فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال هفتم، شمارهٔ ۲۰، زمستان ۱۳۹۶، صص ۱۳۲-۱۰۳
Journal of Educational Measurement & Evaluation Studies
Vol. 7, No. 20, winter 2018

نقش هموارسازی در تبدیل غیرخطی نمره‌های خام به نمره‌های مقیاس نرمال

مجتبی جهانی‌فر *
ابراهیم خدایی **
جلیل یونسی ***
سید امین موسوی ****

تاریخ دریافت مقاله: ۹۶/۰۲/۳۰
تاریخ پذیرش مقاله: ۹۶/۰۹/۱۸

چکیده

آزمون‌های مرکب شامل چند خرده‌آزمون هستند که ممکن است به لحاظ محتوا و تعداد پرسش‌ها متفاوت باشند. برای تفسیرپذیری بهتر و مقایسه‌پذیر کردن نمره خرده‌آزمون‌ها، نمره خام به‌دست آمده از خرده‌آزمون‌ها به مقیاس مشترکی تبدیل می‌شود که به آن نمره مقیاس گفته می‌شود. یکی از روش‌های مرسوم تبدیل نمره‌های خام به نمره‌های مقیاس، تبدیل مقیاس نرمال است. در این تبدیل از فراوانی تراکمی و رتبه درصدی هر نمره برای ساختن نمره مقیاس استفاده می‌شود. هدف این پژوهش، بررسی اثر به‌کارگیری روش پیش‌هموارسازی فراوانی نمره‌ها و پس‌هموارسازی آنها بر میزان خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس است. برای بررسی این اثر از ۱۰۰۰۰ داده شبیه‌سازی شده و ۱۰۰۰۰ داده واقعی آزمون سراسری ایران در گروه آزمایشی ریاضی و فنی در سال ۱۳۹۵ بهره گرفته شد. همچنین از روش‌های هموارسازی دو جمله‌ای کرنل و هموارسازی اسپلاین به ترتیب برای هموار کردن فراوانی نسبی نمره‌ها و نمره‌های مقیاس استفاده شد. برای مقایسه نمره‌های مقیاس ساخته شده از خطای استاندارد اندازه‌گیری شرطی نمره‌ها بهره گرفته شد. نتایج پژوهش، نشان‌دهنده مقدار بالای ضریب پایایی برای همه روش‌ها بود. ضمن اینکه تحلیل نمودار و میانگین خطای استاندارد اندازه‌گیری شرطی نشان داد که در آن دسته از روش‌های تبدیل نمره‌های خام به نمره‌های مقیاس که از پیش‌هموارسازی فراوانی، استفاده شده است، میانگین خطای استاندارد اندازه‌گیری شرطی کمتر بوده و استفاده از پیش‌هموارسازی تا حد مطلوبی نوسان خطا برای سطوح مختلف نمره‌ها را کاهش داده است.

واژگان کلیدی: هموارسازی، نمره‌های مقیاس، تبدیل مقیاس نرمال، خطای استاندارد اندازه‌گیری شرطی

---

* دانشجوی دکتری سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران (نویسنده مسئول: m.jahanifar@ut.ac.ir)
** دانشیار گروه روش‌ها و برنامه‌های آموزشی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران
*** دانشیار گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی
**** استادیار دانشکده آموزش، دانشگاه ساسکاچوان، کانادا

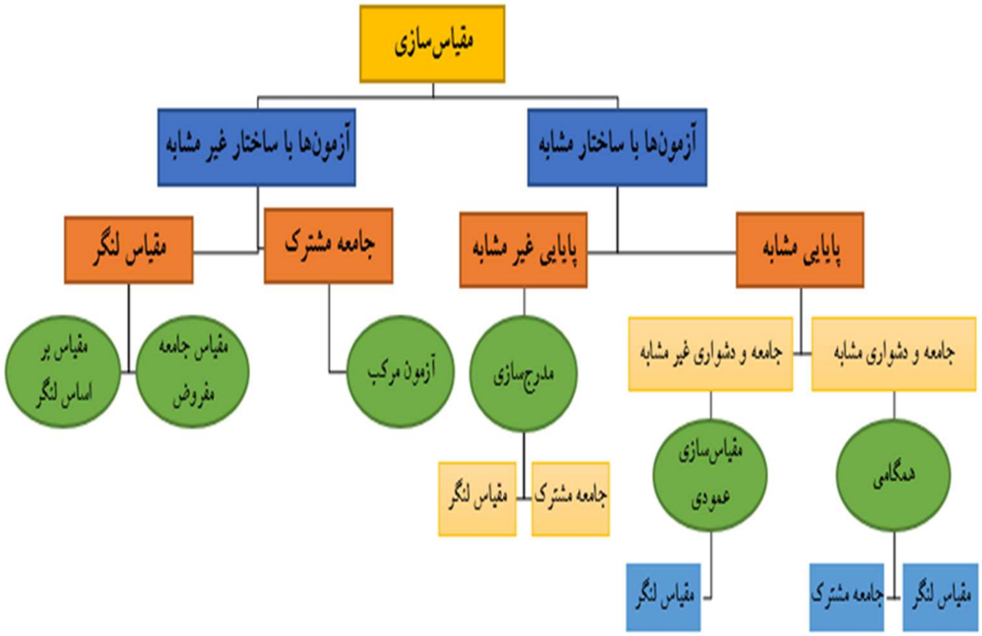
مقدمه

آزمون‌ها با اشکال و روش‌های متفاوت توسط مؤسسه‌ها و شرکت‌های مختلف طراحی و اجرا می‌شوند. این آزمون‌ها در برخی موارد تنها یک موضوع مشخص و گاهی به به‌صورت ترکیبی از موضوع‌های مختلف هستند. به‌عنوان مثال، آزمون ورودی دانشگاه‌ها به‌طور معمول مشتمل بر چند موضوع می‌شود که به آن آزمون‌های مرکب[1] می‌گویند. هر آزمون مرکب شامل چند خرده‌آزمون مختلف است که هر کدام نمره خاصی را به خود اختصاص می‌دهند. در صورتی می‌توان نمره خرده‌آزمون‌های مختلف را باهم جمع کرد که روی یک مقیاس مشترک باشند. برای استخراج نمره کل هر آزمون از نمره‌های خام، ابتدا نمره‌های خام[2] را به نمره‌های مقیاس[3] تبدیل می‌کنند. روش‌های مقیاس‌سازی می‌توانند موجب کاهش اطلاعات و یا ایجاد خطاهای تصادفی و نظام‌دار در نمره‌ها شوند و نمره‌های هر حوزه را تحت تأثیر قرار دهند. هرگاه فرایندهای ساخت آزمون و مقیاس‌سازی[4] برای هر کدام از بخش‌های آزمون مرکب به‌طور مشابه صورت بگیرد، مقایسه نمره‌های آزمون‌شوندگان در همه بخش‌های آزمون و همچنین در نمره مرکب آسان‌تر خواهد بود، ضمن اینکه نمره‌های مرکب معنی‌دارتری تولید خواهد شد (کولن و برنان[5]، ۲۰۱٤).

یکی از قدیمی‌ترین روش‌های پیوند زدن[6] نمره‌ها، مقیاس‌سازی نمره‌ها است. مقیاس‌سازی، همان تبدیل نمره‌های خام به نمره‌های مقیاس است. در روان‌سنجی همواره به مقیاس‌های متفاوت نمره‌ها نیاز بوده و به همین خاطر، قدمت این روش به قدمت علم روان‌سنجی است. هدف از مقیاس‌سازی نمره‌ها تبدیل نمره‌ها از دو آزمون مختلف به مقیاس مشترک[7] است. این روش، پیوند غیرمستقیم بین نمره‌های دو آزمون ایجاد می‌کند (دورانز، پامریچ و هولاند[8]، ۲۰۰۷). در این روش نمره‌های هر آزمون به‌طور جداگانه به مقیاس مشترک برده می‌شوند. در شکل (۱) انواع روش‌های مقیاس‌سازی، نمایش داده شده است.

