

مقاله اصلی

بررسی ضریب پایایی دومین آزمون صلاحیت‌های بالینی دانشجویان پزشکی مشهد با استفاده از تئوری تعمیم‌پذیری

تاریخ دریافت: ۹۶/۰۸/۰۱ - تاریخ پذیرش: ۹۶/۱۲/۰۱

خلاصه

مقدمه

آزمون آسکی یکی از مؤثرترین روش‌ها برای ارزشیابی مهارت‌ها می‌باشد که بدلیل دشواری در اجرای آزمون آسکی سنجش کیفیت آن، امری حیاتی است. یک گام مهم در سنجش کیفیت هر آزمون، تعیین ضریب پایایی آن آزمون می‌باشد، لذا این پژوهش با هدف تعیین ضریب پایایی دومین آزمون صلاحیت‌های بالینی مشهد با استفاده از تئوری تعمیم‌پذیری انجام گردیده است.

روش کار

دومین آزمون صلاحیت‌های بالینی مشتمل بر ۱۵ ایستگاه موازی بود. ۷۵ دانشجو و ۳۰ ممتحن در این آزمون حضور داشتند. چک‌لیست‌های آزمون به تفکیک ایستگاه و ممتحن جمع‌آوری و در بسته نرم‌افزاری اکسل ثبت گردید. مؤلفه‌های واریانس، مجموع مربعات و میانگین مربعات با استفاده از نرم‌افزار SPSS نسخه ۱۶ و آزمون ANOVA محاسبه گردید. سپس واریانس ایستگاه‌ها، دانشجویان و ممتحنین و همچنین ضرایب تعمیم‌پذیری و پایایی آزمون با استفاده از نرم‌افزار EDUG به دست آمد. همچنین برای بررسی رابطه بین جنسیت و نمرات از آزمون‌های آماری U Mann-Whitney و chi square استفاده گردید.

نتایج

ضریب پایایی آزمون ۰/۹۸ محاسبه گردید و مؤلفه تعامل دانشجویان با ممتحنین و ایستگاه‌ها بیشترین سهم را در ایجاد خطای اندازه‌گیری دارا بود. بین جنسیت دانشجویان و نمرات آن‌ها ارتباط معناداری مشاهده نشد، همچنین بین جنسیت ممتحنین و نمرات دانشجویان نیز ارتباط معناداری به دست نیامد.

نتیجه‌گیری

دومین آزمون صلاحیت‌های بالینی مشهد، پایایی بسیار خوبی داشت و با توجه به نتایج حاصله می‌توان گفت، استفاده از تئوری تعمیم‌پذیری برای برآورد ضریب پایایی و مشخص نمودن خطاهای اندازه‌گیری روش مناسبی است اما پژوهش‌های بیشتری باید انجام شود تا بتوان با قطعیت بیشتری ادعا نمود که استفاده از تئوری تعمیم‌پذیری، مزایای بیشتری نسبت به تئوری کلاسیک دارد.

کلیدواژه‌ها

تئوری تعمیم‌پذیری، پایایی، آزمون آسکی، صلاحیت بالینی
پی‌نوشت: این مطالعه فاقد تضاد منافع می‌باشد.

حسن غلامی^۱

سمانه سرگزئی^{۲*}

ملیحه نوبخت^۳

علی عمادزاده^۴

۱- مربی، گروه آموزش پزشکی، دانشکده

پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران

۲- دانشجوی دکتری آموزش پزشکی، گروه

آموزش پزشکی، دانشگاه علوم پزشکی اصفهان،

اصفهان، ایران

۳- کارشناس ارشد آموزش پزشکی، دانشگاه علوم

پزشکی مشهد، مشهد، ایران

۴- استادیار، گروه آموزش پزشکی، دانشکده

پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران

* دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

تلفن: ۰۹۳۷۶۱۲۹۶۰۵

Email: s.sargazi2010@gmail.com

مقدمه

صاحب‌نظران امر آموزش بر این امر که یادگیری از سنجش نشأت می‌گیرد، توافق دارند (۱). در هر برنامه آموزشی، سنجش دانشجویان یکی از مهم‌ترین بخش‌های آموزش را تشکیل می‌دهد و می‌تواند آموزش را از یک حالت ایستا به مسیری پویا هدایت نماید (۲-۳). در طی سال‌های ۱۹۶۰-۱۹۵۰ میلادی، مهارت‌های بالینی عمدتاً از طریق آزمون انشائی و شفاهی سنجیده می‌شد (۴) اما وجود نارضایتی‌هایی از روش‌های سنجش بالینی از سوی دانشجویان و اساتید منجر به تلاش برای یافتن ابزار جدیدی برای سنجش مهارت‌های بالینی گردید (۵). در سال ۱۹۷۵ در اسکاتلند، پروفیسور هاردن و همکاران وی برای اولین بار از آزمون آسکی به‌عنوان روشی برای سنجش مهارت‌های بالینی به‌جای روش‌های قدیمی استفاده کردند که مورد توجه و استقبال اساتید و دانشجویان قرار گرفت (۴، ۷-۶) در یک آزمون آسکی با تعداد زیاد آزمون دهنندگان و ایستگاه‌ها، دانشجویان گروه‌بندی شده و هر کدام از ممتحنین در هر ایستگاه، تعداد مشخصی از دانشجویان را مورد سنجش قرار دهند. در چنین آزمونی، عوامل متعددی وجود دارند که می‌توانند سبب ایجاد خطا شوند از جمله؛ توانائی‌های دانشجویان مختلف، سخت‌گیری یا سهل‌گیری ممتحنین، تفاوت در بیمارناها و تفاوت در دشواری ایستگاه‌ها (۸). به دلیل دشواری در اجرای آزمون آسکی، محدودیت منابع و درگیری افراد زیادی در این آزمون، سنجش کیفیت یک آزمون آسکی امری حیاتی است (۹-۱۰). یکی از گام‌های مهم در سنجش کیفیت هر آزمون، تعیین ضریب پایایی آن آزمون می‌باشد (۱۱) و به دلیل پیچیدگی‌هایی که در امر سنجش عملکرد وجود دارد، پژوهشگرانی که سعی در تعیین پایایی نمرات آزمون‌های عملکردی دارند، با چالش‌های اساسی مواجه گردیده‌اند (۱۲). خصوصیات مختلف یک آزمون از جمله روانی و پایایی آن با

