تاریخ دریافت مقاله: ۹۸/۱۲/۸
تاریخ پذیرش مقاله: ۹۹/۳/۲۴

فصلنامه علمی- پژوهشی روان‌سنجی
دورهٔ هشتم، شمارهٔ ۳۲، بهار ۱۳۹۹
صفحات: ۴۲- ۲۳

مقایسه روش‌های ساختن نمره‌ی کل در آزمون‌های مرکب با بهره‌گیری از خطای استاندارد اندازه‌گیری شرطی

Comparing the efficiency of composite scores constructing methods in the Test Battery by Utilization of conditional standard error of measurement

***مجتبی جهانی فر [۱] حجت اله درفش [۲]**

چکیده

از آزمون‌های مرکب برای تصمیم‌گیری‌های متفاوت در آموزش استفاده می‌شود. تصمیم پذیرش در یک آزمون مرکب بیشتر براساس نمره کل است. روش‌های متفاوتی برای ساختن نمره‌کل می‌توان طرح ریزی کرد و بر اساس نمره‌کل‌های بدست آمده نسبت به پذیرش و یا عدم پذیرش افراد تصمیم گرفت. اما کدام روش نمره‌کل‌سازی می‌تواند با دقت بیشتر و خطای کمتر ساخته شود و از سایر روش‌های نمره‌کل‌سازی کارآمدتر باشد؟ این پژوهش با هدف مقایسه کارآمدی روش‌هایی که برای ساختن نمره‌کل در آزمون‌های مرکب استفاده می‌شود، انجام شده‌است. از ۱۰۰۰۰ نمونه تصادفی آزمون سراسری سال ۱۳۹۵در هفت خرده‌آزمون برای بررسی شش روش نمره‌کل‌سازی استفاده‌شده‌است. برای کاهش آشفتگی در توزیع نمره‌ها از روش هموارسازی کرنل و از طرح‌های وزن‌دهی اسمی و موثر برای ساخت

Abstract

The test is used for decision-making in education. Decision that is taken by results of a test raises sensitivity. Admission decision is taken according to score that is acquired in test. If the test is composed of several sub-tests with different content is named Test Battery and the resulting score is called composite score. Different methods can be designed to construct composite scores, and based on the composite scores, admission decision is made toward accepting or not accepting inviduals. But which constructing technique can be design with more precision and less error? So, it is more efficient than other composite score constructing techniques. This study aimed to compare the effectiveness of ways to make the composite score

۱ . نویسنده مسئول: عضو هیات علمی دانشگاه شهید چمران اهواز، اهواز، ایران
Email: m.jahanifar@scu.ac.ir

۲ . عضو هیات علمی دانشگاه شهید چمران اهواز، اهواز، ایران.

نمره‌ی‌کل استفاده گردید. به منظور بررسی کارآمدی روش‌های نمره‌کل‌سازی، از مفهوم خطای استاندارد اندازه‌گیری شرطی استفاده شد. نتایج حاکی از آن بود که روشهای نمره‌کل‌سازی که از تبدیل نرمال و طرح‌های وزن‌دهی اسمی به همراه پیش‌هموارسازی برای تبدیل نمره‌های خام به نمره‌های مقیاس استفاده کرده‌اند، و یا روش‌هایی که از تبدیل غیرخطی آرک سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس استفاده کرده‌اند، در مقایسه با سایر روشهای استفاده شده در این پژوهش مجذور میانگین خطای استاندارد اندازه‌گیری شرطی کمتری داشته و کارآمدتر بوده اند.

واژه‌های‌کلیدی: آزمون مرکب، نمره‌کل، نمره‌ی مقیاس، طرح وزن‌دهی، خطای استاندارد اندازه‌گیری شرطی

that is used in the Test Battery was conducted. 10,000 random real samples from Iran universities entrance exam in seven sub-tests have used to compare efficiency of 6 methods that was designed to construct composite scores. Raw scores obtained from correct answers, so Normalizing and Arcsine methods was used to construct scale scores. To reduce irregularity in the distribution of scores, kernel smoothing methods and to make composite scores, nominal and effective weighting schemes were used. To evaluate the effectiveness of the composite scores constructing methods, the conditional standard error of measurement was used. Results showed that composite scores that used Arcsine scale scores or normalization scale score with pre-smoothing had less average conditional standard error of measurement and more efficiency in compare to other methods

Keywords: Test Battery, Composite score, Scale score, Weighting scheme, CSEM

مقدمه

کاربرد آزمون‌های روانی و تربیتی در چند سال گذشته، هم در مدارس ابتدایی و متوسطه، هم در بخش‌های خصوصی و هم در سطح کشور از طریق آزمون‌های سراسری و همگانی و هم در سطح بین‌المللی مانند آزمون‌های زبان انگلیسی پیشرفته به نحو سریعی رشد کرده است و در زمینه‌های گوناگون آموزشی برای تصمیم‌گیری‌های مختلف از لحاظ امتحانات دبستانی و دبیرستانی، ارتقا از یک کلاس به کلاس دیگر، طبقه بندی شاگردان، ارزشیابی کارآمدی مدارس، طبقه بندی و ارزشیابی معلمان، پذیرش دانشجو و استخدام کارمند به گونه فزاینده‌ای عمومیت پیدا کرده است. طی دهه‌های اخیر از آزمون‌ها برای روشن ساختن انواع یادگیری، پیدا کردن کج فهمی ها و اختلالات یادگیری، در برنامه ریزی های آموزشی و درسی ، آزمون‌های گروهی مانند آزمون‌های

سراسری و همچنین نگرش سنج ها، مشاهده‌های رفتاری و مشاوره‌های تحصیلی و شغلی استفاده فراوانی شده‌است. با توجه به آنچه که در توصیف و کاربرد آزمونها گفته شد، می‌توان از دیدگاه درجه‌ی اهمیت آزمون در تصمیم‌گیری و سرنوشت افراد، آزمونها را به دو دسته‌ی آزمونهای سرنوشت‌ساز[1] آزمونهای غیرسرنوشت‌ساز[2]تقسیم کرد(ساتن، ۲۰۰۴). هر آزمونی که نتایج آن برای آزمون‌شونده حیاتی و مهم باشد را آزمون سرنوشت‌ساز می‌نامند، به طوری که گذراندن این آزمون با موفقیت برای آزمون‌شونده نتایج سودمندی همچون اخذ مدرک تحصیلی، بورسیه تحصیلی، پذیرش در موسسه یا دانشگاه یا به دست آوردن یک شغل در بر دارد، و به همان نسبت عدم موفقیت در این آزمون پیامدهایی همچون از دست دادن فرصت تحصیل و یا شغل را در پی خواهدداشت. نتایج و نمره‌های حاصل از این آزمونها برای تنبیه، تشویق، تنزیل و یا ارتقا افراد استفاده می‌شود. در مقابل این آزمونها آزمونهای غیر سرنوشت‌ساز نیز وجود دارند که از آنها برای اندازه‌گیری پیشرفت تحصیلی و یا اطلاع پیداکردن از شیوه‌های تدریس و یادگیری استفاده می‌شود(ساتن، ۲۰۰۴). آنچه که این دو نمونه آزمون را از هم جدا می‌کند نه شکل برگزاری و نه نحوه طراحی آنها است، بلکه وجه تمایز بین آزمونهای سرنوشت‌ساز و غیرسرنوشت‌ساز در نحوه‌ی تفسیر و تصمیم‌گیری درباره نتایج این دو آزمون است، عاملی که اهمیت و حساسیت آزمون‌ها را بالا می‌برد خود آزمون نیست، بلکه تصمیم‌هایی است که قرار است بر اساس آن آزمون اخذ شود، اینکه آزمون به صورت چند گزینه ای باشد یا شفاهی آن را خطیر جلوه نمی دهد، مهم نوع تصمیمی است که قرار است براساس نتایج آن آزمون گرفته‌شود.

