

شناخت، رفتار، یادگیری

بازمفهوم سازی روایی در پژوهش های تربیت معلم: بررسی دقت اندازه گیری، شواهد تفسیری و پیامدی

زهره باقرزاده^{۱*}

۱. گروه علوم تربیتی، دانشگاه فرهنگیان، تهران، ایران

* ایمیل نویسنده مسئول: bagherzadehzohreh@cfu.ac.ir

تاریخ ارسال: ۱۴۰۵/۰۳/۰۸

تاریخ بازنگری: ۱۴۰۵/۰۵/۱۲

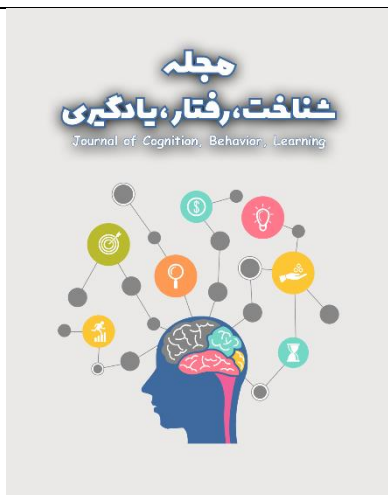
تاریخ پذیرش: ۱۴۰۵/۰۶/۲۴

تاریخ چاپ: ۱۴۰۶/۰۳/۰۱

چکیده

هدف این پژوهش، بازمفهوم سازی روایی در پژوهش های تربیت معلم از طریق بررسی دقت اندازه گیری، شواهد تفسیری و شواهد پیامدی و تبیین نحوه ارتباط این سه سطح در شکل گیری یک استدلال روایی منسجم بود. پژوهش حاضر با طرح مروری تحلیلی انجام شد و مطالعات منتشر شده در فاصله سال های ۲۰۱۰ تا ۲۰۲۵ را در حوزه سنجش و روایی در تربیت معلم بررسی کرد. در جست و جوی اولیه ۲۸۶ رکورد شناسایی شد که پس از حذف موارد تکراری، غربالگری عنوان و چکیده و ارزیابی متن کامل، ۴۴ مطالعه واجد معیارهای ورود به تحلیل نهایی شدند. داده ها با استفاده از فرم ساختاریافته استخراج شواهد روایی گردآوری و در سه محور دقت اندازه گیری، شواهد تفسیری و شواهد پیامدی کدگذاری شدند. تحلیل داده ها با استفاده از تحلیل تو صیفی، تحلیل محتوای جهت دار و سنتز تفسیری انجام گرفت. نتایج نشان داد که شواهد روایی در مطالعات تربیت معلم به طور نامتوازن توزیع شده اند. ۷۹٫۵ درصد مطالعات شاخص های همسانی درونی و ۸۱٫۸ درصد شواهد ساختار درونی را گزارش کرده بودند، در حالی که تنها ۳۱٫۸ درصد ارتباط صریحی میان شواهد گردآوری شده و تفسیر مورد نظر از نمرات برقرار کرده بودند. شواهد مربوط به تغییر ناپذیری اندازه گیری در ۲۰٫۵ درصد، فرایندهای پاسخ در ۱۸٫۲ درصد و عملکرد افتراقی گویه ها در ۱۳٫۶ درصد مطالعات مشاهده شد. مهم تر از همه، ۶۵٫۹ درصد مطالعات فاقد شواهد مستقیم درباره پیامدهای استفاده از نمرات بودند و تنها ۱۸٫۲ درصد مطالعات پوشش نسبتاً جامعی از هر سه حوزه دقت، تفسیر و پیامد ارائه کردند. یافته ها نشان می دهد که روایی در پژوهش های تربیت معلم باید از یک ویژگی ثابت ابزار به یک استدلال انباشتی و زمینه مند درباره دقت نمرات، معنای تفسیرها و مشروعیت کاربرد نتایج بازتعریف شود.

کلیدواژه ها: روایی، تربیت معلم، دقت اندازه گیری، شواهد تفسیری، شواهد پیامدی، سنجش آموزشی.



شیوه استناددهی: باقرزاده، زهره. (۱۴۰۶). بازمفهوم سازی روایی در پژوهش های تربیت معلم: بررسی دقت اندازه گیری، شواهد تفسیری و پیامدی. شناخت، رفتار، یادگیری، ۴ (۳)، ۱-۲۴.



Reconceptualizing Validity in Teacher Education Research: Examining Measurement Precision, Interpretive Evidence, and Consequential Evidence

Zohreh Bagherzadeh^{1*}

1. Department of Educational Sciences, Farhangian University, Tehran, Iran.

*Corresponding Author's Email: bagherzadehzohreh@cfu.ac.ir

Submit Date: 2026-05-29

Revise Date: 2026-08-03

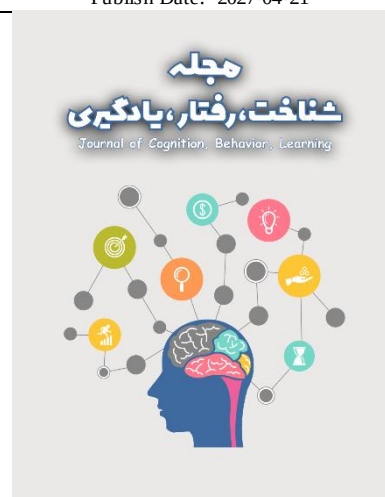
Accept Date: 2026-09-15

Publish Date: 2027-04-21

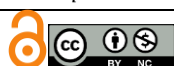
Abstract

This study aimed to reconceptualize validity in teacher education research by examining measurement precision, interpretive evidence, and consequential evidence and clarifying how these three domains contribute to a coherent validity argument. An analytical review design was employed to examine studies published between 2010 and 2025 on measurement and validity in teacher education. The initial search identified 286 records; after duplicate removal, title and abstract screening, and full-text assessment, 44 studies met the inclusion criteria and entered the final analysis. Data were extracted using a structured validity-evidence form and coded across three domains: measurement precision, interpretive evidence, and consequential evidence. Descriptive analysis, directed content analysis, and interpretive synthesis were used to identify dominant patterns, gaps, and inconsistencies in the reviewed studies. The results revealed a marked imbalance in the distribution of validity evidence. Internal consistency indices were reported in 79.5% of studies, and evidence of internal structure appeared in 81.8%. However, only 31.8% of studies explicitly connected the collected evidence to the intended interpretation of scores. Measurement invariance was examined in 20.5% of studies, response-process evidence in 18.2%, and differential item functioning or measurement bias in 13.6%. Most notably, 65.9% of studies provided no direct consequential evidence concerning score use, and only 18.2% demonstrated relatively comprehensive coverage across measurement precision, interpretive evidence, and consequential evidence. Validity in teacher education research should be reconceptualized from a fixed property of an instrument into a cumulative, context-dependent argument concerning score precision, the defensibility of interpretations, and the justification of score use.

Keywords: *validity, teacher education, measurement precision, interpretive evidence, consequential evidence, educational assessment.*



How to cite: Bagherzadeh, Z. (2027). Reconceptualizing Validity in Teacher Education Research: Examining Measurement Precision, Interpretive Evidence, and Consequential Evidence. *Cognition, Behavior, Learning*, 4(3), 1-24.



مقدمه

تربیت معلم یکی از بنیادی‌ترین حوزه‌های نظام آموزشی است؛ زیرا کیفیت آماده‌سازی حرفه‌ای معلمان، به‌طور مستقیم و غیرمستقیم بر کیفیت تدریس، یادگیری دانش‌آموزان، اجرای برنامه‌های درسی، تحقق سیاست‌های آموزشی و توانایی نظام آموزشی برای پاسخ‌گویی به تحولات اجتماعی و فناورانه اثر می‌گذارد. در دهه‌های اخیر، دامنه انتظارات از برنامه‌های تربیت معلم به‌طور چشمگیری گسترش یافته است و دیگر نمی‌توان آماده‌سازی معلم را صرفاً به انتقال دانش موضوعی و روش‌های تدریس محدود کرد. معلم امروز باید مجموعه‌ای از شایستگی‌های پیچیده شامل دانش تخصصی، دانش تربیتی، سواد دیجیتال، توانایی استفاده از فناوری، شایستگی ارزیابی، هویت حرفه‌ای، حساسیت نسبت به تنوع یادگیرندگان، ظرفیت آموزش فراگیر و توانایی انطباق با تغییرات برنامه درسی را در خود توسعه دهد. پژوهش‌های مرتبط با تربیت معلم نیز متناسب با این تحول، طیف گسترده‌ای از سازه‌ها و پیامدها را مورد سنجش قرار داده‌اند؛ از هویت حرفه‌ای و خودکارآمدی گرفته تا شایستگی فناوری، نگرش نسبت به آموزش فراگیر، توانایی تدریس و کیفیت عملکرد حرفه‌ای (Ajani, 2024; Alpaslan et al., 2023; Alshehri & Youssef, 2022). در چنین زمینه‌ای، کیفیت ابزارهای سنجش و اعتبار استنباط‌هایی که بر مبنای نتایج این ابزارها شکل می‌گیرد، به یکی از مسائل مرکزی پژوهش‌های تربیت معلم تبدیل شده است.

یکی از ویژگی‌های برجسته پژوهش‌های جدید تربیت معلم، افزایش وابستگی به ابزارهای اندازه‌گیری برای ارزیابی سازه‌هایی است که غالباً پیچیده، چندبعدی و وابسته به بافت هستند. برای نمونه، مطالعات مختلف به سنجش شایستگی‌های فناورانه و تلفیق فناوری در آموزش پرداخته‌اند و نشان داده‌اند که ارزیابی آمادگی معلمان برای استفاده از فناوری مستلزم توجه هم‌زمان به ابعاد دانشی، نگرشی و عملکردی است (Durusoy & Karamete, 2023; Li, 2023; Saleh et al., 2024). توسعه فناوری‌های جدید نیز این پیچیدگی را افزایش داده است. ظهور هوش مصنوعی مولد موجب شده است سازه‌های تازه‌ای مانند شایستگی هوش مصنوعی، نگرش نسبت به استفاده از هوش مصنوعی و قصد به‌کارگیری آن در تدریس وارد ادبیات تربیت معلم شود (Acquah et al., 2026; Devaraj, 2026). همچنین، مطالعات مربوط به آموزش برنامه‌نویسی، فناوری‌های واقعیت افزوده و فناوری‌های آموزشی دیجیتال نشان می‌دهند که سنجش آمادگی حرفه‌ای دانشجومعلمان بیش از گذشته به متغیرهایی وابسته شده است که در مرز میان دانش، باور، نگرش، تجربه و عملکرد قرار دارند (Baltynova et al., 2023; Mena et al., 2023a, 2023b; Sanusi et al., 2025). در نتیجه، هرگونه ادعای علمی درباره کیفیت دانشجومعلمان یا اثربخشی برنامه‌های تربیت معلم تا حد زیادی به دقت و اعتبار اندازه‌گیری این سازه‌ها وابسته است.

با وجود این وابستگی گسترده به سنجش، مفهوم روایی در بسیاری از پژوهش‌های آموزشی همچنان به‌صورت ساده‌شده و گاه تقلیل‌یافته مورد استفاده قرار می‌گیرد. در رویکردهای سنتی، روایی اغلب به‌عنوان ویژگی خودِ ابزار تلقی شده و با اصطلاحاتی مانند «روایی پرسشنامه»، «روایی سازه» یا «روایی محتوایی» معرفی می‌شود. در چنین رویکردی، پژوهشگر ممکن است پس از گزارش قضاوت چند متخصص، یک تحلیل عاملی و یک ضریب پایایی، ابزار را «رو و پایا» معرفی کند. در حالی که رویکردهای معاصر به سنجش بر این نکته تأکید دارند که روایی اساساً به تفسیر و استفاده از نمرات مربوط است، نه به وجود یک ویژگی ثابت و ذاتی در ابزار. به بیان دیگر، پرسش اصلی آن نیست که آیا یک پرسشنامه «روایی دارد»، بلکه باید پرسید شواهد موجود تا چه اندازه از تفسیر خاصی که پژوهشگر از نمرات ارائه می‌کند و کاربرد مشخصی که برای این نمرات در نظر می‌گیرد حمایت می‌کنند. این مسئله در حوزه تربیت معلم اهمیت مضاعفی دارد؛ زیرا نتایج سنجش ممکن است نه تنها در پژوهش‌های دانشگاهی، بلکه در ارزیابی آمادگی حرفه‌ای، طراحی برنامه‌های آموزشی، تشخیص نیازهای توسعه حرفه‌ای و حتی قضاوت درباره کیفیت برنامه‌های تربیت معلم مورد استفاده قرار گیرند.

نمونه‌های موجود در ادبیات نشان می‌دهد که پژوهشگران به‌طور فزاینده‌ای در حال توسعه و اعتبارسنجی ابزارهای اختصاصی برای سنجش سازه‌های تربیت معلم هستند. برای مثال، ابزارهایی برای ارزیابی نگرش دانشجومعلمان نسبت به آموزش فراگیر توسعه یافته‌اند و تحلیل‌های روان‌سنجی برای بررسی ساختار، ثبات و قابلیت تفسیر آنها انجام شده است (Muhammad, 2025). در پژوهشی دیگر، مقیاسی برای

سنجش ادراک معلمان آینده از آموزش فراگیر طراحی شده است که بر اهمیت سنجش دقیق نگرش‌ها و باورهای مرتبط با پذیرش تفاوت‌های یادگیرندگان تأکید دارد (García et al., 2025). همچنین، توسعه ابزارهای مربوط به شایستگی حرفه‌ای معلمان و ارزیابی برنامه‌های آموزشی نشان می‌دهد که استفاده از سنجش‌های استاندارد به بخش جدایی‌ناپذیری از پژوهش‌های تربیت معلم تبدیل شده است (Erviana et al., 2021; Mujtahid et al., 2023). با این حال، وجود شاخص‌های روان‌سنجی مطلوب به‌تنهایی تضمین نمی‌کند که استنباطی که از نتایج استخراج می‌شود صحیح، موجه و قابل تعمیم باشد.

این مسئله به‌ویژه زمانی آشکار می‌شود که سازه‌های مورد سنجش ماهیتی زمینه‌ای و وابسته به موقعیت دارند. برای مثال، هویت حرفه‌ای معلمان و مدرسان تربیت معلم تحت تأثیر عوامل سازمانی، روان‌شناختی و اجتماعی شکل می‌گیرد و ممکن است در موقعیت‌های مختلف آموزشی معانی متفاوتی پیدا کند (Farwa et al., 2024). همچنین، هویت تدریس در شرایط تغییر برنامه درسی می‌تواند با فرایند تحول در شیوه‌های تدریس مرتبط باشد و سنجش آن نیازمند توجه به زمینه اجرای برنامه و برداشت معلم از نقش حرفه‌ای خود است (Lin et al., 2024). روابط میان ادراک حرفه‌ای، حمایت اجتماعی و هویت حرفه‌ای نیز نشان می‌دهد که حتی سازه‌هایی که با پرسشنامه‌های استاندارد سنجیده می‌شوند، در شبکه‌ای از متغیرهای زمینه‌ای و اجتماعی قرار دارند (Alpaslan et al., 2023). بنابراین، یک ابزار ممکن است در یک گروه یا بافت ساختار آماری مناسبی داشته باشد، اما تفسیر همان نمرات در یک جامعه، زبان، فرهنگ یا مرحله حرفه‌ای دیگر بدون شواهد اضافی قابل دفاع نباشد.

