فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
سال هشتم، شماره ۲۹، پاییز ۹۶، ص ۹۳ تا ۱۲۰

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 8, No. 29, Autumn 2017

مقایسه خطای استاندارد اندازه‌گیری شرطی در روش‌های غیرخطی تبدیل نمره‌های خام به نمره‌های مقیاس

مجتبی جهانی فر*[۱]، ابراهیم خدایی[۲]، جلیل یونسی[۳]، سید امین موسوی[۴]

تاریخ دریافت: ۹۵/۱۱/۲۸
تاریخ پذیرش: ۹۶/۰۷/۱۵

چکیده

برای تفسیرپذیری بهتر و مقایسه پذیر کردن نمره‌ی آزمون‌ها با همدیگر، نمره‌های خام به‌دست‌آمده از خرده‌آزمون‌ها را به مقیاس مشترکی تبدیل می‌کنند که به آن نمره مقیاس گفته می‌شود. روش‌های متفاوت خطی و غیرخطی برای تبدیل نمره‌های خام به نمره‌های مقیاس وجود دارد. روش‌های متداول غیرخطی تبدیل نمره‌های خام به نمره‌های مقیاس، نرمال‌سازی و تبدیل آرک‌سینوس هستند. در این پژوهش که با هدف مقایسه خطای استاندارد اندازه‌گیری تبدیل نرمال و تبدیل آرک‌سینوس انجام‌گرفته از ۱۰۰۰۰ داده شبیه‌سازی‌شده و ۱۰۰۰۰ داده واقعی آزمون سراسری سال ۱۳۹۵ ایران بهره برده‌ایم. به‌منظور مقایسه این دو روش از نمودارها و شاخص‌های آماری و همچنین ویژگی‌های اندازه‌گیری بر اساس نظریه نمره قوی حقیقی ازجمله خطای استاندارد اندازه‌گیری شرطی استفاده شد. نتایج نشان داد که دو روش تبدیل از ویژگی‌های متفاوتی برخوردار هستند. گرچه نمره‌ها در هر دو روش دارای ضریب پایایی بالایی هستند، اما روش آرک‌سینوس ضمن کاهش نوسان خطا برای سطوح مختلف نمره‌ها، دارای میانگین خطای استاندارد اندازه‌گیری شرطی کمتری نسبت به روش نرمال‌سازی بوده است.

واژه‌های کلیدی: تبدیل آرک‌سینوس، خطای استاندارد اندازه‌گیری شرطی، نرمال‌سازی، نمره‌های مقیاس

۱.* دانشجوی دکتری سنجش آموزش، دانشگاه تهران، تهران، ایران. m.jahanifar@ut.ac.ir
۲. دانشیار گروه روش‌ها و برنامه‌های آموزشی و درسی، دانشگاه تهران، تهران، ایران
۳. دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران.
۴. عضو هیئت‌علمی دانشکده آموزش، دانشگاه ساسکاچوان، کانادا.

مقدمه

هر آزمون مرکب شامل چند خرده‌آزمون است که هرکدام نمره خاصی را به خود اختصاص می‌دهند، برای تفسیرپذیری بهتر و مقایسه‌پذیرکردن نمره‌ی آزمون‌ها، نمره‌های خام[1] را به نمره‌های مقیاس[2] تبدیل می‌کنند. در صورتی می‌توان نمره خرده‌آزمون‌های مختلف را باهم جمع کرد که روی یک مقیاس مشترک باشند. روش‌های مقیاس‌سازی می‌توانند موجب کاهش اطلاعات و یا ایجاد خطاهای تصادفی و نظام‌دار در نمره‌های مقیاس شوند و نمره‌های هر حوزه را تحت تأثیر قرار دهند. یکی از قدیمی‌ترین روش‌های پیوند زدن[3] نمره‌ها مقیاس‌سازی نمره‌ها است. مقیاس‌سازی همان تبدیل نمره‌های خام به نمره‌های مقیاس است، در روان‌سنجی همواره به مقیاس‌های متفاوت نمره‌ها نیاز بوده به همین خاطر قدمت این روش به قدمت علم روان‌سنجی است. هدف از مقیاس‌سازی نمره‌ها تبدیل نمره‌ها از دو آزمون مختلف به مقیاس مشترک[4] است. این روش پیوند غیرمستقیم بین نمره‌های دو آزمون ایجاد می‌کند (دورانز، پامریچ و هولاند، ۲۰۰۷). در این روش نمره‌های هر آزمون به‌طور جداگانه به مقیاس مشترک برده می‌شوند. هرگاه دو یا چند آزمون که سازه‌های متفاوتی را اندازه می‌گیرند با هم روی یک جامعه‌ی مشترک از شرکت‌کنندگان اجرا شوند، مقیاس‌سازی بر روی هریک از آزمون‌ها به‌منظور داشتن مقیاس مشترک انجام خواهد گرفت. کولن (۲۰۱۴) این آزمون‌ها را آزمون مرکب[5] نامید. نکته مهم در اینجاست که بردن نمره‌ها به مقیاس مشترک به معنی هم‌تراز کردن آن‌ها نیست و نمی‌توان نمره‌ی آزمون با ساختار مشخص را با آزمون دیگر با ساختار متفاوت تبدیل کرد و مقیاس مشترک تنها آن‌ها را مقایسه پذیر می‌کند نه این‌که راهی را فراهم کند که بتوان آن‌ها را به‌جای هم به کار برد. برای هر آزمون مانند Y که بخواهد مقیاس‌سازی شود، تابع توزیع تراکمی نمره‌ها وجود دارد که روی جامعه مرجع P به‌صورت رابطه (۱) تعریف می‌شود.

1 .raw scores
2. scale scores
3. linking
4. common scale
5. test battery

$$F_{YP}(y) = \Pr\left(Y \le y \middle| P\right) \quad (۱)$$

در رابطه (۱) $F_{YP}(y)$ تابع توزیع تراکمی و $\Pr\left(Y \le y \middle| P\right)$ تابع توزیع احتمال هستند.

نمره‌های آزمون Y به‌وسیله تبدیل S به نمره‌های مقیاس تبدیل می‌شوند.

$$S = S\left(F_{YP}(y)\right) \quad (۲)$$

در رابطه‌ی (۲)، S تابع تبدیل به مقیاس است که می‌تواند خطی و یا غیرخطی باشد. از مزیت‌های این روش، تبدیل نمره خرده‌آزمون‌های یک آزمون مرکب به مقیاسی است که تفسیرپذیری و مقایسه آن‌ها را با همدیگر آسان‌تر می‌کند. نمره بالا در مقیاس مشترک برای یک آزمودنی در یک خرده‌آزمون به معنی عملکرد بهتر وی در آن خرده‌آزمون نسبت به سایر خرده‌آزمون‌ها است (دورانز و همکاران، ۲۰۰۷). تبدیل‌های مقیاس خطی[1]، تبدیل مقیاس نرمال[2] و تبدیل آرک‌سینوس[3] از مشهورترین روش‌های تبدیل مقیاس است، دو روش آخر را روش‌های تبدیل غیرخطی نمره‌های خام به نمره‌های مقیاس می‌نامند.

در روش معمول نرمال‌سازی ابتدا با استفاده از توزیع فراوانی نمره‌های کسب‌شده توسط تمامی شرکت‌کنندگان در آزمون، فراوانی تراکمی و تراکمی نسبی هر نمره محاسبه‌شده و سپس رتبه درصدی برای هر نمره مشخص می‌گردد. با استفاده از رتبه درصدی، نمره z متناظر با رتبه درصدی را از معکوس تابع توزیع تراکمی نرمال مطابق با رابطه (۳) محاسبه می‌کنند.

$$\Phi(z) = \frac{\hat{Q}(y)}{۱۰۰} = \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{z} e^{-\omega^{۲}/۲} d\omega \quad (۳)$$

تابع رابطه (۳) توزیع تراکمی نرمال استاندارد است. ω متغیر انتگرال‌گیری است که دامنه آن از $-\infty$ تا z است و $\hat{Q}(y)$ همان رتبه درصدی است (آنگاف، ۱۹۷۱ و کولن، ۲۰۱۴). به‌عنوان نمونه آزمون SAT[4] برای تبدیل نمره‌های خام به نمره‌های مقیاس از تبدیل

1. linear transformation (LT)
2. normalizing transformation (NT)
3. arcsine transformation (AT)
4. scholastic aptitude test (SAT)