استفاده از تئوری‌های روان‌سنجی^۱، مورد مطالعه قرار می‌گیرد. دو تئوری عمده در روان‌سنجی عبارت‌اند از تئوری نمونه‌گیری تصادفی^۲ و تئوری سؤال- پاسخ^۳. تئوری نمونه‌گیری تصادفی، خود شامل دو تئوری نسبتاً متفاوت است: تئوری کلاسیک یا پایایی کلاسیک^۴ و تئوری تعمیم‌پذیری^۵ (۱۳). به دلیل نواقص تئوری کلاسیک در تخمین ضریب پایایی، انجمن پژوهش‌های آموزشی آمریکا (AERA)^۶، انجمن روانشناسان آمریکا (APA)^۷ و انجمن ملی اندازه‌گیری در آموزش^۸ بر این امر تأکید دارند که برای بررسی روانی و پایایی یک روش سنجش، بهتر است از تئوری تعمیم‌پذیری استفاده شود (۱۴، ۱۵). تئوری تعمیم‌پذیری را می‌توان نمونه کامل شده تئوری کلاسیک در نظر گرفت (۱۶) که برای تعیین قابلیت اطمینان (پایایی) یک آزمون رفتاری بکار می‌رود و در این تئوری، تمامی منابع خطای موجود در اندازه‌گیری مشخص شده و سهم هر کدام در ایجاد خطا تخمین زده می‌شود (۱۷، ۱۸). هنگامی که از تئوری تعمیم‌پذیری برای یک آزمون چهارگزینه‌ای استفاده شود تنها یک منبع خطا وجود دارد که همان سؤالات آزمون هستند اما در رابطه با آزمون‌های عملکردی چندین عامل وجود دارند که ممکن است منجر به ایجاد خطا شوند از جمله فرد امتحان‌دهنده، ممتحن و زمان آزمون (۱۹، ۲۰) که هر کدام از این منابع خطا یک رویه^۹ نامیده می‌شوند. این تئوری برای محققین بسیار مفید است زیرا به آن‌ها امکان می‌دهد تا منابع مختلف خطا در یک آزمون را تعیین کنند. هر چه منابع بیشتری از خطا پیدا شود، بهتر می‌توان آن‌ها را

¹Psychometrics

² Random Sample Theory

³ Item- Response Theory

⁴Classical Test Theory

⁵ Generalizability Theory

⁶American Educational Research Association

⁷American Psychological Association

⁸National Council on Measurement in Education

⁹Facet

شرایط شرکت در دومین آزمون صلاحیت‌های بالینی بوده‌اند، جامعه این پژوهش را تشکیل دادند که تعداد کل آن‌ها ۷۵ نفر بود و با توجه به این تعداد، تمامی افراد جامعه پژوهش، نمونه پژوهش را تشکیل دادند، بنابراین برای نمونه‌گیری، از روش سرشماری استفاده گردید. با توجه به اینکه آزمون صلاحیت‌های بالینی در مرکز مهارت‌های بالینی واقع در بیمارستان قائم (عج) برگزار گردید، این مرکز، محیط پژوهش حاضر می‌باشد. همچنین به دلیلی اینکه بخش ارزیابی چک‌لیست‌های مربوط به آزمون از لحاظ وجود آیتم‌های موردنظر و همچنین وارد کردن اطلاعات موجود در چک‌لیست‌های آزمون صلاحیت‌های بالینی در واحد آزمون دانشکده پزشکی، مستقر در اداره آموزش به مدت ۵ روز انجام گردید، این واحد نیز قسمتی از محیط پژوهش حاضر را شامل می‌شود. اطلاعات تمامی چک‌لیست‌ها بصورت محرمانه در نرم افزار اکسل ثبت گردیدند.

برای به دست آوردن ضریب پایایی آزمون در پژوهش حاضر از تئوری تعمیم‌پذیری بهره گرفته شده است که برای استفاده از این تئوری باید ۵ گام زیر طی شود:

- ۱- تعیین هدف اندازه‌گیری
- ۲- تعیین وجه یا وجه‌های اندازه‌گیری
- ۳- شناسایی طرح مناسب اندازه‌گیری
- ۴- شناسایی و برآورد مؤلفه‌های واریانس و تفسیر آن‌ها
- ۵- برآورد پایایی

در ادامه به تشریح هر یک از مراحل فوق می‌پردازیم:

مرحله اول: تعیین هدف اندازه‌گیری، اولین گام در برآورد پایایی به روش تعمیم‌پذیری است (۲۳). قابل ذکر است هدف اندازه‌گیری باید برآورد پایایی و یا تصمیم‌گیری در مورد پایایی باشد. در این پژوهش، هدف اندازه‌گیری برآورد پایایی دومین آزمون صلاحیت‌های بالینی مشهد با در نظر گرفتن خطاهای وارد بر اندازه‌گیری می‌باشد.