از منظر سنجش، یکی از پیش‌شرط‌های بنیادین برای هر استدلال روایی، دقت اندازه‌گیری است. اگر نمرات با خطای زیاد تولید شوند، ثبات کافی نداشته باشند یا نسبت به تغییرات واقعی سازه حساس نباشند، حتی تفسیر نظری مناسب نیز نمی‌تواند استنباطی معتبر ایجاد کند. با این حال، دقت اندازه‌گیری نباید با روایی یکسان فرض شود. ضرابی مانند آلفای کرونباخ، امگا یا پایایی ترکیبی تنها بخشی از اطلاعات مربوط به کیفیت نمرات را فراهم می‌کنند و به‌خودی‌خود معنای نمره یا مناسبیت کاربرد آن را تعیین نمی‌کنند. مطالعاتی که بر توسعه ابزارها و ارزیابی شایستگی‌های معلمان متمرکز بوده‌اند، اهمیت بررسی هم‌زمان روایی و پایایی را نشان داده‌اند (Erviana et al., 2023; Nabhan & Habók, 2025). برای مثال، تحلیل عاملی اکتشافی می‌تواند ساختار احتمالی یک مقیاس را آشکار کند، اما این تحلیل صرفاً یک گام در زنجیره استدلال روایی است و باید با شواهدی درباره محتوا، روابط با سایر متغیرها، فرایند پاسخ و کاربرد نتایج تکمیل شود.

در این میان، گرایش پژوهش‌های تربیت معلم به استفاده از مقیاس‌های ادراکی و خودگزارشی نیز مسائل خاصی برای تفسیر نمرات ایجاد می‌کند. پژوهش‌های مربوط به ادراک دانشجومعلم از آموزش برنامه‌نویسی، رضایت آنان از برنامه‌های تربیت معلم، دانش ادراک شده درباره آموزش متمایز و نگرش نسبت به فناوری نشان می‌دهد که بخش قابل توجهی از دانش تولیدشده در این حوزه بر گزارش‌های خودارزیابی استوار است (Fang et al., 2024; Li, 2023; Nketsia et al., 2024; Sanusi et al., 2025). اگرچه این سنجش‌ها برای بررسی باورها، ادراکات و نگرش‌ها ضروری هستند، اما تفسیر آنها به‌عنوان شاخص مستقیم شایستگی یا عملکرد حرفه‌ای می‌تواند با مشکل مواجه شود. بالا بودن نمره خودگزارشی لزوماً به معنای عملکرد واقعی بالاتر نیست و استفاده از چنین نتایجی برای استنباط درباره صلاحیت تدریس نیازمند شواهد تکمیلی از منابع دیگر است.

تفاوت میان «آنچه ابزار اندازه‌گیری می‌کند» و «آنچه پژوهشگر از نمره استنباط می‌کند» در مطالعات ارزیابی برنامه‌های تربیت معلم نیز اهمیت دارد. ارزیابی کیفیت برنامه‌های آماده‌سازی معلمان گاه بر اساس ادراک دانشجومعلم، رضایت از برنامه، ارزیابی شایستگی فارغ‌التحصیلان یا میزان انطباق برنامه با استانداردهای حرفه‌ای انجام می‌شود (Ayasrah, 2023; Olvido et al., 2024). چنین داده‌هایی می‌توانند اطلاعات ارزشمندی درباره کیفیت برنامه فراهم کنند، اما هر یک فقط بخشی از سازه پیچیده «کیفیت تربیت معلم» را نشان می‌دهند. مطالعات مربوط به کیفیت معلمان و کیفیت آموزش نیز نشان داده‌اند که عوامل فردی، سازمانی، آموزشی و ساختاری به‌طور هم‌زمان در کیفیت عملکرد حرفه‌ای

نقش دارند (Ekabu, 2023; Guo et al., 2023). از این رو، نسبت دادن یک تفسیر کلی درباره کیفیت برنامه یا شایستگی حرفه‌ای صرفاً بر اساس یک ابزار می‌تواند دامنه استنباط را فراتر از شواهد موجود گسترش دهد.

یکی دیگر از ابعاد مهم در مفهوم جدید روایی، بررسی شواهد تفسیری است. شواهد تفسیری به این پرسش مربوط می‌شوند که آیا داده‌های حاصل از سنجش واقعاً از معنایی که پژوهشگر به نمره نسبت می‌دهد پشتیبانی می‌کنند یا خیر. چنین شواهدی ممکن است از بررسی محتوای ابزار، فرایند پاسخ‌دهی، ساختار درونی، رابطه با متغیرهای بیرونی یا مقایسه میان گروه‌ها به دست آیند. در مطالعات مربوط به TPACK، شایستگی فناوری و توانایی استفاده از رایانه در تدریس، پژوهشگران معمولاً روابط میان دانش فناوریانه، دانش تربیتی و دانش محتوایی را بررسی کرده‌اند تا از ساختار نظری سازه حمایت کنند (Ajani, 2024; Durusoy & Karamete, 2023; Saleh et al., 2024). در حوزه شایستگی هوش مصنوعی نیز تحلیل شبکه‌ای و تحلیل پروفایل پنهان نشان می‌دهد که شیوه‌های جدید تحلیل داده می‌توانند اطلاعات عمیق‌تری درباره ساختار سازه و الگوهای متفاوت دانش‌جو معلمان ارائه کنند (Acquah et al., 2026). با این حال، پیچیده‌تر شدن تکنیک آماری به تنهایی به معنای افزایش روایی نیست؛ بلکه ارزش این تحلیل‌ها زمانی آشکار می‌شود که به‌روشنی مشخص شود هر یافته چگونه از یک تفسیر نظری یا کاربرد مشخص حمایت می‌کند.

همچنین، اعتبار تفسیر نمرات در تربیت معلم باید با توجه به تنوع گروه‌های هدف بررسی شود. توسعه ابزارها برای سنجش نگرش به آموزش فراگیر، دانش آموزش متمایز، شایستگی‌های دیجیتال یا هویت حرفه‌ای نشان می‌دهد که جمعیت‌های تربیت معلم از نظر رشته تحصیلی، مرحله آموزش، تجربه حرفه‌ای، فرهنگ و زمینه نهادی متفاوت‌اند (Farwa et al., 2024; García et al., 2025; Nketsia et al., 2024). این تفاوت‌ها می‌توانند بر نحوه فهم گویه‌ها، الگوهای پاسخ و معنای نمرات اثر بگذارند. در نتیجه، مقایسه گروه‌ها بدون بررسی هم‌ارزی یا تغییرناپذیری اندازه‌گیری می‌تواند به نتیجه‌گیری‌های نادرست منجر شود. همین مسئله در پژوهش‌هایی که برنامه‌های تربیت معلم را در بافت‌های فرهنگی یا ملی مختلف ارزیابی می‌کنند نیز مطرح است؛ زیرا استانداردها و انتظارات حرفه‌ای ممکن است در نظام‌های آموزشی مختلف یکسان نباشند (Ayasrah, 2023; Olvido et al., 2024).

بعد سوم که در بسیاری از مطالعات کمتر مورد توجه قرار گرفته، شواهد پیامدی است. اگر نمرات برای تصمیم‌گیری درباره افراد، برنامه‌ها یا سیاست‌ها استفاده می‌شوند، بررسی صرف دقت و ساختار ابزار کافی نیست و باید پیامدهای استفاده از نتایج نیز مورد ارزیابی قرار گیرد. برای نمونه، استفاده از یک مقیاس برای تشخیص میزان آمادگی دانش‌جو معلمان جهت تدریس فراگیر می‌تواند بر طراحی دوره‌های تکمیلی یا تصمیم‌گیری درباره صلاحیت حرفه‌ای آنان اثر بگذارد. در این شرایط، خطای طبقه‌بندی یا تفاوت عملکرد ابزار در گروه‌های مختلف می‌تواند پیامدهای آموزشی و حرفه‌ای واقعی ایجاد کند. پژوهش‌های مربوط به ارزیابی معلم نشان داده‌اند که ارزیابی کیفیت تدریس نه تنها مسئله‌ای فنی، بلکه فرایندی چندبعدی است که می‌تواند بر توسعه حرفه‌ای، تصمیم‌های مدیریتی و برداشت معلمان از عملکرد خود اثر بگذارد (Wei et al., 2023). از این منظر، روایی باید نه فقط در مرحله تولید نمره، بلکه در مرحله استفاده از نمره نیز مورد بررسی قرار گیرد.

اهمیت توجه به پیامدها زمانی بیشتر می‌شود که یافته‌های پژوهش برای اصلاح برنامه‌های تربیت معلم به کار گرفته شوند. مطالعات مربوط به کیفیت برنامه‌های تربیت معلم، اثربخشی آموزش مبتنی بر شایستگی، آمادگی معلمان برای تدریس خواندن و استفاده از فناوری‌های آموزشی نشان می‌دهد که نتایج پژوهش می‌توانند مبنای تغییر در برنامه درسی، روش آموزش و استانداردهای حرفه‌ای قرار گیرند (Baltynova et al., 2021; Ekabu, 2023; Tortorelli et al., 2023). در چنین شرایطی، اگر تفسیر نتایج بر شواهد ناکافی استوار باشد، پیامدهای تصمیم‌گیری می‌تواند از سطح یک پژوهش فراتر رود و طراحی دوره‌های آموزشی یا ارزیابی کیفیت مؤسسات را تحت تأثیر قرار دهد. مرورهای مربوط به ارزیابی تأثیر برنامه‌های آماده‌سازی معلم نیز نشان داده‌اند که دستیابی به استنتاج‌های علی و قابل اتکا درباره اثربخشی برنامه‌ها با چالش‌های روش‌شناختی قابل توجهی روبه‌رو است (Mancenido, 2023). بنابراین، اعتبار نتیجه‌گیری درباره «اثربخشی» یا «کیفیت» برنامه‌ها مستلزم هماهنگی میان طراحی پژوهش، کیفیت اندازه‌گیری و منطق استنباط است.

در همین راستا، پژوهش‌های مربوط به آموزش فراگیر نیز نشان می‌دهند که اندازه‌گیری در تربیت معلم اغلب با ارزش‌ها و پیامدهای اجتماعی پیوند دارد. آماده‌سازی معلمان برای پاسخ‌گویی به نیازهای یادگیرندگان متنوع، از جمله دانش‌آموزانی که با دشواری‌های زبانی یا یادگیری مواجه‌اند، مستلزم ارزیابی دقیق دانش، نگرش و ظرفیت حرفه‌ای معلمان است (Yildirim, 2026). سنجش نگرش دانشجومعلمان نسبت به آموزش فراگیر نیز تنها زمانی سودمند است که روشن باشد نمرات به چه معنا هستند و چگونه می‌توان از آنها برای بهبود برنامه‌های تربیت معلم استفاده کرد (Muhammad, 2025). اگر چنین مقیاس‌هایی بدون بررسی زمینه فرهنگی، فرایند پاسخ یا پیامدهای کاربردی به کار روند، خطر آن وجود دارد که تفاوت‌های واقعی میان افراد و تفاوت‌های ناشی از ویژگی ابزار با یکدیگر اشتباه شوند.

همین مسئله در سنجش شایستگی‌های نوظهور نیز دیده می‌شود. در حوزه هوش مصنوعی، برای مثال، ممکن است پژوهشگر بر اساس یک نمره خودگزارشی نتیجه بگیرد که دانشجومعلمان از «شایستگی هوش مصنوعی» بالایی برخوردارند، در حالی که چنین نمره‌ای ممکن است بیشتر بازتاب اعتمادبه‌نفس، تجربه قبلی یا نگرش مثبت نسبت به فناوری باشد تا توانایی واقعی استفاده آموزشی از هوش مصنوعی. مطالعات جدید در زمینه شایستگی هوش مصنوعی و قصد استفاده از هوش مصنوعی مولد نشان می‌دهند که عوامل شناختی، عاطفی و زمینه‌ای به‌طور هم‌زمان در شکل‌گیری رفتار دانشجومعلمان نقش دارند (Acquah et al., 2026; Devaraj, 2026). بنابراین، تفکیک دقیق سازه‌ها و جلوگیری از بیش‌تفسیر نمرات یکی از شروط اساسی روایی است.

علاوه بر این، پژوهش‌های تربیت معلم اغلب از مداخلات آموزشی برای ارتقای شایستگی‌های خاص استفاده می‌کنند و سپس تغییر نمرات را به‌عنوان نشانه اثربخشی مداخله تفسیر می‌کنند. برای نمونه، آموزش مهارت‌های رفتاری می‌تواند بر اجرای راهبردهای آموزشی توسط معلمان اثر بگذارد (Aslan, 2022) و چارچوب‌های طراحی آموزشی می‌توانند به ارتقای دانش فناورانه، تربیتی و محتوایی دانشجومعلمان کمک کنند (Durusoy & Karamete, 2023). در چنین مطالعاتی، اعتبار استنباط تغییر بسیار مهم است؛ زیرا افزایش نمره باید واقعاً بازتاب تغییر در سازه هدف باشد، نه اثر تمرین، تغییر در نحوه پاسخ‌دهی یا ویژگی‌های ابزار. از این‌رو، دقت اندازه‌گیری و حساسیت ابزار به تغییر باید به‌صورت جداگانه از روایی تفسیر نتایج بررسی شوند.

حوزه ارزیابی برای یادگیری نیز مثال دیگری از پیوند میان سنجش، تفسیر و پیامد است. استفاده از ارزیابی اصیل و تشخیصی می‌تواند به بهبود فرایند یادگیری کمک کند، اما ارزش این شیوه‌ها به نحوه تفسیر داده‌های حاصل از ارزیابی و تصمیم‌های بعدی وابسته است (Fadlillah & Kusaeri, 2024). بنابراین، سنجش در تربیت معلم نباید صرفاً به تولید نمره محدود شود، بلکه باید بخشی از یک نظام استدلالی در نظر گرفته شود که از جمع‌آوری شواهد آغاز می‌شود، به تفسیر نتایج می‌رسد و در نهایت به تصمیم یا اقدام آموزشی منتهی می‌شود.

از سوی دیگر، مرور پژوهش‌های ارزیابی معلم نشان داده است که تعدد شاخص‌ها و ابزارها لزوماً به معنای کیفیت بیشتر ارزیابی نیست، مگر آنکه رابطه هر شاخص با سازه و تصمیم مورد نظر مشخص باشد (Wei et al., 2023). همین مسئله در پژوهش‌های تربیت معلم نیز قابل مشاهده است. برخی مطالعات ممکن است مجموعه‌ای از شاخص‌های برازش، ضرایب همبستگی و ضرایب پایایی را ارائه کنند، اما بدون یک چارچوب روشن برای پیوند این شواهد با تفسیر نمرات، یافته‌ها صرفاً مجموعه‌ای از نتایج آماری باقی می‌مانند. پژوهش‌های مربوط به ارزیابی برنامه‌های تربیت معلم نیز تأکید می‌کنند که کیفیت استنباط به طراحی پژوهش و وضوح رابطه میان شواهد و ادعاها بستگی دارد (Mancenido, 2023).

مرور ادبیات همچنین نشان می‌دهد که پژوهش‌های مرتبط با فناوری و نوآوری آموزشی در تربیت معلم به‌سرعت در حال گسترش است، اما سرعت توسعه ابزارهای جدید ممکن است از رشد چارچوب‌های اعتبارسنجی آنها پیشی بگیرد. پژوهش‌های مربوط به واقعیت افزوده در آموزش حرفه‌ای معلمان، فناوری‌های آموزشی دیجیتال، سواد نوشتاری دیجیتال و هوش مصنوعی نمونه‌هایی از حوزه‌هایی هستند که ابزارهای سنجش آنها نسبتاً جدید هستند و هنوز نیازمند انباشت شواهد در بافت‌های گوناگون‌اند (Baltynova et al., 2023; Mena et al., 2023a, 2023b; Nabhan & Habók, 2025). در چنین شرایطی، استفاده از اصطلاح «ابزار معتبر» ممکن است تصویری بیش از حد قطعی ایجاد

کند؛ زیرا روایی باید فرایندی انباشتی و وابسته به شواهد تلقی شود که با تغییر جامعه، هدف یا شیوه استفاده از نمره ممکن است نیازمند ارزیابی مجدد باشد.

