تعیین نمره حد نصاب قبولی آزمون عینی ساختارمند بالینی به روش انگوف و ارزیابی تأثیر بحث و بررسی نمرات واقعی

سارا مرتاض هجری، محمد جلیلی*، علی لباف

چکیده

مقدمه: برای تعیین استاندارد قبولی آزمون‌های پزشکی روش‌های مختلفی در سطح جهان به کار می‌رود اما استاندارد اغلب امتحانات داخل کشور، به صورت نمره‌ای ثابت و قراردادی و بدون استفاده از متدلوژی علمی تعیین می‌شود. هدف از انجام این مطالعه، تعیین استاندارد قبولی آزمون عینی ساختارمند بالینی در مقطع پیش‌کارورزی دانشگاه علوم پزشکی تهران به روش انگوف است.

روش‌ها: بعد از طراحی سؤالات آزمون، پانلی از داوران متخصص تشکیل شد و ۱۱ داور به صورت مستقل، احتمال قبولی یک دانشجوی مرزی را در هر یک از ایستگاه‌ها برآورد کردند. میانگین احتمالات تمام داوران در تمام ایستگاه‌ها، استاندارد آزمون محسوب شد. این روند دو بار دیگر، پس از برگزاری جلسه بحث بین داوران و بعد از بررسی نمرات واقعی دانشجویان تکرار شد.

نتایج: استاندارد انفرادی کل آزمون ۴۹/۱۵، استاندارد بعد از بحث ۴۹/۹۰ و استاندارد بعد از بررسی نمرات دانشجویان ۵۱/۵۲ به دست آمد. تغییر استاندارد کل آزمون بعد از بررسی نمرات واقعی نسبت به استاندارد انفرادی معنادار بود (p=۰/۰۲)؛ اما نسبت به استاندارد دوم تفاوت معنادار نداشت. همچنین میزان قبولی بر اساس سه استاندارد به ترتیب ۶۷/۶، ۶۴/۸ و ۵۸/۱ درصد به دست آمد که درصد قبولی سوم نسبت به اول کاهش معنادار داشت (p=۰/۰۰۲).

نتیجه‌گیری: در این تحقیق، روش انگوف برای تعیین استاندارد قبولی آزمون عینی ساختارمند بالینی مورد مطالعه و بررسی قرار گرفت. بر اساس یافته‌های این مطالعه، نتایج روش انگوف مخصوصاً پس از بحث و بررسی نمرات واقعی نسبتاً روا و پایا هستند.

واژه‌های کلیدی: استاندارد قبولی، آزمون عینی ساختارمند بالینی، انگوف، ارزیابی توانمندی، ارزشیابی دانشجو

مجله ایرانی آموزش در علوم پزشکی/ بهمن ۱۳۹۰؛ ۱۱(۷): ۸۸۵ تا ۸۹۴

مقدمه

حد نصاب قبولی یا استاندارد آزمون، حداقل نمره‌ای است که چنانچه دانشجو بتواند آن را کسب کند، در آزمون قبول می‌شود. به روند سیستماتیکی که طی آن افراد حرفه‌ای استاندارد قبولی آزمون را مشخص می‌کنند، «تعیین استاندارد» گفته می‌شود. از آنجا که مسائلی مانند میزان دشواری سؤالات، شرایط آزمون، سطح و خصوصیات فراگیران در آزمون‌های مختلف متفاوت است، برای اجرای آزمونی عادلانه، معقول و منطقی نیست که حد نصاب قبولی همه آزمون‌ها، روش نمره ثابت و از قبل تعیین شده (مثلاً ۱۰ یا ۱۲) باشد. در واقع مهم است که به منظور برقراری عدالت آموزشی، حد نصاب قبولی هر آزمون برای همان آزمون و با توجه به نکات فوق تعیین شود. روش‌های مختلف علمی برای تعیین استاندارد وجود دارد که در حوزه

* **نویسنده مسؤول:** دکتر محمد جلیلی (دانشیار) گروه طب اورژانس دانشگاه علوم پزشکی تهران، مرکز تحقیقات آموزش علوم پزشکی (CERMS)، تهران، ایران.
mjalili@tums.ac.ir
دکتر سارا مرتاض هجری، پزشک عمومی، کارشناس ارشد آموزش پزشکی، دانشجوی دکترای تخصصی آموزش پزشکی دانشگاه علوم پزشکی تهران، کارشناس دفتر توسعه آموزش دانشکده پزشکی دانشگاه علوم پزشکی تهران، تهران، ایران.
(sa_mortazhejri@razi.tums.ac.ir)؛ دکتر علی لباف (استادیار)، گروه طب اورژانس دانشگاه علوم پزشکی تهران، تهران، ایران. (alabaf@tums.ac.ir)
این مقاله حاصل طرح تحقیقاتی است که با شماره ۱۰۱۹۶-۷۶-۰۱-۸۹ در سال ۱۳۸۹ توسط دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران تصویب و هزینه‌های آن پرداخت گردیده است.
تاریخ دریافت مقاله: ۹۰/۲/۲۴، تاریخ اصلاح: ۹۰/۴/۲۶، تاریخ پذیرش: ۹۰/۵/۱۸

امتحانات پزشکی توسط دانشگاه‌های متعدد به کار گرفته شده‌اند. صرف‌نظر از روش مورد استفاده، مهم است که تعیین استاندارد قبولی آزمون با دقت و حساسیت انجام شود تا قضاوت‌هایی که صورت می‌گیرد، از یک طرف با کمترین خطا و هزینه از نظر لطمه به دانشجویان همراه باشد و از طرف دیگر با آسان‌گیری همراه نشود تا گیرندگان خدمت در جامعه به خاطر آن متضرر گردند.

