در پایان‌نامه‌های کارشناسی ارشد» از چک‌لیست فراارزشیابی استافل بیم برای تحلیل ۳۷ پایان‌نامه استفاده کردند. یافته‌ها نشان داد معیارهای سودمندی (میانگین = ۱/۲ از ۵)، قابلیت کاربرد (میانگین = ۲/۳) و درستی (میانگین = ۲/۴) در سطح ضعیف ارزیابی شدند، درحالی‌که معیارهای دقت (میانگین = ۳/۲) و مسئولیت‌پذیری (میانگین = ۳/۱) در سطح متوسط قرار داشتند. این مطالعه نتیجه گرفت که کیفیت ارزشیابی‌های آموزشی در پایان‌نامه‌های تحصیلات تکمیلی نیازمند توجه جدی به تمامی ابعاد استانداردهای ارزشیابی است. سبزواری (۱۳۸۴) در مطالعه نظری خود با عنوان «فراارزشیابی و معیارهای آن» به تحلیل مفهومی ابعاد مختلف فراارزشیابی پرداخت. این پژوهش که با روش کتابخانه‌ای انجام شد، چارچوبی نظام‌مند برای ارزیابی کیفیت ارزشیابی‌های آموزشی ارائه داد. نتایج نشان داد که فراارزشیابی به‌عنوان ابزاری برای بهبود مستمر نظام‌های ارزشیابی، در تمامی حوزه‌های آموزشی ازجمله برنامه‌ها، مؤسسات و ارزیابی‌های فردی کاربرد دارد. مطالعه تأکید کرد که اجرای نظام‌مند فراارزشیابی می‌تواند به ارتقای کیفیت تصمیم‌گیری‌های آموزشی منجر شود.

Alsalamah and Callinan (2022) در مطالعه‌ای با عنوان «مدل کرک‌پاتریک برای ارزشیابی آموزش: تحلیل کتاب‌سنجی پس از ۶۰ سال» به بررسی نظام‌مند ۴۱۶ مطالعه پرداختند. این پژوهش نشان داد که مدل کرک‌پاتریک عمدتاً در آموزش پزشکی (۳۴ درصد) و علوم کامپیوتر (۱۸ درصد) استفاده شده است. یافته‌ها حاکی از آن بود که تنها ۱۸ درصد مطالعات به سطح نتایج سازمانی پرداخته‌اند، درحالی‌که ۶۳ درصد محدود به سطوح واکنش و یادگیری بودند. این مطالعه تأکید کرد که علی‌رغم ظهور مدل‌های جدید، چارچوب کرک‌پاتریک همچنان پرکاربردترین ابزار ارزشیابی آموزشی محسوب می‌شود. Nawaz و همکاران (2022) در فراتحلیلی با عنوان «مدل کرک‌پاتریک و اثربخشی آموزش: فراتحلیل ۱۹۸۲ تا ۲۰۲۱» به بررسی ۴۱ مطالعه پرداختند. نتایج این پژوهش نشان داد میانگین همبستگی بین سطوح یادگیری و رفتار ۰/۴۲ بوده است. مطالعه تأکید کرد که علی‌رغم چهار دهه تحقیق، پیشرفت معناداری در اثربخشی آموزش‌های مدیریتی مشاهده نشده است. پژوهشگران استفاده از روش‌های ارزشیابی ترکیبی را برای سنجش سطوح مختلف مدل پیشنهاد دادند. Sitzmann و همکاران (2020) در پژوهشی با عنوان «مقایسه آموزش مجازی و حضوری: فراتحلیل نتایج بر اساس مدل کرک‌پاتریک» ۹۶ برنامه آموزشی را مقایسه کردند. نتایج

Homepage: https://jem.atu.ac.ir/

نشان داد آموزش‌های مجازی در سطوح واکنش (d=۰/۳۵) و یادگیری (d=۰/۴۱) برتری دارند، درحالی‌که آموزش حضوری در سطوح رفتار (d=۰/۲۸) و نتایج (d=۰/۱۹) مؤثرتر عمل کرده است. این مطالعه بر نیاز به طراحی متفاوت روش‌های ارزشیابی برای محیط‌های آموزشی مختلف تأکید کرد. Powell and Yalcin (2010) در فراتحلیلی با عنوان «اثربخشی آموزش مدیریتی: فراتحلیل ۱۹۵۲-۲۰۰۲» ۸۵ مداخله آموزشی را بررسی کردند. نتایج نشان داد اندازه اثر کلی مدل ۰/۴۶ بوده است. سطح یادگیری با اندازه اثر ۰,۵۱ بیشترین تأثیر را نشان داد، درحالی‌که سطح نتایج با ۰/۲۲ کمترین اثر را داشت. این مطالعه نتیجه گرفت که برنامه‌های آموزشی مدیریت در طول ۵۰ سال پیشرفت محسوسی نداشته‌اند. Alliger و همکاران (1997) در مطالعه پایه‌ای با عنوان «فراتحلیل روابط بین معیارهای آموزش» به بررسی ۳۴ پژوهش تجربی پرداختند. یافته‌ها ماتریس همبستگی بین سطوح مختلف مدل را ارائه کرد: واکنش-یادگیری (۰/۳۸)، یادگیری-رفتار (۰/۴۲) و رفتار-نتایج (۰/۳۵). این پژوهش مبنای نظری مهمی برای روابط بین سطوح مختلف ارزشیابی آموزشی ایجاد کرد و نشان داد که پیش‌بینی نتایج سازمانی صرفاً بر اساس سطوح پایین‌تر با محدودیت همراه است.

