تحلیل کیفیت آزمون عینی ساختارمند بالینی در دانشگاه علوم پزشکی تهران

سارا مرتاض هجری، محمد جلیلی*

چکیده

مقدمه: با رواج آزمون‌های مبتنی بر عملکرد، استفاده از شاخص‌های مناسب برای اطمینان از کیفیت آنها ضروری است. این مطالعه به بررسی متریک‌های آزمون عینی ساختارمند بالینی (Objective Structured Clinical Examination) می‌پردازد.

روش‌ها: در این مطالعه توصیفی مقطعی درده ایستگاه آزمون OSCE پیش‌کارورزی سال ۹۱ دانشگاه علوم پزشکی، نمره حدنصاب به روش رگرسیون مرزی تعیین شد و میزان ردی محاسبه گردید. ریشه میانگین مربعات خطا (Root Mean Square Error) برای تخمین میزان خطای حدنصاب به کار رفت. ثبات درونی آزمون با آلفای کرونباخ محاسبه و سپس مقدار خطای استاندارد اندازه‌گیری (Standard Error of Measurement) تعیین شد. برای هر ایستگاه شاخص‌های آلفا در صورت حذف، R2، تمایز بین درجات، دشواری و تمیز به دست آمد.

نتایج: در این آزمون۲۶۶ نفر از دانشجویان شرکت نمودند.نمره حدنصاب کل ۵۲/۵۵ از ۱۰۰ به دست آمد. ۴ نفر (۱/۵ درصد) در آزمون رد شدند و RMSE معادل ۰/۴۵ بود. نمره SEM ۴/۸۴و آلفای کرونباخ آزمون ۰/۷۰ محاسبه شدند که در صورت حذف هر یک از ایستگاه‌ها بین ۰/۶۴ تا ۰/۷۰ متغیر بود. مقدار R^2 از ۰/۱۶ تا ۰/۸۵ به دست آمد. شاخص تمایز بین درجات از ۰/۶۶ تا ۱/۹۳ بود. دامنه سختی و تمیز به ترتیب ۰/۸۹-۰/۷۱ و ۰/۴۴-۰/۱۲ بود.

نتیجه‌گیری: ثبات درونی، خطای آزمون و خطای حدنصاب در حد قابل قبول بود. تمام ایستگاه‌ها از لحاظ آلفا در صورت حذف آیتم مطلوب بودند. R^2 در دو ایستگاه از حد مطلوب پایین‌تر بود. تمایز بین درجات به غیر از یک ایستگاه خوب بود. ایستگاه‌ها نه چندان دشوار با قابلیت تمیز نه چندان بالا بودند که در آزمون معیاری مطلوب محسوب می‌شود.

واژه‌های کلیدی: تحلیل آزمون، سایکومتریک، آزمون عینی ساختارمند بالینی

مجله ایرانی آموزش در علوم پزشکی/ ۱۳۹۵؛ ۱۷(۶): ۵۰ تا ۵۹

مقدمه

پس از برگزاری آزمون، ضروری است که تحلیل و ارزشیابی آن صورت گیرد تا برگزارکنندگان و شرکت‌کنندگان از کیفیت مناسب آزمون مطمئن گردند. معمولاً هنگامی که از تحلیل آزمون و محاسبه شاخص‌های سایکومتریک صحبت می‌شود، امتحانات چندگزینه‌ای به ذهن متبادر می‌شوند. این سؤالات به علت سهولت تصحیح به صورت رایج در مقاطع و رشته‌های مختلف مورد استفاده قرار می‌گیرند و به همان ترتیب، روش‌های تحلیل مربوط به آنها مانند ضریب دشواری و ضریب تمیز به راحتی در دسترس عموم دست‌اندرکاران ارزیابی قرار دارد. طی چند سال اخیر با توجه به این موضوع که امتحانات چندگزینه‌ای به تنهایی نمی‌توانند وضعیت دانشجو را به صورت کامل و همه جانبه مورد سنجش قرار دهند، استفاده از آزمون‌های مبتنی بر عملکرد (Performance based assessments) به درستی مورد توجه قرار گرفته است(۱). این در حالی است که همزمان با رواج این آزمون‌ها در کشور، استفاده از شاخص‌های سایکومتریک مناسب به منظور تحلیل آنها چندان مورد توجه نبوده است.

* **نویسنده مسؤول:** دکتر محمد جلیلی (استاد)، گروه طب اورژانس و گروه آموزش پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران. mjalili@tums.ac.ir

دکتر سارا مرتاض هجری (استادیار)، گروه آموزش پزشکی و مرکز تحقیقات آموزش علوم پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران. (smortaz@tums.ac.ir)

تاریخ دریافت مقاله: ۹۵/۱/۱۴، تاریخ اصلاحیه: ۹۵/۸/۸، تاریخ پذیرش: ۹۵/۹/۲۸

در حوزه علوم پزشکی، یکی از روش‌های ارزیابی که طی چند سال اخیر به شدت مورد توجه قرار گرفته است و در مقاطع و رشته‌های مختلف برای سنجش مهارت‌های عملی و صلاحیت بالینی (Clinical competence) فراگیران به کار می‌رود، آزمون عینی ساختارمند بالینی (Objective structured Clinical Examination) است. این آزمون از چند ایستگاه تشکیل شده است که دانشجویان به ترتیب آنها را پشت سر می‌گذارند و در هر ایستگاه طی زمان ثابتی باید وظیفه بالینی مشخصی مانند اخذ شرح حال، انجام معاینه، اجرای پروسیجر را در مواجهه با بیمار استاندارد، مانکن یا مولاژ به انجام برسانند. طی این مدت آزمونگر مستقر در ایستگاه به صورت مستقیم عملکرد دانشجو را مشاهده می‌کند و طبق چک‌لیست‌های از پیش تدوین شده به ارزیابی وی می‌پردازد(۲).

