Educational Measurement and Evaluation Studies

Print ISSN: 2476-2865 Online ISSN: 2783-0942

National Organization for Educational Testing

Utilizing the Multi Attribute decision- making method to Rank total score construction methods

Mojtaba Jahanifar[1]

1. Assistant professor, Departement of Educational Sciences, Shahid Chamran University, Ahvaz, Iran., m.jahanifar@scu.ac.ir

Article Info

Article Type:
Research Article

Received
2021/10/10
Received in revised form
2022/02/23
Accepted
2022/03/09
Published online
2022/03/29

ABSTRACT

Objective: The decision of admission in the exams is mostly based on the score obtained in that exam. The exam can be consisting of several sub-exams with different content, which is called a composite exam and the resulting score is called the total score. Different methods of making the total grade cause changes in people's admission decisions. This research aims to rank the methods used to make grades.

Methods: Ten thousand random samples of the national university entrance exam in seven sub-tests have been used to rank the six totalization score methods. The raw score was obtained from the total of correct answers, and also, normalization and arc sine methods were used to convert the scores into scale scores. Nominal, effective and Shannon weighting were used to make the total grade.

In order to rank the total score methods based on the standard error of their conditional measurement, an approach based on multi-indicator decision making was used.

Results: The results showed that those totalization score methods that use the arcsine scale and nominal or Shannon weighting schemes were ranked higher, and if they are used, less errors will be made in the totalization score.

Conclusion: Using the arc sine scale score can contribute to more interpretability and accuracy of composite test scores because of less conversion error and easier use. Different weighting methods do not have much effect on the accuracy of scores and can be chosen and used according to the conditions of the test and the decision of the examiner.

Keywords: scale score, Total score, weighting scheme, CSEM, decision making

Cite this article: Jahanifar, Mojtaba. (2022). Utilizing the Multi Attribute decision- making method to Rank total score construction methods. *Educational Measurement and Evaluation Studies*, 12 (37):101-124 Pages.DOI: 10.22034/EMES.2022.550667.2366

© The Author(s).

Publisher: National Organization of Educational Testing (NOET)

سازمان سنجش آموزش کشور

مطالعات اندازه‌گیری و ارزشیابی آموزشی

شاپا چاپی: ۲۸۶۵-۲۴۷۶ شاپا الکترونیکی: ۰۹۴۲-۲۷۸۳

استفاده از روش تصمیم‌گیری چند شاخصی در رتبه‌بندی روش‌های ساخت نمره کل

مجتبی جهانی فر[۱]

۱. استادیار گروه علوم تربیتی دانشگاه شهید چمران اهواز، اهواز، ایران M.jahanifar@scu.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت
۱۴۰۰/۰۷/۱۸
اصلاح
۱۴۰۰/۱۲/۰۴
پذیرش
۱۴۰۰/۱۲/۱۸
انتشار
۱۴۰۱/۰۱/۰۹

چکیده

هدف: تصمیم پذیرش در آزمون‌ها بیشتر بر اساس نمره‌ای است که در آن آزمون کسب می‌شود. آزمون می‌تواند از چند خرده آزمون با محتوای متفاوت تشکیل شده باشد که به آن آزمون مرکب و نمرهٔ حاصل، نمره‌کل نامیده می‌شود. روش‌های متفاوت نمره‌کل سازی موجب تغییر در تصمیم پذیرش افراد می‌شود. این پژوهش با هدف رتبه‌بندی روش‌هایی که برای ساخت نمره‌کل استفاده می‌شود، انجام شده است.

روش پژوهش: از ده هزار نمونهٔ تصادفی آزمون سراسری در هفت خرده آزمون برای رتبه‌بندی شش روش نمره‌کل سازی بهره گرفته شده است. نمرهٔ خام، از مجموع پاسخ‌های صحیح به دست آمده و از روش‌های نرمال‌سازی و آرک‌سینوس برای تبدیل نمره‌ها به نمره‌های مقیاس بهره برده شده است. از طرح‌های وزن دهی اسمی، مؤثر و شانون برای ساخت نمره‌کل استفاده گردید. به منظور رتبه‌بندی روش‌های نمره‌کل سازی بر اساس خطای استاندارد اندازه‌گیری شرطی آنها از رویکردی مبتنی بر تصمیم‌گیری چند شاخصی استفاده شد.

یافته‌ها: نتایج نشان داد که آن دسته از روش‌های نمره‌کل سازی که از مقیاس آرک‌سینوس و از طرح‌های وزن دهی اسمی و یا شانون بهره می‌برند، حائز رتبه‌های بالاتری شدند و در صورت استفاده از آنها در نمره‌کل سازی، خطای کمتری مرتکب خواهیم شد.

نتیجه‌گیری: استفاده از نمرهٔ مقیاس آرک‌سینوس، به دلیل خطای کمتر، تبدیل و راحت‌تر می‌تواند به تفسیرپذیری و دقت بیشتر نمره‌های آزمون‌های مرکب کمک کند. روش‌های متفاوت وزن دهی تأثیر چندانی بر دقت نمره‌ها ندارد و مطابق با شرایط آزمون و تصمیم آزمون‌ساز می‌تواند انتخاب و استفاده شود.

واژه‌های کلیدی: تصمیم‌گیری چند شاخصی، خطای استاندارد اندازه‌گیری شرطی، طرح وزن دهی، نمره‌کل، نمرهٔ مقیاس

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

استناد: جهانی فر، مجتبی (۱۴۰۱). استفاده از روش تصمیم‌گیری چندشاخصی در رتبه‌بندی روش‌های ساخت نمره‌کل. مطالعات اندازه‌گیری و ارزشیابی آموزشی، ۱۲ (شماره۳۷)، صفحه۱۰۱-۱۲۴. DOI: 10.22034/EMES.2022.550667.2366
ناشر: سازمان سنجش آموزش کشور حق‌مؤلف © نویسندگان.

مقدمه

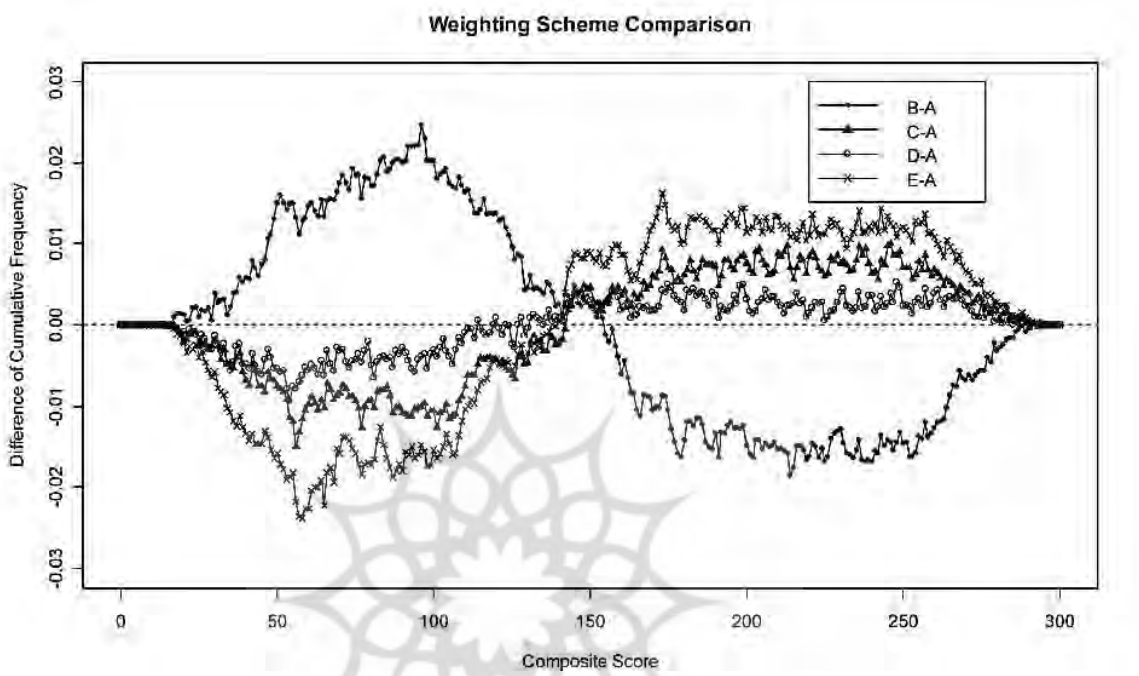
عاملی که اهمیت و حساسیت آزمون‌های سرنوشت ساز را بالا می‌برد خود آزمون نیست، بلکه تصمیم‌هایی است که برای رسیدن به آنها اجرا می‌شود. اینکه آزمون به صورت چند گزینه‌ای باشد یا شفاهی چندان مهم نیست، مهم نوع تصمیمی است که قرار است بر اساس نتایج آزمون گرفته شود. به طور کلی در آزمون‌های سرنوشت ساز به دو مورد مهم اشاره شده است، یکی عواقب آزمون و دیگری پایه‌ای بودن آن برای تصمیم‌گیری (ساتون ، ۲۰۰۴). این آزمون‌ها در برخی موارد تنها به یک موضوع مشخص اختصاص دارند و گاهی به صورت ترکیبی از موضوع‌های مختلف هستند. به عنوان مثال آزمون ورودی دانشگاه‌ها به طور معمول مشتمل بر چند خرده آزمون می‌شود که به آن آزمون‌های مرکب[1] می‌گویند. نمره کل[2] به طور معمول ترکیب خطی از نمرهٔ خام یا نمره‌های مقیاسی بخش‌های مختلف آزمون مرکب است. سهم هر بخش در آزمون مرکب با وزن مشخص می‌شود. این وزن‌ها می‌توانند شامل وزن‌های اسمی و وزن‌های مؤثر باشد. وزن‌های اسمی را سازندگان آزمون بر اساس هدف‌ها و محتوای آزمون تعیین می‌کنند ولی وزن‌های مؤثر سهم آماری هر کدام از بخش‌ها را در واریانس آزمون بر عهده دارند. برای استخراج نمرهٔ کل از نمره‌های خام آزمون، ابتدا نمره‌های خام را به نمره‌های مقیاس[3] تبدیل می‌کنند و پس از اعمال وزن‌های مورد نظر مربوط به هر خرده‌آزمون و ترکیب نمره‌ها، نمره کل ساخته می‌شود. روش‌های متفاوتی برای تبدیل نمره‌های خام به نمره‌های مقیاس وجود دارد و همچنین از طرح‌های وزن‌دهی مختلفی می‌توان برای ساختن نمره کل استفاده کرد. در دنیا و همچنین در ایران آزمون‌های مرکب فراوانی تولید و استفاده می‌شود. به عنوان مثال آزمون‌های "SAT" و "ACT" در ایالات متحده آمریکا و آزمون سراسری ورود به دانشگاه‌های دولتی در ایران نمونه‌ای از آزمون‌های مرکب هستند. هرگاه فرآیندهای ساخت آزمون، نرمال‌سازی و مقیاس‌سازی برای هر کدام از بخش‌های آزمون مرکب به طور مشابه صورت بگیرد، مقایسهٔ نمره‌های آزمون شوندگان در همهٔ بخش‌های آزمون و همچنین در نمره کل آسان‌تر خواهد بود، ضمن اینکه نمره‌های مرکب معنادارتری تولید خواهد شد (کولن و برنان ، ۲۰۱۴). در آزمون‌های مرکب به دلیل تفاوتی که خرده آزمون‌ها در موضوع و تعداد سؤال‌ها دارند، نمره‌های این خرده آزمون‌ها به مقیاس مشترکی برده می‌شود تا تفسیرپذیری بهتری داشته و ترکیب نمره‌ها برای تولید نمره کل امکان پذیر باشد. به طور مثال آزمون»ACT» از چهار خرده آزمون زبان و ادبیات انگلیسی، ریاضیات، مهارت‌های خواندن و علوم تشکیل شده که به صورت چند گزینه‌ای طراحی شده‌اند. مهارت نوشتن هم به عنوان خرده آزمونی اختیاری و به صورت انشائی در این آزمون گنجانده شده است. در این آزمون نمرهٔ خام هر یک از خرده آزمون‌ها به طور مستقیم به مقیاس نمره‌ای بین ۱ تا ۳۶ تبدیل می‌شود (راهنمای فنی آی سی تی ، ۲۰۱۴). آزمون»SAT» یکی از آزمون‌هایی است که برای تبدیل نمره‌های خام به نمره‌های مقیاس از گروه مرجع بهره می‌برد. این آزمون شامل خرده آزمون‌های مهارت‌های خواندن، نوشتن و همچنین ریاضیات است. برای مقیاس سازی در این آزمون، از تبدیل نرمال[4] استفاده می‌شود. در این تبدیل نمره‌ها برحسب فراوانی