بررسی مطالعات داخلی و خارجی نشان می‌دهد که اگرچه مدل کرک‌پاتریک به‌عنوان چارچوبی پرکاربرد در ارزشیابی آموزشی شناخته می‌شود، اما شکاف‌های قابل‌توجهی در نحوه اجرا و گزارش‌دهی سطوح مختلف آن وجود دارد. پژوهش‌های بین‌المللی عمدتاً بر محدودیت‌های مدل در سنجش سطوح بالاتر (رفتار و نتایج) و ضعف همبستگی بین سطوح تأکید داشته‌اند (Alsalamah & Callinan, 2022; Nawaz et al., 2022). در مقابل، مطالعات داخلی نیز حاکی از آن است که استانداردهای ارزشیابی در پژوهش‌های دانشگاهی ایران، به‌ویژه در پایان‌نامه‌های تحصیلات تکمیلی، اغلب در سطح ضعیف تا متوسط قرار دارند و به سطوح کیفی‌تر مدل (مانند مسئولیت‌پذیری و دقت) کمتر توجه شده است (نوری، ۱۳۹۴؛ قائمی پور و همکاران، ۱۳۹۴). با توجه به این چالش‌ها و عدم وجود مطالعه‌ای نظام‌مند که به فراارزشیابی پایان‌نامه‌های ایرانی بر اساس مدل کرک‌پاتریک پرداخته باشد، پژوهش حاضر با هدف تحلیل میزان انطباق این پژوهش‌ها با استانداردهای چهارسطحی مدل و شناسایی نقاط قوت و ضعف روش‌شناختی آن‌ها طراحی شده است. این مطالعه می‌تواند با ارائه چارچوبی بومی، به بهبود کیفیت ارزشیابی‌های آموزشی در پژوهش‌های آتی کمک کند.

Homepage: https://jem.atu.ac.ir/

روش

این پژوهش با هدف بررسی انطباق پایان‌نامه‌های کارشناسی ارشد علوم تربیتی با استانداردهای مدل کرک‌پاتریک، به روش تحلیل محتوای کمی انجام شد. جامعه آماری شامل کلیه پایان‌نامه‌های دانشگاه‌های تهران، علامه طباطبایی، شهید بهشتی و فردوسی مشهد در بازه زمانی ۱۳۸۴ تا ۱۴۰۴ بود که به ارزشیابی آموزشی با مدل کرک‌پاتریک پرداخته بودند. از میان ۳۷ پایان‌نامه شناسایی‌شده، ۲۶ پایان‌نامه بر اساس معیارهای ورود شامل: (۱) تمرکز بر ارزشیابی آموزشی، (۲) استفاده صریح از مدل کرک‌پاتریک، و (۳) دسترسی به متن کامل، به روش سرشماری انتخاب شدند.

ابزار پژوهش چک‌لیست محقق‌ساخته‌ای بود که بر اساس سطوح چهارگانه مدل کرک‌پاتریک طراحی شد. این چک‌لیست شامل ۲۱ نشانگر در چهار سطح بود: سطح واکنش (۷ نشانگر)، سطح یادگیری (۵ نشانگر)، سطح رفتار (۵ نشانگر) و سطح نتایج (۴ نشانگر). مبنای طراحی چک‌لیست، مدل چهارسطحی کرک‌پاتریک به‌عنوان چارچوب اصلی برای اجرایی کردن نشانگرها استوار است، که سطوح آن (واکنش، یادگیری، رفتار، و نتایج) استانداردهای طلایی برای ارزشیابی اثربخشی برنامه‌های آموزشی را فراهم می‌کند. این مدل بر اساس تحلیل کتاب‌سنجی (2022) Alsalamah and Callinanدر ۴۱۶ مطالعه، پرکاربردترین چارچوب (با استفاده در ٪۶۵ سازمان‌ها بر اساس (ATD, 2023) برای فراارزشیابی ارزیابی‌های آموزشی است و اجازه می‌دهد انطباق پایان‌نامه‌ها با هر سطح را به‌صورت نظام‌مند بسنجیم. علاوه بر این، برای تقویت جنبه فراارزشیابی، چک‌لیست با ملاک‌های استانداردهای کمیته مشترک ارزشیابی آموزشی (2011 ,.Yarbrough et al) همخوانی دارد. به‌طور خاص، سطح واکنش با ملاک سودمندی[1] (مفید بودن نتایج برای ذی‌نفعان) مرتبط است، سطوح یادگیری و رفتار با دقت[2] (اعتبار و پایایی داده‌ها)، و سطح نتایج با عملی بودن[3] (امکان‌پذیری اجرا) و تناسب[4] (رعایت اصول اخلاقی و عدالت). این همپوشانی بر اساس فراتحلیل Nawaz و همکاران (2023) و مطالعه داخلی قائمی پور و همکاران (۱۳۹۴) توجیه می‌شود، که ضعف‌های مشابه در پایان‌نامه‌های ایرانی را برجسته می‌کنند و نشان می‌دهند

1. utility
2. accuracy
3. feasibility
4. propriety

Homepage: https://jem.atu.ac.ir/

سطوح کرک‌پاتریک می‌تواند معیارهای «کمیته مشترک استانداردهای ارزشیابی آموزشی»[1] را بدون نیاز به ابزار جداگانه پوشش دهد. روایی ابزار از طریق نظر سه متخصص ارزشیابی آموزشی تأیید شد (CVR=۰/۷۸) و پایایی آن با ضریب کاپای کوهن ۰/۹۲ محاسبه گردید.

برای هر نشانگر، میزان انطباق با استانداردهای مدل کرک‌پاتریک بر اساس مقیاس پنج‌سطحی کیفی ارزیابی شد: عالی (۵)، خیلی خوب (۴)، خوب (۳)، متوسط (۲) و ضعیف (۱). نمره هر سطح از مدل از طریق جمع نمرات تمام نشانگرهای آن سطح محاسبه و سپس به درصد تبدیل شد. درصدهای به‌دست‌آمده در پنج سطح کیفی طبقه‌بندی شدند: عالی (۹۰–۱۰۰ درصد)، خیلی خوب (۷۵–۸۹ درصد)، خوب (۵۵–۷۴ درصد)، متوسط (۳۰–۵۴ درصد) و ضعیف (کمتر از ۳۰ درصد). این روش نمره‌دهی امکان مقایسه نظام‌مند کیفیت پایان‌نامه‌ها در هر سطح را فراهم کرد. جدول ۱ ویژگی‌های چک‌لیست را بر اساس سطوح کرک‌پاتریک و همخوانی با ملاک‌های JCSEE توصیف می‌کند، بدون تغییر در نشانگرها یا مقیاس‌ها، و مبنای آن بر اساس منابع ذکرشده (Yarbrough et al., 2011; Kirkpatrick & Kirkpatrick, 2016) استوار است.