غالباً OSCEهای برگزار شده در کشور یا اساساً مورد تحلیل قرار نمی‌گیرند یا صرفاً به شاخص پایایی که اطلاعاتی در مورد کل آزمون به دست می‌دهد، بسنده می‌شود(۳). در حالی که برای تحلیل OSCE تنها استفاده از یک شاخص کافی نیست و شاخص‌های دیگری وجود دارند که می‌توانند در کنار یکدیگر اطلاعات خوبی در خصوص کل امتحان و همچنین تک‌تک ایستگاه‌های آن ارائه دهند(۴).

با توجه به اهمیت تحلیل آزمون‌های مبتنی بر عملکرد و لزوم ارائه راه‌کارهای پیشنهادی مناسب برای اصلاح آزمون، این مطالعه قصد دارد شاخص‌های سایکومتریک مربوط به OSCE، در سطح هر ایستگاه و در سطح کل آزمون را محاسبه نماید و بر اساس یافته‌ها به بررسی نقاط قوت و ضعف آزمون بپردازد.

روش‌ها

این مطالعه از نوع کمی و توصیفی تحلیلی است و به تحلیل و بررسی آزمون OSCE پیش‌کارورزی اسفند ۱۳۹۱ دانشگاه علوم پزشکی تهران می‌پردازد.

آزمونOSCE پیش‌کارورزی در دانشگاه علوم پزشکی تهران هر ساله برای دانشجویان پزشکی که مقطع کارآموزی را به اتمام رسانده‌اند و قرار است وارد دوره کارورزی شوند، همزمان با آزمون کتبی پیش‌کاروزی برگزار می‌شود. در این آزمون۲۶۶ نفر از دانشجویان شرکت نمودند. OSCE سال ۱۳۹۱ از ۱۳ ایستگاه تشکیل شده بود. سه ایستگاه (شامل تفسیر کلیشه رادیولوژی، معاینه ته چشم و نورولوژی) بدون آزمونگر مستقیم و نمره گلوبال طراحی شده بود؛ از آنجا که نحوه نمره‌دهی آنها مانند سؤالات تشریحی بود، این سه ایستگاه از مطالعه خارج شدند. به این ترتیب اطلاعات ده ایستگاه اطفال، قلب، احیا، تست توبرکولین، پوست، زنان، روانپزشکی، بخیه، فشارخون و ارتوپدی وارد مطالعه شد. دانشجویان در هر یک از ایستگاه‌ها پنج دقیقه فرصت داشتند تا به انجام وظیفه محوله بپردازند. تمام دانشجویانی که در آزمون شرکت کردند، وارد مطالعه شدند. نمرات ایشان بدون نام و با کد مورد تجزیه و تحلیل قرار گرفت و تا انتها نزد نویسندگان محفوظ ماند.

نحوه نمره‌دهی آزمون به این صورت بود که در هر ایستگاه یک آزمونگر مستقر بود که از وی خواسته شده بود تا ضمن تکمیل چک‌لیست مربوطه برای هر دانشجو، عملکرد کلی دانشجو را نیز به صورت گلوبال و فارغ از نمره چک‌لیست در مقیاس لیکرت ۵تایی (رد، مرزی، قابل قبول، خوب، عالی) ارزیابی کند. برای سهولت محاسبه با توجه به این که نمرات چک‌لیست ایستگاه‌های مختلف با یکدیگر برابر نبود، نمرات تمام ایستگاه‌ها از ۱۰ حساب شد و نمره دانشجو در آزمون از حاصل جمع نمرات چک‌لیست وی در تمام ایستگاه‌ها به دست آمد. برای کاهش احتمال تأثیرپذیری نمره گلوبال از نمره چک‌لیست، قبل از آزمون آموزش لازم در خصوص لزوم مستقل بودن دو نمره هم به صورت جمعی و هم به صورت انفرادی برای آزمونگران توضیح داده شد. نمره گلوبال صرفاً برای تعیین حدنصاب آزمون و تعیین شاخص‌های سایکومتریک مورد استفاده قرار گرفت.