افزون بر این، رابطه میان کیفیت سنجش و کیفیت تصمیم‌گیری درباره تربیت معلم را نمی‌توان از زمینه سازمانی جدا کرد. عوامل سازمانی و روان‌شناختی در شکل‌گیری هویت حرفه‌ای مدرسان نقش دارند (Farwa et al., 2024) و کیفیت برنامه‌های تربیت معلم نیز تحت تأثیر شرایط نهادی، ساختار آموزش و قابلیت‌های مدرسان قرار می‌گیرد (Ekabu, 2023; Guo et al., 2023). بنابراین، حتی اگر یک ابزار در سطح روان‌سنجی مطلوب عمل کند، استفاده از نتایج آن برای تصمیم‌گیری سازمانی باید با توجه به زمینه‌ای که نمرات در آن تولید شده‌اند انجام شود. چنین ملاحظه‌ای نشان می‌دهد که روایی ماهیتی موقعیتی دارد و نمی‌توان شواهد حاصل از یک کاربرد را بدون بررسی اضافی به کاربرد دیگر تعمیم داد.

تحول مداوم برنامه‌های تربیت معلم نیز ضرورت بازاندیشی در مفهوم روایی را افزایش داده است. برنامه‌های جدید بر شایستگی‌های قرن بیست‌ویکم، سواد دیجیتال، یادگیری مبتنی بر شایستگی، آموزش فراگیر و توانایی تطبیق با نوآوری‌های فناورانه تأکید دارند (Ajani, 2024; Saleh et al., 2024). هم‌زمان، مطالعاتی که کیفیت فارغ‌التحصیلان مؤسسات تربیت معلم را بررسی کرده‌اند، نشان می‌دهند که ارزیابی پیامدهای تربیت معلم نیازمند نگاه چندبعدی است (Olvido et al., 2024). بنابراین، الگوی اعتبارسنجی نیز باید از مدل‌های محدود مبتنی بر یک ضریب یا یک تحلیل آماری به سمت چارچوبی حرکت کند که در آن دقت اندازه‌گیری، شواهد تفسیری و پیامدهای کاربرد نتایج به‌صورت پیوسته و مرتبط با یکدیگر بررسی شوند.

در مجموع، ادبیات موجود تصویری دوگانه ارائه می‌کند. از یک سو، توسعه ابزارهای تخصصی، استفاده از روش‌های آماری پیشرفته و افزایش توجه به کیفیت اندازه‌گیری در پژوهش‌های تربیت معلم نشانه بلوغ روش‌شناختی این حوزه است (Erviana et al., 2023; García et al., 2025; Nabhan & Habók, 2025). از سوی دیگر، تمرکز مفرط بر شاخص‌های روان‌سنجی بدون بررسی کامل استدلال تفسیری و پیامدهای استفاده از نتایج می‌تواند به نوعی «تقلیل روایی به آزمون آماری» منجر شود. مطالعات مربوط به ارزشیابی معلمان، برنامه‌های آماده‌سازی حرفه‌ای، فناوری، آموزش فراگیر، هویت حرفه‌ای و شایستگی‌های جدید همگی نشان می‌دهند که نمرات فقط زمانی دارای ارزش علمی هستند که رابطه میان نحوه تولید آنها، معنایی که به آنها نسبت داده می‌شود و تصمیمی که بر مبنای آنها اتخاذ می‌شود به‌طور شفاف و مستند مشخص باشد (Mancenido, 2023; Tortorelli et al., 2021; Wei et al., 2023).

بر همین اساس، روایی در پژوهش‌های تربیت معلم باید نه به‌عنوان مهر تأییدی برای یک ابزار، بلکه به‌عنوان یک استدلال مستمر و انباشتی درباره کفایت شواهد برای پشتیبانی از تفسیر و کاربرد نمرات بازفهم شود؛ استدلالی که از دقت اندازه‌گیری آغاز می‌شود، از بررسی معنای نمرات و روابط آنها با شواهد نظری و تجربی عبور می‌کند و در نهایت پیامدهای استفاده از نتایج را در نظر می‌گیرد. چنین بازتعریفی می‌تواند فاصله میان روان‌سنجی فنی و تصمیم‌گیری آموزشی را کاهش دهد و زمینه را برای استفاده مسئولانه‌تر از سنجش در طراحی، ارزیابی و اصلاح برنامه‌های تربیت معلم فراهم کند.

هدف پژوهش حاضر، بازمفهوم‌سازی روایی در پژوهش‌های تربیت معلم از طریق بررسی نظام‌مند دقت اندازه‌گیری، شواهد تفسیری و شواهد پیامدی و تبیین نحوه ارتباط این سه سطح در شکل‌گیری یک استدلال روایی منسجم درباره تفسیر و کاربرد نتایج سنجش است.

روش‌شناسی

پژوهش حاضر با رویکرد مروری تحلیلی و با هدف بازمفهوم‌سازی روایی در پژوهش‌های تربیت معلم، با تأکید بر سه حوزه به‌هم‌پیوسته دقت اندازه‌گیری، شواهد تفسیری و شواهد پیامدی انجام شد. منطق اصلی طراحی پژوهش بر این پیش‌فرض استوار بود که روایی صرفاً یک ویژگی ثابت و فنی ابزار اندازه‌گیری نیست، بلکه کیفیت استنباط‌ها، تفسیرها، تصمیم‌ها و پیامدهایی را نیز دربرمی‌گیرد که بر پایه نمرات و سایر

داده‌های حاصل از سنجش در پژوهش‌های تربیت معلم شکل می‌گیرند. جامعه اسنادی پژوهش شامل مقالات علمی پژوهشی منتشرشده درباره روایی، اعتبار استنباط‌های مبتنی بر سنجش، کیفیت اندازه‌گیری و استفاده از ابزارهای ارزیابی در حوزه تربیت معلم بود. جست‌وجوی منابع با تمرکز بر مطالعات منتشرشده در فاصله سال‌های ۲۰۱۰ تا ۲۰۲۵ انجام گرفت تا ضمن پوشش تحولات نظری معاصر در مفهوم روایی، شواهد تجربی نسبتاً جدید درباره کاربرد رویکردهای روایی در پژوهش‌های تربیت معلم نیز بررسی شود. در مرحله جست‌وجوی اولیه، ۲۸۶ رکورد از پایگاه‌های اطلاعاتی شناسایی شد. پس از حذف ۷۴ رکورد تکراری، ۲۱۲ منبع برای غربالگری عنوان و چکیده باقی ماند. در این مرحله، ۱۲۱ مطالعه به دلیل عدم ارتباط مستقیم با تربیت معلم، تمرکز انحصاری بر پایایی بدون بررسی روایی، فقدان داده یا استدلال مرتبط با تفسیر نمرات، یا ماهیت صرفاً توصیفی و غیرتحلیلی کنار گذاشته شد و متن کامل ۹۱ مقاله مورد بررسی قرار گرفت. پس از ارزیابی متن کامل، ۴۷ مطالعه دیگر به دلیل فقدان اطلاعات کافی درباره شواهد روایی، نامشخص بودن مبنای تفسیر نتایج، تمرکز بر زمینه‌هایی خارج از آموزش و تربیت معلم، یا عدم امکان استخراج داده‌های مورد نیاز حذف شد. در نهایت، ۴۴ مطالعه واجد معیارهای ورود شناخته شدند و نمونه‌هایی مرور را تشکیل دادند. معیارهای ورود شامل انتشار در مجلات علمی داوری‌شده، تمرکز مستقیم یا معنادار بر دانشجومعلم، معلمان در دوره‌های آماده‌سازی حرفه‌ای، مدرسان تربیت معلم یا برنامه‌های آموزش معلم، استفاده از یک یا چند ابزار سنجش یا ارزیابی، و ارائه شواهد قابل استخراج درباره کیفیت اندازه‌گیری، تفسیر نتایج یا پیامدهای استفاده از نتایج بود. مطالعاتی که صرفاً به گزارش ضریب پایایی اکتفا کرده بودند، بدون آنکه شواهدی برای توجیه تفسیر یا کاربرد نمرات ارائه کنند، تنها در صورتی وارد مرور شدند که داده‌های دیگری درباره ساختار ابزار، ارتباط با متغیرهای بیرونی یا شیوه استفاده از نمرات نیز در آنها وجود داشت. واحد تحلیل در این پژوهش نه صرفاً ابزار اندازه‌گیری، بلکه «ادعای روایی» مطرح‌شده در هر مطالعه بود؛ بدین معنا که برای هر پژوهش بررسی شد نویسندگان بر اساس چه نوع شواهدی، چه تفسیری از نمرات یا داده‌ها ارائه کرده‌اند و این تفسیر برای چه نوع کاربرد یا تصمیمی مورد استفاده قرار گرفته است.

گردآوری داده‌ها با استفاده از فرم ساختاریافته استخراج و ارزیابی شواهد روایی انجام شد که متناسب با هدف پژوهش و بر پایه رویکردهای معاصر به اعتبارسنجی طراحی گردید. این فرم به‌گونه‌ای تنظیم شد که علاوه بر ویژگی‌های کتاب‌شناختی و روش‌شناختی هر مطالعه، ماهیت ادعای روایی و شواهد پشتیبان آن را نیز ثبت کند. در بخش مشخصات عمومی، اطلاعاتی از قبیل سال انتشار، کشور یا بافت آموزشی، نوع نمونه، حجم نمونه، نوع برنامه تربیت معلم، طرح پژوهش، نوع ابزار سنجش، تعداد گویه‌ها یا مؤلفه‌های ابزار و نوع نمره یا شاخص مورد استفاده استخراج شد. در بخش مربوط به دقت اندازه‌گیری، شواهد مرتبط با ثبات و دقت نمرات، از جمله ضرایب پایایی، خطای استاندارد اندازه‌گیری، ضرایب همبستگی درون‌طبقه‌ای، شاخص‌های توافق، اطلاعات مربوط به عملکرد گویه‌ها، تحلیل‌های مبتنی بر نظریه کلاسیک آزمون یا نظریه پاسخ به سؤال و، در صورت گزارش، شاخص‌های اطلاعات آزمون ثبت شد. با این حال، در چارچوب مفهومی این پژوهش، پایایی و دقت اندازه‌گیری به‌تنهایی معادل روایی در نظر گرفته نشدند، بلکه به‌عنوان یکی از پیش‌شرط‌های لازم برای حمایت از استنباط‌های معتبر از نمرات تلقی شدند. بخش دوم فرم به شواهد تفسیری اختصاص داشت و شامل اطلاعات مربوط به روایی محتوایی، فرایندهای پاسخ، ساختار درونی، روابط با متغیرهای بیرونی، روایی همگرا و واگرا، روایی ملاکی، تغییرناپذیری اندازه‌گیری میان گروه‌ها و میزان انطباق تفسیر ارائه‌شده با شواهد تجربی بود. همچنین، مشخص شد که آیا مطالعه مورد نظر میان ویژگی‌های روان‌سنجی ابزار و استنباطی که از نمرات حاصل می‌شود تمایز قائل شده است یا روایی را به‌صورت ویژگی ذاتی ابزار گزارش کرده است. بخش سوم فرم برای استخراج شواهد پیامدی طراحی شد و در آن پیامدهای پیش‌بینی‌شده و پیش‌بینی‌نشده استفاده از نمرات، تصمیم‌های آموزشی یا حرفه‌ای مبتنی بر ابزار، آثار استفاده از نتایج بر دانشجومعلم یا معلمان، احتمال طبقه‌بندی نادرست، تفاوت آثار سنجش میان گروه‌های مختلف، مسائل مربوط به عدالت و انصاف و شواهد مرتبط با استفاده مطلوب یا نامطلوب از نتایج ثبت شد. برای افزایش انسجام فرایند استخراج داده‌ها، تعاریف عملیاتی متغیرها پیش از کدگذاری نهایی تدوین شد و فرم اولیه بر روی شش مطالعه خارج از نمونه اصلی به‌صورت آزمایشی اجرا گردید. بر اساس نتایج اجرای آزمایشی، برخی مقوله‌ها ادغام یا تفکیک شدند و دستورالعمل کدگذاری برای مواردی که مطالعات از اصطلاحات متفاوت برای مفاهیم مشابه استفاده کرده

بودند، استانداردسازی شد. افزون بر این، برای هر مقاله یک ماتریس استدلال روایی تهیه شد که در آن رابطه میان نمره یا مشاهده، استنباط مورد نظر، شواهد پشتیبان و کاربرد نهایی نتایج به صورت زنجیره‌ای ثبت شد. این شیوه امکان داد تا ارزیابی روایی از سطح گزارش چند ضریب آماری فراتر رود و سازگاری کل استدلالی که پژوهشگر برای استفاده و تفسیر داده‌ها ارائه کرده است بررسی شود.

تحلیل داده‌ها با تلفیق تحلیل توصیفی، تحلیل محتوای جهت‌دار و سنتز تفسیری انجام شد. در مرحله نخست، ویژگی‌های مطالعات منتخب بر اساس سال انتشار، نوع نمونه، طرح پژوهش، نوع ابزار اندازه‌گیری، رویکرد روان‌سنجی و انواع شواهد روایی توصیف و فراوانی و درصد مطالعاتی که هر یک از انواع شواهد را گزارش کرده بودند محاسبه شد. در مرحله دوم، هر یک از مطالعات در سه حوزه اصلی «دقت اندازه‌گیری»، «شواهد تفسیری» و «شواهد پیامدی» کدگذاری شدند. دقت اندازه‌گیری بر اساس کفایت شواهد مربوط به ثبات، خطای اندازه‌گیری و حساسیت نمرات مورد ارزیابی قرار گرفت؛ شواهد تفسیری بر اساس میزان پشتیبانی داده‌ها از معنای نسبت داده‌شده به نمرات و سازگاری استنباط‌های پژوهش با شواهد محتوایی، ساختاری، فرایندی و بیرونی تحلیل شد؛ و شواهد پیامدی بر اساس میزان توجه مطالعات به آثار ناشی از کاربرد نتایج، عدالت در استفاده، خطاهای تصمیم‌گیری و پیامدهای احتمالی برای افراد یا برنامه‌های تربیت معلم بررسی گردید. در مرحله سوم، میان «گزارش شواهد روان‌سنجی» و «ارائه استدلال روایی» تمایز تحلیلی برقرار شد. مطالعه‌ای ممکن بود شاخص‌های مطلوبی مانند آلفای کرونباخ، پایایی ترکیبی، تحلیل عاملی تأییدی یا همبستگی با ملاک‌های بیرونی گزارش کرده باشد، اما اگر ارتباط منطقی میان این شواهد و تفسیر یا کاربرد نهایی نمرات مشخص نشده بود، این وضعیت به عنوان استدلال روایی ناقص کدگذاری شد. در مقابل، مطالعاتی که به طور روشن مشخص کرده بودند هر نوع شواهد چگونه از یک استنباط معین و کاربرد مشخص حمایت می‌کند، دارای انسجام تفسیری بالاتری در نظر گرفته شدند. برای بررسی الگوهای غالب، کدهای استخراج‌شده از مطالعات به مقوله‌های سطح بالاتر سازمان‌دهی و شباهت‌ها، تفاوت‌ها و خلأهای موجود در استدلال‌های روایی شناسایی شدند. توجه ویژه‌ای به مواردی شد که در آنها روایی به اشتباه مترادف با پایایی، برازش مدل یا تأیید ساختار عاملی تلقی شده بود، زیرا یکی از اهداف اصلی پژوهش، آشکار ساختن فاصله میان سنجش فنی کیفیت ابزار و توجیه استنباط و استفاده از نمرات بود. در مرحله نهایی، سنتز تفسیری یافته‌ها با استفاده از یک چارچوب سه‌لایه انجام گرفت که در آن دقت اندازه‌گیری به عنوان زیرساخت فنی، شواهد تفسیری به عنوان مبنای توجیه معنای نمرات و شواهد پیامدی به عنوان مبنای ارزیابی مشروعیت و کفایت کاربرد نمرات در نظر گرفته شد. بر این اساس، روایی نه به صورت یک ضریب واحد و نه به عنوان ویژگی تغییرناپذیر ابزار، بلکه به عنوان استدلالی انباشتی و وابسته به زمینه درباره مناسب بودن تفسیر و کاربرد داده‌های سنجش مفهوم‌پردازی شد. برای افزایش اعتمادپذیری تحلیل، بخشی از مطالعات به صورت مستقل دوباره کدگذاری شد و موارد اختلاف از طریق بازبینی تعاریف عملیاتی و مقایسه مستقیم شواهد گزارش‌شده در متن مقالات حل شد. همچنین، در مرحله سنتز نهایی، موارد منفی و مطالعاتی که با الگوی غالب همخوانی نداشتند به طور جداگانه بررسی شدند تا چارچوب نهایی صرفاً بر شواهد تأییدکننده مبتنی نباشد و تنوع رویکردهای اعتبارسنجی در پژوهش‌های تربیت معلم را نیز منعکس کند.