آزمون‌ها با شکل‌ها و روش‌های مختلف توسط موسسه ها و شرکت‌های مختلفی طراحی و اجرا می‌شوند، این آزمون‌ها در برخی موارد تنها به یک موضوع مشخص پرداخته و گاهی به صورت ترکیبی از موضوع‌های مختلف هستند، به عنوان مثال آزمون ورودی دانشگاهها به طور معمول مشتمل بر چند خرده‌آزمون می‌شود، که به آن آزمون‌های مرکب[3]می‌گویند. آزمون‌های مرکب شامل چند خرده‌آزمون مختلف است که هر کدام نمره خاصی را به خود اختصاص می‌دهند، در برخی موارد نمره‌ای به نام نمره‌مرکب[4](نمره‌کل) برای آزمون‌های مرکب ساخته می‌شود. نمره‌کل بازتاب‌دهنده عملکرد آزمودنی در چند خرده‌آزمون است. به عنوان مثال نمره‌کل آزمون سراسری ایران بازتاب دهنده عملکرد آزمون‌شونده در دروس عمومی و دروس اختصاصی است، در برخی آزمون‌ها مثل آزمون سراسری نمره‌های مقیاس را با وزن‌های مختلف با هم ترکیب می‌کنند و نمره‌کل را می‌سازند(نقی زاده، ۱۳۹۴) و در برخی دیگر آزمون‌ها مانند ACT[5] از وزن‌های برابر

1. high stakes
2. low stakes
3. Test Battery
4. Composite score
5 American College Testing

برای ساخت نمره‌کل استفاده می‌کنند(راهنمای فنی *ACT*، ۲۰۱۴). وزن‌ها سهم هر خرده‌آزمون را در نمره‌کل مشخص می‌کنند. برای ترکیب‌کردن نمره‌ها و ساختن نمره‌کل از ویژگی‌های مختلف خرده‌آزمون‌ها استفاده می‌شود، که ویژگیهایی از قبیل دشواری سؤالها، واریانس نمره‌های مقیاس، خطای استاندارد اندازه‌گیری نمره‌ها، ضریب پایایی نمره‌ها و همبستگی بین این خرده‌آزمون ها از مهمترین این ویژگی ها هستند(ونگ و استانلی،۱۹۷۰). آزمون *SAT*[1]برای تبدیل نمره‌های خام به نمره‌های مقیاس از گروه مرجع بهره می‌برد. این آزمون که سه ساعت و چهل‌وپنج دقیقه به طول می‌انجامد شامل خرده‌آزمون‌های مهارتهای خواندن، نوشتن و همچنین ریاضیات است. برای مقیاس‌سازی در این آزمون از تبدیل نرمال استفاده می‌شود. در این تبدیل نمره‌های فرمولی[2] برحسب فراوانی تراکمی، به مقیاس نرمال برده می‌شوند، و حاصل این مقیاس‌سازی جدول تبدیلی است که در آن هر نمره بر اساس رتبه و نمره درصدی به مقیاس آزمون *SAT* برده می‌شود (گزارش کالج برد، ۲۰۱۵). آزمون *IBTS* که در دانشگاه آیووا به منظور اندازه‌گیری پیشرفت تحصیلی از سطح مهد کودک تا پایه دوازدهم طراحی شده است، دارای ۱۵ خرده‌آزمون است. البته این خرده‌آزمون‌ها برای پایه های نهم تا دوازدهم شامل نه خرده‌آزمون است. این خرده‌آزمون‌ها شامل مهارتهای خواندن و نوشتن، مهارتهای دستور زبان، مهارتهای شنیداری، مهارتهای محاسباتی و ریاضیات، علوم و مهارتهای اجتماعی می‌باشد. که هر کدام هم در تعداد سؤالهای خرده‌آزمون و هم در زمان پاسخگویی متفاوت هستند. در این آزمون نیز از روش تبدیل نرمال برای ساختن مقیاس نمره‌ها استفاده شده است. (گزارش فنی *IBTS*، ۲۰۱۶). آزمون سراسری ورود به دانشگاههای ایران نیز آزمونی مرکب است که با توجه به رشته تحصیلی دارای خرده‌آزمون‌های متفاوت می‌باشد، در این آزمون نیز پس از محاسبه نمره‌خام، نمره‌ها به روش تبدیل نرمال به مقیاس مشترک برده می‌شوند. نمره های مقیاس در آزمون سراسری ایران به نمره های تراز شهرت دارند و وزنهای اسمی نابرابر برای ساخت نمره کل استفاده می‌شود(نقی زاده، ۱۳۹۴).

روش‌های متفاوتی را می‌توان برای ساختن نمره‌کل پیشنهاد داد که هر کدام می‌توانند ترکیبی از روش‌های مختلف ساختن نمره‌های مقیاس، روش های هموارسازی و روش‌های وزن‌دهی باشند، اما مساله اینجاست که با تغییر روش ساختن نمره‌کل، نمره اختصاص داده‌شده به هر شرکت‌کننده و به تبع آن تصمیم برای پذیرش و یا عدم پذیرش وی در آزمون، دچار تغییر خواهد شد. کدام روش برای ساختن نمره‌کل خطای کمتری دارد و به اصطلاح کارآمدتر است؟ و اختصاص نمره‌کل با کدام روش به شرکت‌کننده دچار خطای اندازه‌گیری کمتری خواهیم شد؟ گالیکسن[3](۱۹۵۰) در فصل بیستم کتاب نظریه آزمون های روانی درباره روش‌های مختلف وزن‌دهی به طور گسترده‌ای

[1] Scholastic Aptitude Test
[2] Formula score
[3] Gulliksen

بحث کرده‌است، هدف گالیکسن از ترکیب نمره‌ها ایجاد نمره‌کل با ضریب پایایی بالا بوده‌است. گالیکسن نشان‌داد که اگر تعداد زیادی آزمون که همبستگی بالایی با هم دارند ترکیب شوند، طرح‌های مختلف وزن‌دهی تغییر چشم‌گیری بر روی پایایی نمره‌کل نخواهدداشت، در صورتی که همبستگی بین این خرده‌آزمون‌ها کم باشد تأثیر ترکیب آنها از طریق روش‌های وزن‌دهی مشهودتر خواهدبود. آلن و ین در سال ۱۹۷۹ کار مشابهی را از طریق وزن‌های رگرسیون چندگانه انجام دادند که با وجود متغیرهای پیش‌بینی‌کننده (خرده‌آزمون) کم و استقلال آنها از همدیگر، طرحهای وزن‌دهی مختلف تفاوت آشکاری را از خود نشان دادند به طوری که انتخاب یک طرح بهینه از طریق کمترین مقدار خطا امکان پذیر می‌شد. خدایی و همکاران(۱۳۹۱) طی پژوهشی بر روی طرحهای مختلف وزن‌دهی نشان دادند که طرحهای مختلف که به ویژگی های آزمون بستگی دارند از جمله طرح وزن‌دهی برحسب ضریب پایایی، واریانس و یا خطای اندازه‌گیری، تأثیر چندانی بر روی ضریب پایایی نمره‌های مرکب ندارند، و این تغییر در ضریب پایایی چندان رضایت بخش نیست بلکه تغییر در مقدار پایایی را در ساخت خرده‌آزمون های خوش‌ساخت و پایا دانسته‌اند. طرحهای استفاده شده در این پژوهش شامل طرحهای: روش متوسط ضریب همبستگی پیرسون، ضریب وزن عاملی و ضرایب رگرسیون بودند. پی و مالر[1](۲۰۰۶) طی یک پژوهش شبیه‌سازی نتایج گالیکسن را تایید کرده و نشان دادند که برخی عوامل مانند تعداد خرده‌آزمون ها و خواص روان‌سنجی آنها بر روایی و پایایی آزمون مرکب تأثیر خواهندداشت. از این گونه پژوهش های شبیه‌سازی برای بررسی تأثیر ویژگی های مختلف خرده‌آزمون ها بر روی نمره‌کل را می‌توان در کین و کیس(۲۰۰۴) و رودنر[2](۲۰۰۱) مشاهده کرد. چانگ (۲۰۰۹) با بررسی پنج طرح وزن‌دهی مختلف برای تولید نمره‌کل نشان داد اگر شاخصه هایی مانند ضریب پایایی را به عنوان شاخص های دقت نمره‌کل در نظر بگیریم، طرحهای مختلف وزن‌دهی تأثیر قابل ملاحظه‌ای بر روی آن نداشته و مقدار ضریب پایایی برای همه طرحهای وزن‌دهی مقدار بالا و قابل قبولی خواهد بود، اما اگر به لحاظ سهم هر خرده‌آزمون در تولید نمره‌کل به عنوان شاخص بنگریم، طرحهای وزن‌دهی مختلف باعث ایجاد سهم های مختلف برای خرده‌آزمون‌ها در تولید نمره‌کل خواهندبود. طرحهای وزن‌دهی در پژوهش چانگ شامل: طرح وزنهای برابر، طرح وزن ضریب پایایی، طرح وزن واریانس نمره‌ها، طرح وزن خطای استاندارد اندازه‌گیری و طرح وزن‌دهی بر حسب طول آزمون برای تولید نمره‌کل بود.