نرمال بهره می‌برد. این آزمون که سه ساعت و چهل‌وپنج دقیقه به طول می‌انجامد شامل خرده‌آزمون‌های مهارت‌های خواندن، نوشتن و همچنین ریاضیات است. در این تبدیل نمره‌های فرمولی[1] برحسب فراوانی تراکمی، به مقیاس نرمال برده می‌شوند و حاصل این مقیاس‌سازی جدول تبدیلی است که در آن هر نمره بر اساس رتبه و نمره درصدی به مقیاس آزمون SAT برده می‌شود (گزارش فنی SAT، ۲۰۱۵). آزمون IBTS که در دانشگاه آیووا به‌منظور اندازه‌گیری پیشرفت تحصیلی از سطح مهدکودک تا پایه دوازدهم طراحی شده است، دارای ۱۵ خرده‌آزمون است. البته این خرده‌آزمون‌ها برای پایه‌های نهم تا دوازدهم شامل نه خرده‌آزمون است. این خرده‌آزمون‌ها شامل مهارت‌های خواندن و نوشتن، مهارت‌های دستور زبان، مهارت‌های شنیداری، مهارت‌های محاسباتی و ریاضیات، علوم و مهارت‌های اجتماعی است؛ که هرکدام هم در تعداد سؤال‌های خرده‌آزمون و هم در زمان پاسخگویی متفاوت هستند. در این آزمون نیز از روش تبدیل نرمال برای ساختن مقیاس نمره‌ها استفاده شده است. (گزارش فنی IBTS، ۲۰۱۶). آزمون سراسری ورود به دانشگاه‌های ایران نیز آزمونی مرکب است که با توجه به رشته تحصیلی دارای خرده‌آزمون‌های متفاوت است، در این آزمون نیز پس از محاسبه نمره خام، نمره‌ها به روش تبدیل نرمال به مقیاس مشترک برده می‌شوند. نمره‌های مقیاس در آزمون سراسری ایران به نمره‌های تراز شهرت دارند (نقی زاده، ۱۳۹۴)

روش دیگر تبدیل خطی نمره‌های خام به نمره‌های مقیاس روش تبدیل آرک‌سینوس است. در تبدیل نمره‌های خام به نمره‌های مقیاس با استفاده از رابطه تبدیل آرک‌سینوس از تعداد پاسخ‌های درست به هر خرده‌آزمون و تعداد سؤال‌های خرده‌آزمون برای ساختن نمره‌ی مقیاس استفاده می‌شود (کولن، ۲۰۱۴).

$$S(X_i) = \frac{1}{2}\left\{\sin^{-1}\sqrt{\frac{X_i}{k+1}} + \sin^{-1}\sqrt{\frac{X_i+1}{k+1}}\right\} \quad (۴)$$

در رابطه (۴)، X_i نمره‌ی خام (تعداد پاسخ‌های صحیح) و k تعداد سؤال‌های خرده‌آزمون هستند. این روش مقیاس‌سازی، نمرات خام با استفاده از تابع تبدیل نمرات

1. formula score

آرک‌سینوس به‌وسیله کولن و هانسون در سال ۱۹۸۹ پیشنهاد شده است. کولن و هانسون (۱۹۸۹) در این پژوهش نشان دادند که در صورت استفاده از این مقیاس خطای نمره‌ها با هم تا حدود زیادی همسان خواهند شد.

آزمون ACT[1]. از چهار خرده‌آزمون زبان و ادبیات انگلیسی، ریاضیات، مهارت‌های خواندن و علوم تشکیل شده است. همگی این خرده‌آزمون‌ها به‌صورت چندگزینه‌ای طراحی شده‌اند. مهارت نوشتن هم به‌عنوان یک خرده‌آزمون اختیاری و به‌صورت انشائی در این آزمون گنجانیده شده است. در این آزمون نمره خام هریک از خرده‌آزمون‌ها به‌طور مستقیم به مقیاس نمره‌ای آرک‌سینوس بین ۱ تا ۳۶ تبدیل می‌شوند (راهنمای ACT، ۲۰۱۴).

برای ساختن نمره مقیاس از روی نمره‌های خام روش‌های متفاوتی وجود دارد به‌گونه‌ای که موسسه‌های مختلف تولید آزمون از مقیاس‌های مختلفی برای تبدیل نمره‌ها بهره می‌برند، اما به‌طورقطع و یقین نمی‌توان اظهار داشت که کدام مقیاس بهترین است. به همین خاطر انتخاب بهترین مقیاس کار دشواری است. مفید بودن هر مقیاس به ویژگی‌هایی است که آن مقیاس برای معنادار بودن نمره‌ها ایجاد کرده و همچنین تمهیداتی است که برای کاهش کج‌فهمی‌ها در تفسیر نمره‌ها ایجاد می‌کند (کولن، ۲۰۱۴).

شون ون چانگ[2] سه روش تبدیل مقیاس یعنی خطی، نرمال‌سازی و تبدیل آرک‌سینوس را با هم مقایسه کرده است. در این مقایسه، سه روش مقیاس‌سازی در شاخصه‌هایی مانند ضریب پایایی نمره‌های مقیاس و تعداد تغییر نمره‌ها در اثر برش نمره‌ها[3] و همچنین شکاف‌های[4] ایجادشده در مقیاس با هم مقایسه شده‌اند. درنتیجه این گزارش چنین آمده است که هر روش دارای معایب و مزایای مربوط به خودش است و هیچ روش همه ویژگی‌های مطلوب را دارا نیست، تصمیم برای انتخاب مقیاس مناسب بستگی هم به خواص اندازه‌گیری و هم به سهولت تفسیر خواهد داشت (چانگ، ۲۰۰۶). در پژوهش چانگ ضریب پایایی خرده‌آزمون‌ها پس از انجام سه نوع تبدیل به هم نزدیک بودند. درحالی‌که در ایجاد

1. american college test
2. Shun-Wen Chang
3. truncation
4. gaps

شکاف‌های بین نمره‌ها روش تبدیل آرک‌سینوس از دو روش دیگر پیشی گرفته است ولی در نمودار خطای استاندارد شرطی این روش از دو روش دیگر دارای خطای کمتری است و روش نرمال‌سازی بیشترین شباهت را به توزیع نمره‌های خام نشان داده است؛ و تبدیل خطی نمره‌ها کمترین شکاف را در مقیاس نمره‌ها نشان داده است. با توجه به منابع و بررسی‌های متفاوت نمی‌توان قطعاً گفت که کدام روش نسبت به سایر روش‌ها ترجیح بیشتری دارد، هرکدام از تبدیل‌ها دارای ویژگی‌های مخصوص به خود هستند و مفید بودن هر روش مقیاس‌سازی بستگی به مفید بودن و تفسیرپذیری بهتر آن داشته و هر روش کاربردهای متفاوتی پیدا کرده است.

علاوه بر پژوهش چانگ (۲۰۰۶) پژوهش‌های کولن و هانسون (۱۹۸۹) کولن و برنان (۲۰۱۴) پترسن، کولن و هاور (۱۹۸۹) نشان دادند که به‌طورقطع و یقین نمی‌توان اظهار داشت که کدام مقیاس بهترین است. به همین خاطر انتخاب بهترین مقیاس کار دشواری است. مفید بودن هر مقیاس به ویژگی‌هایی است که آن مقیاس برای معنادار بودن نمره‌ها ایجاد کرده و همچنین تمهیداتی است که برای کاهش کج‌فهمی‌ها در تفسیر نمره‌ها ایجاد می‌کند بستگی دارد ضمن اینکه بررسی‌ها نشان دادند نمره‌های مقیاس تفاوت آشکاری در ضریب پایایی ندارند، به‌طوری‌که اگر بخواهیم جمع‌بندی از مقایسه نتایج پژوهش‌های مختلف داشته باشیم، این مقایسه را می‌توان چنین خلاصه کرد که این پژوهش‌ها به این نتیجه رسیده‌اند که هیچ برتری شاخصی در انتخاب روش‌های ساختن نمره مقیاس با معیار ضریب پایایی وجود نداشته و ضریب پایایی به شکلی که در این پژوهش‌ها محاسبه شده است نمی‌تواند ملاک مناسبی برای بررسی روش‌های ساخت نمره مقیاس باشد. لذا در این پژوهش به‌منظور مقایسه روش‌های متفاوت ساختن نمره مقیاس به دنبال رویکرد خطای استاندارد اندازه‌گیری شرطی رفته‌ایم.

این پژوهش با هدف بررسی ویژگی‌های اندازه‌گیری در روش‌های تبدیل مقیاس نرمال و تبدیل آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس و همچنین مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌ها بین این دو روش انجام گرفته است و سؤال اصلی آن

این است که این دو روش تبدیل در مقدار خطای استاندارد اندازه‌گیری شرطی چه تفاوتی با هم دارند.

