Educational Measurement and Evaluation Studies

National Organization for Educational Testing

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

Utilizing the Decision-Making Approach to Rank Composite Score Construction Methods

Mojtaba Jahanifar[1]

1. Assistant professor, Education departement, Shahid Chamran University of Ahvaz, Ahvaz, Iran., m.jahanifar@scu.ac.ir

Article Info

Article Type:
Research Article

Received
2022/04/13
Received in revised form
2022/05/21
Accepted
2022/05/24
Published online
2022/08/14

ABSTRACT

Objective: Battery Test is usually used for decision-making in education and Admission decisions. There are several methods to construct composite scores so each method makes a different effect on the admission decision. However, which decision makes fewer errors?

Methods: present research has been conducted to rank different methods of composite score construction based on their CSEM. 10,000 random sample Data from participants of the Iran university entrance exam were used to rank six composite score construction methods. The participants' raw score arises from summing up correct responses. Normalizing and Arcsine transformation methods were used to Construct scale scores, also we used nominal, effective and Shannon weighting schemes to combine subtest scale scores. In order to rank composite score construction methods, a new approach was employed based on the MADM decision-making approach

Results: The results revealed that the methods that use Arcsine to construct scale scores and nominal or Shannon weighting schemes to combine subtest scale scores have taken the higher ranks, and less error will occur at admission decision.

Conclusion: Using the Arc Sine scale score, due to less error and easier conversion, can help the interpretation and accuracy of composite test scores, while different weighting methods do not affect the accuracy of scores and in accordance with the test conditions or test builders' decision can be used.

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

Cite this article: Jahanifar, Mojtaba. (2022). Utilizing the decision-making approach to Rank composite score construction methods. *Educational Measurement and Evaluation Studies*, 12 (37):1-17 Pages.
DOI: 10.22034/EMES.2022.550667.2366

Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی

سازمان سنجش آموزش کشور

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

استفاده از روش تصمیم‌گیری چندشاخصه در رتبه‌بندی روش‌های ساخت نمره‌کل

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
۱۴۰۱/۰۱/۲۴

اصلاح
۱۴۰۱/۰۲/۳۱

پذیرش
۱۴۰۱/۰۳/۰۳

انتشار
۱۴۰۱/۰۵/۲۳

چکیده

هدف: تصمیم پذیرش در آزمونها بیشتر براساس 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 نمره حاصل، نمره کل نامیده می‌شود. روش‌های متفاوت نمره کل سازی موجب تغییر در تصمیم پذیرش افراد می‌شود. این پژوهش با هدف رتبه‌بندی روش‌هایی که برای ساختن نمره کل استفاده می‌شود، انجام شده است.

روش پژوهش: از ۱۰۰۰۰ نمونه تصادفی آزمون سراسری در هفت خرده آزمون برای رتبه‌بندی شش روش نمره کل سازی بهره گرفته شده است. نمره خام از مجموع پاسخ‌های صحیح به دست آمده و از روش‌های نرمال‌سازی و آرک سینوس برای تبدیل نمره‌ها به نمره‌های مقیاس بهره برده شده است. از طرح‌های وزن دهی اسمی، موثر و شانون برای ساخت نمره کل استفاده گردید. به منظور رتبه‌بندی روش‌های نمره کل سازی بر اساس خطای استاندارد اندازه‌گیری شرطی آنها از رویکردی مبتنی بر تصمیم‌گیری چند شاخصه استفاده شد.

یافته‌ها: نتایج نشان داد که آن دسته از روش‌های نمره کل 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 نمره کل سازی ، خطای کمتری مرتکب خواهیم شد.

نتیجه‌گیری: استفاده از نمره مقیاس آرک سینوس، به دلیل خطای کمتر، تبدیل و راحت تر می تواند به تفسیرپذیری و دقت بیشتر نمره های آزمون های مرکب کمک کند، ضمن اینکه روش های متفاوت وزن دهی تاثیر چندانی بر دقت نمره ها نداشته و مطابق با شرایط آزمون و تصمیم آزمون ساز می توانند مورد استفاده قرار بگیرند.

واژه‌های کلیدی: تصمیم‌گیری چند شاخصه، خطای استاندارد اندازه‌گیری شرطی، طرح وزن دهی، نمره کل، نمره مقیاس

استناد: جهانی فر، مجتبی (۱۴۰۱). استفاده از روش تصمیم‌گیری چندشاخصه در رتبه‌بندی روش‌های ساخت نمره‌کل. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱۲ (۳۷)، صفحه۱-۱۷ DOI: 10.22034/EMES.2022.550667.2366

