

Optimal Number of Grading Levels: A Comparison of the Reliability of a 4-Point System with a 20-Point System in Assessing Practical Skills

Tannaz Abdollahzadeh Baghaei¹, Aydin Sohrabi^{2*}

¹Assistant Professor, Department of Orthodontics, Faculty of Dentistry, Tabriz University of Medical Sciences, Tabriz, Iran

²Professor, Department of Orthodontics, Faculty of Dentistry, Tabriz University of Medical Sciences, Tabriz, Iran

Received: 14 June 2026, Accepted: 8 September 2026

Background: Assessing learners' performance is a critical component of any educational system. This study investigated the agreement and reliability of using a simplified 4-point scale (A, B, C, and D) compared to the Iranian conventional 20-point grading scale for evaluating dental students' practical orthodontic skills.

Methods and Materials: This cross-sectional study was conducted in Fall 2023 at the School of Dentistry, Tabriz (Iran). Seventy-five practical tasks performed by 15 fourth-year dental students enrolled in the Orthodontics Practical 1 course were evaluated. Student performance on five orthodontic tasks was independently assessed by two faculty members using both a 4-point and a 20-point grading scale. Inter-assessor agreement was assessed using weighted Cohen's kappa for the 4-point scale and intraclass correlation coefficients (ICC) for the 20-point scale. Agreement between the two grading systems was evaluated using ICC and Bland-Altman analysis. A significance level of 0.05 was considered.

Results: The 4-point scale demonstrated moderate inter-assessor agreement (weighted kappa = 0.50), while the 20-point scale showed moderate agreement (ICC = 0.47). After conversion of the 4-point scores to their corresponding 20-point values, strong agreement was observed between the two grading systems (ICC = 0.944 and 0.969).

Conclusion: A 4-point scale may be a potential alternative to the traditional 20-point scale for assessing dental students' practical skills. This scale offers advantages in terms of simplicity and efficiency while maintaining a level of reliability. The choice between these scales ultimately depends on the specific needs and goals of the educational context.

Keywords: Reliability, Orthodontics, Dental Education, Grade

***Corresponding Author:** asohrabi@tbzmed.ac.ir

➤ **Please cite this paper as:** Abdollahzadeh Baghaei T, Sohrabi A. Optimal number of grading levels: a comparison of the reliability of a 4-point system with a 20-point system in assessing practical skills. *J Mashhad Dent Sch* 2026; 50(3):209-221.

➤ **DOI:** [10.22038/jmds.2026.28346](https://doi.org/10.22038/jmds.2026.28346)



تعداد بهینه سطوح نمره‌دهی: مقایسه پایایی سیستم ۴ رتبه‌ای با ۲۰ نمره‌ای در ارزیابی مهارت‌های عملی

طناز عبدالله زاده بقائی^۱، آیدین سهرابی^{۲*}

^۱استادیار، گروه ارتودنسی، دانشکده دندانپزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران
^۲استاد، گروه ارتودنسی، دانشکده دندانپزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران

تاریخ ارائه مقاله: ۱۴۰۵/۰۳/۲۴ - تاریخ پذیرش: ۱۴۰۵/۰۶/۱۷

چکیده

مقدمه: ارزیابی عملکرد دانشجویان یکی از اجزای حیاتی در هر سیستم آموزشی است. هدف انجام این مطالعه، ارزیابی پایایی و توافق سیستم ساده ۴ رتبه‌ای (A, B, C و D) در مقایسه با سیستم ۲۰ نمره‌ای رایج ایرانی در ارزیابی مهارت‌های ارتودنسی عملی دانشجویان دندانپزشکی بود.

مواد و روش‌ها: این مطالعه مقطعی در پاییز ۱۴۰۲ در دانشکده دندانپزشکی تبریز انجام شد. هفتاد و پنج کار عملی از پانزده دانشجوی سال چهارم دندانپزشکی که واحد ارتودنسی عملی ۱ را گذرانده بودند، مورد بررسی قرار گرفت. عملکرد دانشجویان در پنج کار عملی توسط دو عضو هیات علمی به صورت مستقل و با استفاده از هر دو سیستم ۴ رتبه‌ای و ۲۰ نمره‌ای ارزیابی شد. توافق بین ارزیابان با استفاده از ضریب کاپای وزنی کوهن برای سیستم ۴ رتبه‌ای و ضریب همبستگی درون‌رده‌ای (ICC) برای سیستم ۲۰ نمره‌ای بررسی شد. همچنین توافق بین دو سیستم نمره‌دهی با استفاده از ICC و نمودار Bland-Altman ارزیابی شد. سطح معنی‌داری ۰/۰۵ در نظر گرفته شد.

یافته‌ها: سیستم ۴ رتبه‌ای، توافق متوسط بین ارزیابان را نشان داد ($\text{weighted kappa} = ۰/۵۰$)، در حالی که سیستم ۲۰ نمره‌ای نیز دارای توافق متوسط بود ($\text{ICC} = ۰/۴۷$). پس از تبدیل نمرات سیستم ۴ رتبه‌ای به مقادیر متناظر سیستم ۲۰ نمره‌ای، توافق قوی بین دو سیستم مشاهده شد ($\text{ICC} = ۰/۹۴۴$ و $۰/۹۶۹$).

نتیجه‌گیری: سیستم ۴ رتبه‌ای ممکن است یک گزینه جایگزین بالقوه برای سیستم سنتی ۲۰ نمره‌ای در ارزیابی مهارت‌های عملی دانشجویان دندانپزشکی باشد. این سیستم ضمن داشتن مزایایی مانند سادگی و کارایی، سطح قابل قبولی از پایایی را حفظ می‌کند. انتخاب بین این سیستم‌ها باید بر اساس نیازها و اهداف خاص محیط آموزشی انجام شود.

کلمات کلیدی: پایایی، ارتودنسی، آموزش دندانپزشکی، نمره

مجله دانشکده دندانپزشکی مشهد / سال ۱۴۰۵ / دوره ۵۰ / شماره ۳: ۲۲۱-۲۰۹.