پس از طراحی و اعتبارسنجی ابزار، مراحل اجرای پژوهش به ترتیب زیر انجام شد: ابتدا بررسی مقدماتی ۵ پایان‌نامه برای رفع ابهامات چک‌لیست، سپس اجرای نهایی بر روی ۲۶ پایان‌نامه و درنهایت کدگذاری داده‌ها. تحلیل داده‌ها به‌صورت توصیفی و با استفاده از شاخص‌های فراوانی، درصد و میانگین انجام شد. برای اطمینان از قابلیت اعتماد نتایج، ۲۰ درصد از نمونه‌ها (۵ پایان‌نامه) توسط کدگذار دوم بررسی شد که ضریب توافق ۰/۸۷ به دست آمد. این روش‌شناسی نظام‌مند، امکان ارزیابی دقیق میزان انطباق پایان‌نامه‌ها با استانداردهای مدل کرک‌پاتریک را فراهم آورد، درحالی‌که همخوانی با JCSEE (Yarbrough et al., 2011) فراارزشیابی را بدون نیاز به بازطراحی ابزار، جامع‌تر می‌سازد. مبنای این رویکرد ترکیبی، مطالعات Scriven (2009) در مورد فرا ارزشیابی است که تأکید بر استفاده از مدل‌های خاص (مانند کرک‌پاتریک) برای پوشش استانداردهای کلی دارد.

1 Joint Committee on Standards for Educational Evaluation (JCSEE)

Homepage: https://jem.atu.ac.ir/

جدول ۱. ویژگی‌های چک‌لیست فراارزشیابی مبتنی بر مدل کرک‌پاتریک با همخوانی ملاک‌ها (JCSEE)

سطح	معیارهای کلیدی	شاخص‌های ارزیابی	مقیاس نمره‌دهی	رتبه‌بندی کیفی	ملاک مرتبط با JCSEE
واکنش	رضایت شرکت‌کنندگان	-طراحی پرسشنامه استاندارد -پوشش تمام ابعاد آموزشی -نرخ پاسخ ≥۸۰%	۰-٤ نمره (هر شاخص ۱ نمره)	نمره عالی: ٤ نمره خیلی خوب: ۳ نمره خوب: ۲ نمره متوسط: ۱ نمره ضعیف: ۰ نمره	سودمندی - مفید بودن برای ذی‌نفعان (Yarbrough et al., 2011)
یادگیری	دانش/مهارت کسب‌شده	-آزمون‌های معتبر - مقایسه پیش‌آزمون/پس‌آزمون - گروه کنترل	۰-۱۲ نمره (هر شاخص ٤ نمره)	عالی: ۱۰-۱۲ خیلی خوب: ۷-۹ خوب: ٤-٦ متوسط: ۱-۳ ضعیف: ۰	دقت اعتبار و پایایی داده‌ها (Yarbrough et al., 2011)
رفتار	تغییر عملکرد	-ارزشیابی مکرر (≤۲ بار) -منابع چندگانه سنجش -دوره پیگیری ≥۳ ماه	۰-۱۲ نمره (هر شاخص ٤ نمره)	عالی: ۱۰-۱۲ خیلی خوب: ۷-۹ خوب: ٤-٦ متوسط: ۱-۳ ضعیف: ۰	دقت و عملی بودن امکان‌پذیری و کنترل متغیرها (Nawaz et al., 2023)
نتایج	تأثیر سازمانی	-سنجش ≥۳ شاخص کلیدی -محاسبه- ROI انطباق با اهداف استراتژیک	۰-۱۲ نمره (هر شاخص ٤ نمره)	عالی: ۱۰-۱۲ خیلی خوب: ۷-۹ خوب: ٤-٦ متوسط: ۱-۳ ضعیف: ۰	تناسب و عملی بودن رعایت اخلاقی و امکان‌پذیری (Ghaemipour et al., 2015)

یافته‌ها

در این بخش، یافته‌های توصیفی پژوهش به‌صورت نظام‌مند ارائه می‌گردد. تحلیل‌ها مبتنی بر بررسی میزان انطباق پایان‌نامه‌ها با استانداردهای مدل کرک‌پاتریک در چهار سطح است. داده‌ها با استفاده از شاخص‌های توصیفی شامل فراوانی، درصد، میانگین، و انحراف معیار تحلیل شده‌اند. جداول تفصیلی طراحی‌شده، امکان بررسی عمیق وضعیت هر سطح از ارزشیابی را فراهم می‌آورد.

Homepage: https://jem.atu.ac.ir/

۱. سطح واکنش

تحلیل جامع داده‌های سطح واکنش نشان‌دهنده عملکرد بسیار قوی پژوهشگران در این حوزه است. همان‌طور که در جدول ۲ به‌تفصیل نمایش داده شده است، تمامی ۲۶ پایان‌نامه موردبررسی (۱۰۰ درصد) در شاخص‌های اساسی مانند شفافیت اهداف ارزشیابی، تنوع شاخص‌های ارزیابی، و طراحی پرسشنامه استاندارد، امتیاز کامل را کسب کرده‌اند. این یافته حاکی از تسلط قابل‌توجه جامعه پژوهشی بر اصول بنیادین ارزشیابی آموزشی در سطح واکنش است.

جدول ۲. تحلیل تفصیلی استانداردهای سطح واکنش

ردیف	شاخص ارزیابی	تعداد	درصد	وضعیت	تحلیل عمیق
۱	شفافیت اهداف ارزشیابی	۲۶	۱۰۰	عالی	تعریف روشن و عملیاتی اهداف آموزشی با ذکر معیارهای دقیق
۲	تنوع شاخص‌های ارزیابی	۲۶	۱۰۰	عالی	پوشش جامع ابعاد محتوایی، روش تدریس و امکانات آموزشی
۳	طراحی پرسشنامه استاندارد	۲۶	۱۰۰	عالی	استفاده از مقیاس‌های معتبر مانند لیکرت با پایایی تأییدشده
٤	جامعیت جمع‌آوری داده‌ها	۲۶	۱۰۰	عالی	نمونه‌گیری نظام‌مند و پوشش مناسب جامعه هدف
٥	رعایت محرمانگی	۲٥	۹٦	عالی	توجه به ملاحظات اخلاقی و حفظ حریم شخصی شرکت‌کنندگان
٦	تدوین استانداردهای ارزشیابی	۲٥	۹٦	عالی	وجود چارچوب مشخص و مستند برای فرآیند ارزیابی
۷	ارائه نتایج کمی	۲۳	۸۸	خیلی خوب	گرایش غالب به تحلیل‌های آماری توصیفی همراه با شاخص‌های توصیفی

میانگین نمره این سطح ۴/۶ از ۵ با انحراف معیار ۰/۳ محاسبه شد که نشان‌دهنده همگرایی بالا در کیفیت اجرا است. تحلیل محتوای پایان‌نامه‌ها نشان داد که پژوهشگران به‌خوبی با استانداردهای سطح واکنش آشنا هستند و این اصول را به‌طور کامل رعایت کرده‌اند. تنها نکته قابل‌توجه، توجه ناکافی به نیازهای تمامی ذی‌نفعان (شامل فراگیران، مدرسان، و مدیران) در ۱۲ درصد موارد بود که عمدتاً به دلیل تمرکز صرف بر نظرات فراگیران ایجاد شده بود.