1. Battery Test
2. Composite Score
3. Scale Score
4. .Normal Transformation

تراکمی، به مقیاس نرمال برده می‌شوند و حاصل این مقیاس‌سازی جدول تبدیلی است که در آن هر نمره بر اساس رتبه و نمرهٔ درصدی به مقیاس آزمون»SAT" برده می‌شود. برای محاسبهٔ نمره کل، نمرهٔ خرده آزمون‌ها با هم جمع می‌شود (گـــزارش فنی کالج بـــرد ، ۲۰۱۶). برای ترکیب نمره‌های مقیاس و تولید نمره کل از ویژگی‌های مختلف خرده آزمون‌ها استفاده می‌شود. دشواری سؤال‌ها، واریانس نمره‌های مقیاس، خطای استاندارد اندازه‌گیری نمره‌ها، ضریب پایایی نمره‌ها و همبستگی بین این خرده آزمون‌ها از مهم‌ترین این ویژگی‌ها هستند (ونگ و استانلی، ۱۹۷۰). گالیکسن در فصل بیستم کتاب نظریهٔ آزمون‌های روانی که در سال ۱۹۵۰ منتشر شد دربارهٔ روش‌های مختلف وزن‌دهی به طور گسترده‌ای بحث کرده است. هدف گالیکسن (۱۹۵۰) از ترکیب نمره‌ها ایجاد نمره کل با ضریب پایایی بالا بوده است. گالیکسن نشان داد که اگر تعداد زیادی آزمون که همبستگی بالایی با هم دارند، ترکیب شوند، طرح‌های مختلف وزن دهی، تغییر چشم‌گیری بر پایایی نمره‌کل نخواهد داشت و در صورتی که همبستگی بین این خرده آزمون‌ها کم باشد، تأثیر ترکیب آنها با روش‌های وزن‌دهی مشهودتر خواهد بود. پی و مالر (۲۰۰۶) در پژوهشی شبیه‌سازی شده، نتایج گالیکسن را تأیید کردند و نشان دادند که برخی عوامل مانند تعداد خرده آزمون‌ها و خواص روان‌سنجی آنها بر روایی و پایایی آزمون مرکب تأثیر خواهد داشت. چانگ (۲۰۰۹) با بررسی پنج طرح وزن‌دهی مختلف برای تولید نمره‌کل نشان داد که اگر شاخص‌هایی مانند ضریب پایایی را به عنوان شـــاخص‌های دقت نمره‌کل در نظر بگیریم، طرح‌های مختلف وزن‌دهی تأثیر قابل ملاحظه‌ای بر آن ندارد و مقدار ضریب پایایی برای همهٔ طرح‌های وزن‌دهی مقدار بالا و قابل قبولی خواهد بود. اما اگر به لحاظ سهم هر خرده آزمون در تولید نمره‌کل به عنوان شاخص بنگریم، طرح‌های وزن‌دهی مختلف باعث ایجاد سهم‌های مختلف برای خرده آزمون‌ها در تولید نمره‌کل خواهد بود. شون ون چانگ (۲۰۰۶) سه روش تبدیل مقیاس خطی، نرمال‌سازی و تبدیل آرک سینوس[1] را با هم مقایسه کرده است. در این مقایسه، سه روش مقیاس‌سازی در شاخص‌هایی مانند ضریب پایایی نمره‌های مقیاس بندی شده، نمودار خطای استاندارد شرطی و تعداد تغییر نمره‌ها در اثر هرس کردن [2] و همچنین گاف‌های (شکاف هایی که بین نمرات مقیاس ایجاد می شود) ایجاد شده در مقیاس با هم مقایسه شده‌اند. در نتیجهٔ این گزارش چنین آمده است که هر روش معایب و مزایای مربوط به خودش را دارد و هیچ روشی همهٔ ویژگی‌های مطلوب را دارا نیست و تصمیم برای انتخاب مقیاس مناسب، هم به خواص اندازه‌گیری و هم به سهولت تفسیر بستگی خواهد داشت (چانگ، ۲۰۰۶). در پژوهش چانگ ضریب پایایی خرده آزمون‌ها پس از انجام سه نوع تبدیل، به هم نزدیک بودند. در حالی که در ایجاد گاف‌های بین نمره‌ها، روش تبدیل آرک‌سینوس از دو روش دیگر پیشی گرفته است، ولی نمودار خطای استاندارد شرطی این روش نسبت به دو روش دیگر دارای خطای کمتری است و روش نرمال‌سازی بیشترین شباهت را به توزیع نمره‌های خام نشان داد به طوری که تبدیل خطی نمره‌ها کمترین گاف را در مقیاس نمره‌ها داشت. تا اینجا بررسی پیشینه نشان داد که اگر روش‌های نمره‌کل‌سازی را از بابت ویژگی‌های روانسنجی، خطا و پایایی با هم مقایسه کنیم، روش ترجیحی یافت نمی‌شود و همهٔ روش‌ها در نوع خود کاربردی‌اند و ویژگی‌های مثبت و منفی متفاوتی دارند که موجب پذیرش همگانی و یا کنار گذاشتن آن

1. Arcsine Transformation
2.

روش نمی‌شود و آزمون‌سازها با توجه به ویژگی‌های آزمون و اهداف آن، از روش‌های متفاوت نمره کل سازی بهره برده‌اند. اما نکتهٔ قابل توجه این است که روش‌های نمره کل‌سازی می‌تواند بر تصمیم پذیرش افراد تاثیر بگذارد. این نتیجه را چانگ (۲۰۰۹) با بررسی نمودار فراوانی تراکمی نمره کل‌های حاصل از پنج روش متفاوت، نشان داد (شکل ۱).



شکل ۱. تفاوت فراوانی تراکمی روش‌های نمره کل سازی (چانگ، ۲۰۰۹)

در نمودار شکل (۱)، محور افقی نشان‌دهندهٔ نمره کل و محور عمودی نمایش‌دهندهٔ تفاوت بین فراوانی‌تراکمی روش‌های نمره کل‌سازی است. روش‌های نمره کل‌سازی از A تا E نام‌گذاری شده‌اند و تفاوت بین فراوانی‌تراکمی هر کدام به صورت B-A، C-A و سایر موارد نمایش داده شده است. فراوانی‌تراکمی هر نمره، نسبت افرادی را نشان می‌دهد که نمره‌ای کوچک‌تر یا مساوی با نمرهٔ مورد نظر کسب کرده‌اند. تفاوت فراوانی تراکمی هر نمره در روش‌های مختلف به این معنی است که افراد در روش‌های متفاوت رتبه‌های متفاوتی کسب کرده‌اند. آن گونه که می‌بینید، برای نقاط میانی و نمرهٔ میانگین، روش‌های متفاوت نتوانسته‌اند تفاوت زیادی بین فراوانی تراکمی نمرهٔ افراد ایجاد کنند اما در نقاطی به جز نقاط وسط، روش‌های متفاوت نمره کل‌سازی، فراوانی تراکمی متفاوتی از خود نشان داده‌اند. تفاوت در فراوانی تراکمی هر روش با روش دیگر، به این معنی است که تعداد افرادی که در زیر نمرهٔ خاصی قرار می‌گیرند، در روش‌های متفاوت نمره کل‌سازی با هم فرق دارند. این تفاوت باعث ایجاد گوناگونی در تصمیم پذیرش افراد در روش‌های مختلف نمره کل‌سازی می‌شود (چانگ، ۲۰۰۹). روش‌های متفاوت ساخت مقیاس نیز در تصمیم‌پذیرش افراد موثر است (چانگ، ۲۰۰۶). این گفته بدان معنی است که هرگاه برای ساخت نمرهٔ مقیاس و نمره کل از روش‌های متفاوت تبدیل استفاده شود، تصمیم برای پذیرش و یا