یافته‌ها

پس از اجرای فرایند جست‌وجو، غربالگری و ارزیابی متن کامل مطالعات، در مجموع ۴۴ مطالعه واجد معیارهای ورود به مرحله تحلیل نهایی شدند. بررسی اولیه مطالعات نشان داد که ادبیات مربوط به روایی در پژوهش‌های تربیت معلم از نظر دوره زمانی انتشار، موقعیت جغرافیایی، جامعه مورد مطالعه، طرح پژوهش و نوع ابزارهای اندازه‌گیری تنوع قابل توجهی دارد. با این حال، توزیع مطالعات یکنواخت نبود و بخش عمده پژوهش‌ها در سال‌های اخیر منتشر شده بودند که می‌تواند نشان‌دهنده افزایش توجه به کیفیت اندازه‌گیری، ساختار ابزارها و اعتبار استنباط‌های حاصل از داده‌های سنجش در حوزه تربیت معلم باشد. همچنین، بیشتر مطالعات بر دانشجومعلم‌ان متمرکز بودند و ابزارهای خودگزارشی، پرسشنامه‌ها و مقیاس‌های نگرشی یا ادراکی سهم عمده‌ای از ابزارهای مورد استفاده را تشکیل می‌دادند. در مقابل، مطالعاتی که از ارزیابی‌های عملکردی، مشاهده مستقیم، روبریک‌های حرفه‌ای، پورتفولیو یا سنجش‌های چندمنبعی استفاده کرده بودند، سهم کمتری داشتند. جدول ۱

تصویری کلی از ویژگی‌های اصلی مطالعات واردشده به مرور ارائه می‌کند و مبنایی برای تفسیر الگوهای بعدی مربوط به دقت اندازه‌گیری، شواهد تفسیری و پیامدهای استفاده از نتایج فراهم می‌آورد.

جدول ۱. ویژگی‌های کلی مطالعات واردشده به مرور ($n = ۴۴$)

متغیر	طبقه	فراوانی	درصد
دوره انتشار	۲۰۱۰ تا ۲۰۱۴	۶	۱۳.۶
	۲۰۱۵ تا ۲۰۱۹	۱۴	۳۱.۸
	۲۰۲۰ تا ۲۰۲۵	۲۴	۵۴.۵
منطقه جغرافیایی	آمریکای شمالی	۱۵	۳۴.۱
	اروپا	۱۰	۲۲.۷
	آسیا	۹	۲۰.۵
	اقیانوسیه	۵	۱۱.۴
	خاورمیانه	۳	۶.۸
	چندکشوری	۲	۴.۵
	دانشجومعلم	۲۵	۵۶.۸
	معلمان در دوره‌های آموزش و توسعه حرفه‌ای	۸	۱۸.۲
	مدرسان و اعضای هیئت علمی تربیت معلم	۴	۹.۱
جامعه اصلی مطالعه	نمونه‌های ترکیبی	۷	۱۵.۹
	مقطعی	۲۲	۵۰.۰
	اعتباریابی یا توسعه ابزار	۱۱	۲۵.۰
	طولی	۵	۱۱.۴
	آمیخته	۴	۹.۱
	آزمایشی یا شبه‌آزمایشی	۲	۴.۵
	پرسشنامه یا مقیاس خودگزارشی	۲۱	۴۷.۷
	ارزیابی عملکردی یا مشاهده‌ای	۹	۲۰.۵
	آزمون دانش یا شایستگی	۷	۱۵.۹
	روبریک، پورتفولیو یا ارزیابی مبتنی بر محصول	۴	۹.۱
نوع غالب ابزار	ترکیبی یا چندمنبعی	۳	۶.۸

نتایج جدول ۱ نشان می‌دهد که بیش از نیمی از مطالعات بررسی‌شده، یعنی ۲۴ مطالعه معادل ۵۴.۵ درصد، در فاصله سال‌های ۲۰۲۰ تا ۲۰۲۵ منتشر شده‌اند، در حالی که تنها ۶ مطالعه معادل ۱۳.۶ درصد به دوره ۲۰۱۰ تا ۲۰۱۴ تعلق داشتند. این الگو بیانگر افزایش محسوس توجه به مسائل سنجش و روایی در سال‌های اخیر است. از نظر پراکندگی جغرافیایی، آمریکای شمالی با ۱۵ مطالعه و ۳۴.۱ درصد بیشترین سهم را داشت و پس از آن اروپا با ۲۲.۷ درصد و آسیا با ۲۰.۵ درصد قرار گرفتند. سهم مطالعات انجام‌شده در خاورمیانه تنها ۶.۸ درصد بود و ۴.۵ درصد نیز ماهیت چندکشوری داشتند. از نظر جامعه پژوهش، ۲۵ مطالعه، یعنی ۵۶.۸ درصد کل نمونه، مستقیماً بر دانشجومعلمان متمرکز بودند که نشان می‌دهد مفهوم روایی در ادبیات تربیت معلم عمده‌تأ در ارتباط با سنجش نگرش‌ها، باورها، دانش حرفه‌ای، خودکارآمدی، آمادگی تدریس و شایستگی‌های دانشجومعلمان بررسی شده است. نیمی از مطالعات دارای طرح مقطعی بودند و فقط ۵ مطالعه، معادل ۱۱.۴ درصد، داده‌های طولی را دربرمی‌گرفتند. این یافته از آن جهت اهمیت دارد که بسیاری از سازه‌های حرفه‌ای در تربیت معلم ماهیتی تحولی دارند و اتکا به طرح‌های مقطعی می‌تواند دامنه استنباط درباره ثبات، تغییر و حساسیت ابزارهای سنجش را محدود کند. همچنین، ۴۷.۷ درصد مطالعات از پرسشنامه یا مقیاس‌های خودگزارشی به عنوان ابزار اصلی استفاده کرده بودند، در حالی که فقط ۲۰.۵ درصد بر ارزیابی‌های عملکردی

یا مشاهده‌ای تکیه داشتند. بنابراین، شواهد موجود درباره روایی در این حوزه عمدتاً بر نمرات حاصل از خودگزارشی استوار بود و ارزیابی مستقیم عملکرد حرفه‌ای معلم در مقایسه با سنجش ادراکات و نگرش‌ها حضور محدودتری در ادبیات مورد بررسی داشت.

نخستین محور تحلیل، به دقت اندازه‌گیری اختصاص داشت. در این بخش بررسی شد که مطالعات تا چه اندازه به مسئله ثبات نمرات، خطای اندازه‌گیری، کیفیت گویه‌ها، توافق میان ارزیابان و میزان اطلاعات فراهم‌شده توسط ابزار توجه کرده‌اند. از آنجا که یک ابزار حتی در صورت برخورداری از ساختار عاملی مناسب نمی‌تواند بدون سطح قابل قبول دقت اندازه‌گیری مبنای استنباط معتبر قرار گیرد، شاخص‌های مرتبط با پایایی و خطا به عنوان زیرساخت فنی استدلال روایی ارزیابی شدند. در عین حال، گزارش یک ضریب پایایی به تنهایی به عنوان شواهد کافی برای روایی در نظر گرفته نشد. جدول ۲ فراوانی انواع شواهد مربوط به دقت اندازه‌گیری را در ۴۴ مطالعه نشان می‌دهد. به دلیل آنکه هر مطالعه می‌توانست بیش از یک نوع شاخص را گزارش کند، مجموع درصدهای جدول از ۱۰۰ درصد بیشتر است.

جدول ۲. شواهد مربوط به دقت اندازه‌گیری در مطالعات بررسی‌شده (n = ۴۴)

نوع شواهد دقت اندازه‌گیری	فراوانی مطالعات	درصد مطالعات
آلفای کرونباخ یا شاخص‌های مشابه همسانی درونی	۳۵	۷۹.۵
تحلیل گویه، همبستگی گویه-کل یا شاخص‌های تمایز گویه	۱۸	۴۰.۹
پایایی ترکیبی	۱۲	۲۷.۳
ضریب امگا	۱۰	۲۲.۷
توافق بین ارزیابان یا ضریب همبستگی درون طبقه‌ای	۹	۲۰.۵
پایایی بازآزمایی	۷	۱۵.۹
بررسی اثر سقف، اثر کف یا دامنه توزیع نمرات	۶	۱۳.۶
خطای استاندارد اندازه‌گیری یا شاخص‌های مستقیم خطای نمره	۵	۱۱.۴
شاخص‌های نظریه پاسخ به سؤال یا تابع اطلاعات آزمون	۴	۹.۱
بررسی نظام‌مند دقت نمرات در زیرگروه‌های مختلف	۳	۶.۸

بررسی جدول ۲ نشان می‌دهد که رایج‌ترین رویکرد به دقت اندازه‌گیری، استفاده از آلفای کرونباخ یا شاخص‌های مشابه همسانی درونی بوده است؛ به طوری که ۳۵ مطالعه، معادل ۷۹.۵ درصد، حداقل یک ضریب از این نوع گزارش کرده بودند. با وجود فراوانی بالای این شاخص، تنها ۱۰ مطالعه از ضریب امگا و ۱۲ مطالعه از پایایی ترکیبی استفاده کرده بودند. این تفاوت نشان می‌دهد که در بخش بزرگی از ادبیات تربیت معلم همچنان آلفای کرونباخ به صورت شاخص اصلی یا حتی تنها شاخص پایایی به کار رفته است. تحلیل گویه‌ها در ۱۸ مطالعه، معادل ۴۰.۹ درصد، مشاهده شد که نشان‌دهنده توجه متوسط به عملکرد اجزای ابزار است. با این حال، تنها ۷ مطالعه پایایی بازآزمایی را گزارش کرده بودند و این امر نشان می‌دهد که ثبات زمانی نمرات در بیشتر پژوهش‌ها به طور مستقیم ارزیابی نشده است. این کاستی به ویژه در سنجش سازه‌هایی مانند نگرش به تدریس، خودکارآمدی حرفه‌ای، باورهای تربیتی و آمادگی برای تدریس اهمیت دارد؛ زیرا بدون داده‌های زمانی نمی‌توان به روشنی مشخص کرد که تغییرات مشاهده‌شده بازتاب تحول واقعی سازه هستند یا بخشی از آنها ناشی از خطای اندازه‌گیری است.

یکی از مهم‌ترین یافته‌های این بخش، محدود بودن استفاده از شاخص‌هایی بود که خطای اندازه‌گیری را به طور مستقیم کمی‌سازی می‌کنند. فقط ۵ مطالعه، معادل ۱۱.۴ درصد، خطای استاندارد اندازه‌گیری یا شاخص‌های مشابه را گزارش کرده بودند و فقط ۴ مطالعه از روش‌های مبتنی بر نظریه پاسخ به سؤال یا تابع اطلاعات آزمون بهره گرفته بودند. در نتیجه، هرچند بیشتر پژوهش‌ها درباره «پایایی» ابزار خود اظهار نظر کرده بودند، تعداد بسیار کمتری مشخص کرده بودند که نمرات در کدام دامنه از سازه با دقت بیشتری اندازه‌گیری می‌شوند یا چه مقدار عدم قطعیت در نمره هر فرد وجود دارد. همین مسئله در ارزیابی‌های عملکردی نیز مشاهده شد. ۹ مطالعه از ضریب توافق میان ارزیابان یا همبستگی درون طبقه‌ای استفاده کرده بودند، اما این تعداد نسبت به کل مطالعاتی که از مشاهده، روبریک یا ارزیابی عملکرد حرفه‌ای بهره می‌گرفتند همچنان محدود بود. همچنین، فقط ۳ مطالعه دقت اندازه‌گیری را به تفکیک زیرگروه‌های جمعیتی، آموزشی یا فرهنگی بررسی

کرده بودند. بنابراین، تصویر کلی حاصل از داده‌ها نشان داد که دقت اندازه‌گیری در ادبیات تربیت معلم عمدتاً به معنای «همسانی درونی» مفهوم‌سازی شده است، در حالی که مؤلفه‌های مهم‌تر و کاربردی‌تر آن، از جمله خطای نمره، ثبات زمانی، اطلاعات آزمون، حساسیت به سطوح مختلف توانایی و برابری دقت اندازه‌گیری میان گروه‌ها، کمتر مورد توجه قرار گرفته‌اند.

دومین محور تحلیل به شواهدی اختصاص داشت که مستقیماً از معنای نسبت داده‌شده به نمرات حمایت می‌کنند. در این بخش صرفاً بررسی نشد که مطالعات چه آزمون‌های آماری انجام داده‌اند، بلکه ارزیابی شد که آیا شواهد گزارش‌شده واقعاً با استنباط مورد نظر درباره سازه سنجیده‌شده مرتبط بوده‌اند یا خیر. بر این اساس، شواهد محتوایی، فرایندهای پاسخ، ساختار درونی، روابط با متغیرهای بیرونی، روایی همگرا و واگرا، روایی ملاکی، تغییرناپذیری اندازه‌گیری و ارتباط صریح میان شواهد و تفسیر نمرات مورد بررسی قرار گرفت. جدول ۳ توزیع این شواهد را نشان می‌دهد. همانند جدول قبل، هر مطالعه می‌توانست هم‌زمان چند نوع شواهد تفسیری ارائه کند و بنابراین درصدها مستقل از یکدیگر محاسبه شده‌اند.

