مقایسه مدل‌های کلاسیک و خصیصه مکنون در ارزیابی آزمون‌های تخصصی ورود به دوره‌های کارورزی رشته پزشکی

هادی عباسی*، محمدرضا فلسفی‌نژاد، علی دلاور، نورعلی فرخی، محمدعلی محققی

چکیده

مقدمه: علی‌رغم استفاده گسترده از آزمون پیش‌کارورزی و اهمیت آن در ارزیابی دانشجویان پزشکی، مطالعه‌ای که یک تحلیل روان‌سنجی نظام‌مند از سؤالات این آزمون فراهم نماید وجود نداشت. لذا پژوهش حاضر به ارزیابی آزمون پیش‌کارورزی اسفند ۱۳۸۹ با استفاده از مدل‌های کلاسیک و خصیصه مکنون و مقایسه آن‌ها پرداخته است.

روش‌ها: در این پژوهش توصیفی- مقطعی، تجزیه و تحلیل سؤالات آزمون بر پایه مدل کلاسیک، از طریق محاسبه ضرایب دشواری و تمیز، برآورد اعتبار آزمون به‌روش ضریب آلفای کرونباخ و تحلیل گزینه‌های انحرافی با استفاده از آزمون مقایسه نسبت انتخاب گزینه‌ها از طریق نرم‌افزار EXCEL انجام شد. بررسی بعدیت آزمون با نرم افزار NOHARM4 انجام شد. جهت برآورد پارامترهای سؤال (دشواری، تشخیص، حدس‌پذیری، تابع آگاهی سؤال و آزمون و خطای معیار اندازه‌گیری) از روش‌ها و مدل‌های خصیصه مکنون و از نرم‌افزار BILOG-MG3 استفاده شد.

نتایج: طبق نظریه کلاسیک، ۳۰ سؤال (۱۵ درصد) در بازه قابل قبول ضرایب دشواری و تمیز قرار داشتند. ضریب اعتبار آزمون ۰/۹۱۳ به دست آمد. ۷۳ سؤال (۳۶/۵٪) از لحاظ گزینه‌های انحرافی دارای ایراد بودند. بررسی بعدیت آزمون نشان داد که آزمون تک‌بعدی است. با استفاده از مدل لوجستیک سه‌پارامتری میانگین و انحراف معیار پارامتر دشواری (۰/۳۲۱ ، ۱/۸۷۴)، پارامتر تشخیص (۱/۰۲۱، ۰/۶۶۶) و پارامتر حدس (۰/۲۰۹ ، ۰/۰۸۲) به‌دست آمد. بیشترین میزان آگاهی آزمون در سطوح توانایی ۱/۰ و ۲/۸ قرار داشت و آزمون در نقاط برش از آگاهی کمتری نسبت به سطوح بالاتر توانایی برخوردار بود.

نتیجه‌گیری: یافته‌های پژوهش نشان داد که تحلیل‌های حاصل از مدل‌های خصیصه مکنون می‌توانند در جهت رفع محدودیت‌های نظریه کلاسیک آزمون مورد استفاده قرار گیرند.

واژه‌های کلیدی: آزمون پیش‌کارورزی، مدل کلاسیک، مدل‌های خصیصه مکنون

مجله ایرانی آموزش در علوم پزشکی/ خرداد ۱۳۹۲؛ ۱۳(۳): ۱۶۷ تا ۱۷۸

مقدمه

سنجش یک بخش مهم فرایند آموزشی است. نتایج آزمون حاوی اطلاعات ارزشمندی است که می‌تواند در مراحل مختلف دوره آموزشی مورد استفاده قرار گیرد. تفسیر نمرات آزمون به‌منظور: ۱) بهبود راهبردهای یادگیری، ۲) اصلاح شیوه‌های آموزشی، ۳) بهبود سؤالات آزمون و

*نویسنده مسؤول: هادی عباسی، دانشجوی دکتری سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی، تهران، ایران.
abbasihadi@yahoo.com
دکتر محمدرضا فلسفی‌نژاد (استادیار)، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی، تهران، ایران. (falsafinejad@yahoo.co.uk)؛ دکتر علی دلاور (استاد)، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی، تهران، ایران. (delavar@atu.ac.ir)؛ دکتر نورعلی فرخی (استادیار)، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبایی، تهران، ایران. (farrokhinoorali@yahoo.com)؛ دکتر محمدعلی محققی (استاد)، دانشکده پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران.
(mamohagheghi@yahoo.com)

تاریخ دریافت مقاله: ۹۱/۱۱/۱، تاریخ اصلاحیه: ۹۱/۲/۱، تاریخ پذیرش: ۹۲/۲/۲

شیوه‌های آزمودن و ٤) اتخاذ تصمیمات آگاهانه، بازخوردهای سازنده‌ای در اختیار فراگیران، معلمان، آزمون‌سازان و سیاست‌گذاران قرار می‌دهد. برای رسیدن به این هدف‌ها، اطلاعات مربوط به نمرات آزمون باید به طور مناسب تحلیل و به صورت معتبر تفسیر شود(١).

مرکز سنجش وزارت بهداشت، درمان و آموزش پزشکی جهت تحلیل آزمون‌های مختلف از جمله آزمون پیش‌کارورزی علاوه بر بررسی محتوایی سؤالات، از روش‌های نظریه کلاسیک آزمون (Classical Test Theory) استفاده می‌نماید. شواهد زیادی در رابطه با محدودیت‌آزمون‌هایی که براساس نظریه کلاسیک اندازه‌گیری ساخته‌شده‌اند، به‌دست آمده است. یکی از مهم‌ترین محدودیت‌های نظریه کلاسیک، "وابستگی پارامترهای سؤال به گروه نمونه (Group dependency)" و "وابستگی پارامترهای توانایی به آزمون(Test dependency)" است(٢تا٤). به‌عنوان مثال، دشواری یک سؤال به‌عنوان "نسبتی از آزمودنی‌ها در گروه موردنظر که سؤال را به‌درستی پاسخ داده‌اند" به توانایی آزمودنی‌های مورد اندازه‌گیری و توانایی آزمودنی‌ها، به آسانی یا دشواری سؤالات بستگی دارد. این وابستگی امکان استفاده از مدل کلاسیک را با محدودیت مواجه می‌سازد. همچنین در نظریه کلاسیک، خطای معیار اندازه‌گیری برای تمام آزمودنی‌ها با توانایی‌های مختلف، یکسان است. مسلماً دقت اندازه‌گیری هر سؤال در سطوح مختلف توانایی متفاوت است، که این مسأله در نظریه کلاسیک به‌هیچ وجه قابل مشاهده نیست. محدودیت دیگر نظریه کلاسیک این است که بیشتر آزمون مدار (test oriented) است تا سؤال مدار (item oriented)؛ در نتیجه تعیین سهم سؤال در اعتبار (reliability) و بالطبع خطای معیار اندازه‌گیری آزمون امکان‌پذیر نیست(٥).