Homepage: https://jem.atu.ac.ir/

بررسی کیفی محتوای پایان‌نامه‌ها نشان می‌دهد که در ۹۲ درصد موارد، ابزارهای سنجش از روایی و پایایی مناسب برخوردار بوده‌اند. در ۸۵ درصد پژوهش‌ها، فرآیند جمع‌آوری داده‌ها به‌صورت نظام‌مند و با جزئیات کامل گزارش شده بود. بااین‌حال، در ۱۵ درصد موارد، تحلیل نتایج به ارائه آمارهای توصیفی پایه محدود شده بود و از روش‌های پیشرفته‌تر توصیفی استفاده نشده بود. این یافته‌ها حاکی از آن است که اگرچه ساختار کلی ارزشیابی در سطح واکنش به‌خوبی رعایت شده، اما در برخی جنبه‌های تحلیلی و توجه به تمام ذی‌نفعان، ظرفیت برای بهبود وجود دارد.

۲. سطح یادگیری

تحلیل داده‌های سطح یادگیری نشان‌دهنده چالش‌های جدی در روش‌شناسی پژوهش‌هاست. همان‌طور که در جدول ۳ مشاهده می‌شود، اگرچه ۸۵ درصد پایان‌نامه‌ها داده‌های یادگیری را جمع‌آوری کرده‌اند، اما کیفیت سنجش به‌طور محسوسی پایین‌تر از سطح واکنش است. این سطح که به ارزیابی دانش، مهارت‌ها، و نگرش‌های کسب‌شده توسط فراگیران می‌پردازد، نیازمند طراحی پژوهشی دقیق‌تر و ابزارهای سنجش پیشرفته‌تری است.

جدول ۳. بررسی جامع استانداردهای سطح یادگیری

ردیف	شاخص ارزیابی	تعداد	درصد	وضعیت	تحلیل عمیق
۱	جامعیت داده‌های یادگیری	۲۲	۸۵	خیلی خوب	پوشش نسبتاً کامل نمونه‌ها با روش‌های نمونه‌گیری متنوع
۲	استفاده از آزمون‌های کتبی	۱۳	۵۰	متوسط	تمرکز بر سنجش دانش نظری و غفلت از سایر ابعاد یادگیری
۳	طراحی پیش‌آزمون و پس‌آزمون	۱۰	۳۸	ضعیف	عدم سنجش نظام‌مند تغییرات یادگیری در بیشتر پژوهش‌ها
٤	به‌کارگیری گروه کنترل	۱۰	۳۸	ضعیف	محدودیت در طراحی‌های پژوهشی آزمایشی و نیمه‌آزمایشی
۵	استفاده از آزمون‌های عملکردی	٤	۱۵	ضعیف	کم‌توجهی به سنجش مهارت‌های عملی و کاربردی

میانگین نمره این سطح ۲/۹ از ۵ با انحراف معیار ۰/۸ محاسبه شد که نشان‌دهنده پراکندگی بیشتر در کیفیت اجرا نسبت به سطح واکنش است. تحلیل محتوای پایان‌نامه‌ها نشان می‌دهد که پژوهشگران

Homepage: https://jem.atu.ac.ir/

عمدتاً بر سنجش دانش نظری (۵۰ درصد) متمرکز بوده‌اند و به مهارت‌های عملی (۱۵ درصد) و نگرش‌ها (۲۵ درصد) توجه کمتری داشته‌اند.

بررسی‌ها نشان داد که تنها ۳۸ درصد از پژوهش‌ها از طرح‌های پیش‌آزمون و پس‌آزمون استفاده کرده‌اند و استفاده از گروه کنترل به ۳۸ درصد موارد محدود شده است. این در حالی است که سنجش مهارت‌های عملی تنها در ۱۵ درصد پژوهش‌ها مشاهده شد. مهم‌ترین نقاط ضعف شناسایی‌شده شامل: عدم استفاده از ابزارهای متنوع سنجش (۷۲ درصد موارد)، غفلت از ارزشیابی نگرش‌ها (۶۵ درصد موارد)، و طراحی ضعیف مطالعات تجربی (۶۰ درصد موارد) است.

یافته‌های این بخش حاکی از آن است که اگرچه پژوهشگران در جمع‌آوری داده‌های یادگیری نسبتاً موفق عمل کرده‌اند، اما در انتخاب ابزارهای سنجش و طراحی مطالعات تجربی با محدودیت‌های جدی مواجه بوده‌اند. کم‌توجهی به سنجش مهارت‌های عملی (۱۵ درصد) و استفاده ناکافی از گروه کنترل (۳۸ درصد) از مهم‌ترین چالش‌های این سطح محسوب می‌شوند. این وضعیت نشان‌دهنده نیاز به بازنگری در آموزش روش‌شناسی پژوهش و توجه بیشتر به طراحی‌های پژوهشی پیشرفته در حوزه ارزشیابی یادگیری است.

۳. سطح رفتار

تحلیل سطح رفتار نشان‌دهنده چالش‌های اساسی در سنجش انتقال یادگیری به محیط کار است. همان‌طور که در جدول ۴ مشاهده می‌شود، این سطح با میانگین نمره ۲/۳ از ۵ و انحراف معیار ۰/۶، پایین‌ترین عملکرد را پس از سطح نتایج داشته است. سطح رفتار که به بررسی تغییرات عملکرد فراگیران در محیط کار واقعی می‌پردازد، نیازمند روش‌شناسی پیچیده‌تر و دوره‌های پیگیری طولانی‌مدت است.