روش پژوهش

پژوهش حاضر پژوهشی کمی و از نوع توصیفی (غیرآزمایشی) است و به‌منظور توسعه دانش کاربردی در زمینه ساختن نمره مقیاس انجام گرفته که با این رویکرد، پژوهشی کاربردی محسوب می‌شود. جامعه موردنظر در این پژوهش داوطلبان شرکت‌کننده در آزمون سراسری سال ۱۳۹۵ در گروه آزمایشی ریاضی و فنی هستند، در سال ۱۳۹۵ تعداد ۱۶۲۸۷۹ نفر در آزمون سراسری در رشته ریاضی و فنی شرکت کرده‌اند. در این پژوهش روش‌های مختلف مقیاس‌سازی و تحلیل‌ها بر روی ۱۰۰۰۰ داده شبیه‌سازی‌شده‌ی خرده‌آزمون‌های مختلف عمومی و اختصاصی آزمون سراسری ایران در گروه آزمایشی رشته ریاضی و فنی اجرا شده است. برای تولید داده‌های شبیه‌سازی‌شده از ویژگی هرکدام از خرده‌آزمون‌ها ازجمله تعداد گزینه‌ها، تعداد سؤال‌ها و همچنین دشواری آن‌ها استفاده شده است. هدف از تولید داده‌های شبیه‌سازی در این پژوهش (۱) بررسی اولیه ویژگی‌های آزمون مانند خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس، ضریب پایایی (۲) مبنایی برای بررسی صحت محاسبات و تحلیل داده‌های واقعی است. ازآنجاکه داده‌های واقعی جهت تعیین پارامترهای شبیه‌سازی استفاده شد (به‌منظور تولید داده‌هایی، هرچه بیشتر؛ مبتنی بر شرایط واقعی آزمون)، می‌توان گفت که مطالعه شبیه‌سازی بنیانی برای بررسی نتایج حاصل از تحلیل داده‌های واقعی فراهم می‌نماید که نشأت گرفته از شرایط واقعی آزمون است. مبنای شبیه‌سازی داده‌ها میانگین دشواری سؤال‌های آزمون است، به‌گونه‌ای که میانگین ضریب دشواری را به‌عنوان میانگین نسبی $\bar{P} = \frac{M_x}{K}$ (میانگین تقسیم‌بر تعداد سؤال‌ها) در نظر گرفته سپس بر اساس تعداد گزینه‌ها و تعداد سؤال‌های آزمون داده‌ها تولید شد (بروکس و یوهانسن[1]، ۲۰۰۳). به‌عنوان‌مثال هرگاه بخواهیم ۳۰ سؤال دو ارزشی و چهارگزینه‌ای با ضریب دشواری ۰/۵۰

1. Brooks & Johansson

در سطح آزمون ایجاد کنیم، داده‌هایی تولید خواهد شد که دارای میانگین و واریانس تقریبی ۱۵ (حاصل‌ضرب ۰/۵ در ۳۰) و ۷/۵ (حاصل‌ضرب ۰/۲۵ در ۳۰) خواهند بود. به‌منظور تأیید نتایج شبیه‌سازی و همچنین مبنایی برای تولید داده‌های شبیه‌سازی از طریق نمونه‌گیری تصادفی، نمونه‌ای از داوطلبان آزمون سراسری سال ۹۵ برای بررسی پاسخ‌ها و خرده‌آزمون‌ها انتخاب شدند. با توجه به اینکه در این پژوهش با مقیاس بزرگ[1] سروکار داریم از طریق قاعده سرانگشتی[2] می‌توان حجم نمونه نزدیک به ۱۰۰۰۰ نفر را مناسب دانست. در آزمون سراسری و در گروه آزمایشی ریاضی و فنی 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)، مورداستفاده قرار می‌گیرد.

ابزار اصلی گردآوری داده در این پژوهش همان سؤال‌های آزمون سراسری است، داده‌های این آزمون در اختیار سازمان سنجش آموزش کشور بوده و به‌منظور تحلیل آزمون ۱۳۹۵ گروه آزمایشی ریاضی و فنی از آن‌ها بهره گرفته شده است. ضمناً به‌منظور انجام مطالعه شبیه‌سازی از داده‌های شبیه‌سازی‌شده نیز در کنار آن استفاده می‌شود. برای تولید داده‌های شبیه‌سازی‌شده از نرم‌افزارTAP[3] نسخه ۴، ۷، ۱۴ استفاده شده است (بروکس و یوهانسن، ۲۰۱۴). برای برخی از روش‌ها و تبدیل‌هایی که در این پژوهش از آن‌ها استفاده می‌شود، نرم‌افزار تجاری به بازار عرضه نشده است به همین منظور در طی انجام پژوهش از نرم‌افزار کد نویسی ریاضی و آمارMATLAB[4] برای نوشتن کدهای مربوط به برخی روش‌ها و تبدیل‌ها استفاده شد.

الف) تبدیل نمره‌های خام به نمره‌های مقیاس به روش نرمال‌سازی: برای به دست آوردن نمره مقیاس در این پژوهش به روش نرمال از رویکرد آنگاف (۱۹۷۱) و کولن (۲۰۱۴)

1. large-scale assessment
2. thumbnail rule
3. test analysis program
4. MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language.

استفاده شده است. در این روش پس از تبدیل نمره‌ها به نمره z (رابطه ۳) از تبدیل خطی رابطه (۵) برای تبدیل خطی و محدود کردن دامنه نمره‌ها استفاده شد.

$$sc = \sigma(sc)\,z + \mu(sc) \quad (۵)$$

نمره‌های مقیاس sc[1] هستند. در این پژوهش میانگین تبدیل خطی ۵۰۰۰ و انحراف استاندارد آن ۱۲۵۰ است. هدف از انتخاب این اعداد محدود کردن نمره‌های مقیاس بین اعداد ۰ تا ۱۰۰۰۰ است. این مقادیر اختیاری است و تنها نمره‌ها را در بازه مشخصی نگه می‌دارد، مقادیر میانگین و انحراف استاندارد در تبدیل خطی، نمره‌ها را در طول محور اعداد جابجا و یا در بازه مشخصی محدود می‌کنند. به‌طور مثال: $-۴ < z < +۴ \Rightarrow ۰ < ۱۲۵۰ \times z + ۵۰۰۰ < ۱۰۰۰۰ \Rightarrow ۰ < Sc < ۱۰۰۰۰$. پس از تبدیل خطی نمره‌ها به نزدیک‌ترین عدد صحیح گرد می‌شوند.

ب) تبدیل نمره‌های خام به نمره‌های مقیاس به روش آرک‌سینوس: برای تبدیل مقیاس به روش آرک‌سینوس از رابطه پیشنهادی کولن (۱۹۸۹) و کولن (۲۰۱۴) استفاده شد. در این پژوهش به‌منظور محدود شدن دامنه نمره‌های مقیاس آرک‌سینوس بین ۰ تا ۱۰۰۰۰ از تبدیل خطی $sc(y) = \alpha \times S(X_i) + \beta$ استفاده شد. در این تبدیل خطی $\alpha = ٦٦٩٠$ و $\beta = -٤٦٧$ هستند. $S(X_i)$ تبدیل آرک‌سینوس رابطه (۴) است. این تبدیل دامنه نمره‌های مقیاس را در فاصله ۰ تا ۱۰۰۰۰ محدود می‌کند. در اینجا نیز نمره‌های مقیاس ساخته شده به نزدیک‌ترین عدد صحیح گرد خواهند شد.

روش محاسبه خطای استاندارد اندازه‌گیری شرطی. هم‌اکنون گزارش‌های فنی بیشتر خطای استاندارد اندازه‌گیری را به‌صورت خطای استاندارد اندازه‌گیری کلی[2] ارائه می‌دهند ولی موسسه‌های AERA[3] و NCEME[4] و APA[5] از سال ۱۹۸۵ برای استانداردهایی که به‌منظور تولید آزمون‌های آموزشی و روانی توصیه کرده‌اند، گزارش خطای استاندارد اندازه‌گیری

1. scale scores
2. overall standard error of measurement
3. american educational research association
4. national council on measurement in education
5. american psychological association

شرطی (استاندارد شماره ۲-۱۰ APA) را به همراه گزارش‌های فنی آزمون توصیه کرده‌اند (AERA، ۲۰۱۴). خطای استاندارد اندازه‌گیری شرطی قادر است میزان خطای استاندارد اندازه‌گیری را برای همه سطوح نمره‌ها برآورد کند. این شاخص آماری هم برای نمره‌های خام و هم برای نمره‌های مقیاس و هم برای نمره‌های مرکب قابل‌محاسبه است (وودروف و دیگران، ۲۰۱۳).

بررسی این شاخص آماری نشان می‌دهد که میزان خطای استاندارد اندازه‌گیری برای همه نمره‌ها برابر نیست و سطوح مختلف نمره‌ها دارای خطای استاندارد اندازه‌گیری شرطی متفاوت هستند. خطای استاندارد اندازه‌گیری شرطی تعریف مشابهی با خطای استاندارد اندازه‌گیری دارد، خطای استاندارد اندازه‌گیری شرطی، واریانس نمره‌ی مشاهده‌شده هر شرکت‌کننده در طی برگزاری آزمون‌های موازی در شرایط مشابه است، البته با این فرض که نمره حقیقی او ثابت بماند (هارتل، ۲۰۰۶).

