فصلنامه علمی تدریس پژوهی
Journal of Research in Teaching
سال هفتم– شمارۀ چهارم – زمستان ۱۳۹۸ صفحات: ۶۹-۴۹
Vol 7, No 4, Winter 2020

مقایسه روش آنگوف مبتنی بر IRT و روش بوکمارک در تعیین استاندارد چیرگی آزمون زبان MSRT

محمدآزاد جلالی زاده[۱*]، علی دلاور[۲]، نورعلی فرخی[۳]، محمد عسگری[۴]

M.A. Jalalizadeh[1], A. Delavar[*2], N.A. Farokhi[3], M. Asgari[4]

دریافت مقاله: ۱۳۹۸/۰۲/۱۲ پذیرش مقاله: ۱۳۹۸/۰۹/۱۶

Received Date: 2019/05/02 Accepted Date: 2019/12/07

چکیده

هدف: هدف اصلی این پژوهش مقایسه روش آنگوف مبتنی بر IRT و روش بوکمارک در تعیین استاندارد چیرگی آزمون زبان MSRT بود.

روش: بدین منظور، یکی از نمونه سؤالات آزمون MSRT (آزمون مورخه شهریور ۱۳۹۷) به‌طور تصادفی انتخاب شد و پاسخ‌های داده شده به سؤالات توسط ۵۹۶ آزمودنی از وزارت علوم اخذ شد. آزمون MSRT دارای ۱۰۰ سؤال بوده که ۳۰ سؤال آن را گرامر، ۳۰ سؤال بخش شنیداری و ۴۰ سؤال آن را بخش درک مطلب خواندن تشکیل می‌دهد. دو پنل تخصصی متشکل از ۱۵ متخصص تدریس آزمون تافل تشکیل شد و سپس با استفاده از روش‌های تعیین استاندارد آنگوف مبتنی بر نظریه سؤال- پاسخ و بوکمارک استانداردهای این آزمون در سه بخش جداگانه (گرامر، شنیداری، درک مطلب خواندن) در سه مرحله ارزیابی مشخص شدند.

یافته‌ها: یافته‌های تحقیق گویای این بود نمره برش به‌دست‌آمده از روش آنگوف مبتنی بر نظریه سؤال پاسخ برابر با ۵۳/۶۶ و روش بوکمارک برابر با ۵۴/۲۷ است. هر دو نمره برش به‌دست‌آمده از نمره تعیین شده توسط وزارت علوم بالاتر بود. همچنین، یافته‌های تحقیق نشان داد که در روش معمول و سنتی وزارت علوم که بر پایه نمره ۵۰ به‌عنوان نمره قبولی عمل می‌شود ۷۳/۳ درصد از شرکت‌کنندگان مردود و ۲۶/۷ درصد از شرکت‌کنندگان قبول شده‌اند اما اساس روش آنگوف مبتنی بر نظریه سؤال- پاسخ آمار مردودی برابر ۷۸/۵ درصد و آمار قبولی برابر با ۲۳/۴ درصد می‌شود. در نهایت براساس بوکمارک آمار مردودی برابر با ۲۱/۵ درصد می‌شود. یافته‌های تحقیق دلالت بر این دارند که طراحان آزمون زبان وزارت علوم نیاز به بازنگری در تعیین نمره استاندارد این آزمون دارند.

کلید واژه‌ها: آزمون زبان MSRT، نمره برش، بوکمارک، آنگوف مبتنی بر نظریه سؤال- پاسخ.

۱. دانشجوی دکتری سنجش و اندازه گیری دانشگاه علامه طباطبایی، سنندج، ایران
۲. استاد سنجش و اندازه‌گیری دانشگاه علامه طباطبایی، تهران، ایران
* نویسنده مسئول: **Email:** alidelavar@yahoo.com
۳. دانشیار گروه سنجش و اندازه‌گیری دانشگاه علامه طباطبایی، تهران، ایران
۴. دانشیار گروه سنجش و اندازه‌گیری دانشگاه علامه طباطبایی، تهران، ایران

مقدمه و بیان مسئله

تعیین استاندارد شاید یکی از شاخصه‌های روان‌سنجی است که بیشتر از هر حوزه دیگر روان‌سنجی آمیزه‌ای از هنر، سیاست و فرهنگ است (Cizek, 2001). اکثر کشورها از آموزش استاندارد محور برای اطمینان از اکتساب حداقل سطوح شایستگی بهره می‌گیرند و این کار را برای تسهیل اندازه‌گیری پیشرفت تحصیلی انجام می‌دهند (Fuhrman, 2001; Klieme et al., 2004; Taylor, 2009). استانداردهای عملکرد آموزشی توصیفی از سطوح عملکرد فراهم می‌کنند که باید محقق شوند (Ravitch, 1995). استانداردهای آموزشی احتمالاً ازاین‌رو مؤثر باشند که آن‌ها به‌وضوح به‌عنوان اهداف پیشرفت تعریف می‌شوند (Locke and Latham, 2002). آن‌ها همچنین به مربیان کمک می‌کنند تا انتظارات خود را از دانش‌آموزان یا دانشجویانشان را افزایش دهند که به‌نوبه خود به‌طور مثبت بر عملکرد یادگیرندگان تأثیر می‌گذارد (de Boer, Bosker and Van der Werf, 2010). چه‌بسا، آن‌ها احتمالاً موجب انگیزش مربیان شود تا از این طریق سبک‌های تطبیقی تدریس را به کار ببرند که با نیازهای یادگیرندگان مطابقت داشته باشد (Ledoux, Blok and Boogaard, 2009).

کاربرد مؤثر استانداردهای عملکرد دلالت بر این دارد که آن‌ها در خودشان معتبر و روا هستند. بااین‌حال تعریف استانداردهای عملکرد تکلیفی مستقیم نیست. به‌منظور پاسخ به سؤالاتی نظیر این سؤال که حداقل نمره برای یادگیرنده به‌منظور عملکرد بهینه در آموزش و جامعه چقدر است؟ متخصصان باید دقیقاً تعیین کنند که چه سطح شایستگی باید برای دانش‌آموز یا یادگیرنده انتظار داشت (McGinty, 2005). در کل، تکلیف پیچیده تعریف استانداردهای آموزشی معمولاً توسط پانلی از متخصصان آموزش صورت می‌گیرد که از آن‌ها خواسته می‌شود استانداردهای آموزشی را توسعه دهند که دیدگاه آن‌ها در حد کفایت بالا و درعین‌حال واقع گرانه باشد. این فرایندهای تصمیم‌گیری توسط روش‌های تعیین استاندارد هدایت می‌شود.

نکاتی که در مقیاس اندازه‌گیری نمرات به‌منظور تفکیک این طبقات عملکردی باید مشخص شوند به‌عنوان نمرات استاندارد، نمرات نقطه برش، یا نمره قبولی مشهور هستند و معمولاً از طریق تحقیقات صرف تجربی به‌دست نمی‌آیند. بلکه در عوض، متخصصان آشنا با جامعه آزمودنی‌ها در آزمون موردنظر و آشنا با محتوی آزمون قضاوت‌هایی درباره سطح کفایت محتوی انجام می‌دهند و این متخصصان نمرات حداقل موردقبول یا نمره لازم حداقلی را برای هر عملکرد تعیین می‌کنند (Clauser, 2013). فرایندهای تعیین نمرات نقطه برش به استاندارد گزینی[1] معروف است و معمولاً روشی سیستماتیک و و مکرری برای جایگذاری عقاید متخصصان در مقیاس نمره محسوب می‌شود. به خاطر اینکه نمره قبولی محصول ارزیابی متخصصان است، هیچ نمره قبولی حقیقی وجود ندارد. بلکه، روش‌های استاندارد گزینی روشی سیستماتیکی برای استنباط نمرات قبولی از طریق فهرست گوناگونی از

1. Standard setting

متخصصان محتوی است که اغلب به توسط شواهد تجربی تحت تأثیر قرار می‌گیرد (Reckase, 2000). این تصمیم‌گیری‌های فردی اغلب به‌وسیله روش‌های گوناگون ترکیب شده تا یک نمره قبولی یگانه‌ای را استخراج کند (برای مثال، Cizek, 2001). تعیین استانداردهای عملکرد بخش اساسی فرایندهای توسعه آزمون برای هر آزمون است که برای طبقه‌بندی افراد به کار می‌رود. نمرات نامتناسب قبولی می‌تواند پیامدهای منفی بلندمدتی برای افراد و جامعه در برداشته باشد.