* مؤلف مسئول، نشانی: دانشکده دندانپزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران

مقدمه

ارزیابی عملکرد دانشجویان یک جزء حیاتی در هر سیستم آموزشی است و همانطور که از Cowan نقل شده است "ارزیابی، موتور محرکه یادگیری است". روش‌های ارزیابی می‌توانند در کشورها و در زمینه‌های آموزشی مختلف، متفاوت باشند. هر نهاد و هر سیستم آموزشی ممکن است سیستم نمره دهی خود را بر اساس سیاست‌های آموزشی محلی، هنجارهای فرهنگی رایج و تجربه اعضای هیأت علمی خود اتخاذ کند.^(۱)

اما هر سیستم نمره دهی باید از سطح قابل قبولی از پایایی و روایی برخوردار باشد تا دقت نتیجه ارزیابی را افزایش دهد.^(۲) سیستم‌های ارزیابی قابل اعتماد، ثبات و دقت را در ارزیابی عملکرد دانشجویان ارتقا داده و احتمال سوگیری (bias) و فردیت (subjectivity) در نمره دهی را کاهش می‌دهند. در غیر این صورت، این امر می‌تواند منجر به سوگیری در ارزیابی عملکرد دانشجویان شده و بر اعتبار و قابلیت مقایسه نتایج تأثیر بگذارد.^(۳)

سیستم‌های نمره دهی در کشورهای مختلف متفاوت است. برخی از درصد و برخی از نمرات عددی یا نمرات حروفی استفاده می‌کنند. در حال حاضر اکثر سیستم‌های نمره دهی از جمله سیستم‌های لیکرت و همچنین اغلب مقیاس‌های سنجش نگرش و عملکرد، دارای پنج تا هفت سطح A تا F در ایالات متحده، ۱ تا ۵ در استرالیا، ۱ تا ۶ در آلمان، A تا G در انگلستان و ... هستند، اما در کنار آنها سیستم‌های نمره دهی کاملاً متفاوتی مانند متد ۲۰-۰ در فرانسه وجود دارند، که سیستم استاندارد رایج در کشورمان نیز می‌باشد.^(۴-۷)

علیرغم دهه‌ها تحقیق، مساله تعداد بهینه scaleها در سیستم‌های نمره دهی همچنان محل بحث و پژوهش می‌باشد. به

طور سنتی، سیستم‌های ارزیابی در کشور ما از سیستم ۲۰ نمره‌ای برای ارزیابی جنبه‌های مختلف یادگیری فراگیران استفاده می‌کنند. در حالی که این سیستم‌های جامع، ارزیابی کاملی را ارائه می‌دهند، در عین حال می‌توانند زمان‌بر و پیچیده بوده و منجر به چالش‌هایی در دستیابی به پایایی و قابلیت تکرار بین ارزیابان شوند.^(۸) این شیوه به دلیل دارا بودن سطوح متعدد نمره (۲۱ نمره صحیح و سه جز کسری برای هر نمره) می‌تواند برای نمره دهنده با مقداری عدم قطعیت همراه باشد که گاهی اوقات غیر ضروری به نظر می‌رسد. این امر به ویژه در دوره‌های عملی یا تکالیف عملی که مدرس عملکرد فراگیر را بر اساس ذهنیت و مبتنی بر تجربه ارزیابی می‌کند، محسوس‌تر است.^(۹)

با توجه به این ملاحظات، مدرسین به دنبال روش‌های ارزیابی جایگزین هستند که تعادل بین ارزیابی جامع و سهولت استفاده را برقرار کنند. یکی از این رویکردها می‌تواند اتخاذ یک سیستم ساده ۴ رتبه‌ای باشد. مزیت اصلی سیستم ۴ رتبه‌ای در قابلیت استفاده آن نهفته است، زیرا تعداد دسته‌بندی‌های نمره دهی را به طور قابل توجهی در مقایسه با سیستم‌های ۲۰ نمره‌ای سنتی کاهش می‌دهد. با ارائه تنها چهار گزینه نمره، مدرسان می‌توانند به سرعت در مورد عملکرد دانشجویان تصمیم‌گیری کرده و آن را ارزیابی کنند. استفاده از چهار گزینه می‌تواند یک گزینه قابل کاربرد برای اعضای هیأت علمی دندانپزشکی باشد.

با این حال، سوال اصلی که با استفاده از سیستم ۴ رتبه‌ای مطرح می‌شود این است که آیا می‌تواند قابلیت اعتماد و پایایی سیستم‌های جامع‌تر دارای سطوح نمره دهی بیشتر را حفظ کند. در مطالعه حاضر، تکرارپذیری استفاده از سیستم ۴ رتبه‌ای در ارزیابی عملکرد دانشجویان سال چهار

دندانپزشکی در واحد ارتودنسی ۱ عملی (پره کلینیک) بررسی گردید. هدف از این مطالعه بررسی وجود توافق بین نمرات سیستم ۴ رتبه ای و سیستم ۲۰ نمره ای بود.

دانشجویان را ارزیابی کردند. ارزیابی شامل پنج تکلیف در واحد ارتودنسی پره کلینیک برای هر یک از فراگیران بود که شامل فرم دادن وایر ارتودنسی (wire bending) روی یک برگه کار نمونه (worksheet)، ساخت C-clasp روی کانین کست فک پایین، ساخت Adams-clasp روی مولر اول کست فک بالا، ساخت Ball clasp روی دندانهای قدامی کست فک پایین و تریسینگ سفالومتری لترال روی کاغذ استات بود. اعضای هیات علمی در این ارزیابی، کارهای عملی را با استفاده از هر دو سیستم ۴ رتبه‌ای (A-D-C-B) و سیستم سستی ۲۰ نمره‌ای، ارزیابی و نمره دهی کردند.