رویکردهای متفاوتی توسط کولن (۱۹۹۲)، فلت و کوالس[1] (۱۹۹۶) و برنان و لی[2] (۱۹۹۹) برای محاسبه CSEM پیشنهاد شده است که به دلیل سهولت انجام محاسبه و عدم نیاز روش برنان و لی (۱۹۹۹) به محاسبه خطای استاندارد اندازه‌گیری نمره‌های خام از آن برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس در این پژوهش استفاده شده است. روش برنان و لی برای محاسبه خطای استاندارد اندازه‌گیری شرطی که به روش دوجمله‌ای نیز مشهور است از تعمیم دو روش یعنی نظریه نمره حقیقی قوی لرد (۱۹۶۵) و کولن (۱۹۹۲) استفاده می‌کند و رابطه‌ای را برای خطای استاندارد اندازه‌گیری شرطی ارائه می‌دهد (برنان و لی، ۱۹۹۹). بر اساس نظریه نمره حقیقی قوی[3]، احتمال شرطی اینکه شخصی از مجموع k سؤال در یک آزمون بتواند به y تا از آن‌ها پاسخ صحیح بدهد از رابطه (۶) محاسبه می‌شود:

1. Feldt & Qualls
2. Brenan & Lee
3. strong true score theory

$$p(y\,|\,\pi,k)=\binom{k}{y}\pi^{y}\,(1-\pi)^{y-k} \quad (۶)$$

در رابطه (۶) پارامتر π نمره حقیقی نسبت پاسخ‌های صحیح برای هر شخص است و y متغیر تابع احتمال است. لرد (۱۹۶۵) مقدار خطای اندازه‌گیری را با توجه به توزیع دوجمله‌ای رابطه (۶) برای شخصی که می‌تواند به x سؤال پاسخ صحیح بدهد (x نمره‌ی خام است) طبق رابطه (۷) محاسبه کرده است:

$$\widehat{\sigma}_{E(x)}=\sqrt{\frac{x\,(k-x)}{k-1}}=c_k\sqrt{\frac{x\,(k-x)}{k}} \quad (۷)$$

$c_k=\sqrt{\frac{k}{k-1}}$ همان عامل تصحیح سوگیری برآورد است چون در رابطه (۶) به‌منظور برآورد π از مقدار $\bar{x}=\frac{x}{k}$ استفاده می‌شود. عبارت $c_k=\sqrt{\frac{k}{k-1}}$ باعث می‌شود که $\widehat{\sigma}_{E(x)}$ برآوردگر ناریبی از $\sigma_{E(x)}$ باشد. از همین ایده می‌توان برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره y بهره گرفت، در رابطه (۱۲) عبارت زیر رادیکال واریانس توزیع دوجمله‌ای است، پس می‌توان گفت که عامل تصحیح سوگیری برآورد یعنی c_kدر واریانس شرطی توزیع دوجمله‌ای، یعنی در عبارت:

$$\sqrt{\frac{x\,(k-x)}{k}}=\sqrt{\frac{x}{k}\,k\,(1-\frac{x}{k})}=\sqrt{k\bar{x}\,(1-\bar{x})}$$

ضرب شده است. طبق آنچه در آمار مقدماتی موجود است می‌توان رابطه واریانس شرطی نمره‌های y را به‌صورت زیر نیز نوشت (مود و همکاران، ۲۰۰۸):

$$\sigma^{2}(Y\,|\,X)=E\,((Y-E\,(Y\,|\,X))^{2}\,|\,X) \quad (۸)$$

اگر توزیع شرطی نمره‌ها به‌صورت $p(Y\,|\,X)$ تعریف شده باشد رابطه (۸) به‌صورت رابطه (۹) درخواهد آمد:

$$\sigma^{۲}(Y\,|X\,)=\sum_{y=۰}^{k}\Big((\mathrm{Y}-\mu(Y\,|\mathrm{X}))^{۲}\,p(Y\,|\mathrm{X})\Big) \quad (۹)$$

رابطه (۹) ساده‌تر هم می‌شود به‌گونه‌ای که به میانگین شرطی نیازی نباشد:

$$\sigma^{۲}(Y\,|X\,)=\sum_{y=۰}^{k}Y^{۲}p(Y\,|X\,)-\left(\sum_{y=۰}^{k}Yp(Y\,|X\,)\right)^{۲} \quad (۱۰)$$

خطای استاندارد اندازه‌گیری شرطی نیز واریانس شرطی خطاها به‌شرط هر نمره هستند پس به کمک رابطه (۱۱) خطای استاندارد اندازه‌گیری شرطی قابل‌محاسبه است:

$$\sigma_{E}^{۲}(y\,|x\,)=\frac{k}{k-۱}\left(\sum_{y=۰}^{k}y^{۲}p(y\,|\pi,k\,)-\left(\sum_{y=۰}^{k}yp(y\,|\pi,k\,)\right)^{۲}\right) \quad (۱۱)$$

برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس طبق نظریه نمره حقیقی قوی می‌توان در رابطه (۱۱) به‌جای مقادیر نمره خام y از تبدیل‌شده‌ی غیرخطی آن‌ها (مثلاً نمره‌های نرمال) استفاده کرد:

$$\sigma_{E}^{۲}(s(x)|x\,)=\frac{k}{k-۱}\left(\sum_{y=۰}^{k}f\,(y\,)^{۲}p(y\,|\pi,k\,)-\left(\sum_{y=۰}^{k}f\,(y\,)p(y\,|\pi,k\,)\right)^{۲}\right) \quad (۱۲)$$

برای مقایسه روش‌های مختلف مقیاس‌سازی از میانگین خطای استاندارد اندازه‌گیری شرطی که به شکل رابطه (۱۳) استفاده می‌شود.

$$\overline{\widehat{\sigma}}_{E\,(\mathrm{s(x)|x})}^{۲}=f\,(x\,)\widehat{\sigma}_{E\,(\mathrm{s(x)|x})}^{۲} \quad (۱۳)$$

در رابطه (۱۳)، $f\,(x\,)$ فراوانی نسبی نمره x است $s(x\,)$ همان نمره‌های مقیاس و $\widehat{\sigma}_{E\,(\mathrm{s(x)|x})}$ خطای استاندارد اندازه‌گیری شرطی برای هر نمره مقیاس است (کولن، ۲۰۰۴).

یافته‌ها

در جدول ۱ نتایج تحلیل داده‌های شبیه‌سازی‌شده را برای نمونه تصادفی به حجم ۱۰۰۰۰ نفر مشاهده می‌کنید. این داده‌ها که به‌وسیله نرم‌افزار TAP طراحی شده‌اند به‌صورت امتیاز یک برای پاسخ صحیح و امتیاز صفر برای پاسخ‌های غلط و بی‌پاسخ تهیه شده‌اند و سپس با جمع امتیازها نمره‌های خام به دست آمده است. این ویژگی‌ها شامل شاخص‌های آماری مانند میانگین حسابی، واریانس، چولگی و کشیدگی و شاخص‌های اندازه‌گیری مانند خطای استاندارد اندازه‌گیری و ضریب پایایی کودر ریچاردسون ۲۰[1] است. مقادیر داخل پرانتز میانگین نسبی هستند. خطای استاندارد اندازه‌گیری در جدول ۱ به‌وسیله ضریب پایایی کودر ریچاردسون محاسبه شده و منظور از میانگین نسبی نمره‌های خام در این جدول حاصل تقسیم میانگین نمره‌های خام به تعداد سؤال‌های آن آزمون است.

جدول ۱. شاخص‌های آماری و شاخص‌های اندازه‌گیری برای نمره‌ی خام خرده آزمون‌ها (شبیه‌سازی)

شاخص	خرده آزمون‌ها						
	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
میانگین	۱۲/۶۵ (۰/۵۰)	۹/۹۴ (۰/۳۹)	۱۱/۱۳ (۰/۴۴)	۹/۰۴ (۰/۳۶)	۲۲/۲۳ (۰/۴)	۱۶/۵۵ (۰/۳۶)	۱۳/۵۲ (۰/۳۸)
واریانس	۲۰/۹۵	۲۲/۳۳	۲۰/۵۳	۱۹/۰۳	۹۲/۹۸	۵۶/۳۲	۳۴/۱۵
چولگی	-۰/۰۳۱	۰/۲۹	۰/۱۴	۰/۳۲	۰/۲۹	۰/۳۸	۰/۲۹
کشیدگی	-۰/۵۳	-۰/۴۵	-۰/۴۹	-۰/۳۹	-۰/۴۵	-۰/۳۵	-۰/۴۲
SEM	۲/۱۴۹	۲/۱۸۹	۲/۱۷۳	۲/۱۶۰	۳/۲۵	۲/۹۲	۲/۵۵
KR20	۰/۷۲۸	۰/۷۲۴	۰/۷۷۱	۰/۷۱۴	۰/۸۶۴	۰/۸۴۰	۰/۸۷۷

جدول ۲ شاخص‌های آماری و برخی از ویژگی‌های اندازه‌گیری را برای ۱۰۰۰۰ داده واقعی که به‌صورت تصادفی از میان داوطلبان رشته ریاضی شرکت‌کننده در آزمون سراسری در سال ۱۳۹۵ انتخاب شده‌اند را نمایش می‌دهد.