یکی از مشهورترین روش‌های تنظیم استاندارد که تحقیقات خوبی درباره آن شده است روش Angof (1971) است. این روش روش استاندارد گزینی آزمون‌محور[1] است که ارزیابان به‌جای ارزیابی در باب آزمودنی‌ها در مورد سؤالات آزمون نظر می‌دهند. مثل بقیه روش‌های آزمون محور، متخصصان محتوای آزمون با ملاحظه حداقل توانایی قابل‌قبول برای آزمودنی تصمیم‌گیری انجام می‌دهند.

برخلاف دیگر روش‌های تعیین استاندارد، روند نسبت داده شده به Angof (1971) (و انواع آن) تمرکز اولیه کار اصلی نیست. این روش در ابتدا به‌صورت بخش بسیار کوچکی در ۹۲ صفحه فصل آنگوف در موضوعات درجه‌بندی، هنجاریابی و معادله سازی برای کتاب مرجع معیار آموزشی ویرایش دوم به وجود آمد. فصل اصلی روش آنگوف نامیده می‌شود اگرچه آنگوف این روش را به همکاری در خدمات آزمون آموزشی لدیارد توکر نسبت می‌دهد؛ بنابراین، نام‌گذاری روش به نام روش توکر بسیار مناسب است، اما نمی‌خواهیم برخلاف جریان آب شنا کنیم. روش شرح داده شده توسط (Angof 1971) به‌ندرت دقیقاً به‌صورت که مطرح‌شده استفاده می‌شود. به‌جای آن، پیکربندی جزئی دوباره رویکرد پایه -هر نوع، روش آنگوف اصلاح‌شده نامیده می‌شود- امروزه بسیار رایج است، اگرچه آنچه روش Angof (1971) اصلاح‌شده را دقیقاً تشکیل می‌دهد کمی غیرواضح است؛ اما تقریباً در همه سؤالات در حال حاضر که این روش استفاده می‌شود، قطعاً یک رویکرد Angof (1971) اصلاح‌شده است که بعداً در این فصل شرح داده خواهد شد. علاوه‌بر این رویکرد پایه و Angof (1971) اصلاح‌شده که به اصطلاح روندهای آنگوف توسعه داده شده نامیده می‌شود نیز بسط داده شده است و اخیراً روندی که روش آری/خیر نامیده می‌شود معرفی شده است که بسیار مشابه با این رویکرد اصلی است.

علی‌رغم فقدان شفافیت در نام‌گذاری، مشخص است که روش Angof (1971) (و همه انواع آن) روش بسیار رایج برای قرار دادن استانداردهای عملکردی در استفاده فعلی در زمینه صدور گواهینامه و مدرک است. اگرچه، در زمینه آموزشی بسیار کم استفاده شده است.

اگرچه منطق زیربنایی این روش مستقیم و جذاب است، اما اجرای عملی این روش بسیار پیچیده است (Shepard, 1995; National Research Council, 1999). اولین محدودیت این روش ناتوانی ارزیابان در برآورد معقول از عملکرد آزمودنی‌های دارای حداقل شایستگی در سؤالات آزمون است (Clauser, Harik, Margolis, McManus, Mollon, Chis & Williams, 2008). اگرچه هیچ ملاک

1. Test-centered

مطلقی برای دقت ارزیابان وجود ندارد، ثبات درونی این درجه‌بندی‌ها چارچوب مهمی برای ارزیابی فراهم می‌کند (Kane, 2001). فقدان ثبات درونی معمولاً از طریق عدم توافق بین برآورد احتمالی ارزیابان و دشواری سؤالات نشان داده شده است. علاوه‌بر محدودیت‌های عملی در توانایی ارزیابان برای انجام تکلیف موردنیاز، چندین مسئله نظری در کاربردپذیری روش آنگوف در کاربردهای آزمون‌سازی مدرن وجود دارد. روش تعیین استاندارد آنگوف استانداردهای عملکرد را در چارچوب نظریه آزمون کلاسیک مفهوم‌سازی می‌کند و در نتیجه استانداردهای عملکرد را در سنجه نمره حقیقی تولید می‌کند. در چارچوب نمره حقیقی توانایی مشاهده شده یک آزمودنی به مجموعه خاصی از سؤالات در آزمون بستگی دارد. این نحوه بررسی توانایی آزمودنی به معنای این است که عملکرد نظری آزمودنی با حداقل شایستگی و بنابراین نمره نقطه برش به سؤالات وابسته است. در عمل، تأثیر انتخاب سؤال با تبدیل نقطه برش به مقیاس نمره آزمون در مقیاس شایستگی IRT و از طریق منحنی ویژگی‌های سؤال کاهش پیدا می‌کند، اما این تبدیل نمره قبولی ثابت را بدون توجه به انتخاب سؤالات تضمین نمی‌کند (Ferdous & Plake, 2005). اکثر این محدودیت‌ها می‌تواند به لحاظ نظری با مفهوم‌سازی روش آنگوف در درون چارچوب نظریه سؤال پاسخ کاهش داده شود. مفهوم‌سازی روش آنگوف در درون چارچوب IRT نیاز به هیچ تغییری در فرایند قضاوت ندارند. بلکه، Angof (1971) مبتنی بر IRT به‌سادگی مفاهیم IRT را برای تفسیر درجه‌بندی آنگوف سنتی بکار می‌برد . (Clauser, 2013).

در روش آنگوف توانایی آزمودنی با حداقل شایستگی به‌عنوان مفهوم نظری موجود است و به‌طورکلی از مدل اندازه‌گیری زیربنایی جدا است. اگرچه از ارزیابان انتظار می‌رود تا دیدگاه ثابتی از توانایی آزمودنی را از طریق فرایندهای درجه‌بندی درونی سازند، بااین‌حال هیچ تلاشی برای قرار دادن این صفت زیربنایی در یک مقیاس توانایی صورت نگرفته است. در روش آنگوف قرار گرفته در درون چارچوب IRT، توانایی آزمودنی با حداقل توانایی به‌عنوان نقطه‌ای در مقیاس شایستگی موردنظر قرار می‌گیرد. این فعالیت دلالت بر این ندارد که ارزیابان با مکانیسم‌های IRT یا ویژگی‌های خاص نمره زیربنایی آشنایی دارند. بلکه این روش به‌سادگی نیاز به این دارد که توانایی آزمودنی با حداقل شایستگی می‌تواند در کنار همان مقیاسی قرار بگیرد که توانایی سایر آزمودنی‌های دیگر در ان قرار دارند.کلاسر (Clauser, 2013).

در تلاش برای حل مشکلات مربوط به روش آنگوف روش‌های جدید تعیین نقطه برش در سال‌های اخیر ارائه شده‌اند. یکی از روش‌های استاندارد گزینی که اخیراً موردتوجه محققان قرار گرفته روش بوکمارک است. در واقع روش بوکمارک امر قضاوت را با کاهش و یا تقویت بار شناختی در قضاوت ساده‌سازی می‌کند. بر طبق دیدگاه (Mitzel, 2001) درخواست از اعضای گروه برای برآورد توانایی در سطح سؤال (ویژگی‌های روش آنگوف) از نظر شناختی برای اکثریت اعضای گروه فعالیتی بسیار پیچیده است. توسعه‌دهندگان روش (Lewis, Mitzel, and Green, 1996) بر این امر توافق دارند که اعضای پانل استاندارد گزینی باید مقدمتاً بر روی ابعاد محتوی سنجش تمرکز کنند به‌جای اینکه به برآورد

نحوه عملکرد آزمودنی‌ها در سؤالات آزمون بپردازند. این مباحث به‌طور مستقل توسط (Stone, 2001) در بافت اندازه‌گیری راش انجام شد. توسعه‌دهندگان بوکمارک تأکید بر این دارند که اعضای پانل باید درباره مقدار RP[1] آگاهی پیدا کنند. انتخاب مقدار RP باید تصمیمی مدیریتی باشد که به‌طور پیشرفته در جلسات استاندارد گزینی گرفته باشد.