نقاط ضعف نظریه کلاسیک می‌تواند تفسیر عملکرد آزمون را تحت تاثیر قرار دهد. لذا بهینه‌سازی فرایند گزینش و افزایش کارایی دانش‌آموختگان مستلزم استفاده از مدل‌ها و تکنیک‌هایی در نمره‌گذاری آزمون‌هاست که بتواند به صورت دقیق نقاط قوت و ضعف آن‌ها را نشان دهد. چنین مدل‌ها و روش‌های نمره‌گذاری را می‌توان در مدل‌ها و نظریه‌های جدید اندازه‌گیری که به نظریات خصیصه مکنون (latent trait) یا نظریه پرسش- پاسخ (Item Response Theory) معروف‌اند جستجو کرد. نظریه پرسش- پاسخ (IRT) به کاربرد نظریه یا الگوی صفت مکنون گفته می‌شود. در این نظریه برآوردهای پارامتر مربوط به شخص (توانایی/ خصیصه) و سؤالات، وابسته به گروه نمونه و آزمون نیستند. ویژگی نامتغیر بودن (invariance) نظریه پرسش- پاسخ، این مشکل را مرتفع می‌سازد. هم‌چنین نظریه پرسش-پاسخ، برآوردهای بهتری از توانایی آزمودنی به‌دست می‌دهد و اطلاعات بیشتری درباره آزمودنی و سؤالات فراهم می‌آورد؛ تابع آگاهی (information function) سؤال و آزمون برای سطوح مختلف توانایی مقادیر متفاوتی به‌دست می‌دهد که بیانگر دقت متفاوت سؤال در سطوح مختلف ویژگی مورد اندازه‌گیری است(٥). به همین جهت روش‌های نظریه پرسش- پاسخ، در تحلیل آزمون‌های سرنوشت‌ساز (high-stakes) مانند پیش‌کارورزی ترجیح داده می شود.

علی‌رغم استفاده گسترده از آزمون پیش‌کارورزی و اهمیت فوق‌العاده آن، پژوهشی که با استفاده از مدل‌های نظریه پرسش- پاسخ به تحلیل سؤالات این آزمون اقدام نماید وجود نداشت. تحقیقات انجام شده در خصوص آزمون‌های وزارت بهداشت (در سطح داخلی) را می‌توان به ٥ دسته کلی تقسیم کرد: الف) مطالعاتی که عوامل مؤثر بر نتایج حاصل از آزمون‌ها را بدون انجام هرگونه تحلیل روی سؤالات آزمون مورد بررسی قرار داده‌اند(٦و٧)؛ ب)مطالعاتی که با استفاده از نتایج یک آزمون به دنبال پیش‌بینی نتایج آزمون دیگر بوده‌اند (روایی پیش‌بینی آزمون‌ها) و یا میزان توافق نتایج دو آزمون را بررسی کرده‌اند(٨و٩)؛ ج) مطالعاتی که به بررسی محتوا و

ساختار سؤالات پرداخته‌اند و نه ویژگی‌های روان‌سنجی آن‌ها(۱۰و۱۱)؛ د)مطالعاتی که به بررسی ویژگی‌های روان‌سنجی سؤالات با استفاده از نظریه کلاسیک آزمون پرداخته‌اند و در مواردی ساختار سؤالات را نیز بررسی نموده‌اند(۱۲و۱۳)؛ هـ)مطالعاتی که به بررسی مشخصه‌های روان‌سنجی سؤالات با استفاده از نظریه پرسش-پاسخ پرداخته‌اند. در این زمینه صرفا دو پژوهش(۱٤و۱۵) در مورد آزمون ارتقا و پیش‌کارورزی بورد رشته زنان و زایمان کشور در سال ۱۳۷٤ و درخصوص آزمون دستیاری پزشکی سال ۱۳۸۹ انجام شده است. همان‌طور که ملاحظه می‌شود کلیه تحقیقات انجام شده بر روی آزمون‌های وزارت بهداشت در داخل کشور (به‌استثناء دو مورد) با استفاده از نظریه کلاسیک آزمون به تحلیل سؤالات پرداخته‌اند؛ این در حالی است که مدل‌هایIRT به طور فزاینده‌ای در حوزه‌های سنجش ویژگی‌های روانی و آموزشی(۱٦)، آمار زیستی(۱۷)، سنجش کیفیت زندگی(۱۸) و پژوهش‌های حوزه خدمات سلامت(۱۹و۲۰) مورد استفاده قرار گرفته‌اند.

بررسی عوامل مرتبط با آزمون‌های جامع ورود به دوره انترنی به منظور برنامه‌ریزی جهت حذف یا تعدیل عواملی که موفقیت دانشجویان را کاهش می‌دهد و همچنین تسلط دانشجویان را تضمین می‌نماید، بسیار با اهمیت و ضروری می‌باشد. از جمله این عوامل تاثیرگذار، ویژگی‌های روان‌سنجی آزمون است که از طریق بررسی و تحلیل سؤالات تعیین می‌شود. سوانسون و کیس(۲۱) اشاره می‌کنند که وقتی تصمیمات مهمی براساس نمرات آزمون اتخاذ می‌شوند (مسایل سرنوشت‌سازی که روی زندگی دانشجویان تاثیر می‌گذارند) ویژگی‌های روان‌سنجی نمرات آزمون اهمیت می‌یابند به نحوی که باید به طوریکپارچه مورد بازبینی و بهبود قرارگیرند. لذا می‌بایست به‌منظور اطمینان از منصفانه (عادلانه) بودن آزمون(test fairness)، سؤالات آن مورد تحلیل کمی قرار گیرند تا با استفاده از نتایج حاصل به تجدید نظر در آزمون و بهبود کیفیت سؤالات اقدام شود.