تعیین حدنصاب آزمون

در اکثر آزمون‌ها حدنصاب قبولی با در نظرگرفتن ۵۰ یا ۷۰ درصد نمره کل تعیین می‌شود اما چون امور مهمی مانند درجه سختی سؤالات، سطح دانشجویان و هدف آزمون در

این روش لحاظ نمی‌شوند، توصیه می‌شود از روش‌های علمی تعیین استاندارد (Standard setting) برای آزمون‌های مهم که نتایج سرنوشت‌سازی دارند، استفاده شود. یکی از تکنیک‌های رایج برای تعیین نمره قبولی در آزمون‌های مبتنی بر عملکرد، روش رگرسیون مرزی (Borderline Regression Method) است. برای انجام این روش، برای هر ایستگاه یک معادله رگرسیون طراحی شد. به این ترتیب که نمرات چک‌لیست به عنوان متغیر وابسته و نمرات گلوبال به عنوان متغیر مستقل در نظر گرفته شد(۵). سپس با قرار دادن نمره گلوبال مربوط به دانشجوی مرزی در معادله (۲ در مقیاس لیکرت)، حدنصاب ایستگاه برآورد شد. حدنصاب کل آزمون با پیش گرفتن رویکرد جبرانی (compensatory) از جمع حدنصاب ایستگاه‌ها به دست آمد.

همچنین برای تخمین میزان خطای روش تعیین حدنصاب در کل آزمون از ریشه میانگین مربعات خطا (Root Mean Square Error) طبق فرمول زیر استفاده شد(۶). در این فرمول M تعداد ایستگاه‌ها، n تعداد دانشجویان و Sregr, i انحراف معیار معادله رگرسیون برای ایستگاه i هستند. همچنین MeanG,i و SDG,I به ترتیب میانگین و انحراف معیار نمرات گلوبال برای ایستگاه i هستند. G0 نیز نمره گلوبالی است که برای دانشجوی مرزی در نظر گرفته شده و در تمام ایستگاه‌های این مطالعه یکسان و معادل ۲ بوده است.

$$RMSE_{OSCE} = \sqrt{\frac{1}{M^2}\cdot\frac{1}{n}\cdot\sum_{i=1}^{M}\left\{s^2_{regr,i}\cdot\left(1+\frac{(G_0-Mean_{G,i})^2}{[(n-1)/n]\cdot SD^2_{G,i}}\right)\right\}}$$

شاخص‌های سایکومتریک کل آزمون

در ابتدا شاخص‌های مرکزی و پراکندگی شامل میانگین، میانه، نما و انحراف معیار برای کل آزمون محاسبه شد. همچنین با توجه به حدنصاب کل آزمون، میزان ردی در کل آزمون محاسبه شد. پایایی آزمون از طریق محاسبه ضریب آلفای کرونباخ محاسبه شد. در ادامه، مقدار خطای استاندارد اندازه‌گیری (Standard Error of Measurement) برای کل آزمون از طریق فرمول زیر تعیین شد(۷):

$$SEM = \text{Standard Deviation}\sqrt{1-\text{reliability}}$$

شاخص‌های سایکومتریک ایستگاه‌ها

میانگین، میانه، نما، انحراف معیار و میزان ردی برای هر یک از ایستگاه‌ها محاسبه شد. شاخص آلفا در صورت حذف هر یک از ایستگاه‌ها (Alpha if item deleted) نیز تعیین شد. همچنین از معادله رگرسیونی که برای هر ایستگاه طراحی شده بود، دو شاخص شامل R2 و شاخص تمایز بین درجات (Intergrade discrimination) به دست آمد.

ضریب دشواری با در نظر گرفتن ایستگاه به صورت یک سؤال تشریحی ۱۰ نمره‌ای به دست آمد. بنابراین نمره میانگین ایستگاه بر نمره کل آن (نمره ۱۰) تقسیم شد. برای محاسبه ضریب تمیز ایستگاه نیز از فرمول مربوط به سؤالات تشریحی استفاده شد و از آنجا که OSCE آزمونی برای تعیین میزان رسیدن دانشجویان به حد تسلط در حیطه مهارت‌های بالینی است، از فرمول مربوط به آزمون‌های معیارمحور (Criterion-referenced exams) استفاده شد. بنابراین در هر ایستگاه، تفاضل نمره میانگین دانشجویان در دو گروه رد و قبول کل آزمون بر نمره کل ایستگاه (نمره ۱۰) تقسیم شد.

آنالیز آماری

تجزیه و تحلیل داده‌ها با استفاده از نرم افزار SPSS 15 انجام شد. برای آنالیز نمرات شاخص‌های میانگین، نما، میانه و انحراف معیار بیان شدند. نتایج ردی به صورت تعداد و درصد ارائه شد. پایایی آزمون با استفاده از آلفای کرونباخ محاسبه شد. آنالیز رگرسیون خطی با در نظر گرفتن نمره چک‌لیست و نمره گلوبال به ترتیب به عنوان متغیر وابسته و مستقل انجام شد که بر اساس معادله رگرسیون علاوه بر پیش‌بینی نمره دانشجوی مرزی، ضریب R2 و شیب خط رگرسیون نیز محاسبه شد.

نتایج

تعداد کل دانشجویانی که وارد مطالعه شدند، ۲۶۶ نفر بود که از این میان ۱۷۰ نفر (۶۹/۴ درصد) دختر بودند. با توجه به

این که برای هر ایستگاه دو نمره چک‌لیست و گلوبال منظور شده بود در واقع برای هر دانشجو در ۱۰ ایستگاه ۲۰ حالت و در نتیجه برای ۲۶۶ دانشجو ۵۳۲۰ حالت ایجاد شد. از آنجا که برخی از دانشجویان در بعضی از ایستگاه‌ها نمره چک‌لیست یا گلوبال نداشتند، در این موارد که مجموعاً ۱۷ مورد از مجموع ۵۳۲۰ مورد بودند، نمره میانگین ایستگاه به عنوان نمره ایشان منظور شد.