کنترل کرده و به حداقل رساند (۲۱). رابرت هیتمن و همکاران وی در مطالعه خود به این نتیجه رسیدند که تئوری تعمیم‌پذیری ۳ ویژگی مهم نسبت به سایر روش‌های تعیین پایایی دارد: ۱- بیشتر از دو نمره را برای هر فرد می‌توان در نظر گرفت. ۲- عوامل مختلفی که منجر به ایجاد خطای اندازه‌گیری می‌شوند شناسایی و سهم هر کدام مقابل اندازه‌گیری است. ۳- با استفاده از این تئوری می‌توان تأثیرافزایش یا کاهش تعداد ممتحنین، ایستگاه‌ها و آزمون دهندگان را در مقدار پایایی به دست آورد (۲۲).

یکی از آزمون‌های آسکی که دانشجویان پزشکی دانشگاه علوم پزشکی مشهد در طول دوره تحصیلی خود و قبل از ورود به دوره انترنی آن را تجربه می‌کنند آزمون صلاحیت‌های بالینی است. از آنجا که مدت‌زمان زیادی از راه‌اندازی این آزمون در دانشگاه علوم پزشکی مشهد نمی‌گذرد و با توجه به صرف هزینه، زمان و منابع انسانی قابل توجه برای اجرای آزمون صلاحیت‌های بالینی، این نیاز احساس شد که ضریب پایایی آزمون به عنوان یکی از معیارهای سنجش کیفیت آزمون بررسی شود تا از این طریق بتوان عواملی که بر پایایی آزمون تأثیر می‌گذارند را شناسایی و کنترل نمود و به این ترتیب، کیفیت آزمون‌های بعدی را افزایش داد، لذا این پژوهش با هدف «بررسی ضریب پایایی دومین آزمون صلاحیت‌های بالینی دانشجویان پزشکی مشهد به روش تعمیم‌پذیری» انجام گردید.

روش کار

این پژوهش یک پژوهش توصیفی-تحلیلی از نوع مقطعی است که در سال ۱۳۹۳ در دانشگاه علوم پزشکی مشهد به منظور بررسی عوامل ایجادکننده خطاهای اندازه‌گیری و میزان دخالت هر یک از این عوامل در نمره نهایی آزمون آسکی و همچنین تعیین پایایی دومین آزمون صلاحیت‌های بالینی انجام گرفت. تمامی دانشجویان پزشکی دانشگاه علوم پزشکی مشهد که

مرحله چهارم: در این مرحله پژوهشگر به شناسایی و تفسیر مؤلفه‌های واریانس موجود در طرح اندازه‌گیری می‌پردازد. این پژوهش ۵ مؤلفه واریانس را به صورت همزمان مورد بررسی قرار می‌دهد که این مؤلفه‌ها در جدول شماره (۱-۳) توصیف گردیده‌اند.

جدول ۱- معرفی مؤلفه‌های واریانس موجود در طرح SR:T

ردیف	مؤلفه‌های واریانس	نماد	توصیف مؤلفه
۱	دانشجویان	S	دانشجویان از لحاظ جسمی، روحی و روانی تفاوت‌های زیادی دارند
۲	ایستگاه‌ها	T	ایستگاه‌ها از لحاظ دشواری باهم تفاوت دارند
۳	ممتحنین که درون ایستگاه‌ها آشیانه کرده‌اند	R:T	ممتحنین هر ایستگاه از لحاظ نمره دهی باهم تفاوت دارند و برخی سهل گیر و بر خی سخت گیر هستند
۴	تعامل بین دانشجویان و ایستگاه‌ها	ST	دانشجویان در مواجهه با هر ایستگاه متفاوت عمل می‌کنند
۵	تعامل بین دانشجویان، ایستگاه‌ها و ممتحنین	SR:T	ممتحنین در هر ایستگاه ممکن است در برخورد با دانشجو (به دلیل پیشینه ذهنی نسبت به دانشجو یا عوامل دیگر) متفاوت عمل کنند

مرحله پنجم: در این مرحله پژوهشگر با استفاده از نرم‌افزارهایی از قبیل SPSS، SAS، GENOVA، یا EDUG پایایی طرح اندازه‌گیری را برآورد می‌نماید که در این پژوهش از نرم افزار SPSS برای برآورد مؤلفه‌های واریانس و از نرم افزار EDUG برای تعیین ضریب پایایی استفاده شده است.

نتایج

با توجه به اینکه در این پژوهش هم تحلیل‌های تعمیم‌پذیری انجام گرفته هم ارتباط سنجی بین جنسیت و نمرات، در ارائه نتایج ابتدا نتایج حاصل از تجزیه و تحلیل‌های تعمیم‌پذیری ارائه