با توجه به موارد فوق‌الذکر، مسأله مهمی که مطرح می‌شود این است که آیا سؤالات آزمون پیش‌کارورزی با بهره‌گیری از مدل‌ها و روش‌های اندازه‌گیری به‌ویژه مدل‌های نظریه پرسش- پاسخ از کفایت و مطلوبیت لازم برخوردارند؟ به عبارت دیگر، هدف کلی این پژوهش "ارزیابی و مقایسه مشخصه‌های روان‌سنجی سؤالات آزمون پیش‌کارورزی با استفاده از مدل‌های کلاسیک و خصیصه مکنون" در جهت بهبود و ارتقای کیفی آزمون است. نتایج حاصل از این پژوهش با استفاده از نظریه پرسش- پاسخ؛ به ارتقای کیفیت سؤالات آزمون در سطح ملی، کاهش دغدغه دانشجویان در خصوص کفایت و منصفانه بودن سؤالات آزمون و اتخاذ تصمیمات مناسب‌تر و دقیق‌تر درخصوص عملکرد آزمون‌شوندگان کمک می‌کند.

روش‌ها

در این پژوهش توصیفی- مقطعی، از روش کمی جهت پاسخگویی به سؤالات تحقیق استفاده شد. ماهیت کمی پاسخ‌ها در آزمون‌های چندگزینه‌ای، استفاده از روش‌های ارزیابی کمی سؤالات آزمون در چارچوب نظریه کلاسیک اندازه‌گیری و نظریه پرسش- پاسخ و بهره‌گیری از روش‌های پیشرفته آماری را میسر می‌سازد.

ابزار ارزیابی در این تحقیق، آزمون پیش‌کارورزی است. این آزمون در پایان دوره کارآموزی برگزار می‌شود و قبولی در آن لازمه ورود به دوره کارورزی (انترنی) می‌باشد. آزمون مذکور شامل ۲۰۰ سؤال چهارگزینه‌ای است و موضوعات داخلی، جراحی عمومی، کودکان، زنان و زایمان، مغز و اعصاب، عفونی، رادیولوژی، آسیب‌شناسی، روان‌پزشکی، پوست، ارتوپدی، ارولوژی، چشم‌پزشکی، گوش و حلق و بینی، آمار و اپیدمیولوژی، فارماکولوژی و اخلاق پزشکی را در بر می‌گیرد. موضوعات داخلی، کودکان، جراحی عمومی و زنان و

زایمان به‌ترتیب با ٤٤ (٢٢٪)، ٢٥ (١٢/٥٪)، ٢٣ (١١/٥٪) و ١٨ (٩٪) سؤال، بیشترین تعداد سؤال و اخلاق پزشکی با ٣ (١/٥٪) سؤال، کمترین تعداد سؤال را در این آزمون به خود اختصاص داده‌اند. حد نصاب قبولی در آزمون پیش‌کارورزی، قبل از سال ١٣٩٠، ٧٠ درصد میانگین نمرات ٥ درصد افرادی که بالاترین نمره را در سطح کشور کسب نموده‌اند تعیین شده بود؛ لذا با این آزمون به مثابه یک آزمون هنجار- مرجع رفتار می‌شد؛ در حالی‌که آزمونی ملاک- مرجع است که افراد می‌بایست به حد معینی از چیرگی در حوزه مورد نظر دست یابند. حدنصاب قبولی در این آزمون، از سال ١٣٩٠، ٥٠ درصد نمره کل آزمون تعیین شده است؛ ملاکی مطلق که فارغ از میزان صحت و دقت آن، عملکرد آزمون‌شوندگان با توجه به آن مورد قضاوت قرار می‌گیرد.

جامعه آماری این تحقیق را کلیه شرکت‌کنندگان در آزمون پیش‌کارورزی اسفند ١٣٨٩ (٢١٩٥ نفر) تشکیل می‌دهند. با توجه به دردسترس بودن اطلاعات تمامی داوطلبان و امکان تجزیه و تحلیل پاسخ‌های کلیه افراد، نمونه پژوهش برابر کل جامعه تحقیق بود.

تجزیه و تحلیل سؤالات آزمون بر پایه مدل کلاسیک اندازه‌گیری، از طریق محاسبه ضرایب دشواری و تمیز (همبستگی دورشته‌ای نقطه‌ای)، برآورد اعتبار آزمون به‌روش ضریب آلفای کرونباخ و تحلیل گزینه‌های انحرافی با استفاده از آزمون مقایسه نسبت انتخاب گزینه‌ها در گروه‌های قوی(بالا) و ضعیف (پایین) از طریق نرم‌افزارهای SPSS-11 و EXCEL انجام شده است. همچنین برای تعیین بعدیت آزمون نرم‌افزار NOHARM4 (open-source) به‌کار رفته است. برنامه کامپیوتری NOHARM پارامترهای c (حدس) را برآورد نمی‌کند و با مقادیر وارد شده به‌عنوان ثابت (نامتغیر) رفتار می‌کند(٥). در این تحلیل با توجه به این که سؤالات نمره منفی نداشتند پیش‌بینی می‌شد که آزمون‌شوندگان از حدس و گمان در پاسخ به سؤالات استفاده نمایند؛ در نتیجه مقدار حدس پذیری سؤالات ٠/٢٥ لحاظ شده است. علاوه بر این، با مقادیر ٠/٢٠، ٠/١٥، ٠/١٠، ٠/٠٥ و صفر (به‌عنوان ارزش‌های c) نیز بعدیت آزمون مورد بررسی قرار گرفته است.

جهت برآورد پارامترهای سؤال (دشواری، تشخیص، حدس‌پذیری، تابع آگاهی سؤال و آزمون، خطای معیار اندازه‌گیری) از روش‌ها و مدل‌های نظریه پرسش-پاسخ و از نرم‌افزار BILOG-MG3 استفاده شده است. در این نرم‌افزار، با توجه به نامعلوم بودن پارامترهای سؤال و توانایی، از روش بیشینه درست‌نمایی حاشیه‌ای (Marginal Maximum Likelihood Estimation) جهت برآورد همزمان پارامترها استفاده می‌شود.

نتایج

یافته‌های پژوهش در سه بخش، شامل الف) تجزیه و تحلیل توصیفی، ب) تحلیل سؤالات آزمون براساس مدل کلاسیک و ج) تحلیل سؤالات آزمون با استفاده از مدل‌های پرسش- پاسخ ارائه شده است.