نمره حدنصاب هر ایستگاه که با روش رگرسیون مرکزی محاسبه شد، در جدول ۱ نشان داده شده است. بالاترین حدنصاب مربوط به ایستگاه اطفال (۶/۷۰ نمره) و کم‌ترین حدنصاب مربوط به ایستگاه ارتوپدی (۴/۰۳ نمره) بود. نمره قبولی کل آزمون که از جمع نمرات قبولی ایستگاه‌ها به دست آمد از ۱۰۰ نمره ۵۲/۵۵ محاسبه شد. همچنین مقدار RMSE برای حدنصاب کل آزمون معادل ۰/۴۵ به دست آمد.

جدول ۱: توزیع نمرات ۲۶۶ دانشجو در ایستگاه‌های OSCE پیش‌کارورزی سال ۹۱

شماره ایستگاه	نام ایستگاه	حداقل نمره	حداکثر نمره	میانگین	میانه	نما	انحراف معیار	حدنصاب	میزان ردی (درصد)
۱	روانپزشکی	۰/۰۰	۱۰/۰۰	۷/۲۰	۷/۰۰	۸/۰۰	۱/۷۱	۴/۳۴	۱۹ (۷/۱)
۲	قلب	۳/۰۰	۱۰/۰۰	۸/۱۰	۸/۰۰	۸/۰۰	۱/۳۸	۵/۳۹	۱۷ (۶/۴)
۳	درماتیت	۳/۰۰	۱۰/۰۰	۷/۲۴	۷/۰۰	۷/۰۰	۱/۶۰	۵/۸۳	۴۲ (۱۵/۸)
۴	زنان	۳/۰۰	۱۰/۰۰	۷/۱۴	۷/۰۰	۷/۰۰	۱/۴۴	۴/۶۶	۱۱ (۴/۱)
۵	ارتوپدی	۰/۰۰	۱۰/۰۰	۸/۱۳	۹/۰۰	۹/۰۰	۲/۴۴	۴/۰۳	۲۴ (۹/۰)
۶	توبرکولین	۰/۰۰	۱۰/۰۰	۷/۶۵	۸/۰۰	۸/۰۰	۱/۷۰	۵/۴۸	۲۸ (۱۰/۵)
۷	احیا	۵/۰۰	۱۰/۰۰	۸/۹۵	۹/۰۰	۱۰/۰۰	۱/۲۶	۶/۱۸	۱۵ (۵/۶)
۸	بخیه	۱/۰۰	۱۰/۰۰	۷/۵۳	۸/۰۰	۸/۰۰	۲/۰۰	۵/۰۷	۴۴ (۱۶/۵)
۹	اطفال	۲/۰۰	۱۰/۰۰	۷/۶۷	۸/۰۰	۸/۰۰	۱/۷۲	۶/۷۰	۵۹ (۲۲/۲)
۱۰	فشارخون	۱/۰۰	۱۰/۰۰	۸/۰۰	۸/۰۰	۹/۰۰	۱/۷۰	۴/۸۷	۳ (۱/۱)
	کل آزمون	۴۸/۰۰	۹۳/۰۰	۷۷/۶۲	۷۹/۰۰	۷۹/۰۰	۸/۸۴	۵۲/۵۳	۴ (۱/۵)

نمره هر ایستگاه از ۱۰ و نمره کل آزمون از ۱۰۰ است.

حداقل و حداکثر نمره کل دانشجویان در OSCE از ۱۰۰ نمره به ترتیب ۴۸/۰۰ و ۹۳/۰۰ بود. میانگین نمرات دانشجویان در کل آزمون ۷۷/۶۲ با انحراف معیار ۸/۸۴ نمره به دست آمد. با محاسبه نمره کل تمام دانشجویان و با توجه به نمره حدنصاب کل مشخص شد که ۴ نفر (از ۲۶۶ نفر) در آزمون رد شدند یعنی میزان ردی ۱/۵ درصد بود. ضریب آلفای کرونباخ آزمون معادل ۰/۷۰ به دست آمد. با توجه این مقدار و همچنین انحراف معیار آزمون که معادل ۸/۸۴ نمره بود، SEM آزمون ۴/۸۴ نمره به دست آمد.