مرحله دوم: پس از تعیین هدف اندازه‌گیری، پژوهشگر باید رویه یا وجه‌های موجود در اندازه‌گیری را تعیین نماید. در تئوری تعمیم‌پذیری به هر یک از منابع خطا، وجه یا رویه و به سطوح تشکیل‌دهنده هر وجه، شرایط یا موقعیت اندازه‌گیری می‌گویند (۲۴). در پژوهش حاضر، ۳ رویه شناسایی شد: دانشجویان (S)، ایستگاه‌ها (T) و ممتحنین (R). شایان ذکر است رویه دیگری هم با عنوان بیمارنا شناسایی شد که به دلیل اینکه در تمامی ایستگاه‌ها از بیمارنا استفاده نمی‌شد، امکان وارد کردن آن در پژوهش وجود نداشت. پس از تعیین رویه‌های اندازه‌گیری، نقش هر یک باید تعیین شود. در تئوری تعمیم‌پذیری، رویه‌ها ۲ نوع دارند که در این گام باید مشخص شود: ۱- رویه متمایز^{۱۰} که در واقع به عنوان رویه هدف اندازه‌گیری در نظر گرفته می‌شود ۲- رویه ابزار^{۱۱} که شرایط را برای رسیدن به هدف اندازه‌گیری فراهم می‌نماید (۱۵). اصولاً در بیشتر مطالعات تعمیم‌پذیری، افراد یا آزمودنی‌ها به عنوان رویه هدف اندازه‌گیری محسوب می‌شوند (۲۵). در این پژوهش، رویه دانشجویان (به علت استفاده از نمرات آزمون آن‌ها) به عنوان رویه متمایز و رویه‌های ایستگاه‌ها و ممتحنین به عنوان رویه ابزار در نظر گرفته شده است. مرحله سوم: پس از تعیین وجه اندازه‌گیری، باید طرح مناسب اندازه‌گیری شناسایی شود. شناسایی طرح مناسب نقش تعیین‌کننده‌ای در برآورد پایایی ایفا می‌کند. برای تعیین طرح اندازه‌گیری باید مشخص شود هر یک از رویه‌ها به صورت متقاطع هستند یا آشیانه‌ای یا ترکیبی از این دو (۱۵). در پژوهش حاضر، ارتباط رویه هدف اندازه‌گیری (دانشجویان) با رویه‌های دیگر به صورت متقاطع بود اما رویه ممتحنین و ایستگاه‌ها به صورت آشیانه‌ای ارتباط داشتند، به این معنی که ممتحنین درون ایستگاه‌ها آشیانه کرده بودند و هر ممتحن، تعدادی از دانشجویان را مورد بررسی قرار می‌داد نه تمام آن‌ها را.

¹⁰Differentiation Facet

¹¹Instrumentation Facet

که بر اساس آن ضریب پایایی آزمون ۰/۹۸ به دست آمد. این ضریب پایایی به طرز غیرقابل انتظاری بالاتر از میزانی بود که پژوهشگران توقع داشتند و برای یک آزمون آسکی، ضریب بسیار خوبی است.

می‌شود و سپس نتایج حاصل از تحلیل‌های مربوط به جنسیت و نمره.

برای دستیابی به هدف کلی این پژوهش که تعیین ضریب پایایی دومین آزمون صلاحیت‌های بالینی مشهد با استفاده از تئوری تعمیم‌پذیری است پژوهشگر به جدول تعمیم‌پذیری مراجعه نمود

جدول ۲- نتایج تجزیه و تحلیل‌های تعمیم‌پذیری برای طرح SR:T

درصد	خطای	تغییرپذیری	درصد	خطای	تغییرپذیری	منبع واریانس	رویه	منبع تغییرپذیری
	آشکار			وابسته		متمایزکننده		
۹۸٪/۳	S
.	T
۱٪/۷	(۰)	R:T
.	ST
.	SR:T
٪۱۰۰	مجموع واریانس‌ها

ضریب پایایی = ۰/۹۸

برای تعیین مؤلفه ای که بیشترین تاثیر را در ایجاد خطای اندازه‌گیری دارد از جدول آنالیز واریانس استفاده شد که با توجه به این جدول، مؤلفه تعامل دانشجو با ممتحن و ایستگاه با بیشترین سهم را در ایجاد خطای اندازه‌گیری دارا می‌باشد

و پس از آن دانشجویان با ۳۲/۱٪، در جایگاه دوم قرار دارند، همچنین مؤلفه تعامل ممتحنین و ایستگاه‌ها با ۰/۳٪، کمترین سهم را در ایجاد خطای اندازه‌گیری در نمرات دومین آزمون صلاحیت‌های بالینی مشهد دارد.

جدول ۳- آنالیز واریانس برای طرح SR:T

ردیف	مؤلفه	واریانس	میانگین مربعات	خطای استاندارد برآورد	درصد خطای اندازه‌گیری
۱	S	۶۵/۱۶	۰/۹۱	۰/۰۰۰۳۳	۳۲/۱
۲	T	۳۱/۵۱	۲/۲۵	۰/۰۰۰۳۶	۱۶/۷
۳	R:T	۶/۰۲	۰/۰۱	۰/۰۰۰۰۱	۳
۴	ST	۰/۵۲	۰/۰۰۰۵	۰/۰۰۰۰۱	۰/۳
۵	SR:T	۹۷/۱۷	۰/۰۰۳	۰/۰۰۰۰۲۵	۴۷/۹
۶	درصد کل			۱۰۰	

آزمون کولموگروف-اسمیرنوف^{۱۲} استفاده گردید و مشخص شد که با توجه به جدول حاصل از این آزمون، مشخص شد توزیع نمرات نرمال نیست ($p < ۰/۰۵$). پس از آن نمرات را

برای تجزیه و تحلیل مربوط به تأثیر جنسیت، به بررسی تأثیر جنسیت دانشجویان و ممتحنین بر نمره دانشجویان پرداختیم. در ابتدا برای بررسی نرمال بودن نمرات دانشجویان، از

¹²Colmogrov-Smirnov

دست آوردند. ضریب پایایی در پژوهش لیساکلر (۲۴) و همکارانش ۰/۸ محاسبه شد که این ضریب از بین سایر پژوهش‌ها، به پژوهش حاضر نزدیک‌تر است.

براینیک (۲۷) و همکارانش در پژوهش خود به این نتیجه رسیدند که بهترین راه برای افزایش ضریب پایایی، افزایش تعداد ایستگاه‌ها و همچنین افزایش تعداد ممتحنین است (۲۷)، همچنین هرچه ناهمگنی آزمودنی‌هایی که در آزمون شرکت می‌کنند بیشتر باشد، ضریب پایایی آزمون افزایش خواهد یافت (۲۶). عواملی که به عقیده پژوهشگر منجر به بالا بودن ضریب پایایی دومین آزمون صلاحیت‌های بالینی مشهد گردیده است را می‌توان به صورت ذیل بیان نمود: آشنایی کافی ممتحنین با روش‌های نمره دهی، توجیه بودن ممتحنین در زمینه هدف هر ایستگاه و آیت‌های موردنظر، عدم سوگیری ممتحنین در ارزیابی دانشجو، طراحی صحیح ایستگاه‌های آزمون، مطابقت چک‌لیست‌های طراحی شده برای هر ایستگاه با هدف آن ایستگاه، ایفای نقش خوب توسط بیمارنماها، زمان‌بندی مناسب برای هر ایستگاه و کافی بودن تعداد ایستگاه‌ها.