جزئیات هر سیستم ارزیابی عبارت بودند از:

الف) سیستم ۴ رتبه‌ای: این سیستم شامل چهار سطح عملکرد:

A: عالی،

B: خوب،

C: قابل قبول (قبول لب مرزی) و

D: غیرقابل قبول (مردود) بود.

ب) سیستم ۲۰ نمره‌ای: این سیستم شامل سیستم رایج نمره دهی صفر تا بیست بود.

دو مدرس A و B، به طور مستقل از یکدیگر، عملکرد دانشجویان در پنج تکلیف عملی را با استفاده از سیستم ۴ رتبه‌ای و مجدداً همان پنج تکلیف را با استفاده از سیستم ۲۰ نمره‌ای ارزیابی کردند. بدین منظور، تکلیف wire bending و تمرین ساخت کلاسپ آدامز همه دانشجویان کدگذاری و بدون نام دانشجو در اختیار اساتید قرار داده شد و مدرس A، این تکالیف را با استفاده از سیستم ۴ رتبه‌ای و مدرس B با استفاده از سیستم ۲۰ نمره‌ای، ارزیابی و نمره‌دهی کردند. دو هفته بعد، مدرس A این بار با استفاده از سیستم رایج ۲۰ نمره ای و مدرس B با استفاده از روش

دندانپزشکی در واحد ارتودنسی ۱ عملی (پره کلینیک) بررسی گردید.

هدف از این مطالعه بررسی وجود توافق بین نمرات سیستم ۴ رتبه ای و سیستم ۲۰ نمره ای بود.

مواد و روش‌ها

این مطالعه به تایید کمیته اخلاق دانشگاه علوم پزشکی تبریز (به شماره ۷۰۹۱۷۷-D-۵) رسیده است و به صورت یک پژوهش مشاهده‌ای مقطعی در دانشکده دندانپزشکی دانشگاه علوم پزشکی تبریز در پاییز ۱۴۰۲ اجرا شده است. این پروژه تحقیقاتی از پروتکل و استانداردهای اخلاقی برای تحقیقات آموزشی انسانی پیروی کرد. از تمام شرکت‌کنندگان رضایت برای حضور در این مطالعه، اخذ و محرمانگی آنها و اطلاعات مربوطه در طول مطالعه حفظ گردید. شرکت کنندگان این مطالعه، پانزده دانشجوی سال چهارم دندانپزشکی که در واحد ارتودنسی عملی ۱ (پره کلینیک) حضور داشتند، بودند. واحد مطالعه (unit of study) در این پژوهش، کار عملی ارائه شده توسط دانشجویان در نظر گرفته شده بود.

برای محاسبه حجم نمونه، از رویکرد متداول برای مطالعات توافق و از ضریب همبستگی درون‌رده‌ای (Intraclass correlation coefficient: ICC) به عنوان معیار توافق، استفاده شد. در این فرمول، با فرض ICC مورد انتظار ۰/۷۵، دقت برآورد ۰/۱، سطح معنی داری (α) ۰/۰۵ و در صورت استفاده از دو ارزیاب، حجم نمونه مورد نیاز ۷۵ محاسبه شد. علت تعیین دو نفر به عنوان ارزیاب این بود که این واحد درسی توسط دو عضو هیات علمی ارائه می شد.

این مطالعه با یک طراحی بررسی توافق inter-method و inter-assessor اجرا شد و در آن دو عضو هیئت علمی (مدرس A و B) به طور مستقل از یکدیگر، عملکرد عملی

چهار رتبه‌ای تکالیف ساخت ball clasp ، C-clasp و تریسینگ دانشجویان را ارزیابی نمودند. این فرآیند در هفته سوم و چهارم نیز تکرار شد و در این دورها، دو مدرس به شیوه دوم (اگر در دور اول با سیستم ۴ رتبه‌ای نمره داده بود، با سیستم ۲۰-۰ و بالعکس) کارهای عملی دانشجویان را ارزیابی و نمره دهی کردند.

ترتیب ارائه تکالیف به اساتید به صورت تصادفی انجام شد تا هرگونه اثرات احتمالی ترتیب را به حداقل رساند. بین دو ارزیابی ۲ هفته فاصله بود تا مطمئن شویم نمرات داده شده توسط هر روش از یکدیگر مستقل هستند. ارزیابان نسبت به نام دانشجویان، نمره دانشجویی که ۲ هفته پیش ارزیابی شده بود و همچنین نمرات مدرس بی اطلاع (blind) بودند. در مجموع ۳۰۰ نمره دهی انجام شد: (۱۵) دانشجو X ۵ کار عملی برای هر دانشجو X ۲ مدرس X ۲ روش نمره دهی = ۳۰۰ مشاهده).

توافق بین دو مدرس (inter-assessor agreement) برای بررسی پایایی در هر روش نمره دهی محاسبه شد: برای تعیین میزان توافق inter-assessor در سیستم ۴ رتبه‌ای، آماره weighted Cohen's kappa و فاصله اطمینانی ۹۵٪ آن محاسبه گردید.

تکرارپذیری سیستم ۲۰ نمره‌ای نیز با استفاده از ضریب همبستگی درون‌رده‌ای (ICC) بر اساس یک مدل two-way consistency for single measures همراه با فاصله اطمینانی ۹۵٪ محاسبه شد.

برای ارزیابی توافق بین دو روش نمره‌دهی، توافق درون ارزیاب (intra-assessor agreement) برای هر روش مورد استفاده قرار گرفت:

برای ارزیابی توافق بین دو روش نمره دهی، از آنجایی که در سیستم نمره دهی ۴ رتبه‌ای، از نمرات کیفی رتبه ای (A، B، C، و D) استفاده می‌کند، این چهار حرف را به نمرات متناظر عددی تبدیل کردیم و به این منظور به رتبه A مقدار ۲۰، به B مقدار ۱۶، به C مقدار ۱۲ و به D نمره ۸ را اختصاص دادیم. (در سیستم معمول ۲۰ نمره‌ای، نمرات زیر ۱۲ در دانشکده دندانپزشکی تبریز، مردود محسوب می‌شوند).