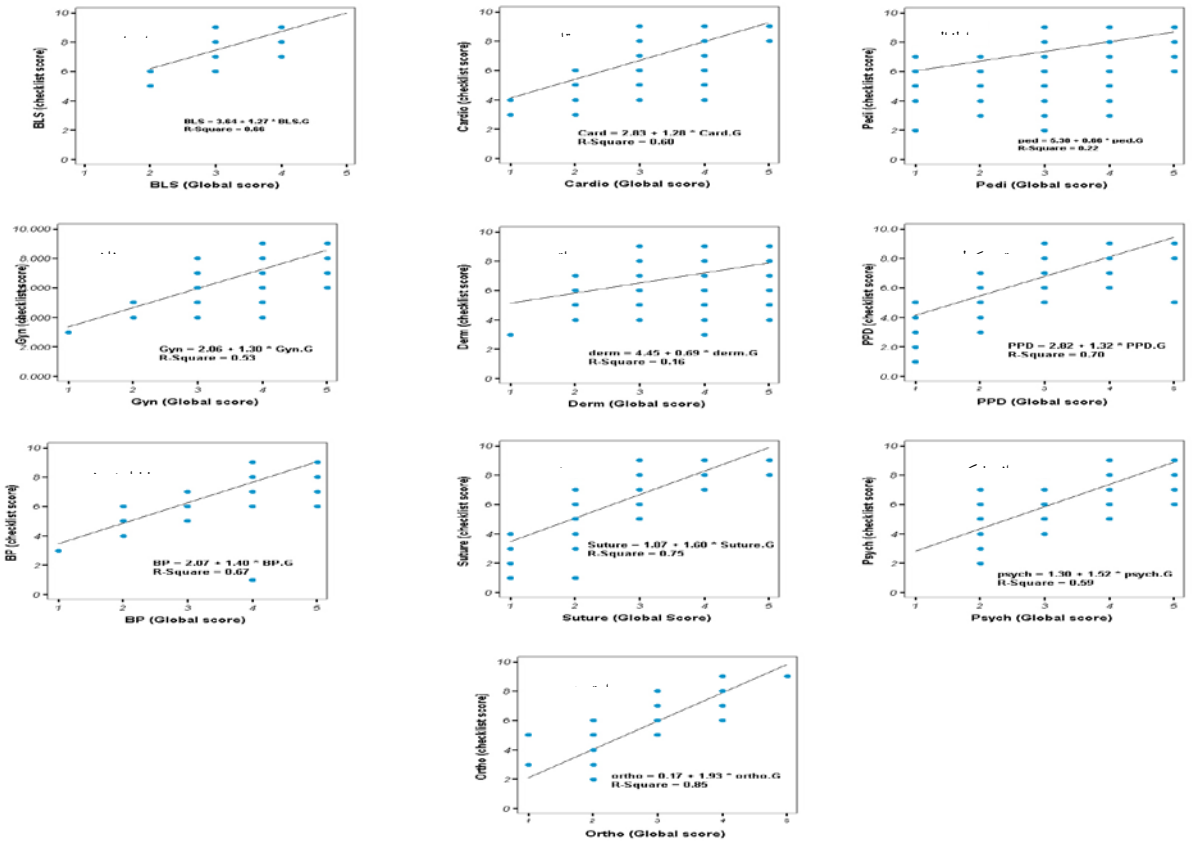
توزیع نمرات دانشجویان در هر یک از ایستگاه‌ها، نمره حد نصاب هر یک از ایستگاه‌ها و همچنین میزان ردی در آنها در جدول ۱ نشان داده شده است. پایین‌ترین نمره میانگین مربوط به ایستگاه زنان (۷/۱۴) و بالاترین نمره میانگین مربوط به ایستگاه احیا (۸/۹۵) بود. بالاترین درصد ردی مربوط به ایستگاه اطفال (۲۲/۲ درصد) و کم‌ترین میزان ردی مربوط به ایستگاه فشارخون (۱/۱ درصد) بود.

نمودار پراکندگی نمرات چک‌لیست در هر ایستگاه بر اساس نمرات گلوبال مربوطه در شکل ۱ و شاخص‌های سایکومتریک هر ایستگاه در جدول ۲ نمایش داده شده است. ضریب آلفای کرونباخ آزمون در صورت حذف هر یک از ایستگاه‌ها از ۰/۶۴ (ارتوپدی) تا ۰/۷۰ (زنان) متغیر بود. همچنین مقدار شاخص R^2 از مقدار ۰/۱۶ (درماتیت) تا میزان ۰/۸۵ (ارتوپدی) به دست آمد. شاخص تمایز بین درجات از ۰/۶۶ (اطفال) تا ۱/۹۳ (ارتوپدی) محاسبه شد. با محاسبه ضریب دشواری مشخص شد که سخت‌ترین و آسان‌ترین ایستگاه به ترتیب عبارت بودند از: معاینه زنان

(۰/۷۱) و احیا (۰/۸۹). پایین‌ترین و بالاترین ضریب تمیز مربوط بود به ایستگاه‌های زنان (۰/۱۲) و ارتوپدی (۰/۴۴).

جدول ۲: متریک‌های مربوط به هر یک از ایستگاه‌های OSCE پیش‌کارورزی سال ۹۱

شماره	نام ایستگاه	آلفا در صورت حذف ایستگاه	تمایز بین درجات	R^2	ضریب دشواری	ضریب تمیز
۱	روانپزشکی	۰/۶۶	۱/۵۲	۰/۵۹	۰/۷۳	۰/۲۰
۲	قلب	۰/۶۸	۱/۲۸	۰/۶۰	۰/۸۱	۰/۱۴
۳	درماتیت	۰/۶۹	۰/۶۹	۰/۱۶	۰/۷۲	۰/۱۴
۴	زنان	۰/۷۰	۱/۳۰	۰/۵۳	۰/۷۱	۰/۱۲
۵	ارتوپدی	۰/۶۴	۱/۹۳	۰/۸۵	۰/۸۱	۰/۴۴
۶	توبرکولین	۰/۶۸	۱/۳۲	۰/۷۰	۰/۷۶	۰/۲۱
۷	احیا	۰/۶۸	۱/۲۷	۰/۶۶	۰/۸۹	۰/۱۳
۸	بخیه	۰/۶۸	۱/۶۰	۰/۷۵	۰/۷۵	۰/۲۵
۹	اطفال	۰/۶۵	۰/۶۶	۰/۲۲	۰/۷۷	۰/۲۳
۱۰	فشارخون	۰/۶۷	۱/۴۰	۰/۶۷	۰/۸۰	۰/۱۹