واریانس نمره آزمون بر اساس ایستگاه‌ها ۳۱/۵۱، بر اساس دانشجویان ۶۵/۱۶ و بر اساس ممتحنین ۶/۰۲ محاسبه گردید. بالا بودن میزان واریانس رویه دانشجویان نشان می‌دهد که این رویه در مقایسه با دو رویه دیگر، بیشترین نقش را در ایجاد خطای اندازه‌گیری داشته است و علت این امر را می‌توان در عوامل زیادی جستجو کرد از جمله: استرس دانشجویان برای شرکت در آزمون، عدم آمادگی دانشجویان برای حضور در آزمون، کافی نبودن زمان برخی از ایستگاه (که این امر منجر به افزایش استرس می‌گردد، شرایط روحی و جسمی دانشجو، مواجهه با ممتحنین ناآشنا، مواجهه دانشجویان با موردهایی که اطلاعات کافی در مورد آن‌ها ندارند، زمان برگزاری آزمون، مکان برگزاری آزمون.

ترنسفورم^{۱۳} نمودیم اما باز هم توزیع نمرات نشد، بنابراین به سراغ آزمون‌های غیرپارمتری رفتیم.

در خصوص ارتباط جنسیت ممتحنین شرکت‌کننده در دومین آزمون صلاحیت‌های بالینی مشهد با نمرات دانشجویان، از آزمون من ویتنی^{۱۴} استفاده گردید که براساس نتایج حاصله مشخص شد بین نمره دانشجویان و جنسیت ممتحنین شرکت‌کننده در دومین آزمون صلاحیت‌های بالینی مشهد ارتباط معناداری وجود ندارد ($\text{sig}=0/854$).

همچنین برای بررسی ارتباط بین جنسیت دانشجویان شرکت‌کننده در دومین آزمون صلاحیت‌های بالینی مشهد و نمره آن‌ها نیز از آزمون من ویتنی بهره گرفته شد که نتایج این آزمون هم نشان داد بین جنسیت دانشجویان شرکت‌کننده در دومین آزمون صلاحیت‌های بالینی مشهد و نمره آن‌ها هم ارتباط معناداری وجود ندارد و بنابراین بین گروه دختران و پسران از لحاظ نمره کسب شده در دومین آزمون صلاحیت‌های بالینی مشهد، اختلاف معناداری وجود ندارد ($\text{sig}=0/83$).

بحث

ضریب پایایی دومین آزمون صلاحیت‌های بالینی دانشجویان پزشکی مشهد با استفاده از تئوری تعمیم‌پذیری ۰/۹۸ به دست آمد که این ضریب در مقایسه با مقالات و پایان‌نامه‌هایی که پژوهشگر آن‌ها را مورد بررسی قرار داده است، بالاترین میزان ضریب می‌باشد. اکثر سازندگان آزمون‌ها، اگر ضریب پایایی ۰/۹ یا بیشتر به دست آورند، احساس رضایت می‌کنند اما از ضریب کمتر از ۰/۷ ناراضی می‌شوند (۲۶). براینیک و همکارانش (۲۷) پایایی آزمون را ۰/۶۶ محاسبه نمودند، داگلاس و لاوسون (۲۸) پایایی را برای روز اول ۰/۶۳ و برای روز دوم ۰/۷۴ به دست آوردند و استیلسون (۲۹) و همکارانش هم ضریب پایایی آزمون موردنظرشان را عدد ۰/۷ به

¹³Transform

¹⁴Mann-Whitney

برای تعیین مؤلفه‌ای که بیشترین خطای اندازه‌گیری را ایجاد می‌کند از جدول آنالیز واریانس استفاده شده است و بر این اساس، مؤلفه تعامل دانشجویان با ممتحنین و ایستگاه‌ها با ۴۷٪، بیشترین سهم را در ایجاد خطای اندازه‌گیری به خود اختصاص داده است و بعد از آن دانشجویان با ۳۲٪ در جایگاه دوم ایجاد خطای اندازه‌گیری قرار دارند، همچنین مؤلفه تعامل دانشجو با ایستگاه (۳٪) و مؤلفه ممتحنین (۳٪) کمترین درصد ایجاد خطای اندازه‌گیری را دارا هستند. تنها پژوهشی که نتایج تقریباً مشابهی با نتایج پژوهش حاضر دارد پژوهش سبک (۳۰) و همکارانش می‌باشد که بر طبق یافته‌های این پژوهش، دومین مؤلفه‌ای که بیشترین خطای اندازه‌گیری را ایجاد می‌کند، تعامل متقاضیان با ممتحنین و ایستگاه‌ها (۴۰٪) می‌باشد که می‌توان گفت درصد تقریباً نزدیکی به درصد مؤلفه در پژوهش حاضر دارد و نتایج سایر مطالعات با پژوهش حاضر همسو نبودند. لیساکلر (۲۴) در پژوهش خود به نتیجه متفاوتی دست یافت و بیشترین سهم خطای اندازه‌گیری را ابتدا در مؤلفه تعامل دانشجویان با موردها (ایستگاه‌ها) (۵۲٪) مشاهده نمود و پس از آن مؤلفه تعامل دانشجویان با موردها و ممتحنین (۱۹٪) بیشترین سهم را دارا بود. در پژوهش داگلاس و لاوسون (۲۸)، ممتحنین با ۴۲/۲۶٪ بیشترین سهم را در خطای اندازه‌گیری ایفا می‌کردند که این نتیجه بر عکس نتیجه پژوهش حاضر است و در پژوهش حاضر، ممتحنین کمترین سهم را در ایجاد خطا دارند. یادگوسکی (۳۱) هم در پژوهش خود به این نتیجه رسید که مؤلفه تعامل بیمارنا و موارد (۲۲٪) رتبه نخست را در ایجاد خطا به خود اختصاص می‌دهد و پس از آن تعامل بیمارناها با موارد و دانشجویان (۱۷٪) در جایگاه دوم قرار دارد. والوند (۳۲) هم در پایان‌نامه خود به این نتیجه رسید که تعامل بین آزمودنی و بیمارناها اولین عامل ایجاد خطا در اندازه‌گیری می‌باشد و پس از آن بیمارناها بیشترین نقش را در خطای اندازه‌گیری دارند. در پژوهش استیلسون (۲۹) هم ممتحنین با ۵۲٪ بیشترین خطای اندازه‌گیری را داشتند.