الف) تجزیه و تحلیل توصیفی

در آزمون پیش‌کارورزی اسفند ١٣٨٩، ٢١٩٥ نفر شامل٧٣٢ دانشجوی مرد (٣٣ درصد) و ١٤٦٣ دانشجوی زن (٦٧ درصد) شرکت نمودند. میانگین و انحراف معیار نمرات آزمون ٢١/١٤±١١٢/٢٦ به‌دست آمد. مقداربالای آماره کولموگروف- اسمیرنوف (٠/٦٩٤) نیز نشان داد که داده‌ها دارای توزیع نرمال است ($p<0/05$).

ب) تحلیل سؤالات آزمون براساس مدل کلاسیک

ضریب اعتبار کل آزمون با استفاده از ضریب آلفای کرونباخ ٠/٩١٣ به‌دست آمد. همچنین، بعد از حذف هر سؤال میزان اعتبار آزمون مجددا مورد بررسی قرار گرفت. نتایج نشان داد که ضریب همبستگی سؤالات ٤٧، ٦٩، ٩٩ و ١٤٨ با کل آزمون منفی بوده و با سایر سؤالات

هم‌خوانی ندارند و با حذف آن‌ها ضریب اعتبار افزایش پیدا می‌کند. خطای معیار اندازه‌گیری آزمون نیز ٦/٢٤ به‌دست آمد. براساس نظریه کلاسیک آزمون، ٧٣ سؤال (٣٦/٥ درصد) دارای دشواری متوسط (٠/٧-٠/٣) و ضریب تمیز ٠/٢ و بالاتر و از این میان ٣٠ سؤال (١٥ درصد) دارای ضریب تمیز بالاتر از ٠/٣ می‌باشند و در اطراف نمره برش آزمون، عملکرد قابل قبولی دارند؛ سایر سؤالات دارای شرایط فوق‌الذکر نمی‌باشند. در جدول١، توزیع فراوانی و درصد سؤالات آزمون برحسب ضرایب دشواری و تمیز حاصل از به‌کارگیری مدل کلاسیک اندازه‌گیری ارائه شده است.

جدول١: توزیع سؤالات آزمون پیش‌کارورزی براساس ضرایب دشواری و تمیز در مدل کلاسیک

شاخص		ضریب دشواری		
		کمتر از ٠/٣	٠/٣-٠/٧	بیشتر از ٠/٧
ضریب تمیز	کم‌تر از ٠/٢	١٧(٪٨/٥)	٥١(٪٢٥/٥)	٧(٪٣/٥)
	٠/٢-٠/٣	٤(٪٢)	٤٣(٪٢١/٥)	٣١(٪١٥/٥)
	بیش از ٠/٣	٢(٪١)	٣٠(٪١٥)	١٥(٪٧/٥)

برای تجزیه و تحلیل گزینه‌های انحرافی سؤالات، از آزمون مقایسه نسبت انتخاب گزینه‌ها در گروه‌های قوی(٢٧٪ آزمون‌شوندگان دارای نمره‌های بالا) و ضعیف (٢٧٪ آزمون‌شوندگان دارای نمره‌های پایین) استفاده گردید. شناسایی اشکالات موجود در گزینه‌های سؤالات با توجه به دو ملاک زیر انجام گرفت: ١) نسبت انتخاب گزینه درست در دو گروه قوی و ضعیف معنادار نبوده و یا در صورت معناداری در گروه ضعیف بیشتر از گروه قوی است، ٢) نسبت انتخاب گزینه انحرافی در دو گروه معنادار نبوده و یا در صورت معناداری در گروه قوی بیشتر از گروه ضعیف است(٢٢). تحلیل گزینه‌های سؤالات آزمون نشان داد که از مجموع ٢٠٠ سؤال، ٧٣ سؤال (٣٦/٥ درصد از کل سؤالات) دارای اشکال هستند. بر این اساس، ٤ سؤال دارای ضریب تمیز منفی می‌باشند که از میان آن‌ها ٣ سؤال در سه گزینه و یک سؤال در دو گزینه دارای ایراد هستند. ضرایب تمیز ١٤ سؤال در سطح ٠/٠١ معنادار نمی‌باشند که از میان آن‌ها ٦ سؤال در یک گزینه، ٦ سؤال در دو گزینه و ٢ سؤال در سه گزینه دارای اشکال هستند. ٥٥ سؤال دیگر دارای ضریب تمیز مثبت می‌باشند که از میان آن‌ها ٤٩ سؤال در یک گزینه و ٦ سؤال در دو گزینه دارای ایراد هستند.

ج) تحلیل سؤالات آزمون با استفاده از مدل‌های پرسش- پاسخ

یک مفروضه اصلی که در مدل‌هایIRT (پرسش- پاسخ) به طور گسترده‌ای مورد استفاده قرار گرفته تک‌بعدی بودن آزمون است و به این معنا است که فقط یک توانایی به‌وسیله مجموعه سؤالات آزمون اندازه‌گیری می‌شود(٥)، لذا قبل از برآورد پارامترهای سؤال، این مفروضه باید بررسی شود. نتایج حاصل از تحلیل عاملی سطح- سؤال به‌وسیله نرم افزار NOHARM نشان داد (جدول شماره ٢) که مقدار Root Mean Square Residual در حالت تک‌بعدی بسیار کوچک و به صفر نزدیک است(٢٣تا٢٥)؛ در عین حال این مقدار از ٤ برابر معکوس ریشه دوم حجم نمونه (که در این تحقیق برابر ٠/٠٨٥ است) کوچک‌تر می‌باشد(٢٦) و از حالت تک بعدی به دو بعدی کاهش ناچیزی را نشان می‌دهد. با توجه به قاعده تیت (Tate) (٢٧)، درصد کاهش RMSR از حالت تک بعدی به دوبعدی (برای حدس‌پذیری ٠/٢٥) کمتر از ١٠ درصد می‌باشد؛ در نتیجه مدل تک‌بعدی برازش بهتری با داده‌ها دارد. همچنین، شاخص تاناکا Tanaka index به مقدار ٠/٩٢ به‌دست آمد که بیانگر برازش قابل قبول مدل با داده‌هاست و در حالت دوبعدی در مقایسه با حالت تک‌بعدی افزایش ناچیز تقریبا ٠/٧ درصدی را نشان می‌دهد(٢٨). بنابراین، با توجه به دلایل مذکور می‌توان گفت که آزمون تک‌بعدی است و می‌توان از مدل‌های تک‌بعدیIRT برای مقیاس‌پردازی آن استفاده کرد. لازم

به ذکر است که با مقادیر ۰/۲۰، ۰/۱۵، ۰/۱۰، ۰/۰۵ و صفر (به‌عنوان ارزش‌هایc) نیز بعدیت آزمون بررسی شد که نتایج حاصل تفاوت ناچیزی با یکدیگر داشتند و بیانگر تک‌بعدی بودن آزمون بودند.