روش بوکمارک می‌تواند به‌صورت جانشین منطقی مجموعه‌ای از استراتژی‌های نگاشت سؤالات که در دهه ۱۹۹۰ در ارتباط با سنجش استاندارد بسط داده شده در نظر گرفته شود که برای ارزیابی ملی پیشرفت آموزشی (NAEP) توسط محققان در آزمایش کالج آمریکایی (ACT) انجام گرفته است. روش‌های نگاشت سؤالات اولیه به‌صورت روند سنجش استاندارد نسبت به کارکرد بازخورد گنجانده شده در روش‌های دیگر کمتر اعمال شده است. در ۱۹۹۶، برای مثال، محققان در ACT از روند نگاشت سؤالات در ارتباط با روشی که تخمین متوسط نامیده می‌شد بهره گرفتند که بسطی از روش آنگوف اصلاح‌شده (Angof, 1971) است. آن روند نگاشت سؤالات برای آزمایش چندین انتخاب و سؤالات پاسخ ساخته شده اعمال شده است. نگاشت سؤالات برای ارائه بازخورد بعد از دور دوم سؤالات دسته‌بندی شده برای ارزیابی علمی ۱۹۹۶ و مدنی NAEP ۱۹۹۸ و ارزیابی نوشتاری استفاده شده است. نگاشت‌ها محل هر سؤال را در رابطه با نمره درجه‌بندی شبیه به NAEP که همراه با توصیفگرهای سطح مختلف به‌دست‌آمده NAEP بود نشان می‌دهد (ALD که امروزه توصیفگر سطح عملکرد نامیده می‌شود). هر سؤال با انتخاب چندگانه براساس احتمال پاسخ صحیح آن برای هر نمره درجه‌بندی شده نگاشته شده است و هر سؤال پاسخ ساخته شده برای هر نمره یک‌بار نگاشته شده است، یعنی، برای احتمال به‌دست آوردن نمره ۱، ۲، ۳ یا بالاتر در هر نمره رتبه‌بندی شده.

در واقع این روش براساس روش‌های طراحی سؤال مبتنی بر نظریه سؤال - پاسخ است که سؤالات را براساس دشواری مرتب می‌کنند و بر تسهیل تکالیف شناختی موردنیاز برای ارزیابی تعیین استانداردها تأکید می‌کند. به این خاطر که ارزیابان پیش‌بینی‌های سطح سؤال را در رابطه با دشواری مطلق سؤال انجام نمی‌دهند و از این طریق خستگی ارزیابی کاسته می‌شود.کلاسر (Clauser, 2013). برای فهم بهتر موضوع خلاصه‌ای از چگونگی دو روش مطالعه، درپی می‌آید:

روش آنگوف مبتنی بر نظریه سؤال پاسخ: یکی از روش‌های تعیین استاندارد یا نمره برش مبتنی بر سؤال است. در این روش با استفاده از نظریه سؤال پاسخ پارامترهای دشواری سؤالات تهیه می‌شود و در کنار هر یک از سؤالات ضرایب دشواری آن‌ها درج می‌شود تا گروهی از متخصصان و ارزیابان در حوزه مربوطه (برای مثال زبان انگلیسی) به برآورد احتمالات پاسخ یک داوطلب مرزی یا داوطلب دارای حداقل شایستگی در آن حوزه برای هر یک از سؤالات بپردازند. سپس نمرات میانگین ارائه‌شده توسط ارزیابان باهمدیگر جمع بسته می‌شود و این فرایند معمولاً در سه مرحله انجام می‌شود و در هر

1. Response probobility

مرحله بازخوردهای لازم به اعضای گروه داده می‌شود و در نهایت در مرحله سوم یک نمره برش کلی که همان نمره استاندارد آزمون با استفاده از روش آنگوف مبتنی بر نظریه سؤال پاسخ است حاصل می‌شود.

روش بوکمارک: یکی از روش‌های سؤال محور برای تعیین نمره برش است که در آن سؤالات ابتدا براساس نظریه سؤال پاسخ به‌ترتیب ضریب دشواری رتبه‌بندی می‌شوند و کتابچه طبقه‌بندی سؤالات به تفکیک از سؤالات آسان به دشوار تهیه و در اختیار گروهی از ارزیابان متخصص در حوزه مرتبط (برای مثال زبان انگلیسی) قرار داده می‌شود. سپس ارزیابان در چندین مرحله یا دور احتمالات پاسخ یک آزمودنی دارای حداقل شایستگی به هر یک از سؤالات را مشخص می‌کند. در نهایت میانگین احتمالات داده شده به هر یک از سؤالات با همدیگر جمع شده و نمره برش برای آزمون حاصل می‌شود.

با توجه به این تاکنون هیچ پژوهشی در جامعه ایرانی در راستای تعیین نقطه برش در زمینه آزمون زبان وزارت علوم صورت نگرفته است و از طرف دیگر هیچ پژوهش دیگری در جهت مقایسه دو روش مدرن تعیین استاندارد صورت نگرفته است، در این تحقیق محقق قصد دارد تا این دو روش استاندارد گزینی نوین را در آزمون MSRT مورد مقایسه قرار دهد.

روش‌شناسی پژوهش

تحقیق فعلی در زیرگروه تحقیقات ارزیابی کیفی ملاک درونی آزمون قرار می‌گیرد. (Hambleton & Pitoniak, 2006). از نظر هدف نیز این تحقیق در حوزه تحقیقات کاربردی قرار می‌گیرد. چراکه نتایج تحقیق می‌تواند مجریان آزمون زبان وزارت علوم را در تعیین نمره برش مناسب برای انتخاب دانشجویان دارای شایستگی زبان یاری کند. از نظر جمع‌آوری داده‌ها، پژوهش فعلی در زیرگروه تحقیقات زمینه‌ای است.

جامعه پژوهش فعلی را کلیه آزمودنی‌هایی تشکیل می‌دهند که به آزمون زبان وزارت علوم در سال گذشته پاسخ داده‌اند. گروه مطالعه در این تحقیق متشکل از ۵۹۶ آزمودنی است که در تاریخ شهریور ماه ۱۳۹۷ در آزمون MSRT شرکت کرده بودند. اعضای پانل تعیین استاندارد متشکل از ۱۵ استاد آزمون تافل بودند که حداقل ۵ سال سابقه تدریس در دوره‌های مرتبط با آزمون تافل و آزمون‌های مشابه نظیر تولیمو، آیلتس، MSRT را داشتند. علت انتخاب ۱۵ نفر ارزیاب به این خاطر است که بر طبق ادبیات نظری تعیین استاندارد، بایستی تعداد ارزیابان حداقل بین ۱۰ تا ۱۵ نفر باشد.

برای روش آنگوف مبتنی بر نظریه سؤال- پاسخ در اولین روز کارگاه، پژوهشگر درباره پژوهش و فرایند آن توضیحات کافی ارائه داد. اگر هر یک از اعضای پنل سؤالی درباره پژوهش یا فرایند آن داشتند، پژوهشگر به سؤالات آن پاسخ می‌داد. سپس پژوهشگر فرم رضایت‌نامه شرکت در پژوهش را

بین شرکت‌کنندگان پخش کرد و از آن‌ها خواست در صورت تمایل به شرکت در پژوهش با آن موافقت کنند. همه شرکت‌کنندگان فرم را امضا کردند.

جلسات آموزشی در اصل از مراحل آموزش استانداردسازی CEFR تبعیت می‌کرد .کنسول اروپایی (Council of European, 2009). پژوهشگر در ابتدا اهداف آزمون را شرح داد و برای اعضای پانل از محتوی آزمون و توصیفات سطح عملکرد توضیحات مفصلی ارائه کرد. سپس اعضایی که در مورد سطح عملکرد سؤالاتی داشتند سؤالات خود را پرسیدند و با اعضای دیگر گروه این مباحث ادامه پیدا کرد تا زمانی که هیچ ابهامی در این زمینه وجود نداشت. بعدازاین گام در مورد تعریف داوطلبان یا دانشجویان دارای سطح مرزی بحث و گفتگو شد. برای آشنایی اعضای گروه با نوع سؤالات و محدوده دشواری سؤالات که آن‌ها در آزمون واقعی خواهند دید، یک فعالیت تمرینی اجرا شد. ۱۰ سؤال تمرینی از نسخه‌های قدیمی آزمون MSRT ارائه شد. برای هر سؤال، اعضای پانل نشان دادند که آیا داوطلب یا دانشجوی مرزی می‌تواند به سؤالات پاسخ صحیح بدهد یا نه. راهنمایی ارائه شده برای ارزیابان در این تکلیف به این صورت ارائه شد:

لطفاً هر یک از سؤالات را بخوانید و درباره این موضوع تصمیم بگیرید که آیا دانشجو یا داوطلب مرزی که شما در ذهن خود متصور می‌شوید می‌تواند به هر یک از این سؤالات پاسخ صحیح بدهد یا نه. اگر شما فکر می‌کنید که می‌تواند پاسخ صحیح بدهد نمره ۱ را به ایشان بدهید و اگر فکر می‌کنید که او قادر به پاسخ صحیح نیست نمره ۰ را به او بدهید.