به این صورت تمامی حروف رتبه ای به اعداد متناظرشان عددگذاری شدند. سپس اختلاف نمرات هر تکلیف (تفاوت نمره اخذ شده از سیستم ۴ رتبه‌ای recode شده با نمره اخذ شده در سیستم ۲۰ نمره‌ای) محاسبه شد. همچنین، دو نمره معدل تکالیف برای هر دانشجو با استفاده از هر دو شیوه نمره دهی (سیستم ۲۰ نمره‌ای و اعداد متناظر با نمره اخذ شده با سیستم ۴ رتبه‌ای) و تفاوت این دو معدل محاسبه شد.

تمام تحلیل داده ها در MedCalc Statistical Software (MedCalc Software Ltd, Ostend, Belgium) نسخه ۲۲ و Microsoft Excel انجام شد. سطح معنی داری آماری ۰/۰۵ مفروض گردید.

یافته ها

مقایسه قابلیت اعتماد هر روش (توافق inter-assessor برای هر روش):

توافق بین مدرس A و B در استفاده از سیستم ۴ رتبه‌ای با آماره kappa برآورد گردید. فراوانی مشاهده شده از نمره های داده شده توسط هر یک از مدرسین در جدول ۱ ارائه شده است.

همبستگی درون‌رده‌ای اندازه‌گیری شد و نتایج، نشان دهنده توافقی برابر با ۰/۹۴۴ (با فاصله اطمینانی ۰/۹۵، ۰/۸۴۲) الی ۰/۹۸۱ (برای مدرس A و ۰/۹۶۹ (با فاصله اطمینانی ۰/۹۵، ۰/۹۰۹) الی ۰/۹۸۱) برای مدرس B (جدول ۴) بود، که مطابق با طبقه‌بندی Koo & Li در محدوده توافق‌های خوب تا عالی و عالی قرار می‌گیرند.^(۱۰)

نتایج توافق بین دو مدرس در سیستم ۴ رتبه‌ای نشان دهنده ضریب weighted kappa ۰/۵ با خطای استاندارد ۰/۰۷ و فاصله اطمینانی ۰/۹۵ در محدوده ۰/۳۶ تا ۰/۶۴ بود و این نشان دهنده سطح متوسطی از توافق بین دو مدرس در به کارگیری سیستم ۴ رتبه‌ای برای ارزیابی عملکرد دانشجویان دندانپزشکی است.

جدول ۱. فراوانی نمرات داده شده در سیستم ۴ رتبه‌ای به تفکیک دو مدرس

مدرس B					
کل	D	C	B	A	
۲۷ (۰/۳۶/۰)	۰	۳ (۰/۴/۰)	۱۲ (۰/۱۶/۰)	۱۲ (۰/۱۶/۰)	A
۲۴ (۰/۳۲/۰)	۴ (۰/۵/۳)	۵ (۰/۶/۶)	۱۳ (۰/۱۷/۳)	۲ (۰/۲/۶)	B
۱۲ (۰/۱۶/۰)	۲ (۰/۲/۶)	۶ (۰/۸/۰)	۴ (۰/۵/۳)	۰	C
۱۲ (۰/۱۶/۰)	۸ (۰/۱۰/۶)	۳ (۰/۴/۰)	۱ (۰/۱/۳)	۰	D
۷۵ (۰/۱۰۰/۰)	۱۴ (۰/۱۸/۶)	۱۷ (۰/۲۲/۶)	۳۰ (۰/۴۰/۰)	۱۴ (۰/۱۸/۶)	کل

مدرس A

جدول ۲. توافق بین ارزیاب برای دو روش (سیستم ۴ رتبه‌ای در مقابل سیستم ۲۰ نمره‌ای)

سیستم ۴ رتبه‌ای	سیستم ۲۰ نمره‌ای	
۰/۵۰	۰/۴۷	ضریب توافق
متوسط	متوسط	سطح توافق
۰/۳۶-۰/۶۴	۰/۳۶-۰/۶۴	فاصله اطمینانی ۰/۹۵

در نمودار Bland-Altman (شکل ۱) تفاوت نمره میانگین هر دانشجو در دو سیستم نمره‌دهی در برابر میانگین نمره مایخوده در دو سیستم، به تفکیک دانشجویان و مدرسین ترسیم شده (شکل ۱). حداکثر اختلاف بین سیستم ۴ رتبه‌ای کمی شده (converted to 20-point)، و نمره اخذ شده از سیستم ۲۰ نمره‌ای همواره کمتر از ۲ نمره و عمدتاً در محدوده ۱- تا ۱+ نمره اختلاف بوده است.

ضریب همبستگی درون‌رده‌ای (ICC) برای ارزیابی توافق بین دو عضو هیئت علمی در استفاده از سیستم ۲۰ نمره‌ای برابر با ۰/۴۷ با فاصله اطمینان ۰/۹۵ در بازه ۰/۳۶ تا ۰/۶۴ به دست آمد که در محدوده توافق نزدیک به متوسط قابل تفسیر است (جدول ۲).

مقایسه توافق بین روش‌ها:

معدل هر دانشجو در هر دو سیستم ۴ رتبه‌ای (بعد از عددگذاری متناظر) و سیستم ۲۰ نمره‌ای محاسبه شد (جدول ۳). توافق بین نمرات سیستم ۴ رتبه‌ای و نمرات سیستم ۲۰ نمره‌ای برای هر ارزیاب با استفاده از ضریب