---

1. Battery tests
2. Raw scores
3. Scale scores
4. Scaling
5. Kolen & Brenan
6. Linking
7. Common scale
8. Dorans, Pommerich & Holland



شکل (۱) طبقه‌بندی انواع روش‌های مقیاس‌سازی (دورانز و همکاران، ۲۰۰۷)

هرگاه دو یا چند آزمون که سازه‌های متفاوتی را اندازه می‌گیرند با هم در جامعه‌ای مشترک از شرکت‌کنندگان اجرا شوند، مقیاس‌سازی روی هریک از آزمون‌ها به‌منظور داشتن مقیاس مشترک انجام خواهد گرفت. کولن و برنان (۲۰۱٤) این آزمون‌ها را آزمون مرکب نامیدند. نکته مهم در اینجاست که بردن نمره‌ها به مقیاس مشترک به معنی هم‌تراز کردن آنها نیست و نمی‌توان نمره آزمون با ساختار مشخص را با آزمون دیگر با ساختار متفاوت تبدیل کرد و مقیاس مشترک تنها آنها را مقایسه‌پذیر می‌کند نه اینکه راهی را فراهم کند که بتوان آنها را به جای هم به کار برد. برای هر آزمون مانند Y که بخواهد مقیاس‌سازی شود، تابع توزیع تراکمی نمره‌ها وجود دارد که روی جامعه مرجع P به صورت رابطه (۱) تعریف می‌شود.

$$F_{YP}(y) = \Pr\left(Y \le y \middle| P\right) \quad (1)$$

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 به وسیله تبدیل S به نمره‌های مقیاس تبدیل می‌شوند.

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

در رابطه (۲)، S تابع تبدیل به مقیاس است که می‌تواند خطی 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 خرده‌آزمون‌هاست (دورانز و همکاران، ۲۰۰۷). تبدیل‌های مقیاس خطی[۱]، تبدیل مقیاس نرمال[۲] و تبدیل آرک سینوس[۳] از مشهورترین روش‌های تبدیل مقیاس است. دو روش آخر را روش‌های تبدیل غیرخطی نمره‌های خام به نمره‌های مقیاس می‌نامند.

تبدیل مقیاس نرمال روش غیرخطی تبدیل نمره‌ها است. نمره‌های خام در هر خرده‌آزمون با روش غیرخطی نرمال‌سازی به یک مقیاس مشترک برده می‌شوند. در روش معمول نرمال‌سازی ابتدا با استفاده از توزیع فراوانی نمره‌های کسب شده توسط تمامی شرکت‌کنندگان در آزمون، فراوانی تراکمی و تراکمی نسبی هر نمره، محاسبه شده و سپس رتبه درصدی برای هر نمره مشخص می‌شود. سپس با استفاده از رتبه درصدی، نمره z متناظر با رتبه درصدی از جدول تبدیل توزیع نرمال، استخراج و به‌عنوان نمره مقیاس در نظر گرفته می‌شود (آنگوف[۴]، ۱۹۷۱). به‌عنوان نمونه، آزمون *SAT*[۵] برای تبدیل نمره‌های خام به نمره‌های مقیاس از گروه مرجع بهره می‌برد. زمان این آزمون سه ساعت و ٤٥ دقیقه و شامل خرده‌آزمون‌های مهارت‌های خواندن، نوشتن و همچنین ریاضیات است. برای مقیاس‌سازی در این آزمون از تبدیل نرمال استفاده می‌شود. در این تبدیل، نمره‌های فرمولی[٦] برحسب فراوانی تراکمی به مقیاس نرمال برده می‌شوند و حاصل این مقیاس‌سازی، جدول تبدیلی است که در آن هر نمره بر اساس رتبه و نمره درصدی به مقیاس آزمون *SAT* برده می‌شود (گزارش کالج برد، ۲۰۱۵). آزمون ***IBTS*** که در دانشگاه آیووا[۷] به‌منظور اندازه‌گیری پیشرفت تحصیلی از سطح مهد کودک تا پایه

---

[1]. Linear Transformation (LT)
[2]. Normalizing Transformation (NT)
[3]. Arcsine Transformation (AT)
[4]. Angoff
[5]. Scholastic Aptitude Test
[6]. Formula score
[7]. Iowa

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 هر کدام هم در تعداد سؤال‌های خرده‌آزمون و هم در زمان پاسخگویی متفاوت هستند. در این آزمون نیز از روش تبدیل نرمال برای ساختن مقیاس نمره‌ها استفاده شده است (گزارش فنی *IBTS*، ۲۰۱۶). آزمون سراسری ورود به دانشگاه‌های ایران نیز آزمونی مرکب است که با توجه به رشته تحصیلی دارای خرده‌آزمون‌های متفاوت 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.

چانگ[1] (۲۰۰۶) سه روش تبدیل مقیاس یعنی خطی، نرمال‌سازی و تبدیل آرک سینوس را با هم مقایسه کرده است. در این مقایسه، سه روش مقیاس‌سازی در شاخصه‌هایی مانند ضریب پایایی نمره‌های مقیاس، نمودار خطای استاندارد شرطی و تعداد تغییر نمره‌ها در اثر برش نمره‌ها[2] و همچنین شکاف‌های[3] 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 ندارد. تصمیم برای انتخاب مقیاس مناسب هم به خواص اندازه‌گیری و هم به سهولت تفسیر بستگی خواهد داشت. در پژوهش چانگ، ضریب پایایی خرده‌آزمون‌ها پس از انجام سه نوع تبدیل به هم نزدیک بودند. در حالی که در ایجاد شکاف‌های بین نمره‌ها روش تبدیل آرک سینوس از دو روش دیگر پیشی گرفته است. در نمودار خطای استاندارد شرطی، این روش از دو روش دیگر خطای کمتری دارد. روش نرمال‌سازی بیشترین شباهت را به توزیع نمره‌های خام و تبدیل خطی نمره‌ها کمترین شکاف را در مقیاس نمره‌ها نشان داده است. با توجه به منابع و بررسی‌های متفاوت نمی‌توان قطعا گفت که کدام روش به سایر روش‌ها ترجیح بیشتری دارد. هرکدام از تبدیل‌ها دارای ویژگی‌های مخصوص به خود هستند و مفید بودن هر روش مقیاس‌سازی به مفید بودن و

---

[1]. Shun-Wen Chang
[2]. Truncation
[3]. Gaps

تفسیرپذیری بهتر آن بستگی داشته و هر روش کاربردهای متفاوتی پیدا کرده است (چانگ، ۲۰۰۶).

پس از تبدیل غیرخطی نمره‌های خام به نمره‌های مقیاس، توزیع نمره‌ها تغییر خواهد کرد، ضمن اینکه فراوانی توزیع نمره‌های مقیاس شامل بی‌نظمی و همچنین گسستگی‌هایی در توزیع است. این موضوع هنگام بررسی نمودار توزیع فراوانی نمره‌های خام بیشتر دیده می‌شود. یکی از روش‌های آماری که از آن برای پیوستگی نمره‌ها و کاهش خطاهای تصادفی در تبدیل توزیع نمره‌ها استفاده می‌شود هموارسازی[1] است. هموارسازی برای تصحیح پیوستگی توزیع نمره‌ها نیز به کار می‌رود و از آن به‌منظور ساختن یک پیوستار از نمره‌ها به جای یک توزیع گسسته استفاده می‌شود (کولن، ۱۹۹۱). برای آنکه هدف هموارسازی دقیق‌تر بیان شود چند تعریف در ادامه ارائه خواهد شد.

اگر $f(i)$ فراوانی نسبی جامعه در نمره i و $\hat{f}(i)$ فراوانی نسبی نمونه تصادفی از جامعه در نمره i باشند، آنگاه مقدار $\hat{f}(i)$ برآوردی از $f(i)$است، به‌گونه‌ای که[2] $E\left(\hat{f}(i)\right)=f(i)$. همچنین اگر $\hat{f}_S(i)$ فراوانی نسبی هموارشده نمونه در نمره i و $f_S(i)$ فراوانی نسبی هموارشده جامعه باشند آنگاه خواهیم داشت: $E\left(\hat{f}_S(i)\right)=f_S(i)$.

همچنین اگر $E\left(\hat{f}(i)-f(i)\right)^2$ واریانس خطای برآورد فراوانی نسبی جامعه از روی فراوانی نسبی نمونه باشد و عبارت $E\left(\hat{f}_S(i)-f(i)\right)^2$واریانس خطای برآورد فراوانی نسبی جامعه از روی فراوانی نسبی هموارشده باشد که به آن خطای مجذور میانگین نیز گفته می‌شود. روش هموارسازی مطلوب، روشی است که فراوانی نسبی هموارشده نسبت به فراوانی نسبی هموار نشده، برآوردگر بهتری برای فراوانی نسبی جامعه باشد یا به زبان ریاضی رابطه (۳) برقرار باشد (کولن، ۱۹۹۱):

$$E\left(\hat{f}_S(i)-f(i)\right)^2 < E\left(\hat{f}(i)-f(i)\right)^2 \qquad (۳)$$

می‌توان گفت روش هموارسازی مطلوب آن است که بیشترین شباهت را بین نمره‌های هموارشده و هموار نشده ایجاد کند (کولن، ۱۹۹۱). روش‌های متعددی برای هموارسازی وجود دارند. این روش‌ها هم می‌توانند روی فراوانی نمره‌ها و هم می‌توانند

1. Smoothing
2. Expectation

روی نمره‌های مقیاس اجرا شوند. معروف‌ترین روش‌های هموارسازی عبارت‌اند از روش هموارسازی کرنل[۱]، روش هموارسازی اسپلاین[۲]، چندجمله‌ای لگاریتم خطی[۳] و بتای چهار پارامتری[٤].