عدم‌پذیرش افراد نیز تغییر پیدا می‌کند. طبق آنچه که تاکنون به آن اشاره شد، روش‌های متفاوتی را می‌توان برای ساختن نمره‌کل پیشنهاد داد که هر کدام می‌توانند ترکیبی از روش‌های ساخت نمره‌های مقیاسی و روش‌های وزن‌دهی باشند، اما مسئله اینجاست که با تغییر روش ساخت نمره‌کل، نمرهٔ اختصاص داده شده به هر شرکت‌کننده و به دنبال آن تصمیم برای پذیرش و یا عدم پذیرش وی در آزمون دچار تغییر می‌شود. مسئلهٔ اصلی این پژوهش آن است که کدام روش نمره‌کل‌سازی می‌تواند خطای کمتری داشته باشد و اینکه چگونه می‌توان این روش‌ها را از نظر خطای اندازه‌گیری رتبه‌بندی کرد؟ هدف عمدهٔ این پژوهش ارائه روشی است تا بتوان روش‌های متفاوت نمره‌کل سازی را برحسب اینکه کمترین خطا یا بیشترین خطا را برای سطوح توانایی مختلف مرتکب می‌شوند، رتبه‌بندی کرد.

روش پژوهش

در این بخش ضمن بیان روش نمونه‌گیری و ساختار داده‌ها، روش‌هایی همچون مقیاس‌سازی، مراحل مختلف ساخت نمره‌کل، بررسی دقت نمره‌ها و نحوهٔ استفاده از روش تصمیم‌گیری چند شاخصی به طور کامل شرح داده شده‌اند.

روش نمونه‌گیری و ابزار پژوهش

جامعهٔ مورد نظر در این پژوهش، داوطلبان شرکت کننده در آزمون سراسری سال ۱۳۹۵ در گروه آزمایشی ریاضی و فنی هستند. طبق گزارش روابط عمومی سازمان سنجش آموزش کشور در سال ۱۳۹۵ تعداد ۱۶۲۸۷۹ نفر در آزمون سراسری در گروه آزمایشی ریاضی و فنی شرکت کرده‌اند. به منظور بررسی پاسخ‌ها و خرده آزمون‌ها، نمونه‌ای از این داوطلبان به روش نمونه‌گیری تصادفی انتخاب شدند. با توجه به اینکه در این پژوهش با مقیاس بزرگ[1] سر و کار داریم طبق قاعده سر انگشتی[2] می‌توان حجم نمونه نزدیک به ده هزار نفر را مناسب دانست. در این پژوهش روش‌های مختلف نمره‌کل سازی و مقیاس سازی و سایر تحلیل‌ها برای خرده آزمون‌های مختلف عمومی و اختصاصی آزمون سراسری ایران در گروه آزمایشی رشته ریاضی و فنی اجرا شد. در آزمون سراسری ۱۳۹۵ و در گروه آزمایشی ریاضی و فنی، داده‌های چهار درس عمومی زبان و ادبیات فارسی (۲۵ سؤال)، زبان و ادبیات عربی (۲۵ سؤال)، معارف اسلامی (۲۵ سؤال) و زبان انگلیسی(۲۵ سؤال) و سه درس اختصاصی ریاضیات (۵۵ سؤال) ، فیزیک (۴۵ سؤال) و شیمی(۳۵ سؤال) استفاده شد.

روش ساختن نمره‌های خام و نمره‌های مقیاسی

برای ساخت نمره‌های خام، به پاسخ درست هر سؤال نمرهٔ یک و به پاسخ‌های نادرست و سفید نمرهٔ صفر تعلق گرفت و بابت پاسخ نادرست جریمه‌ای تعلق نگرفت، یعنی آزمون نمره منفی نداشت. نمرهٔ خام هر فرد در هر خرده آزمون از رابطه (۱) محاسبه می‌شود:

$$X_i = \sum_{j=}^{k} u_j \qquad (۱)$$

در رابطه (۱)، k تعداد سؤال‌ها در هر خرده آزمون است و u_j نمرهٔ هر سؤال است که می‌تواند یکی از مقادیر صفر یا

1. Large-scale assessment
2. Thumbnail rule

یک را بپذیرد. X_i نمرهٔ خام شخص iام در هر خرده آزمون است. در این پژوهش برای تبدیل نمره‌های خام به نمره‌های مقیاسی، از روش‌های نرمال‌سازی و روش تبدیل آرک سینوس استفاده شده است.

روش نرمال سازی: برای به دست آوردن نمرهٔ مقیاسی به روش نرمال سازی مراحل تبدیل به شرح زیر است (کولن، ۲۰۱۴):

مرحلهٔ اول: توزیع فراوانی نسبی نمره‌ها محاسبه می‌شود. مرحلهٔ دوم: با استفاده از فراوانی نسبی و توزیع تراکمی نمره‌ها، رتبهٔ درصدی نمره‌ها محاسبه می‌شود. مرحلهٔ سوم: نمرهٔ z مربوط به هر رتبهٔ درصدی از روی معکوس رابطه ۲ محاسبه می‌شود:

$$\Phi(z) = \frac{\hat{Q}(y)}{۱۰۰} = \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{z} e^{-\omega^{۲}/۲} d\omega \qquad (۲)$$

رابطهٔ (۲) توزیع تراکمی نرمال استاندارد است. ω متغیر انتگرال‌گیری است که دامنهٔ آن از $-\infty$ تا z است و $\hat{Q}(y)$ رتبهٔ درصدی است. مرحلهٔ چهارم: تبدیل خطی نمرات z به دست آمده در مرحلهٔ سوم با استفاده از رابطه (۳) است:

$$sc(y) = \sigma(\mathrm{sc})\, z + \mu(\mathrm{sc}) \qquad (۳)$$

Sc نمرات مقیاس‌بندی شده هستند. در این پژوهش میانگین این تبدیل خطی ۵۰۰۰ و انحراف استاندارد آن ۱۲۵۰ است. این تبدیل دامنه نمره‌های مقیاس را در فاصله ۰ تا ۱۰۰۰۰ محدود خواهد کرد. مرحلهٔ پنجم: تبدیل نمره‌های به دست آمده در مرحلهٔ چهارم به نمره‌های صحیح است که با $sc_{\mathrm{int}}(y)$ نشان داده می‌شود. در این روش عدد به دست آمده به نزدیک‌ترین نمرهٔ صحیح گرد می‌شود.

روش تبدیل آرک سینوس: در تبدیل نمره‌های خام به نمره‌های مقیاس‌بندی شده با استفاده از تبدیل آرک سینوس، از تعداد پاسخ‌های درست به هر خرده آزمون و تعداد سؤال‌های خرده آزمون برای ساختن نمره مقیاس استفاده می‌شود.

$$S(X_i) = \frac{۱}{۲}\left\{ \sin^{-۱}\sqrt{\frac{X_i}{k+۱}} + \sin^{-۱}\sqrt{\frac{X_i+۱}{k+۱}} \right\} \qquad (۴)$$

در رابطهٔ (۴)، X_i نمرهٔ خام (تعداد پاسخ‌های صحیح) و k تعداد سؤال‌های خرده آزمون هستند (کولن، ۲۰۱۴). در این پژوهش به منظور محدود شدن دامنهٔ نمره‌های مقیاس آرک سینوس بین ۰ تا ۱۰۰۰۰ از تبدیل خطی $Sc(y) = aS(X_i) + b$ استفاده شده است. در این رابطه $a = ۶۶۹۰$ و $b = -۴۶۷$ هستند، این تبدیل دامنه، نمره‌های مقیاس را در فاصله ۰ تا ۱۰۰۰۰ محدود خواهد کرد تا با نمره‌های مقیاس نرمال در محدودهٔ مشترکی باشند. در پایان، نمره‌های مقیاسی ساخته شده به نزدیک‌ترین عدد صحیح گرد خواهد شد.

طرح‌های وزن‌دهی و ساخت نمرهٔ کل

از سه طرح وزن‌دهی در این پژوهش استفاده شده است، که عبارتند از: الف) طرح وزن‌دهی اسمی که بر اساس اهمیت هر

درس در ســـاخت نمره کل به دست می‌آید، ب) طرح وزن‌دهی مؤثر بر اساس واریانس و کواریانس هر آزمون و ج) طرح وزن‌دهی شانون. در ادامه به چگونگی محاسبهٔ این وزن‌ها و استفاده آنها در این پژوهش پرداخته شده است.

طرح وزن‌دهی اسمی: در این طرح ، که با نشان A مشخص شده، آزمون‌ساز بر اساس اهمیتی که هر درس در تشکیل نمره‌کل خواهد داشـــت، اقدام به وزن‌دهی می‌کند. در این پژوهش به هفت خرده آزمون مطابق جدول (۱) وزن داده شده است.