مقدمه

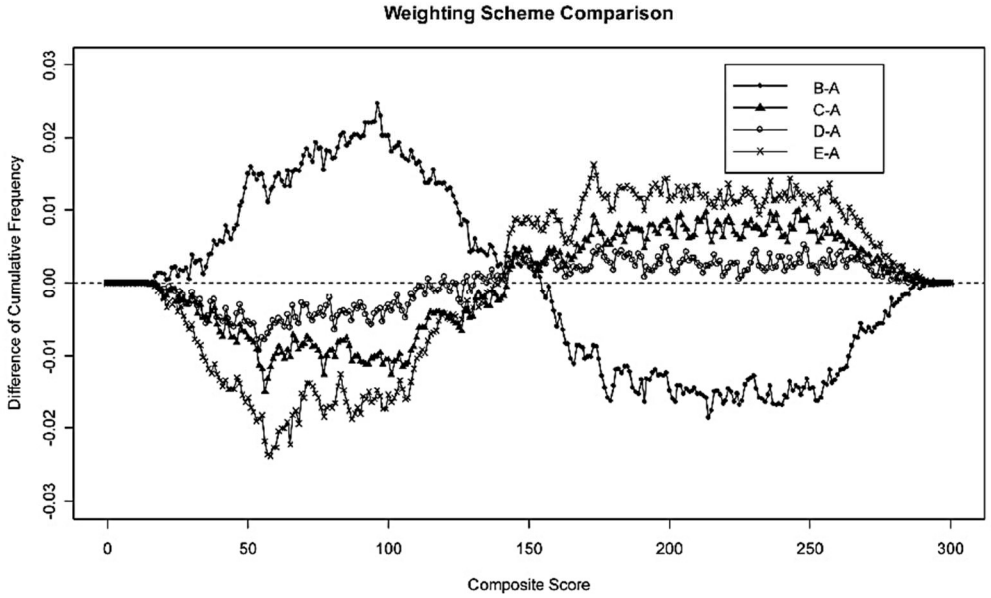
عاملی که اهمیت و حساسیت آزمون‌های سرنوشت ساز را بالا می‌برد خود آزمون نیست، بلکه تصمیم‌هایی است که برای رسیدن به آنها، آزمون اجرا میشود . اینکه آزمون به صورت چند گزینه‌ای باشد یا شفاهی چندان مهم نیست، مهم نوع تصمیمی است که قرار است براساس نتایج آزمون گرفته شود. به طور کلی در آزمون‌های سرنوشت ساز به دو مورد مهم اشاره شده: یکی عواقب آزمون و دیگری پایه‌ای بودن آن برای تصمیم‌گیری (ساتون ، ۲۰۰۴). 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 خرده آزمون می‌شود، که به آن آزمون‌های مرکب[1] می‌گویند. نمره‌کل[2] به طور معمول ترکیب خطی از نمره 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 باشند. 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 برعهده دارند. برای استخراج نمره‌کل 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، نمره‌کل ساخته می‌شود. روش‌های متفاوتی برای تبدیل نمره‌های خام به نمره‌های مقیاس وجود دارد و همچنین از طرح‌های وزن‌دهی مختلفی می‌توان برای ساختن نمره‌کل استفاده کرد. در دنیا و همچنین در ایران آزمون‌های مرکب فراوانی تولید و استفاده می‌شوند، 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. هر گاه فرآیندهای ساخت آزمون، نرمال سازی ، و مقیاس‌سازی برای هر کدام از بخش‌های آزمون مرکب به طور مشابه صورت بگیرد، مقایسه نمره‌های آزمون شوندگان در همه 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 که از چهار خرده آزمون زبان و ادبیات انگلیسی، ریاضیات، مهارت‌های خواندن و علوم تشکیل شده است. این خرده آزمون‌ها به صورت چند گزینه‌ای طراحی شده‌اند. مهارت نوشتن هم به عنوان یک خرده آزمون اختیاری و به صورت انشائی در این آزمون گنجانیده شده است. در این آزمون نمره خام هریک از خرده آزمون‌ها به طور مستقیم به مقیاس نمره‌ای بین ۱ تا ۳۶ تبدیل می‌شوند (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 مقیاس‌سازی در این آزمون از تبدیل نرمال[4] استفاده می‌شود. در این تبدیل نمره‌ها برحسب فراوانی ترا کمی، 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. برای محاسبه نمره کل، نمره خرده آزمون¬ها با هم جمع می‌شوند (گزارش فنی کالج برد ، ۲۰۱۶). برای ترکیب کردن 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 نظریه آزمون‌های روانی که در سال ۱۹۵۰ منتشر یافت درباره روش‌های مختلف وزن دهی 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 روی پایایی نمره کل نخواهد داشت، در صورتی که همبستگی بین این خرده آزمون ها کم باشد تأثیر ترکیب آنها از طریق روش های وزن دهی مشهود تر خواهد بود. پی و مالر (۲۰۰۶) طی یک پژوهش شبیه‌سازی شده نتایج گالیکسن را تایید کرده و نشان دادند که برخی عوامل مانند تعداد خرده آزمون‌ها و خواص روان‌سنجی آنها بر روایی و پایایی آزمون مرکب تأثیر خواهند داشت. چانگ (۲۰۰۹) با بررسی پنج طرح وزن دهی مختلف برای تولید نمره کل

---

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

نشان داد، اگر شاخص‌هایی مانند ضریب پایایی را به عنوان شاخص‌های دقت نمره کل در نظر بگیریم، طرح‌های مختلف وزن دهی تأثیر قابل ملاحظه‌ای بر روی آن نداشته و مقدار ضریب پایایی برای همه طرح‌های وزن دهی مقدار بالا و قابل قبولی خواهد بود، اما اگر به لحاظ سهم هر خرده آزمون در تولید نمره کل به عنوان شاخص بنگریم، طرح‌های وزن‌دهی مختلف باعث ایجاد سهم‌های مختلف برای خرده آزمون‌ها در تولید نمره کل خواهند 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 نتیجه این گزارش چنین آمده است که هر روش دارای معایب و مزایای مربوط به خودش است، و هیچ روشی همه 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، روش نرمال‌سازی بیشترین شباهت را به توزیع نمره‌های خام نشان داد. و تبدیل خطی نمره‌ها دارای کمترین گاف در مقیاس نمره‌ها بود. تا اینجا بررسی پیشینه نشان داده که اگر روش‌های نمره کل‌سازی را از بابت ویژگی‌های روانسنجی، خطا و پایایی با هم مقایسه کنیم، روش ترجیحی یافت نمی‌شود و همه روش‌ها در نوع خود کاربردی بوده و ویژگی‌های مثبت و منفی متفاوتی دارند که موجب پذیرش همگانی و یا کنار گذاشتن آن روش نمی‌شود و آزمون‌سازها با توجه به ویژگی‌های آزمون و اهداف آن، از روش‌های متفاوت نمره کل سازی بهره برده‌اند.اما نکته قابل توجه این است که روش‌های نمره‌کل‌سازی می‌توانند بر تصمیم پذیرش افراد تاثیر بگذارند. این نتیجه را چانگ (۲۰۰۹) با بررسی نمودار فراوانی تراکمی نمره‌های کل پنج روش متفاوت نمره‌کل‌سازی نشان داد. شکل ۱ این موضوع را به خوبی نمایش می‌دهد .