شکل ۱: نمودار پراکندگی نمرات چک‌لیست بر اساس ارزیابی گلوبال ایستگاه‌های OSCE پیش‌کارورزی ۹۱ و معادله رگرسیون مربوطه

بحث

در این مطالعه شاخص‌های سایکومتریک آزمون OSCE پیش‌کارورزی دانشگاه علوم پزشکی تهران در سطح کل آزمون و در سطح هر یک از آیتم‌ها مورد بررسی قرار گرفت.

بر اساس نمره حدنصابی که با روش رگرسیون مرزی به دست آمد و بالاتر از نصف نمره کل بود، میزان ردی در آزمون ۱/۵ درصد محاسبه شد که با توجه به معیاری بودن آزمون چندان دور از ذهن نیست. در آزمون سال ۸۸ که با شرکت ۱۰۵ دانشجو و ۹ ایستگاه برگزار شد، این میزان ۴/۸ درصد محاسبه شده بود(۶). نکته جالب توجه این که چنانچه قرار بود حدنصاب OSCE پیش‌کارورزی مانند آزمون کتبی پیش‌کارورزی به صورت مقدار ثابت ۵۰ درصد کل نمره در نظر گرفته شود، میزان ردی از این هم کم‌تر می‌شد.

همچنین مقدار RMSE حدنصاب که میزان خطا در برآورد نمره حدنصاب را مشخص می‌کند، معادل ۰/۴۵ به دست آمد که ناچیز است. این میزان برای OSCE سال ۱۳۸۸ با ۹ ایستگاه معادل ۰/۵۵ به دست آمده بود(۶). در مطالعه‌ای که هومر(Homer) و همکاران انجام دادند تا میزان خطای روش رگرسیون مرزی را در حجم نمونه‌های متفاوت بررسی کنند، دریافتند که میزان خطا در OSCEهایی که با شرکت تعداد زیاد دانشجویان (حدود ۲۵۰ تا ۳۰۰ نفر) برگزار می‌شوند، چندان زیاد نیست. این نویسندگان میزان RMSE برای یک آزمون ۲۰ ایستگاهه را حدود ۰/۳ برآورد کردند و ذکر نمودند که با کاهش تعداد ایستگاه‌ها و کاهش تعداد دانشجویان این مقدار افزایش می‌یابد به طوری که در حجم نمونه کم‌تر از ۵۰ نفر، میزان خطا بیش از یک و غیر قابل قبول است(۸).

آلفای کرونباخ که اطلاعاتی در مورد ثبات درونی آزمون به دست می‌دهد، در این مطالعه ۰/۷۰ بود که با توجه به این که سه ایستگاه آزمون از این مطالعه حذف شده بود و همچنین این مسأله که دامنه مطلوب پایایی OSCE بین ۰/۷ تا ۰/۹ در نظر گرفته می‌شود(۷)، پایایی قابل قبولی محسوب می‌شود. در مطالعه مروری نظام‌مند که توسط برانیک (Brannick) و همکاران در خصوص پایایی OSCE انجام شد، ۳۹ مقاله مورد بررسی قرار گرفت و میانگین آلفا به صورت کلی ۰/۶۶ (بازه اطمینان ۹۵ درصد ۰/۶۲ تا ۰/۷۰) به دست آمد. همان‌طور که انتظار می‌رفت مقدار آلفا با تعداد ایستگاه‌های آزمون ارتباط مستقیم داشت و نویسندگان این مقاله مروری گزارش کردند میانگین آلفا در OSCEهایی با کم‌تر از ۱۰ ایستگاه معادل ۰/۵۶ و در OSCEهایی با بیش از ۱۰ ایستگاه مساوی ۰/۷۴ بود. هرچند عوامل دیگری نیز به جز تعداد ایستگاه روی پایایی موثر بودند به گونه‌ای که برخی مطالعات با تعداد کم‌تر از ۱۰ ایستگاه پایایی بالای ۰/۸۰ گزارش کرده بودند و برخی برعکس حتی با ۲۵ ایستگاه پایایی پایینی داشتند(۹).