یکی از روش‌های متداول برای تعیین استاندارد، روش موسوم به انگوف است(۱). در این روش پس از طراحی سؤالات، پانلی از داوران متخصص تشکیل می‌شود و هر داور به صورت مستقل به این سؤال پاسخ می‌دهد که: «چقدر احتمال دارد یک دانشجوی مرزی بتواند به یک سؤال پاسخ صحیح بدهد». سپس میانگین احتمالاتی که داورهای مختلف به هر سؤال داده‌اند محاسبه می‌شود و به این ترتیب استاندارد آن سؤال مشخص می‌شود. برای تعیین استاندارد کل آزمون، میانگین استانداردهای همه سؤالات محاسبه می‌گردد.

طی سال‌های گذشته، تغییراتی توسط پژوهشگران مختلف در روش انگوف اعمال شده که همه آنها تحت عنوان کلی روش انگوف تغییریافته مشهورند. از جمله این تغییرات، بحث بین داوران در فواصل کار و همچنین در انتهای جلسه است که بر اساس آن، داوران می‌توانند در نمراتی که داده‌اند تغییراتی اِعمال کنند(۲). در نوع دیگری از روش انگوف، پیشنهاد شده که قبل از جلسه بحث انتهایی، نمرات واقعی دانشجویان در اختیار داوران قرار گیرد. گفته می‌شود در این روش چون داوران در فضایی واقعی‌تر کار می‌کنند، راحت‌تر می‌توانند استاندارد را برآورد نمایند(۳).

در مورد سابقه تعیین استاندارد قبولی در کشور باید گفت که در «راهنمای برگزاری آزمون بالینی ساختاردار عینی» که توسط دبیرخانه شورای آموزش پزشکی و تخصصی تدوین شده، تعیین استانداردها به روشی درست بسیار با اهمیت عنوان شده است و استفاده از دو روش انگوف تغییریافته و رگرسیون مرزی توصیه شده است(۴). همچنین در یکی از مقالات خبرنامه شماره هفده دبیرخانه شورای آموزش پزشکی و تخصصی در مورد تعیین حدنصاب آزمون ذکر شده که علی‌رغم اهمیت تعیین استاندارد هر آزمون به صورت جداگانه و اختصاصی برای همان آزمون، با توجه به مشکلات اجرایی و سیاست‌گذاری، انجام روش‌های علمی برای تعیین استاندارد عملاً امکان‌پذیر نیست و حتی استاندارد آزمون‌های حساس و مهمی مانند ارتقا و دانشنامه، ثابت و از پیش تعیین شده است(۵). در حد اطلاع مجریان این طرح، تاکنون برای امتحانی در دانشگاه‌های علوم پزشکی کشور، تعیین استاندارد با متدلوژی علمی صورت نگرفته است و در اغلب امتحانات، استاندارد آزمون به صورت قراردادی و ثابت، نمره ۱۰ یا ۱۲ در نظر گرفته می‌شود.

از سال ۱۳۸۷ تاکنون از دانشجویان پزشکی دانشگاه علوم پزشکی تهران OSCE پیش‌کارورزی به عمل می‌آید. نمره این آزمون به دلیل محدودیت‌های موجود، در کنار نمره امتحان کتبی پیش‌کارورزی جنبه تشویقی داشته و آزمون دارای نمره حدنصاب قبولی نیست؛ بنابراین دانشجویان در آن رد یا قبول نمی‌شوند. از طرفی با توجه به مصوبه وزارت بهداشت، درمان و آموزش پزشکی که دانشگاه‌ها موظف هستند در آینده نزدیک توان‌مندی دانشجویان خود را در آستانه فارغ‌التحصیلی با OSCE مورد سنجش قرار دهند و بر اساس نتایج حاصل از آن در مورد صلاحیت آنان برای شروع به کار مستقل تصمیم‌گیری کنند، کسب تجربه برای تصمیم‌گیری در مورد صلاحیت دانشجویان اهمیت فراوانی پیدا می‌کند. در این حالت که هدف از برگزاری OSCE، تصمیم‌گیری در مورد ورود دانشجویان به مقطع کارورزی بر اساس صلاحیت بالینی آنها است، باید یک آزمون معیارمحور با نمره حدنصاب قبولی در نظر گرفته شود و بنابراین تعیین استاندارد آن حائز اهمیت است. هدف ما از انجام این مطالعه، تعیین استاندارد برای آزمون OSCE پیش‌کارورزی به عنوان پایلوتی برای تعیین استاندارد OSCE پس‌کارورزی بود. در واقع در مطالعه ما نمره قبولی با روش‌های مختلفی استخراج شده و در این مقاله نتایج حاصل از تعیین استاندارد OSCE پیش‌کارورزی به شیوه انگوف ارائه خواهد شد.

روش‌ها

این پژوهش یک مطالعه توصیفی است که هدف آن تعیین استاندارد نمره قبولی آزمون OSCE است. برای انجام روش انگوف، باید پانلی از داوران متخصص تشکیل می‌شد.

برای این کار لیستی شامل اعضای هیأت‌علمی دانشکده پزشکی که با کارآموزان سال‌های چهارم و پنجم پزشکی کار کرده بودند و در زمینه برگزاری OSCE تجربه داشتند، آماده شد. در انتخاب ایشان تلاش شد تا بیشترین تنوع از لحاظ تخصص، سابقه آموزشی، سن و جنس در نظر گرفته شود. همه اعضای کمیته قبلاً پزشک عمومی بوده‌اند پس می‌توانستند در تعیین حداقل نمره آزمون نظر دهند. پس از این که کمیته آزمون، ایستگاه‌های OSCE را طراحی و چک‌لیست‌های مربوط را آماده کرد، جلسه تعیین استاندارد به صورت کارگاهی یک روزه (سه ساعت و نیم) با حضور ۱۵ نفر از اعضای هیأت‌علمی منتخب برگزار شد (جدول ۱).