شکل۱. تفاوت فراوانی تراکمی روش¬های نمره‌کل سازی (چانگ، ۲۰۰۹)

در نمودار شکل ۱، محور افقی نشان‌دهنده نمره‌کل و محور عمودی نمایش‌دهنده تفاوت بین فراوانی‌تراکمی روش‌های نمره‌کل‌سازی است. روش‌های نمره‌کل‌سازی از A تا E نام‌گذاری شده‌اند و تفاوت بین فراوانی‌تراکمی هرکدام به صورت B-A، C-A و ... نمایش داده شده است. فراوانی‌تراکمی هر نمره نسبت افرادی است که نمره‌ای کوچک‌تر یا مساوی نمره مورد نظر را کسب کرده‌اند را نشان می‌دهد و تفاوت فراوانی تراکمی هر نمره در روش‌های مختلف به این معنی است که افراد در روش‌های متفاوت رتبه‌های متفاوتی کسب کرده‌اند. آنگونه که می‌بینید، برای

[1]. Arcsine Transformation

نقاط میانی و نمره میانگین، روش‌های متفاوت نتوانسته‌اند تفاوت زیادی بین فراوانی تراکمی نمره افراد ایجاد کنند، اما در نقاطی به جز نقاط وسط، روش‌های متفاوت نمره‌کل‌سازی، 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 نمره خاصی قرار می‌گیرند، در روش‌های متفاوت نمره‌کل‌سازی با هم فرق دارند. این تفاوت باعث ایجاد گوناگونی در تصمیم‌پذیرش افراد در روش‌های مختلف نمره‌کل‌سازی می‌شود (چانگ، ۲۰۰۹). روش‌های متفاوت ساخت مقیاس نیز در تصمیم‌پذیرش افراد موثر هستند (چانگ، ۲۰۰۶). این گفته بدان معنی است که هرگاه برای ساخت نمره‌مقیاس 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 مساله اینجاست که با تغییر روش ساخت نمره‌کل، نمره اختصاص داده‌شده به هر شرکت‌کننده و به دنبال آن تصمیم برای پذیرش و یا عدم پذیرش وی در آزمون دچار تغییر می‌شود. مساله 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 رتبه¬بندی کرد؟ هدف عمده این پژوهش ارائه روشی است تا بتوان روش‌های متفاوت نمره کل 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 نمره کل، بررسی دقت نمره‌ها و نحوه استفاده از روش تصمیم‌گیری چند شاخصه به طور کامل شرح داده شده‌اند.

روش نمونه‌گیری و ابزار پژوهش

جامعه 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 سال۱۳۹۵ تعداد ۱۶۲۸۷۹ نفر در آزمون سراسری در رشته ریاضی و فنی شرکت کرده‌اند. از طریق نمونه‌گیری تصادفی، نمونه‌ای از این داوطلبان به منظور بررسی پاسخ‌ها و خرده آزمون‌ها انتخاب شدند. با توجه به اینکه در این پژوهش با مقیاس بزرگ[1] سر و کار داریم از طریق قاعده سر انگشتی[2] می‌توان حجم نمونه نزدیک به ۱۰۰۰۰ نفر را مناسب دانست. در این پژوهش روش‌های مختلف نمره کل 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 شده است. در آزمون سراسری ۱۳۹۵ و در گروه آزمایشی ریاضی و فنی، داده‌های 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 استفاده قرار گرفت .

روش ساختن نمره‌های خام و نمره‌های مقیاسی

برای ساخت نمره‌های خام، به پاسخ درست هر سوال نمره یک و به پاسخ‌های نادرست و سفید نمره صفر تعلق می‌گیرد، و بابت پاسخ نادرست جریمه‌ای صورت نخواهد گرفت، یعنی آزمون نمره منفی نخواهد داشت. نمره خام هر فرد در هر خرده آزمون از رابطه ۱ محاسبه می‌شود:

$$X_i = \sum_{j=1}^{k} u_j \qquad (1)$$

در رابطه ۱، k تعداد سوال‌ها در هر خرده آزمون است، و u_j نمره هر سوال است، که می‌تواند یکی از مقادیر صفر و یا یک را بپذیرد. X_i نمره خام شخص i ام 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 نمره مقیاسی به روش نرمال سازی مراحل تبدیل به شرح زیر است (کولن، ۲۰۱۴):

1. Large-scale assessment
2. Thumbnail rule

مرحله اول: توزیع فراوانی نسبی نمره‌ها محاسبه می‌شود. مرحله دوم: با استفاده از فراوانی نسبی و توزیع تراکمی نمره‌ها، رتبه درصدی نمره‌ها محاسبه می‌گردد. مرحله سوم: نمره z مربوط به هر رتبه درصدی از روی معکوس رابطه ۲ محاسبه می‌شود:

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

رابطه ۲ توزیع ترا کمی نرمال استاندارد است. ω متغیر انتگرال‌گیری است که دامنه آن از $-\infty$ تا z است و $\hat{Q}(y)$ رتبه درصدی است. مرحله چهارم: تبدیل خطی نمرات z بدست آمده در مرحله سوم با استفاده از رابطه ۳ است:

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

Sc نمرات مقیاس بندی 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. مرحله پنجم: تبدیل نمرات بدست آمده در مرحله چهارم به نمرات صحیح است که با $sc_{\mathrm{int}}(y)$ نشان داده می‌شوند. در این روش عدد به دست آمده به نزدیک‌ترین نمره صحیح گرد می‌شود.

روش تبدیل آرک سینوس: در تبدیل نمره‌های خام به نمره‌های مقیاس بندی 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 رابطه ۴، X_i نمره خام (تعداد پاسخ‌های صحیح) و k تعداد سؤال‌های خرده آزمون هستند (کولن، ۲۰۱۴). در این پژوهش به منظور محدود شدن دامنه 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 محدوده مشترکی باشند. در پایان نمره‌های مقیاسی ساخته شده به نزدیک‌ترین عدد صحیح گرد خواهند شد.

طرح‌های وزن‌دهی و ساخت نمره 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 ساخت نمره کل به دست می‌دهد ب) طرح وزن‌دهی موثر براساس واریانس و کواریانس هر آزمون ج) طرح وزن‌دهی شانون. در ادامه به چگونگی محاسبه این وزن‌ها و استفاده آن‌ها در این پژوهش پرداخته شده است.