1. KR20

جدول ۲. شاخص‌های آماری و شاخص‌های اندازه‌گیری
برای نمره‌ی خام خرده آزمون‌ها (داده‌های واقعی)

شاخص	خرده آزمون‌ها						
	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

کمترین نمره خام برای دروس عمومی صفر و بیشترین نمره خام برای آن‌ها ۲۵ است. کمترین نمره خام برای خرده‌آزمون ریاضیات صفر و بیشترین آن ۵۵ است. همچنین کمترین نمره خام برای خرده‌آزمون‌های فیزیک و شیمی صفر و بیشترین نمره خام برای آن‌ها به ترتیب ۴۵ و ۳۵ است. از ضریب کودر ریچاردسون برای محاسبه پایایی نمره‌های خام استفاده گردید که طبق داده‌های جدول‌های ۱ و ۲ ضریب پایایی نمره‌های خام بین ۰/۷۱۴ تا ۰/۹۰۱ متغیر بوده است، تغییر مقادیر پایایی بین خرده‌آزمون‌ها ارتباط نزدیکی با تعداد سؤال‌های خرده‌آزمون داشته است، به‌طوری‌که چه در داده‌های شبیه‌سازی‌شده و چه در داده‌های واقعی آن دسته از خرده آزمون‌هایی که تعداد سؤال‌های بیشتری داشته‌اند، داری ضریب پایایی بزرگ‌تر بودند و خرده آزمون‌هایی که داری تعداد سؤال‌های کمتر بودند حائز ضریب پایایی کمتری شده‌اند. تأثیر دیگری که تعداد سؤال‌ها بر ویژگی نمره‌های خام داشته تغییر در واریانس نمره‌ها بود، به‌طوری‌که خرده آزمون‌هایی که تعداد سؤال بیشتری داشتند نسبت به خرده آزمون‌هایی که تعداد سؤال کمتری داشته‌اند، واریانس بزرگ‌تری را در بین نمره‌ها ایجاد کردند. البته این گفته تا حدودی بستگی به دشواری سؤال‌های هر خرده‌آزمون نیز دارد، به‌طوری‌که در داده‌های واقعی جدول ۲ خرده‌آزمون‌های دشوار مانند ریاضیات، فیزیک و شیمی علی‌رغم تعداد سؤال‌های بیشتر نتوانستند واریانس را خیلی نسبت به

خرده‌آزمون‌های آسان‌تر مثل ادبیات فارسی افزایش دهند. با دقت در جدول ۲ می‌بینید که خرده‌آزمون شیمی به دلیل دشواری زیاد با وجود تعداد سؤال بیشتر نسبت به ادبیات فارسی واریانس کمتری را در نمره‌های خام ایجاد کرده است.

از دو روش نرمال‌سازی و آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس در این پژوهش بهره گرفته شده است. جدول ۳ و جدول ۴ برخی شاخص‌های آماری به همراه ضریب پایایی را برای نمره‌های شبیه‌سازی‌شده و واقعی ۱۰۰۰۰ نمونه تصادفی برای هفت خرده‌آزمون نمایش می‌دهند. برای محاسبه ضریب پایایی گزارش شده از رابطه پایایی کولن (۲۰۱۲) استفاده شد. این ضریب با فرض اینکه توزیع نمره‌های خام دوجمله‌ای است محاسبه شده است و برای محاسبه ضریب پایایی نمره‌های مقیاس به کار می‌رود (کولن، ۲۰۱۲). ضرایب پایایی مانند کرونباخ، کودر ریچاردسون و...تنها برای نمره‌های خام کاربرد داشته و بدون اطلاع از چگونگی پاسخ افراد به همه سؤال‌ها قادر به محاسبه پایایی نیستند، این در حالی است که ضریب پایایی کولن بدون دسترسی به پاسخ همه سؤال‌ها توسط افراد و تنها با داشتن توزیع نمره‌ها می‌تواند پایایی نمره‌های مقیاس را محاسبه کند.

جدول ۳. برخی شاخص‌های آماری به همراه ضریب پایایی نمره‌های نرمال (داده‌های شبیه‌سازی‌شده)

خرده‌آزمون	گشتاورهای اول تا چهارم				ضریب پایایی
	میانگین	واریانس	چولگی	کشیدگی	
فارسی	۴۹۹۹/۸	۱۵۴۵۰۰۰	-/۰۳	-۰/۶۶	۰/۹۵۷
عربی	۵۰۰۱/۳	۱۵۴۱۶۰۰	۰/۸۵	-۰/۵۳	۰/۹۵۶
معارف	۵۰۰۰/۶	۱۵۴۴۰۰۰	۰/۴۵	-۰/۶۲	۰/۹۵۵
زبان خارجه	۵۰۰۱/۶	۱۵۳۸۸۰۰	۱/۰۶	-۰/۶۲	۰/۹۵۳
ریاضی	۵۰۰۰/۲	۱۵۵۷۴۰۰	۰/۶۱	-۰/۶۱	۰/۹۷۰
فیزیک	۵۰۰۰/۷	۱۵۵۳۱۰۰	۰/۷۷	-۰/۶۵	۰/۹۶۳
شیمی	۵۰۰۰/۹	۱۵۵۰۰۰۰	۰/۸۱	-۰/۶۶	۰/۹۵۹

جدول ٤. برخی شاخص‌های آماری به همراه ضریب پایایی نمره‌های نرمال (داده‌های واقعی)

خرده‌آزمون	گشتاورهای اول تا چهارم				ضریب پایایی
	میانگین	واریانس	چولگی	کشیدگی	
فارسی	۵۰۰۵/۳	۱۵۱۴۱۰۰	۱/۳۷	-۰/۵۳	۰/۹۵۲
عربی	۵۰۲۰/۳	۱۴۳۳۵۰۰	۱/۴۲	-۰/۷۱	۰/۹۴۰
معارف	۵۰۰۴/۶	۱۵۲۰۷۰۰	۱/۲۵	-۰/۰۷	۰/۹۵۷
زبان خارجه	۵۰۳۸/۶	۱۳۴۰۰۰۰	۱/۵۴	-۰/۲۰	۰/۹۴۰
ریاضی	۵۰۳۹/۷	۱۳۳۷۹۰۰	۱/۲۷	-۱/۳۶	۰/۹۲۲
فیزیک	۵۰۴۰/۸	۱۳۳۳۵۰۰	۱/۳۴	-۱/۱۳	۰/۹۲۴
شیمی	۵۰۵۳/۳	۱۲۶۳۴۰۰	۱/۲۹	-۱/۳۸	۰/۹۱۲

جدول ۵ برخی شاخص‌های آماری به همراه ضریب پایایی را برای تبدیل آرک‌سینوس نمایش می‌دهد. داده‌های این جدول حاصل از ۱۰۰۰۰ داده شبیه‌سازی شده است که به‌صورت تصادفی برای هفت خرده‌آزمون انتخاب شده‌اند.

جدول ۵. برخی شاخص‌های آماری و ضریب پایایی نمره‌های آرک‌سینوس (شبیه‌سازی)

خرده‌آزمون	گشتاورهای اول تا چهارم				ضریب پایایی
	میانگین	واریانس	چولگی	کشیدگی	
فارسی	۵۰۴۱/۸	۱۷۲۰۷۰۰	۰/۰۵	۰/۶۴	۰/۹۵۹
عربی	۴۲۵۴/۹	۱۹۳۴۳۰۰	۰/۸۸	-۰/۵۱	۰/۹۶۲
معارف	۴۶۰۶/۵	۱۷۰۱۱۰۰	۰/۵۰	-۰/۶۰	۰/۹۵۸
زبان خارجه	۳۹۹۲/۲	۱۶۸۳۱۰۰	۱/۱۰	-۰/۴۶	۰/۹۵۶
ریاضی	۴۲۹۷/۹	۱۶۸۰۱۰۰	۰/۸۲	-۰/۶۲	۰/۹۷۲
فیزیک	۴۰۳۴	۱۵۳۲۹۰۰	۱/۰۵	-۰/۵۴	۰/۹۶۵
شیمی	۴۱۷۷/۲	۱۴۹۶۹۰۰	۰/۹۴	-۰/۵۴	۰/۹۵۹

جدول ۶ برخی شاخص‌های آماری به همراه ضریب پایایی را برای نمره‌های واقعی ۱۰۰۰۰ نمونه تصادفی داوطلبان آزمون سراسری در سال ۱۳۹۵ را نمایش می‌دهد که در آن‌ها از تبدیل آرک‌سینوس برای تولید مقیاس استفاده‌شده است.

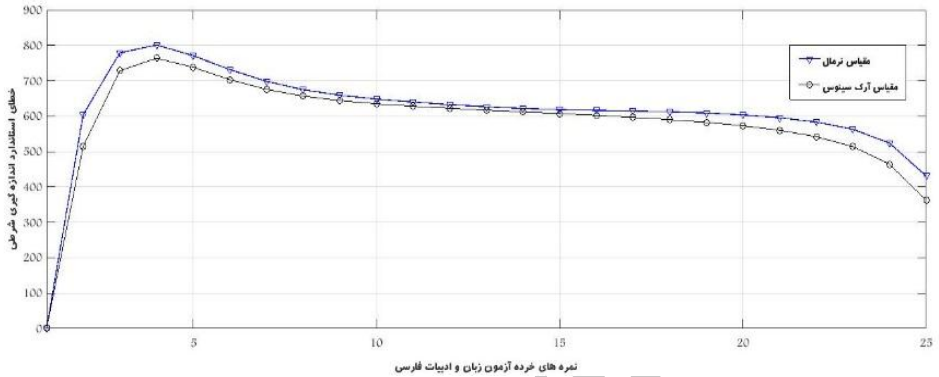
جدول ٦. برخی شاخص‌های آماری به همراه ضریب پایایی نمره‌های آرک‌سینوس (داده‌های واقعی)