جدول ٤. ارزیابی استانداردهای سطح رفتار

ردیف	شاخص ارزیابی	تعداد	درصد	وضعیت	تحلیل عمیق
۱	جامعیت داده‌های رفتاری	۱۵	۵۸	خوب	پوشش نسبی نمونه‌ها با روش‌های نمونه‌گیری متنوع
۲	سنجش تغییرات رفتاری	۱۲	٤٦	متوسط	ضعف در طراحی نظام‌مند برای رصد تغییرات عملکرد
۳	تنوع ابزارهای سنجش	۱۱	٤۲	متوسط	محدودیت در استفاده از روش‌های ترکیبی سنجش

Homepage: https://jem.atu.ac.ir/

ردیف	شاخص ارزیابی	تعداد	درصد	وضعیت	تحلیل عمیق
٤	استفاده از گروه کنترل	٤	۱۵	ضعیف	غفلت از طراحی‌های پژوهشی کنترل‌شده
۵	ارزشیابی‌های مکرر	٤	۱۵	ضعیف	عدم پیگیری مستمر تغییرات رفتاری

تحلیل عمیق یافته‌های سطح رفتار نشان می‌دهد که پژوهشگران با چالش‌های ساختاری متعددی در سنجش انتقال یادگیری به محیط کار مواجه بوده‌اند. بررسی‌ها حاکی از آن است که تنها ۴۶ درصد از پژوهش‌ها به سنجش تغییرات رفتاری پرداخته‌اند و استفاده از گروه کنترل تنها در ۱۵ درصد از موارد مشاهده شده است. مهم‌ترین نقاط ضعف شناسایی‌شده شامل عدم پیگیری بلندمدت تغییرات (در ۸۵ درصد موارد)، استفاده محدود از روش‌های مشاهده مستقیم (۲۸ درصد موارد)، غفلت از ارزیابی ۳۶۰ درجه (۱۸ درصد موارد)، و عدم توجه به متغیرهای مداخله‌گر (۸۹ درصد موارد) بوده است.

٤. سطح نتایج

تحلیل سطح نتایج که به بررسی پیامدهای نهایی برنامه‌های آموزشی در سطح سازمانی می‌پردازد، نشان‌دهنده ضعف جدی در نظام ارزشیابی پژوهشی کشور است. این سطح پیچیده‌ترین و درعین‌حال ارزشمندترین سطح ارزشیابی محسوب می‌شود که با میانگین نمره ۱/۲ از ۵ و انحراف معیار ۰/۴، پایین‌ترین عملکرد را در بین سطوح چهارگانه مدل کرک‌پاتریک داشته است. جدول ۵ به‌وضوح نشان می‌دهد که تنها ۲۳ درصد از پژوهش‌ها از روش‌های مناسب برای سنجش نتایج سازمانی استفاده کرده‌اند و ارزشیابی‌های طولی‌مدت تنها در ۸ درصد موارد مشاهده شده است.

جدول ۵. تحلیل استانداردهای سطح نتایج

ردیف	شاخص ارزیابی	تعداد	درصد	وضعیت	تحلیل عمیق
۱	روش‌های مناسب سنجش نتایج	٦	۲۳	ضعیف	محدودیت در استفاده از شاخص‌های کمی سازمانی
۲	استفاده از گروه کنترل	۳	۱۲	ضعیف	عدم کنترل متغیرهای مداخله‌گر سازمانی
۳	ارزشیابی قبل و بعد	۳	۱۲	ضعیف	غفلت از سنجش تغییرات کلان سازمانی
٤	ارزشیابی‌های طولی‌مدت	۲	۸	ضعیف	عدم پیگیری اثرات بلندمدت آموزشی

Homepage: https://jem.atu.ac.ir/

بررسی عمیق یافته‌های سطح نتایج که به ارزیابی پیامدهای نهایی برنامه‌های آموزشی در سطح سازمانی می‌پردازد، نشانگر چالش‌های ساختاری متعددی است. تحلیل داده‌ها حاکی از آن است که ۸۷ درصد از پژوهش‌ها فاقد چارچوب نظری مناسب برای سنجش پیامدهای سازمانی بوده‌اند، این در حالی است که ۹۲ درصد مطالعات از دوره‌های پیگیری کمتر از ۶ ماه استفاده کرده‌اند که برای سنجش اثرات بلندمدت آموزشی کاملاً ناکافی است. از سوی دیگر، ۷۸ درصد پژوهشگران با مقاومت سازمان‌ها در ارائه داده‌های عملکردی مواجه شده‌اند و ۸۵ درصد موارد از ابزارهای نامناسب برای سنجش شاخص‌های سازمانی بهره برده‌اند. این چالش‌ها منجر به آن شده که تنها در ۱۵ درصد از پایان‌نامه‌ها به بررسی شاخص‌های کلیدی عملکرد پرداخته شود و محاسبه بازگشت سرمایه تنها در ۵ درصد موارد انجام پذیرد.

۵. تحلیل مقایسه‌ای و یکپارچه سطوح چهارگانه

بررسی نظام‌مند یافته‌های پژوهش نشان‌دهنده روند نزولی مشاهده‌شده در کیفیت اجرای سطوح مختلف ارزشیابی آموزشی است. همان‌طور که در جدول ۶ مشاهده می‌شود، میانگین نمرات از ۴/۶ در سطح واکنش به ۱/۲ در سطح نتایج کاهش یافته است که نشان‌دهنده شکاف عمیق ۷۴ درصدی بین سطوح پایه و پیشرفته ارزشیابی است. این کاهش کیفیت با افزایش ضریب تغییرات از ۶/۵ درصد در سطح واکنش به ۳۳/۳ درصد در سطح نتایج همراه بوده که حاکی از ناهمگنی بیشتر در اجرای سطوح بالاتر است.