جدول ۳. انواع شواهد تفسیری گزارش‌شده در مطالعات بررسی‌شده ($n = 44$)

نوع شواهد تفسیری	فراوانی مطالعات	درصد مطالعات
شواهد ساختار درونی، شامل تحلیل عاملی اکتشافی یا تأییدی	۳۶	۸۱.۸
رابطه با متغیرهای بیرونی	۲۵	۵۶.۸
شواهد محتوایی و قضاوت خبرگان	۲۳	۵۲.۳
روایی همگرا	۲۱	۴۷.۷
روایی واگرا یا تمایزی	۱۷	۳۸.۶
ارتباط با ملاک یا پیامد بیرونی مشخص	۱۱	۲۵.۰
تغییرناپذیری اندازه‌گیری بین گروه‌ها	۹	۲۰.۵
شواهد مربوط به فرایندهای پاسخ	۸	۱۸.۲
بررسی تفاوت عملکرد گویه‌ها یا سوگیری اندازه‌گیری	۶	۱۳.۶
پیوند صریح میان شواهد گردآوری‌شده و تفسیر مورد نظر از نمرات	۱۴	۳۱.۸

یافته‌های جدول ۳ نشان می‌دهد که تحلیل ساختار درونی، با فاصله زیاد، رایج‌ترین نوع شواهد تفسیری در مطالعات تربیت معلم بوده است. در مجموع، ۳۶ مطالعه معادل ۸۱.۸ درصد از تحلیل عاملی اکتشافی، تحلیل عاملی تأییدی یا روش‌های مشابه برای بررسی ساختار ابزار استفاده کرده بودند. این فراوانی بالا نشان می‌دهد که در ادبیات مورد بررسی، روایی غالباً با تأیید ساختار عاملی ابزار پیوند خورده است. با این حال، تنها ۱۴ مطالعه، یعنی ۳۱.۸ درصد، به‌صورت روشن توضیح داده بودند که چگونه ساختار مشاهده‌شده از تفسیر خاصی از نمرات حمایت می‌کند. بنابراین، در بخش مهمی از مطالعات، مدل عاملی مناسب به‌عنوان نتیجه نهایی اعتبارسنجی ارائه شده بود، بی‌آنکه رابطه میان برازش ساختاری و ادعای تفسیری درباره سازه به‌طور کامل توضیح داده شود. این تمایز یکی از یافته‌های محوری پژوهش حاضر بود، زیرا نشان داد وجود شاخص‌های برازش مطلوب لزوماً به معنای کامل بودن استدلال روایی نیست.

شواهد مربوط به روابط با متغیرهای بیرونی در ۲۵ مطالعه، معادل ۵۶.۸ درصد، مشاهده شد. با این حال، ماهیت این روابط متفاوت بود. در برخی مطالعات، ارتباط میان نمرات ابزار و متغیرهایی مانند خودکارآمدی، دانش حرفه‌ای، کیفیت تدریس، نگرش به یادگیری یا عملکرد تحصیلی از پیش بر مبنای نظری مشخص شده بود و الگوی مشاهده‌شده برای حمایت از تفسیر نمرات مورد استفاده قرار می‌گرفت. در مقابل، در شماری از مطالعات صرفاً همبستگی‌های معنادار آماری گزارش شده بود، بدون آنکه جهت، شدت یا مبنای نظری مورد انتظار رابطه از پیش مشخص شده باشد. شواهد محتوایی در ۲۳ مطالعه گزارش شده بود، اما در بسیاری از این موارد، بررسی محتوا به تأیید گویه‌ها توسط تعدادی متخصص محدود شده بود و اطلاعات کاملی درباره فرایند انتخاب خبرگان، معیارهای داوری یا نحوه اصلاح گویه‌ها ارائه نشده بود.

روایی همگرا در ۲۱ مطالعه و روایی واگرا در ۱۷ مطالعه بررسی شده بود. با وجود این، یافته‌ها نشان داد که روایی همگرا بیشتر از تمایز تجربی میان سازه‌ها مورد توجه قرار گرفته است. این مسئله می‌تواند در پژوهش‌های تربیت معلم اهمیت ویژه‌ای داشته باشد، زیرا بسیاری از سازه‌های مورد سنجش، مانند خودکارآمدی، نگرش حرفه‌ای، تعهد، هویت معلمی و آمادگی تدریس از نظر مفهومی نزدیک به یکدیگرند و فقدان شواهد کافی برای روایی واگرا می‌تواند موجب تداخل در تفسیر نمرات شود. فقط ۹ مطالعه، معادل ۲۰.۵ درصد، تغییرناپذیری اندازه‌گیری را میان گروه‌های مختلف بررسی کرده بودند و تنها ۶ مطالعه شواهدی درباره عملکرد افتراقی گویه‌ها یا احتمال سوگیری اندازه‌گیری ارائه داده بودند. بنابراین، در بخش بزرگی از مطالعات مشخص نبود که آیا ابزار برای مقایسه دانشجومعلم در گروه‌های جنسیتی، فرهنگی، زبانی، رشته‌ای یا مراحل مختلف آموزش معلم به صورت هم‌ارز عمل می‌کند یا خیر.

یکی دیگر از کاستی‌های برجسته، سهم پایین شواهد مربوط به فرایندهای پاسخ بود. تنها ۸ مطالعه، معادل ۱۸.۲ درصد، بررسی کرده بودند که پاسخ‌دهندگان چگونه گویه‌ها، موقعیت‌های آزمون یا تکالیف سنجشی را درک می‌کنند و آیا فرایند ذهنی آنان با سازه‌ای که ابزار قصد اندازه‌گیری آن را دارد سازگار است یا خیر. این یافته نشان می‌دهد که ادبیات اعتبارسنجی در تربیت معلم عمدتاً بر ساختار آماری داده‌ها متمرکز بوده و سازوکار شناختی یا تفسیری تولید پاسخ‌ها کمتر مورد توجه قرار گرفته است. در مجموع، الگوی حاصل نشان داد که شواهد تفسیری در این حوزه فراوان‌تر از شواهد مربوط به پیامدها هستند، اما حتی در این سطح نیز غالب مطالعات به مجموعه محدودی از شاخص‌های روان‌سنجی تکیه می‌کنند و تنها حدود یک‌سوم آنها استدلال صریح و منسجمی میان شواهد، معنای نمرات و نوع کاربرد مورد نظر برقرار کرده‌اند.

سومین محور تحلیل بر شواهد پیامدی استفاده از نمرات و نتایج سنجش تمرکز داشت. هدف از این تحلیل مشخص کردن آن بود که مطالعات تربیت معلم تا چه اندازه فراتر از کیفیت آماری ابزار، به این پرسش پرداخته‌اند که استفاده از نتایج سنجش چه آثار آموزشی، حرفه‌ای، تصمیمی یا توزیعی برای افراد و برنامه‌ها ایجاد می‌کند. در این بخش، پیامدهای مورد انتظار و ناخواسته، عدالت در تصمیم‌گیری، احتمال طبقه‌بندی نادرست، اثرات استفاده از نمرات بر دانشجومعلم یا معلمان، آثار مربوط به تصمیم‌های برنامه‌ای و ملاحظات اخلاقی مورد بررسی قرار گرفت. جدول ۴ توزیع شواهد پیامدی را در مطالعات نشان می‌دهد.

جدول ۴. شواهد مربوط به پیامدهای تفسیر و استفاده از نتایج سنجش در مطالعات بررسی‌شده (n = ۴۴)

نوع شواهد پیامدی	فراوانی مطالعات	درصد مطالعات
توجه به پیامدهای مورد انتظار استفاده از نمرات	۱۲	۲۷.۳
بررسی عدالت یا تفاوت پیامدها میان زیرگروه‌ها	۱۰	۲۲.۷
بحث درباره ملاحظات اخلاقی یا محدودیت‌های کاربرد نمرات	۸	۱۸.۲
بررسی تأثیر نمرات بر تصمیم‌های آموزشی یا حرفه‌ای	۷	۱۵.۹
بررسی آثار سنجش بر دانشجومعلم، معلمان یا ذی‌نفعان	۶	۱۳.۶
بررسی پیامدهای ناخواسته استفاده از نتایج	۵	۱۱.۴
تحلیل احتمال طبقه‌بندی یا تصمیم‌گیری نادرست	۴	۹.۱
ارزیابی تجربی هزینه‌ها و منافع استفاده از نتایج	۲	۴.۵
فقدان هرگونه شواهد پیامدی مستقیم	۲۹	۶۵.۹

جدول ۴ آشکار می‌سازد که شواهد پیامدی ضعیف‌ترین بخش استدلال روایی در پژوهش‌های بررسی‌شده بوده است. از ۴۴ مطالعه، ۲۹ مطالعه، معادل ۶۵.۹ درصد، هیچ‌گونه شواهد تجربی مستقیمی درباره پیامدهای استفاده از نمرات ارائه نکرده بودند. در بسیاری از این مطالعات، پس از گزارش ساختار عاملی مناسب و ضرایب پایایی قابل قبول، پیشنهادهایی برای استفاده از ابزار در انتخاب، ارزیابی، تشخیص نیازهای آموزشی یا بررسی شایستگی‌های حرفه‌ای ارائه شده بود، اما پیامدهای چنین کاربردهایی مورد آزمون قرار نگرفته بود. فقط ۱۲ مطالعه، معادل ۲۷.۳ درصد،

به صورت مشخص به پیامدهای مورد انتظار استفاده از نتایج پرداخته بودند. حتی در این گروه نیز بخش عمده شواهد به سطح توصیفی محدود می شد و تنها تعداد معدودی از مطالعات بررسی کرده بودند که استفاده واقعی از نمرات چگونه بر تصمیم های آموزشی یا حرفه ای اثر می گذارد. عدالت در استفاده از نتایج در ۱۰ مطالعه، معادل ۲۲٫۷ درصد، مورد توجه قرار گرفته بود. این مطالعات عمدتاً تفاوت آثار یا عملکرد ابزار را در گروه های جنسیتی، فرهنگی، زبانی یا مراحل مختلف تربیت معلم بررسی کرده بودند. با این حال، فقط ۵ مطالعه به پیامدهای ناخواسته اشاره داشتند و تنها ۴ مطالعه احتمال طبقه بندی نادرست افراد را تحلیل کرده بودند. این موضوع به ویژه در مواردی اهمیت داشت که ابزار برای تصمیم های سطح بالاتر، از جمله تشخیص آمادگی حرفه ای، شناسایی نیاز به مداخله آموزشی، تأیید صلاحیت یا طبقه بندی سطح شایستگی مورد استفاده قرار می گرفت. در چنین شرایطی، یک خطای نسبتاً کوچک در اندازه گیری می تواند به تصمیم متفاوتی درباره فرد منجر شود، اما این جنبه در غالب مطالعات مورد بررسی قرار نگرفته بود.

ملاحظات اخلاقی و محدودیت های کاربرد نمرات در ۸ مطالعه گزارش شده بود و ۶ مطالعه به آثار سنجش بر دانشجومعلم، معلمان یا سایر ذی نفعان پرداخته بودند. با این حال، فقط ۲ مطالعه ارزیابی نسبتاً مستقیم و تجربی از هزینه ها و منافع استفاده از نتایج ارائه کرده بودند. بنابراین، شواهد نشان می دهد که حوزه تربیت معلم در زمینه اعتبارسنجی همچنان عمدتاً در سطح «آیا ابزار یک سازه را به صورت منسجم اندازه گیری می کند؟» باقی مانده است و پرسش پیچیده تر «آیا استفاده از این نمرات برای هدف مورد نظر موجه، عادلانه و از نظر پیامدها قابل دفاع است؟» کمتر مورد توجه قرار گرفته است. این فاصله میان کیفیت روان سنجی و مشروعیت کاربردی نمرات یکی از اصلی ترین خلأهای شناسایی شده در مرور حاضر بود.

در یک جمع بندی کلی از یافته ها، بررسی ۴۴ مطالعه نشان داد که سه سطح دقت اندازه گیری، شواهد تفسیری و شواهد پیامدی از پوشش متوازی برخوردار نیستند. در سطح نخست، بیشتر مطالعات حداقل یک شاخص پایایی، به ویژه آلفای کرونباخ، گزارش کرده بودند، اما شاخص های مستقیم خطای اندازه گیری، ثبات زمانی و دقت در زیرگروه ها به ندرت استفاده شده بود. در سطح دوم، ساختار درونی ابزارها به طور گسترده بررسی شده بود و بیش از چهارپنجم مطالعات از تحلیل عاملی استفاده کرده بودند، اما تنها حدود یک سوم مطالعات رابطه روشنی میان مجموعه شواهد گردآوری شده و تفسیر مشخص نمرات برقرار کرده بودند. در سطح سوم، یعنی پیامدهای استفاده از نتایج، خلأ بسیار بزرگ تری مشاهده شد و تقریباً دو سوم مطالعات فاقد هرگونه شواهد پیامدی مستقیم بودند. این الگو بیانگر نوعی عدم توازن در مفهوم پردازی روایی در پژوهش های تربیت معلم است؛ به گونه ای که هرچه از سطح فنی و آماری اندازه گیری به سمت تفسیر، کاربرد و پیامدهای تصمیم گیری حرکت می شود، حجم شواهد کاهش می یابد.

تحلیل مقایسه ای مطالعات همچنین نشان داد که پژوهش هایی که صرفاً به گزارش همسانی درونی و تحلیل عاملی اکتفا کرده بودند، معمولاً از واژه «روایی ابزار» به صورت کلی استفاده می کردند و روایی را همچون ویژگی ثابتی از پرسشنامه یا مقیاس تلقی می کردند. در مقابل، مطالعاتی که چند منبع شواهد را با یکدیگر ترکیب کرده بودند، بیشتر بر محدود بودن تفسیر نمرات به جامعه، زمینه و هدف خاص تأکید داشتند. این دسته از پژوهش ها به طور معمول از زبان دقیق تری استفاده می کردند و به جای بیان اینکه یک ابزار «دارای روایی است»، توضیح می دادند که شواهد موجود تا چه اندازه از یک تفسیر یا کاربرد مشخص حمایت می کند. چنین تفاوتی نشان می دهد که مسئله اصلی در ادبیات تربیت معلم صرفاً کمبود آزمون های آماری پیشرفته نیست، بلکه نحوه صورت بندی مفهوم روایی و ارتباط منطقی میان داده های تجربی، استنباط ها و کاربردهای مورد نظر نیز اهمیت بنیادی دارد.

از منظر ترکیب کلی شواهد، تنها ۸ مطالعه از ۴۴ مطالعه، معادل ۱۸٫۲ درصد، پوشش نسبتاً جامعی از سه حوزه اصلی مورد بررسی ارائه کرده بودند؛ به این معنا که علاوه بر دقت اندازه گیری، چند نوع شواهد تفسیری را گزارش کرده و دست کم بخشی از پیامدهای استفاده از نتایج را نیز بررسی کرده بودند. ۱۷ مطالعه، معادل ۳۸٫۶ درصد، در سطح میانی قرار گرفتند و عمدتاً شواهد قابل قبول درباره دقت اندازه گیری و ساختار یا روابط بیرونی ارائه کردند، اما بررسی پیامدها در آنها محدود بود. ۱۹ مطالعه، معادل ۴۳٫۲ درصد، دارای استدلال روایی محدود بودند و عمدتاً

روایی را از طریق یک یا دو شاخص روان‌سنجی، مانند آلفای کرونباخ و تحلیل عاملی، گزارش کرده بودند. این توزیع نشان می‌دهد که کمتر از یک‌پنجم ادبیات بررسی شده به الگوی نزدیک شده است که در آن روایی به‌عنوان استدلالی انباشتی درباره تفسیر و استفاده از نمرات در نظر گرفته شود.