یکی از مواردی که به صورت بارز و مشخص در کارگاه مورد توجه قرار گرفت، تعریف سطوح مختلف عملکرد دانشجو شامل عالی، رضایت‌بخش، مرزی و افتاده، و همچنین مفهوم دانشجوی مرزی بود که تلاش شد با بحث گروهی، برداشت اعضای هیأت‌علمی در این خصوص به یکدیگر نزدیک شود. بر اساس بحث‌های صورت گرفته در نهایت جمع به تعریف زیر در مورد عملکرد دانشجوی مرزی رسید: دانشجویی که وظیفه مورد نظر را در سطح قابل قبول ارائه نمی‌دهد و از طرفی عملکرد او تا آن اندازه هم بد نیست که او را افتاده محسوب کنیم.

جدول ۱: برنامه کارگاه تعیین استاندارد به روش آنگوف و جدول زمان‌بندی

عنوان	جزییات	روش	زمان
مقدمه	اهمیت ارزیابی و انواع آزمون مفهوم و ضرورت تعیین استاندارد آزمون سابقه OSCE در دانشکده و برنامه‌های آینده	سخنرانی	۵ دقیقه
روش آنگوف	انواع روش‌های تعیین استاندارد روش آنگوف روش کار در این کارگاه	سخنرانی	۵ دقیقه
سطوح عملکردی و دانشجوی مرزی	سطوح عملکردی و توانمندی نمونه‌ای از یک مهارت در سطوح مختلف مفهوم دانشجوی مرزی	سخنرانی	۱۵ دقیقه
تمرین	تعیین استاندارد برای دو ایستگاه OSCE قبلی	انفرادی/بحث گروهی	۳۰ دقیقه
اجرای روش انگوف	تعیین استاندارد برای نیمی از ایستگاه‌های OSCE	کار عملی انفرادی	۴۰ دقیقه
	بحث گروهی در مورد استاندارد هر ایستگاه و بازبینی استاندارد	بحث گروهی/انفرادی	۴۰ دقیقه
	تعیین استاندارد برای نیمه دوم ایستگاه‌های OSCE	کار عملی انفرادی	۴۰ دقیقه
	بحث گروهی در مورد استاندارد هر ایستگاه و بازبینی استاندارد	بحث گروهی/انفرادی	۴۰ دقیقه

همچنین به منظور کسب آمادگی لازم، قبل از شروع تعیین استاندارد برای این آزمون، دو ایستگاه از آزمون مشابه سال قبل انتخاب شد تا داوران به صورت آزمایشی استاندارد آنها را تعیین کنند. با شروع روند تعیین استاندارد برای آزمون، از داوران خواسته شد احتمال قبول شدن یک دانشجوی مرزی در هر یک از ایستگاه‌ها را به صورت مقداری از صفر تا ۱۰۰ به صورت انفرادی بیان کنند که در نهایت «استاندارد اول» از آنها استخراج شد.

برای افزایش میزان توافق بین داوران و در نتیجه افزایش پایایی تعیین استاندارد، دو جلسه بحث در کارگاه هم در نظر گرفته شد. به این صورت که بعد از اتمام نیمی از ایستگاه‌ها (هفت ایستگاه) جلسه متوقف شد، استانداردهای اول جمع‌آوری شد و استاندارد کل هر ایستگاه و استاندارد کل آزمون بر اساس استانداردهای اول محاسبه شد. از داوران خواسته شد تا بر اساس این داده‌ها با یکدیگر تبادل نظر کرده و در صورت صلاحدید در استاندارد اول خود تجدید نظر کنند. به این ترتیب «استاندارد دوم» به دست آمد.

پس از برگزاری OSCE و تعیین نمرات دانشجویان، از هیأت‌علمی مجدداً برای شرکت در جلسه‌ای دعوت به عمل آمد که در آن، میانگین نمرات دانشجویان در هر ایستگاه، حداقل و حداکثر نمره هر ایستگاه و درصد قبولی دانشجویان بر اساس استانداردی که داوران در کارگاه قبلی تعیین کرده بودند، در اختیار ایشان قرار گرفت و از آنان خواسته شد تا در صورت صلاحدید در استانداردی که تعیین کرده بودند، تجدیدنظر کنند.

در این جلسه ۱۱ نفر از داوران حضور داشتند. این استاندارد به عنوان «استاندارد سوم» محسوب شد.

استانداردهای اول (انفرادی)، دوم (بعد از بحث) و سوم (بعد از بررسی نمرات واقعی) به صورت دو به دو توسط آزمون آماری paired sample T-test و در سطح معناداری آماری ۰/۰۵ با یکدیگر مقایسه شدند. مقایسه میزان قبولی ناشی از استانداردها با تست McNemar انجام شد. برای سنجش پایایی تعیین استاندارد، دو شاخص بازه اطمینان استانداردها و توافق بین داوران (از طریق Inter Class ICC (Correlation) محاسبه شد. برای بررسی روایی، میزان قبولی دانشجویان در آزمون OSCE را با نمرات کتبی آزمون جامع پیش‌کارورزی ایشان مقایسه کردیم. همچنین میزان همبستگی میان استانداردهای تعیین شده با درصد قبولی مربوطه توسط ضریب همبستگی پیرسون به عنوان شاخصی از روایی محاسبه شد.