از روش‌های هموارسازی بیشتر به‌منظور برآورد توزیع‌های تجربی و همچنین هموارسازی روش صدک‌های یکسان برای دستیابی به توزیع هموار با ویژگی‌هایی نزدیک به ویژگی‌های نمونه اصلی استفاده شده است. هموارسازی به امید دستیابی به توزیعی دقیق‌تر نسبت به توزیع اصلی انجام می‌شود اما همیشه خطر دستیابی به توزیعی که برآورد ضعیفی از توزیع اصلی باشد وجود داشته است (کولن و برنان، ۲۰۱٤). کار روی روش‌های تحلیلی هموارسازی به صورت تجربی، در کارهای لیو (۲۰۱۱) و موسس و هولاند[۵] (۲۰۰۹) دیده می‌شود. در پژوهش لیو[٦] (۲۰۱۱) برای انتخاب روش مناسب هموارسازی از شاخص‌های آماری استفاده شده و بیشتر بر شباهت توزیع هموارشده و توزیع هموار نشده، تکیه شده است. پژوهش موسس و هولاند (۲۰۰۹) به‌منظور انتخاب راهبردی برای گزینش روش هموارسازی مناسب و همچنین اثر هموارسازی بر دقت توابع همترازسازی انجام گرفته است.

این پژوهش با هدف کلی بررسی نقش هموارسازی در میزان خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس به روش تبدیل نرمال طرح‌ریزی شده است که خود شامل خرده‌اهداف زیر است:

- بررسی نقش پیش‌هموارسازی فراوانی نمره‌ها و پس‌هموارسازی آنها در تبدیل نمره‌های خام به نمره‌های مقیاس نرمال و تأثیر آن بر خطای استاندارد اندازه‌گیری شرطی نمره‌ها
- تعیین روش تبدیل مقیاس نرمال کارآمد بر اساس خطای استاندارد اندازه‌گیری شرطی نمره‌ها

با توجه به هدف‌های پژوهش، پرسش‌های زیر مطرح می‌شوند:

---

1. Kernel Smoothing
2. Spline Smoothing
3. Log linear Smoothing
4. Four parameter Beta
5. Moses & Holland
6. Liu

۱. استفاده از روش‌های هموارسازی (پیش و پس هموارسازی) چه تأثیری بر دقت نمره‌های مقیاس نرمال خواهد داشت؟

۲. کارآمدترین تبدیل مقیاس نرمال در میان روش‌های تبدیل مقیاس نرمال، با توجه به میزان خطای استاندارد اندازه‌گیری شرطی هر روش کدام است؟

روش پژوهش

پژوهش حاضر، پژوهشی کمی و از نوع توصیفی (غیر آزمایشی) است و به‌منظور توسعه دانش کاربردی در زمینه ساختن نمره مقیاس انجام گرفته که با این رویکرد، پژوهشی کاربردی محسوب می‌شود. جامعه این پژوهش، داوطلبان شرکت‌کننده در آزمون سراسری سال ۱۳۹۵ در گروه آزمایشی ریاضی و فنی به تعداد ۱۶۲۸۷۹ نفر هستند. در این پژوهش، روش‌های مختلف مقیاس‌سازی و تحلیل‌ها بر روی ۱۰۰۰۰ داده شبیه‌سازی شده خرده‌آزمون‌های مختلف عمومی و اختصاصی آزمون سراسری ایران در گروه آزمایشی رشته ریاضی و فنی اجرا شده است. برای تولید داده‌های شبیه‌سازی شده از ویژگی هر کدام از خرده‌آزمون‌ها ازجمله تعداد گزینه‌ها، تعداد پرسش‌ها 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 و همچنین رابطه آنها و روش‌های هموارسازی؛ ۲- تعیین مقدار بهینه پارامترهای هموارسازی مانند h؛ ۳- مبنایی برای بررسی صحت محاسبات و تحلیل داده‌های واقعی است. از آنجاکه داده‌های واقعی برای 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 می‌کند که برگرفته از وضعیت واقعی آزمون است. مبنای شبیه‌سازی داده‌ها میانگین دشواری سؤال‌های آزمون است، به‌گونه‌ای که میانگین ضریب دشواری را به عنوان میانگین نسبی $\overline{P} = \frac{M_x}{K}$ (میانگین تقسیم بر تعداد سؤال‌ها) در نظر گرفته سپس بر اساس تعداد گزینه‌ها و تعداد سؤال‌های آزمون داده‌ها تولید شد (بروکس و یوهانسن[1]، ۲۰۰۳). به عنوان مثال هرگاه بخواهیم ۳۰ سؤال دو ارزشی و چهارگزینه‌ای با ضریب دشواری ۰/۵۰ در سطح آزمون ایجاد کنیم، داده‌هایی تولید خواهد شد که دارای میانگین

[1]. Brooks & Johansson

و واریانس تقریبی ۱۵ (حاصل ضرب ۰/۵ در ۳۰) و ۷/۵ (حاصل ضرب ۰/۲۵ در ۳۰) خواهد 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] سر و کار داریم از طریق قاعده سرانگشتی[2] می‌توان حجم نمونه نزدیک به ۱۰۰۰۰ نفر را مناسب دانست. در آزمون سراسری در گروه آزمایشی ریاضی‌وفنی شامل چهار درس عمومی زبان و ادبیات فارسی (۲۵ پرسش)، زبان و ادبیات عربی (۲۵ پرسش)، معارف اسلامی (۲۵ پرسش) و زبان انگلیسی (۲۵ پرسش) و سه درس اختصاصی ریاضیات (۵۵ پرسش)، فیزیک (٤۵ پرسش) و شیمی (۳۵ پرسش) است.

ابزار پژوهش

ابزار اصلی گردآوری داده‌ها در این پژوهش همان پرسش‌های 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 شبیه‌سازی شده نیز در کنار آن استفاده می‌شود. برای تولید داده‌های شبیه‌سازی شده از نرم‌افزار *TAP*[3] نسخه ٤,۷,۱٤ استفاده شده است (بروکس و یوهانسن، ۲۰۱٤). 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 اجرای پژوهش از نرم‌افزار کدنویسی ریاضی و آمار *MATLAB*[4] برای نوشتن کدهای مربوط به برخی روش‌ها و تبدیل‌ها استفاده شد.

---

[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.

روش‌های تبدیل مقیاس نرمال

الف- تبدیل مقیاس نرمال بدون استفاده از هموارسازی

برای به دست آوردن نمره مقیاس به روش تبدیل مقیاس نرمال مراحل به شرح زیر است (آنگوف، ۱۹۷۱):

مرحله اول، توزیع فراوانی نسبی نمره‌ها محاسبه می‌شود. مرحله دوم، با استفاده از فراوانی نسبی و توزیع‌تراکمی نمره‌ها رتبه درصدی نمره‌ها محاسبه می‌شود. مرحله سوم، نمره z مربوط به هر رتبه درصدی از روی معکوس تابع (٤) محاسبه می‌شود:

$$\Phi(z)=\frac{\hat{Q}(y)}{۱۰۰}=\frac{۱}{\sqrt{۲\pi}}\int_{-\infty}^{z}e^{-\omega^{۲}/۲}d\,\omega \qquad (٤)$$

تابع رابطه (٤) یعنی $\Phi(z)$ 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. مرحله چهارم، تبدیل خطی نمره‌های z به دست آمده در مرحله سوم با استفاده از رابطه (۵) است:

$$sc=\sigma(\mathrm{sc})\,\mathrm{z}+\mu(\mathrm{sc}) \qquad (۵)$$

نمره‌های مقیاس sc[۱] 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 جابه‌جا و یا در بازه مشخصی محدود می‌کنند. به‌طور مثال:

$$-۴<z<+۴\Rightarrow ۰<۱۲۵۰\times z+۵۰۰۰<۱۰۰۰۰\Rightarrow ۰<Sc<۱۰۰۰۰$$

مرحله پنجم، تبدیل نمره‌های به دست آمده در مرحله چهارم به نمره‌های صحیح یعنی $sc_{\mathrm{int}}(y)$ است. در این مرحله، عدد به دست آمده به نزدیک‌ترین نمره صحیح گرد می‌شود.