یکی از کاربردهای پایایی، تخمین میزان خطای اندازه‌گیری در آزمون‌های سطح بالا و مهم است تا با افزودن آن به نمره حدنصاب از قبولی دانشجویان غیرتوانمندی که احتمالاً به خاطر خطای آزمون قبول اعلام شده‌اند، جلوگیری شود(۱۰). هرچه پایایی آزمون بالاتر باشد، SEM کم‌تر خواهد بود. دراین مطالعه میزان خطا در برآورد نمرات دانشجویان نزدیک به ۵ درصد بود که قابل قبول محسوب می‌شود. در مطالعه هومر(Homer) و همکاران خطای آزمون ۳-۲ درصد برآورد شد که در واقع از خطای حدنصاب در آن مطالعه بیش‌تر بود که این امر موجب اطمینان از تأثیر

اندک خطای تعیین استاندارد در نتایج رد و قبول می‌گردد(۸). در مطالعه ما نیز خطای حدنصاب کم‌تر از خطای آزمون به دست آمد.

شاخص آلفا در صورت حذف آیتم اطلاعاتی در خصوص هر ایستگاه به دست می‌آید. انتظار می‌رود که با حذف هر ایستگاه چون تعداد کل ایستگاه‌ها کم شده است، میزان آلفا کاهش یابد یا تغییر چندانی نکند. اگر غیر از این باشد، به این معناست که عملکرد ایستگاه هم‌راستا با بقیه آزمون نبوده و مشکلی ایجاد کرده است. در این مطالعه تمام ایستگاه‌ها از این لحاظ مطلوب بودند. همین یافته در مطالعه پل (Pell)و همکاران نیز که به بررسی OSCE 20 ایستگاهه با شرکت ۲۵۰ دانشجو پرداخته بودند، وجود داشت(۴).

شاخص R2 (ضریب تعیین- coefficient of determination) میزان تغییر در متغیر وابسته (نمره چک‌لیست) را نسبت به تغییراتی که در متغیر مستقل (نمره گلوبال) ایجاد شده است، نشان می‌دهد. به صورت منطقی انتظار داریم در نمرات گلوبال بالا، نمرات چک‌لیست هم بالا باشند و برعکس. باید دقت شود که آزمونگر حاضر در ایستگاه برای دادن نمره گلوبال، نمره چک‌لیست را ترجمه نکرده باشد. در غیر این صورت مقدار این شاخص به صورت غیرطبیعی بالا خواهد بود. اما به صورت کلی مقدار شاخص R2 متوسط یعنی بالای ۰/۵ نشان دهنده رابطه منطقی بین نمرات گلوبال و چک‌لیست است(۴). به عنوان مثال در مطالعه ما، شاخص R2ایستگاه توبرکولین ۰/۷ بود که مشخص می‌کند ۷۰ درصد تغییرات در نمرات گلوبال دانشجویان به دلیل تغییرات در نمره چک‌لیست آنها است و به عبارت دیگر با تفاوت‌های موجود در نمرات چک‌لیست توجیه می‌شود که خوب در نظر گرفته می‌شود. اما شاخص مربوط به دو ایستگاه درماتیت (۰/۱۶) و اطفال (۰/۲۲) از حد مطلوب پایین‌تر است. در این موارد با مراجعه به نمودار پراکندگی نمرات و خط رگرسیون ایستگاه باید تحلیل دقیق انجام شود. همان‌طور که در شکل ۱ مشاهده می‌شود، نمرات چک‌لیست دانشجویان در این دو ایستگاه در هر یک از رده‌های ارزیابی گلوبال طیف وسیعی داشته است. به عبارت دیگر برخی از دانشجویان علیرغم این که نمره چک‌لیست خوبی گرفته‌اند، در ارزیابی گلوبال رضایتبخش نبوده‌اند و برعکس. این مسأله می‌تواند نشان دهنده مشکلی در آزمونگر یا چک‌لیست باشد و از تفاوت برداشت آزمونگر و طراح چک‌لیست ناشی شده باشد(۴و۱۱). فولر و همکاران نشان دادند که R2 سه ایستگاهی که در آزمون سال ۲۰۰۷ دارای مقادیری پایین‌تر از نیم بود، با اعمال تغییراتی در چک‌لیست در استفاده مجدد سال بعد بهبودی نشان داد(۳).

شاخص تمایز بین درجات معادل شیب خط رگرسیون است و نشان می‌دهد که به طور متوسط افزایش چند واحد در نمره چک‌لیست هر ایستگاه منجر به ارتقا یک واحد نمره در مقیاس گلوبال می‌گردد. هرچند که مقدار ایده‌آلی برای این شاخص در OSCE وجود ندارد، یک‌دهم نمره کل چک‌لیست به عنوان حد مطلوب این شاخص ذکر شده است(۴). مقدار پایین این شاخص معمولاً با وجود مشکل در شاخص‌های دیگر همراه است؛ مانند پایین بودن R2 که رابطه ضعیف بین نمرات چک‌لیست و گلوبال را نشان می‌دهد یا واریانس بالای بین آزمونگران (در مواردی که برای هر ایستگاه بیش از یک آزمونگر وجود دارد) که نشان دهنده عدم توافق بین آنها است. مقدار بالای این شاخص می‌تواند با نمره حدنصاب پایین همراه باشد(۴). در مطالعه حاضر که نمره کامل چک‌لیست ۱۰ بود، مقدار مطلوب این شاخص معادل ۱ در نظر گرفته شد. همان طور که

در جدول ۲ مشخص است، تقریباً تمام ایستگاه‌ها از نظر این شاخص در وضعیت خوبی بود به غیر از ایستگاه ارتوپدی که مقدار تمایز بین درجات آن نزدیک به ۲ به دست آمد. جالب این که این ایستگاه پایین‌ترین نمره حدنصاب (۴/۰۳) را نیز دارا بود.