جدول ۱. ضریب خرده آزمون‌ها در طرح وزن‌دهی A

آزمون	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
وزن اسمی	٤	۲	۳	۲	٤	۳	۲

طرح وزن دهی مؤثر: در این طرح، که در پژوهش حاضر B نام دارد، از وزن‌های مؤثر برای تشکیل نمره‌کل استفاده شده است. این وزن‌ها با استفاده از ضرایب اسمی ساخته شده و به واریانس و کواریانس بین خرده آزمون‌ها وابسته هستند. رابطهٔ (۵) وزن مؤثر نسبی را نشان می‌دهد که بر اساس واریانس خرده آزمون و کواریانس آن با سایر خرده آزمون‌ها تعریف شده است:

$$ew_i = \frac{w_i^{۲}\sigma_i^{۲} + w_i \sum_{j \neq i} w_j \sigma_{ij}}{\sum_i \left[w_i^{۲}\sigma_i^{۲} + w_i \sum_{j \neq i} w_j \sigma_{ij} \right]} \qquad (۵)$$

در رابطهٔ (۵)، $\sigma_i^{۲}$ واریانس نمرات خرده آزمون iام، σ_{ij} کواریانس بین نمرات خرده آزمون i ام و j ام، و w_i و w_j هم به ترتیب وزن اسمی که در نمرات خرده آزمون i ام و j ام ضرب می‌شود (کولن، ۲۰۱۴).

طرح وزن دهی بر اساس آنتروپی: روش آنتروپی شانون[1]، که در این پژوهش با نشان C نامگذاری شده است، بر اساس نظریه بی‌نظمی شانون در علم اطلاعات[2] طرح‌ریزی شده است. آنتروپی مفهومی بسیار با اهمیت در علوم اجتماعی، فیزیکی و نیز در نظریهٔ اطلاعات است. آنتروپی در نظریهٔ اطلاعات معیار عدم اطمینان در خصوص یک پیشامد یا متغیر است که با توزیع احتمال آن مشخص می‌شود. اندازه‌گیری این عدم اطمینان توسط شانون به صورت زیر بیان شده است (آذر و رجب‌زاده، ۱۳۹۳).

$$E = S(P_۱, P_۲, ..., P_n) = -K \sum_{i=۱}^{n} P_i \ln P_i \qquad (۶)$$

1. Shannon Entropy
2. Information theory

در رابطهٔ (۶)، K مقداری ثابت است. این رابطه به رابطهٔ آنتروپی توزیع احتمال P مشهور است. در این پژوهش از مفهوم آنتروپی برای محاسبهٔ بی‌نظمی، از توزیع احتمال نمره‌های مقیاسی در هر خرده آزمون استفاده می‌شود. نحوهٔ محاسبهٔ وزن‌های آنتروپی در این پژوهش به شرح زیر است:

مرحلهٔ اول: ابتدا ماتریس تصمیم، تشکیل داده می‌شود. ستون‌های این ماتریس خرده آزمون‌ها و سطرهای آن افراد را تشکیل می‌دهد، به طوری که درآیه‌های ماتریس تصمیم، نمره‌های مقیاسی هر فرد در هر خرده آزمون هستند. فرض کنید که n شخص در m خرده آزمون شرکت کرده و نمرهٔ مقیاسی شخص i ام در خرده آزمون j ام به صورت S_{ij} تعریف شده باشد.

$$S_D = \left[S_{ij} \right]_{n \times m} \quad (۷)$$

مرحلهٔ دوم: برای هر یک از نمره‌های ماتریس تصمیم، احتمال P_{ij} به صورت رابطه (۱۰) محاسبه می‌شود.

$$P_{ij} = \frac{S_{ij}}{\sum_{i=1}^{n} S_{ij}}, j = 1, 2, \ldots, m \quad (۸)$$

مرحلهٔ سوم: آنتروپی E_j را برای هر خرده آزمون به صورت رابطه (۱۱) محاسبه می‌شود.

$$E_j = -K \sum_{i=1}^{n} P_{ij} \ln P_{ij} \quad (۹)$$

در رابطهٔ (۹) مقدار K برابر با مقدار $\frac{1}{\ln(n)}$ است. این کار مقدار E_j را بین صفر تا یک نگه می‌دارد.

مرحلهٔ چهارم: برای نشان دادن درجهٔ مفید بودن اطلاعاتی که از هر خرده آزمون استخراج می‌شود، می‌توان از درجه انحراف[1] استفاده کرد. مقدار درجهٔ انحراف از رابطهٔ $d_j = 1 - E_j$ محاسبه می‌شود. برای هر خرده آزمون یک درجه انحراف محاسبه می‌شود. نزدیکی درجهٔ انحراف دو خرده آزمون‌ها به یکدیگر به معنی عدم تفاوت افراد بین آن دو خرده آزمون است. پس نقش آن خرده آزمون‌ها در اندازه‌گیری باید به همان اندازه کم شود.

مرحلهٔ پنجم: وزن هر خرده آزمون با استفاده از درجه انحراف آن خرده آزمون محاسبه می‌شود.

$$w_j^* = \frac{d_j}{\sum_{j=1}^{m} d_j} \quad (۱۰)$$

1. Degree of diversification

می‌توان از وزن‌های اســـمی در تشکیل وزن‌های آنتروپی نیز استفاده کرد به گونه‌ای که آزمون‌ساز از قبل برای هر خرده آزمون، وزن λ_j را تعریف کرده باشد. وزن نهایی آنتروپی برای هر خرده آزمون به صورت رابطهٔ (۱۱) تعریف می‌شود.

$$w_j = \frac{\lambda_j w^*_j}{\sum_{j=1}^{m} \lambda_j w^*_j} \qquad (۱۱)$$

روش‌های ساختن نمره‌کل: برای ساختن نمره‌کل، وزن مربوط به هر خرده آزمون در نمرهٔ مقیاسی آن خرده آزمون ضرب می‌شود و سپس حاصل ضرب‌ها با هم جمع و بر مجموع وزن‌ها تقسیم می‌شود. در جدول (۲) روش‌های مختلف ساختن نمره‌کل را که در این پژوهش استفاده شد، نمایش داده شده است. این جدول ترکیب روش‌های مختلف هموارسازی، ساختن نمرهٔ مقیاسی و همچنین وزن‌دهی را برای ساخت نمره‌کل نمایش می‌دهد. علامت ستاره در هر خانه، به این معنی است که از آن روش خاص برای ساختن نمره‌کل بهره گرفته شده است.

جدول ۲. روش‌های مختلف طراحی شده در پژوهش برای ساختن نمره کل

روش	طرح وزن دهی			نمره‌های مقیاس	
	A	B	C	NT[1]	AT[2]
A: مقیاس نرمال و طرح وزن دهی NA	*	–	–	*	–
B: مقیاس نرمال و طرح وزن دهی NB	–	*	–	*	–
C: مقیاس نرمال و طرح وزن دهی NC	–	–	*	*	–
A: مقیاس آرک و طرح وزن دهی AA	*	–	–	–	*
B: مقیاس آرک و طرح وزن دهی AB	–	*	–	–	*
C: مقیاس آرک و طرح وزن دهی AC	–	–	*	–	*

خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاسی و نمره‌کل

استانداردهای حرفه‌ای برای برگزاری آزمون‌های آموزشی همواره توجه کرده‌اند که در گزارش‌های مربوط به آزمون، هم از خطای اســـتاندارد اندازه‌گیری کلی (SEM[3]) و هم از خطای استاندارد اندازه‌گیری شرطی (CSEM[4]) استفاده شود. هم اکنون گزارش‌های فنی خطای اســـتاندارد اندازه‌گیری را بیشتر به صورت خطای استاندارد اندازه‌گیری کلی ارائه می‌دهند

1. Normal Transformation
2. Arcsine Transformation
3. Standard Error of Measurement
4. Conditional Standard Error of Measurement

ولی مؤسسه‌های NCEME، AERA و APAاز سال ۱۹۸۵ برای استانداردهایی که به منظور تولید آزمون‌های آموزشی و روانی توصیه کرده‌اند، گزارش خطای اســـتاندارد اندازه‌گیری شرطی را به همراه گزارش‌های فنی آزمون همواره توصیه کرده‌اند (اســـتاندارد شماره ۲-۱۰ APA). خطای استاندارد اندازه‌گیری شرطی که در این پژوهش با نماد CSEM نمایش داده شده است، قادر است میزان خطای استاندارد اندازه‌گیری را برای همهٔ سطوح نمره‌ها برآورد کند. این شاخص آماری هم برای نمره‌های خام و هم برای نمره‌های مقیاس‌بندی شده و هم برای نمره‌های ترکیبی قابل محاسبه است (وودروف و همکاران، ۲۰۱۳). بررســـی این شاخص آماری نشان می‌دهد که میزان خطای استاندارد اندازه‌گیری برای همهٔ نمره‌ها برابر نیســـت و سطوح مختلف نمره‌ها خطای استاندارد اندازه‌گیری شرطی متفاوتی دارند (کولن، ۲۰۱۴). خطای استاندارد اندازه‌گیری شرطی، تعریفی شبیه خطای اســـتاندارد اندازه‌گیری دارد. خطای استاندارد اندازه‌گیری شرطی، واریانس نمرهٔ مشاهده شدهٔ هر شرکت کننده در طی برگزاری آزمون‌های موازی در شرایط مشابه است، البته با این فرض که نمرهٔ حقیقی او ثابت بماند (هارتل، ۲۰۰۶). روش‌های متعددی برای محاسبهٔ خطای استاندارد اندازه‌گیری شرطی پیشنهاد شده است که در این پژوهش از روش برنان و لی (۱۹۹۹) برای محاسبهٔ این خطا استفاده شد.