ب- روش نرمال‌سازی با استفاده از هموارسازی

در این روش، همه مراحل بیان شده در قسمت (الف) اجرا خواهند شد. تفاوت این روش تبدیل نرمال با روش قبل در اضافه کردن یک مرحله دیگر با عنوان هموارسازی است. در این روش پس از انجام مرحله اول و به دست آوردن فراوانی نسبی نمره‌ها، با روش هموارسازی دوجمله‌ای کرنل[۲] فراوانی نمره‌ها هموار می‌شوند (در ادامه به

1. Scale scores
2. Kernel Binomial Scaling

چگونگی هموارسازی اشاره خواهد شد). سایر مراحل همچون روش (الف) خواهد بود و این روش پیش‌هموارسازی[1] نام دارد. روش دیگر هموارسازی، پس‌هموارسازی[2] نام دارد، در این روش و در قسمت (الف) پس از مرحله پنجم نمره‌های به دست آمده از طریق توابع اسپلاین[3]، هموار خواهند شد که در ادامه در مورد توابع اسپلاین بیشتر توضیح داده می‌شود.

روش‌های هموارسازی

روش پیش‌هموارسازی دوجمله‌ای کرنل: در این پژوهش از روش هموارسازی کرنل دوجمله‌ای برای پیش‌هموارسازی استفاده شده است. ایده‌ای که در پس این روش پنهان شده این است که برای هر نمره تابع چگالی احتمال معرفی می‌شود که به آن تابع چگالی احتمال کرنل یا هسته گفته می‌شود. معمولا از داده‌های پیوسته برای هموارسازی استفاده می‌شود و از توزیع نرمال برای چگالی بهره می‌برند ولی نمره‌های خام معمولا گسسته هستند؛ به همین خاطر از کرنل دو جمله‌ای برای هموارسازی استفاده می‌شود. در این روش از پارامتر h که عددی زوج است برای کنترل درجه هموارسازی استفاده می‌شود. به عنوان مثال برای یک آزمون که دارای k سؤال است و توزیع نمونه‌ای $\hat{f}(i)$ را دارد و در آن $i = ۰,۱,۲,...,k$، برآوردگر کرنل به صورت رابطه (٦) است:

$$\hat{f}_h(i) = \begin{cases} \sum_{j=i-h/۲}^{i+h/۲} B(j-i+h/۲ \mid h)\hat{f}(j) \\ B(m \mid h) = \binom{h}{m}(\frac{۱}{۲})^m(۱-\frac{۱}{۲})^{h-m} \end{cases} \qquad (٦)$$

در رابطه (٦)، B تابع توزیع دوجمله‌ای با احتمال پیروزی $\frac{۱}{۲}$ است. h پارامتر هموارسازی است و $m = ۰,۱,۲,...,h$، اگر مقدار h صفر باشد مانند آن است که هموارسازی نشده است و مقدار آن به شباهت بیشتر نمونه هموارشده با اصلی بستگی دارد. به عنوان مثال برای $h = ۲$ هموارسازی کرنل به صورت زیر خواهد بود:

$$\hat{f}(i) = \frac{۱}{۴}\hat{f}(i-۱) + \frac{۱}{۲}\hat{f}(i) + \frac{۱}{۴}\hat{f}(i+۱)$$

---

1. Pre-Smoothing
2. Post-Smoothing
3. Spline Functions

تعیین مقدار h از طریق آزمایش و خطا است. در روش کرنل به جای فراوانی هر نمره، میانگین وزن‌دار نمره‌های قبل و بعد هر نمره قرار داده می‌شود تا توزیع هموار شود (کولن و برنان، ۲۰۱٤ و کولن، ۱۹۹۱). برای اجرای روش هموارسازی دوجمله‌ای کرنل، نویسندگان مقاله، برنامه‌ای را در نرم‌افزار *MATLAB* توسعه داده‌اند. در آن نرم‌افزار می‌توان با وارد کردن مقادیر دلخواه h به هموارترین حالت برای فراوانی نمره‌ها رسید که هم به چهار گشتاور اول داده‌های مشاهده شده نزدیک‌تر باشند و هم فراوانی‌ها هموارتر شوند. در این پژوهش برای انتخاب بهترین h برای هموارسازی، ضمن مقایسه چهار گشتاور اول نمره‌های مشاهده شده و نمره‌های هموار شده، نمودار فراوانی آنها نیز مقایسه می‌شود. ضمنا از وارسی مقطعی اعتبار[1] نیز برای انتخاب مناسب‌ترین h استفاده می‌شود. در روش وارسی مقطعی اعتبار نمره‌ها برای h های زوج مختلف هموار می‌شوند و مقدار مطلوب h مقداری خواهد بود که عبارت M_h رابطه (۷) را به کمترین مقدار خود برساند.

$$M_h = \sum_{i=-h/۲}^{k+\frac{h}{۲}} \hat{f}_h(i) - \frac{۲}{n-۱}\left\{\left[\sum_{۰}^{k} \hat{f}(i)\hat{f}_h(i)\right] - \frac{B(h/۲|h)}{n}\right\} \qquad (۷)$$

روش پس‌هموارسازی اسپلاین: در این روش، منحنی هموار را که به وسیله توابع اسپلاین تولید می‌شود با نمره‌های مشاهده شده برازش می‌دهند. در روش اسپلاین ابتدا تابع اسپلاین رابطه (۸) را تشکیل می‌دهند:

$$\sum_{i=۱}^{n}(f(\mathrm{x_i}) - \hat{f}(x_i)^۲) + \lambda\int_{x_۱}^{x_n} \hat{f}''(x)^۲ dx \qquad (۸)$$

در رابطه (۸)، $\mathrm{x_i}$ ها برداری از اعداد صحیح هستند که به وسیله تابع $f(\mathrm{x_i})$ معرفی می‌شوند و $\hat{f}(x_i)$ برآورد تابع اسپلاین برای تابع $f(\mathrm{x_i})$ است که می‌توانند کمینه‌کننده رابطه (۸) باشد و $\hat{f}''(x)$ مشتق دوم تابع $\hat{f}$ است. در این رابطه λ که مقداری غیر منفی است می‌تواند میزان هموار بودن و همچنین نزدیکی داده‌های هموار شده به داده‌های مشاهده شده را نمایش دهد. بیشترین استفاده از مقدار λ برای مقادیر نزدیک به صفر است که تابع اسپلاین مکعبی[2] ساخته می‌شود و مقادیر خیلی بزرگ λ موجب خطی شدن تابع اسپلاین می‌شود. رویکردی که در این پژوهش برای

---

[1]. Cross-Validation
[2]. Cubic Spline Function

هموارسازی نمره‌ها پس از نرمال‌سازی استفاده شده، رویکرد پیشنهادی دی‌بور[1] (۲۰۰۱) است. در این روش، دی‌بور برای نزدیک‌تر شدن مقادیر هموارشده به مقادیر واقعی و همچنین هموارتر شدن نمره‌ها تابع رابطه (۹) را پیشنهاد داده است:

$$p\sum_{1}^{n}\left(\frac{f(x_i)-\hat{f}(x_i)}{\delta_i}\right)+(1-p)\int \left(\hat{f}^{(m)}(x)\right)^2 dx \qquad (۹)$$

در رابطه (۹)، پارامتر p عامل هموارسازی است که مقادیر بین صفر تا یک اختیار می‌کند و از مشتق m ام تابع اسپلاین برای محاسبه رابطه (۹) استفاده می‌شود. در این برآورد δ_i برای کنترل هموارسازی استفاده می‌شود. مقادیر δ_i از ۱ تا n که همان تعداد داده‌ها است گسترش می‌یابد (دی‌بور، ۲۰۰۱).