بعدازاینکه اعضای پنل قضاوت‌هایشان در مورد آزمون تمرینی را تکمیل کردند، هر سؤال شرح داده شد. شرح و بسط در مورد سؤال شامل دلایل اعضای پانل برای پاسخ ۱ یا صفر به سؤالات بود. از اعضای پانل خواسته شد تا توضیح دهند چرا نمره ۱ یا صفر به توصیف اولیه‌ای که از ویژگی‌های داوطلبان مرزی دارند داده‌اند. در مرحله بعد به ارزیابان گفته شد که احتمال پاسخ درست توسط داوطلب مرزی را برای هر یک از سؤالات مشخص کنند و تمرینات مختلفی بر روی سؤالات آزمون تمرینی انجام دادند.

بعد از جلسه آموزشی، اعضای پانل کتابچه‌ای با ۱۰۰ سؤال آزمون MSRT را دریافت کردند. علاوه‌بر این سؤالات، نسبت پاسخ درست و دشواری سؤالات که با استفاده از مدل نظریه سؤال پاسخ سه پارامتری فراهم شده بودند برای هر سؤال به‌طور مجزا برای داوطلبان فراهم شد. پژوهشگر توضیح داد که دشواری سؤال احتمالاً تحت تأثیر عامل حدس زدن باشد؛ بااین‌حال، اعضای پنل باید عمدتاً بر روی محتوی سؤالات آزمون متمرکز شوند. براساس سطح عملکرد، اعضای پنل مشخص کردند که دانشجویان یا داوطلبان مرزی می‌توانند به این سؤالات پاسخ صحیح بدهند. در اولین مرحله تعیین استاندارد، ارزیابان همه ۱۰۰ سؤال آزمون را برای تکمیل قضاوت‌های دور اول درجه‌بندی کردند.

بعد از تکمیل همه قضاوت‌ها، ورقه درجه‌بندی همه اعضای پنل برای سطح عملکرد مرزی جمع‌آوری شد. میانگین بزرگ قضاوت‌های تعیین شده توسط اعضای پنل موردمحاسبه قرار گرفت و

این مقدار مساوی با سهم نسب درست همه سؤالات بود. داوطلبان با مقدار نسبت درست در داده‌های آزمون واقعی مشخص شدند.

براساس تحلیل برگه‌های کاری، به اعضای پنل بازخوردی در مورد درجه‌بندی‌هایشان در مرحله اولیه داده شد که شامل درجه‌بندی کلی توسط اعضای پنل برای هر سؤال بود، میانگین و میانه نمرات برش، مقادیر دشواری سؤال و واریانس سؤالات نیز به آن‌ها بازخورد داده شد. اعضای پنل به‌مرور این بازخوردها پرداختند و یک ساعت بحث و گفتگو در این مورد داشتند. در طی توضیحات، پژوهشگر از اعضای پنل خواست که بار دیگر قضاوت‌هایشان را براساس سطح عملکرد ارائه دهند. فرایند مشابه در سه مرحله تکرار شد. البته در مراحل دوم و سوم علاوه بر بازخوردهای فراهم شده در مرحله اول، داده‌های تأثیرگذار، شامل درصدهای آزمودنی‌ها در هر طبقه عملکردی براساس داده‌های آزمون واقعی به ارزیابان ارائه شد. محاسبه نمرات برش پیشنهاد شده براساس داده‌های به‌دست‌آمده در مرحله سوم بود.

مراحل انجام شده برای روش بوکمارک با معرفی روش بوکمارک برای اعضای پنل شروع شد و به دنبال آن یک جلسه تمرینی که سؤالات براساس دشواری سؤالات از ساده‌ترین به سخت‌ترین سؤالات از طریق مدل نظریه سؤال پاسخ سه پارامتری درجه‌بندی شده بود. این سؤالات همان سؤالاتی بودند که در روش آنگوف مبتنی بر نظریه سؤال- پاسخ در مورد آن‌ها بحث و گفتگو شده بود. بااین‌حال در روش بوکمارک سؤالات براساس دشواری آن‌ها درجه‌بندی شده بود.

بعد از جلسه تمرینی، اعضای پنل یک برگ کپی از OIB (کتابچه رتبه‌بندی سؤالات) را دریافت کردند، درحالی‌که مفهوم‌سازی گروه مرزی برای سطح عملکرد داوطلب دارای شایستگی زبانی توضیح داده شد، اعضای پنل با سؤالات ساده شروع کردند و تا آنجایی در کتابچه رتبه‌بندی شده سؤالات پیش رفتند که اعتقاد داشتند دانشجویان یا داوطلبان مرزی احتمالاً شانس ۶۷٪ درصدی برای پاسخ به همه سؤالات باقیمانده دارند. به متخصصان گفته شد که شماره صفحه آن سؤال را یادداشت کنند. تمرین کوتاه قبل از اولیه مرحله اجرا شد. راهنمایی ارائه شده به این صورت بود:

لطفاً تا زمانی بر روی کتابچه کار کنید که به صحفه ای برسید که در آن سؤالی در کتابچه از دیدگاه شما، دانشجوی مرزی دارای ۶۷ درصد شانس پاسخ به سؤالات قبل از آن سؤال را دارا باشد. برای سؤالات بعد از آن سؤال مشخص، شانس پاسخ صحیح به زیر ۶۷ درصد افت خواهد کرد. صفحه سؤال مشخصه را در برگه خود یادداشت فرمایید.

بعد از مشخص شدن نمرات برش قضاوت‌های اعضای پنل برای هر سطح، متخصصان بازخوردهای مشابه با بازخوردهای روش آنگوف دریافت کردند، شامل نمودارهایی که توزیع هر مکان بوکمارک اعضای پنل را نشان می‌داد و درصد تجمعی داوطلبان در هر سطح عملکرد (بین مرحله دوم و سوم). بحث گروهی بین هر مرحله صورت گرفت. مباحث بر دلایل مکان‌های بوکمارک و ویژگی‌های سؤالات

و دانش و مهارت داوطلبان مرزی متمرکز بود. فرایندهای مشابه همچنین در سه مرحله بوکمارک تکرار شد.

در روش آنگوف معمولی (سنتی)، همانند دو روش دیگر روال معارفه و کار تمرینی با ۱۰ سؤال MSRT صورت گرفت و سپس ارزیابان در سه دور ارزیابی به قضاوت درباره سؤالات و احتمال پاسخ‌دهی داوطلبان مرزی یا داوطلبان با حداقل شایستگی پاسخ به سؤالات مشخص شد. تنها تفاوت این روش با دو روش دیگر این بود که در روش آنگوف سنتی کتابچه آزمون یا ضرایب دشواری سؤالات روشن نبود و بلکه فرایند قضاوت مبتنی بر ارزیابی خود ارزیابان و مباحث صورت گرفته در جلسات استوار بود.

در ابتدا، روش آنگوف- مبتنی بر IRT اجرا شد. اعضای پانل به‌مرور هر یک از سؤالات به‌صورت انفرادی پرداختند و آن را به‌صورت فردی مطالعه کردند. بعد از تصمیم‌گیری فردی، اعضای پانل تصمیم‌گیری‌های خود را در دورهای (راند) چندگانه قبل از محاسبه درجه‌بندی نهایی مورد تعدیل قرار دادند. در واقع، نتیجه مطلوب این است که یک نمره برش از طریق تکرارهای مختلف به دست بیاید. درجه‌بندی در سه دور انجام شد. در هر مرحله درجه‌بندی، اعضای پانل نتایج درجه‌بندی را بر روی فرم‌های درجه‌بندی تهیه شده ثبت کردند که مورد جمع‌آوری و ثبت قرار گرفت. اطلاعات به‌دست‌آمده بعد از هر جلسه درجه‌بندی فراهم شد و اعضای پانل آن را موردبحث قرار دادند. نتایج جلسه درجه‌بندی قبلی، نرخ قبولی حاصله و اطلاعات سؤال واقعی به‌عنوان یک بازخورد قبل از جلسه بعدی داده شد و به‌عنوان اساسی برای تصمیم‌گیری بکار رفت.