جدول ۲: توزیع نمرات دانشجویان شرکت‌کننده در OSCE پیش‌کارورزی به تفکیک ایستگاه

شماره ایستگاه	نام ایستگاه	پایین‌ترین نمره	بالاترین نمره	میانگین	انحراف معیار	میانه	نما	چارک	
								یک	سه
۱.	شرح حال سردرد	۲۵	۹۵	۶۶/۳۶	۱۳/۸۰	۶۵	۶۵	۶۰	۷۵
۲.	مهارت ارتباطی	۲۶	۹۶	۶۲/۵۹	۱۶/۶۰	۶۲	۷۲	۵۰	۷۳
۳.	معاینه قلب و ریه	۰	۷۱/۴۳	۲۸/۱۶	۱۹/۶۸	۲۸/۵۷	۱۴/۲۹	۱۴/۲۹	۴۲/۸۶
۴.	معاینه پستان	۳۵/۲۷	۱۰۰	۷۹/۶۷	۱۳/۹۲	۸۱/۸۲	۱۰۰	۷۰	۸۸
۵.	اداره سوختگی	۵	۱۰۰	۳۴/۷۶	۱۸/۳۹	۳۵	۴۰	۲۰	۵۰
۶.	تشخیص ضایعه پوستی	۰	۹۰	۴۸/۱۴	۲۶/۸۷	۵۵	۰	۲۵	۷۵
۷.	معاینه ژنیتال	۰	۱۰۰	۵۱/۳۱	۲۴/۴۵	۵۰	۶۲/۵۰	۲۰	۶۰
۸.	آتل‌گیری	۰	۱۰۰	۱۱/۲۴	۱۹/۲۵	۰	۰	۰	۲۰
۹.	انجام سوناژ	۲۸/۵۷	۱۰۰	۷۷/۸۲	۱۶/۹۵	۷۸/۵۷	۵۸/۷۱	۷۱/۴۳	۹۱/۰۷
۱۰.	معاینه شکم	۸/۳۳	۹۵/۸۳	۶۵/۱۵	۱۵/۵۸	۵۴/۱۷	۵۴/۱۷	۴۵/۸۳	۶۶/۶۷
۱۱.	بخیه زدن	۰	۱۰۰	۷۸/۵۰	۲۱/۰۸	۸۵	۱۰۰	۷۰	۹۵
۱۲.	آزمون تیروئید	۰	۱۰۰	۴۸۰/۱۰	۳۳/۳۸	۵۰	۷۵	۲۵	۷۵
۱۳.	رفلکس نوزادی	۰	۱۰۰	۶۰/۴۶	۲۱/۲۹	۶۰	۳۷/۱۴	۴۲/۸۶	۷۷/۱۴
۱۴.	شرح حال اسهال	۲۰	۹۶	۵۴/۲۵	۱۷/۳۳	۵۲	۵۲	۴۰	۶۶
	کل ازمون	۳۱/۴۵	۷۵/۶۵	۵۴/۱۱	۸/۸۰	۵۳/۵۷	۴۹/۸۱	۴۷/۵۰	۶۰/۲۲

نتایج

در تاریخ ۱۸ اسفند ۸۸، OSCE پیش‌کارورزی در دو نوبت صبح و عصر برگزار شد که در آن ۱۰۵ دانشجو به صورت داوطلبانه شرکت کردند. این آزمون شامل ۱۴ ایستگاه بود. جدول۲ توزیع نمرات دانشجویان در ایستگاه‌های مختلف را نشان می‌دهد.

برای تعیین استاندارد به روش انگوف، در جلسه اول ۱۵ داور حضور داشتند که رشته تخصصی ایشان به ترتیب ۴ نفر داخلی، ۱ نفر عفونی، ۲ نفر زنان و ۴ نفر طب اورژانس بود. میانگین و انحراف معیار سن داوران ۳۹/۰۹±۵/۲۰ سال بود. همچنین میانگین سابقه کار ایشان به عنوان هیأت‌علمی ۷/۹۵ سال بود و دو نفر

(۱۸/۲ درصد) خانم بودند. از آنجا که ۱۱ نفر از این داوران در جلسه دوم تعیین استاندارد شرکت کردند، تمام آنالیزهای انجام شده بر اساس نظرات ۱۱ داور صورت پذیرفت.

میانگین نظر ۱۱ داور در هر ایستگاه به صورت انفرادی، بعد از بحث و بعد از بررسی نمرات واقعی، یعنی استانداردهای اول، دوم و سوم در جدول ۳ نشان داده شده‌اند. میانگین استاندارد همه ایستگاه‌ها یعنی نمره قبولی کل آزمون، به صورت انفرادی ۴۹/۱۵ (بازه اطمینان ۵۱/۸۱-۴۶/۴۹)، بعد از بحث ۴۹/۹۰ (بازه اطمینان ۵۳/۵۸-۴۶/۲۱) و بعد از بررسی نمرات دانشجویان ۵۱/۵۲ (بازه اطمینان ۵۴/۷۴-۴۸/۳۰) به دست آمد. تغییر استاندارد کل آزمون بعد از بررسی نمرات واقعی نسبت به استاندارد انفرادی معنادار بود (p=۰/۰۲)؛ اما نسبت به استاندارد بعد از بحث تفاوت معناداری نداشت.