در عمل از مشتق دوم برآورد تابع اسپلاین برای محاسبه رابطه (۹) استفاده می‌شود که موجب تبدیل رابطه (۹) به تابع اسپلاین مکعبی خواهد شد. از آنجایی که برای محاسبه مقدار مناسب p از روش آزمایش و خطا استفاده می‌شود همواره از معیاری به نام S برای پایان دادن به الگوریتم برآورد استفاده می‌شود. طبق آنچه دی‌بور (۲۰۰۱) بیان کرده ابتدا برآوردی برای δ_i تعیین می‌شود که معمولا برابر با انحراف استاندارد داده‌ها است، سپس S انتخاب می‌شود. انتخاب مقادیر بزرگ S به معنی هموارتر شدن نمره‌ها است و هرگاه شرط رابطه (۱۰) برقرار شود هموارسازی متوقف خواهد شد:

$$\sum_{1}^{n}\left(\frac{f(x_i)-\hat{f}(x_i)}{\delta_i}\right)\le S \qquad (۱۰)$$

برای محاسبه روش هموارسازی اسپلاین، نویسندگان مقاله، برنامه‌ای را در نرم‌افزار *MATLAB* توسعه داده‌اند که در آن می‌توان با تعیین دلخواه مقادیر p، نمره‌ها را پس از نرمال‌سازی به روش اسپلاین هموارسازی کرد. در این برنامه از وارسی مقطعی اعتبار برای رسیدن به هموارترین نمره در کنار بررسی گشتاورهای اول تا چهارم استفاده خواهد شد. وارسی مقطعی اعتبار به دنبال مقداری از p خواهد بود که در آن ضمن هم‌خوانی بیشتر گشتاورهای اول تا چهارم بیشترین مقدار ممکن را برای شاخص برازش R^2 و کمترین مقدار را برای شاخص برازش *RMSE*[2] دربرداشته باشد.

---

[1]. De Boor
[2]. Root Mean Square Error

روش محاسبه خطای استاندارد اندازه‌گیری شرطی

هم‌اکنون، گزارش‌های فنی بیشتر خطای استاندارد اندازه‌گیری را به صورت خطای استاندارد اندازه‌گیری کلی[۱] ارائه می‌دهند ولی مؤسسه‌های *AERA*[۲] و*NCEME*[۳] و*APA*[٤] از سال ۱۹۸۵ برای استانداردهایی که به‌منظور تولید آزمون‌های آموزشی و روانی توصیه کرده‌اند، گزارش خطای استاندارد اندازه‌گیری شرطی[٥] (استاندارد شماره ۲-۱۰ *APA*) را به همراه گزارش‌های فنی آزمون توصیه کرده‌اند (*AERA*، ۲۰۱٤). خطای استاندارد اندازه‌گیری شرطی، می‌تواند 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 قابل محاسبه است (وودروف و همکاران[٦]، 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 نمره مشاهده شده هر شرکت‌کننده در طی برگزاری آزمون‌های موازی در شرایط مشابه است، البته با این فرض که نمره حقیقی او ثابت بماند (هارتل، ۲۰۰٦).

کولن (۱۹۹۲)، فلت و کوالس[۷] (۱۹۹٦) و برنان و لی[۸] (۱۹۹۹) رویکردهای متفاوتی را برای محاسبه *CSEM* پیشنهاد کرده‌اند که به دلیل سهولت محاسبه و نیاز نداشتن 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

1. Overall standard error of measurement
2. American Educational Research Association
3. National Council on Measurement in Education
4. American Psychological Association
5. Conditional standard error of measurement (CSEM)
6. Woodruff et al
7. Feldt & Qualls
8. Brenan & Lee

قوی لرد[1] (۱۹۵۵ و ۱۹٦٥) 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[2]، احتمال شرطی اینکه شخصی از مجموع k سؤال در یک آزمون بتواند به y تا از آنها پاسخ صحیح بدهد با رابطه (۱۱) محاسبه می‌شود:

$$p(y\,|\,\pi,k)=\binom{k}{y}\pi^{y}(1-\pi)^{y-k} \qquad (11)$$

در رابطه (۱۱) پارامتر π نمره حقیقی نسبت پاسخ‌های درست برای هر شخص است و y متغیر تابع احتمال است. لرد (۱۹٦٥) مقدار خطای اندازه‌گیری را با توجه به توزیع دوجمله‌ای رابطه (۱۱) برای شخصی که می‌تواند به x سؤال پاسخ درست بدهد (x نمره خام است) طبق رابطه (۱۲) محاسبه کرده است:

$$\hat{\sigma}_{E(x)}=\sqrt{\frac{x(k-x)}{k-1}}=c_k\sqrt{\frac{x(k-x)}{k}} \qquad (12)$$

$c_k=\sqrt{\frac{k}{k-1}}$ همان عامل تصحیح سوگیری برآورد است چون در رابطه (۱۱) به‌منظور برآورد π از مقدار $\bar{x}=\frac{x}{k}$ استفاده می‌شود. عبارت $c_k=\sqrt{\frac{k}{k-1}}$ باعث می‌شود که $\hat{\sigma}_{E(x)}$ برآوردگر نااریبی 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_k در 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 به صورت زیر نیز نوشت (مود، گری بیل و بوز[3]، ۲۰۰۸):

$$\sigma^{2}(Y\,|\,X)=E((Y-E(Y\,|\,X))^{2}\,|\,X) \qquad (13)$$

اگر توزیع شرطی نمره‌ها به صورت $p(Y\,|\,X)$ تعریف شده باشد رابطه (۱۳) به صورت رابطه (١٤) درخواهد آمد:

---
1. Lord
2. Strong true score theory
3. Mood, Gray bill & Boes

$$\sigma^{2}(Y|X)=\sum_{y=0}^{k}\left((Y-\mu(Y|X))^{2}p(Y|X)\right) \quad (۱۴)$$

رابطه (۱٤) ساده‌تر هم می‌شود به‌گونه‌ای که به میانگین شرطی نیازی نباشد:

$$\sigma^{2}(Y|X)=\sum_{y=0}^{k}Y^{2}p(Y|X)-\left(\sum_{y=0}^{k}Y p(Y|X)\right)^{2} \quad (۱۵)$$

خطای استاندارد اندازه‌گیری شرطی نیز واریانس شرطی خطاها به شرط هر نمره هستند پس به کمک رابطه (۱٦) خطای استاندارد اندازه‌گیری شرطی قابل محاسبه است:

$$\sigma_{E}^{2}(y|x)=\frac{k}{k-1}\left(\sum_{y=0}^{k}y^{2}p(y|\pi,k)-\left(\sum_{y=0}^{k}yp(y|\pi,k)\right)^{2}\right) \quad (۱٦)$$

برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس طبق نظریه نمره حقیقی قوی می‌توان در رابطه (۱٦) به جای مقادیر نمره خام y از تبدیل شده‌ی غیرخطی آنها (مثلا نمره‌های نرمال) استفاده کرد:

(۱۷)

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

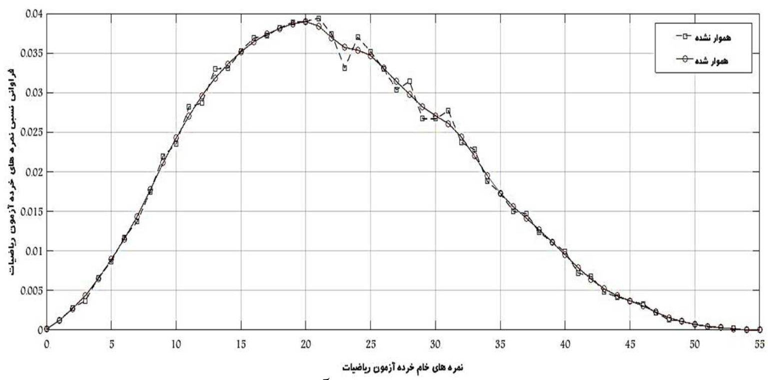
پژوهشگران این مقاله برای محاسبه خطای استاندارد اندازه‌گیری شرطی نمره‌های خام و نمره‌های مقیاس در نرم‌افزار MATLAB برنامه‌ای را توسعه داده‌اند. در این برنامه برای مقایسه روش‌های مختلف مقیاس‌سازی از میانگین خطای استاندارد اندازه‌گیری شرطی که به شکل رابطه (۱۸) محاسبه می‌شود، برای بررسی کارآمدی روش‌ها استفاده می‌شود.

$$\bar{\sigma}^{2}_{E(s(x)|x)}=f(x)\,\sigma^{2}_{E(s(x)|x)} \quad (۱۸)$$

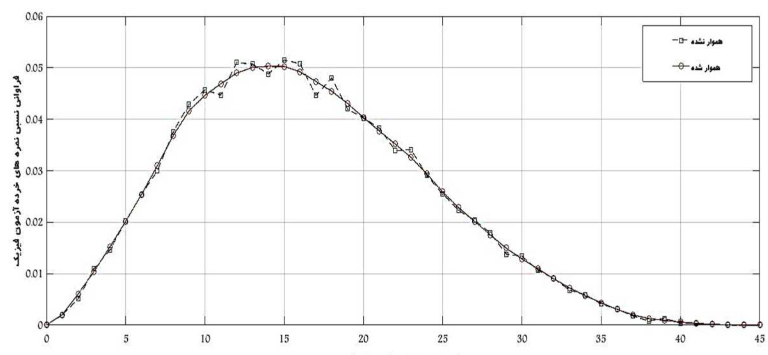
در رابطه (۱۸)، $f(x)$ فراوانی نسبی نمره x است، هر روشی که کمترین میانگین خطای استاندارد اندازه‌گیری شرطی را داشته باشد، کارآمدتر است. $s(x)$ همان نمره‌های مقیاس و $\sigma_{E(s(x)|x)}$ خطای استاندارد اندازه‌گیری شرطی برای هر نمره مقیاس است (کولن و برنان، ۲۰۱٤).