طرح وزن‌دهی اسمی: در این طرح ، که با A مشخص شده، آزمون ساز بر اساس اهمیتی که هر درس در تشکیل نمره کل خواهد داشت، 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 موثر: در این طرح، که در پژوهش حاضر B نام دارد، از وزن‌های مؤثر برای تشکیل نمره کل استفاده شده است. این وزن‌ها با استفاده از ضرایب اسمی ساخته شده و به واریانس و کواریانس بین خرده آزمون‌ها وابسته هستند. رابطه ۵ وزن مؤثر نسبی را نشان می‌دهد که براساس 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 رابطه ۵، σ_i^2 واریانس نمرات خرده آزمون i ام، 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 می‌شوند (کولن، ۲۰۱۴).

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

$$E = S(P_1, P_2, ..., P_n) = -K \sum_{i=1}^{n} P_i \ln P_i \qquad (۶)$$

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

مرحله اول: ابتدا ماتریس تصمیم تشکیل داده می‌شود. ستون‌های این ماتریس خرده آزمون‌ها و سطرهای آن افراد را تشکیل می‌دهند، 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 نمره مقیاسی شخص i ام در خرده آزمون j ام به صورت S_{ij} تعریف شده باشد.

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

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

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

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

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

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

مرحله چهارم: برای نشان دادن درجه مفید بودن اطلاعاتی که از هر خرده آزمون استخراج می‌شود می‌توان از درجه انحراف[3] استفاده کرد. مقدار درجه انحراف از رابطه $d_j = 1 - E_j$ محاسبه می‌شود، برای هر خرده آزمون یک درجه انحراف محاسبه می‌شود، نزدیکی درجه انحراف دو خرده آزمون به همدیگر به معنی عدم تفاوت افراد بین آن دو خرده آزمون است. پس نقش آن خرده آزمون‌ها در اندازه‌گیری باید به همان اندازه کم شود.

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

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

---

1. Shannon Entropy
2. Information theory
3. Degree of diversification

می‌توان از وزن‌های اسمی 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 رابطه ۱۱ تعریف می‌شود.

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

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

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

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

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 استاندارد اندازه‌گیری کلی (SEM[3]) و هم از خطای استاندارد اندازه‌گیری شرطی (CSEM[4]) استفاده شود. هم اکنون گزارش‌های فنی بیشتر خطای استاندارد اندازه‌گیری را به صورت خطای استاندارد اندازه‌گیری کلی ارایه می‌دهند ولی موسسه های AERA، NCEME و APA از سال ۱۹۸۵ برای استانداردهایی که به منظور تولید آزمون‌های آموزشی و روانی توصیه کرده‌اند، گزارش خطای استاندارد اندازه‌گیری شرطی را به همراه گزارش‌های فنی آزمون همواره توصیه کرده‌اند (استاندارد شماره ۲-۱۰ APA). خطای استاندارد اندازه‌گیری شرطی که در این پژوهش با نماد

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

CSEM نمایش داده شده است، قادر است میزان خطای استاندارد اندازه‌گیری را برای همه سطوح نمرات برآورد کند. این شاخص آماری هم برای نمرات خام و هم برای نمرات مقیاس‌بندی شده و هم برای نمرات ترکیبی قابل محاسبه است (وودروف و همکاران، ۲۰۱۳). بررسی این شاخص آماری نشان می‌دهد که میزان خطای استاندارد اندازه‌گیری برای همه نمرات برابر نیست، و سطوح مختلف نمره‌ها دارای خطای استاندارد اندازه‌گیری شرطی متفاوتی هستند (کولن، ۲۰۱۴). خطای استاندارد اندازه‌گیری شرطی تعریف مشابهی با خطای استاندارد اندازه‌گیری دارد، خطای استاندارد اندازه‌گیری شرطی، واریانس نمره مشاهده شده هر شرکت کننده در طی برگزاری آزمون‌های موازی در شرایط مشابه است، البته با این فرض که نمره حقیقی او ثابت بماند (هارتل، ۲۰۰۶). روش‌های متعددی برای محاسبه خطای استاندارد اندازه‌گیری شرطی پیشنهاد شده است، که در این پژوهش از روش برنان و لی (۱۹۹۹) برای محاسبه این خطا استفاده می‌شود.

روش برنان و لی برای محاسبه خطای استاندارد اندازه‌گیری شرطی که به روش دو جمله‌ای نیز مشهور است از تعمیم دو نظریه، یعنی نظریه نمره حقیقی قوی[1] لرد در سال ۱۹۵۵و ۱۹۵۷ و کولن در سال ۱۹۹۲استفاده 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 ۱۲ محاسبه می‌شود:

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

در رابطه ۱۲ پارامتر π نمره حقیقی نسبت پاسخ‌های صحیح برای هر شخص است. طبق آنچه که در آمار مقدماتی موجود است می‌توان رابطه واریانس شرطی نمره‌های y را به صورت زیر نیز نوشت:

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

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

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

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

$$\sigma_{E}{}^{2}(y\,|x)=\frac{k}{k-1}\left\{\sum y^{2}p(y\,|\pi,k)-\left(\sum yp(y\,|\pi,k)\right)^{2}\right\} \qquad (۱۵)$$

برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس طبق نظریه نمره حقیقی قوی می‌توان در رابطه ۱۵ به جای مقادیر نمره خام y از تبدیل شده غیرخطی آنها (مثلا نمره‌های نرمال) استفاده کرد.

$$\sigma_{E}{}^{2}(s(x)|x)=\frac{k}{k-1}\left\{\sum f(y)^{2}p(y\,|\pi,k)-\left(\sum f(y)p(y\,|\pi,k)\right)^{2}\right\} \qquad (۱۶)$$

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

$$Y_i=\sum_{j=1}^{n}W_jS_j(X_i) \qquad (۱۷)$$

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

$$SEM_{Y_i}=\sqrt{\sum_{j=1}^{m}w_j^{2}\hat{\sigma}_{E(s(x_{ij})|x_{ij})}} \qquad (۱۸)$$

---

1 Strong True Score

در رابطه ۲۰ SEM_{Y_i} خطای استاندارد شرطی برای نمره کل شخص i ام است. m تعداد خرده آزمون‌ها است. برای محاسبه ضریب پایایی نمره‌های مرکب از رابطه ۱۹ استفاده خواهد شد، این رابطه توسط وانگ و استانلی (۱۹۷۰) و کولن (۲۰۰۶) پیشنهاد شده است.