با بررسی که در پژوهشهای گذشته صورت گرفته بیشتر آنها به منظور مقایسه روشهای متفاوت ساختن نمره کل، از ضریب پایایی استفاده کرده اند، در این بررسی‌ها روشهای نمره کل سازی ضریب پایایی تفاوت آشکاری نداشته است، به طوری که اگر بخواهیم جمع بندی از مقایسه نتایج

[1] Pei and maller
[2] Rudner

پژوهش‌های مختلف داشته باشیم، این مقایسه را می‌توان چنین خلاصه‌کرد که همه پژوهش‌ها به این نتیجه رسیده اند که هیچ برتری شاخصی در انتخاب روشهای ساختن نمره کل با معیار ضریب پایایی وجود نداشته و ضریب پایایی به شکلی که در این پژوهش‌ها محاسبه شده است نمی‌تواند ملاک مناسبی برای بررسی کارآمدی روشهای ساخت نمره‌کل باشد. لذا در این پژوهش به منظور مقایسه روشهای متفاوت ساختن نمره‌کل به دنبال رویکرد خطای استاندارد اندازه گیری شرطی رفته‌ایم.

آزمون سراسری بزرگترین آزمون انتخابی برای پذیرش دانشجو در دانشگاهها و موسسه های دولتی و غیردولتی در ایران است. چندین دهه است که سازمان سنجش آموزش کشور مسؤولیت اجرای این آزمون را به عهده دارد. تقریبا تمام کسانی که داوطلب ورود به آموزش عالی در ایران هستند باید کنکور بدهند. کنکور یا همان آزمون سراسری هر ساله به طور متوسط با یک میلیون ایرانی ارتباط مستقیم دارد که البته با در نظر گرفتن خانواده های داوطلبان بالای چند میلیون ایرانی هر ساله به نوعی با این پدیده اجتماعی در ارتباط هستند. یکی از مراحل مهم برگزاری آزمون سراسری تولید نمره‌کل و سپس پذیرش افراد بر اساس آن می‌باشد، حساسیت نتایج آزمون سراسری و پذیرش افراد در موسسات آموزش عالی و یا عدم پذیرش آنها تحلیل نتایج آزمون و تفسیر نمره‌ها با خطای کمتر و تفسیرپذیری بیشتر را ضروری و با اهمیت جلوه می دهد، همچنین برگزاری آزمون استاندارد و خطای کم به تحقق عدالت آموزشی هر چه بیشتر خواهد انجامید. در این پژوهش کارآمدی شش روش نمره‌کل‌سازی معرفی خواهد شد و هدف کلی آن بررسی روش‌های مختلف نمره‌کل‌سازی است که برای پذیرش افراد به کار می رود. با بهره‌گیری از نتایج این بررسی به دنبال روشی کارآمد برای تشکیل نمره‌کل هستیم. روشی که ضمن داشتن ویژگی‌های مناسب آماری و روان‌سنجی، کمترین خطا را متوجه شرکت‌کنندگان در آزمون مرکب کند. سوال اصلی این پژوهش این است که در میان روشهای متفاوت ساختن نمره‌کل کدام از همه کارآمدتر است؟

روش

روش پژوهش، جامعه، حجم نمونه و روش نمونه گیری: پژوهش حاضر پژوهشی کمی و از نوع توصیفی (غیر آزمایشی) است و به منظور توسعه دانش کاربردی در زمینه ساختن نمره مقیاس انجام گرفته که با این رویکرد، پژوهشی کاربردی محسوب می‌شود. جامعه مورد نظر در این پژوهش داوطلبان شرکت کننده در آزمون سراسری سال ۱۳۹۵ در گروه آزمایشی ریاضی و فنی هستند، در سال۱۳۹۵تعداد ۱۶۲۸۷۹ نفر در آزمون سراسری در رشته ریاضی و فنی شرکت کرده‌اند. در این پژوهش روش‌های مختلف مقیاس‌سازی و تحلیل‌ها بر روی ۱۰۰۰۰ شرکت کننده در

خرده‌آزمون‌های مختلف عمومی و اختصاصی آزمون سراسری ایران در گروه آزمایشی رشته ریاضی و فنی اجرا شده است. با توجه به اینکه در این پژوهش با مقیاس بزرگ[1]سر و کار داریم از طریق قاعده سرانگشتی[2]می‌توان حجم نمونه نزدیک به ۱۰۰۰۰ نفر را مناسب دانست که البته به صورت تصادفی انتخاب شده اند. در آزمون سراسری و در گروه آزمایشی ریاضی‌وفنی چهار درس عمومی زبان و ادبیات فارسی (۲۵ سوال)، زبان و ادبیات عربی (۲۵ سوال)، معارف اسلامی (۲۵ سوال) و زبان انگلیسی (۲۵ سوال) و سه درس اختصاصی ریاضیات (۵۵ سوال)، فیزیک (۴۵ سوال) و شیمی (۳۵ سوال)، مورد آزمون قرار می‌گیرند.

ابزار پژوهش: ابزار اصلی گردآوری داده در این پژوهش همان سؤال‌های آزمون سراسری است، داده‌های این آزمون در اختیار سازمان سنجش آموزش کشور بوده و به منظور تحلیل آزمون ۱۳۹۵ گروه آزمایشی ریاضی و فنی از آنها بهره گرفته شده است. برای برخی از روش ها و تبدیل هایی که در این پژوهش از آنها استفاده می‌شود، نرم افزار تجاری به بازار عرضه نشده است به همین منظور در طی انجام پژوهش از نرم افزار کد نویسی ریاضی و آمار **MATLAB**[3] برای نوشتن کدهای مربوط به برخی روش‌ها و تبدیل ها استفاده شد.

روند انجام پژوهش

تبدیل نمره‌های خام به نمره‌های مقیاس

برای به دست‌آوردن نمره‌مقیاس به دو روش عمل شده است :

الف) روش تبدیل مقیاس نرمال: در این روش نمره های خام با استفاده از تابع توزیع تجمعی نرمال به نمره هایz تبدیل و سپس با استفاده از تبدیل خطی به مقیاسی بین ۰ تا ۱۰۰۰۰ برده می شوند(کولن، ۲۰۱۴ و آنگوف، ۱۹۷۱).

ب) روش تبدیل مقیاس آرک سینوس: در این روش نمره‌های خام با استفاده از رابطه تبدیل آرک سینوس به نمره‌های مقیاس برده می‌شوند، در اینجا از تعداد پاسخ های درست به هر خرده‌آزمون و تعداد سؤالهای خرده‌آزمون برای ساختن نمره‌مقیاس استفاده می‌شود(کولن، ۲۰۱۴). این مقیاس نیز با استفاده از تبدیل خطی به دامنه ای بین ۰ تا ۱۰۰۰۰ برده می‌شود.