روش برنان و لی برای محاسبهٔ خطای استاندارد اندازه‌گیری شرطی که به روش دو جمله‌ای نیز مشهور است، از تعمیم دو نظریه، یعنی نظریهٔ نمرهٔ حقیقی قوی[1] لرد در سال ۱۹۵۵و ۱۹۵۷ و کولن در سال ۱۹۹۲استفاده می‌کند و رابطه‌ای را برای خطای استاندارد اندازه‌گیری شرطی ارائه می‌دهد (برنان و لی، ۱۹۹۹). براساس نظریهٔ نمرهٔ حقیقی قوی، احتمال شرطی اینکه شخصی از مجموع k سؤال در یک آزمون بتواند به y تا از آنها پاسخ صحیح بدهد از رابطهٔ (۱۲) محاسبه می‌شود:

$$p(y \mid \pi, k) = \binom{k}{y} \pi^{y} (1-\pi)^{y-k} \qquad (۱۲)$$

در رابطهٔ (۱۲) پارامتر π نمرهٔ حقیقی نســـبت پاسخ‌های صحیح برای هر شخص است. طبق آنچه که در آمار مقدماتی موجود است می‌توان رابطهٔ واریانس شرطی نمره‌هایy را به صورت زیر نیز نوشت:

$$\sigma^{2}(Y \mid X) = E\left((Y - E(Y \mid X))^{2} \mid X\right) \qquad (۱۳)$$

که اگر تابع توزیع شرطی نمره‌ها موجود باشد رابطه به صورت زیر تغییر خواهد کرد:

$$\sigma^{2}(Y \mid X) = \sum Y^{2} p(Y \mid X) - \left(\sum Y p(Y \mid X)\right)^{2} \qquad (۱۴)$$

1. Strong True Score

خطاهای استاندارد اندازه‌گیری شرطی همان واریانس شرطی خطاها به شرط هر نمره است (کولن، ۱۹۹۲)، پس به کمک رابطهٔ (۱۴) خطای استاندارد اندازه‌گیری شرطی محاسبه می‌شود.

$$\sigma_E^{۲}(y|x)=\frac{k}{k-۱}\left\{\sum y^{۲}p(y|\pi,k)-\left(\sum yp(y|\pi,k)\right)^{۲}\right\} \quad (۱۵)$$

برای محاسبهٔ خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس، طبق نظریهٔ نمرهٔ حقیقی قوی می‌توان در رابطهٔ ۱۵ به جای مقادیر نمرهٔ خام y از تبدیل شده غیرخطی آنها (مثلا نمره‌های نرمال) استفاده کرد.

$$\sigma_E^{۲}(s(x)|x)=\frac{k}{k-۱}\left\{\sum f(y)^{۲}p(y|\pi,k)-\left(\sum f(y)p(y|\pi,k)\right)^{۲}\right\} \quad (۱۶)$$

هر گاه نمرهٔ مقیاسی شخص i ام برای خرده آزمون jام را با $S_j(X_i)$ و وزن هر خرده آزمون را با W_j نمایش دهیم، آنگاه نمرهٔ کل هر فرد، Y_i، از رابطهٔ ۱۷ محاسبه می‌شود.

$$Y_i=\sum_{j=۱}^{n}W_jS_j(X_i) \quad (۱۷)$$

برای محاسبهٔ خطای استاندارد اندازه‌گیری شرطی نمره‌های کل از رابطه‌ای که لاری پرایس و همکارانش پیشنهاد داده‌اند اســتفاده می‌شود، این رابطه از ترکیب خطای استاندارد اندازه‌گیری شرطی هر کدام از خرده آزمون‌ها به دست آمده است (پرایس، ۲۰۰۶).

$$SEM_{Y_i}=\sqrt{\sum_{j=۱}^{m}w_j^{۲}\hat{\sigma}_{E(s(x_{ij})|x_{ij})}} \quad (۱۸)$$

در رابطهٔ (۲۰) SEM_{Y_i} خطای اســتاندارد شرطی برای نمره‌کل شخص i ام است. m تعداد خرده آزمون‌ها است. برای محاســبهٔ ضریب پایایی نمره‌های مرکب از رابطه (۱۹) استفاده خواهد شد، این رابطه را وانگ و استانلی (۱۹۷۰) و کولن (۲۰۰۶) پیشنهاد داده‌اند.

$$\rho_c=۱-\frac{\sum_i w_i^{۲}\sigma_i^{۲}(۱-\rho_{ii})}{\sum_i\left(w_i^{۲}\sigma_i^{۲}+w_i\sum_{k\neq j}w_k\sigma_{ik}\right)} \quad (۱۹)$$

روش تصمیم‌گیری چند شاخصی

تحقیق در عملیات، رویکردی علمی است که در صدد حل مسائل مدیریتی و هدف آن کمک به مدیران برای تصمیم‌گیری بهتر اســـت. تحقیق در عملیات معمولا در قالب عناوینی همچون علم مدیریـــت، روش‌های کمّی، تحلیل کمّی، و علم تصمیم‌گیری نیز بیان می‌گردد. در علم مدیریت تصمیم‌گیری، نتیجهٔ فرآیند انتخاب گزینه‌ای بهتر از بین دو یا چند گزینهٔ متفاوت و یا پاســـخ مثبت یا منفی به یک موضوع می‌باشـــد که ما را در رسیدن به مقصود (آرمان) یاری می‌دهد. به طور کلی اگر بخواهیم تعریف جامعی از فنون تصمیم‌گیری ارائه دهیم می‌توان گفت فنون تصمیم‌گیری به مجموعه فنون و روش‌هایی اطلاق می‌شود که برای ارزیابی راه حل‌های ممکن موجود (گزینه‌های رقیب) و انتخاب بهترین راه حل به کار می‌رود (آذر و رجب زاده، ۱۳۹۳). فنون تصمیم‌گیری را می‌توان با توجه به ماهیت، شرایط و معیارهای مؤثر در تصمیم‌گیری طبقه‌بندی کرد. فنون تصمیم‌گیری بر اساس ماهیت به دو دسته تصمیم‌گیری‌های کیفی (ذهنی) و تصمیم‌گیری‌های کمّی (عینی) تقسیم می‌شود. همچنین فنون تصمیم‌گیری بر مبنای شرایط تصمیم‌گیری به چهار دسته تقسیم می‌شود: الف) تصمیم‌گیری در شرایط اطمینان کامل ب) تصمیم‌گیری در شرایط احتمالی (مثلا تصمیم بیز[1]) ج) تصمیم‌گیری در شرایط فازی د) تصمیم‌گیری در شرایط عدم اطمینان کامل (آذر و رجب زاده، ۱۳۹۳). تقسیم‌بندی دیگر مورد توجه در این پژوهش، طبقه‌بندی تصمیم‌گیری بر اساس معیارهای مورد ارزیابی است. این فنون تصمیم‌گیری به دو دستهٔ مهم تصمیم‌گیری تک شاخصی و تصمیم‌گیری چند شاخصی تقسیم می‌شود. فنون تصمیم‌گیری تک معیاره، به مجموعه فنونی گفته می‌شود که به دنبال ارزیابی راه حل‌های ممکن موجود و انتخاب بهترین راه حل بر اســـاس معیار ارزیابی است. مثال بارز این نوع تصمیم‌گیری، برنامه‌ریزی‌های خطی و غیرخطی در مدیریت است. فنون تصمیم‌گیری که به دنبال ارزیابی راه حل‌های ممکن موجود بر اساس چند شاخص برای انتخاب بهترین راه حل می‌باشد را فنون تصمیم‌گیری چند شاخصی می‌نامند (ایشیزاکا و نمری، ۲۰۱۳). به طور کلی اصول تصمیم‌گیری چند شاخصی[2] (MCDM) به دو دسته کلی تقسیم می‌شوند: الف) الگوهای تصمیم‌گیری چند هدفی[3] (MODM)، ب) الگوهای تصمیم‌گیری چند شاخصی[4] (MADM). در الگوهای چند هدفی، چندین هدف به طور هم‌زمان برای بهینه شدن، مورد توجه قرار می‌گیرد. ممکن است مقیاس سنجش برای هر هدف با هدف دیگر متفاوت باشد. به عنوان مثال یک هدف ممکن است حداکثر کردن سود بر حسب واحد پول باشد، ولی هدف دیگر حداقل اســـتفاده از ساعت کار برحسب ساعت باشد (ایشیزاکا و نمری، ۲۰۱۳). گاهی این اهداف در یک جهت نیستند و به صورت متضاد عمل می‌کنند. به عنوان مثال تصمیم‌گیرنده از یک طرف تمایل دارد رضایت کارکنان را افزایش دهد و از طرف دیگر می‌خواهد هزینه‌های حقوق و دستمزد را کاهش دهد. در الگوهای تصمیم‌گیری چند شاخصی، گزینه‌های مختلف را بر حسب چند شاخص اولویت‌گذاری یا ارزیابی می‌کنند. فنون تصمیم‌گیری چند شاخصی متفاوتی وجود دارد که در این پژوهش از رویکرد مبتنی بر فن تصمیم‌گیری «TOPSIS" برای رتبه‌بندی روش‌های نمره کل‌سازی

1. Bayesian Decision
2. Multiple Criteria Decision Making (MCDM)
3. Multiple Objective Decision Making (MODM)
4. Multiple Attribute Decision Making (MADM)

استفاده شده است. عبارت «TOPSIS" مخفف عبارت «تکنیک شباهت ترجیح ترتیب با جواب ایده آل[1]» است. این روش با دریافت تعداد کمی ورودی از کاربر، خروجی ساده و قابل فهمی دارد و تنها پارامترهای ذهنی آن وزن‌ها و شاخص‌ها هستند. ایدهٔ بنیادین «TOPSIS" آن است که بهترین جواب، جوابی است که کمترین فاصله را از جواب ایده‌آل و بیشترین فاصله را از جواب ضد ایده‌آل داشته باشد (ایشیزاکا و نمری، ۲۰۱۳). در ادامه ابتدا روش رایج «"TOPSIS که در علم مدیریت کاربرد بیشتری دارد شرح داده می‌شود و سپس ایده این پژوهش که مبتنی بر «"TOPSIS است توضیح داده خواهد شد.