سپس مشخص کردیم که با توجه به نمرات دانشجویان در ایستگاه‌های مختلف و همچنین با توجه به سه استاندارد تعیین شده برای آن ایستگاه، چه تعداد از دانشجویان حداقل نمره قبولی را کسب کرده‌اند و در آن ایستگاه قبول محسوب می‌شوند (جدول ۴). میزان قبولی بر اساس استاندارد اول، دوم و سوم به ترتیب ۶۷/۶، ۶۴/۸ و ۵۸/۱ درصد به دست آمد که میزان قبولی سوم نسبت به اول، کاهش معنادار داشت (p<۰/۰۱).

جدول ۳: میانگین و انحراف معیار استانداردهای اول، دوم و سوم آزمون OSCE پیش‌کارورزی تعیین شده به ترتیب با روش انگوف انفرادی، بعد از بحث و بعد از بررسی نمرات واقعی به تفکیک ایستگاه

شماره ونام ایستگاه	استاندارد اول	استاندارد دوم	استاندارد سوم
۱- شرح حال سردرد	۵۸/۶۴±۱۳/۲۴	۶۱/۳۶±۹/۲۴	۶۳/۱۸±۷/۱۶
۲- مهارت ارتباطی	۵۷/۲۷±۱۵/۵۵	۵۲/۲۷±۱۵/۰۶	۵۹/۵۵±۹/۶۰
۳- معاینه قلب و ریه	۳۹/۵۵±۱۳/۸۶	۴۱/۳۶±۱۰/۷۴	۳۹/۰۹±۸/۰۰
۴- معاینه پستان	۴۳/۶۴±۹/۵۱	۴۱/۳۶±۷/۱۰	۵۹/۵۵±۱۰/۱۱
۵- اداره سوختگی	۳۹/۰۹±۸/۳۱	۴۰/۴۵±۱۲/۳۴	۳۸/۱۸±۷/۱۶
۶- تشخیص ضایعه پوستی	۴۵/۹۱±۱۲/۶۱	۵۱/۳۶±۷/۷۷	۵۰/۰۰±۶/۷۰
۷- معاینه ژنیتال	۴۵/۰۰±۱۵/۱۶	۴۲/۷۳±۱۰/۳۳	۴۲/۲۷±۸/۷۶
۸- آتل‌گیری	۴۰/۰۰±۱۲/۸۴	۳۷/۲۷±۹/۰۴	۳۳/۱۸±۱۱/۶۷
۹- انجام سونداژ	۶۳/۱۸±۱۳/۴۶	۵۹/۰۹±۱۰/۴۴	۶۲/۷۳±۹/۳۱
۱۰- معاینه شکم	۵۰/۹۱±۸/۰۰	۵۴/۰۹±۷/۰۰	۵۳/۱۸±۵/۱۳
۱۱- بخیه زدن	۵۲/۷۳±۱۲/۹۱	۵۴/۵۵±۷/۲۳	۶۲/۷۳±۸/۷۶
۱۲- آزمون تیروئید	۴۹/۵۵±۱۳/۵۰	۵۲/۲۷±۱۲/۷۲	۵۰/۰۰±۱۰/۰۰
۱۳- رفلکس نوزادی	۴۵/۹۱±۱۱/۱۴	۵۳/۱۸±۸/۴۴	۵۳/۶۴±۵/۰۴
۱۴- شرح حال اسهال	۵۶/۸۲±۱۱/۲۴	۵۷/۲۷±۸/۴۷	۵۴/۰۹±۸/۳۱
کل آزمون	۴۹/۱۵±۳/۹۶	۴۹/۹۰±۵/۴۸	۵۱/۵۲±۴/۷۹

بازه اطمینان برای استاندارد اول، دوم و سوم به ترتیب ۵/۳۲، ۷/۳۷ و ۶/۴۴ به دست آمد. میزان توافق بین داوران به ترتیب برای استانداردهای اول تا سوم ۰/۷۷، ۰/۸۸ و ۰/۹۵ به دست آمد.

میزان قبولی دانشجویان در آزمون جامع پیش‌کارورزی ۹۶/۲ درصد بود. همچنین میزان همبستگی میان استاندارد تعیین شده به هر یک از روش‌های مورد مطالعه با درصد قبولی مربوطه (به عبارت دیگر درجه دشواری آزمون) توسط ضریب همبستگی پیرسون، به عنوان شاخصی از روایی، محاسبه شد که به ترتیب برای استانداردهای اول، دوم و سوم ۰/۴۱(p=۰/۱۴)، ۰/۱۴(p=۰/۱۵) و ۰/۸۱(p<۰/۰۰۱) به دست آمد.

جدول ۴: میزان (درصد) قبولی شرکت‌کنندگان در OSCE پیش‌کارورزی بر اساس استانداردهای تعیین شده به روش انگوف انفرادی (استانداردداول)، بعد از بحث (استاندارد دوم) و بعد از بررسی نمرات واقعی (استاندارد سوم) به تفکیک ایستگاه