جدول ۳. معدل نمره تکالیف هر دانشجو به تفکیک سیستم نمره دهی و به تفکیک مدرسین

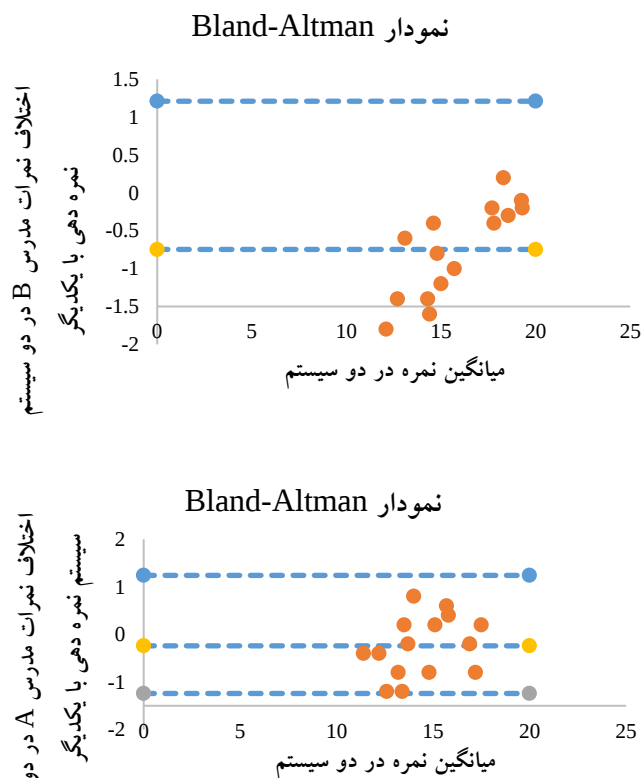
کمى شده سیستم چهار رتبه‌ای*	سیستم بیست نمره‌ای	کمى شده سیستم چهار رتبه‌ای*	سیستم بیست نمره‌ای	کمى شده سیستم چهار رتبه‌ای*	سیستم بیست نمره‌ای
۱۲/۸۰	۱۳/۵۰	۱۲/۸۰	۱۳/۶۰	۱۲/۸۰	۱۳/۴۰
۱۲/۸۰	۱۳/۷۰	۱۲/۰۰	۱۲/۴۰	۱۳/۶۰	۱۵/۰۰
۱۶/۸۰	۱۶/۷۰	۱۶/۰۰	۱۵/۴۰	۱۷/۶۰	۱۸/۰۰
۱۴/۴۰	۱۴/۲۰	۱۴/۴۰	۱۳/۶۰	۱۴/۴۰	۱۴/۸۰
۱۸/۲۰	۱۸/۴۰	۱۷/۶۰	۱۷/۴۰	۱۹/۲۰	۱۹/۴۰
۱۴/۲۰	۱۴/۸۰	۱۳/۶۰	۱۳/۴۰	۱۵/۲۰	۱۶/۲۰
۱۶/۸۰	۱۶/۸۰	۱۵/۲۰	۱۵/۰۰	۱۸/۴۰	۱۸/۲۰
۱۱/۲۰	۱۲/۳۰	۱۱/۲۰	۱۱/۶۰	۱۱/۲۰	۱۳/۰۰
۱۷/۲۰	۱۶/۲۵	۱۶/۰۰	۱۵/۶۰	۱۸/۴۰	۱۸/۷۰
۱۳/۶۰	۱۴/۶۰	۱۲/۸۰	۱۴/۰۰	۱۴/۴۰	۱۵/۲۰
۱۲/۰۰	۱۳/۳۰	۱۲/۰۰	۱۳/۲۰	۱۲/۰۰	۱۳/۴۰
۱۴/۴۰	۱۵/۴۰	۱۴/۴۰	۱۵/۲۰	۱۴/۴۰	۱۵/۶۰
۱۳/۶۰	۱۴/۵۰	۱۳/۶۰	۱۳/۸۰	۱۳/۶۰	۱۵/۲۰
۱۷/۲۰	۱۷/۷۰	۱۶/۸۰	۱۷/۶۰	۱۷/۶۰	۱۷/۸۰
۱۸/۰۰	۱۸/۱۵	۱۶/۸۰	۱۷/۰۰	۱۹/۲۰	۱۹/۳۰

* نمرات سیستم ۴ رتبه‌ای بر اساس الگوی زیر کمى (Recoded) شده‌اند: D=8, C=12, B=16, A=20

جدول ۴. میزان توافق بین دو روش به تفکیک مدرسین

میانگین نمره در سیستم ۴ رتبه‌ای*	میانگین نمره در سیستم ۲۰ نمره‌ای	میانگین اختلاف بین دو متد (انحراف معیار)	ضریب توافق ICC بین نمرات دو روش (فاصله اطمینانی ۹۵٪)
مدرس A	۱۴/۳۵ (۱/۹۹)	۱۴/۵۹ (۱/۸)	۰/۵۶ (۰/۳۶)
مدرس B	۱۵/۴۶ (۲/۶۹)	۱۶/۲۱ (۲/۲)	۰/۷۷ (۰/۵۸)

* نمرات سیستم ۴ رتبه‌ای بر اساس الگوی زیر کمى (Recoded) شده‌اند: D=8, C=12, B=16, A=20



شکل ۱. نمودار Bland-Altman برای توافق بین نمره های سیستم بیست نمره ای و نمره های تبدیل شده (کمی شده) از سیستم ۴ رتبه ای به تفکیک دو مدرس A و B

بحث

از سوی دیگر، برای سیستم ۲۰ نمره ای، مقدار ICC برابر با ۰/۴۷ (۹۵% CI، ۰/۳۶-۰/۶۴) به دست آمد. با مقایسه این دو سیستم، دیده می شود که در سیستم ۴ رتبه ای، ضریب توافق بین مدرسین (inter-assessor) اندکی بالاتر بوده هر چند این تفاوت نه از نظر آماری و نه از نظر کاربردی معنی دار نمی باشد.