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

روش تصمیم‌گیری چند شاخصه

تحقیق در عملیات، رویکردی علمی می‌باشد که در صدد حل مسائل مدیریتی است و هدف آن کمک به مدیران جهت تصمیم‌گیری بهتر است. تحقیق در عملیات معمولا در قالب عناوینی همچون علم مدیریت، روش‌های کمی، تحلیل کمی، و علم تصمیم‌گیری نیز بیان می‌گردد. در علم مدیریت تصمیم‌گیری نتیجه فرآیند انتخاب یک گزینه بهتر از بین دو یا چند گزینه متفاوت و یا پاسخ مثبت یا منفی به یک موضوع می‌باشد، 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 جهت 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 موثر در تصمیم‌گیری طبقه‌بندی کرد. فنون تصمیم‌گیری بر اساس ماهیت به دو دسته تصمیم‌گیری‌های کیفی(ذهنی) و تصمیم‌گیری‌های کمی (عینی) تقسیم می‌شوند. 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[۱]) ج) تصمیم‌گیری در شرایط فازی د) تصمیم‌گیری در شرایط عدم اطمینان کامل (آذر و رجب زاده، ۱۳۹۳). تقسیم‌بندی دیگری که در این پژوهش مد نظر است طبقه‌بندی تصمیم‌گیری بر اساس معیارهای مورد ارزیابی می‌باشد. این فنون تصمیم‌گیری به دو دسته مهم تصمیم‌گیری تک شاخصه و تصمیم‌گیری چند شاخصه تقسیم می‌شوند. فنون تصمیم‌گیری تک معیاره مجموعه فنونی را می‌گویند که به دنبال ارزیابی راه حل‌های ممکن موجود و انتخاب بهترین راه حل براساس یک معیار ارزیابی هستند. 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 می‌باشند را فنون تصمیم‌گیری چند شاخصه می‌نامند (ایشیزاکا و نمری، ۲۰۱۳). به طور کلی اصول تصمیم‌گیری چند شاخصه[۲] (MCDM) به دو دسته کلی تقسیم می‌شوند: الف) مدل‌های تصمیم‌گیری چند هدفه[۳] (MODM)، ب) مدل‌های تصمیم‌گیری چند شاخصه[۴] (MADM). در مدل‌های چند هدفه، چندین هدف به طور همزمان جهت بهینه شدن، مورد توجه قرار می‌گیرند. مقیاس سنجش برای هر هدف ممکن است با مقیاس سنجش برای هدف دیگر متفاوت باشد. مثلا یک هدف ممکن است حداکثر کردن سود بر حسب واحد پول باشد، ولی هدف دیگر حداقل استفاده 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 مدل‌های تصمیم‌گیری چند شاخصه، گزینه‌های مختلف را بر حسب چند شاخص اولویت‌گذاری یا ارزیابی می‌کنند، فنون تصمیم‌گیری چند شاخصه متفاوت هستند که در این پژوهش از رویکرد مبتنی بر فن تصمیم‌گیری TOPSIS برای رتبه‌بندی روش‌های نمره کل‌سازی استفاده شده است. عبارت TOPSIS مخفف عبارتی به معنای "تکنیک شباهت ترجیح ترتیب با جواب ایده آل[۵]" است. این روش تعداد کمی ورودی از کاربر دریافت می‌کند و خروجی این روش به سادگی قابل فهم است تنها پارامترهای ذهنی این روش وزن‌ها و شاخص‌ها هستند. ایده بنیادین TOPSIS آن است که بهترین جواب، جوابی است که کمترین فاصله را از جواب ایده آل و بیشترین فاصله را از جواب ضد ایده آل 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 تاپسیس است توضیح داده خواهد شد.

---

1 . Bayesian Decision
2 . **M**ultiple **C**riteria **D**ecision **M**aking (MCDM)
3 . **M**ultiple **O**bjective **D**ecision **M**aking (MODM)
4 . **M**ultiple **A**ttribute **D**ecision **M**aking (MADM)
5 . **T**echnique of **O**rder **P**reference **S**imilarity to the **I**deal **S**olution

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

مرحله اول: عملکرد گزینه‌ها در شاخص‌های متفاوت نرمال می‌شوند 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 ایده آل (اگر معیار مدنظر معیار 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.

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

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

مرحله دوم: محاسبه وزن‌ها برای هر شاخص و تشکیل ماتریس نرمال موزون به وسیله ضرب ماتریس وزن‌ها در ماتریس تصمیم.

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

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

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

الف) از بهترین عملکرد گزینه‌ها برای هر شاخص در تشکیل بردار ایده آل یعنی $A^{+} = (v_1^{+}, v_2^{+}, ..., v_m^{+})_{1 \times m}$ و از بدترین عملکرد گزینه ها در هر شاخص برای تشکیل بردار ضد ایده آل یعنی $A^{-} = (v_1^{-}, v_2^{-}, ..., v_m^{-})_{1 \times m}$ 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.

مرحله چهارم: فاصله هر گزینه با گزینه ایده‌آل و ضد ایده‌آل به وسیله فاصله اقلیدسی محاسبه می‌شود.