هموارسازی توزیع نمره‌های خام

در این پژوهش از روش هموارسازی کرنل دوجمله‌ای برای پیش‌هموارسازی استفاده‌شده‌است. ایده‌ای که در پس این روش پنهان شده این است که برای هر نمره تابع چگالی احتمال معرفی می‌شود، که به آن تابع چگالی احتمال کرنل یا هسته می‌گویند. معمولا از داده‌های پیوسته برای

[1] Large-scale assessment
[2] Thumbnail rule
[3] MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language.

هموارسازی استفاده می‌شود و از توزیع نرمال برای چگالی بهره می‌برند. ولی نمره‌های خام معمولا گسسته هستند به همین خاطر از کرنل دوجمله‌ای برای هموارسازی استفاده می‌شود. در این روش از پارامتر h برای کنترل درجه هموارسازی استفاده می‌شود. به عنوان مثال برای یک آزمون که دارای k سؤال است، و توزیع نمونه‌ای $\hat{f}(i)$ را دارد که در آن $i = ۰,۱,۲,...,k$ برآوردگر کرنل به صورت زیر است:

$$\hat{f}_h(i) = \begin{cases} \sum_{j=i-h/۲}^{i+h/۲} B(j-i+h/۲|h)\hat{f}(j) \\ B(m|h) = \binom{h}{m}(\frac{۱}{۲})^m(۱-\frac{۱}{۲})^{h-m} \end{cases} \quad (۱)$$

در رابطه(۱)، h عددی زوج و پارامتر هموارسازی است و $m = ۰,۱,۲,...,h$. اگر مقدار h صفر باشد مانند آن است که هموارسازی نشده‌است، و مقدار آن بستگی به شباهت بیشتر نمونه هموارشده با اصلی دارد، به عنوان مثال برای $h = ۲$ هموارسازی کرنل به صورت $\hat{f}(i) = \frac{۱}{۴}\hat{f}(i-۱) + \frac{۱}{۲}\hat{f}(i) + \frac{۱}{۴}\hat{f}(i+۱)$ است. تعیین مقدار h از طریق آزمایش و خطا است. در روش کرنل به جای فراوانی هر نمره میانگین وزن‌دار نمره‌های قبل و بعد هر نمره قرار داده‌می‌شود تا توزیع هموار شود (کولن، ۲۰۱۴و کولن ۱۹۹۱). در این پژوهش برای انتخاب بهترین h برای هموارسازی ضمن مقایسه چهار گشتاور اول نمره‌های مشاهدشده و نمره‌های هموارشده به مقایسه نمودار فراوانی آنها نیز پرداخته می‌شود. در این روش نمره‌ها برای h های مختلف هموار می‌شوند و مقدار مطلوب h مقداری خواهد بود که عبارت رابطه (۲) را به کمترین مقدار خود برساند.

$$M_h = \sum_{i=-h/۲}^{k+\frac{h}{۲}} \hat{f}_h(i) - \frac{۲}{n-۱}\left\{\left[\sum_{۰}^{k}\hat{f}(i)\hat{f}_h(i)\right] - \frac{B(h/۲|h)}{n}\right\} \quad (۲)$$

طرح‌های وزن‌دهی برای ساختن نمره‌کل

برای محاسبه نمره‌کل در این پژوهش از دو طرح وزن‌دهی استفاده می‌شود، . طرح اول A است که طرح وزن‌دهی اسمی نام دارد. در این وزن‌دهی بر اساس اهمیتی که هر درس در تشکیل نمره‌کل خواهدداشت، آزمون‌ساز اقدام به وزن‌دهی می‌کند. در این پژوهش(مشابه وزن‌دهی سازمان سنجش آموزش کشور) به هفت خرده‌آزمون مطابق جدول ۱وزن داده خواهد شد.

جدول ۱. ضریب خرده‌آزمون‌ها در طرح وزن‌دهی A

	خرده‌آزمون‌ها						
	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
وزن اسمی	۴	۲	۳	۲	۴	۳	۲

طرح دوم که در پژوهش حاضر، طرح B نام دارد و در آن از وزن‌های مؤثر برای تشکیل نمره‌کل استفاده‌شده‌است. این وزن‌ها با استفاده از ضرایب اسمی ساخته شده و به واریانس و کواریانس بین خرده‌آزمون‌ها وابسته هستند(کولن، ۲۰۱۴). رابطه (۳) وزن مؤثر نسبی را نشان می‌دهد که براساس واریانس خرده‌آزمون و کواریانس آن با سایر خرده‌آزمون‌ها تعریف شده‌است:

$$ew_i = \frac{w_i^2\sigma_i^2 + w_i\sum_{j\neq i} w_j\sigma_{ij}}{\sum_i\left[w_i^2\sigma_i^2 + w_i\sum_{j\neq i} w_j\sigma_{ij}\right]} \qquad (۳)$$

در رابطه(۳)، σ_i^2 واریانس نمرات خرده‌آزمون i ام، σ_{ij} کواریانس بین نمرات خرده‌آزمون i ام و j ام، همچنین w_i و w_j هم به ترتیب وزن اسمی که در نمرات خرده‌آزمون i ام و j ام ضرب شده‌اند را نشان می‌دهند.

در آزمون‌هایی که به صورت مرکب برگزار می‌شوند و نیاز به یک نمره‌کل برای تصمیم‌گیری دارند، برای ساختن نمره‌کل در هر آزمون مرکب نمره‌مقیاس هر خرده‌آزمون را در وزن مربوط به هر خرده‌آزمون ضرب‌کرده و سپس حاصل باهم جمع می‌شود. نمره حاصل همان نمره‌کل است. هر گاه نمره‌مقیاس شخص i ام برای خرده‌آزمون j ام را با $S_j(X_i)$ و وزن هر خرده‌آزمون را با W_j نمایش دهیم، آنگاه نمره‌کل هر فرد از رابطه(۴) محاسبه خواهد شد. در اینجا m تعداد خرده آزمون ها است.

$$Y_i = \sum_{j=1}^{m} w_j S_j(X_i) \qquad (۴)$$

در جدول ۲ روشهای مختلف ساختن نمره‌کل را که در این پژوهش مورد استفاده قرار خواهند گرفت نمایش داده‌شده‌اند. این جدول ترکیب روش‌های مختلف هموارسازی، ساختن نمره‌مقیاس و همچنین وزن‌دهی را برای ساختن نمره‌کل نمایش می‌دهد. در هر خانه که علامت ستاره استفاده‌شده باشد، به این معنی است که از آن روش خاص برای ساختن نمره‌کل بهره گرفته شده‌است.

جدول ۲. روش‌های مختلف طراحی شده در پژوهش برای ساختن نمره‌کل

نام روش	پیش هموارسازی[۱]	طرح وزن‌دهی		تبدیل به نمرات مقیاس	
		A	B	NT[۲]	AT[۳]
NA	-	❋	-	❋	-
NA_S	❋	❋	-	❋	-
NB	-	-	❋	❋	-
NB_S	❋	-	❋	❋	-
AA	-	❋	-	-	❋
AB	-	-	❋	-	❋

بررسی دقت نمره‌های مقیاس و نمره کل

هم اکنون گزارش‌های فنی بیشتر خطای استاندارد اندازه‌گیری را به صورت خطای استاندارد اندازه‌گیری کلی[۴] ارایه می دهند ولی موسسه های AERA[۵] و NCEME[۶] و APA[۷] از سال ۱۹۸۵ برای استانداردهایی که به منظور تولید آزمونهای آموزشی و روانی توصیه کرده‌اند، گزارش خطای استاندارد اندازه‌گیری شرطی[۸] را به همراه گزارش‌های فنی آزمون همواره توصیه‌کرده‌اند (استاندارد شماره ۲-۱۰ APA). خطای استاندارد اندازه‌گیری شرطی که در این پژوهش با نماد CSEM نمایش داده خواهد شد، قادر است میزان خطای استاندارد اندازه‌گیری را برای همه سطوح نمره‌ها برآوردکند. این شاخص آماری هم برای نمره‌های خام و هم برای نمرات مقیاس و هم برای نمره‌های مرکب قابل محاسبه است(وودروف و دیگران، ۲۰۱۳). بررسی این شاخص آماری نشان می دهد که میزان خطای استاندارد اندازه‌گیری برای همه نمرات برابر نیست، و سطوح مختلف نمره‌ها دارای خطای استاندارد اندازه‌گیری شرطی متفاوتی هستند. در این پژوهش نیز از CSEM برای بررسی دقت نمره‌ها استفاده شده‌است. رویکردهای متفاوتی توسط کولن(۱۹۹۲)، فلت و کوالس[۹](۱۹۹۶) و برنان و لی[۱۰](۱۹۹۹) برای محاسبه CSEM پیشنهاد شده‌است، که به دلیل سهولت انجام محاسبه و عدم نیاز روش برنان و لی(۱۹۹۹) به محاسبه خطای استاندارد

---

[۱] پیش هموارسازی (Pre smoothing): یعنی هموارسازی نمره های خام پیش از تبدیل به نمره های مقیاس.