همان‌طور که در بالا بیان گردید، مؤلفه تعامل دانشجویان با ممتحنین و ایستگاه‌ها با ۴۷٪، بیشترین سهم را در ایجاد خطای اندازه‌گیری به خود اختصاص داده است که پژوهشگر در این باب، عوامل زیر را مؤثر می‌داند: دانشجو در هر ایستگاه با یک ممتحن جدید مواجه می‌شود و این امر می‌تواند بر روی عملکرد دانشجو تأثیر می‌گذارد، مخصوصاً این امر که ممکن است دانشجو با تعدادی از ممتحنین، از قبل آشنا نباشد و یا تجربه حضور در برخی از ایستگاه‌ها را نداشته باشد از عوامل بالقوه ایجاد خطای اندازه‌گیری محسوب می‌شود. از سوی دیگر دانشجویان در زمان‌های مختلف در هر ایستگاه حاضر شده‌اند و ممکن است خستگی ناشی از مدت آزمون، بر روی عملکرد آن‌ها و همچنین بر روی نحوه ارزیابی ممتحن از دانشجو تأثیرگذار باشد، بنابراین احتمال دارد دانشجویی که زودتر در یک ایستگاه حاضر شده باشد، نسبت به دانشجویی که دیرتر حاضر شده است، عملکرد بهتری داشته باشد و به همان نسبت، خستگی ممتحن هم می‌تواند منجر شود دانشجوی اول را با دقت بیشتری نسبت به دانشجوی آخر ارزیابی نماید.

در این پژوهش، کمترین سهم خطای اندازه‌گیری مربوط به مؤلفه تعامل دانشجویان با ایستگاه‌ها می‌باشد. داگلاس و لاوسون در پژوهش خود به این نتیجه رسیدند که بالا بودن میزان خطای این مؤلفه نمایان گر تفاوت ایستگاه‌ها از لحاظ دشواری است و بر این اساس می‌توان اذعان نمود در پژوهش حاضر، ایستگاه‌ها از لحاظ دشواری تفاوت چندانی نداشتند و این یکی از نقاط قوت دومین آزمون صلاحیت‌های بالینی مشهد به شمار می‌رود. در خصوص ارتباط بین جنسیت دانشجویان و نمرات دومین آزمون صلاحیت‌های بالینی، پس از تجزیه و تحلیل‌ها مشخص شد بین جنسیت دانشجو و نمره وی ارتباط معناداری وجود ندارد. عدم وجود تفاوت بین نمرات دانشجویان دختر و پسر مبین این امر است که جنسیت دانشجویان در عملکرد آن‌ها تأثیر چندانی نداشته است و دانشجویان دختر و پسر در تمام ایستگاه‌ها

ممتحنین با دانشجویان تعامل نداشتند و فقط ناظر بر کار آنها بودند اما در پژوهش چامبرز (۳۶)، بیمارناها هم با پزشک تعامل داشتند و هم عملکرد آنها را به عنوان بررسی می‌کردند و این تعامل، ممکن است موجب برخی سوگیری‌ها شده باشد که در نهایت در نمره دانشجو تأثیرگذار بوده است، علاوه بر این، ممتحنینی که در پژوهش حاضر دانشجویان را نمره دهی می‌نمودند همگی پزشک متخصص و یا دارای مدرک دکترای تخصصی بودند اما بیمارناهایی که در پژوهش چامبرز، عملکرد دانشجویان را بررسی می‌کردند، افراد آموزش‌دیده (غیرپزشک) بودند.

نتیجه‌گیری

پایایی یکی از شاخص‌های اساسی در تعیین کیفیت هر آزمون است و با توجه به مطالب اخیر، به‌طور کلی می‌توان گفت دومین آزمون صلاحیت‌های بالینی مشهد از پایایی بسیار خوبی برخوردار می‌باشد، اما نکته‌ای که باید مورد توجه قرار گیرد آن است که این ضریب بدون در نظر گرفتن تأثیر رویه بیمارناها بر روی خطای اندازه‌گیری محاسبه شده است که با توجه به مقالات یادگوسکی، والوند و لیساکلر (که در این مقالات، بیمارناها و موردها جزو مهم‌ترین عوامل ایجادکننده خطای اندازه‌گیری بودند)، می‌توان گفت این رویه، جزو عوامل بالقوه در ایجاد خطای اندازه‌گیری به شمار می‌رود و پیشنهاد می‌شود برای انجام پژوهش‌های مشابه، بیمارنا هم به عنوان یک رویه مورد بررسی قرار گیرد. همچنین نتیجه استفاده از تئوری تعمیم‌پذیری برای برآورد ضریب پایایی در این آزمون آسکی نشان داد این تئوری روشی مفید برای تعیین ضریب پایایی آزمون است که بر اساس آن می‌توان مشخص نمود چه عواملی در پایایی نقش دارند و با کنترل کردن این عوامل می‌توان پایایی را در آزمون‌های بعدی افزایش داد.