فاصله هر گزینه تا ایده‌آل (ایده‌آل مثبت)

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

فاصله هر گزینه تا ضد ایده‌آل (ایده‌آل منفی)

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

مرحله پنجم: محاسبه ضریب نسبی نزدیکی برای هر گزینه. از رابطه ۲۴ برای محاسبه 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.

$$c_i = \frac{d_i^-}{d_i^- + d_i^+} \quad (۲۴)$$

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

تصمیم‌گیری چند شاخصه و رتبه‌بندی روش‌های نمره کل‌سازی

در اینجا قصد داریم با تلفیق مفاهیم خطای استاندارد اندازه‌گیری شرطی و نظریه تصمیم با رویکرد روش TOPSIS ، راه کار تازه‌ای برای بررسی کارآمدی نمره‌های کل ارائه شود، به طوری که انتخاب آن روش نمره کل‌سازی نسبت به بقیه کارآمدتر بوده و موجب زیان کمتر به داوطلبان شود. در این رویکرد آن روش نمره کل‌سازی انتخاب می‌شود که در آن نسبت به بقیه روش‌ها برای همه 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. ماتریس تصمیم در اینجا قدری متفاوت با ماتریس تصمیم در روش TOPSIS است، در روش TOPSIS ماتریس تصمیم یک ماتریس شامل نمره‌های خام هر گزینه (اینجا شخص) در هر شاخص (اینجا خرده آزمون) بوده است. اما در این روش ماتریس تصمیم به صورت زیر تعریف می‌شود:

$$CSEM_D = \left[SEM_{Y_{ij}} \right]_{n \times m} \quad (۲۵)$$

همان طور که ملاحظه می‌شود، ماتریس تصمیم ماتریسی است که دارای n سطر و m ستون است.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 j ام است. برای هر شخص در هر روش نمره کل‌سازی یک نمره کل تولید شده است که خطای استاندارد اندازه‌گیری شرطی آن $SEM_{Y_{ij}}$ است. 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، مستقیما ایده‌آل‌ها و ضد ایده‌آل‌ها تعریف می‌شوند.

فرض می‌شود که n شخص هر کدام دارای m نوع نمره کل باشند. پس هر شخص دارای m خطای استاندارد اندازه‌گیری شرطی است. پس از تشکیل ماتریس تصمیم، بردارهای ایده‌آل (ایده‌آل مثبت) و ضد ایده‌آل (ایده‌آل منفی) به این صورت تعریف می‌شوند. **بردار ایده آل (ایده آل مثبت)**، برداری که شامل کمترین خطا در بین همه روش‌ها برای نمره کل هر شخص است، یا به زبان ساده‌تر، هر شخص برای هر روش نمره کل‌سازی دارای یک خطای استاندارد اندازه‌گیری شرطی است، که کمترین آن به عنوان ایده‌آل مثبت در نظر گرفته می‌شود، چون n نفر در آزمون شرکت کرده‌اند بردار حاصل داری یک سطر و n ستون خواهد بود 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 که شامل بیشترین خطا در بین همه روش‌ها برای نمره کل هر شخص است، یا به زبان ساده‌تر، هر شخص برای هر روش نمره کل‌سازی دارای یک خطای استاندارد اندازه‌گیری شرطی است، که بیشترین آن به عنوان ایده‌آل منفی در نظر گرفته می‌شود، چون n نفر در آزمون شرکت کرده‌اند بردار حاصل داری یک سطر و n ستون است 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 فاصله بین ایده‌آل‌های مثبت و منفی تا خطای استاندارد اندازه‌گیری شرطی افراد محاسبه می‌شود.

فاصله هر روش نمره کل سازی تا ایده‌آل مثبت: $$d_j^+ = \left\{ \sum_{i=1}^{n} (SEM_{Y_{ij}} - SEM^+_{Y_j})^2 \right\}^{\frac{1}{2}}$$

فاصله هر روش نمره کل‌سازی 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):

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

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

یافته‌ها

در جدول ۳ نتایج تحلیل داده‌ها را برای نمونه تصادفی به حجم ۱۰۰۰۰ نفر مشاهده می‌کنید. در ابتدا نمره‌های خام به صورت امتیاز یک برای پاسخ صحیح و امتیاز صفر برای پاسخ‌های نادرست و بی‌پاسخ تهیه شده‌اند و سپس با جمع امتیازها نمره‌های خام به دست آمده‌اند. 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 هستند. خطای استاندارد اندازه‌گیری در جدول ۳ به وسیله ضریب پایایی کودر ریچاردسون محاسبه شده و منظور از میانگین نمره‌های خام نسبت در این جدول، میانگین حاصل تقسیم نمره خام هر فرد به تعداد سوال‌های آن آزمون می‌باشد.

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

شاخص	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
میانگین	۷/٦۸ (۰/۳۰)	۵/۲۹ (۰/۲۱)	۹/۰۵ (۰/۳٦)	۵/۹۹ (۰/۲۳)	٤/۷۲ (۰/۰۸)	۵/۱۹ (۰/۱۱)	۳/۰۳ (۰/۰۸)
واریانس	۱۸/٤۳	۲۰/۳۷	۳۱/۸٤	٤۰/۰۰	٤۱/۷۸	٤٦/۵۰	۱۵/۸۵
چولگی	۰/۳۷	۱/۲۰	۰/٤۱	۱/۰۳	۲/٤۳	۱/۹۷	۲/۰۱
کشیدگی	-۰/۱۲	۱/٦۳	-۰/٦۹	۰/۱۳	۷/٦۳	٤/۲۹	۵/۲٦
SEM	۲/۰۵۳	۱/۸٤۳	۲/۲٤۱	۲/۰۵٤	۲/۰۲٤	۲/۰۹۵	۱/٦۱۲
KR20	۰/۷۷۱	۰/۸۳۳	۰/۸٤۲	۰/۸۹٤	۰/۹۰۱	۰/۹۰۵	۰/۸۳٦

آنگونه که گفته شد شش روش برای ساختن نمره کل طراحی شده است، تنوع این روش–ها به خاطر ترکیب روش‌های متفاوت مقیاس‌سازی و طرح‌های وزن دهی است. برای ساختن این نمره کل‌ها از دو روش مقیاس‌سازی و سه طرح وزن دهی استفاده شد. جدول ۴ برخی شاخص‌های آماری به همراه ضریب پایایی هر روش نمره کل‌سازی را نمایش می‌دهد. در اینجا برای محاسبه ضریب پایایی نمره‌های کل از رابطه ۱۹ استفاده شد.