[2] Normal Transformation

[3] Arcsine Transformation

[4] Overall standard error of measurement

[5] American Educational Research Association

[6] National Council on Measurement in Education

[7] American Psychological Association

[8] Conditional standard error of measurement

[9] Feldt & Qualls

[1] Brenan & Lee [0]

اندازه‌گیری نمره‌های خام از آن برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس در این پژوهش استفاده شده‌است.

روش برنان و لی(۱۹۹۹) برای محاسبه خطای استاندارد اندازه‌گیری شرطی که به روش دوجمله‌ای نیز مشهور است از تعمیم دو روش یعنی نظریه نمره حقیقی قوی لرد(۱۹۶۸) و کولن (۱۹۹۲) استفاده می‌کند و رابطه‌ای را برای خطای استاندارد اندازه‌گیری شرطی ارایه می‌دهد (برنان و لی، ۱۹۹۹). براساس نظریه نمره حقیقی قوی، احتمال شرطی اینکه شخصی از مجموع k سوال در یک آزمون بتواند به y تا از آنها پاسخ صحیح بدهد از رابطه(۵) محاسبه می‌شود:

$$p(y\,|\,\pi,k)=\binom{k}{y}\pi^{y}(1-\pi)^{k-y} \qquad (۵)$$

در رابطه (۶) پارامتر π نمره حقیقی نسبت پاسخ‌های صحیح برای هر شخص است. برنان و لی (۱۹۹۹) برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس طبق نظریه حقیقی قوی از رابطه(۶) استفاده کرده‌اند.

(۶)

$$\sigma_{E}^{2}(s(x)|x)=\frac{k}{k-1}\left(\sum_{y=0}^{k}f(y)^{2}p(y\,|\,\pi,k)-\left(\sum_{y=0}^{k}f(y)p(y\,|\,\pi,k)\right)^{2}\right)$$

تبدیل مقیاس نمره خام y به صورت $f(y)$ است. برای مقایسه روشهای مختلف مقیاس‌سازی از میانگین خطای استاندارد اندازه‌گیری استفاده می شود، این رابطه توسط کولن(۲۰۱۴) ارائه شده و در اینجا معیاری برای بررسی کارآمدی نمره‌های مقیاس و نمره های‌کل خواهد بود.

$$\bar{\hat{\sigma}}^{2}_{E(s(x)|x)}=f(x)\hat{\sigma}^{2}_{E(s(x)|x)} \qquad (۷)$$

در رابطه (۷)، $f(x)$ فراوانی نسبی نمره x است، هر روشی که کمترین میانگین خطای استاندارد اندازه‌گیری شرطی را دارا باشد کارآمدتر خواهد بود. $s(x)$ همان نمره‌های مقیاس و $\hat{\sigma}_{E(s(x)|x)}$ خطای استاندارد اندازه‌گیری شرطی برای هر نمره‌مقیاس است. برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مرکب از رابطه ای که لاری پرایس و همکارانش(۲۰۰۶) پیشنهاد داده اند استفاده می کنیم این رابطه از ترکیب وزنی خطای استاندارد اندازه‌گیری شرطی هر کدام از خرده‌آزمون ها به دست آمده است.

$$SEM_{Y_i}=\sqrt{\sum_{j=1}^{m}w_j^{2}\hat{\sigma}^{2}_{E(s(x_{ij})|x_{ij})}} \qquad (۸)$$

در رابطه(۸)، SEM_{Y_i} خطای استاندارد شرطی برای نمره‌کل شخص i ام است. تعداد خرده‌آزمون‌ها m است. برای محاسبه ضریب پایایی نمره‌های کل از ضریب پایایی نمره‌های مرکب استفاده خواهدشد، رابطه (۹) رابطه ضریب پایایی نمره‌های مرکب است(کولن، ۲۰۰۶):

$$\rho_{cs} = ۱ - \frac{\sum_i w^{۲}_i \sigma^{۲}_i (۱ - \rho_{ii})}{\sum_i \left(w^{۲}_i \sigma^{۲}_i + w_i \sum_{k \neq j} w_k \sigma_{ik} \right)} \quad (۹)$$

یافته ها

پس از انتخاب نمونه ۱۰۰۰۰ نفری از میان شرکت کنندگان، در ابتدا به بررسی نمره های خام پرداختیم. این نمره خام همان تعداد پاسخهای صحیحی بوده که داوطلب آزمون سراسری به هر خرده آزمون داده اند. جدول ۳ شاخص‌های آماری و برخی از ویژگی‌های اندازه‌گیری را برای ۱۰۰۰۰ داده که به صورت تصادفی از میان داوطلبان رشته ریاضی شرکت‌کننده در آزمون سراسری در سال ۱۳۹۵انتخاب شده‌اند را نمایش می‌دهد. این ویژگی‌ها شامل شاخص‌های آماری مانند میانگین حسابی، واریانس، چولگی و کشیدگی و شاخص‌های اندازه‌گیری مانند خطای استاندارد اندازه‌گیری و ضریب پایایی کودر ریچاردسون ۲۰[1]می‌باشد. مقادیر داخل پرانتز میانگین نسبی هستند منظور از میانگین نسبی نمره‌های خام در این جدول حاصل تقسیم میانگین نمره‌های خام به تعداد سوالهای آن آزمون می‌باشد. خطای استاندارد اندازه‌گیری در جدول ۳ به وسیله ضریب پایایی کودر ریچاردسون محاسبه شده است.

جدول ۳. شاخص‌های آماری و شاخص‌های اندازه‌گیری برای نمره‌ی خام خرده‌آزمونها

شاخص	خرده‌آزمونها						
	فارسی	**عربی**	**معارف**	**زبان**	**ریاضی**	**فیزیک**	**شیمی**
میانگین	۷/۶۸ (۰/۳۰)	۵/۲۹ (۰/۲۱)	۹/۰۵ (۰/۳۶)	۵/۹۹ (۰/۲۳)	۴/۷۲ (۰/۰۸)	۵/۱۹ (۰/۱۱)	۳/۰۳ (۰/۰۸)
واریانس	۱۸/۴۳	۲۰/۳۷	۳۱/۸۴	۴۰/۰۰	۴۱/۷۸	۴۶/۵۰	۱۵/۸۵
چولگی	۰/۳۷	۱/۲۰	۰/۴۱	۱/۰۳	۲/۴۳	۱/۹۷	۲/۰۱
کشیدگی	-۰/۱۲	۱/۶۳	-۰/۶۹	۰/۱۳	۷/۶۳	۴/۲۹	۵/۲۶
SEM	۲/۰۵۳	۱/۸۴۳	۲/۲۴۱	۲/۰۵۴	۲/۰۲۴	۲/۰۹۵	۱/۶۱۲
KR20	۰/۷۷۱	۰/۸۳۳	۰/۸۴۲	۰/۸۹۴	۰/۹۰۱	۰/۹۰۵	۰/۸۳۶