در مجموع، یافته‌های این مرور نشان داد که بازتعریف روایی در پژوهش‌های تربیت معلم مستلزم عبور از الگوی سنتی «آزمون روایی ابزار» به سوی ارزیابی زنجیره‌ای از شواهد است. در این زنجیره، نخست باید مشخص شود که نمرات با چه میزان دقت و تحت چه میزان خطایی تولید می‌شوند؛ سپس باید نشان داده شود که این نمرات واقعاً از چه تفسیری حمایت می‌کنند؛ و در نهایت باید ارزیابی شود که استفاده از این تفسیر برای تصمیم‌مورد نظر چه پیامدهایی برای افراد، گروه‌ها و برنامه‌های تربیت معلم دارد. داده‌های این پژوهش نشان داد که در ادبیات موجود، حلقه نخست این زنجیره نسبتاً بیشتر مورد توجه قرار گرفته، حلقه دوم به‌طور نامتوازن و عمدتاً از طریق تحلیل ساختار درونی پوشش داده شده و حلقه سوم همچنان به میزان زیادی مغفول مانده است. از این‌رو، مسئله اصلی در بازمفهوم‌سازی روایی نه افزودن یک شاخص آماری دیگر به مجموعه آزمون‌های متداول، بلکه انتقال واحد تحلیل از «ابزار» به «استنباط و استفاده» و ارزیابی منسجم تمام شواهدی است که از آن استنباط و استفاده پشتیبانی می‌کنند.

بحث و نتیجه‌گیری

یافته‌های پژوهش حاضر نشان داد که الگوی غالب روایی در پژوهش‌های تربیت معلم هنوز به‌طور عمده بر شاخص‌های فنی و روان‌سنجی متمرکز است و میان سه سطح دقت اندازه‌گیری، شواهد تفسیری و شواهد پیامدی توازن کافی وجود ندارد. در میان ۴۴ مطالعه بررسی‌شده، شواهد مربوط به دقت اندازه‌گیری و ساختار درونی ابزارها بیشترین فراوانی را داشتند، در حالی که شواهد مربوط به پیامدهای استفاده از نتایج به‌مراتب کمتر گزارش شده بودند. این یافته نشان می‌دهد که در بخش قابل توجهی از پژوهش‌های تربیت معلم، روایی همچنان با کیفیت فنی ابزار، به‌ویژه پایایی و برازش ساختار عاملی، هم‌معنا یا دست‌کم بسیار نزدیک تلقی می‌شود. چنین الگویی را می‌توان در شماری از مطالعات توسعه و اعتباریابی ابزار مشاهده کرد که تمرکز اصلی آنها بر ساختار عاملی، همسانی درونی و کیفیت گویه‌ها قرار دارد (Erviana et al., 2023; Muhammad, 2025; Nabhan & Habók, 2025). اگرچه این شواهد برای ارزیابی کیفیت اندازه‌گیری ضروری‌اند، اما نتایج پژوهش حاضر نشان داد که اتکای غالب به آنها می‌تواند موجب محدود شدن مفهوم روایی به ویژگی‌های ابزار شود و رابطه میان نمره، تفسیر و کاربرد را در حاشیه قرار دهد.

در محور دقت اندازه‌گیری، آلفای کرونباخ یا شاخص‌های مشابه همسانی درونی در ۷۹٫۵ درصد مطالعات گزارش شده بود و به این ترتیب، رایج‌ترین شاخص مورد استفاده در مجموعه مطالعات بررسی‌شده محسوب می‌شد. در مقابل، شاخص‌هایی مانند ضریب امگا، پایایی بازآزمایی، خطای استاندارد اندازه‌گیری، نظریه پاسخ به سؤال و بررسی دقت نمرات در زیرگروه‌های مختلف فراوانی بسیار کمتری داشتند. این یافته نشان می‌دهد که مفهوم دقت اندازه‌گیری در بسیاری از پژوهش‌های تربیت معلم به ثبات درونی نمرات تقلیل یافته است. این الگو با پژوهش‌هایی همسو است که در توسعه ابزارهای تربیت معلم، پایایی درونی را یکی از شاخص‌های اصلی کیفیت ابزار در نظر گرفته‌اند (Erviana et al., 2023; Mujtahid et al., 2021). همچنین، در مطالعات مربوط به سنجش نگرش، شایستگی یا سواد حرفه‌ای دانشجومعلمان، استفاده از مقیاس‌های چندگویه‌ای و تحلیل همسانی درونی بسیار رایج است (Muhammad, 2025; Nabhan & Habók, 2025). با این حال، یافته حاضر نشان می‌دهد که گزارش همسانی درونی نمی‌تواند جایگزین بررسی خطای واقعی نمرات، ثبات زمانی یا دقت اندازه‌گیری در سطوح مختلف سازه شود.

این موضوع در پژوهش‌های تربیت معلم اهمیت ویژه‌ای دارد، زیرا بسیاری از سازه‌های مورد مطالعه ماهیت تحولی و زمینه‌ای دارند. برای مثال، هویت حرفه‌ای، نگرش نسبت به فناوری، خودکارآمدی و آمادگی برای تدریس ممکن است در طول دوره تربیت معلم تغییر کنند و بنابراین،

ثبات زمانی پایین همیشه به معنای ضعف ابزار نیست؛ بلکه ممکن است منعکس کننده تغییر واقعی سازه باشد. از سوی دیگر، بدون داده‌های طولی یا شاخص‌های مستقیم خطا، امکان تمایز این دو وضعیت دشوار می‌شود. مطالعات مرتبط با هویت حرفه‌ای دانشجویان و مدرسان نشان داده‌اند که این سازه تحت تأثیر حمایت اجتماعی، عوامل سازمانی و تجربه حرفه‌ای قرار دارد (Alpaslan et al., 2023; Farwa et al., 2024). همچنین، پژوهش‌های مربوط به هویت تدریس در مواجهه با تغییر برنامه درسی نشان می‌دهند که این سازه در طول زمان و در پاسخ به زمینه آموزشی تغییرپذیر است (Lin et al., 2024). در نتیجه، استفاده انحصاری از شاخص‌های همسانی درونی برای ارزیابی کیفیت چنین سازه‌هایی نمی‌تواند تصویر کاملی از دقت اندازه‌گیری ارائه دهد.

یافته مهم دیگر این بود که در میان شواهد تفسیری، ساختار درونی با ۸۱٫۸ درصد بیشترین سهم را داشت. این امر نشان می‌دهد که تحلیل عاملی اکتشافی و تأییدی در بسیاری از مطالعات به ابزار اصلی برای پشتیبانی از ادعاهای روایی تبدیل شده است. این نتیجه با روند مشاهده‌شده در مطالعات جدید تربیت معلم همسوست. برای مثال، توسعه مقیاس‌های مربوط به نگرش به آموزش فراگیر یا سواد نوشتاری دیجیتال به‌شدت بر تحلیل ساختار درونی متکی بوده است (Muhammad, 2025; Nabhan & Habók, 2025). همچنین، پژوهش‌هایی که سازه‌های پیچیده‌ای مانند شایستگی هوش مصنوعی را بررسی کرده‌اند، از روش‌های چندمتغیره و تحلیل ساختاری برای آشکار کردن الگوهای پنهان در داده‌ها استفاده کرده‌اند (Acquah et al., 2026). این مطالعات نشان می‌دهند که تحلیل ساختار درونی برای فهم ابعاد سازه اهمیت زیادی دارد، اما یافته پژوهش حاضر آشکار ساخت که در بسیاری از موارد، تحلیل عاملی پایان فرایند اعتبارسنجی تلقی شده است، نه یکی از اجزای یک استدلال گسترده‌تر درباره معنای نمره.

در همین راستا، تنها ۳۱٫۸ درصد مطالعات رابطه صریحی میان شواهد گردآوری‌شده و تفسیر مورد نظر از نمرات برقرار کرده بودند. این یافته از نظر مفهومی یکی از مهم‌ترین نتایج پژوهش حاضر است، زیرا نشان می‌دهد مسئله اصلی نه فقدان شاخص‌های آماری، بلکه ضعف در منطق استنباط است. ممکن است یک ابزار دارای برازش عاملی مناسب، پایایی بالا و همبستگی‌های قابل قبول با متغیرهای دیگر باشد، اما اگر مشخص نباشد این شواهد دقیقاً از چه تفسیری حمایت می‌کنند، استدلال روایی ناقص باقی می‌ماند. این مسئله در مطالعات مربوط به TPACK، شایستگی دیجیتال و استفاده از فناوری اهمیت ویژه دارد، زیرا در این حوزه‌ها مرز میان دانش ادراک‌شده، نگرش، مهارت و عملکرد واقعی همیشه روشن نیست (Ajani, 2024; Durusoy & Karamete, 2023; Li, 2023). برای نمونه، بالا بودن نمره دانشجویان در یک مقیاس خودگزارشی فناوری الزاماً به معنای توانایی بالای او در استفاده مؤثر از فناوری در موقعیت واقعی تدریس نیست.

این مسئله در حوزه‌های نوظهوری مانند هوش مصنوعی حتی برجسته‌تر است. نتایج مطالعات جدید نشان می‌دهد که شایستگی یا قصد استفاده از هوش مصنوعی حاصل تعامل عوامل شناختی، عاطفی و زمینه‌ای است (Acquah et al., 2026; Devaraj, 2026). بنابراین، تفسیر یک نمره واحد به‌عنوان «شایستگی هوش مصنوعی» ممکن است دامنه‌ای گسترده‌تر از آنچه ابزار واقعاً اندازه‌گیری می‌کند داشته باشد. یافته‌های پژوهش حاضر در این زمینه تأکید می‌کنند که روایی باید بر تطابق میان ادعای نظری و دامنه واقعی شواهد استوار باشد. به عبارت دیگر، اگر ابزار عمدتاً ادراک یا نگرش را می‌سنجد، استفاده از نتایج آن برای استنتاج درباره عملکرد یا توانایی عملی نیازمند شواهد اضافی است.

در زمینه روابط با متغیرهای بیرونی، بیش از نیمی از مطالعات شواهدی از این نوع ارائه کرده بودند، اما کیفیت این شواهد یکسان نبود. بخشی از مطالعات روابط نظری معناداری میان سازه مورد سنجش و متغیرهای مرتبط بررسی کرده بودند، در حالی که در بخشی دیگر، همبستگی‌های آماری بدون صورت‌بندی روشن فرضیه‌های نظری گزارش شده بود. این یافته با پیچیدگی سازه‌های تربیت معلم سازگار است. برای مثال، ارتباط میان دانش ریاضی برای تدریس و خودکارآمدی ریاضی معلمان از نظر نظری قابل انتظار است (Alshehri & Youssef, 2022)، همان‌گونه که رابطه میان هویت حرفه‌ای، حمایت اجتماعی و ادراک حرفه‌ای نیز دارای منطق نظری مشخصی است (Alpaslan et al., 2023). با این حال، صرف وجود یک رابطه معنادار آماری برای پشتیبانی از یک تفسیر روایی کافی نیست؛ شدت، جهت و منطق نظری رابطه نیز باید از پیش مشخص و سپس با داده‌ها ارزیابی شود.

شواهد محتوایی در حدود نیمی از مطالعات مشاهده شد، اما یافته‌ها نشان داد که این بخش نیز اغلب به داوری خبرگان محدود می‌شود. این الگو با بسیاری از مطالعات توسعه ابزار در آموزش و تربیت معلم همسواست که برای بررسی پوشش محتوایی گویه‌ها از نظر متخصصان استفاده می‌کنند (Eriviana et al., 2023; Mujtahid et al., 2021). با وجود اهمیت قضاوت تخصصی، اعتبار محتوایی زمانی قوی‌تر است که انتخاب متخصصان، معیارهای داوری، فرایند اصلاح گویه‌ها و ارتباط هر گویه با تعریف نظری سازه به‌طور شفاف گزارش شود. این موضوع در سنجش سازه‌هایی مانند آموزش فراگیر، شایستگی دیجیتال یا کیفیت برنامه‌های تربیت معلم اهمیت بیشتری دارد، زیرا این مفاهیم چندبعدی و تا حد زیادی وابسته به بافت هستند (Ayasrah, 2023; García et al., 2025).

یکی از یافته‌های قابل توجه دیگر، سهم پایین شواهد مربوط به فرایندهای پاسخ بود. تنها ۱۸,۲ درصد مطالعات بررسی کرده بودند که پاسخ‌دهندگان چگونه گویه‌ها را تفسیر می‌کنند یا چه فرایند ذهنی در شکل‌گیری پاسخ نقش دارد. این کمبود در پژوهش‌های تربیت معلم مسئله مهمی است، زیرا بسیاری از ابزارها بر سنجش ادراک، نگرش، باور یا خودارزیابی استوارند. برای مثال، مطالعات مربوط به نگرش دانش‌جو معلمان به برنامه‌نویسی، رضایت از برنامه‌های تربیت معلم یا دانش ادراک‌شده درباره آموزش متمایز عمدتاً به داده‌های خودگزارشی وابسته‌اند (Fang et al., 2024; Nketsia et al., 2024; Sanusi et al., 2025). در چنین ابزارهایی، روشن شدن اینکه پاسخ‌دهنده یک اصطلاح یا موقعیت را چگونه می‌فهمد، برای تفسیر نمره اهمیت اساسی دارد. فقدان شواهد فرایند پاسخ می‌تواند سبب شود تفاوت در نمرات به تفاوت در سازه نسبت داده شود، در حالی که بخشی از آن ممکن است ناشی از تفاوت در فهم گویه‌ها یا چارچوب مرجع پاسخ‌دهندگان باشد.

نتایج همچنین نشان داد که تغییرناپذیری اندازه‌گیری و بررسی سوگیری گویه‌ها در تعداد محدودی از مطالعات مورد توجه قرار گرفته است. تنها ۲۰,۵ درصد مطالعات تغییرناپذیری اندازه‌گیری و ۱۳,۶ درصد عملکرد افتراقی گویه‌ها را بررسی کرده بودند. این یافته در حوزه تربیت معلم مهم است، زیرا بسیاری از مطالعات به مقایسه گروه‌های مختلف دانش‌جو معلمان، رشته‌ها، کشورها یا بافت‌های فرهنگی می‌پردازند. پژوهش‌های مرتبط با آموزش فراگیر، فناوری و هویت حرفه‌ای نشان داده‌اند که زمینه فرهنگی و نهادی می‌تواند بر برداشت و پاسخ افراد اثر بگذارد (Farwa et al., 2024; García et al., 2025; Lin et al., 2024). بنابراین، مقایسه میانگین‌ها بدون بررسی هم‌ارزی اندازه‌گیری ممکن است تفاوت‌های ظاهری ایجاد کند که بخشی از آن ناشی از عملکرد متفاوت ابزار در گروه‌ها باشد.

مهم‌ترین شکاف شناسایی‌شده در پژوهش حاضر به شواهد پیامدی مربوط بود. در ۶۵,۹ درصد مطالعات هیچ شواهد مستقیم درباره پیامدهای استفاده از نتایج گزارش نشده بود. این یافته نشان می‌دهد که بخش بزرگی از ادبیات تربیت معلم هنوز اعتبارسنجی را پیش از مرحله کاربرد متوقف می‌کند. این در حالی است که نتایج سنجش در تربیت معلم می‌توانند برای اصلاح برنامه درسی، ارزیابی کیفیت، تشخیص نیازهای آموزشی، تصمیم درباره صلاحیت حرفه‌ای و طراحی مداخلات آموزشی مورد استفاده قرار گیرند. مرورهای مربوط به ارزیابی معلم نیز نشان داده‌اند که ارزیابی می‌تواند پیامدهای گسترده‌ای برای تصمیم‌های حرفه‌ای، توسعه شغلی و سیاست‌های آموزشی داشته باشد (Wei et al., 2023). همچنین، مطالعات ارزیابی برنامه‌های تربیت معلم تأکید کرده‌اند که نتیجه‌گیری درباره اثربخشی برنامه‌ها با چالش‌های جدی طراحی و استنباط همراه است (Mancenido, 2023).