خرده‌آزمون	گشتاورهای اول تا چهارم				ضریب پایایی
	میانگین	واریانس	چولگی	کشیدگی	
فارسی	۳۵۵۰/۵۰	۱۹۱۶۴۰۰	۱/۳۵	−۰/۳۹	۰/۹۶۲
عربی	۲۷۰۰/۱	۲۵۴۹۲۰۰	۱/۵۰	−۰/۵۳	۰/۹۷۲
معارف	۳۹۳۹	۳۰۴۹۷۰۰	۱/۱۴	−۰/۴۳	۰/۹۷۶
زبان خارجه	۲۷۷۰/۳	۴۸۶۸۹۰۰	۱/۵۰	−۰/۵۱	۰/۹۸۷
ریاضی	۱۳۷۲/۸	۱۶۱۷۰۰۰	۱/۳۸	−۱/۰۳	۰/۹۷۴
فیزیک	۱۶۴۲/۹	۲۳۰۶۰۰۰	۱/۴۱	−۰/۹۴	۰/۹۸۰
شیمی	۱۳۹۵/۳	۱۶۰۱۱۰۰	۱/۴۰	−۱	۰/۹۶۸

تفاوت بین ویژگی‌های آماری و اندازه‌گیری در تبدیل نرمال و آرک‌سینوس را می‌توان در مقایسه جداول ۳ و ۴ با جداول ۵ و ۶ مشاهده کرد، ضریب پایایی نمره‌ها در هر دو روش مقدار بالایی را نشان داده است، دقت بالای نمره‌ها در تبدیل نمره‌های خام به نمره‌های مقیاس در همه خرده‌آزمون‌ها مشاهده می‌شود، هرچند که مقدار ضریب پایایی برای تبدیل آرک‌سینوس در نمره‌های واقعی و شبیه‌سازی شده مقداری بیشتر از ضریب پایایی در نمره‌های نرمال را نشان داده است ولی این تفاوت در بیشترین موارد به چند صدم نمی‌رسد. اگر چهار گشتاور اول (میانگین، واریانس، چولگی و کشیدگی) هر دودسته نمره را با هم مقایسه کنیم می‌بینیم که روش تبدیل آرک‌سینوس توانسته است، واریانس بیشتری را بین نمره‌ها ایجاد کند. در مقایسه واریانس خرده‌آزمون‌ها مشاهده می‌کنید که واریانس در روش تبدیل آرک‌سینوس از روش نرمال‌سازی در همه آن‌ها بیشتر بوده که نشان می‌دهد این روش قادر است تفکیک بهتری بین سطوح نمره‌ها ایجاد کند. میزان چولگی و کشیدگی نمره‌ها در هر دو روش تفاوت زیادی نداشته و نشان می‌دهد که دو روش بر روی چولگی و کشیدگی نمره‌ها تغییر چندانی ایجاد نمی‌کنند.

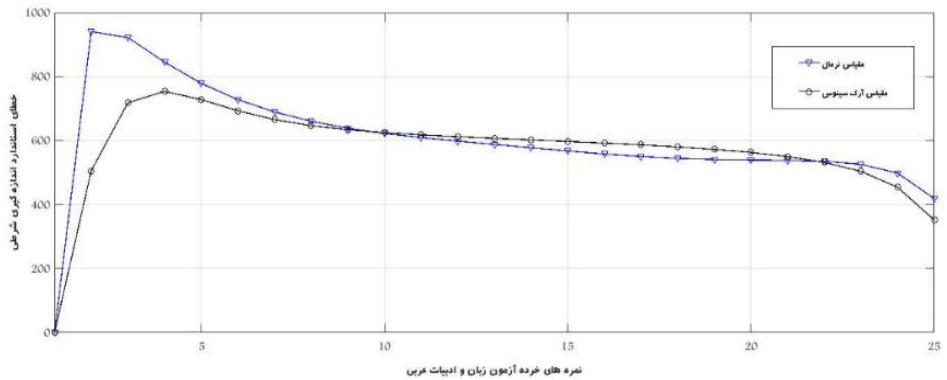
در این پژوهش از دو روش مقیاس‌سازی برای تبدیل نمره‌های خام به نمره‌های مقیاس استفاده شده است. روش اول که به روش نرمال‌سازی مشهور است و روش دوم که به تبدیل آرک‌سینوس مشهور است در این قسمت قصد بر آن است که با مقایسه نمودارهای خطای

استاندارد اندازه‌گیری شرطی و همچنین مقادیر میانگین این خطا برای هر روش، این دو نوع تبدیل با هم مقایسه شوند. نمودار شکل ۱ خطای استاندارد اندازه‌گیری شرطی برای نمره‌های خرده‌آزمون زبان و ادبیات فارسی را برای نمره‌های مقیاس نرمال و آرک‌سینوس با هم مقایسه کرده است.



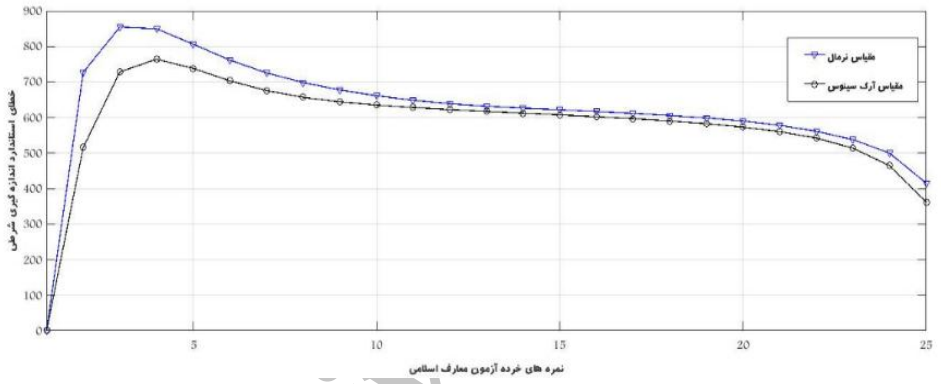
شکل ۱. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون فارسی

همان‌گونه که در شکل ۱ می‌بینید، خطای استاندارد اندازه‌گیری شرطی برای مقیاس آرک‌سینوس از مقیاس نرمال در بیشتر سطوح کمتر بوده هرچند در میانه طیف نمره‌ها این خطاها به هم نزدیک هستند. شکل ۲ خطای استاندارد اندازه‌گیری شرطی برای نمره‌های خرده‌آزمون زبان و ادبیات عربی را برای نمره‌های مقیاس نرمال و آرک‌سینوس را با هم مقایسه کرده است.



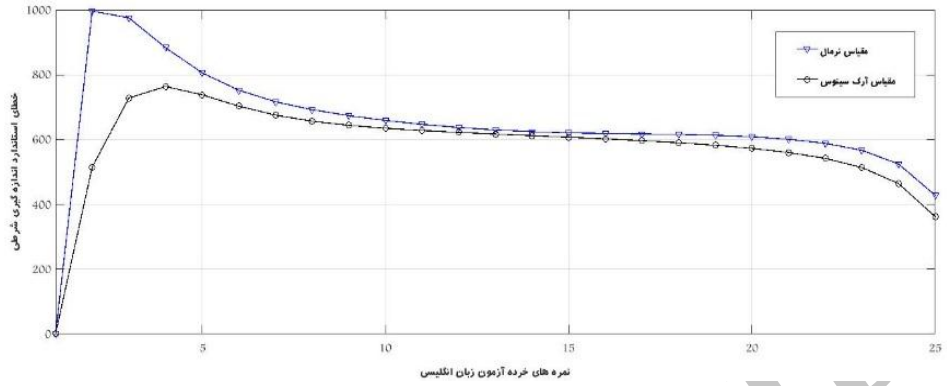
شکل ۲. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون عربی

در نمودار شکل ۲ می‌بینید، خطای استاندارد اندازه‌گیری شرطی برای مقیاس آرک‌سینوس از مقیاس نرمال در بیشتر سطوح کمتر بوده هرچند در میانه طیف نمره‌ها این خطاها به هم نزدیک هستند و در نقاطی هم این خطای مقیاس نرمال است که کمتر شده ولی نوسان خطاها در مقیاس آرک‌سینوس نسبت به مقیاس نرمال کمتر بوده به‌طوری‌که نمودار آن افقی‌تر از نمودار مقیاس نرمال به نظر می‌رسد که خود نشان‌دهنده‌ی هم‌ترازی خطاها برای بیشتر سطوح نمره‌ها است. شکل ۳ همین نمودار را برای خرده‌آزمون معارف اسلامی نشان داده است.