شماره و نام ایستگاه	استاندارد اول	استاندارد دوم	استاندارد سوم
۱- شرح حال سردرد	۷۷/۱	۷۲/۴	۷۲/۴
۲- مهارت ارتباطی	۶۱/۰	۶۷/۶	۵۷/۱
۳- معاینه قلب و ریه	۳۴/۳	۳۴/۳	۳۴/۳
۴- معاینه پستان	۹۹/۰	۹۹/۰	۹۱/۴
۵- اداره سوختگی	۴۹/۵	۳۱/۴	۴۹/۵
۶- تشخیص ضایعه پوستی	۶۳/۸	۵۵/۲	۶۳/۸
۷- معاینه ژنیتال	۶۱/۹	۶۱/۹	۶۱/۹
۸- آتل‌گیری	۱۱/۴	۱۱/۴	۱۱/۴
۹- انجام سوندژ	۸۳/۸	۸۴/۸	۸۳/۸
۱۰- معاینه شکم	۶۱/۰	۴۸/۶	۶۱/۰
۱۱- بخیه زدن	۸۷/۶	۸۷/۶	۸۲/۹
۱۲- آزمون تیروئید	۶۵/۷	۴۰/۰	۶۵/۷
۱۳- رفلکس نوزادی	۷۰/۵	۶۱/۰	۶۱
۱۴- شرح حال اسهال	۳۷/۱	۳۷/۱	۴۷/۶
کل آزمون	۶۷/۶	۶۴/۸	۵۸/۱

بحث

هرچند برای تعیین استاندارد آزمون‌های مبتنی بر عملکرد مطالعات‌های زیادی انجام شده اما به عقیده پژوهشگران هیچ روش تعیین استانداردی برای OSCE کامل و بی‌نقص نیست (۶و۷). از طرفی استفاده از روش‌های مختلف، موجب تعیین استانداردهای مختلف می‌شود (۸و۹). ما در این مطالعه روش انگوف را مورد بررسی قرار دادیم. برای کاهش محدودیت‌های روش انگوف، دو مدل روش انگوف یعنی بحث و بررسی نمرات واقعی پیشنهاد شده است که ما در این مطالعه به ارزیابی هر دو مدل پرداختیم.

پس از تعیین استاندارد هر ایستگاه به صورت انفرادی، بحث بین افراد به آنها این امکان را می داد که از نظرات و استدلال یکدیگر آگاه شوند و در صورت لزوم، استاندارد خود را تغییر دهند. همچنین بعد از امتحان جلسه مجدد انگوف برگزار و نمرات دانشجویان در اختیار داوران قرار داده شد. بر اساس نتایج به دست آمده، استاندارد دوم تفاوت معنادار با استاندارد اول نداشت. در مطالعه Stern برای تعیین استاندارد OSCE با ۱۰ ایستگاه، به روش آنگوف تغییریافته، استاندارد انفرادی با استاندارد بعد از بحث مقایسه شد که تقریباً یکسان بود اما تفاوت بین داورها کم شده بود(۱۰). در مطالعه ما بررسی نمرات واقعی باعث شد که استاندارد سوم نسبت به اول افزایش معنادار داشته باشد. مطالعه Kramer نشان داده بود که با بررسی نمرات واقعی، استاندارد به طور معنادار کاهش یافت (۷۳/۴ نسبت به ۶۶/۳، p<۰/۰۰۱)(۱۱). البته این مسأله در مطالعه Schoonheim-Klein تأیید نشد. در واقع در برخی از ایستگاه‌ها استاندارد پس از بررسی نمرات واقعی افزایش یافت. نویسندگان توضیح داده‌اند که از نظر داوران، کاهش استاندارد یک ایستگاه همیشه بهترین راه حل نیست(۱۲). بر اساس مطالعه Busch در این روش توافق بین داوران افزایش می‌یابد(۱۳). بر اساس یک مطالعه دیگر، داوران پس از دیدن نمرات واقعی، در ۲۵ درصد موارد برآورد خود را تغییر داده بودند. همچنین تغییرات اعمال شده به ازای هر آیتم بسیار کوچک بودند. علاوه برآن معمولاً آیتم‌هایی تغییر پیدا کردند که در ابتدا ضریب دشواری بسیار بالا یا بسیار پایینی داشتند(۱۴).

بر پایه مطالعات صورت گرفته، نشان داده شده که پایایی و روایی روش‌های تعیین استاندارد هم متفاوت بوده است(۳).

توافق بین داوران بعد از بحث و بعد از بررسی نمرات واقعی افزایش یافت. این امر، یافته‌های مطالعات دیگر را مبنی بر این که بحث و بررسی نمرات واقعی در راستای افزایش توافق بین داوران و افزایش پایایی روش تعیین استاندارد مفید هستند، تأیید می‌کند(۱۵و۱۶).

روش استاندارد طلایی برای سنجش روایی روش‌های تعیین استاندارد وجود ندارد. در برخی از مطالعات از میزان سخت‌گیرانه بودن یا آسان‌گیرانه بودن استانداردها، به عنوان ملاکی به منظور تخمین روایی استفاده شده است. در مطالعه ما بیشترین و کمترین درصد قبولی به ترتیب بر اساس استاندارد اول (انفرادی) و سوم (بررسی نمرات واقعی) به دست آمد. همچنین اگر استاندارد قبولی این آزمون را مانند سایر امتحانات مقطع

مربوطه عدد ثابت ۱۲ یعنی ۶۰ درصد در نظر بگیریم، مشخص است که اعداد به دست آمده با روش انگوف، نسبتاً آسان‌گیرانه‌تر بوده‌اند.

رویکرد دیگری که برای سنجش روایی روش تعیین استاندارد مورد استفاده قرار می‌گیرد، مقایسه درصد قبولی حاصل از این روش با سایر آزمون‌هایی است که از همین دانشجویان در همین مقطع به عمل آمده است. بر اساس مقایسه میزان قبولی دانشجویان در آزمون جامع پیش‌کارورزی مشاهده شد که درصد قبولی بر اساس استانداردهای حاصل از هر سه مدل انگوف، بسیار کمتر از درصد قبولی دانشجویان در آزمون کتبی پیش‌کارورزی (۹۶/۸ درصد) بود. البته هدف کلی از هر دو آزمون، سنجش آمادگی دانشجویان برای ورود به مقطع کارورزی است اما اساساً دو چیز متفاوت را اندازه‌گیری می‌کنند: آزمون کتبی، بیشتر به سنجش محفوظات و دانش شرکت‌کنندگان می‌پردازد در حالی که حیطه ارزیابی آزمون OSCE، مهارت‌های عملی و بالینی دانشجویان است. به نظر می‌رسد به علت تأکید مکرر بر آزمون‌های کتبی و سنجش دانش در مقاطع مختلف، طبیعی است که دانشجویان در حوزه دانشی عملکرد بهتری از خود نشان دهند(۱۷).