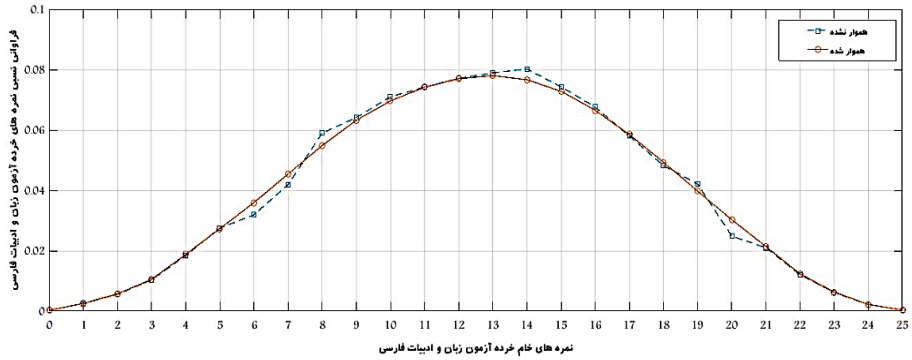
1. KR20

بررسی نمودارهای فراوانی نمره‌های خام نشان دهنده‌ی بی نظمی و آشفتگی در توزیع نمره‌ها است، برای کاهش بی‌نظمی‌ها می‌توان از روش هموارسازی دوجمله‌ای کرنل استفاده کرد. جدول۴ ضمن ارائه گشتاورهای اول تا چهارم برای نمره‌های پیش‌هموارشده برای همه هفت خرده‌آزمون ، مقادیری از h را که به ازاء آنها توزیع نمره‌ها، هموارترین حالت خود را نشان می‌دهد نمایش داده‌است. مقدار h مطلوب برای حالتی محاسبه شده که مقدار رابطه (۲) را کمینه کند.

جدول ۴. مقادیر چهار گشتاور اول برای نمره‌های خام پیش‌هموارشده

خرده‌آزمون	گشتاورهای اول تا چهارم				پارامتر پیش هموارسازی	
	میانگین	واریانس	چولگی	کشیدگی	h[1]	M_h
فارسی	(۰/۳۰)۷/۶۶	۱۸/۸۳	۰/۳۸	-۰/۱۱	۲	-۰/۰۶۵۶
عربی	(۰/۲۰)۵/۲۳	۲۰/۵۱	۱/۲۳	۱/۶۹	۲	-۰/۰۸۰۹
معارف	(۰/۳۶)۹/۰۷	۳۲/۰۲	۰/۴۰	-۰/۶۹	۲	-۰/۰۴۹۹
زبان خارجه	(۰/۲۳)۵/۹۲	۳۹/۳۴	۱/۰۶	۰/۱۸	۴	-۰/۰۹۸۲
ریاضی	(۰/۰۸)۴/۷۵	۴۱/۵۳	۲/۴۲	۷/۵۸	۶	-۰/۱۱۹
فیزیک	(۰/۱۱)۵/۱۰	۴۵/۶۳	۲/۰۱	۴/۳۹	۲	-۰/۱۱۶
شیمی	(۰/۰۸)۳/۰۷	۱۵/۷۲	۱/۹۹	۵/۲۱	۴	-۰/۱۷۰

برای مقایسه بهتر فراوانی‌های هموارشده با فراوانی‌های هموارنشده، فراوانی نسبی نمره‌های خام در شکل ۱ نمایش داده‌شده‌است. نمودار شکل ۱برای h=۲ که توانسته مطلوب ترین و هموارترین فراوانی را برای نمره‌های خرده‌آزمون زبان‌وادبیات‌فارسی نمایش دهد رسم شده است. خرده‌آزمون



ادبیات‌فارسی داری ۲۵ سوال است .

شکل ۱. نمودار فراوانی نسبی نمره‌های خرده‌آزمون زبان‌وادبیات‌فارسی

[1] که توانسته مقدار رابطه ۲ را به کمترین مقدار برساند.h آن مقدار از

در این مطالعه شش روش برای ساختن نمره‌کل طراحی شده‌است، تنوع این روش‌ها بـه خـاطر ترکیب روش‌های متفاوت مقیاس‌سازی، هموارسازی، و طرح‌های وزن‌دهی است. برای سـاختن ایـن نمره‌کل‌ها از دو روش مقیاس‌سازی، یک روش هموارسازی و دو طرح وزن‌دهی استفاده‌شـد. جـدول ۵ برخی شاخص‌های آماری به همراه ضریب پایایی هر روش نمره‌کل‌سازی را نمایش می‌دهد.

جدول ۵ : برخی شاخص‌های آماری به همراه ضریب پایایی در روش‌های نمره‌کل‌سازی

نام روش	گشتاورهای اول تا چهارم				ضریب پایایی[1]
	میانگین	واریانس	چولگی	کشیدگی	
NA	۵۰۳۴/۶	۹۳۳۶۹۰	۰/۶۲۰	۳/۰۹۰	۰/۹۷۴
NA_S	۵۰۲۴/۴	۹۳۴۹۴۰	۰/۶۲۳	۳/۰۹۵	۰/۹۷۷
NB	۵۰۳۹/۷	۹۶۶۶۶۰	۰/۶۴۰	۳/۰۴۴	۰/۹۶۵
NB_S	۵۰۲۸/۷	۹۷۲۹۷۰	۰/۶۴۴	۳/۰۴۸	۰/۹۵۴
AA	۲۰۰۸/۴	۱۴۳۴۴۰	۱/۰۶۲	۴/۳۶۰	۰/۹۷۵
AB	۱۷۲۵/۵	۱۴۷۸۱۰	۱/۱۶۴	۴/۵۷۷	۰/۹۸۱

جدول ۶ مقدار وزن‌ها را برای هر خرده‌آزمون و برای هر طرح وزن‌دهی نمـایش داده‌انـد. توجـه کنید که مقادیر طرح‌های وزن‌دهی A به دلیلی عدم وابستگی وزن‌هـا بـه ویژگـی نمره‌هـا و ثابـت بودن ، در این جدول ذکر نشده‌اند(جدول ۱ را ببینید).

جدول ۶ : مقادیر وزن‌ها برای هر خرده‌آزمون و در هر روش نمره‌کل‌سازی

روش	خرده‌آزمون‌ها						
	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
NB	۰/۰۲۶۴	۰/۰۲۳۴	۰/۰۵۲۸	۰/۰۴۳۲	۰/۳۴۰	۰/۲۹۳	۰/۲۲۰
NB1	۰/۰۲۵۷	۰/۰۲۳۱	۰/۰۵۲۵	۰/۰۴۳۲	۰/۳۴۱	۰/۲۹۳	۰/۲۲۱
AB	۰/۰۲۱۷	۰/۰۲۱۸	۰/۰۵۱۷	۰/۰۴۶۹	۰/۳۳۹	۰/۳۰۰	۰/۲۱۸

با ضرب هر کدام از وزن‌ها در خرده‌آزمون مربوط و جمع کردن آنها برای هر فرد یـک نمره‌کـل ساخته می‌شود، که طبق آنچه که گفته شد برای هر فرد شش نمره‌کـل متفـاوت سـاخته می‌شـود، جدول ۵ گشتاورهای اول تا چهارم نمره‌های کل را به همراه ضریب پایایی آنها نمایش می‌دهنـد. بـا توجه به داده‌های این جدول می‌توان روش‌های نمره‌کل‌سازی را با توجه به نوع مقیاس به کار رفتـه در آنها به دو دسته تقسیم کرد، دسـته اول آن روش‌هـایی هسـتند کـه از روش نرمال‌سـازی بـرای ساختن مقیاس استفاده کردند(بدون توجه بـه طـرح وزن‌دهـی و روش هموارسـازی) و دسـته دوم

[1] ضریب پایایی نمره های کل (کولن، ۲۰۰۶) که در رابطه ۹ معرفی شده است.