ابتدا عملکرد nگزینه در m شاخص در یک ماتریس تصمیم به صورت $X = \left[x_{ij} \right]_{n \times m}$ جمع‌آوری می‌شود. «"TOPSIS بر اساس پنج مرحله محاسبه بنا نهاده شده است (ایشیزاکا و نمری، ۲۰۱۳):

مرحلهٔ اول: عملکرد گزینه‌ها در شاخص‌های متفاوت نرمال می‌شود تا بتوان اندازه‌های موجود با واحدهای مختلف را با هم مقایسه کرد. برای این منظور چند راه نرمال‌سازی وجود دارد:

الف) نرمال‌سازی توزیعی با تقسیم عملکردها بر جذر مجموع عناصر ستون در ماتریس تصمیم به دست می‌آید:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{n} x_{ij}^{2}}} \qquad (۲۰)$$

ب) نرمال‌سازی ایده‌آل (اگر معیار مد نظر سود و فایده باشد) عملکرد را بر بیشترین مقدار موجود در هر ستون تقسیم می‌کند. اگر معیار کمترین باشد، هر عملکرد بر کمترین مقدار در ستون خود تقسیم می‌شود.

$$r_{ij} = \frac{x_{ij}}{\max(x_{ij})} \quad \text{برای معیار بیشترین} \qquad (۲۱)$$

$$r_{ij} = \frac{x_{ij}}{\min(x_{ij})} \quad \text{برای معیار کمترین} \qquad (۲۲)$$

مرحلهٔ دوم: محاسبهٔ وزن‌ها برای هر شاخص و تشکیل ماتریس نرمال موزون با ضرب ماتریس وزن‌ها در ماتریس تصمیم

$$V = X \times W \Rightarrow \left[v_{ij} \right]_{n \times m} = \left[x_{ij} \right]_{n \times m} \times \left[w_{ij} \right]_{m \times m} \qquad (۲۳)$$

1. Technique of Order Preference Similarity to the Ideal Solution

برای تشکیل ماتریس وزن‌ها ابتدا برای هر شاخص، وزن جداگانه‌ای به یکی از روش‌های ساختن وزن (آنتروپی شانون، AHP یا سایر روش‌های مشابه) ساخته می‌شود. در این تحلیل چون mشاخص موجود است، پس m وزن نیز محاسبه می‌شود و ماتریس $\left[w_{ij} \right]_{m\times m}$ به گونه‌ای ساخته می‌شود که وزن شاخص‌ها روی قطر اصلی قرار می‌گیرد و سایر درآیه‌های ماتریس صفر است.

مرحلهٔ سوم: تشکیل بردارهای ایده‌آل (ایده‌آل مثبت) و ضد ایده‌آل (ایده‌آل منفی). سه راه مختلف برای تشکیل بردارهای ایده‌آل وجود دارد:

الف) از بهترین عملکرد گزینه‌ها برای هر شاخص در تشکیل بردار ایده‌آل یعنی $A^{+}=(\mathrm{v}_{1}^{+},\mathrm{v}_{2}^{+},...,\mathrm{v}_{m}^{+})_{1\times m}$ و از بدترین عملکرد گزینه‌ها در هر شاخص برای تشکیل بردار ضد ایده‌آل یعنی $A^{-}=(\mathrm{v}_{1}^{-},\mathrm{v}_{2}^{-},...,\mathrm{v}_{m}^{-})_{1\times m}$ استفاده می‌شود.

ب) استفاده از ایده‌آل‌های مطلق برای ساختن ایده‌آل و ضد ایده‌آل، به طوری که بردار ایده‌آل به صورت $A^{+}=(1,1,...,1)_{1\times m}$ و بردار ضد ایده‌آل به صورت $A^{-}=(0,0,...,0)_{1\times m}$ تعریف می‌شود.

ج) نقاط ایده‌آل و ضد ایده‌آل توسط تصمیم‌گیرنده مشخص شود.

مرحلهٔ چهارم: فاصلهٔ هر گزینه با گزینهٔ ایده‌آل و ضد ایده‌آل از راه فاصلهٔ اقلیدسی محاسبه می‌شود.

فاصلهٔ هر گزینه تا ایده‌آل (ایده‌آل مثبت)

$$d_{i}^{+}=\left\{\sum_{j=1}^{m}(v_{ij}-v_{j}^{+})^{2}\right\}^{\frac{1}{2}}$$

فاصلهٔ هر گزینه تا ضد ایده‌آل (ایده‌آل منفی)

$$d_{i}^{-}=\left\{\sum_{j=1}^{m}(v_{ij}-v_{j}^{-})^{2}\right\}^{\frac{1}{2}}$$

مرحلهٔ پنجم: محاسبهٔ ضریب نسبی نزدیکی برای هر گزینه. از رابطه (۲۴) برای محاسبهٔ ضریب مجاورت استفاده می‌شود. مقادیر نزدیک به یک، به معنی نزدیک بودن گزینه به ایده‌آل و مقادیر نزدیک به صفر، به معنی نزدیک بودن گزینه به ضد ایده‌آل است.

$$c_{i}=\frac{d_{i}^{-}}{d_{i}^{-}+d_{i}^{+}} \qquad (۲۴)$$

گزینه‌ها را می‌توان بر حسب مقدار c_i به دست آمده اولویت‌گذاری و یا رتبه‌بندی کرد (ایشیزاکا و نمری، ۲۰۱۳).

تصمیم‌گیری چند شاخصی و رتبه‌بندی روش‌های نمره کل‌سازی

در اینجا قصد داریم با تلفیق مفاهیم خطای استاندارد اندازه‌گیری شرطی و نظریهٔ تصمیم با رویکرد روش «TOPSIS"،

راهکار تازه‌ای برای بررسی کارآمدی نمره‌های کل ارائه شود، به طوری که انتخاب آن روش نمره کل‌سازی نسبت به بقیه کارآمدتر باشد و موجب زیان کمتری به داوطلبان شود. در این رویکرد آن روش نمره‌کل سازی انتخاب می‌شود که در آن نســـبت به بقیهٔ روش‌ها برای همهٔ افراد شرکت کننده کمترین خطا را انجام داده باشیم. در ادامه این روش به طور کامل شرح داده خواهد شد.

ابتدا ماتریس تصمیم تشکیل می‌شود. در اینجا ماتریس تصمیم با ماتریس تصمیم در روش «TOPSIS" قدری متفاوت است. در روش TOPSIS"" ماتریس تصمیم ماتریسی شامل نمره‌های خام هر گزینه (اینجا شخص) در هر شاخص (اینجا خرده آزمون) است. اما در این روش ماتریس تصمیم به صورت زیر تعریف می‌شود:

$$CSEM_{D} = \left[SEM_{Y_{ij}} \right]_{n\times m} \qquad (۲۵)$$

همان طور که ملاحظه می‌شود، ماتریس تصمیم، ماتریسی است که n سطر و m ستون دارد.n تعداد افراد شرکت کننده در آزمون مرکب و m تعداد روش‌های نمره کل‌سازی است. درایه‌های این ماتریس یعنی $SEM_{Y_{ij}}$ دیگر نمره‌های خام یا نمره‌های مقیاس نیستند، بلکه خطای استاندارد اندازه‌گیری شرطی نمره کل شخص i ام در روش نمره کل‌سازی jام است. برای هر شخص در هر روش نمره‌کل سازی یک نمره‌کل تولید شده است که خطای استاندارد اندازه‌گیری شرطی آن $SEM_{Y_{ij}}$ اســـت. در این روش از روش‌های نرمال‌سازی و یا وزن دهی که در روش "TOPSIS" به آنها اشاره شد، خبری نیست و پس از تشکیل ماتریس تصمیم، ایده‌آل‌ها و ضد ایده‌آل‌ها به طور مستقیم تعریف می‌شود.

فرض می‌شود که n شخص، هر کدام m نوع نمره‌کل داشته باشند. پس هر شخص m خطای استاندارد اندازه‌گیری شرطی دارد. پس از تشکیل ماتریس تصمیم، بردارهای ایده‌آل (ایده‌آل مثبت) و ضد ایده‌آل (ایده‌آل منفی) به این صورت تعریف می‌شود. **بردار ایده‌آل (ایده‌آل مثبت)**، برداری است که شامل کمترین خطا در بین همه روش‌ها برای نمره کل هر شخص است یا به زبان ساده‌تر، هر شخص برای هر روش نمره‌کل سازی یک خطای استاندارد اندازه‌گیری شرطی دارد که کمترین آن به عنوان ایده‌آل مثبت در نظر گرفته می‌شود. چون n نفر در آزمون شرکت کرده‌اند بردار حاصل یک سطر و n ستون خواهد داشت که به صورت $A_j^{+} = (SEM^{+}_{Y_1}, SEM^{+}_{Y_2}, ..., SEM^{+}_{Y_n})_{1\times n}$ تعریف می‌شود. **بردار ضد ایده‌آل (ایده‌آل منفی)**، برداری است که شامل بیشترین خطا در بین همهٔ روش‌ها برای نمره‌کل هر شخص است و به زبان ساده‌تر، هر شخص برای هر روش نمره‌کل سازی خطای استاندارد اندازه‌گیری شرطی دارد که بیشترین آن به عنوان ایده‌آل منفی در نظر گرفته می‌شود. چون n نفر در آزمون شرکت کرده‌اند بردار حاصل یک سطر و n ستون دارد و به صورت $A_j^{-} = (SEM^{-}_{Y_1}, SEM^{-}_{Y_2}, ..., SEM^{-}_{Y_n})_{1\times n}$ تعریف می‌شود. پس از تعیین ایده‌آل‌ها برای هر روش نمره کل‌سازی، فاصلهٔ بین ایده‌آل‌های مثبت و منفی تا خطای استاندارد اندازه‌گیری شرطی افراد محاسبه می‌شود.

فاصلهٔ هر روش نمره‌کل سازی تا ایده‌آل مثبت:

$$d_j^{+} = \left\{ \sum_{i=1}^{n} (SEM_{Y_{ij}} - SEM^{+}_{Y_j})^{2} \right\}^{\frac{1}{2}}$$

فاصلهٔ هر روش نمره‌کل سازی تا ایده‌آل منفی:

$$d_j^- = \left\{ \sum_{i=1}^{n} (SEM_{Y_{ij}} - SEM^-_{Y_j})^2 \right\}^{\frac{1}{2}}$$

پس از تعیین فاصله‌ها نوبت به محاسبه شاخص مجاورت می‌شود که به صورت زیر تعریف می‌شود (مشابه روش TOPSIS):

$$c_j = \frac{d_j^-}{d_j^- + d_j^+} \quad j = 1, 2, \ldots, m \qquad (۲۶)$$

حال می‌توان روش‌های نمره‌کل سازی را بر حسب شاخص دوری یا نزدیکی رتبه‌بندی کرد. هر روش نمره‌کل سازی که شاخص آن به یک نزدیک‌تر باشد به ایده‌آل مثبت نزدیک‌تر و کارآمدتر است و هر روش که شاخص دوری یا نزدیکی آن به صفر نزدیک‌تر باشد به ایده‌آل منفی نزدیک‌تر و ناکارآمدتر است.