نتایج پژوهش

نمودارهای (۲ و۳) فراوانی نمره‌های خام را در کنار فراوانی هموار شده برای خرده‌آزمون‌های ریاضی و فیزیک نمایش می‌دهند. این دو نمودار از روی داده‌های شبیه‌سازی شده ترسیم شده‌اند. نمودار هموار شده برای هر دو خرده‌آزمون ریاضی و فیزیک به ازای $h = ۱۰$ بهترین هموارشدگی را دارند. برای رسیدن به هموارترین توزیع فراوانی از شاخص M_h استفاده شد، توزیع‌های مختلف به ازای مقادیر زوج پارامتر h محاسبه شده و سپس توزیع‌هایی که کمترین M_h را دارند به‌عنوان هموارترین توزیع انتخاب شدند. مقادیر h برای خرده‌آزمون‌های مختلف، متفاوت بوده، به‌طوری که برخی با مقدار ۲ و برخی با مقادیر٤ و یا ٦ به هموارترین حالت خود دست یافته‌اند.



نمودار (۱) فراوانی نسبی نمره‌های خام خرده‌آزمون ریاضی (شبیه‌سازی شده)



نمودار (۲) فراوانی نسبی نمره‌های خام خرده‌آزمون فیزیک (شبیه‌سازی شده)

در جدول (۱) برخی ویژگی‌های آماری برای نمره‌های خام خرده‌آزمون‌ها پیش از هموارسازی ارائه شده است. این ویژگی‌ها شامل شاخص‌های آماری مانند میانگین، واریانس و شاخص‌های اندازه‌گیری مانند خطای استاندارد اندازه‌گیری و ضریب پایایی کودر- ریچاردسون ۲۰ است. مقادیر داخل پرانتز میانگین نمره‌های خام نسبت هستند. جدول (۲) مربوط به داده‌های واقعی است.

جدول (۱) شاخص‌های آماری و شاخص‌های اندازه‌گیری برای نمره خام خرده‌آزمون‌ها (داده‌های شبیه‌سازی)

	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
میانگین	۱۵/۰۸ (۰/٦۰)	۱۳/۱٦ (۰/۵۲)	۱۳/٤۰ (۰/۵۳)	۱۱/۷٤ (۰/٤٦)	۲۷/۷٤ (۰/۵۰)	۲۰/۷۵ (۰/٤٦)	۱۷/٤۲ (۰/٤۹)
واریانس	۱۹/۸۵	۲۱/۰۹	۲۰/۵۸	۱۹/۸۱	۹۰/۳۰	٦۲/۵۷٦	۳۷/۲۰
SEM	۲/۱٤۹	۲/۱۸۹	۲/۱۷۳	۲/۱٦۰	۳/۲۵	۲/۹۲	۲/۵۵
KR20	۰/۷۲۸	۰/۷۲٤	۰/۷۷۱	۰/۷۱٤	۰/۸٦٤	۰/۸٤۰	۰/۸۷۷

جدول (۲) شاخص‌های آماری و شاخص‌های اندازه‌گیری برای نمره خام خرده‌آزمون‌ها (داده‌های واقعی)

	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
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

کمترین نمره خام برای خرده‌آزمون‌های زبان و ادبیات فارسی، زبان و ادبیات عربی، معارف اسلامی، زبان و ادبیات انگلیسی، صفر و بیشترین نمره خام برای این خرده‌آزمون‌ها 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 ٤۵ و ۳۵ است. از ضریب کودر- ریچاردسون

برای محاسبه پایایی نمره‌های خام استفاده شد که بر اساس داده‌ها در جدول‌های (۱ و ۲) ضریب پایایی نمره‌های خام بین ۰/۷۰ تا ۰/۹ متغیر بوده است. تغییر مقادیر پایایی بین خرده‌آزمون‌ها ارتباط نزدیکی با تعداد پرسش‌های خرده‌آزمون داشته است. به‌طوری که چه در داده‌های شبیه‌سازی شده و چه در داده‌های واقعی آن دسته از خرده‌آزمون‌هایی که تعداد پرسش‌های بیشتری داشته‌اند، داری ضریب پایایی بزرگ‌تر هستند. در روش نرمال‌سازی بدون استفاده از هموارسازی، همان‌طور که در قسمت‌های پیش هم گفته شد، فراوانی نمره‌های خام هموار نشده است و همان‌گونه که در نمودارهای (۱ و ۲) دیده می‌شود نمره‌ها دارای بی‌نظمی و آشفتگی هستند. جدول‌های (۳ و۴) ویژگی‌های آماری و اندازه‌گیری مربوط به این نمره‌ها را برای داده‌های شبیه‌سازی شده و واقعی نمایش می‌دهد. در این جدول‌ها از ویژگی‌های آماری میانگین، واریانس، چولگی و کشیدگی به‌منظور مقایسه گشتاورهای اول تا چهارم نمره‌های هموارشده و هموار نشده استفاده شده است.

جدول (۳) ویژگی‌های آماری و ضریب پایایی نمره‌های مقیاس نرمال بدون استفاده از روش هموارسازی (داده‌های شبیه‌سازی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی[1]
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

[1]. ضریب پایایی در اینجا از رابطه $\rho = 1 - \frac{\sigma_E^2}{\sigma_{s(x)}^2}$ محاسبه می‌شود. این ضریب برای پایایی نمره‌های مقیاس که توزیع مشخصی دارند محاسبه می‌شود (کولن، ۲۰۱۲).

جدول (٤) ویژگی‌های آماری و ضریب پایایی نمره‌های مقیاس نرمال بدون استفاده از روش هموارسازی (داده‌های واقعی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
فارسی	۵۰۰۵/۳	۱۵۱٤۱۰۰	۱/۳۷	−۰/۵۳	۰/۹۵۲
عربی	۵۰۲۰/۳	۱٤۳۳۵۰۰	۱/٤۲	−۰/۷۱	۰/۹٤۰
معارف	۵۰۰٤/٦	۱۵۲۰۷۰۰	۱/۲۵	−۰/۰۷	۰/۹۵۷
زبان خارجه	۵۰۳۸/٦	۱۳٤۰۰۰۰	۱/۵٤	−۰/۲۰	۰/۹٤۰
ریاضی	۵۰۳۹/۷	۱۳۳۷۹۰۰	۱/۲۷	−۱/۳٦	۰/۹۲۲
فیزیک	۵۰٤۰/۸	۱۳۳۳۵۰۰	۱/۳٤	−۱/۱۳	۰/۹۲٤
شیمی	۵۰۵۳/۳	۱۲٦۳٤۰۰	۱/۲۹	−۱/۳۸	۰/۹۱۲

جدول (۵) ویژگی‌های آماری و ضریب پایایی را برای نمره‌های نرمال‌سازی شده با استفاده از روش هموارسازی دوجمله‌ای کرنل در داده‌های شبیه‌سازی نمایش می‌دهد. برای پیش‌هموارسازی نمره‌های خام به ترتیب برای درس‌های ادبیات فارسی، عربی، معارف اسلامی، زبان انگلیسی، ریاضی، فیزیک و شیمی، در روش وارسی مقطعی اعتبار مقادیر ۸،٦،۸،۸،۱۰،۱۰،۱۰ برای پارامتر پیش‌هموارسازی h انتخاب شده است.