همچنین در پنل دیگر، از روش بوکمارک برای تعیین نقطه برش استفاده شد. اعضای پانل نقطه‌ای را نشان دادند که شخصی با حداقل شایستگی می‌تواند به سؤالات پاسخ دهد. بعد از تأیید نتایج فردی، اعضای پانل نمرات خود را از طریق مباحث گروهی مورد تعدیل قرار دادند. بعد از تأیید نمرات تعدیل‌یافته و تأیید نرخ قبولی، دور (راند) با این عقیده که نمره برش دیگر مورد تعدیل قرار نخواهد گرفت پایان یافت.

یافته‌های پژوهش

سؤال اول: نمره برش آزمون زبان وزارت علوم با استفاده از روش آنگوف مبتنی بر نظریه سؤال پاسخ چه نمره‌ای است؟

در جدول (۱) نمرات استاندارد چیرگی سه بخش آزمون MSRT (گرامر، درک مطلب خواندن، شنیداری) استخراج شده از طریق روش آنگوف مبتنی بر نظریه سؤال پاسخ گزارش شده است.

جدول (۱): نمرات استاندارد چیرگی سه بخش آزمون MSRT (گرامر، درک مطلب خواندن، شنیداری) استخراج شده از طریق روش آنگوف مبتنی بر نظریه سؤال پاسخ

	مرحله اول		مرحله دوم		مرحله سوم	
	میانگین (انحراف معیار)	نمره برش	میانگین (انحراف معیار)	نمره برش	میانگین (انحراف معیار)	نمره برش
گرامر	۵۷/۸۴ (۳/۴۵)	۱۷/۳۵	۵۸/۰۴ (۲/۵۵)	۱۷/۴۱	۵۸/۰۶ (۱/۹۰)	۱۷/۴۲
درک مطلب	۵۱/۲۶ (۳/۰۸)	۲۰/۵۰	۵۱/۸۴ (۲/۵۱)	۲۰/۷۳	۵۲/۱۷ (۲/۲۲)	۲۰/۸۶
شنیداری	۴۹/۰۰۸ (۱/۸۱)	۱۴/۷۱	۵۰/۱۰ (۱/۴۶)	۱۵/۰۳	۵۰/۷۴ (۱/۰۳)	۱۵/۲۲
نمره کل آزمون MSRT	۵۲/۷۰ (۲/۲۷)		۵۳/۳۲ (۱/۷۵)		۵۳/۶۶ (۱/۳۹)	

با توجه به یافته‌های گزارش شده در این جدول مشخص است که نمره برش برای بخش گرامر در مرحله اول برابر با ۱۷/۳۵، مرحله دوم ۱۷/۴۱ و در مرحله سوم برابر با ۱۷/۴۲ است؛ نمره برش برای بخش درک مطلب خواندن در مرحله اول برابر با ۲۰/۵۰، مرحله دوم برابر با ۲۰/۷۳ و در مرحله سوم برابر با ۲۰/۸۶ است. همچنین، نمره برش برای بخش شنیداری در مرحله اول برابر با ۱۴/۷۱، در مرحله دوم برابر با ۱۵/۰۳ و در مرحله سوم برابر با ۱۵/۲۲ است.

در نهایت یافته اصلی این پژوهش این بود که با استفاده از روش آنگوف مبتنی بر نظریه سؤال پاسخ نمره برش کل آزمون MSRT در مرحله اول برابر با ۵۲/۷۰، مرحله دوم برابر با ۵۳/۳۲ و مرحله سوم نیز برابر با ۵۳/۶۶ بود. در نهایت، می‌توان نتیجه گرفت که با استفاده از روش آنگوف مبتنی بر نظریه سؤال پاسخ نمره برش آزمون MSRT برابر با ۵۳/۶۶ است.

سؤال دوم: نمره برش آزمون زبان وزارت علوم به‌دست‌آمده در روش بوکمارک چه نمره‌ای است؟

در جدول (۲) نمرات استاندارد چیرگی سه بخش آزمون MSRT (گرامر، درک مطلب خواندن، شنیداری) استخراج شده از طریق روش بوکمارک گزارش شده است.

جدول (۲): نمرات استاندارد چیرگی سه بخش آزمون MSRT (گرامر، درک مطلب خواندن، شنیداری) استخراج شده از طریق روش بوکمارک

	مرحله اول		مرحله دوم		مرحله سوم	
	میانگین (انحراف معیار)	نمره برش	میانگین (انحراف معیار)	نمره برش	میانگین (انحراف معیار)	نمره برش
گرامر	۵۶/۶۱ (۱/۶۷)	۱۶/۹۸	۵۸/۰۶ (۱/۷۰)	۱۷/۴۲	۵۸/۵۰ (۱/۲۷)	۱۷/۵۵
درک مطلب	۵۰/۸۶ (۱/۹۰۵)	۲۰/۳۴	۵۱/۳۹ (۲/۳۴)	۲۰/۵۵	۵۲/۱۰ (۰/۸۳۷)	۲۰/۸۴
شنیداری	۵۰/۱۴ (۲/۵۱)	۱۵/۰۴	۵۱/۶۹ (۲/۵۵)	۱۵/۵۰	۵۲/۲۰ (۰/۹۸۲)	۱۵/۶۶
نمره کل آزمون MSRT	۵۲/۵۴ (۱/۳۶)		۵۳/۷۱ (۱/۶۸)		۵۴/۲۷ (۰/۷۵۶)	

با توجه به یافته‌های گزارش شده در این جدول مشخص است که نمره برش برای بخش گرامر در مرحله اول برابر با ۱۶/۹۸، مرحله دوم ۱۷/۴۲ و در مرحله سوم برابر با ۱۷/۵۵ است؛ نمره برش برای بخش درک مطلب خواندن در مرحله اول برابر با ۲۰/۳۴، مرحله دوم برابر با ۲۰/۵۵ و در مرحله سوم برابر با ۲۰/۸۴ است. همچنین، نمره برش برای بخش شنیداری در مرحله اول برابر با ۱۵/۰۴، در مرحله دوم برابر با ۱۵/۵۰ و در مرحله سوم برابر با ۱۵/۶۶ است.

در نهایت یافته اصلی این پژوهش این بود که با استفاده از روش بوکمارک نمره برش کل آزمون MSRT در مرحله اول برابر با ۵۲/۵۴، مرحله دوم برابر با ۵۳/۷۱ و مرحله سوم نیز برابر با ۵۴/۲۷ بود. در نهایت، می‌توان نتیجه گرفت که با استفاده از روش بوکمارک نمره برش آزمون MSRT برابر با ۵۴/۲۷ است.

سؤال سوم: آیا طبقه‌بندی افراد به دو مقوله قبول/ مردود با روش تعیین استاندارد آنگوف، آنگوف مبتنی بر نظریه سؤال- پاسخ، روش بوکمارک و روش سنتی وزارت علوم به نتایج متفاوتی می‌انجامد. در گام بعدی این پژوهش به طبقه‌بندی داوطلبان مردودی و قبولی براساس استانداردهای به‌دست‌آمده از روش آنگوف مبتنی بر سؤال- پاسخ و بوکمارک و روش مرسوم وزارت علوم پرداخته شد. با توجه به یافته‌های پژوهش فعلی که نمره برش آزمون MSRT برای روش آنگوف مبتنی بر نظریه سؤال- پاسخ برابر با ۵۳/۶۶، برای روش بوکمارک برابر با ۵۴/۲۷ به‌دست آمد و ازآنجاکه نمره برش تعیین شده توسط وزارت علوم برای این آزمون برابر با ۵۰ است. داده‌ها ۵۹۶ نفر از آزمودنی‌ها بر طبق این نمرات برش در هر یک از روش‌های تعیین استاندارد به دو بخش قبولی و مردودی تقسیم شد. نتیجه آزمون خی دو و توزیع فراوانی قبولی/ مردودی براساس این چهار روش در جدول (۳) گزارش شده است.