جدول ٤. برخی شاخص‌های آماری به همراه ضریب پایایی در روش‌های نمره کل‌سازی

نام روش	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
NA	۵۰۳٤/٦	۹۳۳٦۹۰	۰/٦۲۰	۳/۰۹۰	۰/۹۰۳
NB	۵۰۳۹/۷	۹٦٦٦٦۰	۰/٦٤۰	۳/۰٤٤	۰/۹۰۳
NC	۵۰۳٤/٦	۹۳۳۷۱۰	۰/٦۲۰	۳/۰۹۰	۰/۹۱٤
AA	۲۰۰۸/٤	۱٤۳٤٤۰	۱/۰٦۲	٤/۳٦۰	۰/۹۲۸
AB	۱۷۲۵/۵	۱٤۷۸۱۰	۱/۱٦٤	٤/۵۷۷	۰/۹۲۸
AC	۲۰۰۸/۵	۱٤۳٤۳۰	۱/۰٦۲	٤/۳۵۹	۰/۹۲۸

جدول ۵ مقدار وزن‌ها را برای هر خرده آزمون و برای هر طرح وزن دهی نمایش داده است. توجه کنید که مقادیر طرح‌های وزن‌دهی A به دلیلی عدم وابستگی وزن‌ها به ویژگی نمره‌ها و ثابت بودن، در جدول ذکر نشده‌اند.

جدول ۵. مقادیر وزن‌ها برای هر خرده آزمون در روش های نمره کل‌سازی

روش	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
NB	۰/۰۲٦٤	۰/۰۲۳٤	۰/۰۵۲۸	۰/۰٤۳۲	۰/۳٤۰	۰/۲۹۳	۰/۲۲۰
NC	۰/۱۰۵۲	۰/۰۵۲٦	۰/۰۷۸۹	۰/۰۵۲٦	۰/۳۱٦	۰/۲۳۷	۰/۱۵۸
AB	۰/۰۲۱۷	۰/۰۲۱۸	۰/۰۵۱۷	۰/۰٤٦۹	۰/۳۳۹	۰/۳۰۰	۰/۲۱۸
AC	۰/۱۰۵۳	۰/۰۵۲٦	۰/۰۷۹۰	۰/۰۵۲٦	۰/۳۱٦	۰/۲۳۷	۰/۱۵۸

طبق آنچه که گفته شد، به دنبال روشی هستیم تا با آن بتوان کارآمدترین روش را که در آن به افراد در تصمیم پذیرش کمترین ضرر وارد می‌شود را انتخاب کرد. در این روش کمترین خطاها به عنوان ایده‌آل مثبت و بیشترین خطاها به عنوان ایده‌آل منفی در نظر گرفته خواهند شد. هر روشی که کمترین فاصله را با ایده‌آل مثبت داشته باشد، روش مطلوب و کارآمدی خواهد بود. جدول ۶ شاخص‌های دوری و نزدیکی (مجاورت) همه شش روش نمره کل‌سازی را نمایش می‌دهد. رتبه‌بندی روش‌ها بر اساس شاخص c انجام گرفته 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 c (شاخص مجاورت)	رتبه بر اساس شاخص مجاورت
NA	۰/۲۷۷۹۲۳	ششم
NB	۰/۱۲٦۷۰۰	نهم
NC	۰/۲۷۷۸۰۸	هفتم
AA	۰/۸٦۷٤۷۸	اول
AB	۰/۷۹۳۹۵۰	سوم
AC	۰/۸٦۷٤۷۲	دوم

با بررسی جدول ۶ مشاهده می‌شود که رتبه‌های اول و دوم متعلق به دو روش نمره کل‌سازی AC و AA است. این دو روش به کمترین خطای استاندارد اندازه‌گیری ایده‌آل نزدیک هستند.

بحث و نتیجه گیری

آنچه گذشت پژوهشی توصیفی با رویکرد کاربردی است که به منظور بررسی روش‌های مختلف ساختن نمره کل در آزمون‌های مرکب انجام شده است. در این پژوهش از نمونه تصادفی ۱۰۰۰۰ نفری شرکت کنندگان آزمون سراسری ایران سال ۱۳۹۵ در رشته ریاضی استفاده شد. خرده آزمون‌های هفت گانه آزمون سراسری ۱۳۹۵ گروه ریاضی فنی شامل: زبان و ادبیات فارسی، زبان و ادبیات عربی، معارف اسلامی، زبان و ادبیات انگلیسی، ریاضیات، فیزیک و شیمی هستند. برای ساختن نمره خام از مجموع پاسخ‌های صحیح استفاده شد و از روش‌های نرمال‌سازی و آرک‌سینوس برای ساختن مقیاس استفاده گردید. به منظور ساختن نمره کل از سه طرح وزن دهی اسمی، موثر و شانون استفاده شده و برای همه نمره‌های ساخته شده خطای استاندارد اندازه‌گیری شرطی محاسبه و به منظور رتبه‌بندی نمره‌های ساخته شده از روش تصمیم‌گیری چندشاخصه استفاده شد. در این پژوهش سعی شد با تلفیق مفاهیم خطای استاندارد اندازه‌گیری شرطی و روش تصمیم‌گیری رویکردی جدید برای بررسی کارآمدی نمره‌های کل ارائه شود.