جدول (۵) ویژگی‌های آماری و ضریب پایایی نمره‌های نرمال‌سازی با استفاده از روش پیش‌هموارسازی (شبیه‌سازی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
فارسی	٤۹۹۹/۹	۱۵٤۵۰۰۰	−۰/۹٤	−۰/٦٤	۰/۹۵۷
عربی	۵۰۰۱/۳	۱۵٤۱٦۰۰	۰/۸٤	−۰/۵۰	۰/۹۵٦
معارف	۵۰۰۰/۳	۱۵٤۳۹۰۰	۰/٤۵	−۰/۵۹	۰/۹۵۵
زبان خارجه	۵۰۰۱/۷	۱۵۳۸۹۰۰	۱/۰٦	۰/٦۰	۰/۹۵۳
ریاضی	۵۰۰۰/۲	۱۵۵۷٤۰۰	۰/٦۱	−۰/٦۲	۰/۹۷۰
فیزیک	۵۰۰۰/۸	۱۵۵۳۱۰۰	۰/۷٦	−۰/٦۵	۰/۹٦٤
شیمی	۵۰۰۰/٦	۱۵۵۰۱۰۰	۰/۸۱	−۰/٦۱	۰/۹۵۹

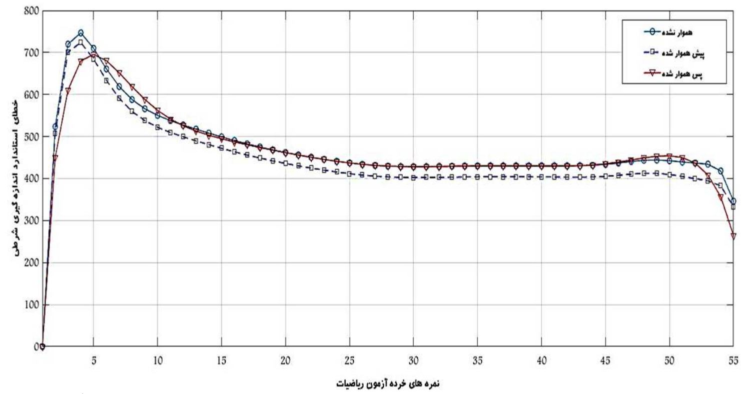
جدول (٦) ویژگی‌های آماری و ضریب پایایی را برای نمره‌های نرمال‌سازی شده با استفاده از روش هموارسازی دوجمله‌ای کرنل در داده‌های واقعی نمایش می‌دهد. برای

پیش‌هموارسازی نمره‌های خام به ترتیب برای درس‌های ادبیات فارسی، عربی، معارف اسلامی، زبان انگلیسی، ریاضی، فیزیک و شیمی، در روش وارسی مقطعی اعتبار مقادیر ۲،۲،٤،۲،٦،٦،٤ برای پارامتر پیش‌هموارسازی h انتخاب شده است.

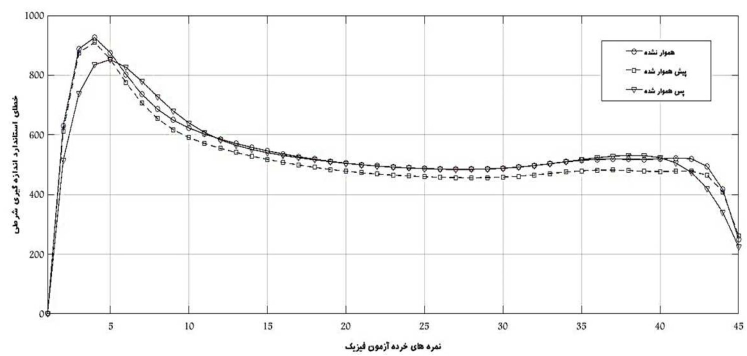
جدول (٦) ویژگی‌های آماری و ضریب پایایی نمره‌های نرمال‌سازی با استفاده از روش پیش‌هموارسازی (واقعی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
فارسی	۵۰۰۵/۵	۱۵۱٤۰۰۰	۱/۳۹	۰/٤۵−	۰/۹۵۲
عربی	۵۰۲۰/۱	۱٤۳٤٤۰۰	۱/٤۲	۰/۷۲−	۰/۹۳۸
معارف	۵۰۰٤/۵	۱۵۲۱۱۰۰	۱/۲۵	۰/۰٦−	۰/۹۵۸
زبان خارجه	۵۰۳٦/۹	۱۳٤۸۱۰۰	۱/۵۳	۰/۲۷−	۰/۹۳۵
ریاضی	۵۰۳۸/۸	۱۳٤۲۳۰۰	۱/۲۷	۱/۳۵−	۰/۹۲۲
فیزیک	۵۰٤۰/۱	۱۳۳٦۱۰۰	۱/۳٤	۱/۱۲−	۰/۹۲۵
شیمی	۵۰۵۲/۲	۱۲۷۰۱۰۰	۱/۲۹	۱/۳۸−	۰/۹۱۳

نتایج جدول‌های (۳، ٤، ۵ و ٦) نشان‌دهنده تغییر ویژگی نمره‌ها نسبت به نمره‌های خام است. به‌طوری که ضمن تغییر در گشتاورهای اول تا چهارم نمره خرده‌آزمون‌ها، پایایی نمره‌های مقیاس نیز نسبت به نمره‌های خام به‌طور محسوسی افزایش پیدا کرد. نکته اول که باید به آن اشاره کرد تغییر در چهار گشتاور اول نمره‌ها نسبت به نمره‌های خام است، به‌طوری که میانگین نمره‌ها و واریانس آنها تغییر محسوسی داشته است و این امر به خاطر تبدیل خطی است که پس از تبدیل نمره‌ها به نمره مقیاس نرمال انجام می‌گیرد. همان‌طور که گفته شد، پس از نرمال‌سازی نمره‌های خام به تبدیل خطی نمره‌های به دست آمده با میانگین ۵۰۰۰ و انحراف استاندارد ۱۲۵۰ اقدام و پس از آن، نمره به دست آمده به نزدیک‌ترین عدد صحیح گرد شده است. تفاوت دیگری که در تبدیل‌های نرمال مشاهده می‌شود، افزایش ضریب پایایی نمره‌های مقیاس نسبت به نمره‌های خام است. این موضوع که هم در داده‌های شبیه‌سازی شده و هم در داده‌های واقعی مشهود است به دلیل تبدیل غیرخطی نمره‌ها پیش آمده است و یکی از نتایج تبدیل غیرخطی نمره‌ها افزایش ضریب پایایی آنها است که این امر در تبدیل خطی نمره‌ها امکان‌پذیر نیست. نمودارهای (۳ و ٤) خطای استاندارد اندازه‌گیری شرطی را برای سه روش و خرده‌آزمون‌های ریاضی و فیزیک نمایش می‌دهند.



نمودار (۳) مقایسه خطای استاندارد اندازه‌گیری شرطی مقیاس نرمال در خرده‌آزمون ریاضی



نمودار (٤) مقایسه خطای استاندارد اندازه‌گیری شرطی مقیاس نرمال در خرده‌آزمون فیزیک

در نمودارهای (۳ و ٤) دیده می‌شود که پیش‌هموارسازی نمره‌های خام موجب کاهش سطح خطای استاندارد اندازه‌گیری شرطی نمره‌ها شده است. در بیشتر سطوح نمره‌ها پیش‌هموارسازی توانسته موجب کاهش خطا شود، البته این گفته برای ابتدا و انتهای طیف نمره‌ها همواره صادق نیست و پیش‌هموارسازی نتوانسته به‌خوبی پس‌هموارسازی باعث کاهش خطای نمره‌ها شود. اما اگر ابتدا و انتهای طیف نمره‌ها را استثنا کنیم در نقاط میانی توزیع نمره‌ها پیش‌هموارسازی نمره‌ها توانسته عملکرد بهتری نمایش دهد به‌طوری که نسبت به روش‌های پس‌هموارسازی و ناهموار خطای نمره‌ها کمتر است. نکته دیگر اینکه، روش پس‌هموارسازی نتوانسته تغییرات عمده‌ای را در خطای استاندارد اندازه‌گیری شرطی نمره‌ها ایجاد کند و حتی در برخی موارد از روش

هموار نشده هم عملکرد ضعیف‌تری داشته است. نمودار خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس نشان می‌دهد که روش پس‌هموارسازی در بیشتر موارد در ابتدا و انتهای طیف توانسته عملکرد مناسبی از خود نشان داده و موجب کاهش خطای استاندارد اندازه‌گیری شرطی شود ولی در میانه طیف این روند ادامه پیدا نکرده است به‌گونه‌ای که نه‌تنها موجب کاهش خطای استاندارد اندازه‌گیری شرطی نشده بلکه در برخی موارد خطای استاندارد اندازه‌گیری شرطی را به نسبت حالت هموار نشده افزایش هم داده است. این نتیجه، نشان از ناکارآمدی روش پس‌هموارسازی برای کاهش خطای استاندارد اندازه‌گیری شرطی نمره‌های نرمال دارد. برخلاف روش پیش‌هموارسازی که توانسته به نحو مطلوبی خطای استاندارد اندازه‌گیری شرطی را نسبت به حالت هموار نشده کاهش دهد و نتایج نشان داده که استفاده از پیش‌هموارسازی رویکرد مطلوبی برای کاهش خطا در تبدیل مقیاس نرمال است.