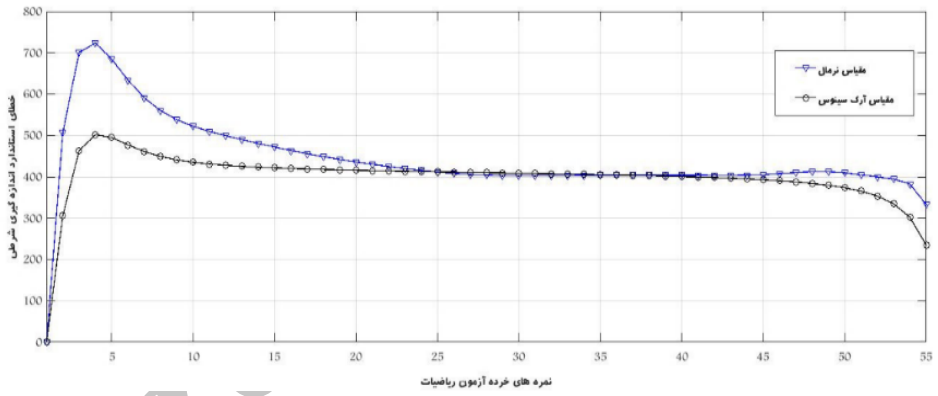
شکل ۳. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون معارف اسلامی

بازهم در شکل ۳ می‌توانید کاهش خطای استاندارد اندازه‌گیری شرطی را برای مقیاس آرک‌سینوس و مسطح شدن نمودار آن را نسبت به نمودار مقیاس نرمال مشاهده کنید. شکل ۴ نمودار را برای خرده‌آزمون زبان و ادبیات انگلیسی نشان داده است.



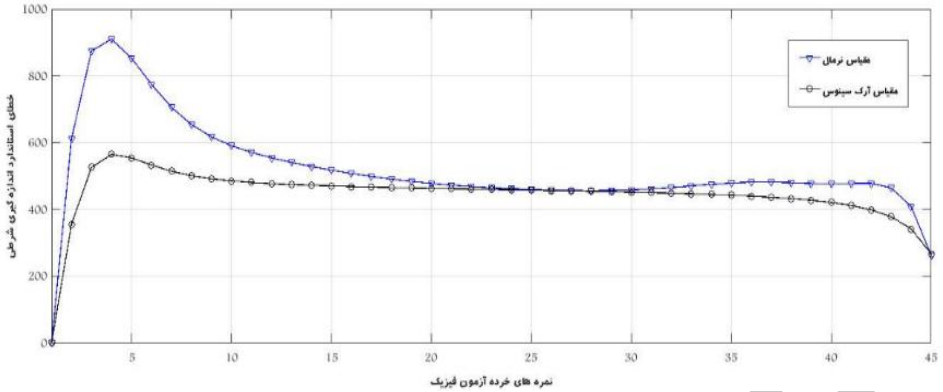
شکل ٤. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون زبان و ادبیات انگلیسی

شکل ۵ نمودار را برای خرده‌آزمون ریاضی نمایش داده است.



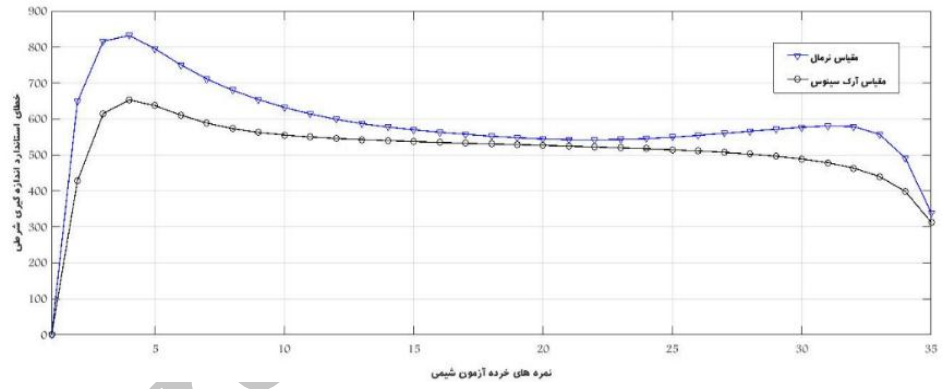
شکل ۵. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون ریاضی

در نمودار ۵ به کاهش چشم‌گیر خطا و همچنین مسطح شدن نمودار به‌خصوص در میانه طیف دقت کنید. نمودار ۶ مربوط به خرده‌آزمون فیزیک است.



شکل ٦. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون فیزیک

در نمودار ۶ نیز میانه طیف نمره‌ها، در هر دو مقیاس خطای مشابهی را به نمایش گذاشته‌اند. شکل ۷ نمودار را برای خرده‌آزمون شیمی نشان داده است.



شکل ۷. مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس خرده‌آزمون شیمی

نمودارهای ۱ تا ۷ مقایسه بین خطای استاندارد اندازه‌گیری شرطی بین نمره‌های خرده آزمون‌های مختلف را در دو نوع تبدیل نشان می‌دهند، دقت در این نمودارها دو نکته اساسی را نمایش می‌دهد، اول اینکه استفاده از روش آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس باعث کاهش خطای استاندارد اندازه‌گیری شرطی برای کلیه سطوح نمره‌ها گردیده است به‌طوری‌که به‌عنوان‌مثال بیشینه خطای استاندارد اندازه‌گیری شده در درس ریاضی به روش نرمال‌سازی عدد ۵۹۶/۰۳۲ است، ولی بیشینه خطای استاندارد اندازه‌گیری

شرطی در درس ریاضی برای روش آرک‌سینوس ۵۶۱/۸۳ است، دومین نکته اینکه روش آرک‌سینوس در تبدیل نمره‌های خام به نمره‌های مقیاس نوسان کمتری داشته و برای همه سطوح نمره‌ها خطا تقریباً یکسان است. نمودار در ابتدا اعداد بالایی را برای خطای استاندارد اندازه‌گیری شرطی در هر دو تبدیل نشان می‌دهند که وقتی به نقاط میانی توزیع نزدیک می‌شویم این مقدار کاهش می‌یابد، این کاهش خطا در روش آرک‌سینوس در میانه توزیع بیشتر بوده ضمن اینکه نمودار روش تبدیل آرک‌سینوس در نقاط میانی نوسان نداشته و به خط راست نزدیک‌تر است که این ویژگی نشان از تمایل این روش به همسان‌سازی خطاهای استاندارد اندازه‌گیری شرطی برای همه سطوح نمره‌ها دارد، یا به زبان ساده در این روش برای همه نمره‌ها تقریباً به یک اندازه مرتکب خطای استاندارد اندازه‌گیری می‌شویم. در روش نرمال‌سازی ضمن اینکه مقدار خطای استاندارد اندازه‌گیری شرطی بالایی را برای بیشتر سطوح نمره‌ها مرتکب می‌شویم، برای همه سطوح نمره‌ها به یک اندازه خطا نمی‌کنیم و این مقدار دچار نوسان است که در نمودارهای ۱ تا ۷ این نوسان را مشاهده می‌کنید.

جدول ۷ مقادیر مجذور میانگین خطای استاندارد اندازه‌گیری شرطی برای هفت خرده‌آزمون در دو مقیاس استفاده‌شده در این پژوهش را نشان می‌دهد.

جدول ۷. مجذور میانگین خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس

نوع داده / مقیاس	خرده آزمون‌ها						
	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
	نرمال						
واقعی	۲۷۰/۴۷	۲۹۳/۰۷	۲۵۴/۲۰	۲۸۲/۹۸	۳۲۳/۶۰	۳۱۷/۳۰	۳۳۳/۲۳
شبیه‌سازی	۲۵۸/۳۷	۲۵۹/۹۵	۲۶۲/۹۷	۲۷۰/۱۴	۲۱۷/۴۴	۲۳۸/۳۵	۲۵۲/۳۴
	آرک‌سینوس						
واقعی	۲۵۳/۹۱	۲۴۷/۷۸	۲۵۲/۲	۲۳۵/۹۵	۱۹۳/۱۸	۲۰۳/۶۸	۲۱۴/۲۴
شبیه‌سازی	۲۴۸/۵۷	۲۵۱/۲۳	۲۵۱/۲۱	۲۵۴/۶۴	۲۰۴/۱۳	۲۱۷/۵۶	۲۳۳/۹۱

با توجه به جدول ۷ مقادیر مجذور میانگین مربوط به روش آرک‌سینوس در همه خرده‌آزمون‌ها از روش نرمال‌سازی کمتر است، میانگین خطای استاندارد اندازه‌گیری شرطی

در روش نرمال‌سازی چه در داده‌های شبیه‌سازی شده و چه در داده‌های واقعی مقدار بیشتری را نشان می‌دهد که دلیل آن‌هم به خاطر مقدار بالای خطای استاندارد اندازه‌گیری شرطی در ابتدا و انتهای توزیع و هم به دلیل نوسان‌هایی است که این خطا در میانه‌های توزیع از خود نشان داده است.