ضعف شواهد پیامدی زمانی اهمیت بیشتری پیدا می‌کند که سنجش با تصمیم‌های سطح بالا همراه باشد. برای مثال، استفاده از ابزارهای سنجش نگرش به آموزش فراگیر یا شایستگی حرفه‌ای برای تعیین نیاز به آموزش تکمیلی می‌تواند پیامدهایی مستقیم برای دانش‌جو معلمان داشته باشد (García et al., 2025; Muhammad, 2025). به همین ترتیب، ارزیابی کیفیت فارغ‌التحصیلان یا برنامه‌های تربیت معلم می‌تواند بر قضاوت درباره عملکرد دانشگاه‌ها و مؤسسات آموزشی اثر بگذارد (Olvido et al., 2024). پژوهش حاضر نشان داد که در بسیاری از مطالعات، چنین کاربردهایی پیشنهاد شده‌اند، اما احتمال خطای طبقه‌بندی، پیامدهای ناخواسته یا عدالت در استفاده از نتایج به‌طور مستقیم بررسی نشده است.

این نتیجه با یافته مربوط به ارزیابی عدالت نیز همسو است. تنها ۲۲٫۷ درصد مطالعات به عدالت یا تفاوت پیامدها میان زیرگروه‌ها پرداخته بودند. در حالی که تربیت معلم به‌طور فزاینده‌ای بر آموزش فراگیر، پاسخ‌گویی به تنوع یادگیرندگان و عدالت آموزشی تأکید دارد (García, 2026; Yildirim, 2026; et al., 2025)، خود ابزارهای سنجش مورد استفاده در این حوزه نیز باید از منظر عدالت ارزیابی شوند. اگر یک ابزار در گروه‌های مختلف فرهنگی، زبانی یا آموزشی عملکرد متفاوتی داشته باشد، استفاده از نمرات آن برای تصمیم‌گیری ممکن است به نتایج نابرابر منجر شود. بنابراین، عدالت نه یک موضوع جانبی، بلکه بخشی از اعتبار کاربرد نمرات است.

یکی دیگر از یافته‌های مهم پژوهش حاضر این بود که تنها ۱۸٫۲ درصد مطالعات پوشش نسبتاً جامعی از سه حوزه دقت اندازه‌گیری، شواهد تفسیری و شواهد پیامدی ارائه کرده بودند. این نتیجه نشان می‌دهد که رویکرد جامع به روایی هنوز در ادبیات تربیت معلم فراگیر نشده است. بخش بزرگی از مطالعات یا بر کیفیت فنی ابزار تمرکز داشته‌اند یا حداکثر به برخی شواهد تفسیری افزوده‌اند، اما کمتر به چرخه کامل تولید نمره، تفسیر و کاربرد پرداخته‌اند. این یافته با مطالعاتی که بر چندبعدی بودن کیفیت تربیت معلم و ضرورت توجه به طیف گسترده‌ای از شواهد تأکید دارند همسو است (Ekabu, 2023; Guo et al., 2023; Olvido et al., 2024). همچنین، مرورهای مربوط به آموزش حرفه‌ای معلمان نشان داده‌اند که اثربخشی برنامه‌های تربیت معلم را نمی‌توان صرفاً بر پایه یک شاخص یا یک منبع داده تعیین کرد (Mancenido, 2023; Tortorelli et al., 2021).

در مجموع، نتایج پژوهش حاضر از ضرورت بازتعریف روایی به‌عنوان یک استدلال انباشتی و زمینه‌مند حمایت می‌کنند. بر اساس این دیدگاه، دقت اندازه‌گیری باید به‌عنوان شرط لازم اما ناکافی در نظر گرفته شود؛ شواهد تفسیری باید مشخص کنند که نمرات چه معنایی دارند و تا چه حد با سازه نظری مورد نظر سازگارند؛ و شواهد پیامدی باید روشن کنند که استفاده از این نمرات برای یک هدف خاص تا چه اندازه موجه، عادلانه و قابل دفاع است. این رویکرد به‌ویژه در شرایطی اهمیت دارد که تربیت معلم با حوزه‌های جدیدی مانند هوش مصنوعی، فناوری‌های دیجیتال، آموزش مبتنی بر شایستگی و آموزش فراگیر روبه‌رو است (Acquah et al., 2026; Ajani, 2024; Devaraj, 2026; Yildirim, 2026). در چنین زمینه‌هایی، سرعت توسعه ابزارهای جدید نباید از سرعت انباشت شواهد روایی پیشی بگیرد.

پژوهش حاضر با وجود ارائه تصویری نسبتاً جامع از نحوه مفهوم‌پردازی و گزارش روایی در پژوهش‌های تربیت معلم، با چند محدودیت همراه بود. نخست، نمونه‌نمایی به ۴۴ مطالعه محدود شد و بنابراین نتایج باید با توجه به دامنه منابع بررسی‌شده تفسیر شوند. دوم، ناهمگونی قابل توجهی از نظر نوع سازه، جامعه پژوهش، ابزار اندازه‌گیری، بافت فرهنگی و طراحی پژوهش میان مطالعات وجود داشت و این ناهمگونی امکان مقایسه مستقیم همه شاخص‌ها را محدود می‌کرد. سوم، برخی مطالعات اطلاعات کافی درباره فرایند طراحی ابزار، خطای اندازه‌گیری، نحوه تصمیم‌گیری بر اساس نمرات یا پیامدهای استفاده از نتایج ارائه نکرده بودند و در نتیجه، نبود یک نوع شواهد در مقاله لزوماً به معنای آن نیست که پژوهشگران آن را در عمل بررسی نکرده‌اند. چهارم، ارزیابی پژوهش حاضر بر گزارش‌های منتشرشده استوار بود و داده‌های خام، مستندات فنی ابزارها یا گزارش‌های غیرمنتشرشده در دسترس نبود. همچنین، به دلیل تنوع زیاد شاخص‌ها و طرح‌ها، انجام فراتحلیل کمی برای همه ابعاد روایی امکان‌پذیر نبود و تحلیل اصلی بر سنتز تفسیری و مقایسه الگوهای گزارش‌شده استوار شد.

پیشنهاد می‌شود پژوهش‌های آینده روایی را از مرحله طراحی مطالعه و پیش از جمع‌آوری داده‌ها به‌عنوان یک استدلال چندمرحله‌ای تعریف کنند و برای هر تفسیر یا کاربرد مورد نظر، شواهد لازم را از پیش مشخص سازند. استفاده از طرح‌های طولی، چندمنبعی و چندروشی می‌تواند امکان تفکیک تغییر واقعی سازه از خطای اندازه‌گیری را افزایش دهد. همچنین، لازم است مطالعات بیشتری به بررسی خطای استاندارد اندازه‌گیری، دقت نمرات در سطوح مختلف سازه، تغییرناپذیری اندازه‌گیری و عملکرد افتراقی گویه‌ها بپردازند. استفاده از مصاحبه شناختی، روش‌های «تفکر با صدای بلند» و تحلیل فرایند پاسخ نیز می‌تواند به روشن شدن معنای واقعی پاسخ‌ها کمک کند. پژوهش‌های آینده باید پیامدهای واقعی استفاده از نمرات را نیز به‌صورت تجربی بررسی کنند، به‌ویژه در مواردی که نتایج برای طبقه‌بندی افراد، ارزیابی صلاحیت حرفه‌ای، اصلاح برنامه‌های تربیت معلم یا تصمیم‌های سازمانی به کار می‌روند. انجام مطالعات بین‌فرهنگی همراه با آزمون هم‌ارزی اندازه‌گیری،

بازاعتبارسنجی ابزارها در بافت‌های متفاوت و بررسی نقش زبان، فرهنگ و نوع برنامه تربیت معلم نیز می‌تواند دامنه تعمیم استنباط‌های روایی را روشن‌تر کند.

در عمل، پژوهشگران تربیت معلم باید از به‌کارگیری عبارت‌های کلی مانند «ابزار دارای روایی است» پرهیز کنند و به‌جای آن مشخص سازند که چه نوع شواهدی از چه تفسیر و کاربردی حمایت می‌کند. گزارش یک ضریب پایایی یا یک مدل عاملی مناسب نباید به‌عنوان پایان فرایند اعتبارسنجی تلقی شود. برای ابزارهای خودگزارشی، توصیه می‌شود در صورت امکان از داده‌های عملکردی، مشاهده، ارزیابی استاد یا سایر منابع مستقل نیز استفاده شود تا تفسیر نمرات محدود به ادراک فردی نباشد. هنگام مقایسه گروه‌ها، لازم است پیش از تحلیل تفاوت میانگین‌ها، هم‌ارزی اندازه‌گیری بررسی شود. در مطالعاتی که نتایج می‌توانند بر تصمیم‌های آموزشی یا حرفه‌ای اثر بگذارند، پژوهشگران باید پیامدهای احتمالی خطای اندازه‌گیری، طبقه‌بندی نادرست و تفاوت آثار در زیرگروه‌ها را گزارش کنند. همچنین، مجلات و داوران علمی می‌توانند از پژوهشگران بخواهند به‌جای ارائه مجموعه‌ای از شاخص‌های پراکنده روان‌سنجی، منطق ارتباط میان شواهد، تفسیر و کاربرد نمرات را به‌صورت صریح بیان کنند؛ رویکردی که می‌تواند کیفیت استنباط‌ها و استفاده مسئولانه‌تر از نتایج سنجش را در پژوهش‌های تربیت معلم تقویت کند.

مشارکت نویسندگان

در نگارش این مقاله تمامی نویسندگان نقش یکسانی ایفا کردند.

تعارض منافع

در انجام مطالعه حاضر، هیچ‌گونه تضاد منافی وجود ندارد.

Extended Abstract

Introduction

Validity is one of the most fundamental concepts in educational measurement and teacher education research because the quality of conclusions drawn from empirical data depends not only on the technical characteristics of an instrument but also on the defensibility of the interpretations and uses assigned to the resulting scores. Contemporary teacher education research increasingly relies on questionnaires, rating scales, competency assessments, performance measures, observation protocols, and technology-related instruments to evaluate a wide range of constructs, including professional identity, self-efficacy, inclusive education attitudes, technological pedagogical content knowledge, digital literacy, artificial intelligence competence, and readiness for professional practice. The expansion of these constructs reflects the increasingly multidimensional nature of teacher preparation and the growing expectations placed on preservice teachers and teacher educators in rapidly changing educational contexts (Acquah et al., 2026; Ajani, 2024; Alpaslan et al., 2023).

Recent studies illustrate the diversification of measurement in teacher education. For example, professional identity has been examined in relation to social support and professional perceptions (Alpaslan et al., 2023), while teacher educators' identity development has been associated with organizational and psychological factors (Farwa et al., 2024). Research has also focused on mathematical self-efficacy (Alshehri & Youssef, 2022), technological pedagogical content knowledge (Ajani, 2024; Durusoy & Karamete, 2023), technology integration attitudes (Li, 2023), competency-based teacher education (Ekabu, 2023), and the quality of teacher education programs and graduates (Ayasrah, 2023; Olvido et al., 2024). At the same time, emerging areas such as artificial intelligence competence, generative AI adoption, programming education, digital academic writing, and augmented reality have introduced new constructs that require increasingly sophisticated measurement and interpretation procedures (Acquah et al., 2026; Devaraj, 2026; Mena et al., 2023a, 2023b; Nabhan & Habók, 2025; Sanusi et al., 2025).

Despite this methodological expansion, validity is still frequently treated as if it were an intrinsic and permanent property of an instrument. In many empirical studies, an instrument is labeled "valid" after evidence

of internal consistency, expert judgment, or factorial structure has been reported. However, such evidence addresses only part of the broader validity argument. Measurement precision is necessary because unstable or error-prone scores cannot support defensible interpretations, yet reliability alone cannot determine what scores mean or whether they are appropriate for a particular use. Similarly, a satisfactory factor structure can support claims about internal organization, but it does not automatically justify broader inferences about professional competence, instructional effectiveness, or decision-making. Studies focused on instrument development and psychometric evaluation in teacher education demonstrate the importance of reliability and factorial evidence, but they also reveal the need to move beyond narrow psychometric confirmation (Erviana et al., 2023; Muhammad, 2025; Nabhan & Habók, 2025).

The interpretive dimension of validity is particularly important in teacher education because many constructs are contextual, developmental, and partly self-reported. Measures of professional identity, inclusive education attitudes, technology competence, perceived knowledge, and professional readiness may be sensitive to cultural background, institutional context, prior experience, and respondents' interpretation of item wording (García et al., 2025; Lin et al., 2024; Nketsia et al., 2024). Consequently, the same score may not necessarily have an equivalent meaning across groups or settings. This problem becomes more pronounced when self-reported measures are interpreted as direct indicators of actual professional competence. For example, a high score on a technology-related self-report scale may reflect confidence or positive attitudes rather than demonstrated classroom performance. Similar concerns arise in the assessment of AI competence and intention to use generative AI, where cognitive, affective, and contextual factors interact in shaping measured responses (Acquah et al., 2026; Devaraj, 2026).

A further issue concerns the consequences of score use. Teacher education assessments may inform decisions about program improvement, professional readiness, curriculum revision, training needs, inclusion, and institutional quality. Therefore, the validity of score use cannot be evaluated solely by examining internal psychometric properties. Teacher evaluation research has demonstrated that assessment outcomes may influence professional development and institutional decision-making (Wei et al., 2023), while research on teacher preparation programs has emphasized methodological challenges in drawing defensible conclusions about program effectiveness (Mancenido, 2023). Studies on inclusive education and differentiated instruction further demonstrate that teacher-related assessments may have implications for equity, diversity, and educational access (García et al., 2025; Nketsia et al., 2024; Yildirim, 2026). Consequently, a comprehensive validity framework should incorporate not only measurement precision and interpretive evidence but also evidence concerning the intended and unintended consequences of score use.

The growing use of authentic, diagnostic, competency-based, digital, and performance-oriented assessment further reinforces the need for a broader understanding of validity (Baltynova et al., 2023; Fadlillah & Kusaeri, 2024). Teacher education research increasingly evaluates complex professional capacities rather than isolated knowledge components. This development requires validity to be reconceptualized as an accumulating, contextualized argument about the adequacy of evidence supporting score interpretation and use rather than as a fixed characteristic assigned to an instrument. Accordingly, the present study aimed to reconceptualize validity in teacher education research by systematically examining measurement precision, interpretive evidence, and consequential evidence and by clarifying how these three domains collectively support defensible interpretations and uses of assessment results.

Methods and Materials

The study employed an analytical review design focused on contemporary validity practices in teacher education research. The review covered studies published between 2010 and 2025 that directly addressed measurement, assessment, instrument development, score interpretation, or validity-related evidence in teacher education contexts. The initial search identified 286 records. After removing 74 duplicates, 212 records

remained for title and abstract screening. At this stage, 121 studies were excluded because they were not directly related to teacher education, focused exclusively on reliability without relevant validity evidence, lacked interpretable measurement information, or did not provide sufficient empirical or methodological detail. Full texts of 91 studies were then assessed. A further 47 studies were excluded because they lacked adequate validity-related information, did not clearly identify the basis for score interpretation, focused on contexts outside teacher education, or did not provide extractable data. The final review therefore included 44 studies. A structured data extraction form was developed to record publication characteristics, geographical context, participant group, study design, type of measurement instrument, and types of validity evidence. The extraction framework was organized around three principal domains: measurement precision, interpretive evidence, and consequential evidence. Measurement precision included internal consistency, composite reliability, omega coefficients, test-retest reliability, inter-rater agreement, item analysis, standard error of measurement, item response theory indices, and subgroup-specific precision. Interpretive evidence included content evidence, response processes, internal structure, convergent and discriminant evidence, relationships with external variables, criterion-related evidence, measurement invariance, differential item functioning, and explicit connections between evidence and intended score interpretations. Consequential evidence included intended and unintended consequences, fairness, classification error, effects on educational or professional decisions, stakeholder consequences, ethical considerations, and evaluation of costs and benefits.