تقدیر و تشکر: پژوهشگر در پایان لازم می‌داند از همکاری دانشکده پزشکی مشهد، مرکز مهارت‌های بالینی بیمارستان قائم

تقریباً به یک نسبت (از لحاظ نمره) ظاهر شده‌اند. کارسون (۳۳) و همکارانش نیز در پژوهش خود در دانشگاه کالگری به این نتیجه رسیدند در ایستگاه‌های عمومی بین عملکرد دختران و پسران تفاوت معناداری وجود ندارد که این نتیجه با پژوهش حاضر همخوانی دارد و فقط در ایستگاه‌های حساس به جنسیت نظیر ایستگاه زنان، دانشجویان دختر نسبت به دانشجویان پسر عملکرد بهتری داشتند؛ اما کادی (۳۴) و همکارانش در مطالعه خود نتیجه گرفتند که در حیطه مهارت‌های بین فردی، جنسیت دانشجویان در نمره آنها تأثیر داشته و آزمودنی‌های زن نمرات بهتری نسبت به مردان کسب نموده بودند. نتایج پژوهش تاچی (۳۵) هم در ایستگاه معاینه کمر با مطالعه حاضر مشابهت داشت و تفاوتی بین جنسیت بیمارنا و دانشجو مشاهده نگردیده اما در ایستگاه معاینه غدد لنفاوی، بین عملکرد دانشجویان دختر و پسر تفاوت معناداری مشاهده شد و دانشجویان دختر، نمره بهتری نسبت به دانشجویان پسر کسب نمودند. نتایج بررسی‌های چامبرز (۳۶) نیز با پژوهش ما تفاوت دارد و در آن زنان نسبت به مردان نمره بهتری کسب نمودند و این امر بدون در نظر گرفتن جنسیت بیمارنا بود. همچنین برایلوسکی (۳۷) و همکارانش در سال ۱۹۹۷ در پژوهش خود به این نتیجه رسیدند که در سال‌های ۱۹۹۳ و ۱۹۹۴ دانشجویان دختر نسبت به دانشجویان پسر نتایج بهتری کسب نمودند و در سال ۱۹۹۵ دانشجویان پسر نسبت به دانشجویان دختر، بهتر ظاهر شدند و نتایج این پژوهش نیز با پژوهش حاضر هم‌راستا نیست.

نتایج تجزیه و تحلیل در خصوص ارتباط بین جنسیت ممتحنین و نمرات دومین آزمون صلاحیت‌های بالینی، نشان داد تفاوت معناداری بین جنسیت ممتحنین و نمرات دانشجویان وجود ندارد که این نتایج با نتایج پژوهش چامبرز و همکارانش همخوانی ندارد و در پژوهش آنها مشخص گردید بیمارناهای مرد، نسبت به بیمارناهای زن نمرات بهتری به پزشکان دادند، به عقیده پژوهشگر علت این امر این است که در پژوهش حاضر،

(عج) و کلیه افرادی که در این پژوهش مشارکت داشتند این مقاله مستخرج از پایان نامه دانشجویی نویسنده مسئول می
 تقدردانی نماید. باشد.

References

- Gupta P, Dewan P, Singh T. Objective structured clinical examination (OSCE) revisited. *Indian pediatrics*. 2010;47(11):911-20.
- Karimi-moonaghi H. *practical guide to medical sciences student assessment (focused on new methods in clinical assessment)*. Mashhad: Mashhad university of medical sciences; 2011.
- Koohpaye-zade J, Hemmati A, Baradaran H, Mirhosseini F, Akbari H, Sarvieh M. Survey on validity and reliability of Direct Observation of Procedural skills (DOPS) in clinical skills assessment of Midwifery students of Kashan School of Nursing and Midwifery. *Quarterly Journal of Sabzevar University of Medical Sciences*. 2014;21(1):145-54.
- Stilson F. *Psychometrics of OSCE standardized patient measurements*: University of South Florida; 2009.
- Barman A. Critiques on the objective structured clinical examination. *ANNALS-ACADEMY OF MEDICINE SINGAPORE*. 2005;34(8):478.
- Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of competency-based evaluation. *Archives of pediatrics & adolescent medicine*. 2000;154(7):736-41.
- El-Nemer A, Kandeel N. Using OSCE as an assessment tool for clinical skills: nursing students' feedback. *Australian Journal of basic and Applied sciences*. 2009;3(3):2465-72.
- Lawson D, Douglas M. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. *Journal of manipulative and physiological therapeutics*. 2006;29(6):463-7.
- Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics—AMEE guide no. 49. *Medical teacher*. 2010;32(10):802-11.
- Harden R. What is an OSCE? *Medical teacher*. 1988;10(1):19-22.
- Wang Z. Estimating reliability under a generalizability theory model for writing scores in c-base: University of Missouri-Columbia; 2005.
- Brennan RL. *Generalizability theory: Statistics for social science and public policy*. New York: Springer-Verlag Retrieved March. 2001;30.
- Kiamanesh A. Generalizability theory in Educational measurement. *Psychology and Educational Sciences Journal*. 2014;52(1):12-35.
- Alkahtani SF. *Oral performace scoring using generalizability theory and many-facet Rasch measurement: A comparison study*: The Pennsylvania State University; 2012.
- Cardinet J, Johnson S, Pini G. *Applying generalizability theory using EduG*: Taylor & Francis; 2011.
- Brennan RL. Coefficients and indices in generalizability theory. *Center for Advanced Studies in Measurement and Assessment, CASMA Research Report*. 2003;1:1-44.
- O'Brian N, O'Brian S, Packman A, Onslow M. Generalizability Theory I Assessing Reliability of Observational Data in the Communication Sciences. *Journal of Speech, Language, and Hearing Research*. 2003;46(3):711-7.
- Webb NM, Shavelson RJ. Generalizability theory: overview. *Wiley StatsRef: Statistics Reference Online*. 2005.
- Lengh CJ. *Generalizability Theory: Measuring the Dependability of Selected Methods for Scoring Classroom Assessments*: ERIC; 2010.
- Narayanan A, Greco M, Campbell JL. Generalisability in unbalanced, uncrossed and fully nested studies. *Medical education*. 2010;44(4):367-78.
- Ribar RJ. GENERALIZABILITY THEORY AND THE SOCIAL COMPETENCE PERFORMANCE CHECKLIST2005.
- Heitman RJ, Kovaleski JE, Pugh SF. Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. *Journal of athletic training*. 2009;44(1):48.
- Alizadeh E, Falsafi nejad Mr, Delavar A, Farrokhi Na, Abbaspoor A. Survey on reliability of Job Analysis Data be mans of Generalizability Theory. *quarterly journal of Educational Measurement*. 2014;14(4):1-23.
- Keller LA, Clauser BE, Swanson DB. Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in health sciences education*. 2010;15(5):717-33.
- Kiamanesh A. Generalizability theory in Educational measurement. *Psychology and Educational Sciences Journal*. 1993;2.
- Biaban gard E. *research methods in Psychology and Educational sciences*. Tehran: Doran; 2012. 488 p.