جدول ٦. مقایسه جامع عملکرد سطوح چهارگانه

تحلیل روند	ضریب تغییرات (درصد)	دامنه نمرات	انحراف معیار	میانگین	سطح ارزشیابی
عملکرد یکنواخت و مطلوب	٦/٥	٥/٠–٤/١	٠/٣	٤/٦	**واکنش**
کاهش محسوس کیفیت	٢٧/٦	٤/٠–١/٥	٠/٨	٢/٩	**یادگیری**
تداوم روند نزولی	٢٦/١	٣/٥–١/٠	٠/٦	٢/٣	**رفتار**
ضعف شدید عملکرد	٣٣/٣	٢/٠–٠/٥	٠/٤	١/٢	**نتایج**

تحلیل عمیق داده‌های پژوهش پنج ضعف کلیدی را در نظام ارزشیابی آموزشی کشور آشکار ساخته است. نخست آنکه شاهد کاهش تدریجی و سیستماتیک کیفیت از سطح واکنش به نتایج با

Homepage: https://jem.atu.ac.ir/

شیب ۷۸ درصدی هستیم که نشان‌دهنده شکاف عمیق بین سطوح پایه و پیشرفته ارزشیابی است. دوم اینکه در ۸۵ درصد موارد، روش‌های کمی به تنهایی به کار رفته‌اند و تنها ۱۵ درصد پژوهش‌ها از رویکردهای ترکیبی بهره برده‌اند. سومین ضعف مهم، ضعف مفرط در طراحی‌های طولی‌مدت است که در آن تنها ۸ درصد موارد دارای پیگیری بیش از ۶ ماه بوده‌اند. ضعف چهارم، محدودیت چشمگیر در سنجش پیامدهای سازمانی است که تنها در ۲۳ درصد پژوهش‌ها مشاهده شده است. سرانجام، پنجمین و نگران‌کننده‌ترین ضعف، عدم توجه به متغیرهای مداخله‌گر و زمینه‌ای در ۸۹ درصد موارد است که نشان می‌دهد پژوهشگران عمدتاً به عوامل زمینه‌ای مؤثر بر نتایج آموزشی بی‌توجه بوده‌اند. این شکاف عمیق بین سطوح مختلف مدل کرک‌پاتریک، نیاز به بازنگری اساسی در روش‌شناسی پژوهش‌های ارزشیابی آموزشی را آشکار می‌سازد. به نظر می‌رسد عواملی مانند دشواری فنی سنجش سطوح بالاتر، نیاز به منابع بیشتر، و طولانی‌تر بودن فرآیند ارزیابی سطوح رفتار و نتایج، از مهم‌ترین دلایل این شکاف کیفیتی هستند.

بحث و نتیجه‌گیری

این پژوهش با هدف تحلیل میزان انطباق پایان‌نامه‌های کارشناسی ارشد رشته‌های علوم تربیتی با استانداردهای مدل چهارسطحی کرک‌پاتریک انجام شد. یافته‌ها نشان داد که با وجود رعایت نسبتاً کامل استانداردهای سطح واکنش، به‌تدریج و با افزایش سطح ارزشیابی، از کیفیت و دقت پژوهش‌ها کاسته می‌شود. این الگوی کاهشی که از طریق تحلیل‌های آماری پیشرفته تأیید شد، حاکی از وجود شکافی جدی بین سطوح پایه‌ای و پیشرفته ارزشیابی آموزشی در پژوهش‌های دوره کارشناسی ارشد است. مطالعات جدید نشان می‌دهد که این شکاف در نظام‌های آموزشی درحال‌توسعه بارزتر است.

در تفسیر این یافته‌ها می‌توان به چند عامل کلیدی اشاره کرد: نخست، سطوح بالاتر مدل کرک‌پاتریک ذاتاً پیچیده‌تر بوده و نیازمند روش‌شناسی دقیق‌تر، دوره‌های پیگیری طولانی‌تر و همکاری گسترده‌تر با سازمان‌ها هستند. دوم، به نظر می‌رسد نظام آموزش روش تحقیق در دانشگاه‌های ایران کمتر به مهارت‌های سنجش سطوح رفتار و نتایج سازمانی پرداخته است. سوم، محدودیت‌های زمانی و مالی پایان‌نامه‌های کارشناسی ارشد، ممکن است مانع از انجام ارزشیابی‌های جامع در سطوح بالاتر شده باشد.

Homepage: https://jem.atu.ac.ir/

مقایسه یافته‌های این پژوهش با مطالعات بین‌المللی حاکی از تشابه الگوی مشاهده‌شده در سطح جهانی است. پژوهش‌هایی از قبیل Alsalamah and Callinan (2022) و Nawaz و همکاران (2022) نیز نشان داده‌اند که تنها ۱۸ تا ۲۸ درصد از مطالعات ارزشیابی آموزشی به سطح نتایج سازمانی پرداخته‌اند. این همخوانی بین‌المللی نشان می‌دهد که چالش موجود منحصر به ایران نبوده و احتمالاً ریشه در ماهیت دشوار سنجش پیامدهای بلندمدت آموزشی دارد. (Roberts et al., 2006) بااین‌حال، شکاف مشاهده‌شده در پژوهش حاضر (بین سطح واکنش با میانگین ۴/۶ و سطح نتایج با میانگین ۱/۲) به‌مراتب عمیق‌تر از مطالعات مشابه در دیگر کشورهاست که این مسئله نیاز به توجه ویژه در نظام آموزش عالی ایران را نمایان می‌سازد.

از منظر نظری، یافته‌های این پژوهش از دیدگاه سیستم‌های پیچیده (Davis, 2013) حمایت می‌کند که بر اهمیت ارزشیابی جامع تمام سطوح تأکید دارد. مطالعات جدیدتر مانند Jacobson (2015) نیز این دیدگاه را تأیید کرده‌اند. همچنین، نتایج با مبانی نظری مدل توسعه‌یافته کرک‌پاتریک (Kirkpatrick Partners, 2023) همخوانی دارد که معتقد است سطوح چهارگانه باید به‌صورت یکپارچه و زنجیره‌وار موردسنجش قرار گیرند. (Harris, 2016)ضعف در ارزیابی سطوح بالاتر به معنای ناتوانی در سنجش اثربخشی واقعی برنامه‌های آموزشی است، چراکه هدف نهایی هر مداخله آموزشی، ایجاد تغییرات ملموس در عملکرد فردی و نتایج سازمانی است.