یافته‌ها

در جدول (۳) نتایج تحلیل داده‌ها برای نمونهٔ تصادفی به حجم ده هزار نفر مشاهده می‌شود. در ابتدا نمره‌های خام به صورت امتیاز یک برای پاسخ صحیح و امتیاز صفر برای پاسخ‌های نادرست و بی‌پاسخ تهیه شده است و سپس با جمع امتیازها نمره‌های خام به دست آمده است. جدول (۳) شامل شاخص‌های آماری میانگین، واریانس، چولگی و کشیدگی و شاخص‌های اندازه‌گیری مانند خطای استاندارد اندازه‌گیری و ضریب پایایی کودر ریچاردسون ۲۰ (KR20) می‌باشد. مقادیر داخل پرانتز، میانگین نمره‌های خام نسبت است. خطای استاندارد اندازه‌گیری در جدول (۳) با ضریب پایایی کودر ریچاردسون محاسبه شده و منظور از میانگین نمره‌های خام نسبت، در این جدول، میانگین حاصل تقسیم نمرهٔ خام هر فرد به تعداد سؤال‌های آن آزمون می‌باشد.

جدول ۳. شاخص‌های آماری و شاخص‌های اندازه‌گیری برای نمرهٔ خام خرده آزمون‌ها

شاخص	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

آن گونه که گفته شد، شش روش برای ساختن نمره‌کل طراحی شده است. تنوع این روش‌ها به خاطر ترکیب روش‌های متفاوت مقیاس‌سازی و طرح‌های وزن دهی است. برای ساخت این نمره‌کل‌ها از دو روش مقیاس‌سازی و سه طرح وزن دهی استفاده شد. جدول (۴) برخی شاخص‌های آماری به همراه ضریب پایایی هر روش نمره‌کل سازی را نشان می‌دهد. در اینجا برای محاسبهٔ ضریب پایایی نمره‌کل‌ها از رابطهٔ (۱۹) استفاده شد.

جدول ٤. برخی شاخص‌های آماری به همراه ضریب پایایی در روش‌های نمره‌کل سازی

نام روش	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
NA	۵۰۳۴/۶	۹۳۳۶۹۰	۰/۶۲۰	۳/۰۹۰	۰/۹۰۳
NB	۵۰۳۹/۷	۹۶۶۶۶۰	۰/۶۴۰	۳/۰۴۴	۰/۹۰۳
NC	۵۰۳۴/۶	۹۳۳۷۱۰	۰/۶۲۰	۳/۰۹۰	۰/۹۱۴
AA	۲۰۰۸/۴	۱۴۳۴۴۰	۱/۰۶۲	۴/۳۶۰	۰/۹۲۸
AB	۱۷۲۵/۵	۱۴۷۸۱۰	۱/۱۶۴	۴/۵۷۷	۰/۹۲۸
AC	۲۰۰۸/۵	۱۴۳۴۳۰	۱/۰۶۲	۴/۳۵۹	۰/۹۲۸

جدول (۵) مقدار وزن‌ها را برای هر خرده آزمون و برای هر طرح وزن دهی نشان می‌دهد. توجه شود که مقادیر طرح‌های وزن‌دهی Aبه دلیل عدم وابستگی وزن‌ها به ویژگی نمره‌ها و ثابت بودن، در جدول نیامده است.

جدول ۵. مقادیر وزن‌ها برای هر خرده آزمون در روش‌های نمره‌کل سازی

روش	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
NB	۰/۰۲۶۴	۰/۰۲۳۴	۰/۰۵۲۸	۰/۰۴۳۲	۰/۳۴۰	۰/۲۹۳	۰/۲۲۰
NC	۰/۱۰۵۲	۰/۰۵۲۶	۰/۰۷۸۹	۰/۰۵۲۶	۰/۳۱۶	۰/۲۳۷	۰/۱۵۸
AB	۰/۰۲۱۷	۰/۰۲۱۸	۰/۰۵۱۷	۰/۰۴۶۹	۰/۳۳۹	۰/۳۰۰	۰/۲۱۸
AC	۰/۱۰۵۳	۰/۰۵۲۶	۰/۰۷۹۰	۰/۰۵۲۶	۰/۳۱۶	۰/۲۳۷	۰/۱۵۸

طبق آنچه که گفته شـــد، به دنبال انتخاب کارامدترین روشـــی هستیم که تصمیم پذیرش بر اساس آن، کمترین ضرر به افراد وارد شود. در این روش کمترین خطاها به عنوان ایده‌آل مثبت و بیشترین خطاها به عنوان ایده‌آل منفی در نظر گرفته می‌شـــود. هر روشی که کمترین فاصله را با ایده‌آل مثبت داشته باشـــد، روش مطلوب و کارآمدی خواهد بود. جدول (۶) شاخص‌های دوری و نزدیکی (مجاورت) همهٔ شش روش نمره‌کل سازی را نمایش می‌دهد. رتبه‌بندی روش‌ها بر اساس

شاخص c انجام شده و مقادیر نزدیک به عدد یک رتبه‌های بالاتر و مقادیر نزدیک به عدد صفر رتبه‌های پایین‌تری را کسب کرده‌اند. این تحلیل‌ها به صورت کد نویسی در نرم افزار «MATLAB" صورت گرفته است.

جدول ٦. شاخص دوری و نزدیکی (مجاورت) برای روش‌های نمره کل‌سازی

روش	شاخص (شاخص مجاورت)c	رتبه بر اساس شاخص مجاورت
NA	۰/۲۷۷۹۲۳	ششم
NB	۰/۱۲۶۷۰۰	نهم
NC	۰/۲۷۷۸۰۸	هفتم
AA	۰/۸۶۷۴۷۸	اول
AB	۰/۷۹۳۹۵۰	سوم
AC	۰/۸۶۷۴۷۲	دوم

با بررسی جدول (۶) مشاهده می‌شود که رتبه‌های اول و دوم متعلق به دو روش نمره‌کل سازی AC و AA است. این دو روش به کمترین خطای استاندارد اندازه‌گیری ایده‌آل نزدیک‌اند.

بحث و نتیجه گیری

آنچه گذشت پژوهشی توصیفی با رویکرد کاربردی است که به منظور بررسی روش‌های مختلف ساخت نمره‌کل در آزمون‌های مرکب انجام شده است. در این پژوهش از نمونه تصادفی ده هزار نفری شرکت کنندگان آزمون سراسری ایران در سال ۱۳۹۵ در گروه آزمایشی ریاضی استفاده شد. خرده آزمون‌های هفت گانهٔ آزمون سراسری سال ۱۳۹۵ گروه ریاضی و فنی شامل زبان و ادبیات فارسی، زبان و ادبیات عربی، معارف اسلامی، زبان و ادبیات انگلیسی، ریاضیات، فیزیک و شیمی بود. برای ساخت نمرهٔ خام، از مجموع پاسخ‌های صحیح و از روش‌های نرمال‌سازی و آرک‌سینوس برای ساخت مقیاس استفاده شد. به منظور ساخت نمره‌کل از سه طرح وزن دهی اسمی، مؤثر و شانون استفاده شد. ضمن محاسبهٔ خطای استاندارد اندازه‌گیری شرطی برای همهٔ نمره‌های ساخته شده، به منظور رتبه‌بندی نمره‌های ساخته شده از روش تصمیم‌گیری چند شاخصی استفاده شد. در این پژوهش برای بررسی کارآمدی نمره‌های کل، سعی شد با تلفیق مفاهیم خطای استاندارد اندازه‌گیری شرطی و روش تصمیم‌گیری، رویکردی جدید ارائه شود.

طرح وزن‌دهی A که به هیچ کدام از ویژگی‌های نمره‌ها وابستگی ندارد و مطابق جدول (۱) برای خرده آزمون‌ها ارائه شد. مقادیر وزنی در طرح‌های وزن‌دهی B وC در جدول (۵) ارائه شد. از میان این روش‌ها، برای وزن‌دهی، در دو روش از طرح B و دو روش از طرح C استفاده شد. نتایج نشان داد که هر کدام از خرده آزمون‌ها که واریانس بزرگ‌تری داشت، وزن آن نیز بزرگ‌تر بود. این تفاوت در واریانس‌ها، هم به دلیل تعداد سؤال‌ها و هم به دلیل وزن اسمی، به وجود آمد. خرده آزمون‌های