27. Brannick MTE-K, H Tugba; Prewett, Matthew. A systematic review of the reliability of objective structured clinical examination scores. *Medical Education*. 2011;45(12):1181-9.
28. Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. *Journal of manipulative and physiological therapeutics*. 2006;29(6):463-7.
29. Stilson FR. Psychometrics of OSCE standardized patient measurements: University of South Florida; 2009.
30. Sebok SS, Luu K, Klinger DA. Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. *Advances in health sciences education*. 2014;19:71-84.
31. Iramaneerat C, Yudkowsky R, Myford CM, Downing SM. Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement. *Advances in health sciences education*. 2008;13(4):479-93.
32. Vallevand ALC. Reliability, validity and sources of errors in assessing physician performance in an objective structured clinical examination: A generalizability theory analysis [Ph.D.]. Ann Arbor: University of Calgary (Canada); 2008.
33. Carson JA, Peets A, Grant V, McLaughlin K. The effect of gender interactions on students' physical examination ratings in objective structured clinical examination stations. *Acad Med*. 2010;85(11):1772-6.
34. Cuddy MM, Swygert K, Jobe A, Swanson DB. A Multi-level Analysis of Examinee Gender, Standardized Patient Gender and USMLE Step 2 Clinical Skills Communication and Interpersonal Skills Scores. 2009.
35. Touchie C, Wood T, Smees S, Humphrey-Murto S. Does the gender of the standardised patient influence candidate performance in an objective structured clinical examination? *Med Educ*. 2009;43(6):153-7.
36. Furman G, Colliver JA, Galofre A. Effects of student gender and standardized-patient gender in a single case using a male and a female standardized patient. *Academic Medicine*. 1993;68(4):301-3.
37. Brailovsky C, Grand'Maison P, Miller F, Rainsberry P. Detection of gender differences in high-stakes performance SP-based examinations in family medicine. *Advances in Medical Education: Springer*; 1997. p. 131-4.

Original Article

reliability of second clinical skills OSCE in Mashhad by use of Generalizability theory

Received:23/10/2017 - Accepted: 21/01/2018

Gholami Hasan¹
Sargazi samaneh^{2*}
Nobakht Maliheh³
Emadzadeh Ali⁴

1- instructor, medical education department, medical faculty, Mashhad university of medical sciences, Mashhad, Iran

2- Ph.D. student of medical education, Medical education department, Isfahan University of medical sciences, Isfahan, Iran

3- Master of medical education, Mashhad university of medical sciences, Mashhad, Iran

4- assistant professor, medical education department, medical faculty, Mashhad university of medical sciences, Mashhad, Iran

* Isfahan University of medical sciences, Isfahan, Iran.

Tel: 09376129605

Email: s.sargazi2010@gmail.com

Abstract

Introduction: OSCE is a combination of simulation tasks to assess medical students' skills in diagnosis and treatment of diseases. it is vital to measure the quality of an OSCE. One of the important steps in assessing the quality of a test is to determine it's reliability. This study aimed to determine the reliability coefficient of the second clinical skills test in Mashhad by using generalizability theory.

Subjects & Methods: The OSCE was conducted in two days and consists of 15 parallel stations. 75 students and 30 examiners participated in this test. Check lists collected and recorded in Excel. Variances and mean squares (MS) calculated by using SPSS. Variance for stations, students and examiners and generalizability and reliability coefficients calculated by using EDUG. Mann-Whitney U and Chi square tests were used for assessing relation between students' gender and score, raters' gender and score and relation between students gender and raters gender.

Results: There was no significant relationship between gender of students and their scores, gender of examiners and students' scores, and between gender of students and gender of examiners. Reliability coefficient was calculated 0/98 and interaction between students, examiners and station component had the largest effect on creating measurement error.

Conclusion: this exam had a good reliability coefficient and according to results we can say, using Generalizability Theory for estimating the reliability coefficient, measurement error, and to identify impact of any facets on measurement error is a logical way. However, more research needs to be done.

Keywords: OSCE, Generalizability theory, reliability, clinical skills

Acknowledgement: *There is no conflict of interest.*