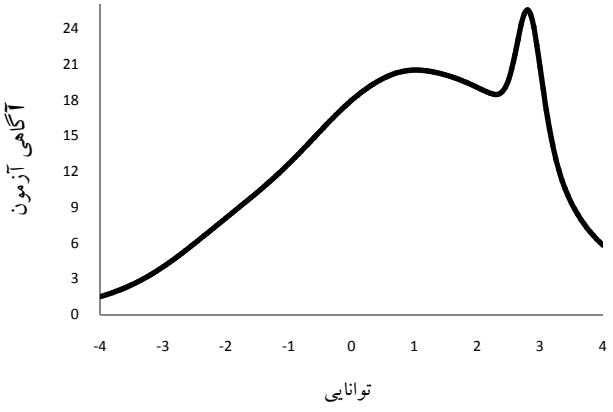
جدول ۲: بررسی بعدیت آزمون پیش‌کارورزی با استفاده از راه حل اکتشافیNOHARM

		تک بعدی		دو بعدی	
شاخص	حدس پذیری	RMSR	Tanaka index	RMSR	Tanaka index
مقدار	۰/۲۵	۰/۰۰۵۷۲۱٤	۰/۹۲۳۷۰۱۷	۰/۰۰٥٤٦٧٣	۰/۹۳۰۳۲۹۷

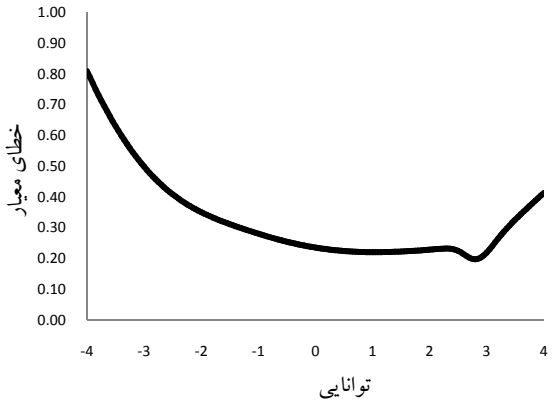
درصد کاهش RMSR حاصل از بکارگیری مدل دو بعدی با حدس پذیری (۰/۲۵): ٤/٤٤ (۱۰٪ >)

به منظور برآورد پارامترهای سؤال و توانایی، با توجه به دلایل فنی و آماری زیر از مدل لجستیک سه‌پارامتری (Three-Parameter Logistic Model) استفاده شد: ۱)عدم لحاظ نمره منفی و استفاده از عامل حدس به‌عنوان یک راهبرد برای پاسخ‌گویی به سؤالات و ۲)آماره خی‌دو (chi-square) که جهت بررسی برازش مدل با داده‌ها مورد استفاده قرارگرفت، نشان داد که مدل سه‌پارامتری نسبت به مدل‌های یک و دوپارامتری برازش بهتری با داده‌های آزمون دارد. پس از برآورد پارامترهای سؤال میزان برازندگی تک‌تک سؤالات با مدل سه‌پارامتری محاسبه گردید. نتایج نشان داد که ۹ سؤال (۲٦، ٤۷، ٦۹، ۸۰، ۸٤، ۹۳، ۹۹، ۱٤۸ و ۱۷۲) با این مدل برازش ندارند. میانگین و انحراف معیار پارامترهای برآورد شده سؤالات آزمون نیز به‌ترتیب عبارتند از: پارامتر دشواری (۰/۳۲۱، ۱/۸۷٤)، پارامتر تشخیص (۱/۰۲۱، ۰/٦٦٦) و پارامتر حدس (۰/۲۰۹ ، ۰/۰۸۲). لازم به ذکر است که با توجه به عدم برازش ۹ سؤال با مدل سه‌پارامتری، یافته‌های بخش (ج) با احتساب ۱۹۱ سؤال و حداکثر نمره ۱۹۱ به‌دست آمده است.

در نظریه پرسش–پاسخ، پس از برآورد پارامترهای سؤال، مقادیر آگاهی برای تک‌تک سؤالات و برای کل آزمون محاسبه می‌شود. نمودارهای (۱) و (۲) تابع آگاهی و خطای معیار اندازه‌گیری آزمون پیش‌کارورزی را نشان می‌دهند. در جدول۳ نیز مقادیر آگاهی و خطای معیار اندازه‌گیری آزمون در سطوح مختلف توانایی نشان داده شده است. مشاهده و بررسی جدول مذکور و منحنی آگاهی آزمون نشان می‌دهد که بیشترین میزان آگاهی‌بخشی این آزمون در سطح توانایی ۲/۸ (معادل نمره ۱٦٦) و سپس در توانایی ۱/۰ (معادل نمره ۱۳۲) و در حالت کلی‌تر در فاصله ۱/۰ و ۲/۸ قرار دارد. بیشتر بودن مقدار آگاهی در این بازه به این معنا است که می‌توان توانایی افرادی را که در این سطوح قرار دارند در مقایسه با سایر افراد با دقت بیشتری برآورد کرد. به‌عبارت دیگر، توانایی افراد قوی‌تر در مقایسه با افراد ضعیف‌تر با دقت بیشتر و خطای کمتری اندازه‌گیری می‌شود.



نمودار ۱: منحنی آگاهی آزمون پیش کارورزی



نمودار ۲: منحنی خطای معیار آزمون پیش کارورزی

جهت تعیین نمره برش (ملاک قبولی) روی مقیاس توانایی، نمرات برش (۷۰درصد از میانگین نمرات ۵درصد بالای آزمون‌شوندگان و ۵۰درصد نمره کل آزمون) روی مقیاس نمره حیطه‌ای (domain score) تنظیم شدند (به‌ترتیب ٥٦ و ٥٠ برحسب درصد)؛ بدین شکل که تمامی مقادیر توانایی، به نمره حیطه‌ای تبدیل شدند و سپس مقادیر توانایی مربوط به نمرات حیطه‌ای ٥٦ و ٥٠ به‌عنوان نمره برش روی مقیاس توانایی شناسایی و میزان آگاهی و خطای معیار اندازه‌گیری آن‌ها تعیین گردید. بدین ترتیب افراد دارای توانایی ۰/۰۷- و بالاتر (براساس ملاک ٥٦درصد) و افراد دارای توانایی ۰/٥٩- و بیشتر (طبق ملاک ٥٠درصد) پذیرفته و سایر افراد ردمی‌شوند. در جدول ٤ نتایج این بررسی دیده می‌شود.