بحث و نتیجه‌گیری

پژوهش حاضر به‌منظور بررسی ویژگی‌های اندازه‌گیری و آماری در روش‌های نرمال‌سازی و آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس و همچنین مقایسه خطای استاندارد اندازه‌گیری شرطی نمره‌ها بین این دو روش طرح ریزی شده است. در این پژوهش نمره‌های خام حاصل از خرده‌آزمون‌های ادبیات فارسی، عربی، معارف اسلامی، زبان انگلیسی، ریاضی، فیزیک و شیمی که مربوط به ۱۰۰۰۰ نمونه شبیه‌سازی شده و ۱۰۰۰۰ نمونه واقعی بود را به دو روش نرمال‌سازی و تبدیل آرک‌سینوس به نمره‌های مقیاس تبدیل شدند. شاخص‌های مختلف آماری و شاخص‌های اندازه‌گیری مانند خطای استاندارد اندازه‌گیری و ضریب پایایی هم برای نمره‌های خام و هم برای نمره‌های مقیاس محاسبه و گزارش شدند. از نظریه نمره حقیقی قوی برای بررسی خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس و همچنین محاسبه پایایی نمره‌ها استفاده شد و با استفاده از مجذور میانگین خطای استاندارد اندازه‌گیری شرطی و نمودارهای خطای استاندارد اندازه‌گیری شرطی اقدام به مقایسه دقت روش‌های ساخت نمره مقیاس (روش‌های نرمال‌سازی و تبدیل آرک‌سینوس) نمودیم. نتایج حاکی از برتری روش تبدیل آرک‌سینوس در کاهش خطای استاندارد اندازه‌گیری شرطی برای توزیع نمره‌ها داشت، به‌طوری‌که با وجود خطای استاندارد اندازه‌گیری شرطی بالایی که هر دو روش برای ابتدا و انتهای توزیع نمره‌ها از خود نشان دادند، روش تبدیل آرک‌سینوس توانسته ضمن کاهش میزان خطای استاندارد اندازه‌گیری شرطی برای میانه توزیع موجب همسان‌سازی محسوس در خطا گردیده به‌طوری‌که در نقاط میانی توزیع شاهد تبدیل این منحنی به یک خط تقریباً راست بودیم. میانگین خطای استاندارد اندازه‌گیری شرطی روش تبدیل آرک‌سینوس از روش نرمال‌سازی کمتر بوده است و به نظر می‌رسد

استفاده از تبدیل آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس می‌تواند باعث کاهش خطای استاندارد اندازه‌گیری برای همه سطوح نمره‌ها شده و به همسانی خطا برای همه سطوح نمره‌ها کمک کند.

در این پژوهش با بررسی که در انواع روش‌های تبدیل مقیاس صورت گرفت دیده شده که هرچند نوسان خطا در مقیاس نرمال زیادتر از مقیاس آرک‌سینوس بوده است، ولی روش مقیاس نرمال نیز در برخی موارد مقدار خطای کمتری را نسبت آرک‌سینوس نشان داده است این خود نشان می‌دهد که نباید نتیجه‌گیری مطلقی را گزارش داد و برترین مقیاس را به مقیاس خاصی اطلاق کرد، همچنین با تمرکز بر روی تحلیل نتایج مقیاس آرک‌سینوس مشخص شد که این مقیاس دارای دو برتری نسبت به مقیاس نرمال است، برتری اول کاهش مقدار خطا و برتری دوم کاهش نوسان خطاهاست، مقیاس آرک‌سینوس توانسته خطاها را برای سطوح مختلف نمره‌ها همانند کند، نتایج به‌دست‌آمده در این پژوهش با نتایج حاصل از پژوهش‌های کولن و همکاران (۱۹۸۹) همخوانی دارد. کولن (۲۰۱۴) بر همان نتایج در کتاب هم‌تراز سازی، پیوند زدن و مقیاس‌سازی اشاره مجدد کرده است، در پژوهش کولن و همکاران (۱۹۸۹) از آرک‌سینوس به‌عنوان مقیاسی نام برده شده که می‌تواند نوسان خطاها را کاهش داده و به همانندسازی خطاها برای همه سطوح نمره‌ها کمک کند، پژوهش حاضر نیز نتایج گزارش‌شده توسط آن پژوهش را تأیید کرده و نشان داد هم در داده‌های شبیه‌سازی و هم در داده‌های واقعی نوسان خطاها در مقیاس آرک‌سینوس کاهش یافته و می‌توان به‌طور تقریبی چنین بیان کرد که خطا تقریباً برای همه سطوح نمره‌ها یکسان است. همچنین در آن پژوهش به این موضوع اشاره شده که مقیاس خوب و برتری وجود ندارد و هرکدام از مقیاس‌ها داری ویژگی‌های خاص خودش است و مطلوب بودن هر مقیاس در معناداری و تفسیرپذیری بهتر آن است. همچنین نتایج این پژوهش نشان داد که هیچ‌کدام از روش‌های مقیاس‌سازی نتوانسته تفاوت آشکاری را در مقدار ضریب پایایی ایجاد کنند و برای هرکدام از روش‌ها مزایایی بیان شد، به‌عنوان‌مثال مقیاس نرمال در نمودار خطای استاندارد اندازه‌گیری دچار نوسان است و یا اینکه خطاها در مقیاس آرک‌سینوس به هم نزدیک هستند، چانگ در سال ۲۰۰۶ برای بررسی انواع روش‌های مقیاس‌سازی خطی و غیرخطی از

ضریب پایایی آن‌ها استفاده کرده است، در پژوهش چانگ (۲۰۰۶) از روش خطی تبدیل نمره‌ها، روش آرک‌سینوس و نرمال بهره برده شده، نتایج این پژوهش با پژوهش چانگ (۲۰۰۶) در برآورد مقدار ضریب پایایی به هم نزدیک بوده و هر دو پژوهش روی این موضوع تأکید دارند که روش‌های مقیاس سازی نتوانستند تأثیر چشم‌گیری بر مقدار پایایی داشته باشند. این گفته بیانگر این موضوع است که ضریب پایایی معیار مناسبی برای مقایسه نمره‌های مقیاس نیست. این پژوهش نشان داد که هرگاه معیار برای انتخاب تبدیل مقیاس را میانگین خطای استاندارد اندازه‌گیری شرطی قرار دهیم، می‌توانیم بین روش‌های متفاوت مقیاس سازی تفکیک و تفاوت قائل شویم. به‌گونه‌ای که جدول ۷ نمایش داد مجذور میانگین خطای استاندارد اندازه‌گیری شرطی در بین روش‌های متفاوت مقیاس سازی مقادیر متفاوتی خواهد داشت و نشان می‌دهد که می‌تواند معیار مناسبی برای مقایسه ایجاد کند.

به‌عنوان پیشنهاد کاربردی باید چنین گفت که استفاده از مقیاس آرک‌سینوس، باعث همانند شدن خطای استاندارد اندازه‌گیری برای همه سطوح نمره‌ها می‌شود، به همین منظور درصورتی‌که بخواهید نمره عادلانه‌تری به افراد اختصاص دهید از مقیاس آرک‌سینوس برای تبدیل نمره‌های خام به نمره‌های مقیاس استفاده کنید. ضمن اینکه برای تفسیرپذیری بهتر آزمون‌ها توصیه می‌شود که خطای استاندارد اندازه‌گیری شرطی هر نمره محاسبه شود و آزمون‌ساز در گزارش فنی مربوط به آزمون به این خطاها اشاره کند، این موضوع باعث می‌شود که دید بهتری نسبت به نمره‌ها داشته و در مواردی که قرار است از برش نمره‌ها استفاده شود با دقت مطلوب‌تری این کار صورت بگیرد.

در این پژوهش از روش‌های هموارسازی برای ساختن نمره‌های مقیاس استفاده نشد و همه نمره‌ها به‌صورت هموار نشده در تحلیل‌ها مورداستفاده قرار گرفته‌اند. برای هموار شدن توزیع نمره‌ها و کمک به کاهش خطاهای تصادفی می‌توان از یکی از روش‌های هموارسازی توزیع بتای چهار پارامتری[1] (لرد ۱۹۶۵) و یا دوجمله‌ای کرنل (کولن ۱۹۹۱) نیز استفاده کرد. نکته آخر اینکه در این پژوهش از توزیع دوجمله‌ای نمره حقیقی قوی برای توزیع شرطی

1. Beta4

نمره‌های مشاهده شده و حقیقی بهره برده‌ایم که با بسط این توزیع می‌توان از توزیع دوجمله‌ای گسترش یافته (کولن ۱۹۹۲) نیز برای بررسی دقیق‌تر خطای استاندارد اندازه‌گیری شرطی استفاده کرد.

منابع

سازمان سنجش آموزش کشور (۱۳۹۵). *کارنامه آماری آزمون سراسری سال ۱۳۹۵*. تهران: انتشارات سازمان سنجش آموزش کشور (دفتر طرح و آمار)

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

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 RL. Thorndike (Ed.).

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

Brennan, R. L., & Lee, W. C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement*, 59(1), 5-24.

Brooks, G. P., & Johnson, G. A. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*. 27(4), 303-304.

Brooks, G. P., & Johnson, G. A. (2014). *TAP: Test Analysis Program* version (14.7.4) [computer software]. Retrieved from http://www.ohio.edu/people/brooksg/software.htm.

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

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

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

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

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

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

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

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

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

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

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling and Linking (2rd Ed.)*. New York: Springer.

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.

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

Lee, W. C., Brennan, R. L. & Kolen, M. J. (2000), Estimators of Conditional Scale-Score Standard Errors of Measurement: A Simulation Study. *Journal of Educational Measurement*, 37, 1–20.

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

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

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

Mood, M. A., Gray bill, A. F. & Boes, C. D. (2008). *Introduction to the Theory of Statistics*. C.A: McGraw-Hill.

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 Education; and Macmillan.

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

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

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*, 2013 (7). ACT, Inc.