جدول (۳): توزیع فراوانی قبولی/ مردودی براساس چهار روش تعیین استاندارد

ملاک	سنتی (روش معمول وزارت علوم)	آنگوف مبتنی بر نظریه سؤال-پاسخ	بوکمارک	خی دو
مردود (درصد)	۴۳۶ (٪ ۷۳/۳)	۴۵۶ (٪ ۷۸/۵)	۴۶۷ (٪۷۸/۵)	$X^2(3)=4.86;$ $p>.05$
قبولی (درصد)	۱۵۹ (٪ ۲۶/۷)	۱۳۹ (٪ ۲۳/۴)	۱۲۸ (٪ ۲۱/۵)	

با توجه به یافته‌های گزارش شده در جدول (۳) مشخص است که در روش معمول و سنتی وزارت علوم که بر پایه نمره ۵۰ به‌عنوان نمره قبولی عمل می‌شود ۷۳/۳ درصد از شرکت‌کنندگان مردود و ۲۶/۷ درصد از شرکت‌کنندگان قبول شده‌اند؛ اما براساس روش آنگوف مبتنی بر نظریه سؤال- پاسخ آمار مردودی برابر ۷۸/۵ درصد و آمار قبولی برابر با ۲۳/۴ درصد می‌شود. در نهایت براساس بوکمارک آمار مردودی برابر با ۲۱/۵ درصد می‌شود.

آزمون خی دو برای بررسی مقایسه‌ای توزیع فراوانی مردودی‌ها و قبولی‌ها با توجه به ۳ روش حاکی از عدم معنی‌داری بود. این نتیجه نشان می‌دهد که بین سه روش تعیین استاندارد در میزان قبولی و مردودی داوطلبان آزمون زبان MSRT تفاوت معنی‌داری وجود ندارد.

سؤال چهارم: آیا قضاوت‌های ارزیابان در مراحل سه‌گانه تعیین استاندارد از همسانی (توافق) مطلوبی برخوردار است؟ به‌منظور بررسی همسانی (توافق) قضاوت‌های هر یک از ارزیابان در سه مرحله تعیین استاندارد و توافق بین ارزیابان در هر یک از مراحل اجرای تکنیک‌های بوکمارک و آنگوف مبتنی بر نظریه سؤال پاسخ محاسبه گردید.

در جدول (۴) ضریب همسانی کاپای کوهن درون ارزیابان برای هر یک از روش‌ها گزارش شده است. با توجه به یافته‌های گزارش شده در این جدول مشخص است که ضرایب همسانی درون ارزیابان برای هر دو روش مشابه با هم است.

جدول (۴): مقادیر ضرایب کاپای کوهن درون ارزیابان برای هر یک از روش‌ها

عضو پنل	مرحله	آنگوف مبتنی بر نظریه سؤال پاسخ		بوکمارک	
		اول	دوم	اول	دوم
ارزیاب ۱	دوم	۰/۷۷		۰/۵۵	
	سوم	۰/۸۲	۰/۸۹	۰/۶۲	۰/۷۵
ارزیاب ۲	دوم	۰/۶۶		۰/۵۲	
	سوم	۰/۶۹	۰/۷۸	۰/۴۸	۰/۷۲
ارزیاب ۳	دوم	۰/۶۸		۰/۷۰	
	سوم	۰/۷۳	۰/۷۷	۰/۷۴	۰/۸۰
ارزیاب ۴	دوم	۰/۷۱		۰/۸۲	
	سوم	۰/۷۴	۰/۸۱	۰/۸۵	۰/۸۶
ارزیاب ۵	دوم	۰/۷۲		۰/۶۹	
	سوم	۰/۷۶	۰/۷۹	۰/۷۲	۰/۷۷
ارزیاب ۶	دوم	۰/۸۱		۰/۸۰	
	سوم	۰/۷۸	۰/۸۵	۰/۷۶	۰/۸۴
ارزیاب ۷	دوم	۰/۷۶		۰/۸۳	
	سوم	۰/۷۹	۰/۸۸	۰/۸۴	۰/۸۸
ارزیاب ۸	دوم	۰/۶۹		۰/۷۲	
	سوم	۰/۷۶	۰/۸۰	۰/۷۷	۰/۹۰
ارزیاب ۹	دوم	۰/۶۹		۰/۵۵	
	سوم	۰/۷۰	۰/۸۳	۰/۵۹	۰/۷۲
ارزیاب ۱۰	دوم	۰/۶۱		۰/۷۴	
	سوم	۰/۶۶	۰/۷۳	۰/۶۹	۰/۷۸
ارزیاب ۱۱	دوم	۰/۸۲		۰/۷۸	
	سوم	۰/۸۵	۰/۸۶	۰/۸۰	۰/۸۴
ارزیاب ۱۲	دوم	۰/۸۹		۰/۷۹	
	سوم	۰/۹۱	۰/۹۰	۰/۷۵	۰/۸۸
ارزیاب ۱۳	دوم	۰/۶۹		۰/۶۹	
	سوم	۰/۷۴	۰/۸۳	۰/۷۱	۰/۷۷
ارزیاب ۱۴	دوم	۰/۸۳		۰/۷۵	
	سوم	۰/۸۴	۰/۸۵	۰/۷۲	۰/۸۵
ارزیاب ۱۵	دوم	۰/۸۰		۰/۷۴	
	سوم	۰/۸۱	۰/۸۹	۰/۶۹	۰/۸۶

سؤال پنجم: آیا استانداردهای تعیین شده از نظر شواهد روایی رویه‌ای از مقبولیت برخوردار است؟ در انتهای جلسات تعیین استاندارد برای هر یک از روش‌های تعیین استاندارد (آنگوف مبتنی بر سؤال- پاسخ و بوکمارک) از اعضای پنل های تخصص خواسته شد تا نظرات خود را در قالب یک فرم ارزیابی مطابق جدول (۵) و در یک طیف پاسخ کاملاً مخالفم تا کاملاً موافقم پنج‌درجه‌ای تکمیل کنند. در این فرم آیتم‌هایی در خصوص هدف مطالعه تعیین استاندارد، روشن بودن آموزش استاندارد، مفید بودن بازخورد و بحث بین مراحل گنجانده شده است.

جدول (۵): نظرات اعضای پنل های ارزیابی پیرامون تعیین استاندارد به تفکیک روش‌ها

ردیف	موضوع	روش	
		روش آنگوف مبتنی بر نظر سؤال پاسخ	روش بوکمارک
۱	هدف مطالعه تعیین استاندارد قابل‌درک بود.	(۰/۴۲)۴/۰۵	(۰/۵۱)۴/۱۵
۲	توضیحات ارائه شده درباره تعیین نمره برش روشن و مناسب بود.	(۰/۵۱)۴/۱۹	(۰/۴۸)۴/۱۷
۳	ارائه بازخورد و بحث بین مراحل مفید و سودمند بود.	(۰/۷۲)۴/۳۶	(۰/۶۹)۴/۴۲
۴	فرایند قضاوت در خصوص تعیین نمره برش آسان بود.	(۰/۷۹)۳/۷۸	(۰/۸۵)۳/۵۲
۵	نمره برش تعیین شده معقول، مناسب و قابل دفاع است.	(۰/۸۰)۴/۲۹	(۰/۷۲)۴/۳۶

برای بررسی این سؤال، از فرم ارزیابی تعیین استاندارد آزمون استفاده شد، این فرم توسط عباسی (Abbasi2013) توسعه داده شده است و شامل ۵ آیتم است که در یک طیف لیکرت ۵ گزینه‌ای از کاملاً مخالفم تا کاملاً موافقم نمره‌گذاری می‌شود و به سنجش روایی رویه‌ای می‌پردازد. سؤالات این فرم قابل‌درک بودن روش تعیین استاندارد، روشن بودن توضیحات ارائه شده درباره تعیین نمره برش، سودمندی ارائه بازخورد و بحث بین مراحل، آسان بودن فرایند قضاوت در خصوص تعیین نمره برش و معقول بودن نمره برش را موردسنجش قرار می‌دهد. با توجه به یافته‌های گزارش شده در جدول (۵) مشخص است که از نظر قابل‌درک بودن هدف مطالعه، تعیین استاندارد، توضیحات ارائه شده درباره تعیین نمره برش، فرایند قضاوت و نمره برش معقول هر دو روش تقریباً نمرات میانگین مشابهی کسب کردند. با توجه به یافته‌های گزارش شده در این جدول مشخص است که از نظر شواهد رویه‌ای در اکثر موارد بین دو روش آنگوف مبتنی بر نظریه سؤال – پاسخ و روش بوکمارک تفاوت معناداری وجود ندارد.