مقایسه جداول۳ و ٤ نشان می‌دهد که آگاهی آزمون در نمرات توانایی ۱/۰ (معادل نمره ۱۳۲) و بالاتر، بیشتر از آگاهی آزمون در اطراف نمرات برش (دامنه ۹۵/۵ تا ۱۰۷) است؛ در حالی که انتظار می‌رود به‌منظور جداسازی افراد دارای حداقل صلاحیت از سایر افراد، آگاهی آزمون در اطراف نمرات برش به مقدار زیادی قله‌ای بوده و از بیشترین میزان دقت در اندازه‌گیری توانایی افراد برخوردار باشد.

جدول ۳: مقادیر آگاهی و خطای معیار اندازه‌گیری برآورد شده در سطوح مختلف توانایی

θ	آگاهی	خطای معیار اندازه‌گیری
-۳	٤/۰۱۸٦	۰/٤۹۸۸
-۲	۸/۱۱٥۰	۰/۳٥۱۰
-۱	۱۲/٦٥۹۱	۰/۲۸۱۱
۰	۱۷/۹۸۲۹	۰/۲۳٥۸
۱	۲۰/٥۲۳۹	۰/۲۲۰۷
۲	۱۹/۰۷۲٥	۰/۲۲۹۰
۳	۲۰/۷۷۳٥	۰/۲۱۹٤

جهت تجزیه و تحلیل گزینه‌های انحرافی سؤالات، گروه‌های بالا (قوی) و پایین (ضعیف) براساس توانایی برآورد شده آزمون‌شوندگان (طبق نظریه پرسش- پاسخ) مشخص شدند. با توجه به ضریب همبستگی بالای بین نمره خام و توانایی (۰/۹۸) انتظار می‌رود که جایگاه افراد در دو گروه براساس این دو کمیت تغییر اندکی داشته باشد. بر این اساس به‌ترتیب ۹۳ و ۹٤ درصد آزمون‌شوندگان در دو گروه بالا و پایین (با توجه به نمره خام و توانایی) یکسان می‌باشند؛ لذا نتایج حاصل از تحلیل گزینه‌های انحرافی سؤالات برمبنای این دو نمره، تفاوتی با یکدیگر ندارند.

جدول ۴: مقادیر آگاهی و خطای معیار اندازه‌گیری در نمرات برش برحسب پارامتر توانایی

نمره برش	نمره حیطه‌ای	θ (توانایی)	آگاهی	خطای معیار اندازه‌گیری
۷۰ درصد میانگین نمرات ۵ درصد بالا	٥٦	-۰/۰٦۷۹	۱۷/٦٦٥۹	۰/۲۳۷۹
۵۰ درصد نمره کل آزمون	۵۰	-۰/۵۸۷۷۱	۱٤/۸۹۹۵	۰/۲۵۹۱

بحث

علی‌رغم استفاده گسترده از آزمون‌های عینی در فرایند آموزش و ارزیابی دانشجویان پزشکی، مطالعه‌ای که یک تحلیل روان‌سنجی نظام‌مند از عملکرد دانشجویان به‌ویژه در آزمون پیش‌کارورزی فراهم نماید وجود نداشت. لذا هدف این پژوهش ارزیابی سؤالات آزمون پیش‌کارورزی اسفند ۱۳۸۹ با استفاده از تکنیک‌های متداول در مدل‌های کلاسیک و خصیصه مکنون (پرسش- پاسخ) بوده است.

برطبق مدل کلاسیک اندازه‌گیری، ضریب اعتبار برآورد شده آزمون که با استفاده از فرمول آلفای کرونباخ به‌دست آمد حاکی از همسانی درونی و اعتبار مطلوب آزمون است. این یافته با مطالعه صبوری و همکاران هم-خوانی زیادی دارد(۱۳). بررسی شاخص‌های کلاسیک آزمون نشان می‌دهد که ۷۳ سؤال (۳٦/۵ درصد) دارای دشواری متوسط (۰/۷-۰/۳) و قدرت تشخیص ۰/۲ و بالاتر و از این میان ۳۰ سؤال (۱۵ درصد) دارای قدرت تشخیص بالاتر از ۰/۳ می‌باشند و در اطراف نمره برش آزمون، عملکرد قابل قبولی دارند که بیانگر وضعیت مطلوب‌تر نسبت به برخی مطالعات مشابه دارد(۱٤و۱۵). لذا می‌بایست تعداد سؤالات بیشتری با دشواری متوسط و با ضریب تمیز بالاتر از ۰/۳ به آزمون اضافه نمود و یا در جهت بهبود کیفیت سؤالات دارای دشواری متوسط و ضریب تمیز ۰/۳-۰/۲ اقدام کرد. تحلیل گزینه‌های انحرافی سؤالات نیز نشان می‌دهد که گزینه‌های ۷۳ سؤال (۳٦/۵ درصد) از کارایی لازم برخوردار نیستند. بنابراین لازم است جهت ارتقای کیفی آزمون، گزینه‌های انحرافی معنادارتری به آن اضافه شود تا به افزایش قدرت تشخیص آزمون نیز کمک کنند.