ضریب همبستگی پیرسون بین استاندارد تعیین شده و میزان قبولی مرتبط با آن، به عنوان شاخص دیگری از روایی محاسبه شد که در تمام مدل‌ها مثبت بود. این ضریب در مدل اول و دوم تفاوت چندانی نکرد اما بعد از بررسی نمرات واقعی، افزایش یافت که نشان می‌دهد استانداردها نسبت به سختی ایستگاه حساس بوده‌اند. در مطالعه Kramer نیز روایی پس از بررسی نمرات واقعی افزایش نشان داده بود (از ۰/۶۹ به ۰/۸۸)(۱۱).

یکی از محدودیت‌های مطالعه این بود که تعیین استاندارد با روش انگوف برای اولین بار در دانشکده پزشکی انجام می‌شد و هیچ یک از داوران تجربه قبلی در این زمینه نداشتند. در اجرای روش انگوف، داوران معمولاً برای در نظر گرفتن و تصور کردن خصوصیات دانشجوی مرزی و تبدیل این مفهوم مبهم به عدد مشکل دارند(۱۸و۱۹ و۲۰). گفته شده که فضای مجازی و فرضی حاکم بر جلسه می‌تواند درک داور را از عملکرد دانشجوی مرزی تحت تأثیر قرار دهد به طوری که استاندارد بیش از حد بالا و سخت‌گیرانه باشد(۲۱و۲۲). یکی از راه‌حل‌های عنوان شده، اهمیت دادن به انتخاب داورها و همچنین آموزش دادن به آنهاست(۱۵و۲۰و۲۳و۲۴). به گونه‌ای که یکی از علل نتایج ضعیف تعیین استاندارد را فقدان آموزش و تمرین داوران می‌دانند(۱۵). هرچند پروتکل آموزشی مشخص و استانداردی برای داوران وجود ندارد. اما ذکر شده که دادن بازخوردهای سریع مخصوصاً طی تمرین‌ها، کمک‌کننده است(۲۵). در مطالعه ما در ابتدای کارگاه، روش کار برای داوران توضیح داده شد و بعد از آنان خواستیم در رابطه با روش، نحوه نمره دادن و مفهوم دانشجوی مرزی با یکدیگر بحث کنند تا به دید مشترک برسند. همچنین به صورت تمرینی استاندارد دو ایستگاه از آزمون سال قبل را تعیین کردند و بازخورد گرفتند. سپس فرایند تعیین استاندارد را شروع کردند. ما تلاش کردیم تا با برگزاری کارگاه آموزشی قبل از اجرای روش این نقص را جبران کنیم.

محدودیت دیگر، عدم حضور ۴ نفر از داوران در جلسه بررسی نمرات واقعی بود که ما آنها را از مطالعه حذف کردیم. این امر ممکن است نتایج ما را مخدوش کرده باشد. به همین دلیل ما میانگین استانداردهای این داوران را با بقیه مقایسه کردیم که تفاوتی بین آنها وجود نداشت. همچنین وجود اعضای هیأت‌علمی با تخصص‌ها و سوابق متفاوت، ضمن این که مطابق مستندات علمی موجود، مزایایی دارد؛ ممکن است هنگام بحث باعث ایجاد تورش ناشی از غلبه نظرات افراد مطلع یا با تجربه‌تر شود.

نتیجه‌گیری

با استفاده از روش انگوف می‌توان استانداردی روا و پایا برای آزمون عینی بالینی ساختارمند به دست آورد خصوصاً اگر بحث در فواصل کار و بررسی نمرات واقعی به آن افزوده شود. از این طریق می‌توان از اجرای عدالت در سنجش و آزمون که یکی از مسائل مهم و کلیدی است، اطمینان حاصل کرد. این روش‌ها قضاوت صحیح‌تری از عملکرد فراگیران و معیار منصفانه‌تری

برای قضاوت به دست می‌دهند.

قدردانی

این مقاله حاصل بخشی از پایان‌نامه تحت عنوان «تعیین استاندارد آزمون عینی ساختارمند بالینی به سه روش انگوف تغییریافته، انگوف سه سطحی و رگرسیون مرزی» در مقطع کارشناسی ارشد آموزش پزشکی در سال ۱۳۹۰ می‌باشد که با حمایت دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران اجرا شده است. این مقاله حاصل طرح تحقیقاتی با عنوان مذکور مصوب دانشگاه علوم پزشکی و خدمات بهداشتی درمانی تهران در سال ۱۳۸۹ به کد ۱۰۱۹۶-۷۶-۰۱-۸۹ است.