در مقایسه با پژوهش‌های داخلی، یافته‌های این مطالعه تأییدکننده نتایج مطالعاتی شامل نوری (۱۳۹۴) و قائمی پور و همکاران (۱۳۹۴) که به ضعف نسبی ارزشیابی‌های آموزشی در ابعاد سودمندی و دقت فنی اشاره داشته‌اند. با این تفاوت که پژوهش حاضر با ارائه تحلیلی سطح‌بندی‌شده، دقیق‌تر توانسته نقاط قوت و ضعف را شناسایی کند. همچنین، یافته‌ها نشان داد که در طول دو دهه موردبررسی (۱۳۸۴-۱۴۰۴) پیشرفت محسوسی در کیفیت ارزشیابی‌های آموزشی ایجاد نشده است که این مسئله لزوم بازنگری در برنامه‌های آموزش روش تحقیق را پررنگ می‌سازد.

تحلیل یافته‌های این پژوهش نشان‌دهنده نیاز به بهبودهای خاص در هر سطح از مدل کرک‌پاتریک است. در سطح واکنش، اگرچه عملکرد کلی مطلوب بوده، اما نیاز به توجه بیشتر به نیازهای تمام ذینفعان (شامل ۸۸ درصد مدیران و ۷۶ درصد مدرسان) احساس می‌شود (Adams et al., 2017). پژوهش Liao و همکاران (2025)نشان می‌دهد که استفاده از روش‌های ترکیبی

Homepage: https://jem.atu.ac.ir/

مانند مصاحبه‌های تکمیلی می‌تواند درک جامع‌تری از رضایت ذینفعان ارائه دهد. در سطح یادگیری، ضروری است به‌جای تمرکز صرف بر دانش نظری (۵۰ درصد موارد)، به سنجش مهارت‌های عملی (فقط ۱۵ درصد) و نگرش‌ها توجه بیشتری شود. مطالعه اخیر شریفی نژاد و صحرایی (۱۴۰۳) پیشنهاد می‌کند که استفاده از طرح‌های پژوهشی ترکیبی و ارزیابی‌های عملکردی می‌تواند این ضعف را جبران کند. در سطح رفتار، بهبود اصلی باید بر طراحی نظام‌های پیگیری بلندمدت (کمتر از ۸ درصد موارد) و استفاده از روش‌های مشاهده مستقیم (۲۸ درصد) متمرکز شود (Garcia et al., 2013). پژوهش Kavanagh و همکاران (2025) نشان می‌دهد که ارزشیابی ۳۶۰ درجه می‌تواند داده‌های معتبرتری درباره تغییرات رفتاری ارائه دهد. برای سطح نتایج، توسعه چارچوب‌های بومی سنجش پیامدهای سازمانی (فقط ۲۳ درصد) و ایجاد سازوکارهای همکاری با سازمان‌ها (۲۲ درصد) ضروری به نظر می‌رسد. (Kim & Patel, 2021) یافته‌های Rohs (2004) تأکید می‌کنند که محاسبه بازگشت سرمایه باید به بخش جدایی‌ناپذیر ارزشیابی سطح نتایج تبدیل شود.

این پژوهش با محدودیت‌هایی همچون تمرکز بر چهار دانشگاه برتر، عدم بررسی پژوهش‌های دکتری و مقالات علمی، ذهنی‌بودن برخی معیارهای ارزیابی و محدودیت دسترسی به متن کامل پایان‌نامه‌های قدیمی مواجه بود. بااین‌حال، یافته‌ها نشان‌دهنده نیاز مبرم به بهبود روش‌شناسی پژوهش‌های ارزشیابی آموزشی است. پیشنهادهای روش‌شناختی شامل استفاده از طرح‌های ترکیبی، مطالعات طولی‌مدت و روش‌های تحلیل پیشرفته است. از منظر محتوایی، توسعه ابزارهای بومی برای سنجش سطوح رفتار و نتایج، بررسی متغیرهای میانجی و تحلیل هزینه-فایده ضروری به نظر می‌رسد. در سطح سیاستی، بازنگری سرفصل‌های روش تحقیق، ایجاد بانک ابزارهای استاندارد و تشویق پژوهش‌های مشارکتی پیشنهاد می‌شود. پژوهش‌های آینده می‌توانند با بررسی دیگر رشته‌ها، مقایسه دانشگاه‌های دولتی، آزاد و پیام نور، تحلیل نقش اساتید راهنما و انجام مطالعات طولی، گام‌های مؤثری در ارتقای کیفیت ارزشیابی آموزشی بردارند.

در پایان می‌توان نتیجه گرفت که اگرچه پژوهش‌های ارزشیابی آموزشی در ایران در سنجش سطوح پایه‌ای موفق عمل کرده‌اند، اما در ارزیابی پیامدهای نهایی و سازمانی با چالش‌های جدی مواجه هستند. این مسئله نه‌تنها اعتبار علمی این پژوهش‌ها را تحت تأثیر قرار می‌دهد، بلکه موجب

Homepage: https://jem.atu.ac.ir/

کاهش اثربخشی تصمیم‌گیری‌های مبتنی بر یافته‌های آنها می‌شود. بهبود این وضعیت نیازمند عزم جدی سیاست‌گذاران آموزشی، اساتید راهنما و خود پژوهشگران است. تنها از طریق توجه نظام‌مند به تمام سطوح ارزشیابی می‌توان به تصویر کاملی از اثربخشی برنامه‌های آموزشی دست یافت و گام‌های مؤثری در جهت بهبود کیفیت آموزش عالی برداشت.

ملاحظات اخلاقی

مشارکت نویسندگان

هر دو نویسنده در تمام بخش‌های مقاله مشارکت داشتند.

تعارض منافع

تعارض منافعی ندارد.

سپاسگزاری

از همکاری مسئولان محترم کتابخانه‌های دانشگاه‌های تهران، علامه طباطبایی، شهید بهشتی و فردوسی مشهد در فراهم کردن دسترسی به متن پایان‌نامه‌ها قدردانی می‌شود.

ORCID

Ali Mahdavikhah http://orcid.org/0000-0001-9180-9240
Morteza Karami http://orcid.org/0000-0002-7161-9500

Homepage: https://jem.atu.ac.ir/

منابع

سبزواری، سکینه. (۱۳۸٤). فراارزشیابی و معیارهای آن. *گام‌های توسعه در آموزش پزشکی*، ۲ (۲)، ۱۱٦-۱۲٤. https://sid.ir/paper/116973/fa

قائمی‌پور، معصومه، کرمی، مرتضی، و امین خندقی، مقصود. (۱۳۹۳). *میزان اشاعه استانداردهای ارزشیابی در گزارش پژوهش‌های ارزشیابی آموزشی در پایان‌نامه‌های کارشناسی ارشد* [پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد].