طرح وزن‌دهی A که به هیچکدام از ویژگی‌های نمره‌ها وابستگی نداشته و مطابق جدول ۱ برای خرده آزمون‌ها ارائه می‌شوند. مقادیر وزنی در طرح‌های وزن‌دهی B و C در جدول ۵ ارائه شده‌اند. از میان این روش‌ها دو روش از طرح B و دو روش هم از طرح C برای وزن‌دهی استفاده کرده‌اند. نتایج نشان می‌دهند که هر کدام از خرده آزمون‌ها که دارای واریانس بزرگ‌تری باشند وزن آنها نیز بزرگ‌تر است، این تفاوت در واریانس‌ها هم به دلیل تعداد سوال‌ها و هم به دلیل وزن اسمی، به وجود آمده است. خرده آزمون‌های ریاضی، فیزیک و شیمی هم به لحاظ تعداد سوال‌ها و هم به لحاظ وزن اسمی از خرده آزمون‌های دیگر بالاتر هستند، و واریانس بزرگ‌تری نیز دارند، 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 شده که مقدار وزن هر کدام از روش‌ها برای یک خرده آزمون مشخص متفاوت باشد. با ضرب هر کدام از وزن‌ها در خرده آزمون مربوط و جمع کردن آنها برای هر فرد یک نمره کل ساخته می‌شود، که طبق آنچه که گفته شد برای هر فرد شش نمره کل متفاوت ساخته می‌شود، جدول ۴ گشتاورهای اول تا چهارم نمره‌های کل را به همراه ضریب پایایی آنها نمایش می‌دهد. با توجه به داده‌های این جدول می‌توان روش‌های نمره کل‌سازی را با توجه به نوع مقیاس به کار رفته در آنها به دو دسته تقسیم کرد. دسته اول آن روش‌هایی هستند که از روش نرمال‌سازی برای ساختن مقیاس استفاده می‌کنند (بدون توجه به طرح وزن دهی) و دسته دوم روش‌هایی که از آرک‌سینوس برای ساختن مقیاس استفاده می کنند (بدون توجه به طرح وزن دهی). در روش‌های نمره کل‌سازی که در آنها از مقیاس نرمال استفاده می‌شود، میانگین نمره‌های کل بین ۵۰۳۴ تا ۵۰۳۹ تغییر پیدا کرده‌اند. تنها طرح‌هایی که داری طرح وزن دهی B بوده‌اند واریانس‌های بزرگ‌تری را ایجاد کرده‌اند و واریانس نمره‌ها برای سایر طرح‌های وزن‌دهی به طور تقریبی تفاوت زیادی نداشته است. مقادیر پایایی در جدول ۴ نشان می‌دهند که هیچ کدام از روش‌های نمره کل‌سازی نتوانسته‌اند تفاوت آشکاری را 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 آرک–سینوس برای نمره کل‌سازی استفاده شده، میانگین نمره‌ها بین ۱۷۰۰ تا ۲۰۰۸ است و مشابه دسته اول، روش‌های شامل طرح وزن‌دهی B واریانس بیشتری در میان نمره‌ها ایجاد کرده‌اند، انواع روش‌های نمره کل‌سازی که شامل مقیاس آرک‌سینوس بوده‌اند نتوانستند تفاوت آشکاری در مقدار ضریب پایایی ایجاد کنند. نکته‌ای که باید در اینجا به آن اشاره کرد تفاوت ضریب پایایی بین روش‌هایی است که شامل مقیاس نرمال بودند و روش‌هایی که شامل آرک سینوس هستند، ضریب پایایی نمره کل‌هایی که با مقیاس آرک‌سینوس ساخته شده‌اند، از مقیاس نرمال بیشتر هستند. 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. که این نتایج حاکی از برتری هر چند اندک روش‌هایی است که از مقیاس آرک‌سینوس برای ساخت نمره کل استفاده می‌کنند.

یافته های پژوهش نشان داد که در میان شش روش برای نمره کل‌سازی روش نمره کل سازی AA با شاخص دوری و نزدیکی (مجاورت) ۰/۸۶۷۴۷۸رتبه اول را کسب کرده است. این به این معنی است که نمره کل‌های ساخته شده در این روش کمترین فاصله را تا خطای کمینه تعیین شده داشته و در صورتی که افراد با استفاده از نمره کل‌های ساخته شده از 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 است، این اختلاف در حد ۰/۰۰۰۰۶ است و می‌توان این دو روش را تقریبا معادل در نظر گرفت. نتایج استفاده از روش تصمیم‌گیری چندشاخصه حاکی از برتری دو روش AC و AA داشته است، البته در رتبه‌بندی این روش‌ها تفاوت زیادی نشان نداده‌اند. این نشان می‌دهد که آزمون‌ساز می‌تواند با توجه به مقتضیات آزمون و شرایط اجرا و تفسیر آزمون، با اطمینان برابر از هرکدام از این روش‌ها بهره ببرد. در پایان می‌توان چنین گفت که طبق روش تصمیم و معیار انتخاب روشی که کمترین خطا را برای همه افراد مرتکب می‌شوند، روش‌های نمره کل‌سازی AC و AA را می‌توان به طور تقریبی معادل دانست و هیچکدام در ساختن نمره کل بر دیگری برتری نشان نداده است.

با توجه به نتایجی که از تحلیل تصمیم به دست آمد به سازندگان آزمون‌های مرکب می‌توان چنین توصیه کرد، که برای رسیدن به کمترین خطای ممکن از یکی از این دو روش AC و AA استفاده کنند، البته چون روش تبدیل آرک‌سینوس به همراه وزن‌های اسمی بیشتر برای آزمون‌هایی که اهمیت و محتوای دروس برای پذیرش اهمیت قابل توجهی دارد کاربرد دارد، به آن منظور از آن استفاده شود، این کاربرد به خاطر ماهیت وزن دهی به شیوه A می‌باشد. اما برای مواردی که پراکندگی نمرات و حداکثر کردن استفاده از آنتروپی اطلاعات برای آزمون‌ساز اهمیت داشته باشد، توصیه می‌شود، روش AC را بر روش AA ترجیح بدهد. تنها شرایط آزمون و تفاسیری که قرار است از آن بشود در اینجا اهمیت دارد زیرا نشان دادیم به لحاظ دقت هر دو روش رتبه یکسانی دارند.

برای ساختن نمره‌کل روش‌های متعددی وجود دارد، به عنوان مثال نمره‌های نرمال‌سازی شده را می‌توان پس یا پیش از تبدیل به مقیاس نرمال هموارسازی کرد، و 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 سایر

روش‌های وزن‌دهی می‌باشد که در این پژوهش به آنها پرداخته نشده‌است. روش تصمیم‌گیری چندشاخصه قادر است همه انواع روش‌های نمره کل‌سازی 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 داده های آزمون سراسری با این پژوهش داشتند، تقدیر و تشکر می نمایم.

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.