منابع

1. Angoff WH. Scales, norms and equivalent scores. In: Thorndike RL, ed. Educational Measurement. 2nd ed. Washington, DC: American Council on Education; 1971.
2. Hambleton RK, Plake BS. Using an extended Angoff procedure to set standards on complex performance assessments. Applied Measurement in Education. 1995; 8(1): 41-55.
3. Cusimano MD. Standard setting in medical education. Acad Med. 1996; 71(10 Suppl): S112-20.
4. Malekanrad E, Einollahi B. [Rahnemaye sadeh baraye bargozaryeh azmoon balinyeh sakhtardare eini]. [cited 2011 6 Mar]; Available from: http://rds.sem-ums.ac.ir/edc/downloads/simple%20help%20for%20OSCE.pdf. [Persian]
5. Iranian Council for Graduate Medical Education. [Tarrahye Azmoon]. [cited 2011 6 Mar]; Available from: http://dme.hbi.ir/cgme/newsletter/17/pdf17/MAGHALEH-TARAHY.pdf. [Persian]
6. Chesser AM, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD. Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Med Educ. 2004; 38(8): 825-31.
7. Davison I, Bullock AD. Evaluation of the Introduction of the Objective Structured Public Health Examination. Birmingham: The University of Birmingham; 2007
8. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000; 75(3): 267-71.
9-Cusimano MD, Rothman AI. The effect of incorporating normative data into a criterion-referenced standard setting in medical education. Acad Med. 2003; 78(10 Suppl): S88-90.
10. Stern DT, Ben-David MF, De Champlain A, Hodges B, Wojtczak A, Schwarz MR. Ensuring global standards for medical graduates: a pilot study of international standard-setting. Med Teach. 2005; 27(3): 207-13.
11. Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. Med Educ. 2003; 37(2): 132-9.
12. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. Eur J Dent Educ. 2009; 13(3): 162-71.
13. Busch JC, Jaeger RM. Influence of Type of Judge, Normative Information, and Discussion on Standards Recommended for the National Teacher Examinations. Journal of Educational Measurement. 1990; 27(2): 145-63.
14. Norcini JJ, Shea JA, Kanya DT. The effect of various factors on standard setting. Journal of Educational Measurement. 1988; 25(1): 57-65.
15. Yudkowsky R, Downing SM, Wirth S. Simpler standards for local performance examinations: the Yes/No Angoff and whole-test Ebel. Teach Learn Med. 2008; 20(3): 212-7.
16. Wayne DB, Barsuk JH, Cohen E, McGaghie WC. Do baseline data influence standard setting for a clinical skills examination? Acad Med. 2007; 82(10 Suppl): S105-8.
17. Wilkinson TJ, Newble DI, Frampton CM. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Med Educ. 2001; 35(11): 1043-9.
18. Impara JC, Plake BS. Standard setting: An alternative approach. Journal of Educational Measurement.

1997; 34(4): 353-66.
19. Boursicot K. Setting Standards in a Professional Higher Education Course: Defining the Concept of the Minimally Competent Student in Performance Based Assessment at the Level of Graduation from Medical School. Higher Education Quarterly. 2006; 60(1): 74-90.
20. Ricker KL. Setting cut-scores: A critical review of the Angoff and modified Angoff methods. Alberta journal of educational research. 2006; 52(1).): 53-64
21. Norcini JJ. Research on standards for professional licensure and certification examinations. Evaluation & the Health Professions. 1994; 17(2): 160-77.
22. Fehrmann ML, Woehr DJ, Arthur W. The Angoff cutoff score method: The impact of frame-of-reference rater training. Educational and psychological measurement. 1991; 51(4): 857-72.
23. Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. Advances in Health Sciences Education. 1996; 1(3): 215-9.
24. Norcini JJ. Setting standards on educational tests. Med Educ. 2003; 37(5): 464-9.
25. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach. 2003; 25(3): 245-9.

Setting Standard Threshold Scores for an Objective Structured Clinical Examination using Angoff Method and Assessing the Impact of Reality Chacking and Discussion on Actual Scores

Sara Mortaz Hejri[1], Mohammad Jalili[2], Ali Labaf[3]

Abstract

Introduction: *A variety of standard setting methods are used worldwide for medical examination acceptance scores while standards of most exams in our country are pre-determined fixed scores which are set without any scientific methodolgy. The aim of this study is to determine minimum pass level for a pre-internship objective structured clinical examination using Angoff method in Tehran University of Medical Sciences.*
Methods: *After designing the questions for examination, a panel of eleven faculty members was formed. These judges were asked to individually estimate the probability that a borderline student would pass each station. The mean of all stations estimated by judges was considered as the standard for the whole exam. This procedure was repeated twice more after sessions of discussion between judges and checking students' real scores.*
Results: *The individual standard for the whole test was 49.15 while it turned to 49.90 after discussion and finally 51.52 after checking the real scores of students. The change of standard of the whole test after checking real scores was significant compared to individual standard (p=0.02). It showed no significant difference compared to the second standard. The rates of passing students according to the three standards were respectively 67.6%, 64.8% and 58.1% which showed a significant reduction in the third compared to the first one.*
Conclusion: *Angoff method was used in this study to set standard for an OSCE. According to the findings of the study, it seems to be a credible and reliable procedure, especially when group discussion and reality check are used.*

Keywords: Standard setting, Angoff, objective structured clinical examination, OSCE, Competence assessment, Student Assessment

Addresses:
[1] MD, MSc, PhD Student of Medical Education, Educational Development Office, Tehran University of Medical Sciences, Tehran, Iran. E-mail: sa_mortazhejri@razi.tums.ac.ir
[2] (✉) Associate Professor, Department of Emergency Medicine, Center for Educational Research in Medical Sciences (CERMS),Tehran University of Medical Sciences, Tehran, Iran. E-mail: mjalili@tums.ac.ir
[3] Assistant Professor, Department of Emergency Medicine, Tehran University of Medical Sciences, Tehran, Iran. E-mail: alabaf@tums.ac.ir