روش‌هایی که از آرک‌سینوس برای ساختن مقیاس استفاده کرده‌اند(بدون توجه به طرح وزن‌دهی). در روش‌های نمره‌کل‌سازی که در آنها از مقیاس نرمال استفاده‌شده، میانگین نمره‌های کل بین ۵۰۲۴/۴ تا ۵۰۳۹/۷ تغییر پیدا کرده‌اند. تنها طرح‌هایی که داری طرح وزن‌دهی B بوده اند واریانس های بزرگ‌تری را ایجاد کرده‌اند و واریانس نمره‌ها برای سایر طرح‌های وزن‌دهی به طور تقریبی تفاوت زیادی نداشته است. مقادیر پایایی در جدول های ۵ نشان می‌دهند که هیچ کدام از روش‌های نمره‌کل‌سازی نتوانسته‌اند تفاوت آشکاری را در ضریب پایایی ایجاد کنند به طوری که در روش‌های شامل مقیاس نرمال تفاوت ضریب پایایی بین بالاترین مقدار و کمترین مقدار ۰/۰۲۳ است. در روشهایی که در آنها از مقیاس آرک‌سینوس برای نمره‌کل‌سازی استفاده‌شده مشابه دسته اول روش‌های شامل طرح وزن‌دهی B واریانس بیشتری در میان نمره‌ها ایجاد کرده‌اند، اما بازهم انواع روش‌های نمره‌کل‌سازی که شامل مقیاس آرک‌سینوس بوده اند نیز نتوانستند تفاوت آشکاری در مقدار ضریب پایایی ایجاد کنند. به طوری که تفاوت کمترین مقدار پایایی در روش‌های شامل مقیاس آرک‌سینوس با بیشترین مقدار پایایی ۰/۰۰۶ است. نکته ای که باید در اینجا به آن اشاره کرد تفاوت ضریب پایایی بین روش‌هایی است که شامل مقیاس نرمال بودند و روش‌هایی که شامل آرک‌سینوس هستند، ضریب پایایی نمره‌کل هایی که با مقیاس آرک‌سینوس ساخته شده‌اند، از مقیاس نرمال کمی بیشتر هستند. میانگین ضریب پایایی برای نمره‌هایی که از مقیاس نرمال استفاده کرده‌اند برابر ۰/۹۶۷ و برای نمره‌هایی که از مقیاس آرک‌سینوس استفاده کرده‌اند برابر ۰/۹۷۸ است. که این نتایج حاکی از برتری هر چند اندک روش‌هایی است که از مقیاس آرک‌سینوس برای ساخت نمره‌کل استفاده می‌کنند.

جدول ۷ میانگین خطای استاندارد اندازه‌گیری شرطی را برای شش روش نمره‌کل‌سازی نمایش داده‌است. برای محاسبه این شاخص ابتدا در هر روش خطای استاندارد اندازه‌گیری شرطی را برای هر نمره‌کل محاسبه و سپس از همه خطاها، میانگین گرفته شده‌است.

جدول ۷. مجذور میانگین خطای استاندارد اندازه‌گیری شرطی روش‌های نمره‌کل‌سازی

نام روش نمره کل سازی	$\overline{SEM_{Y_i}}$
NA	۱۵۰/۸۵
NA_S	۱۴۶/۷۲
NB	۱۹۱/۷۶
NB_S	۱۸۷/۳۶
AA	۷۴/۲۶
AB	۸۹/۰۴

از مقادیر جدول ۷ برای مقایسه کارآمدی روشهای متفاوت نمره کل سازی که در این پژوهش باهم مقایسه شده اند، استفاده می شود. هر چند استفاده از هیچکدام از روشهای نمره کل سازی تفاوت محسوسی را در مقدار ضریب پایایی ایجاد نکرده ولی اگر شاخص مجذور میانگین خطای استاندارد اندازه‌گیری شرطی را برای روش های مختلف نگاه کنیم شاهد تفاوت در روش های مختلف نمره کل سازی خواهیم بود.

بحث و نتیجه گیری

این پژوهش با هدف بررسی ویژگی‌ها و همچنین کارآمدی روش‌های مختلف نمره‌کل‌سازی انجام شده‌است. در این پژوهش با تلفیق روشهای هموارسازی فراوانی نمره‌های خام، روش‌های غیرخطی تبدیل نمره‌های خام به نمره‌های مقیاس و همچنین طرحهای متفاوت وزن‌دهی ، شش روش برای ساختن نمره‌کل ارائه شده‌است. از ۱۰۰۰۰ داده آزمون سراسری ایران برای بررسی کارآمدی روشهای نمره‌کل‌سازی بهره گرفتیم. شاخص اصلی در این پژوهش برای بررسی کارآمدی روشهای نمره‌کل‌سازی مجذور میانگین خطای استاندارد اندازه‌گیری شرطی است. هر روش که در این شاخص مقدار کمتری کسب کند، کارآمد تر خواهد بود.

نتایج نشان دادند که استفاده از وزن‌های موثر هر چند تا حدودی توانسته واریانس نمره‌ها را افزایش دهد و مقدار کمی روی پایایی تاثیر بگذارد ولی نتوانسته بخوبی وزن های اسمی باعث کاهش خطای استاندارد اندازه‌گیری شرطی شود و روش های نرمال سازی که در آنها از وزن های اسمی استفاده می‌شود دارای خطای استاندارد اندازه‌گیری شرطی کمتری هستند. در روشهایی که از وزنهای اسمی استفاده می شود، روش تبدیل مقیاس آنها نرمال سازی است، و با پیش هموارسازی نمره ها همراه است میانگین خطای استاندار اندازه‌گیری شرطی کمتری گزارش شده است. ضمنا در مقایسه روشهای نمره کل سازی، آن دسته از روشهایی که از تبدیل آرک سینوس استفاده کرده اند از روش هایی که از نرمال سازی بهره برده اند میانگین خطای استاندار اندازه‌گیری شرطی کمتری داشته اند. با بررسی یافته ها می‌توان به دو نکته دست یافت، نکته اول اینکه بالا بودن وزن اسمی به طور خودکار موجب بالا رفتن سهم هر خرده‌آزمون و افزایش وزن موثر آن خرده‌آزمون شده‌است و نکته دوم آنکه روش‌های هموارسازی تاثیری بر وزن‌های موثر ندارند و با تغییر نوع هموارسازی تغییر زیادی در اندازه‌ی وزن هر خرده‌آزمون مشاهده نمی‌شود و آنچه باعث تغییر در وزن‌ها شده، یا نوع مقیاس‌سازی ، یا نوع طرح وزن‌دهی و یا مقدار وزن اسمی هر خرده‌آزمون است. این نتیجه بخاطر عدم وابستگی روش های هموارسازی به وزن خرده آزمون ها است. همچنین نتایج نشان می‌دهند که هر کدام از خرده‌آزمون‌ها که دارای واریانس بزرگ‌تری باشند وزن آنها نیز بزرگ‌تر است، این تفاوت در واریانس ها هم به دلیل تعداد سوالها و هم به دلیل

وزن اسمی، به وجود آمده است. خرده‌آزمون‌های ریاضی، فیزیک و شیمی هم به لحاظ تعداد سوالها و هم به لحاظ وزن اسمی از خرده‌آزمون‌های دیگر بالاتر هستند، و واریانس بزرگ‌تری نیز دارند، و در طرح‌های وزن‌دهی وزن بزرگ‌تری را نیز به خود اختصاص داده اند.

آن گونه که نتایج نشان داد، در میان روشهای وزن‌دهی، روش وزن‌دهی اسمی به طور کلی در کاهش خطای استاندارد اندازه‌گیری شرطی از روش وزن‌دهی موثر موفق‌تر عمل کرده به طوری که میانگین خطای روشهایی که از طرح وزن‌دهی A استفاده کرده‌اند به مراتب از روش‌هایی که از طرح وزن‌دهی B استفاده کرده اند کمتر است، هر چند که طرح وزن‌دهی B توانسته است تا حدود زیادی به افزایش واریانس نمره‌های کل کمک‌کند. در میان روش‌های نمره‌کل‌سازی که از طرح وزن‌دهی A و نرمال‌سازی به طور همزمان استفاده کرده اند، هرگاه نمره‌های خام پیش هموارسازی شده اند میانگین خطای استاندارد اندازه‌گیری شرطی کاهش یافته‌است. بیشترین کاهش برای خطای استاندارد اندازه‌گیری شرطی در میان روشهای بررسی شده مربوط به روشی است که در آن ضمن استفاده از طرح A از آرک سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس استفاده شده‌است.