بحث و نتیجه‌گیری

با توجه به اینکه آزمون زبان MSRT یکی از آزمون‌های سرنوشت‌سازی است که سالیانه هزاران نفر شرکت‌کننده در آن شرکت می‌کنند، این آزمون نیاز به بررسی ویژگی‌های روان‌سنجی ویژه‌ای دارد. متأسفانه، مطالعه ادبیات پژوهشی نشانگر این بود که تاکنون هیچ مطالعه‌ای در زمینه تعیین نقطه برش این آزمون صورت نگرفته است. این ضعف ادبیات موجب شد که این پژوهش برای اولین بار به‌عنوان یک پژوهش پیشگام در جهت تعیین نمره برش آزمون MSRT انجام شود. یافته‌های پژوهش فعلی نشان داد که استفاده از روش‌های تعیین استاندارد بوکمارک، آنگوف و آنگوف مبتنی بر نظریه سؤال-پاسخ به برآورد نمرات بالاتری از مقدار تعیین شده توسط وزارت علوم می‌انجامد (نمره ۵۰). این یافته در حقیقت گویای این است که طراحان آزمون زبان وزارت علوم نیاز به بازنگری در روش‌های تعیین استاندارد این آزمون هستند. هرچند که از طریق یافته‌های تحقیق فعلی نمی‌توان مقدار دقیقی برای نمره برش این آزمون تعیین کرد ولی این تحقیق به‌عنوان یک پژوهش پیشگام می‌تواند روش استاندارد تعیین شده توسط وزارت علوم را به چالش بکشد و طراحان این آزمون را برای انتخاب روش‌هایی نوین در جهت انتخاب یک روش بهینه و مدرن یاری کند. در زمینه تفاوت بین روش‌های بکار رفته در این تحقیق باید گفت که نتایج تحقیق فعلی نشان داد که روش بوکمارک نسبت به دو روش دیگر نمره برش بالاتری را پیشنهاد کرد. همچنین یافته مهم این تحقیق این بود که روش‌های بوکمارک و آنگوف مبتنی بر نظریه سؤال- پاسخ همسانی بین ارزیابان بالاتری نسبت به روش آنگوف دارد. در کل، یافته‌های این پژوهش گویای این است که وزارت علوم نیاز به بهبود و ارتقای روش‌های تعیین نمره برش برای آزمون‌های سرنوشت‌ساز دارد. البته باید به این واقعیت اذعان کرد که در خارج از کشور نیز هنوز تحقیقات خیلی کمی در زمینه تعیین نمره برش آزمون‌هایی نظیر تافل صورت گرفته است که نشان می‌دهد تعیین استانداردهای آزمون‌های سرنوشت‌ساز نظیر آزمون MSRT باید موردتوجه قرار بگیرد.

این پژوهش همانند همه پژوهش‌ها دارای محدودیت‌هایی است. ازجمله، به علت محدودیت‌های اجرایی پژوهش فقط یک نمونه سؤال آزمون MSRT موردبررسی قرار گرفت، بنابراین از تعمیم شتاب‌زده برای تعیین استاندارد نمره زبان MSRT باید اجتناب شود. همچنین، به علت محدودیت‌های اجرایی تنها سه روش تعیین نمره برش موردبررسی قرار گرفت و این در حالی است که روش‌های نوین دیگری در این زمینه در ادبیات پژوهش مطرح هستند. از طرف دیگر، در این پژوهش به علت هزینه‌های اجرایی بالا تنها ۱۵ ارزیاب برای هر پنل استفاده شد، این در حالی است که تقریباً از حداقل تعداد اعضای پنل برای هر روش استفاده شد.

این پژوهش به‌عنوان یک پژوهش پیشگام در زمینه تعیین نقطه برش برای آزمون زبان MSRT است، باید توجه کرد که نمی‌توان از طریق نتایج پژوهش فعلی اقدام به تصمیم‌گیری در زمینه تعیین

نمره برش این آزمون کرد بلکه توصیه می‌شود پژوهش‌های مشابهی با تحقیق فعلی برای حصول یک نمره دقیق‌تر استاندارد برای این آزمون استفاده شود.

تعیین نمره برش یک آزمون یکی از مهم‌ترین اقدامات طراحان آزمون در راستای روایی ابزار محسوب می‌شود، همچنین بررسی روایی درونی و بیرونی معیار روایی یک نمره بسیار مهمی است که باید موردتوجه قرار بگیرد.کین(Kane,1994). بررسی‌های درونی در مورد روایی نمره برش به همسانی نتایج توجه می‌کند، به‌ویژه همسانی اعضای پنل در فرایند تصمیم‌گیری و قضاوت‌هایشان معطوف می‌شود. بررسی روایی بیرونی نتایج را با تصمیم‌های اتخاذ شده با استفاده نمرات برش مشابه با انواع تصمیم‌های گرفته شده از طریق فرایندهای جایگزین موردبررسی قرار می‌دهد، به‌ویژه این کار را با فرایندهای تعیین استاندارد معادل انجام می‌دهد.

با توجه به یافته‌های تحقیق فعلی، مقایسه نمرات برش حاصل از روش‌های بوکمارک و آنگوف در تحقیق فعلی نتایج ضدونقیضی با یافته‌های تحقیقات گذشته به همراه داشت. برای مثال، یافته‌های (Wang, 2003) دال بر این بود که نمره برش به‌دست‌آمده از روش بوکمارک در مقایسه با نمرات به‌دست‌آمده از روش آنگوف پایین‌تر است و یا پژوهش‌های انجام شده توسط (Peterson, 2013)، (Buckendahl, Smith, Impara & Plake, 2002.) نشان داد که نمرات برش روش‌های بوکمارک و آنگوف با همدیگر تفاوتی ندارند. بااین‌حال یافته‌های پژوهش فعلی با یافته‌های پژوهش انجام شده توسط هسیه (Hsieh,2013) که در مورد مقایسه نمرات برش دو روش آنگوف و بوکمارک در زمینه آزمون زبان انگلیسی بود نتایج مشابهی به همراه داشت. به‌طوری‌که یافته‌های تحقیق فعلی همانند پژوهش انجام شده توسط هسیه (Hsieh, 2013) نشان داد که نمرات برش روش بوکمارک بالاتر از روش آنگوف است. در تبیین این یافته می‌توان گفت که روش آنگوف نسبت به روش بوکمارک از قضاوت‌های کمتر عینی و بیشتر مبتنی بر برداشت ذهنی ارزیابان استفاده می‌کند. در واقع، در روش بوکمارک با توجه به اینکه بازخوردهای داده شده به ارزیابان بیشتر مبتنی بر واقعیات پاسخ‌ها بوده به همین خاطر ارزیابان اطلاعات آماری قوی‌تری نسبت به روش انگوف دارند. ثانیاً، تحقیقات گذشته در زمینه مقایسه روش‌های (1971) Angofحوزه‌های دیگری نظیر ریاضیات را که عمدتاً آزمونی تک‌بعدی محسوب می‌شود موردمطالعه قرار داده‌اند و این در حالی است که پژوهش فعلی از آزمون زبان که دارای سه بعد شنیداری، خواندن و شنیداری بودند استفاده کرده است و وقتی از این آزمون در یک کتابچه سؤال رتبه‌بندی شده استفاده می‌شود احتمالاً موجب سردرگمی ارزیابان در روش بوکمارک می‌شود.

در زمینه روایی درونی و بیرونی نمرات برش نتایج یافته‌های تحقیق با نتایج یافته‌های پژوهش دیگر تا حدود زیادی همسو است. به‌عبارت‌دیگر، یافته‌های تحقیق گویای این بود که (1971) Angof مبتنی بر نظریه سؤال پاسخ به علت بازخورد واقعی که از آمار مربوط به سؤالات آزمون برای ارزیابان فراهم می‌سازد با احتمال زیادی منجر به این می‌شود که همسانی پاسخ‌های بین ارزیابان و درون

ارزیابان بالاتر از روش آنگوف ساده باشد. همچنین همین بازخوردهای واقعی از آمار مربوط به سؤالات آزمون موجب می‌شود که آن‌ها رضایت بیشتری از قضاوت‌های خود در تعیین استانداردها و نمرات برش داشته باشند. در کل، می‌توان این‌گونه استنباط کرد که روش‌های بوکمارک و آنگوف مبتنی بر نظریه سؤال پاسخ با توجه به بازخوردهای واقعی از ویژگی‌های سؤالات کار و تکلیف ارزیابان را برای درجه‌بندی و تعیین نمرات برش با سهولت همراه می‌سازد و موجب می‌شود که ارزیابان نمرات برش عینی‌تری را برای سؤالات و آزمون ارائه دهند. همچنین مشخص است که استفاده از هر سه روش تعیین نمره برش نسبت به روش سنتی وزارت علوم باید ترجیح داده شود و یافته‌های این پژوهش دارای مهم‌ترین دلالت بر تغییر نگرش متصدیان وزارت علوم در زمینه جایگزینی روش تعیین نمره برش آزمون MSRT با یکی از روش‌های معرفی شده در تحقیق فعلی را روش می‌سازد.