ریاضی، فیزیک و شیمی هم به لحاظ تعداد سؤال‌ها و هم به لحاظ وزن اسمی از خرده آزمون‌های دیگر بالاتر بودند، واریانس بزرگ‌تری داشتند و در طرح‌های وزن‌دهی، وزن بزرگ‌تری را نیز به خود اختصاص داده‌اند. نوع مقیاس بر روی مقادیر این وزن‌ها موثر بوده است، به عنوان نمونه طرح‌های NB و AB در نوع وزن‌دهی مشترک ولی در مقیاس متفاوت هستند، این تفاوت در مقیاس، باعث شد که مقدار وزن هر کدام از روش‌ها برای هر خرده آزمون مشخص، متفاوت باشد. با ضرب هر کدام از وزن‌ها در خرده آزمون مربوط و جمع کردن آنها برای هر فرد، یک نمره‌کل ساخته می‌شود که طبق آنچه که گفته شد، برای هر فرد شش نمره‌کل متفاوت ساخته شد. جدول (۴) گشتاورهای اول تا چهارم نمره‌های کل را به همراه ضریب پایایی آنها نشان می‌دهد. با توجه به داده‌های این جدول می‌توان روش‌های نمره‌کل سازی را با توجه به نوع مقیاس به کار رفته در آنها به دو دسته تقسیم کرد. دستهٔ اول روش‌هایی است که از روش نرمال‌سازی برای ساخت مقیاس استفاده می‌شود (بدون توجه به طرح وزن دهی) و دسته دوم روش‌هایی که از آرک‌سینوس برای ساخت مقیاس استفاده می‌شود (بدون توجه به طرح وزن دهی). در روش‌های نمره‌کل سازی که در آنها از مقیاس نرمال استفاده می‌شود، میانگین نمره‌های کل بین ۵۰۳۴ تا ۵۰۳۹ متغیر بود. تنها طرح‌هایی که دارای طرح وزن دهی B بودند واریانس‌های بزرگ‌تری ایجاد کردند و واریانس نمره‌ها برای سایر طرح‌های وزن‌دهی به طور تقریبی تفاوت زیادی نداشت. مقادیر پایایی در جدول (۴) نشان می‌دهد که هیچ کدام از روش‌های نمره کل‌سازی نتوانسته‌اند تفاوت آشکاری در ضریب پایایی ایجاد کنند. پژوهش‌های پیشین نیز از جمله چانگ (۲۰۰۶)، پی و مایر (۲۰۰۷) و ذوالفقارنسب، خدایی، و یادگارزاده (۱۳۹۱) نیز همین نتیجه را ارائه کرده‌اند، به طوری که در روش‌های شامل مقیاس نرمال، تفاوت ضریب پایایی بین بالاترین مقدار و کمترین، نزدیک ۰/۰۱۱ است. در روش‌هایی که در آنها از مقیاس آرک‌سینوس برای نمره‌کل سازی استفاده شده، میانگین نمره‌ها بین ۱۷۰۰ تا ۲۰۰۸ است و مشابه دستهٔ اول، روش‌های شامل طرح وزن‌دهی B واریانس بیشتری در میان نمره‌ها ایجاد کرده است. انواع روش‌های نمره‌کل سازی که شامل مقیاس آرک‌سینوس بوده‌اند نتوانستند تفاوت آشکاری در مقدار ضریب پایایی ایجاد کنند. نکته‌ای که باید در اینجا به آن اشـــاره کرد تفاوت ضریب پایایی بین روش‌هایی است که شامل مقیاس نرمال بودند و روش‌هایی که شـــامل آرک‌سینوس هستند. ضریب پایایی نمره کل‌هایی که با مقیاس آرک‌سینوس ساخته شده‌اند، از مقیاس نرمال بیشتر است. میانگین ضریب پایایی برای نمره‌هایی که از مقیاس نرمال استفاده کرده‌اند برابر ۰/۹۰۶ و برای نمره‌هایی که از مقیاس آرک‌سینوس استفاده کرده‌اند برابر ۰/۹۲۸ است. این نتایج حاکی از برتری، هر چند اندک، روش‌هایی است که از مقیاس آرک‌سینوس برای ساخت نمره‌کل استفاده می‌شود.

یافته‌های پژوهش نشان داد که در میان شش روش برای نمره‌کل سازی، روش نمره‌کل سازی AA با شاخص دوری و نزدیکی (مجاورت) ۰/۸۶۷۴۷۸ رتبهٔ اول را کسب کرد. این به این معنی است که نمره‌کل‌های ساخته شده به این روش، کمترین فاصله را تا خطای کمینهٔ تعیین شده دارد و در صورتی که افراد با استفاده از نمره‌کل‌های ساخته شده به این روش رتبه‌بندی شوند با کمترین خطای ممکن رتبه‌بندی خواهند شد. البته اختلاف شاخص دوری و نزدیکی این روش با روش بعد از آن یعنی AC بسیار ناچیز و در حد ۰/۰۰۰۰۶ است که می‌توان این دو روش را تقریباً معادل در نظر گرفت. نتایج استفاده از

روش تصمیم‌گیری چند شاخصی حاکی از برتری دو روش AC و AA است گرچه در رتبه‌بندی این روش‌ها تفاوت زیادی مشاهده نشد. این نشان می‌دهد که آزمون‌ساز می‌تواند با توجه به مقتضیات آزمون و شرایط اجرا و تفسیر آزمون، با اطمینان برابر از هر کدام از این روش‌ها بهره ببرد. در پایان می‌توان چنین گفت که طبق روش تصمیم و معیار انتخاب، روشی که کمترین خطا را برای همهٔ افراد مرتکب می‌شود، روش‌های نمره‌کل سازی AC و AA را می‌توان به طور تقریبی با هم معادل دانست و هیچ کدام در ساخت نمره کل بر دیگری برتری نداشت.

با توجه به نتایجی که از تحلیل تصمیم به دست آمد، به سازندگان آزمون‌های مرکب می‌توان توصیه کرد که برای رسیدن به کمترین خطای ممکن، یکی از دو روش AC یا AA را انتخاب کنند. البته چون کاربرد روش تبدیل آرک‌سینوس به همراه وزن‌های اسمی، بیشتر در آزمون‌هایی است که محتوای درس‌ها برای پذیرش، اهمیت زیادی دارد، در چنین مواردی از آن استفاده شود. این کاربرد به خاطر ماهیت وزن دهی به شیوهٔ A می‌باشد. اما برای مواردی که پراکندگی نمره‌ها و استفادهٔ حداکثری از آنتروپی اطلاعات برای آزمون‌ساز اهمیت داشته باشد، روش AC بر روش AA ترجیح دارد. در اینجا، تنها شرایط آزمون و تفاسیری که قرار است از آن بشود، اهمیت دارد زیرا نشان داده شد که هر دو روش به لحاظ دقت، رتبهٔ یکسانی دارند.

برای ساخت نمره‌کل روش‌های متعددی وجود دارد. به عنوان مثال نمره‌های نرمال‌سازی شده را می‌توان پس یا پیش از تبدیل به مقیاس نرمال هموارسازی کرد یا نمره‌های مقیاس آرک‌سینوس را می‌توان پس از تبدیل به مقیاس آرک‌سینوس هموارسازی کرد، ضمن اینکه روش‌های وزن‌دهی همچون وزن‌دهی با عکس خطای استاندارد اندازه‌گیری، وزن‌دهی بر اساس پایایی نمره‌ها و وزن‌دهی بر اساس طول آزمون از دیگر روش‌های وزن‌دهی است که در این پژوهش به آنها پرداخته نشد. روش تصمیم‌گیری چند شاخصی قادر است همهٔ انواع روش‌های نمره‌کل‌سازی را با هم مقایسه کند.

تقدیر و تشکر

از سازمان سنجش آموزش کشور به خاطر حمایت معنوی که از این اثر و همکاری که در ارسال داده‌های آزمون سراسری با این پژوهش داشتند، تقدیر و تشکر می‌نمایم.

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.)

American Educational Research Association, American Psychological Association, & National

Council on Measurement in Education, (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Azar, A., Rajabzade A. (2015). *Applied Decision Making MADM Approach*. Tehran: Negah Danesh.

Brennan, Robert L., Lee, Won-Chan. (1999) Conditional Scale-Score Standard Errors of Measurement under Binomial and Compound Binomial Assumptions, *Educational and Psychological Measurement*, Vol 59, Issue 1, pp. 5 – 24.

Brooks, G. P., Johnson, G. A.(2014). *TAP: Test Analysis Program* [computer software]. Chicago.

Chang, S. W. (2009). Choice of weighting schemes in forming the composites, *bulletin of educational psychology*,40(3), 489-510, national Taiwan normal university, Taipei, Taiwan, R.O.C.

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.

De Boor, C. (2001). *A Practical Guide to Splines* (Revised Edition). pp. 207–214, New York: Springer.

Feldt, L. S. (2004). Estimating the reliability of a test battery composite or a test score based on weighted item scoring. *Measurement and Evaluation in Counseling and Development*, 37(3), 184-190.

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

Gronlund. N. E. & Linn R. T. (1990), *measurement and evaluation in teaching*. New York: Macmillan.

Haertel, H. E. (2006). *Reliability*. In R. L. Brennan (Ed.), Educational measurement (4rd. ed., pp. 65-86). CT: American Council on Education and Praeger.

Iowa Assessment (2016). *Iowa Tests of Basic Skills*, Retrieved itp.education.uiowa.edu

Ishizaka, A., Nemery, P. (2013). *Multi-criteria Decision Analysis: Methods and Software*, New York: John Wiley & sons.

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., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, 33, 129-140.

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

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

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

Kolen, M. J, Wang, T., Lee, W. Chon. (2012), Conditional Standard Errors of Measurement for Composite Scores Using IRT, *International Journal of Testing*, 12, 1-20.

Lord, F. M., & Novick, M. R. (1967). Statistical theory of mental test scores. MA: Adisson-wesley.

Nunnally, J. c., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Magnusson, D. (1967). *Test theory*. MA: Addison-Wesley.

Nitko, A. J. (2001), *Educational assessment and evaluation* (3rd Ed.). New Jersey: Merrill prentice-hall.

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.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). *Scaling, norming, and equating*. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 221-262). New York: American Council on Education, and Macmillan.

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.

Testing, National Organization. (2015, Sep 01). NOET web page. Retrieved from www.sanjesh.org

The ACT, The ACT technical manual (2014), Retrieved www.act.org

The SAT, SAT technical manual (2015), Retrieved collegereadiness.collegeboard.org.

Wang, T. (1998). Weights that maximize reliability under a congeneric model. *Applied psychological measurement*, 22(2), 179-187.

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. Retrieved from www.act.org.

Zolfagharnasab, S., Khodaei, E., Yadegarzadeh, G. (2013). Optimum Weighting to Entrance Subtests and Their Items to Make Composite Score. *Educational Measurement and Evaluation Studies*, 3(4), 79-104.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی