زمستان ۱۳۹۸، دوره ۲۸، شماره ۴، صفحه ۱۳ تا ۲۵

دارای رتبه علمی پژوهشی
فصلنامه علمی پژوهشی طب و تزکیه
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

مقایسه تأثیر دو نوع شیوه ارزیابی کاغذی و مبتنی بر رایانه بر نتایج و عملکرد داوطلبان آزمون زبان وزارت بهداشت، درمان و آموزش پزشکی (MHLE)

آبتین حیدرزاده[۱]، سمانه پنجه علی بیک[۲*]، پگاه درخشان[۳]، مجید نوایی[۴]، حسن تورانی[۵]

تاریخ دریافت : ۱۳۹۸/۱۰/۱۹
تاریخ پذیرش : ۱۳۹۸/۱۲/۱۸

چکیده

زمینه و هدف: پژوهش حاضر به بررسی تأثیر دو نوع روش ارزیابی کاغذی و مبتی بر رایانه بر نتایج ارزیابی فرگیران در بُعد نمره نهایی فراگیران، پرداخته است.

روش بررسی: نظر به برگزاری آزمون زبان وزارت بهداشت، درمان و آموزش پزشکی، با هر دو نوع ابزار ارزیابی در یک زمان و با سؤالات یکسان، آزمون مذکور مورد تحلیل و بررسی قرار داده‌ایم. بدین منظور و برای کاهش خطای نمونه‌گیری، پنج دوره از این آزمون را که هم زمان در هر دو نوع کاغذی و مبتنی بر رایانه برگزار شده است، انتخاب و شاخص‌های روانسنجی همچون ضریب دشواری، ضریب تمیز و شاخص پایایی کودر ریچارسون، محاسبه و در بخش آمار استنباطی با استفاده از آزمون‌های من-ویتنی، تحلیل واریانس و شاخص اندازه اثر (η^2) با استفاده از نرم افزار SPSS و در سطح معنی داری، ۰،۰۵ به تحلیل، مقایسه نتایج داوطلبان پرداخته‌ایم.

یافته‌ها: نتایج نشان می‌دهند شاخص‌های روانسنجی آزمون‌ها در دوره‌های مختلف برگزاری آزمون در نوع مبتنی بر رایانه نسبت به کاغذی برتری دارد. همچنین میانگین نمرات شرکت‌کنندگان از ۲،۲ تا ۰،۵۴ به ترتیب در دوره‌های ۳ و ۵ بین آزمون مبتنی بر رایانه و کاغذی تفاوت داشت. به جز آزمون دوره ۳ نمرات سایر آزمون‌ها ، از توزیع یکسانی در نوع مبتنی بر رایانه و کاغذی پیروی می‌کردند.

نتیجه‌گیری: نتایج تحلیل مبین آن است که با سؤالات یکسان آزمون، در آزمون‌های مبتنی بر رایانه در مقایسه با همتای کاغذی آن، به طور کلی نتایج بهتری را به دست آورده‌اند و در عین حال ضریب افتراق آزمون در نوع مبتنی بر رایانه، بهتر بوده است.

کلمات کلیدی: ارتقاء صلاحیت‌های حرفه‌ای، صلاحیت‌های حرفه‌ای، مدیران آموزشی.

۱. گروه پزشکی اجتماعی، دانشکده پزشکی، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی گیلان، رشت، ایران
۲*. نویسنده مسئول، دکترای تخصصی (Ph.D) ریاضیات کاربردی، مرکز سنجش آموزش پزشکی، وزارت بهداشت درمان و آموزش پزشکی، تهران، ایران
۳. دکترای حرفه ای پزشکی، دانشکده پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران
۴. کارشناس ارشد آمار کاربردی، مرکز سنجش آموزش پزشکی، وزارت بهداشت درمان و آموزش پزشکی، تهران، ایران
۵. کاربرد در کامپیوتر، مرکز سنجش آموزش پزشکی، وزارت بهداشت درمان و آموزش پزشکی، تهران، ایران

دارای رتبه علمی پژوهشی
طب و تزکیه
فصلنامه علمی - پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

مقدمه

نظام آموزشی به عنوان یکی از شاخص‌های رشد و توسعه هر کشور، نیازمند بازنگری و بروزرسانی در ابعاد گوناگونی است. نگاهی گذرا به شاخصه‌های توسعه آموزش حاکی از آن است که ارزیابی‌های آموزشی پلی است میان درونداد و برونداد آموزشی. نتایج به دست آمده از هر آزمون هدایتگر مدرسان و سیاست‌گذاران حوزه آموزش در سنجش میزان و کنترل درک فرایند یادگیری بوده و فراتر از آن در پیش‌بینی عملکرد فراگیران در آینده نقشی تعیین کننده دارد[۱] بنابراین بدیهی است برای نیل به هدف نهایی، اصلاح ساختار میانی ضروری است.

بر اساتید و فعالان حوزه آموزش پوشیده نیست که نمره کسب شده در آزمون به تنهایی درخور اهمیت نیست. بلکه انتخاب روش‌های ارزیابی، نوع ارزیابی و پایش آن نیز تعیین کننده پیامدهای نظام آموزشی خواهد بود. در باب تنوع ارزیابی‌های آموزشی و تأثیر موارد مختلف بر نتایج آن در ادبیات موضوع، مقالات متعددی وجود دارد. در این میان آنچه کمتر در مورد آن بحث شده است، تأثیر شیوه اجرای ارزیابی نمرات آزمون فراگیران است.

در طی چهار دهه گذشته و با افزایش فناوری‌های برخط و رسانه‌ای، حرکت فرایندهای آموزشی به سوی ماشینی شدن در هر دوحیطه یادگیری و ارزیابی، گسترش یافته است. تاریخچه استفاده از آزمون‌های مبتنی بر رایانه به سال ۱۹۹۰ و کاربری آن در ارتش باز می‌گردد[۲]. امروزه اما سهولت در اجرا، سرعت در بازخورد، امکان ایجاد تنوع در آیتم‌های آزمون از نکات برجسته این روش نوین آزمون در مقایسه با نوع کاغذی و سنتی آن است.[۳] بالطبع راه‌اندازی و تجهیز یک مرکز آزمون مبتنی بر رایانه بسیار هزینه بردار است، لذا قبل از حرکت در هر مسیری مستلزم آن است که آیا رسیدن به مقصد نهایی ارزش سختی راه را دارد یا خیر؟

آزمون‌های مبتنی بر رایانه و مداد کاغذی را می‌توان در ابعاد مختلفی از هزینه تا اجرا مورد کنکاش قرار داد [۴-۶]. با این حال موضوع اصلی این پژوهش در تأثیر این دو شیوه بر عملکرد و نمرات آزمون‌دهندگان است که با پیشرفت روز افزون استفاده از فناوری‌های دیجیتال در آزمون، بسیار مورد توجه قرار گرفته است. در مطالعات اخیر میزان اضطراب آزمون رایانه‌ای بر داوطلبان بررسی و در نهایت نه تنها تفاوت معنا داری دیده نشده است [۷] بلکه حتی دانشجویانی باکمترین میزان آشنایی با رایانه نیز در این خصوص عملکرد مثبتی از خود نشان داده‌اند. نتایج سایر مطالعات مهر تأییدی بر عدم تأثیر اضطراب در این دو شیوه ارزیابی است [۸]. نمرات اکتسابی داوطلبان در این آزمون‌ها در مطالعات اخیر مورد بررسی قرار گرفته است که نتایج تحلیل‌ها اثبات در برخی از آنان نتایج آزمون‌های مبتنی بر رایانه بر کاغذی، الویت داشته است. [۷، ۹]

در پژوهش حاضر، نتایج ۵ دوره از آزمون زبان وزارت بهداشت، درمان و آموزش پزشکی که تا سال ۱۳۹۸ پیش‌شرط ورود به دوره های دکتری تخصصی این وزارتخانه بوده است، مورد بررسی و تحلیل قرار گرفته است. این آزمون‌ها از نیمه دوم سال ۱۳۹۶ به هر دو روش کاغذی و مبتنی بر رایانه، برگزار شده‌اند. در بررسی نتایج آزمون، داوطلبان در دو دسته تقسیم بندی شده‌اند و نتایج هر دوره در هر دو بعد شاخص‌های کیفیت آزمون و نوع توزیع نمرات، مورد واکاوی قرار گرفته است.

روش پژوهش

تحلیل سؤالات آزمون از مهمترین مؤلفه‌های ارزیابی است که با بررسی هر سؤال و تعیین میزان دقت آن در

یاران رشته علمی پژوهشی
طب و تزکیه
فصلنامه علمی پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

تمیز و نارسایی‌های آن در نهایت مبین نقاط قوت، ضعف و کیفیت هر ارزیابی است. یافته‌های این تحقیق در دو بخش توصیفی و استنباطی ارایه شده‌اند. به منظور بررسی شاخص‌های روانسنجی سؤالات آزمون از محاسبه سنجه‌های ضریب تمیز درجه دشواری با استفاده از روش کلاسیک استفاده نموده‌ایم و سپس عملکرد داوطلبان را در هر دو نوع آزمون بر اساس نمره نهایی ایشان مورد تحلیل قرار داده‌ایم.

بدین منظور با بکارگیری آماره‌های توصیفی شامل (میانگین و انحراف معیار) و در بخش استنباطی، با استفاده از آزمون تحلیل واریانس به بررسی تفاوت میان نتایج آزمون‌های مبتنی بر رایانه و کاغذی پرداخته‌ایم. در بررسی پایایی آزمون از روش کودر ریچاردسون استفاده شده است. برای مقایسه دو نوع شیوه ارزیابی، داوطلبان در دو گروه شرکت‌کنندگان آزمون کاغذی، با نام‌گذاری گروه ۱ و شرکت کنندگان آزمون مبتنی بر رایانه، گروه ۲ و در ۵ دوره متوالی، مورد تحلیل قرار گرفته‌اند.

آزمون MHLE:

کسب حداقل توانمندی‌ها در حیطه‌های مختلف زبان انگلیسی یکی از ملزومات اصلی در دوره‌های دکتری تخصصی(Ph.D.) وزارت بهداشت، درمان و آموزش پزشکی است. تا سال ۹۸ لازم بود متقاضیان از آزمون‌های زبان معتبر که در سطح کشور در حال برگزاری است همانند MSRT، TOFEL،IELTS و TOLIMOکف حداقلی برای شرکت در آزمون و ادامه تحصیل در مقاطع دکتری تخصصی، کسب نمایند. در این بین آزمون اختصاصی زبان وزارت بهداشت، درمان و آموزش پزشکی (MHLE)[1] به دلیل مقرون به صرفه بودن آن و ارتباط مستقیم با حوزه وزارت بهداشت، یکی از آزمون‌هایی که بیشترین تعداد متقاضیان را دارا می‌باشداین آزمون در هر سال حداقل شش دوره برگزار می‌گردد. داوطلبان در مدت ۹۰ دقیقه به ۱۰۰ پرسش در سطح آزمون استاندارد تافل (۳۰ پرسش درك مطلب شنیداری، ۴۰ پرسش واژگان، ساختار و دستور زبان و ۳۰ پرسش درك مطلب نوشتاری) پاسخ می‌دهند و نمره حداکثر برای این آزمون ۱۰۰ (بدون نمره منفی) است.

با توجه به زیر ساخت‌ها و ظرفیت کنونی مراکز آزمون مبتنی بر رایانه دانشگاه‌های علوم پزشکی کشور و تعداد کثیر داوطلبان، این آزمون به طور هم زمان به دو صورت کاغذی و مبتنی بر رایانه به خصوص در برخی دانشگاه‌های مادر کلان مناطق آمایشی، برگزار می‌گردد. شایان ذکر است بدون در نظر گرفتن تفاوت ابزار دو آزمون با توجه به نام گذاری، از مهمترین تفاوت‌های میان دو آزمون (کاغذی و مبتنی بر رایانه) می‌توان به تفاوت در اجرای بخش شنیداری اشاره نمود.

شاخص‌های تحلیل و روانسنجی سؤالات آزمون:

هدف از تحلیل و روانسنجی بررسی یک به یک سؤالات به منظور تعیین میزان دقت و نارسایی‌های هرسؤال است. به وسیله شاخص‌های روانسنجی، نقاط قوت و ضعف هر آزمون و کیفیت کلیه سؤالات آن، معین خواهد شد. بعد از اجرای هر آزمون، با نتایج به دست آمده از تحلیل سؤالات و نتایج حاصل اساتید قادر به تجدید نظر در آزمون و بهبود کیفیت سؤالات ارزیابی‌های آتی خواهند بود.

به طور خلاصه شاخص‌های زیر به عنوان پیامد نهایی در بررسی آزمون با تعاریف زیر بررسی می‌گردد و در بخش‌های بعدی پژوهش حاضر، مورد استفاده قرار می‌گیرند.

1 Ministry of Health Language Examination

دارای رتبه علمی پژوهشی
طب و توانبخشی
فصلنامه علمی - پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

- **ضریب دشواری**: ضریب‌های دشواری بین ۰/۳ تا ۰/۷ حداکثر اطلاع را درباره تفاوت بین آزمودنی‌ها به دست می‌دهند و نشان از مناسب بودن سؤال می‌باشد. هر اندازه ضریب دشواری یك سؤال بزرگ‌تر (به یک نزدیکتر) باشد، آن سؤال آسان‌تر است و هر اندازه که این ضریب کوچك‌تر (به صفر نزدیك‌تر) باشد سؤال دشوارتر است. بنابراین افزایش ضریب دشواری، با افزایش سهولت آزمون رابطه‌ای مستقیم دارد. در نتیجه به جای ضریب دشواری می‌توان از ضریب آسانی یا سهولت نام برد.
- **ضریب تمیز**: ضریب تمیز قدرت سؤال را در تمایزگذاری بین گروه قوی و گروه ضعیف آزمون‌دهندگان مشخص می‌کند (دامنه تغییرات این ضریب بین ۱- تا ۱+است)، ضریب تمیز بزرگ‌تر نشانگر بهینه بودن قوه تمیز آن سؤال است و برعکس، ضریب تمیز صفر حاکی از آن است که آن سؤال نتوانسته بین دو گروه قوی و ضعیف، تمایزی قائل شود. ضریب تمیز منفی نشانگر این است که داوطلبان ضیف تر بهتر از داوطلبان قوی‌تر به آن سوال پاسخ داده‌اند.
- **کودر ریچاردسون**: برای ارزیابی پایایی آزمون از روش کودرریچاردسون استفاده می‌شود. این روش، که به اسم مبدع آن شهرت دارد در سال ۱۹۳۷ ابداع شد برای متغیرهای اسمی‌دو وجهی با کدهای صفر و یک طراحی شده است. این روش میزان همسازی درونی آزمون، یعنی میزان تداخل همه سؤالات از لحاظ سنجش یک ویژگی مشترک را ارزیابی می کند. در این روش بر اساس همبستگی درونی سؤالات، شاخص کودر ریچاردسون استخراج می‌شود و چنانچه مقدار این شاخص بیش از ۰،۷ باشد، می‌توان گفت ابزار سنجش (کل آزمون) پایایی قابل قبول دارد.

شاخص عملکرد دانشجویان:

در ادامه توزیع نمرات در آزمون‌های برگزار شده و تفاوت در نحوه توزیع آن نیز مورد بررسی قرار گرفته است. بدین منظور برای مقایسه توزیع نمرات هر دو گروه، از آزمون من-ویتنی استفاده شده است.

علاوه بر آن، برای مقایسه متوسط عملکرد داوطلبان هر دو گروه آزمون‌دهندگان، از آزمون تحلیل واریانس و شاخص اندازه اثر[1] مجذور اتا (η^2) و نرم افزار SPSS استفاده شده است. ابتدا کل داوطلبان شرکت کننده در هر دو نوع روش آزمون مورد بررسی قرار گرفته اند و سپس برای بررسی بهتر و دقیق تر هر دو گروه آزمون دهنده در هر دوره نیز مقایسه شده اند.

یافته‌ها و نتایج

در ۵ دوره آزمون بررسی شده، تعداد شرکت کنندگان در نوع کاغذی از دوره ۱ تا ۵ به ترتیب، ۱۹۸۷، ۱۱۶۴، ۶۰۶، ۱۵۸۷ و ۲۲۹۲ و تعداد شرکت کنندگان در آزمون مبتنی بر رایانه از دوره ۱ تا ۵ به ترتیب برابر؛ ۸۶۵ ، ۵۶۸، ۷۶۳، ۹۷۷ و ۱۲۲۳ است.

شاخص‌های روانسنجی سؤالات آزمون:

در جدول ۱ تحلیل شاخص‌های روانسنجی برای ۵ دوره از هر دو روش آزمون به تفکیک ارائه شده است. بررسی یافته‌های شاخص ضریب دشواری در جدول ۱ در تطبیق با تعاریف شاخص‌های روانسجی ارائه شده در بخش قبل، نشان می‌دهد که سؤالات یکسان آزمون برای آزمون‌دهندگان گروه ۲، در مقایسه با گروه ۱ آسانتر بوده است. همچنین از تعریف شاخص ضریب تمیز، می‌توان نتیجه گرفت در این ۵ دوره آزمون، آزمون‌های رایانه‌ای

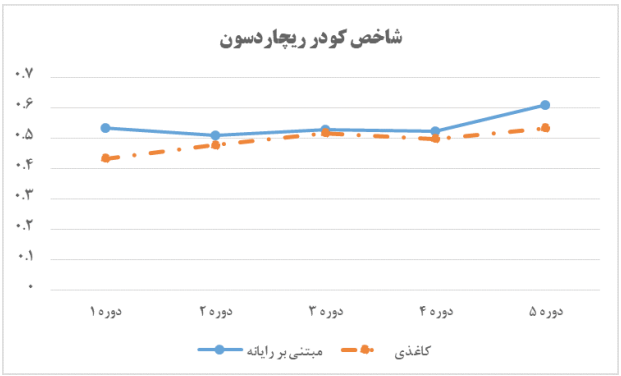
1 Effect Size

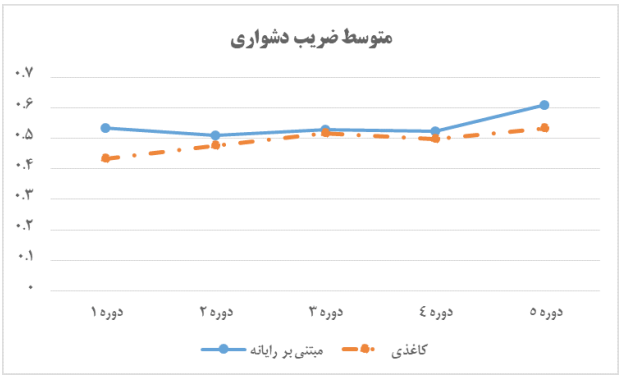
طب و زکیه
فصلنامه علمی - پژوهشی
دارای رتبه علمی پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

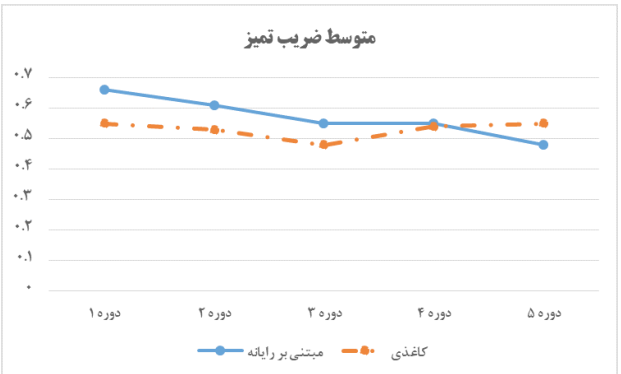
قدرت بیشتری در افتراق داوطلبان گروه ضعیف از قوی دارند. نتایج حاصله از شاخص کودر ریچاردسون برای هر دو نوع آزمون طی دوره‌های مختلف، نشان از عدم پایایی آزمون‌ها در هر ۵ دوره مورد بررسی می‌باشد. نمودار ۱ به مقایسه نتایج به دست آمده در هر دو گروه از آزمون‌دهندگان، پرداخته است.

جدول ۱. نتایج ضرایب دشواری، تمیز و شاخص پایایی کودر ریچاردسون

گروه آزمون	دوره آزمون	متوسط ضریب تمیز	متوسط ضریب دشواری	شاخص کودر ریچاردسون
گروه ۲ مبتنی بر رایانه	دوره ۱	۰,۶۶	۰,۳۷	۰,۵۳۳
	دوره ۲	۰,۶۱	۰,۳۳	۰,۵۰۹
	دوره ۳	۰,۵۵	۰,۳۴	۰,۵۲۹
	دوره ۴	۰,۵۵	۰,۳۲	۰,۵۲۲
	دوره ۵	۰,۴۸	۰,۳۶	۰,۶۰۹
گروه ۱ کاغذی	دوره ۱	۰,۵۵	۰,۲۹	۰,۴۳۳
	دوره ۲	۰,۵۳	۰,۲۸	۰,۴۷۷
	دوره ۳	۰,۴۸	۰,۲۹	۰,۵۱۷
	دوره ۴	۰,۵۴	۰,۲۷	۰,۴۹۸
	دوره ۵	۰,۵۵	۰,۳۱	۰,۵۳۳







نمودار ۱. شاخص‌های کیفیت آزمون در دوره‌های مختلف آزمون‌های کاغذی و مبتنی بر رایانه

فصلنامه علمی - پژوهشی طب و تزکیه
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

مقایسه توزیع نمرات

برای مقایسه توزیع نمرات داوطلبان، از آزمون من- ویتنی استفاده و نتایج آن در هر دوره، در جدول ۲، گزارش شده است. بررسی این جدول و نتایج آزمون من- ویتنی نشان می‌دهد، تفاوت معناداری بین توزیع نمرات آزمون‌دهنگان در دو شیوه اجرایی در هر یک از دوره‌های ۱، ۲، ۴ و ۵ وجود ندارد بنابراین دو روش از یک نوع توزیع برخوردارند (P> ۰،۰۵).

در دوره ۳ تفاوت معناداری بین توزیع نمرات داطلبان در دو روش مشهود بوده و دو روش آزمون‌گیری، دارای توزیع نمرات یکسان نیستند. بر اساس آزمون کای دو[1] می‌توان گفت توزیع داده‌های کاغذی (Statistics=۰،۰۸) نسبت به مبتنی بر رایانه(Statistics=۰،۱۱) به نرمال نزدیک‌تر است.

جدول۲. نتایج آزمون Mann-Whitney U

دوره	Mann-Whitney U	Z	P
دوره ۱	۸۳۷۲۴۴،۵	-۰،۹۰۶	۰،۳۶۵
دوره ۲	۳۱۹۹۵۲	-۱،۰۸۸	۰،۲۷۷
دوره ۳	۲۰۹۹۴۷،۵	-۲،۹۲۵	۰،۰۰۳
دوره ۴	۷۵۹۳۹۰	-۰،۸۷۲	۰،۳۸۳
دوره ۵	۱۳۷۷۸۵۷،۵	-۰،۸۲۷	۰،۴۰۸

مقایسه عملکردی داوطلبان در دوگروه آزمون کاغذی و مبتنی بر رایانه

در این بخش به منظور مقایسه متوسط عملکرد داوطلبان، از تحلیل واریانس استفاده شده است. بدین منظور ابتدا نتایج کل شرکت‌کنندگان در هر دوره از آزمون را تحلیل و یافته‌ها در جدول ۳گزارش شد.

جدول ۳. نتایج تحلیل واریانس در شرکت‌کنندگان به تفکیک دوره در دوگروه آزمون‌دهندگان

دوره	گروه	کل شرکت‌کنندگان					
		میانگین	انحراف معیار	تعداد	f	Sig	اندازه اثر (η^2)
۱	رایانه (گروه۲)	۴۵،۲۰	۱۴،۰۷	۸۶۵	۳،۴۶۴	۰،۰۶۳	۰،۰۳۵
	کاغذی(گروه ۱)	۴۴،۲۱	۱۲،۵۲	۱۹۷۸			
۲	رایانه(گروه ۲)	۳۴،۹۵	۱۱،۰۹	۵۶۸	۵،۲۷۶	۰،۰۲۲	۰،۰۵۵
	کاغذی (گروه ۱)	۳۳،۸۱	۸،۹۵	۱۱۶۴			
۳	رایانه (گروه ۲)	۴۲،۹۱	۱۱،۶۰	۷۶۳	۱۳،۹۱۰	۰،۰۰۰	۰،۱۰۰
	کاغذی (گروه ۱)	۴۰،۷۱	۹،۷۵	۶۰۶			
۴	رایانه (گروه ۲)	۴۴،۱۰	۱۲،۴۳	۹۷۷	۰،۶۴۷	۰،۴۲۱	۰،۰۱۶
	کاغذی (گروه ۱)	۴۳،۷۰	۱۲،۲۸	۱۵۸۷			
۵	رایانه (گروه ۲)	۴۳،۷۸	۱۱،۶۱	۱۲۲۳	۱،۸۹۶	۰،۱۶۹	۰،۰۲۳
	کاغذی(گروه ۱)	۴۳،۲۴	۱۰،۸۲	۲۲۹۲			
کل	رایانه (گروه ۲)	۴۲،۸۴	۱۲،۶۳	۴۳۹۶	۱۵،۰۸۵	۰،۰۰۰	۰،۰۳۲
	کاغذی (گروه ۱)	۴۱،۹۵	۱۱،۸۱	۷۶۲۷			

1 Chi-Square

دارای رتبه علمی پژوهشی
طب و تزکیه
فصلنامه علمی، پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

در ادامه با توجه به تأثیر بیشتر نمرات پذیرفته‌شدگان بر نتایج آزمون، همگن‌سازی نتایج و بررسی صحت نتایج به گونه‌ای دقیق‌تر، نمرات پذیرفته‌شدگان هر دو روش در هر دوره نیز مورد آزمایش قرار گرفت تا شرایط کمی همگن‌تر گردد و نتایج آن در جدول ۵ گزارش شد.

نتایج به دست آمده از تحلیل واریانس برای کل شرکت‌کنندگان (جدول ۳) نشان می‌دهد تفاوت نمرات آزمون‌های هر دو گروه آزمون‌دهندگان در دوره‌های ۲ و ۳ معنادار بوده است ($P< ۰,۰۵$). بر اساس میانگین نمرات در هر دو دوره ۲ و ۳ میانگین نمرات داوطلبان آزمون مبتنی بر رایانه بالاتر از میانگین نمرات داوطلبان آزمون کاغذی است. دوره ۱، ۴ و ۵ میانگین نمرات در دو گروه تفاوت معناداری نداشته است. ($P> ۰,۰۵$). شاخص η^2 نشان می‌دهد، تفاوت میان متوسط میانگین نمرات دو گروه شرکت‌کنندگان دارای بیشترین میزان اختلاف در دوره ۳ و با ضریب ۱۰ درصد بوده است. در مجموع ۵ دوره نیز تفاوت بین نمرات دو گروه شرکت کننده در آزمون معنادار شده است ($P < ۰,۰۵$) و میانگین نمرات داوطلبان آزمون مبتنی بر رایانه بالاتر از میانگین نمرات داوطلبان آزمون کاغذی است. میزان اندازه اثر نشان می‌دهد در مجموع ۵ دوره اختلاف میان دو گروه حدود ۳ درصد بوده است. بنابراین با توجه به تفسیر شاخص η^2 در آزمون تحلیل واریانس[۱۰] که در جدول ۴ گزارش شده است نوع آزمون در دوره ۳، تأثیری متوسط و در سایر دوره‌ها و مجموع ۵ دوره، تأثیر کمی بر تغییر نتایج نمرات شرکت‌کنندگان داشته‌است.

جدول ۴. کلاس‌های اندازه اثر [۱۰]

ردیف	مقدار η^2	اندازه اثر
۱	۰,۰۱	کم
۲	۰,۰۶	متوسط
۳	۰,۱۴	زیاد

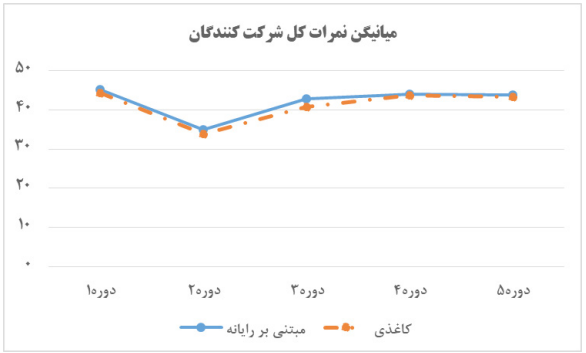
جدول ۵. نتایج تحلیل واریانس در پذیرفته‌شدگان به تفکیک دوره در دو گروه آزمون‌دهندگان

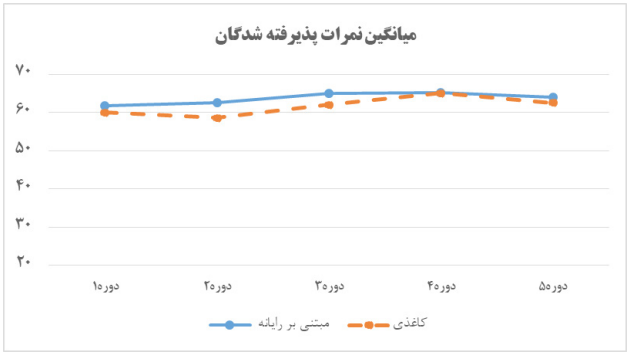
دوره	گروه	کل پذیرفته‌شدگان					
		میانگین	انحراف معیار	تعداد	f	Sig	اندازه اثر (η^2)
۱	رایانه(گروه ۲)	۶۱,۸۰	۱۱,۶۸	۲۶۴	۵,۶۱	۰,۰۱۸	۰,۲۸۸
	کاغذی (گروه۱)	۵۹,۹۶	۹,۶۲	۵۴۶			
۲	رایانه(گروه ۲)	۶۲,۵۴	۱۰,۶۲	۴۸	۴,۵۳۲	۰,۰۳۶	۰,۲۰۲
	کاغذی (گروه ۱)	۵۸,۵۴	۸,۹۹	۶۱			
۳	رایانه(گروه ۲)	۶۴,۹۵	۸,۰۸	۱۰۷	۴,۹۹۸	۰,۰۲۷	۰,۱۷۶
	کاغذی (گروه۱)	۶۲,۰۶	۶,۷۱	۵۲			
۴	رایانه(گروه۲)	۶۵,۲۴	۹,۵۲	۱۶۰	۰,۰۴۱	۰,۸۳۹	۰,۰۱
	کاغذی (گروه۱)	۶۴,۰۸	۸,۸۶	۲۵۵			
۵	رایانه(گروه ۲)	۶۴,۰۸	۸,۳۲	۱۸۳	۵,۴۰۷	۰,۰۲	۰,۱۰۳
	کاغذی (گروه۱)	۶۲,۴۹	۶,۷۵	۳۲۰			
کل	رایانه(گروه ۲)	۶۳,۵۶	۱۰,۰۴	۷۶۲	۱۸,۸۶۱	۰,۰۰۱	۰,۰۹۵
	کاغذی (گروه ۱)	۶۱,۶۹	۸,۸۹	۱۲۳۴			

طب و تزکیه
فصلنامه علمی - پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

ستون دوم جداول ۳ و ۵ تحلیل واریانس جامعه آزمون را کل پذیرفته‌شدگان در نظر گرفته است. این کار به نوعی داده‌های پرت را کنار گذاشته و نتیجه به دست آمده صحت فرضیه‌های محقق را بهتر نشان می‌دهد چراکه می‌توان گفت مقایسه پذیرفته‌شدگان هر دو آزمون همگن‌تر می‌باشد. همچنین نتایج به دست آمده از تحلیل واریانس برای گروه پذیرفته شدگان نشان می‌دهد تفاوت نمرات پذیرفته‌شدگان هر دو گروه در دوره‌های ۱، ۲، ۳ و ۵ معنادار بوده است (P<۰،۰۵). بر اساس میانگین نمرات در هر ۴ دوره میانگین نمرات داوطلبان گروه ۲، بالاتر از میانگین نمرات پذیرفته‌شدگان آزمون کاغذی است. دوره ۴ میانگین نمرات در دو گروه تفاوت معناداری نداشته است. (P>۰،۰۵).

شاخص اندازه اثر نشان می‌دهد، تفاوت میان دو گروه پذیرفته‌شدگان بیشترین میزان اختلاف را در دوره ۱ و با ضریب ۲۸/۸ درصد دارا بوده که با توجه به جدول ۴، تغییر نوع آزمون بر نتایج داوطلبان در این دوره و دوره‌های ۲ و ۳، تأثیر زیاد، در دوره ۴ تأثیر کم و در دوره ۵ تأثیر متوسط در تغییر نتایج نمرات داوطلبان، داشته است. در مجموع ۵ دوره نیز تفاوت بین نمرات دو گروه پذیرفته‌شدگان آزمون مبتنی بر رایانه و کاغذی در آزمون معنادار شده است (P<۰،۰۵) و میانگین نمرات پذیرفته شدگان گروه۲ بالاتر از میانگین نمرات پذیرفته شدگان گروه ۱ است. میزان اندازه اثر نشان می‌دهد در مجموع ۵ دوره، اختلاف میان دو گروه حدود ۱۰ درصد بوده است. در نمودار ۲، میانگین نمرات پذیرفته شدگان و شرکت‌کنندگان در ۵ دوره آزمون به تفکیک کاغذی و رایانه‌ای داده شده است.





نمودار ۲. میانگین نمرات پذیرفته‌شدگان و شرکت‌کنندگان در ۵ دوره آزمون به تفکیک کاغذی و رایانه‌ای

بحث و نتیجه‌گیری

با رشد سریع اقتصاد و تکنولوژی، ساخت روش‌های جدید آموزشی که از نظر اقتصادی و کارایی برتر باشد، بسیار ضروری است[۱۱].

آزمون‌های رایانه‌ای را می‌توان انقلابی در تاریخ آزمون‌سازی دانست که از سال ۱۹۹۰ آغاز و با سرعت رشد و توسعه یافته است. امروزه در کشور ما نیز به طور وسیعی آزمون‌های مبتنی بر رایانه در حال برگزاری و فضای خالی برای پژوهش در خصوص کارایی و تفاوت‌هایآن با نوع سنتی وجود دارد[۱۲، ۱۳]. در این پژوهش با توجه به برگزار شدن آزمون MHLE به صورت همزمان و با سؤالات یکسان در هر دو حالت رایانه‌ای و مداد-کاغذی، این فرصت در اختیار ما قرار داد تا به مقایسه این دو روش، در ۵ دوره مختلف بپردازیم.

برای آنکه بتوان نتایج دو گروه آزمون دهندگان را بایکدیگر مقایسه کرد، می‌توان از توزیع آماری نمرات کسب شده از آزمون کمک گرفت. بدین صورت که اگر دو آزمون از توزیع یکسانی پیروی کنند، سنجش آنها نیز

یاران رتبه علمی پژوهشی
طب و تزکیه
فصلنامه علمی - پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

یکسان می‌باشد.

در یافته‌ها، ما بدین نتیجه رسیدیم که در توزیع نمرات هر دو گروه از آزمون‌دهندگان تفاوت معنا داری وجود نداشته است و تنها آزمونی که توزیع متفاوت داشته است، آزمون دوره ۳ بوده است که توزیع نمرات دو گروه از آزمون‌دهندگان با یکدیگر تفاوت معنادار نشان داده است. در این دوره نتایج آزمون‌دهندگان گروه ۲ نزدیکی بیشتری به توزیع نرمال داشته است. هر چند که پیروی از توزیع نرمال را نمی‌توان به عنوان برتری یک آزمون دانست، اما دوره ۳ در سایر قسمت‌های آنالیز مربوط به میانگین و تحلیل آزمون تفاوت بین دو شیوه آزمون مشهودی دیده می‌شد که نیاز به بررسی بیشتری دارد.

با وجود اینکه مطالعاتی که به بررسی تفاوت توزیع نمرات در بین دو نوع آزمون کاغذی و میتنی بر رایانه پرداخته‌اند بسیار کم است و یا وجود ندارد؛ اما برخی نویسندگان به بررسی همبستگی این دو نوع آزمون پرداخته‌اند که از بین ۳۲ مطالعه مورد بررسی در این زمینه، ۳۰مورد همبستگی بالای ۰،۷۵ نشان داده‌اند[۱۴]. به طور کلی میانگین وزن داده شده همبستگی در مطالعات بررسی شده ۰،۹ بوده است که این نشان دهنده وضعیت نسبی ثابت توزیع نمرات بعد از تبدیل آزمون به رایانه‌ای است که با نتایج ما همراستا می‌باشد. پژوهش حاضر، همانند سایر مطالعات مبین آن است که آزمون رایانه‌ای و کاغذی در توزیع نمرات توافق بالایی از خود نشان می‌دهند.

از سوی دیگر آزمون‌های مبتنی بر رایانه، نیازمند پیش نیازهایی از نظرآشنایی با رایانه و تجربه استفاده از آن را دارد. با توجه به این موضوع، مطالعات سعی در یکسان‌سازی متغیرهایی ازجمله سن، تجربه استفاده از رایانه و پلتفورم رایانه و سپس انجام آنالیز در بین گروه‌ها کرده‌اند. شواهد اندکی وجود دارد که پلتفورم کوچکتر باعث همخوانی کمتر با مدل کاغذ مدادی می‌شود[۱۵]. از سوی دیگر با افزایش سن کاهش مختصری در همخوانی بین دو روش وجود داشته که این میزان معنا دار نبوده است[۱۴]. در مطالعه ما با توجه به این آزمون MHLE جز آزمون‌های پیش نیاز آزمون دکتری تخصصی وزارت بهداشت، درمان و آموزش پزشکی می‌باشد و تمام شرکت‌کنندگان حداقل مدرک کارشناسی ارشد را دارا بوده و از نظر سنی و سطح تحصیلات، در یک سطح هستند. همچنین با توجه به کاربری گسترده رایانه در دانشگاه‌ها در دوره کارشناسی و کارشناسی ارشد، بین گروه‌ها از نظر سطح آشنایی با رایانه، تفاوت معناداری وجود ندارد. مراکز آزمون مبتنی بر رایانه همگی تحت نظارت مرکز سنجش آموزش پزشکی مجهز شده و از نظر تکنیکال و نوع پلتفروم استاندارد ارزیابی دوره‌ای می‌گردند.

نتایج ما نشان داده است که میانگین نمرات کسب شده در بین شرکت‌کنندگان دو گروه آزمون دهنده فقط در دو نوبت با یکدیگر تفاوت معنا دار داشته‌اند که در هر دو نوبت آزمون‌دهندگان گروه ۲ میانگین بالاتری کسب کرده‌اند. بیشترین تفاوت بین دو گروه در آزمون ۳ با ۲،۲ اختلاف و کمترین تفاوت در آزمون دوره ۵ بوده که این تفاوت به ۰،۵۴، تقلیل یافته است. همچنین شاخص اندازه‌گیری شده اندازه اثر، در همین دو دوره بالاترین مقدار بین سایر دوره‌ها، بوده که به ترتیب ۰،۱۰۰ و ۰،۰۵ است. در میان مطالعات مختلف در حوزه آزمون‌های پزشکی تفاوت بین میانگین نمرات آزمون مبتنی بر رایانه و کاغذی بین ۰،۱ درصد تا ۵،۸ درصد می‌باشد[۱۶، ۱۷]. در مطالعه متا آنالیز انجام شده به طور میانگین ۰،۲ درصد تفاوت بین هر دو آزمون، عنوان شده است که به بیان دیگر در یک آزمون ۱۰۰ سوالی،

دارای رتبه علمی پژوهشی
طب و توانبخشی
فصلنامه علمی - پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

میانگین نمرات در آزمون رایانه‌ای، به میزان ۰،۲ درصد بالاتر از کاغذی بوده، که این اختلاف بین دو آزمون معنادار نمی‌باشد. و در ۹۳٪ مطالعات این میانگین بین ۵- تا ۵+ درصد بوده است[۱۴]. این تفاوت اندک بین دو آزمون نشان دهنده آن است که تفاوت در تکنیک برگزاری آزمون اثر مثبت یا منفی چشمگیری بر نتایج آزمون تمامی شرکت کنندگان نخواهد گذاشت. این در حالی است که اکثر مطالعات صرفاً بر روی نتایج شرکت کنندگان هر آزمون انجام شده است و با توجه به اینکه هتروژنیسیتی زیادی بین شرکت‌کنندگان هر آزمون وجود دارد، بررسی میانگین نمرات در پذیرفته‌شدگان این امکان را به ما می‌دهد که میانگین نمرات را درگروه همگن‌تری، بررسی کنیم. در طی این بررسی مشاهده می‌کنیم که میانگین نمرات در بین گروه‌های آزمون رایانه‌ای و کاغذی در همه نوبت‌های آزمون، به جز نوبت چهارم تفاوت معنی‌دار دارد و ارجحیت میانگین، با نمرات آزمون رایانه‌ای می‌باشد. این موضوع با وجود اینکه تفاوت بین دو نوع آزمون را نشان می‌دهد ولی بالا بودن میانگین نمرات آزمون دهندگان گروه ۲، نسبتی به گروه ۱، در همه نوبت‌های برگزار شده آزمون، نشان دهنده آن است که متغیرهایی مانند استرس و عدم آشنایی با رایانه و سایر موارد، نمی‌تواند اثر منفی بر آزمون بگذارند. مطالعات در خصوص آزمون‌های مبتنی بر رایانه، بیان داشته‌اند، که کاهش میزان استرس داوطلبان در پاسخگویی به سؤالات و مسائل فنی مانند سرعت نمایش سؤالات، علائم گرافیکی و متنی، کیفیت نوشتاری، تنظیمات صفحه نمایش، سرعت تجهیزات سخت افزار، نرم‌افزارهای اصلی و جانبی مربوطه، زمان اولیه ورود به آزمون به عنوان بخش مهمی از هر آزمون مبتنی بر رایانه در ارتقاء کیفیت آزمون تأثیر زیادی دارد[۴، ۱۸]. همچنین در آزمون‌های مداد کاغذی در برخی از موارد، پاسخنامه داوطلب مخدوش شده و یا به دلیل خطای انسانی پاسخنامه اشتباه به داوطلب داده می‌شود[۶، ۱۹، ۲۰]. که همه این موارد می‌توانند علتی برای برتری نمرات کسب شده در این نوع آزمون باشد. از سوی دیگر بررسی ما در تحلیل آزمون‌ها و با توجه به اینکه سؤالات درهر شیوه آزمون‌گیری یکسان بوده است، به این نتیجه رسید که ارزیابی مبتنی بر رایانه‌ای، در شرایط یکسان شاخص‌های روانسجی، ضریب دشواری و ضریب تمیز بالاتری نسبت به همتای مداد کاغذی خود در تمام دوره‌ها نشان داده است که این موضوع نشان می‌دهد آزمون‌های رایانه‌ای، علاوه بر کاهش اضطراب دانشجویان، بر کیفیت و سنجش خود آزمون نیز می‌افزاید.

نتایج ما نشان می‌دهد که آزمون مبتنی بر رایانه از نظر شاخص‌های تحلیل آزمون شامل ضریب دشواری و ضریب تمیز برتری نسبت به آزمون مداد کاغذی دارد. میانگین نمرات شرکت‌کنندگان در آزمون مبتنی بر رایانه، بالاتر بوده است، که می‌تواند نشان از کاهش استرس و افزایش میزان پاسخگویی آنها باشد. علاوه بر این توزیع نمرات در بین دو آزمون مبتنی بر رایانه و کاغذی در اکثر نوبت‌های آزمون، یکسان می‌باشد.

فصلنامه علمی - پژوهشی طب و تزکیه
دارای رتبه علمی پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

منابع

1. Jalili M, K.M.M., Gandomkar R, Mortaz Hejri S. , *Principles and Methods of student assessment in health professions.* 1nd ed Tehran: Iranian Academy of Medical Sciences. ; 2017. **[Persian]**.
2. Sands, W.A., B.K. Waters, and J.R. McBride, *Computerized adaptive testing: From inquiry to operation*. 1997: American Psychological Association.
3. Thurlow, M., et al., *Computer-Based Testing: Practices and Considerations. Synthesis Report 78.* National Center on Educational Outcomes, University of Minnesota, 2010.
4. Clariana, R. and P. Wallace, *Paper–based versus computer–based assessment: key factors associated with the test mode effect.* British Journal of Educational Technology, 2002. 33(5): p. 593-602.
5. Mandel, A., et al., *Cost analysis for computer supported multiple-choice paper examinations.* GMS Zeitschrift für Medizinische Ausbildung, 2011. 28(4).
6. Piaw, C. *Comparisons Between Computer-Based Testing and Paper-Pencil Testing: Testing Effect, Test Scores. Testing Time and Testing Motivation*. in *Proceedings of the Informatics Conference at: University of Malaya*. 2011.
7. Chin, C.H., J.S. Donn, and R.F. Conry, *Effects of computer-based tests on the achievement, anxiety, and attitudes of grade 10 science students.* Educational and Psychological Measurement, 1991. 51(3): p. 735-745.
8. Kolagari, S., et al., *The effect of computer-based tests on nursing students' test anxiety: a quasi-experimental study.* Acta Informatica Medica, 2018. 26(2): p. 115.
9. Boevé, A.J., et al., *Introducing computer-based testing in high-stakes exams in higher education: Results of a field experiment.* PloS one, 2015. 10(12): p. e0143616.
10. Paul D. Ellis, *The Essential Guide to Effect Sizes : Statistical Power, Meta-Analysis, and the Interpretation of Research Results.* 1nd ed Cambridge University Press; 2010.
11. Balaban, M.A., et al., *An approach to teaching port management aided by a constructive modelling and simulation environment.* International Journal of Service and Computing Oriented Manufacturing, 2016. 2(2): p. 155-178.
12. Qureshi, J. and M. Rizwan, *A PROPOSAL OF ELECTRONIC EXAMINATION SYSTEM TO EVALUATE DESCRIPTIVE ANSWERS.* Science International, 2015. 27(3).
13. Yu, Y., *Construction of electronic examination and education platform for financial management.* International Journal of Emerging Technologies in Learning (iJET), 2016. 11(09): p. 14-19.
14. Gwaltney, C.J., A.L. Shields, and S. Shiffman, *Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: a meta-analytic review.* Value in health, 2008. 11(2): p. 322-333.
15. Palmblad, M. and B. Tiplady, *Electronic diaries and questionnaires: designing user interfaces that are easy for all patients to use.* Quality of

Life research, 2004. 13(7): p. 1199-1207.

16. Hufford, M.R. and S. Shiffman. *Correspondence between paper and electronic visual analog scales among adult asthmatics*. in *Drug Information Association Statistics Conference. Hilton Head, South Carolina*. 2002.
17. Kleinman, L., et al., *A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) questionnaire.* Medical care, 2001. 39(2): p. 181-189.
18. Washburn, S., J. Herman, and R. Stewart, *Evaluation of performance and perceptions of electronic vs. paper multiple-choice exams.* Advances in physiology education, 2017. 41(4): p. 548-555.
19. Jeong, H., *A comparative study of scores on computer-based tests and paper-based tests.* Behaviour & Information Technology, 2014. 33(4): p. 410-422.
20. Russell, M., A. Goldberg, and K. O'connor, Computer-based testing and validity: A look back into the future. Assessment in education: principles, policy & practice, 2003. 10(3): p. 279-293.

طب و تزکیه
فصلنامه علمی - پژوهشی
دارای رتبه علمی پژوهشی
زمستان ۱۳۹۸- دوره ۲۸ - شماره ۴

Received: 2020/01/09
Accepted: 2020/03/08

فصلنامه علمی - پژوهشی طب و تزکیه
زمستان ۱۳۹۸ - دوره ۲۸ - شماره ۴

Comparison the effect of two types of paper-based and computer-based assessment methods on the results and performance of MHLE language test candidates

Abtin Heidarzadeh [1],Samaneh Panjeh Ali Beik [2]*, Pegah Derakhshan [3],
Majid Navaee [4], Hasan Turani [5]

Abstract

Background and objective: The purpose of this study was to investigate the effect of two types of paper-based and computer-based assessment methods on learners' assessment results in the final score of learners.

Methods and Materials: Considering the language test of the Ministry of Health and Medical Education, we analyzed and evaluated the test with both types of assessment tools at the same time and with the same questions. For this purpose, and to reduce sampling error, select five periods of this test, both simultaneous on paper and computer-based, and use psychometric indices such as difficulty coefficient, clean coefficient and Richardson Coder reliability index, Calculate and inferential statistics using Mann-Whitney tests, analysis of variance and effect size index (2η) using SPSS software and at the significant level of 052 we analyze, compare the results of the volunteers.

Results: The results show that the psychometric indices of the exams are better than the paper-based type in different exam periods. Also, the mean scores of participants ranged from 2.2 to 0.54 in periods 3 and 5, respectively, between the computer-based and paper-based tests. Except for period 3 tests, the scores of other tests followed the same distribution in computer-based and paper-based types.

Conclusion: The results of the analysis show that with the same test questions, in computer-based tests, compared to its paper-based counterpart, they generally achieved better results, while differentiating the coefficient of type-based tests. On the computer, it's been better.

Keywords: Ministry of Health Language Test ; Computer Based ; Paper Based ; Validity ; Reliability

1. Associate Prof, Department of Community Medicine, Faculty of Medicine, Guilan University of Medical Sciences, Rasht, Iran.
2*. Corresponder Author,Ph.D, Applied Mathematics, National Center of Medical Education Assessment, Ministry of Health, Tehran, Iran.
3. Medical Doctor, National Center of Medical Education Assessment, Ministry of Health, Tehran, Iran.
4. MSc, Applied Statistics, National Center of Medical Education Assessment, Ministry of Health, Tehran, Iran
5. Applied in Computer BSc, National Center of Medical Education Assessment, Ministry of Health, Tehran, Iran.