یکی از مزایای اصلی سیستم ۴ رتبه ای، سادگی و کاربردی بودن آن است. بدین معنی که برای مدرسین، تصمیم گیری و نمره دهی به یک دانشجوی از بین چهار رتبه، آسان تر و سریع تر از انتخاب نمره مناسب از بین ۲۱ نمره خواهد بود.^(۱۲) علی رغم داشتن سطوح کمتر، سیستم ۴ رتبه ای سطح خوبی از توافق را بین ناظران نشان داد و می تواند

ضریب کاپای وزنی بین دو عضو هیئت علمی در سیستم نمره دهی ۴ رتبه ای برابر با ۰/۰۵ بود که نشان دهنده ی توافق متوسط است.^(۱۰) بر این اساس، فرضیه صفر مبنی بر عدم وجود توافق بین دو سیستم نمره دهی رد شد. این یافته نشان می دهد که هر دو مدرس، برداشت نسبتاً مشابه و متناظری از سطوح سیستم ۴ رتبه ای در ارزیابی عملکرد دانشجویان داشته اند. همچنین، خطای استاندارد ۰/۰۷ بیانگر پراکندگی اندک در برآورد ضریب کاپا است و این موضوع، تکرارپذیری توافق میان دو مدرس را در سیستم ۴ رتبه ای تأیید می کند.^(۱۱)

ابزاری قابل اعتماد و تکرارپذیر برای ارزیابی عملکرد دانشجویان، به ویژه در دوره های عملی باشد.

سیستم آموزشی در کشورمان به طور سنتی از سیستم ۲۰ نمره‌ای استفاده می‌کند.^(۱۳) این سیستم بیش از یک قرن پیش از سیستم نمره دهی کشور فرانسه اقتباس شده است. این سیستم نمره دهی همچنین در بلژیک، لبنان، پرتغال (برای آموزش متوسطه و عالی) و گاهی در برخی از موسسات کشورهای دیگر مانند دانشگاه سنت اندروز در اسکاتلند استفاده می‌شود.^(۱۴، ۱۵)

تقریباً یک قرن پیش، دکتر رنسیس لیکرت^(۱۶) سیستم ۵ رتبه ای را معرفی و گسترش داد. با این حال، این احتمال وجود دارد که سیستم لیکرت اگر ۷ رتبه یا بیشتر داشته باشد می‌تواند پایایی و اعتبار سنجش‌ها را افزایش دهد، همچنان که سیستم نمره دهی سنتی و متداول در کشورمان، دارای ۲۱ سطح (از نمره صفر تا ۲۰) می‌باشد و می‌تواند ارزیابی دقیق تری از عملکرد فراگیران ارائه داده و امکان بازخورد بهتری را فراهم می‌کند. در این مطالعه، مقادیر ICC برای سیستم ۲۰ نمره‌ای نشان دهنده توافق در محدوده متوسط بود و با مقادیر متناظر به دست آمده برای سیستم ۴ رتبه‌ای قابل مقایسه بود. لازم به ذکر است که برخی از محققان استدلال می‌کنند که افراد نمی‌توانند، در آن واحد و به طور همزمان، بیش از هفت گزینه متمایز را در مورد یک موضوع در ذهن تمایز دهند و ضرائب پایایی در فراتر از هفت سطح بهبود نمی‌یابد. در یک مطالعه عنوان شد که در مورد میزان اطلاعاتی که می‌توانیم همزمان پردازش کنیم محدودیت‌هایی وجود دارد.^(۱۷) طبق این نظریه، ما فقط می‌توانیم حدود هفت دسته مختلف را تشخیص دهیم، حدود هفت مورد را در حافظه کوتاه مدت به خاطر بسپاریم و همزمان

روی حدود شش چیز تمرکز کنیم. این بدان معنی است که استفاده از بیش از شش یا هفت گزینه ممکن است موثر نباشد، زیرا فراتر از توانایی‌های شناختی ما بوده و نیاز به تلاش ذهنی بیشتر و تولید داده‌های دقیق‌تر دارد.^(۱۸)

کاهش سطوح نمره دهی به صورت قبول/مردود در خیلی از مطالعات مطرح شده و از آن دفاع شده است. برای مثال Bloodgood و همکاران^(۱۹)، نشان دادند این میزان سطوح نمره دهی در کاهش اضطراب دانشجویان و نیز بدون تاثیر بر عملکرد نهایی آنها در آزمون USMLE موثر است. در چند مطالعه دیگر همین نتیجه بدست آمده و حتی در یک مرور سیستماتیک این مساله تایید شده است.^(۲۰-۲۳) این میزان تغییر در سطوح نمره دهی به نظر نویسندگان این مطالعه باید با احتیاط بیشتری انجام شود و کاهش سطوح نمره دهی ابتدا در سطوح کمتر بررسی و ارزیابی شود. در واقع گذار شتاب زده به نمره دهی دو سطحی که تعداد بسیار زیاد مطالعات منجر به مرور سیستماتیک هم همین را نشان می‌دهند، باید با ارزیابی‌های محتاطانه تری نظیر مطالعه حاضر کنترل شود.

Taherdoost و همکاران^(۲۴)، نشان دادند که سیستم هفت سطحی به نظر می‌رسد متضمن بهترین نتیجه باشد. مطالعه تأثیرگذار دیگری توسط Revilla و همکاران با کار بر روی داده‌های چهار آزمایش نظرسنجی اجتماعی اروپا (European Social Survey experiments) نشان داد که پاسخ دهندگان هنگام پاسخ دادن به نظرسنجی‌های طولانی که هر آیت از تعداد سطوح زیادی برخوردار است، دچار افت و کاهش توجه و تمرکز می‌شوند و پیشنهاد کردند که یک سوال ۵ سطحی بهتر از ۷ یا ۱۱ سطح است.^(۲۵)

داشتند (یک نمره در بیست نمره). این امر به ویژه برای ارزیابی‌هایی که در آنها فراگیران به جای یک مورد در چندین ماژول مورد ارزیابی قرار میگیرند، مفید خواهد بود (مانند این مطالعه که در آن از ۵ ماژول و تکلیف عملی برای ارزیابی دانشجویان استفاده شده بود). وجود چند ماژول در یک ارزیابی، با استفاده از اصل برگشت به میانگین (regression toward the mean) به کاهش نوسانات در نمرات کمک می‌کند و در نهایت منجر به ارزیابی معتبرتری می‌شود. به طور کلی، ارزیابی‌هایی با ماژول‌هایی دارای طراحی بهینه، قابل اعتمادتر بوده و تصویر دقیق‌تری از یادگیری فراگیران ارائه می‌دهند.^(۲۶)