بر اساس نظریه پرسش- پاسخ و به‌منظور برآورد پارامترهای سؤال و توانایی، از مدل لجستیک سه‌پارامتری استفاده شد. این در حالی است که در تحقیقات مشابه از مدل‌های یک‌پارامتری(۱۵)، دوپارامتری(۱) و سه‌پارامتری(۱٤) با توجه به توجیهات منطقی و آماری استفاده شده است. نتایج کاربرد مدل سه‌پارامتری نشان داد که ۹ سؤال با این مدل برازش ندارند. انحراف معیار پارامتر دشواری بیانگر این است که سؤالات در طول پیوستار توانایی توزیع شده‌اند؛ در حالی‌که در آزمون‌های تعیین صلاحیت انتظار می‌رود که میزان دشواری سؤالات با نمره برش آزمون انطباق داشته باشند. متوسط مقادیر پارامتر حدس (۰/۲۰۹) حاکی از تاثیر عامل حدس در پاسخ‌گویی آزمون‌شوندگان به سؤالات آزمون است که با توجه به عدم اعمال نمره منفی در نمره‌گذاری آزمون قابل توجیه است. بررسی منحنی تابع آگاهی آزمون و مقادیر آگاهی برآورد شده در سطوح مختلف توانایی، نشان داد که آزمون برای افرادی که در محدوده توانایی ۱/۰ و ۲/۸ قرار دارند دارای آگاهی بیشتر و دقت اندازه‌گیری بالاتری است. با توجه به دو ملاک قبولی آزمون (۷۰ درصد میانگین نمرات ۵ درصد بالا و ۵۰ درصد نمره کل آزمون)، نمرات برش متناظر با آن‌ها در مقیاس توانایی تعیین شدند. در صورتی که فاصله توانایی صفر تا ۰/٦- به‌عنوان بازه‌ای که دو نمره برش مذکور در آن قرار دارند مشخص شوند؛ از ۲۳ سؤالی که در این محدوده قرار دارند ۱۸ سؤال (۹ درصد) دارای ضریب تشخیص ۰/٦۵ و بالاتر هستند که بیانگر کمبود تعداد سؤالات متناسب با این فاصله با قدرت تشخیص متوسط و بالاتر است؛

موضوعی که در مطالعه‌ای مشابه(۱٤) نیز به آن اشاره گردیده است. فقدان سؤالاتی که به طور مناسب با سطح خاصی از توانایی انطباق داشته باشند می‌تواند اندازه‌گیری در آن سطح را با مشکل مواجه سازد. در این خصوص، دو اقدام پیشنهاد می‌شود: ۱) افزودن تعدادی سؤال با قدرت تشخیص بالا که دشواری آن‌ها در دامنه مورد نظر قرار دارد و ۲) تغییر جایگاه نمره برش به سطوح بالاتر پیوستار توانایی، با توجه به این که آگاهی‌بخشی آزمون در آن محدوده بیشتر است. با توجه به حساسیت حرفه پزشکی در حفظ و حراست سلامت جامعه و دغدغه تسلط دانشجویان بر مباحث درسی و عدم لحاظ نمره منفی در آزمون، پیشنهاد دوم منطقی‌تر به‌نظر می‌رسد.

نظریه پرسش- پاسخ در هر سطح توانایی برآوردی از خطای معیار اندازه‌گیری فراهم می‌کند (جدول۳)؛ در حالی‌که در نظریه کلاسیک آزمون تنها یک برآورد خطا برای تمامی آزمودنی‌ها (و در تمامی سطوح توانایی) ارائه می‌گردد(۵). در نتیجه استفاده از مدل‌های پرسش- پاسخ به میزان زیادی کارایی و دقت اندازه‌گیری را بهبود می‌بخشد(۱٤) و می‌تواند در جهت رفع محدودیت‌های نظریه کلاسیک آزمون و تکمیل نتایج حاصل از آن مورد استفاده قرار گیرد(۲۹). بررسی ادبیات شیوه‌های سنجش در علوم پزشکی نیز حاکی از وجود یک نیاز اساسی برای امکان‌سنجی کاربردهای نظریه پرسش- پاسخ در سطوح مختلف سنجش در آموزش پزشکی است(۱).

این پژوهش بر روی آزمون پیش‌کارورزی و در یک مقطع زمانی خاص انجام شده است؛ لذا می‌بایست تعمیم‌پذیری نتایج حاصل با احتیاط صورت پذیرد. همچنین، پیشنهاد می‌گردد علاوه بر تحلیل‌های متداول کلاسیک، از روش‌ها و تکنیک‌های جدید اندازه‌گیری (IRT) در سایر آزمون‌های وزارت بهداشت، درمان و آموزش پزشکی استفاده شود.

نتیجه‌گیری

مقایسه مدل‌های کلاسیک و خصیصه مکنون (پرسش- پاسخ) نشان می‌دهد که تحلیل‌های حاصل از مدل‌های پرسش- پاسخ می‌تواند در جهت رفع محدودیت‌های نظریه کلاسیک آزمون مورد استفاده قرار گیرد. نامتغیر بودن پارامترهای سؤال و توانایی، توصیف رابطه میان عملکرد و توانایی از طریق خم ویژه سؤال، محاسبه آگاهی سؤال و آزمون و نمودارهای آن‌ها، برآورد خطای معیار پارامترهای سؤال و گزینش سؤالات مناسب برای ساخت آزمون با دقت مطلوب اندازه‌گیری در هر سطح از توانایی (در اطراف نقطه برش)، از جمله این موارد است(۵و۲۹و۳۰).

قدردانی

از مساعدت همکاران معاونت محترم آموزشی وزارت بهداشت، درمان و آموزش پزشکی و همکاری مرکز سنجش آموزش پزشکی تشکر و قدردانی می‌گردد.

منابع

1. Kokinova M. Psychometric analysis of students performance on basic medical science tests. [dissertation]. Kent: Kent State University; 2004.
2. Hambleton RK, Swaminathan H, Rogers HJ.Fundamentals of item response theory. Newbury Park, CA: Sage; 1991.
3. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates; 1980.
4. Wright BD, Stone MH. Best test design. Chicago: MESA; 1979.
5. Falsafinejad MR, (Translator). Fundamentals of Item Response Theory.Hambleton RK, Swaminathan H, Rogers HJ, (Author). Tehran: Allameh Tabatabaei University; 2010. [Persian]