پیشنهادات کاربردی

این پژوهش به‌عنوان یک پژوهش مقدماتی در زمینه تعیین نمره برش برای آزمون زبان MSRT است، باید توجه کرد که نمی‌توان از طریق نتایج پژوهش فعلی اقدام به تصمیم‌گیری در زمینه تعیین نمره برش این آزمون کرد بلکه توصیه می‌شود پژوهش‌های مشابهی با تحقیق فعلی برای حصول یک نمره دقیق‌تر استاندارد برای این آزمون استفاده شود. در واقع، یافته‌های تحقیق فعلی بازخوردهای مفید و مناسبی در زمینه استفاده از روش‌های مدرن تعیین نمره استاندارد برای آزمون‌های سرنوشت‌ساز نظیر MSRT برای متصدیان اجرای این آزمون به همراه دارد.

References:

ACT Inc. (2005d). Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: *Pilot study report to COSDAM. Iowa City, IA: Author.*

ACT Inc. (2007a). Developing achievement levels on the 2006 National Assessment of Educational Progress in grade 12 economics: Process report. Iowa City, IA: Author.

Andrich, D. (1978a). Rating formulation for ordered response categories. *Psychometrika*, *49*(4), 561–573.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.; pp. 508-600). Washington, DC: American Council on Education.

Bechger, T. M., Kuijper, H., & Maris, G. (2009). Standard setting in relation to the common European framework of reference for languages: The case of the state examination of Dutch as a second language. *Language Assessment Quarterly*, *6*, 126–150.

Behuniak, P., Archambault, F. X., & Gale, R. K. (1982). Angoff and Nedelsky standard setting procedures. Implications for the validity of proficiency score interpretation.

Educational and Psychological Measurement, 42(1), 247–255. doi:10l1177/0013164482421031.

Berk, R. A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). Applied Measurement in Education, 9(3), 215–225.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4*(2), 219–240. doi:10.1177/014662168000400209.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4(2), 219–240

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253–263. doi:10.1111/j.1745-3984.2002.tb01177.x.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clauser, B. E., Harik, P., Margolis, M. J., McManus, I., Mollon, J., Chis, L., & Williams, S. (2008). An empirical examination of the impact of group discussion and examinee performance information on judgments made in the Angoff standard-setting procedure. Applied Measurement in Education, 22(1), 1–21.

Clauser, B. E., Mee, J., & Margolis, M. J. (2011, April). The effect of data format on integration of performance data into Angoff judgments. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Clauser, B.E., Harik, P., Margolis, M.J., McManus, I.C., Mollon, J., Chis, L., & Williams, S. (2009). Empirical evidence for the evaluation of performance standards estimated using the Angoff procedure. *Applied Measurement in Education, 22,* 1-21.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J., & Dillon, G.F. (2009). Judges' use of examinee performance data in an Angoff standard-setting exercise for a medical licensing examination: An experimental study. *Journal of Educational Measurement,* 46(4), 390-407.

Clauser, B.E., Swanson, D.B., & Harik, P. (2002). A multivariate generalizability analysis of the impact of training and examinee performance information on judgments made in an Angoff-style standard-setting procedure. *Journal of Educational Measurement, 39, 269-290.*

Clauser, J.C. (2013). Examination of the Application of Item Response Theory to the Angoff Standard Setting Procedure. Onpuplished *Dissertations. University of Massachusetts – Amherst.*

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centred method for setting standards on achievement test. Applied Measurement in Education, 12(4), 343-366.

Dawber, T., & Lewis, D. M. (2002). The cognitive experience of bookmark standard setting participants. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved from http:// www2.education.ualberta.ca/educ/psych/crame/files/standard_setting.pdf

Ferdous, A. A., & Plake, B. S. (2005). Understanding the factors that influence decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18(3), 257–267.*

Ferdous, A. A., & Plake, B. S. (2008). Item response theory-based approaches for computing minimum passing scores from Angoff-based standard-setting study. *Educational and Psychological Measurement. 68 (5),778-796.*

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18,* 223–232.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18(3), 223–232*.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22–32. doi:10.1111/j.1745-3992.2003.tb00113.x.

Halpin, G., Sigmon, G, & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. *Educational and Psychological Measurement, 43(1), 185–196. doi:10.1177/001316448304300126.*

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.) *Educational measurement* (4th ed.). Westport, CT: American Council on Education & Praeger.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement (4th ed.) (pp. 433–470). Westport, CT: Praeger Publishe*rs.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24(4), 355-366.*

Harsch, C., & Rupp, A. (2011). Designing and scaling level-specific CEFR Writing Tasks. *Language Assessment Quarterly, 8*, 1–33.

Hein, S. F., & Skaggs, G. E. (2009). A qualitative investigation of panelists' experiences of standard setting using two variations of the bookmark method. Applied Measurement in Education, 22(3), 207–228.

Hsieh, M. (2013). An application of Multifaceted Rasch measurement in the Yes/No Angoff standard setting procedure. Language Testing. 30(4). 112-132.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63(4), 584–601.*

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34(4), 353–366*

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64(3), 425–461.

Kane, M. (1994a, October). *Examinee-centered vs. task-centered standard setting*. Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments, Washington, DC.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). Mahwah, NJ: Lawrence Earlbaum Associates, Publishers.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Lewis, D. M., Mitzel, H. C., and Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Livingstone, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. Applied Measurement in Education, 2(2), 121-141.

M. Alimirzaie , A. Moghadam zadeh,A. Minaei , B. Ezanloo , K. Salehi .(2019). Sources of theDifferential ItemFunctioning and its Application inEducation, Journal of Research in Teaching(Vol 7, No 1, Spring 2019).

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. Applied Measurement in Education, 18(3), 269–287

McGinty, D. (2005). Illuminating the "black box" of standard setting: An exploratory qualitative study. Applied Measurement in Education, 18(3), 269–287.

Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.

National Research Council. (1999). Setting reasonable and useful performance standards. In J. W. Pelligrino, L. R. Jones,. & K. J. Mitchell (Eds.), Grading the nation's report card: *Evaluating NAEP and transforming the assessment of educational progress (pp. 162–184). Washington, DC: National Academy Press.*

N. Yousofi , S. Ebadi , M. Saedi Dovaise . Investigating the L2 Motivation of theUndergraduat Students from thePerspective of the "L2 Motivational Self System". Journal of Research in Teaching, Vol 5, No 3, Autumn 2017

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4*, 295–317.

of Applied Measurement, 2, 187-201.

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35(2–3), 95–101.*

Pant, H. A., Rupp, A. A., Tiffin-Richards, S. P., & Köller, O. (2009). Validity issues in standard-setting studies. *Studies in Educational Evaluation, 35(2–3), 95–101.*

Pellegrino, J. W., Jones, L. R., & Mitchell, K. J. (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Shepard, L. A., Glaser, R., Linn, R. L., & Bohrnstedt, G. (1993). Setting performance standards for student achievement. A report of the National Academy of Education panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992

achievement levels. Stanford, CA: Stanford University, National Academy of Education.

Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Stanford, CA: National Academy of Education.

Shepard, L.A. (1995). Implications for standard setting of the National Academy of

Skorupski, W. P., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard-setting studies? Applied Measurement in Education, 18(3), 233–256.

Stone, G. E. (2001). Objective standard setting (or truth in advertising). *Journal*

Stone, G. E., Beltyukova, S., & Fox, C. M. (2008). Objective standard setting for judge-mediated examinations. *International Journal of Testing*, *8*, 180–196. doi: 10.1080/15305050802007083

Tannenbaum, R. J., & Caroline Wylie, E. (2005). Mapping English Language Proficiency Test Scores Onto the Common European Framework. ETS Research Report Series.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, *40*(3), 231–253. doi:10.1111/j.1745-3984.2003. tb01106.x.