یک نقطه قوت کلیدی این مطالعه، استفاده از روش‌های آماری قوی برای ارزیابی توافق بین روش‌های اندازه‌گیری بود. محققین این زمینه اغلب مایل اند که نتایج حاصل از دو (یا بیشتر) مدرس یا تکنیک مختلف را با هم مقایسه نمایند. روش‌های ساده مانند روش‌های مورد استفاده برای تجزیه و تحلیل جداول توافقی ۲x۲ (برای داده‌های کیفی) و یا همبستگی (در مورد داده‌های کمی) جهت این منظور روشهای معمول هستند، اما محدودیت‌های بارزی نیز دارند.^(۲۷) در این مطالعه، با استفاده از آزمون‌های توافق kappa و ICC، که درک دقیق‌تری از میزان توافق بین روش‌ها را ارائه می‌دهند، از این محدودیت‌ها اجتناب کردیم.

ضمناً برای به دست آوردن تصویر جامع‌تری از الگو، مقدار و جهت توافق بین دو روش نمره‌دهی، در کنار آزمون‌های توافق kappa و ICC، از نمودار بلاند-آلتمن استفاده نمودیم. این نمودار به طور ساده و بصری کمک کرد تا سوگیری سیستماتیک (تفاوت میانگین) بین روش‌ها را بررسی و مشاهده کنیم. این نمودار همچنین دامنه (محدوده توافق) را به صورت ۹۵٪ تفاوت‌های حاصل از روش دوم در قیاس

در مطالعه دیگری، محققان تأثیر گزینه‌های پاسخ بر اثربخشی نظرسنجی‌ها را بررسی کردند. شرکت کنندگان از پرسشنامه‌ها برای نظردهی در خصوص اجزای ارائه خدمات در یک فروشگاه یا در یک رستوران استفاده کردند. این پرسشنامه‌ها دارای سوالات متنوعی با تعداد متفاوتی از گزینه‌ها برای انتخاب (از ۲ الی ۱۱ گزینه) بودند. نشان داده شد که در حالی که همه انواع سوالات (از لحاظ تعداد گزینه‌ها) انسجام داخلی مناسبی را نشان می‌دادند، در آنهایی که بیش از ۱۰ گزینه پاسخ داشتند، در طول زمان ثبات پاسخ کمتری مشاهده گردید و نتیجه‌گیری شد برای ارتقای تکرارپذیری پاسخ، بهتر است تعداد سطوح سنجش زیر ۱۰ مورد نگاه داشته شود.^(۱۸)

به طور کلی، در این مطالعه، هر دو سیستم نمره‌دهی، سطح معقولی از تکرارپذیری و توافق (Internal consistency) در ارزیابی مهارت‌های ارتودنسی عملی دانشجویان دندانپزشکی نشان دادند. انتخاب بین سیستم ۴ رتبه‌ای و یا سیستم ۲۰ نمره‌ای به زمینه آموزشی (context)، اهداف ارزیابی و امکانات و منابع موجود بستگی دارد. مطالعه ما تفاوت‌های اندکی بین سیستم‌های ۲۰ نمره‌ای و ۴ رتبه‌ای نشان داد و سیستم ۴ رتبه‌ای می‌تواند به خاطر سادگی و سهولت کاربردی‌اش، به ویژه در سناریوهایی که ارزیابی سریع و کارآمد ضروری است، با اطمینان خاطر مورد ترجیح مدرسین واقع شود. از سوی دیگر، سیستم ۲۰ نمره‌ای ارزیابی‌های دقیق‌تر و با تقسیمات ظریف‌تری را ارائه می‌دهد که آن را برای موقعیت‌هایی که ارزیابی‌های جامع و تمیز بیشتر مورد نیاز است، مناسب می‌سازد. قابل توجه است که این نتیجه‌گیری در نمودار بلاند-آلتمن (شکل ۱) نیز مورد تأیید قرار گرفت. در این نمودار مشاهده می‌شود که نقاط نشان دهنده توافق بین دو سیستم نمره‌دهی به طور کلی در محدوده یک نمره از یکدیگر اختلاف

نتیجه‌گیری

هر دو سیستم نمره‌دهی، پایایی مناسبی در ارزیابی مهارت‌های ارتودنسی عملی دانشجویان دندانپزشکی نشان دادند. انتخاب بین سیستم چهار رتبه‌ای و یا سیستم ۲۰ نمره‌ای به زمینه آموزشی (context) و اهداف ارزیابی بستگی دارد. سیستم ۲۰ نمره‌ای، ارزیابی دقیق‌تری از عملکرد دانشجویان، همراه با تکرارپذیری متوسط تا بالا ارائه می‌دهد و در عین حال، برای نمره‌دهی دقیق، نیاز به تمرکز و زمان دارد. سیستم چهار رتبه‌ای می‌تواند انتخاب اعضای هیات علمی در مواقعی که اولویت با سادگی و سهولت اجرا است، باشد و سیستم ۲۰ نمره‌ای در ارزیابی‌های نیازمند جزئیات و تمیز بیشتر ترجیح داده شود.

تشکر و قدردانی

این تحقیق هیچ‌گونه کمک مالی از ارگان‌های دانشگاهی، تجاری یا غیرانتفاعی دریافت نکرده است. نویسندگان از تمامی دانشجویان شرکت‌کننده صمیمانه تقدیر می‌نمایند.

تضاد و منافع

هیچ موردی برای اعلام وجود ندارد.

با روش اول نشان می‌دهد.^(۲۸) با این حال، ضروری است توجه شود که نمودار بلاند-آلتمن فقط فاصله‌های بین روشها را نشان می‌دهد و تفسیر این که آیا این اختلاف، از نظر کاربردی قابل قبول می‌باشد یا خیر را بر عهده خواننده می‌گذارد. در این مطالعه، اختلاف بیش از ۱-۰/۷۵ نمره (از بیست نمره) به عنوان اختلاف معنی‌دار کاربردی مفروض شد.