6. Roudbari M, Sheibak A. The survey of the result and the associated factors in the preinternship examination at Zahedan University of Medical Sciences in September & Mars 2001.Zahedan Journal of Research in Medical Sciences. 2006; 8(1): 19-27.
7. Roudbari M, Movahhed S. [The related factors on the pre-internship scores and the exam result of Zahedan Medical students in 2002]. Strides in Development of Medical Education. 2005; 1(2): 94-103. [Persian]
8. Rezaei M, Roshani D, Iranfar SH, Shakibaei D. [Mizane tavafoghe azmoonhaye jamee oloome paye va pishkarvarzi dar rotbebandie daneshgahe oloom pezeshkie Kermanshah (1369-81)]. Journal of Kermanshah University of Medical Sciences. 2005; 9(3):48-54. [Persian]
9. Mohamadi M, Ahmadi J. [Predictive validity of the comprehensive basic science examination (CBSE) for success assessment of comprehensive preintership examination (CPIE) in medical students]. Iranian Journal of Medical Education.2002;2: 107. [Persian]
10. Shakurnia A, Khosravi Boroujeni A, Mozafari A, Elhampour H. [An Evaluation of Exam Questions Designed by Faculty Members, Emphasizing on the Multiple Choice Question Structure, Ahvaz Jundishapur University of Medical Sciences, 2007]. Strides in Development of Medical Education. 2010; 6(2): 129-138. [Persian]
11. Haghshenas MR, Vahidshahi K, Mahmoudi M, Shahbaznejad L, Parvinnejad N, Emadi A. [Evaluation of Multiple Choice Questions in the School of Medicine, Mazandaran University of Medical Sciences, the First Semester of 2007]. Strides in Development of Medical Education. 2010; 5(2): 119-126. [Persian]
12. Garakyaraghi M, Avizhgan M, Ebrahimi A, Esfandiari E, Esmaeili A, Shayan Sh,et al. [Assessment of qualitative and quantitative indexes of clerkship tests in general medicine]. Iranian Journal of Medical Education. 2011; 10(5): 533-543. [Persian]
13. Saburi M, Shayan Sh, Salehi A, Honarpisheh H. [Review Psychometric Parameters of the 29th Residency Test (1380) According to the Classic Test Theory (CTT)]. Iranian Journal of Medical Education.2002; 2: 49-50. [Persian]
14. Salehi A. [Barrasye moshakhasehaye ravansanjie azmoone ertegha va preborde reshteye zanan va zayemane keshvare Iran barasase nazariye soal pasokh (IRT)]. [dissertation]. Tehran: Allameh Tabatabaei University; 1998. [Persian]
15. Izanloo A. [Analytical study of medical residency exam by classical test theory and item response theory (2010-11)]. [dissertation]. Mashhad: Mashhad University of Medical Sciences; 2012. [Persian]
16. Scot SL, Edward HI. Empirical bayes and item-clustering efforts in a latent variable hierarchical model: a case study from the national assessment of educational progress.J Am Stat Assoc. 2002; 97(459): 409-419.
17. Bock RD. Gibbons,RD.High-dimentional multivariate probit analysis. Biometrics. 1996; 52(4): 1183-1194.
18. Douglas JA. Item response models for longitudinal quality of life data in clinical trials. Stat Med. 1999; 18(21): 2913-2917.
19. Gibbons RD, Wilcox-Gok V. Health service utilization and insyrance coverage: a multivariate probit analysis. J Am Stat Assoc. 1998; 93(441): 63-72.
20. Landrum MB, Bronskill SE, Normand SLT. Analytic methods for constructing cross-sectional profiles of health care providers. Health Services & Outcomes Research Methodology. 2000; 1(1): 23-47.
21. Swanson DB, Case SM. Assessment in basic science instruction: directions for practice and research. Advances in Health Sciences Education.1997; 2: 71-84.
22. Sharifi HP. [Osoule ravansanji va ravanazmayi]. Tehran: Roshd; 2002. [Persian]
23. Fraser C, McDonald RP.NOHARM: Least squares item factor analyses.Multivariate Behavioral Research. 1988; 23: 267-269.
24. Waller NG, Tellegen A, McDonald RP, Lykken DT. Exploring nonlinear models in personality assessment: Development and preliminary validation of a negative emotionality scale. Journal of Personality. 1996; 64(3): 545–576.
25. Browne MW, Cudeck R.Alternative ways of assessing model fit. In: KA Bollen & JS Long, editors. Testing structural equation models. Newbury Park, CA: Sage; 1993: 136–162.
26. McDonald RP, Mok MMC. Goodness of fit in item response models. Multivariate Behavioral Research. 1995; 30(1): 23-40.

27. Tate R.A comparison of selected empirical methods for assessing the structure of response to test items.Applied Psychological Measurement. 2003; 27(3): 159-203.
28. McDonald RP.Normal-ogive multidimensional model. In: WJ van der Linden & RK Hambleton, editors. Handbook of Modern Item Response Theory. New York: Springer; 1997.
29. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003; 37(9): 830-837.
30 Delavar A, (Translator).Introduction to measurement theory.Allen MJ, Yen WM, (Author). Tehran: Samt; 1995. [Persian]

The Comparison of Two Models for Evaluation of Pre-internship Comprehensive Test: Classical and Latent Trait

Hadi Abbasi[1], Mohammad Reza Falsafinejad[2], Ali Delavar[3], Noor Ali Farrokhi[4], Mohammad Ali Mohagheghi[5]

Abstract

Introduction: *Despite the widespread use of pre-internship comprehensive test and its importance in medical students' assessment, there is a paucity of the studies that can provide a systematic psychometric analysis of the items of this test. Thus, the present study sought to assess March 2011 pre-internship test using classical and latent trait models and compare their results.*
Methods: *In this cross-sectional descriptive research, item analysis based on classical model was conducted by calculation of difficulty and discrimination coefficients, estimation of reliability by Cronbach's alpha coefficient method and distracters analysis by comparing distracter proportions using EXCEL software. NOHARM4 software was also used to assess test dimensionality. To assess items parameters (difficulty, discrimination, guessing, information functions, and standard error of measurement) latent trait models and BILOG-MG3 software were used.*
Results: *According to the classical test theory, 30 items (15%) were in the acceptable range of difficulty and discrimination coefficients. Test reliability coefficient was 0.913. Seventy three items (36.5%) had problems with the distracters. The analysis of test dimensionality indicated that the test was unidimensional. Using Three-Parameter Logistic Model, we obtained the mean and standard deviation of items parameters, i.e. item difficulty (0.321, 1.874), item discrimination (1.021, 0.666), and item guessing (0.209, 0.082). Maximum test information function was between 1.0 to 2.8 ability levels, and the test had lower information compared to the higher levels of ability at cutting scores.*
Conclusion: *The findings indicate that analyses of latent trait models can be used to overcome the limitations of the classical test theory.*

Keywords: Pre-internship test, Classical model, Latent trait models

Addresses:
[1](✉) PhD Student in Assessment and Measurement, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University,Tehran, Iran. E-mail:abbasihadi@yahoo.com
[2]Assistant Professor, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran , Iran. E-mail: falsafinejad@yahoo.co.uk
[3]Professor, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran , Iran. E-mail: delavar@atu.ac.ir
[4]Assistant Professor, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran , Iran. E-mail: farrokhinoorali@yahoo.com
[5]Professor, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran. E-mail: mamohagheghi@yahoo.com