در مقایسه با پژوهش های گذشته، یک نقطه اشتراک و یک نقطه اختلاف می توان در نتایج این پژوهش یافت، نقطه اشتراک نتایج این پژوهش با پژوهش های کین و کیس(۲۰۰۴)، چانگ(۲۰۰۹)، خدایی و همکاران(۱۳۹۱) و پی و مالر (۲۰۰۶) در عدم اختلاف روش های وزن دهی و نمره کل سازی در تاثیرگذاری آنها در میزان ضریب پایایی نمره ها است، به طوری که در این پژوهش و همچنین در پژوهش های گذشته نشان داده شده که انواع روش های ساخت نمره کل نتوانسته تغییر فاحشی در ضریب پایایی نمره ها ایجاد کند و از این جهت این نمره ها با هم تفاوتی نداشتند. اما وجه تمایز نتایج این پژوهش با سایر پژوهش های مشابه صورت گرفته در توجه این روش به میانگین خطای استاندارد اندازه گیری شرطی است، در این پژوهش بین انواع روش نمره کل سازی در میزان این خطا تفاوت های آشکاری مشاهده شده که خود مبنایی برای مقایسه کارآمدی آن روش ها بوده است.

با توجه به نتایج می توان چنین گفت که در صورت استفاده از تبدیل آرک سینوس برای ایجاد نمره های مقیاس می توان به کمترین میانگین خطا دست یافت و روش وزن دهی اسمی در این کاهش خطا از روش وزن دهی موثر، کارآمدتر است. همچنین در روش تبدیل مقیاس نرمال کارآترین روش مربوط به حالتی است که هم نمره ها هموار می شوند و هم از وزن دهی اسمی استفاده شده است. ماهیت کاربردی پژوهش اقتضا می کند که به سازندگان آزمون های مرکب توصیه کنیم که برای رسیدن به کمترین خطای ممکن از یکی از این دو روش استفاده کنند، که البته استفاده از روش تبدیل آرک سینوس به دلیل سهولت در تبدیل و کاهش بیشتر سطح خطا بر روش تبدیل نرمال برتری دارد. ضمن اینکه به آزمون سازان توصیه می شود برای کاهش خطای

استاندارد نمره ها، در جایی که واریانس خرده آزمون ها و کواریانس بین آنها اهمیتی ندارد، روش وزن دهی اسمی را بر روش وزن دهی موثر ترجیح دهند.

برای ساختن نمره‌کل روشهای متفاوت دیگری نیز وجود دارند، به عنوان مثال نمره‌های نرمال سازی شده را می‌توان پس از تبدیل به مقیاس نرمال هموارسازی کرد(پس هموارسازی)[1] ، و یا نمره‌های مقیاس آرک سینوس را می‌توان پس از تبدیل به مقیاس آرک سینوس هموارسازی کرد، ضمن اینکه روش‌های وزن‌دهی همچون: وزن دهی با عکس خطای استاندارد اندازه‌گیری، وزن‌دهی بر اساس پایایی نمره‌ها و وزن‌دهی بر اساس طول آزمون از سایر روش‌های وزن‌دهی می باشد که در این پژوهش به آنها پرداخته نشده‌است.

منابع

خدایی، ابراهیم، ذوالفقار نسب، سلیمان و یادگار زاده، غلام رضا. (۱۳۹۱). وزن‌دهی بهینه به سؤال‌های خرده‌آزمونهای ورودی برای ساخت نمره‌کل ترکیبی. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۳(۴) ، ۱۰۴-۷۹.

مری جی. آلن، وندی ام. ین، **مقدمه ای بر نظریه های اندازه‌گیری**، ترجمه علی دلاور (۱۳۹۰). انتشارات سمت، تهران .

نقی زاده، سیما. (۱۳۹۴). **نمره‌کل‌سازی آزمون سراسری در گروه آزمایشی علوم ریاضی و فنی سال ۱۳۹۱ براساس توزیع واقعی نمرات و مقایسه آن با روش فعلی**. تهران: مرکز تحقیقات ارزشیابی، اعتبار سنجی و تضمین کیفیت آموزش عالی (سازمان سنجش آموزش کشور).

References:

Allen, M. J., & Wendy, Y. M. (1979). **Introduction to Measurement** Theory. California: Cole publishing company.

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. (Reprinted as 'W. A. Angoff, Scales, norms, and equivalent scores'. Princeton, NJ: Educational Testing Service, 1984.).

Brennan, R. L., & Lee, W. C. (1999). Conditional Scale-Score Standard Errors of Measurement under Binomial and Compound Binomial Assumptions. *Educational and Psychological Measurement*, 59(1), 5 – 24.

[1] Post smoothing

Chang, S. W. (2009). Choice of weighting schemes in forming the composites. *buletin of educational psychology*, 40(3), 489-510.

Chang, S. W. (2006). .Methods in Scaling the Basic Competence Test. *Educational and Psychological Measurement* , 66(6), 907-929.

Dorans N. J., Pommerich, M. & Holland P. W. (2007). A Framework and History for Score Linking. In Holland P. W. (Eds.). *Linking and Aligning Scores and Scales* (pp. 5-30). New York: Springer.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.). *Educational measurement* (3rd ed., 105-146). New York, NY: Macmillan.

Feldt, L. S., & Quails, A. L. (1996). Estimation of measurement error variance at specific score levels. *Journal of Educational Measurement*, 33, 141-156. 156.

Gulliksen, H. (1950). *Theory of mental test*. New York: John Wiley & sons.

Iowa Assessment (2016). **Iowa Test Of Basic Skills**. Iowa City: Author Retrieved : www.itp.education.uiowa.edu

Kane, M., & Case, S. M. (2004). The reliability and validity of weighted composite scores. *Applied Measurement in Education*, 17, 221-240.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement of scale scores. *Journal of Educational Measurement*, 29, 285-307.

Kolen, M. J., & Hanson, B. A. (1989). Scaling the ACT Assessment. In R. L. Brennan (Ed.). *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 35-55). Iowa City, IA: American College Testing Program.

Kolen, M.J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, 28, 257-282.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.). *Educational measurement* (4rd ed., 236-241). CT: American Council on Education, and Praeger.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling and Linking, 3rd Ed*. New York: Springer.

Lord, F. M. (1955), Estimating Test Reliability. *ETS Research Bulletin Series*, 1955: i–17.

Lord, F. M. (1965). A strong true-score theory with applications. *Psychometrika*, 30,239-270.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. MA: Adisson-wesley.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika*, 34, 259-299

Pei, L. K., & Maller, S. J. (2006). Monte Carlo simulation study of differential weights on composite reliability and validity. Paper presented at the *annual meeting of the National Council on Measurement in Education*, San Francisco.

Price, R. L., Raju, N., Lurrie, A. Wilkins, C. & Zhu, J. (2006). Conditional standard errors of measurement for composite scores on the Wechsler

Preschool and Primary Scale of Intelligence-Third Edition. *Psychological Reports*,98,237-252

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, 20(1), 16-19.

Sutton, R. (2004). Teaching under high-stakes testing: Dilemmas and decisions of a teacher educator. *Journal of Teacher Education*, 55(5), 463-475.

The ACT, ACT *assessment technical manual* (2014). Iowa City: Author. Retrieved from : http://www.act.org/research/researchers/techmanuals.html.

The SAT, SAT *technical manual* (2016). New York: Author. Retrieved fromcollegereadiness.collegeboard.org.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. Review of Educational Research, 4, 663- 705.

Woodruff, D., Traynor, A., Cui, Z., Fang, Y., (2013). A Comparison of Three Methods for Computing Scale Score Conditional Standard Errors of Measurement. *ACT Research report series*, no.7.