از نظر ضریب دشواری می‌توان مشاهده کرد که ایستگاه‌ها تقریباً مشابه یکدیگر و حدوداً در دامنه آسان قرار گرفته‌اند. همچنین ضریب تمیز ایستگاه‌ها نیز چندان بالا نبود ضمن این که هیچ ایستگاهی ضریب تمیز منفی نداشت. با توجه به معیاری بودن آزمون که انتظار می‌رود اغلب دانشجویان در اغلب ایستگاه‌ها عملکرد خوبی از خود به نمایش بگذارند، ایستگاه‌های نه چندان دشوار با قابلیت تمیز نه چندان بالا، مطلوب محسوب می‌شود.

از محدودیت‌های این مطالعه می‌توان به این موضوع اشاره کرد که تمام شاخص‌ها مبتنی بر نظریه تست کلاسیک بودند. بررسی OSCE با استفاده از شاخص‌های مربوط به تئوری سؤال پاسخ (Item Response Theory) و تئوری تعمیم‌پذیری (Generalizability Theory) می‌تواند در مطالعات بعدی مد نظر قرار گیرد(۱۲).

نتیجه‌گیری

محاسبه شاخص‌های سایکومتریک متناسب با آزمون‌های مبتنی بر عملکرد برای تحلیل آیتم‌ها و نیز تحلیل کل آزمون قابل اجرا است و اطلاعات مفیدی در خصوص نقاط مثبت و منفی آزمون در اختیار می‌گذارد که برای ارائه راه‌کارهای اصلاحی حائز اهمیت است.

قدردانی

این مقاله حاصل بخشی از پایان‌نامه PhD در مقطع دکترای آموزش پزشکی است که با حمایت دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران اجرا شده است.

منابع

1. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001; 357(9260): 945-9.
2. Marks M, Humphrey-Murto S. Performance assessment. In: Dent J ,Harden R, editors. A Practical Guide for Medical Teachers. 3rd Ed. Amsterdam, Netherlands: Elsevier; 2009. Section 6, chapter 44.
3. Fuller R, Homer M, Pell G. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. Med Teach. 2013; 35(6): 515-7.
4. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics: AMEE guide no. 49. Med Teach. 2010; 32: 802–811
5. Mckinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No.85. Med Teach. 2014; 36(2): 97-110
6. MortazHejri S, Jalili M, Muijtjens AMM, Van Der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci. 2013; 18(10): 887–891.
7. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011; 33(6): 447-58.
8. Homer M, Pell G, Fuller R, Patterson J. Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. Med Teach. 2016; 38(2):

181-8.
9. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ. 2011; 45(12: 1181-9.
10. Tavakol M, Dennick R. Making sense of Cronbach's alpha .Int J Med Educ. 2011; 2: 53–55.
11. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 2015; 37(12): 1106-13.
12. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Med Teach. 2012; 34(3): e161-75.

Analyzing the Quality of Objective Structured Clinical Examination in Tehran University of Medical Sciences

Sara Mortaz Hejri[1], Mohammad Jalili[2]

Abstract

Introduction: *With the increasing popularity of performance-based assessments, it is necessary to employ proper indicators to ensure their quality. The present study examines the metrics of Objective Structured Clinical Examination (OSCE).*

Methods: *In this descriptive, cross-sectional research in 10 pre-internship OSCE stations of Tehran University of Medical Sciences in 2012, the cut-off score was determined by borderline regression method and the failure rate was calculated. Root Mean Square Error (RMSE) was used to estimate error rate threshold. Internal consistency was calculated using Cronbach's alpha and then, the Standard Error of Measurement (SEM) was determined. Alpha if item deleted, R2 coefficient, intergrade discrimination, difficulty and discrimination indices were calculated for each station. A total of 266 students participated in this exam.*

Results: *A total of 266 students participated in this exam. The OSCE total cut-off score was 52.55 (out of 100). Four students (1.5%) failed the exam and the RMSE equaled 0.45. SEM was 4.84 and Cronbach's alpha was calculated at 0.70 where the alpha if item deleted scores varied from 0.64 to 0.70. The R2 coefficient ranged from 0.16 to 0.85 and the intergrade discrimination ranged between 0.66 and 1.93. The ranges of difficulty and discrimination indices were 0.71-0.89 and 0.12-0.44 respectively.*

Conclusion: *The internal consistency, SEM and threshold error rate were all acceptable. The alpha if item deleted was in the acceptable range in all stations. In two stations, the R2 value was lower than the desired range. The intergrade discrimination value was appropriate in all stations except one. The stations were not too difficult or highly discriminative which is considered favorable in criterion-referenced exams.*

Keywords: Test analysis, psychometrics, Objective Structured Clinical Examination

Addresses:

[1.] Assistant Professor, Department of Medical Education, Health Professions Education research Center, Tehran University of Medical Sciences, Tehran, Iran. Email: Sa_mortazhejri@razi.tums.ac.ir

[2.] (✉) Professor, Department of Emergency Medicine, Department of Medical Education, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran. Email: mjalili@tums.ac.ir