تحقیقات بیشتری برای بررسی قابلیت تعمیم این یافته‌ها به سایر فیلدهای آموزش دندانپزشکی و مقایسه عملکرد سیستم ۲۰ نمره‌ای با سایر روش‌های ارزیابی ضروری است. از آنجا که مقدار ICC پیش از مطالعه و در مرحله طراحی و محاسبه حجم نمونه قابل پیش‌بینی نبود، محاسبه حجم نمونه بر اساس بهترین برآورد و شواهد موجود انجام شد. پایین‌تر بودن ICC مشاهده‌شده نسبت به مقدار مورد انتظار ممکن است به دلیل کاهش توان واقعی مطالعه نسبت به توان برنامه‌ریزی‌شده باشد. بنابراین، ضرورت دارد که عدم مشاهده تفاوت آماری معنی‌دار بین دو سیستم نمره‌دهی با احتیاط تفسیر شود و نیز انجام مطالعات با حجم نمونه بالاتر برای تأیید یافته‌های حاضر پیشنهاد می‌گردد.

منابع

1. Baker A, Young K, Potter J, Madan I. A review of grading systems for evidence-based guidelines produced by medical specialties. Clin Med 2010;10(4):358-63.
2. McMorran C, Ragupathi K, Luo S. Assessment and learning without grades? Motivations and concerns with implementing gradeless learning in higher education. Assess Eval High Educ 2017 3;42(3):361-77.
3. McMorran C, Ragupathi K. The promise and pitfalls of gradeless learning: responses to an alternative approach to grading. J Furth High Educ 2020 8;44(7):925-38.
4. Nepal KP. An approach to assign individual marks from a team mark: the case of Australian grading system at universities. Assess Eval High Educ 2012 ;37(5):555-62.

5. Mukhtoraliyeva Mukhtasar, Sharopov Begyor, Kazadayev Aleksandr, Dedaxanov Farrukh, Umarov Isroil, Khakimov Sodiqjon, et al. Analysis of the effectiveness of the development of the German education system in our country. *J New Century Innov* 2022;18(1):168–73.
6. Atkinson A, Burgess S, Croxson B, Gregg P, Propper C, Slater H, et al. Evaluating the impact of performance-related pay for teachers in England. *Labour Econ* 2009;16(3):251–61.
7. Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. *Acad Med* 2012;87(8):1070–6.
8. Ramesh D, Sanampudi SK. An automated essay scoring system: a systematic literature review. *Artif Intell Rev* 2022;55(3):2495–527.
9. Isaacson JJ, Stacy AS. Rubrics for clinical evaluation: objectifying the subjective experience. *Nurse Educ Pract* 2009;(2):134–40.
10. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15(2):155–63.
11. Smith G. Simple regression. In: *Essential Statistics, Regression, and Econometrics*. 2nd ed. Amsterdam: Academic Press; 2015. P. 219–59.
12. Devlin SJ, Dong HK, Brown M. Selecting a scale for measuring quality. *Mark Res* 1993;5(3):12–17.
13. Dolton P, Marcenaro-Gutierrez O, Navarro L. The effective use of student achievement data: International comparisons and grading practices. *Educ Econ* 2013;21(5):502–18.
14. Kirsch M, Beernaert Y, Nørgaard B. The French higher education system and the Bologna Process. *Eurn J Educ* 2003;38(3):261–78.
15. Wagenaar R. Educational structures, learning outcomes, workload and the calculation of ECTS credits. *Tuning J High Educ* 2014;1(2):223–46.
16. Sullivan GM, Artino AR. Analyzing and interpreting data from Likert-type scales. *J Grad Med Educ* 2013;5(4):541–2.
17. Krishna A. An integrative review of sensory marketing: Engaging the senses to affect perception, judgment and behavior. *J Consum Psychol* 2012;22(3):332–51.
18. Preston CC, Colman AM. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychol (Amst)* 2000;104(1):1–15.
19. Bloodgood RA, Short JG, Jackson JM, Martindale JR. A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Acad Med* 2009;84(5):655–62.
20. Rohe DE, Barrier PA, Clark MM, Cook DA, Vickers KS, Decker PA. The benefits of pass-fail grading on stress, mood, and group cohesion in medical students. *Mayo Clin Proc* 2006;81(11):1443–8.
21. Spring L, Robillard D, Gehlbach L, Simas TA. Impact of pass/fail grading on medical students' well-being and academic outcomes: a systematic review. *Med Educ* 2011;45(9):867–77.
22. Modrau B, Christensen AG, Norregaard O. Pass/fail versus tiered grades and academic performance in undergraduate medical education: A crossover study. *JMIR Med Educ* 2025;11:e74975.

23. Iyer AA, Hayes C, Chang BS, Farrell SE, Fladger A, Hauer KE, et al. Should medical school grading be tiered or pass/fail? A scoping review of conceptual arguments and empirical data. *Acad Med* 2025;100(8):975-85.
24. Taherdoost H. What is the best response scale for survey and questionnaire design; review of different lengths of rating scale / attitude, scale / Likert scale. *Int J Acad Res Manag* 2019;8(1):1–10.
25. Revilla MA, Saris WE, Krosnick JA. Choosing the number of categories in agree–disagree scales. *Sociol Methods Res* 2013;43(1):73–97.
26. de Vet HCW, Mokkink LB, Mosmuller DG, Terwee CB. Spearman–Brown prophecy formula and Cronbach’s alpha: different faces of reliability and opportunities for new applications. *J Clin Epidemiol* 2017;85: 45–9.
27. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Measures of agreement. *Perspect Clin Res* 2017;8(4):187–91.
28. Giavarina D. Understanding Bland Altman analysis. *Biochem Med (Zagreb)* 2015;25(2):141–51.