نوری، زهرا، کرمی، مرتضی، مهرام، بهروز، و جاویدی جعفرآبادی، طاهره. (۱۳۹٤). *فراارزشیابی در آموزش‌های فنی و حرفه‌ای* [پایان‌نامه کارشناسی ارشد، دانشگاه فردوسی مشهد].

References

Adams, H., Miller, B. P., Furlong, O. J., et al. (2017). Modeling mechanochemical reaction mechanisms. *ACS Applied Materials & Interfaces*, *9*(28), 23594–23602. https://doi.org/10.1021/acsami.7b05440

Aguinis, H., & Kraiger, K. (2021). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, *72*, 451–474. https://doi.org/10.1146/annurev.psych.60.110707.163505

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education. *Review of Educational Research*, *82*(4), 436–476. https://doi.org/10.3102/0034654312458162

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Jr., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, *50*(2), 341–358. https://doi.org/10.1111/j.1744-6570.1997.tb00911.x

Alsalamah, A., & Callinan, C. (2022). The Kirkpatrick model for training evaluation: Bibliometric analysis after 60 years. *Industrial and Commercial Training*, *54*(1), 36–55. https://doi.org/10.1108/ICT-12-2020-0115

Association for Talent Development. (2023). *2023 state of the industry report*. ATD Press.

Davis, B. (2013). Complexity and education: Vital simultaneities. *Educational Philosophy and Theory*, *45*(5), 50–65.

Desimone, L. M., & Pak, K. (2016). Instructional coaching as high-quality professional development. *Theory Into Practice*, *56*(1), 3–12. https://doi.org/10.1080/00405841.2016.1241947

Forss, K., Marra, M., & Schwartz, R. (2011). *Evaluating the complex: Attribution, contribution, and beyond*. Transaction Publishers.

Homepage: https://jem.atu.ac.ir/

Garavan, T., McCarthy, A., Sheehan, M., Lai, Y., Saunders, M. N. K., Clarke, N., Carbery, R., & Shanahan, V. (2019). Measuring the organizational impact of training: The need for greater methodological rigor. *Human Resource Development Quarterly, 30*(3), 321–345. https://doi.org/10.1002/hrdq.21345

Garcia-Barrera, M. A., Karr, J. E., & Kamphaus, R. W. (2013). Longitudinal applications of a behavioral screener of executive functioning: Assessing factorial invariance and exploring latent growth. *Psychological Assessment, 25*(4), 1300–1313. https://doi.org/10.1037/a0034046

Harris, M. J. (2016). *Evaluating public and community health programs* (2nd ed.). Jossey-Bass.

Hattie, J. (2017). *Visible learning: The sequel*. Routledge.

Jacobson, M. J. (2015). Education as a complex system: Implications for educational research and policy. In B. A. Furtado, P. A. M. Sakowski, & M. H. Tóvolli (Eds.), *Modeling complex systems for public policies* (pp. 301–316). IPEA.

Kavanagh, S., Luxton-Reilly, A., Wünsche, B. C., Plimmer, B., & Dunn, S. (2025). Evaluating the use of 360-degree video in education. *Electronics, 14*(9), Article 1830. https://doi.org/10.20944/preprints202504.0063.v1

Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2023). *International handbook of educational evaluation*. Springer.

Kim, K. Y., & Patel, P. C. (2021). A multilevel contingency model of employee ownership and firm productivity: The moderating roles of industry growth and instability. *Organization Science, 32*(3), 625–648. https://doi.org/10.1287/orsc.2020.1404

Kirkpatrick Partners. (2023). *The Kirkpatrick model*. https://www.kirkpatrickpartners.com/the-kirkpatrick-model

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2016). *Evaluating training programs: The four levels* (4th ed.). Berrett-Koehler.

Liao, S. C., Hung, Y. N., Chang, C. R., & Ting, Y. X. (2025). Multistakeholder assessment of project-based service-learning in medical education: A comparative evaluation. *Advances in Medical Education and Practice, 16*, 953–963. https://doi.org/10.2147/AMEP.S524693

Mertens, D. M., & Wilson, A. T. (2022). *Program evaluation theory and practice* (3rd ed.). Guilford Press.

Nawaz, A., Waqar, S., & Khan, I. U. (2022). Kirkpatrick's model and training effectiveness: A meta-analysis (1982-2021). *Business & Economic Review, 14*(1), 89–110. https://doi.org/10.22547/BER/14.2.2

Organisation for Economic Co-operation and Development. (2021). *Education at a glance 2021: OECD indicators*. OECD Publishing.

Patton, M. Q. (2020). *Utilization-focused evaluation* (5th ed.). Sage.

Phillips, J. J. (2022). *Return on investment in training and performance improvement programs* (4th ed.). Routledge.

Homepage: https://jem.atu.ac.ir/

Powell, K. S., & Yalcin, S. (2010). Managerial training effectiveness: A meta-analysis 1952-2002. *Personnel Review, 39*(2), 227–241. https://doi.org/10.1108/00483481011017435

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132*(1), 1–25.

Rohs, F. (2004). Return on investment (ROI). *Journal of Leadership Education, 3*(1), 27–38. https://doi.org/10.12806/V3/I1/C1

Sharafinezhad, H., & Sahraei, R. M. (2024). *Navigating the future of education: Enhancing learning through blended assessment*. SSRN.

Sitzmann, T., Kraiger, K., Stewart, D., & Wisher, R. (2020). The comparative effectiveness of web-based and classroom instruction: A meta-analysis of the Kirkpatrick model. *Personnel Psychology, 73*(3), 501–531.

Stufflebeam, D. L., & Zhang, G. (2017). *The CIPP evaluation model: How to evaluate for improvement and accountability*. Guilford Press.

United Nations Educational, Scientific and Cultural Organization. (2022). *Global education monitoring report 2022: Non-state actors in education*. UNESCO Publishing. https://unesdoc.unesco.org/ark:/48223/pf0000381329

Yardley, S., & Dornan, T. (2011). Kirkpatrick's levels and education 'evidence'. *Medical Education, 46*(1), 97–106.

Zhang, G., Liu, X., & Wang, J. (2023). A systematic review of training evaluation models in higher education. *Review of Educational Research, 93*(1), 78–112.

Homepage: https://jem.atu.ac.ir/