Data were analyzed using descriptive analysis, directed content analysis, and interpretive synthesis. Frequencies and percentages were calculated for study characteristics and different forms of validity evidence. Each study was coded across the three major analytical domains. Particular attention was given to distinguishing the reporting of psychometric statistics from the construction of an explicit validity argument. Studies were also categorized according to the comprehensiveness of their validity evidence, and recurring gaps, patterns, and inconsistencies in score interpretation and use were synthesized.

Findings

Of the 44 studies included in the review, 24 studies (54.5%) were published between 2020 and 2025, 14 (31.8%) between 2015 and 2019, and 6 (13.6%) between 2010 and 2014. Preservice teachers were the primary participants in 25 studies (56.8%), while 8 studies (18.2%) focused on teachers participating in professional preparation or development programs, 4 studies (9.1%) focused on teacher educators, and 7 studies (15.9%) used mixed samples. Cross-sectional designs were most common, appearing in 22 studies (50.0%). Self-report questionnaires or scales were the dominant form of measurement in 21 studies (47.7%).

Measurement precision was most frequently represented by internal consistency coefficients. Cronbach's alpha or comparable internal consistency indices were reported in 35 studies (79.5%). Item analysis or item-total statistics were reported in 18 studies (40.9%), composite reliability in 12 studies (27.3%), omega coefficients in 10 studies (22.7%), and inter-rater agreement or intraclass correlation coefficients in 9 studies (20.5%). Only 7 studies (15.9%) examined test-retest reliability, 5 studies (11.4%) reported standard errors of measurement or equivalent direct error indices, and 4 studies (9.1%) used item response theory or test information approaches. Only 3 studies (6.8%) systematically examined measurement precision across subgroups.

Interpretive evidence was dominated by internal structure analysis. Factor analytic or comparable structural evidence was reported in 36 studies (81.8%). Relationships with external variables were examined in 25 studies (56.8%), while 23 studies (52.3%) reported content-related evidence. Convergent evidence was reported in 21 studies (47.7%) and discriminant evidence in 17 studies (38.6%). Only 11 studies (25.0%) linked scores to an external criterion, 9 studies (20.5%) examined measurement invariance, 8 studies (18.2%) addressed response processes, and 6 studies (13.6%) considered differential item functioning or possible measurement bias. Most importantly, only 14 studies (31.8%) explicitly connected the collected evidence to a specific intended interpretation of scores.

Consequential evidence was the least developed domain. Only 12 studies (27.3%) addressed intended consequences of score use, 10 studies (22.7%) considered fairness or differential consequences across groups, and 8 studies (18.2%) discussed ethical issues or limitations associated with score use. Seven studies (15.9%) examined the influence of scores on educational or professional decisions, 6 studies (13.6%) considered effects on preservice teachers, teachers, or other stakeholders, and 5 studies (11.4%) addressed unintended consequences. Classification or decision error was directly examined in only 4 studies (9.1%), and only 2 studies (4.5%) evaluated the costs and benefits of score use. Overall, 29 studies (65.9%) provided no direct consequential evidence.

The integrative analysis showed that only 8 studies (18.2%) provided relatively comprehensive evidence across measurement precision, interpretive evidence, and consequential evidence. Seventeen studies (38.6%) demonstrated moderate coverage, generally combining measurement precision with some interpretive evidence but offering limited examination of consequences. Nineteen studies (43.2%) presented limited validity arguments and relied mainly on one or two psychometric indicators, most commonly internal consistency and factor analysis.

Discussion and Conclusion

The findings indicate a substantial imbalance in the way validity is conceptualized and documented in teacher education research. The strongest emphasis remains on technical properties of scores, especially internal consistency and internal structure, while direct evidence concerning the meaning and consequences of score use remains comparatively limited. This pattern suggests that validity is often operationalized as a checklist of psychometric indicators rather than as an integrated argument linking evidence, interpretation, and intended use.

The dominance of internal consistency coefficients is particularly noteworthy. Although such coefficients are important for evaluating score stability within a measure, they provide limited information about temporal stability, individual score error, precision across different levels of a construct, or equivalence across groups. The very low frequency of standard error measures, item response theory approaches, and subgroup-specific precision demonstrates that uncertainty in individual and comparative score interpretation is rarely addressed directly.

A similar pattern was observed for interpretive evidence. Internal structure was extensively examined, yet only about one-third of studies explicitly explained how the collected evidence supported the intended meaning of scores. This distinction is crucial because statistical model fit cannot substitute for a validity argument. An instrument may demonstrate an acceptable factorial solution while still failing to justify broader interpretations regarding professional competence, readiness, or instructional quality. The limited attention to response processes and measurement invariance further suggests that assumptions about how respondents understand items and whether scores function equivalently across groups often remain untested.

The greatest weakness concerned consequential evidence. Nearly two-thirds of the studies provided no direct evidence concerning what happens when scores are used for educational, professional, or programmatic decisions. This omission is especially important in teacher education because assessment outcomes may influence training needs, professional judgments, curriculum revision, program evaluation, and institutional policy. Without examining fairness, classification error, unintended effects, or stakeholder consequences, technically sound measurement may still lead to questionable or inequitable decisions.

The findings therefore support a reconceptualization of validity as a cumulative and context-dependent argument rather than a fixed property of an instrument. Measurement precision should be treated as the technical foundation of the validity argument, interpretive evidence should establish what scores can reasonably mean, and consequential evidence should evaluate whether the intended use of those interpretations

is justified and defensible. Under this framework, validity is not established once and for all; rather, it must be continually supported whenever the population, context, construct, or purpose of score use changes.

In conclusion, teacher education research would benefit from moving beyond the conventional practice of declaring instruments “valid and reliable” on the basis of a small set of psychometric statistics. A stronger approach requires explicit articulation of intended score interpretations, systematic collection of multiple sources of evidence, direct examination of measurement error and group comparability, and assessment of the consequences associated with score use. Such a reconceptualization can strengthen methodological rigor, improve the defensibility of research conclusions, and promote more responsible use of assessment evidence in teacher education.

References

- Acquah, B. Y. S., Arthur, F., Gyedu, F. O., Quayson, E., Boateng, E., Quaye, S. A., & Nortey, S. A. (2026). Preservice Teachers' Artificial Intelligence Competence: A Network and Latent Profile Analyses. *European Journal of Education*, 61(3). <https://doi.org/10.1111/ejed.70735>
- Ajani, O. A. (2024). Technological Pedagogical Content Knowledge for Twenty-First Century Learning Skills. *International Journal of Research in Business and Social Science* (2147-4478), 13(4), 468-476. <https://doi.org/10.20525/ijrbs.v13i4.3355>
- Alpaslan, M. M., Ulubey, Ö., & Yıldırım, K. (2023). Exploring the Relationship Between Preservice Teachers' Professional Identity, Professional Perceptions, and Social Support. *Psycho-Educational Research Reviews*, 12(3). https://doi.org/10.52963/perr_biruni_v12.n3.02
- Alshehri, K. A., & Youssef, N. H. (2022). The Influence of Mathematical Knowledge for Teaching Towards Elementary Teachers' Mathematical Self-Efficacy. *Eurasia Journal of Mathematics Science and Technology Education*, 18(6), em2118. <https://doi.org/10.29333/ejmste/12086>
- Aslan, Y. G. (2022). Effects of Behavioral Skills Training on Teachers' Implementation of Power Card Strategy. *Khazar Journal of Humanities and Social Sciences*, 25(4), 5-20. <https://doi.org/10.5782/2223-2621.2022.25.4.5>
- Ayasrah, M. N. (2023). Quality Evaluation of Special Education Teacher Preparation Programs at the Northern Governorates Universities in the Hashemite Kingdom of Jordan, in Light of CAEP Standards. *Journal of Higher Education Theory and Practice*, 23(3). <https://doi.org/10.33423/jhetp.v23i3.5838>
- Baltynova, A., Kamariyash, K., Muzdbaeva, T., Bolat, Y., Beleukhanova, K., Zharikova, D., & Mollakuqe, E. (2023). Pedagogical Conditions for the Training of Future Teachers Based on Digital Educational Technologies. *International Journal of Emerging Technologies in Learning (Ijet)*, 18(18), 121-137. <https://doi.org/10.3991/ijet.v18i18.43209>
- Devaraj, M. (2026). Understanding Pre-Service Science Teachers' Intentions to Integrate Generative AI: Cognitive, Affective, and Contextual Predictors. <https://doi.org/10.21203/rs.3.rs-10970681/v1>
- Durusoy, O., & Karamete, A. (2023). Enhancing Pre-Service Teachers' Technological Pedagogical Content Knowledge (TPACK) Through the Learning by Design Framework: A Fink Taxonomy-Based Study. *Necatibey Eğitim Fakültesi Elektronik Fen Ve Matematik Eğitimi Dergisi*, 17(1), 174-210. <https://doi.org/10.17522/balikesirnef.1262115>
- Ekabu, P. K. (2023). Teacher Educators Pedagogical Effectiveness in Implementing Competency-Based Teacher Education and Competence-Based Assessment Programs: A Case of Primary Diploma Teacher Training Colleges in Kenya. *Jep*. <https://doi.org/10.7176/jep/14-33-07>
- Erviana, V. Y., Ghufon, A., & Retnawati, H. (2023). The Validity and Reliability of IoJEPD Model Instruments to Improve Elementary School Teacher Competence. *Pegagog*, 13(02). <https://doi.org/10.47750/pegagog.13.02.24>
- Fadlillah, N., & Kusaeri, K. (2024). Optimizing Assessment for Learning in Islamic Education Through Authentic and Diagnostic Assessment : A Systematic Literature Review. *Jurnal Kependidikan Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan Pengajaran Dan Pembelajaran*, 10(2), 654. <https://doi.org/10.33394/jk.v10i2.11555>
- Fang, L., Poon, C. H., & Wing, C. K. (2024). Pre-Service Music Teachers' Satisfaction With the Teacher Education Curriculum in Northern China. *International Journal of Education and Practice*, 12(3), 667-683. <https://doi.org/10.18488/61.v12i3.3747>
- Farwa, U. E., Li, Q., Ye, J., Khan, M. K., & Zulfiqar, S. (2024). Organizational and Psychological Factors in Teacher Educators' Professional Identity Development: An Empirical Analysis. *British Educational Research Journal*, 51(2), 869-896. <https://doi.org/10.1002/berj.4106>

- García, M. M., Ortiz-Marcos, J. M., Invernón-Gómez, A. I., & Higuera-Rodríguez, L. (2025). Educating for Inclusion: PERKIN, the Scale That Measures Future Teachers' Perceptions. *Sage Open*, 15(3). <https://doi.org/10.1177/21582440251367584>
- Guo, Q., Azar, A. S., & Ahmad, A. (2023). Factors Affecting Teachers Quality in Higher Vocational Colleges in Maoming City, China. *International Journal of Professional Business Review*, 8(8), e02994. <https://doi.org/10.26668/businessreview/2023.v8i8.2994>
- Li, M. (2023). Chinese Mathematics Teachers' TPACK and Attitudes Toward ICT Integration in the Post-Pandemic Era. *Eurasia Journal of Mathematics Science and Technology Education*, 19(7), em2301. <https://doi.org/10.29333/ejmste/13346>
- Lin, T. C., Lee, Y.-S., & Ye, J. H. (2024). A Study on Taiwan's Vocational Senior High School Teachers' Teaching Identity and Teaching Transformation When Facing a New Competency-Based Curriculum. *Frontiers in psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1290551>
- Mancenido, Z. (2023). Impact Evaluations of Teacher Preparation Practices: Challenges and Opportunities for More Rigorous Research. *Review of Educational Research*, 94(2), 268-307. <https://doi.org/10.3102/00346543231174413>
- Mena, J., Molina, O. E., & Pérez-Calvo, E. (2023a). Teachers' Professional Training Through Augmented Reality: A Literature Review. <https://doi.org/10.20944/preprints202304.0988.v1>
- Mena, J., Molina, O. E., & Pérez-Calvo, E. (2023b). Teachers' Professional Training Through Augmented Reality: A Literature Review. *Education Sciences*, 13(5), 517. <https://doi.org/10.3390/educsci13050517>
- Muhammad, A. E. (2025). Validating the Teachers' Attitudes Towards Inclusive Education Scale: A Psychometric Study With Pre-service Teachers. *British Journal of Special Education*, 52(4), 523-538. <https://doi.org/10.1111/1467-8578.70041>
- Mujtahid, I. M., Berlian, M., Vebrianto, R., & Thahir, M. (2021). Developing Training Evaluation Instrument for Teachers of Elementary School and Early Childhood Education in Pekanbaru, Riau. *İlköğretim Online*, 20(1). <https://doi.org/10.17051/ilkonline.2021.01.81>
- Nabhan, S., & Habók, A. (2025). The Digital Literacy Academic Writing Scale: Exploratory Factor Analysis. *Sage Open*, 15(1). <https://doi.org/10.1177/21582440241311709>
- Nketsia, W., Opoku, M. P., Amponteng, M., & Mprah, W. K. (2024). Exploring the Perceived Knowledge of Teacher Educators and Pre-Service Teachers on the Differentiated Instruction Practices of Teacher Educators. *Frontiers in Education*, 9. <https://doi.org/10.3389/feduc.2024.1356675>
- Olvido, M. M. J., Dayagbil, F. T., Alda, R., Uytico, B. J., & Rodriguez, K. F. R. (2024). An Exploration of the Quality of Graduates of Philippine Teacher Education Institutions. *Frontiers in Education*, 9. <https://doi.org/10.3389/feduc.2024.1235261>
- Saleh, S. G., AlAli, R. M., & Mashal, A. A. (2024). Competencies of Using Computers in Teaching: An Evaluative Study on Students of the Faculty of Education, Sohag University. *Journal of Ecohumanism*, 3(7), 2884-2905. <https://doi.org/10.62754/joe.v3i7.4422>
- Sanusi, I. T., Cudjoe, E. S., Ayanwale, M. A., & Adepoju, B. (2025). Pre-Service Teachers' Perception of Programming Education. *Sage Open*, 15(1). <https://doi.org/10.1177/21582440251327019>
- Tortorelli, L. S., Lupo, S. M., & Wheatley, B. C. (2021). Examining Teacher Preparation for Code-Related Reading Instruction: An Integrated Literature Review. *Reading Research Quarterly*, 56(S1). <https://doi.org/10.1002/rrq.396>
- Wei, X., Chow, M.-K., Huang, L., Huang, X., & Cheng, G. (2023). Teacher Evaluation in Primary and Secondary Schools: A Systematic Review of SSCI Journal Publications From 2012 to 2022. *Sustainability*, 15(9), 7280. <https://doi.org/10.3390/su15097280>
- Yildirim, Ş. S. (2026). Elevating ELT Pedagogy: Bridging Teacher Training Gaps for Inclusive Support of Specific Language Learner Difficulties. *Support for Learning*, 41(3). <https://doi.org/10.1111/1467-9604.70064>