در این پژوهش، نمره‌های نرمال‌سازی شده را که بدون هموارسازی تولید شدند با استفاده از توابع اسپلاین آن‌گونه که در قسمت‌های قبل توضیح داده شده هموارسازی می‌شوند. این روش که به خاطر انتخاب مقادیر کوچک برای p شباهت زیادی نیز به توابع اسپلاین مکعبی دارد (نحوه انتخاب مقادیر p در بخش روش‌ها توضیح داده شده است) به روش پس‌هموارسازی مشهور است. در این روش نمره‌ها با یک تابع اسپلاین برازش داده می‌شوند، مقدار p انتخاب شده برای برازش داده‌ها مقدار ۰/۰۰۳۹۹۹ انتخاب شده است که برای همه خرده‌آزمون‌ها موجب به دست آمدن مقادیر مطلوب (این شاخص برای همه خرده‌آزمون‌ها مقادیر بالای ۰/۹۶۵ را به دست می‌دهد) شده است. همچنین با انتخاب ۰/۰۰۳۹۹۹ برای پارامتر p، مقدار RMSE ها برای هیچ‌کدام از خرده‌آزمون‌ها از عدد ٦١/١٧ تجاوز نکرده است که این مقادیر نشان‌دهنده برازش تابع‌های اسپلاین برآورد شده با داده‌های تجربی هستند.

جدول‌های (۷ و۸) ویژگی‌های آماری و ضریب پایایی را برای نمره‌های نرمال‌سازی شده با استفاده از روش پس‌هموارسازی اسپلاین در داده‌های شبیه‌سازی و واقعی نمایش می‌دهند.

جدول (۷) ویژگی‌های آماری و ضریب پایایی نمره‌های نرمال‌سازی با استفاده از روش پس‌هموارسازی (داده‌های شبیه‌سازی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
فارسی	۵۰۰۰/۱	۱۵۶۱۷۰۰	-۰/۰۳	-۰/۷۶	۰/۹۵۷
عربی	۴۹۹۷/۵	۱۵۵۸۱۰۰	۰/۸۵	-۰/۶۱	۰/۹۵۷
معارف	۴۹۹۹/۳	۱۵۶۵۳۰۰	۰/۴۶	-۰/۷۲	۰/۹۵۵
زبان خارجه	۴۹۹۷/۳	۱۵۵۸۴۰۰	۱/۰۷	-۰/۶۴	۰/۹۵۳
ریاضی	۵۰۰۰/۸	۱۵۵۷۴۰۰	۰/۶۱	-۰/۶۶	۰/۹۷۰
فیزیک	۵۰۰۱/۸	۱۵۶۰۷۰۰	۰/۷۷	-۰/۷۲	۰/۹۶۴
شیمی	۵۰۰۱/۳	۱۵۶۰۷۰۰	۰/۸۰	-۰/۷۲	۰/۹۵۹

جدول (۸) ویژگی‌های آماری و ضریب پایایی نمره‌های نرمال‌سازی با استفاده از روش پس‌هموارسازی (داده‌های واقعی)

خرده‌آزمون	میانگین	واریانس	چولگی	کشیدگی	ضریب پایایی
فارسی	۵۰۰۶/۱	۱۵۰۶۱۰۰	۱/۳۶	-۰/۵۹	۰/۹۵۱
عربی	۵۰۳۰/۵	۱۳۸۲۹۰۰	۱/۴۱	-۰/۷۷	۰/۹۳۸
معارف	۵۰۰۱/۶	۱۵۳۱۳۰۰	۱/۲۵	-۰/۱۶	۰/۹۵۸
زبان خارجه	۵۰۸۲/۷	۱۲۰۶۵۰۰	۱/۵۴	-۰/۲۶	۰/۹۳۴
ریاضی	۵۰۷۴/۲	۱۱۷۵۳۰۰	۱/۲۷	-۱/۳۵	۰/۹۱۱
فیزیک	۵۰۸۱/۴	۱۱۶۴۵۰۰	۱/۳۴	-۱/۱۲	۰/۹۱۳
شیمی	۵۱۱۳/۷	۱۰۵۴۰۰۰	۱/۲۹	-۱/۳۷	۰/۸۹۵

همان‌طور که دیده می‌شود روش پس‌هموارسازی واریانس نمره‌های مشاهده شده را به‌طور محسوسی افزایش داده و تا حدودی تغییرات جزئی در اندازه پایایی هم به وجود آمده است.

جدول (۹) مجذور میانگین خطای استاندارد اندازه‌گیری شرطی را برای همه خرده‌آزمون‌ها و برای همه روش‌ها نمایش می‌دهد. برای محاسبه میانگین خطای استاندارد اندازه‌گیری شرطی از رابطه (۲۲) استفاده شده است، روش‌هایی که مقدار میانگین خطای استاندارد اندازه‌گیری شرطی آنها کمتر است، کارآمدتر بوده و دقیق‌تر هستند.

جدول (۹) مجذور میانگین خطای استاندارد اندازه‌گیری شرطی برای همه خرده‌آزمون‌ها

نوع مقیاس	فارسی	عربی	معارف	زبان	ریاضی	فیزیک	شیمی
هموار نشده- شبیه‌سازی شده	۲۵۴/۹۲	۲۵۶/۱۳	۲۵۸/۲۹	۲۶۳/۰۱	۲۱۶/۹۹	۲۲۸/۶۵	۲۴۳/۶۴
هموار نشده- واقعی	۲۷۰/۴۷	۲۹۳/۰۷	۲۵۴/۲	۲۸۲/۹۸	۳۲۳/۶۰	۳۱۷/۳۰	۳۳۳/۲۳
پیش هموارشده- شبیه‌سازی شده	۲۳۷/۹۷	۲۳۷/۶۵	۲۳۹/۵۶	۲۴۴/۴۱	۱۹۴/۹۴	۲۰۷/۸۵	۲۲۳/۹۴
پیش هموار -واقعی	۲۶۲/۳۰	۲۸۷/۲۵	۲۴۸/۰۲	۲۷۸/۵۶	۳۱۸/۱۳	۳۱۲/۴۸	۳۲۷/۶۸
پس هموارشده- شبیه‌سازی شده	۲۶۳/۱۷	۲۶۲/۷۷	۲۶۵/۱۴	۲۶۸/۷۷	۲۱۷/۸۲	۲۳۰/۹۶	۲۸۴/۰۹
پس هموارشده- واقعی	۲۷۶/۳۷	۲۹۷/۶۷	۲۵۵/۳۹	۲۸۵/۹۱	۳۲۸/۳۸	۳۲۲/۴۶	۳۳۷/۴۶

جدول (۹) مجذور میانگین خطای استاندارد اندازه‌گیری شرطی را برای همه خرده‌آزمون‌ها و هر سه روش نمایش می‌دهد. مقادیر مجذور میانگین مربوط به روش پیش‌هموارسازی در همه خرده‌آزمون‌ها از روش‌های هموار نشده و پس‌هموارشده، کمتر است. میانگین خطای استاندارد اندازه‌گیری شرطی ارتباط نزدیکی با فراوانی خطای استاندارد شرطی دارد و این کاهش میانگین در روش پیش‌هموارسازی نسبت به روش‌های دیگر ناشی از کاهش بی‌نظمی در این روش است. روش پس‌هموارسازی در این جدول بیشترین مقادیر میانگین خطا را نشان داده است. این موضوع از نمودارهای (۳ و ٤) هم قابل بررسی است و نوسان ناگهانی نمودار پس‌هموارسازی در میانه توزیع باعث افزایش این میانگین شده است، هرچند که ضریب پایایی این روش بالا بوده است که این بالا بودن را می‌توان به افزایش واریانس نمره‌ها در این روش هم نسبت داد. روش پیش‌هموارسازی به علت پایین‌تر بودن خطای استاندارد اندازه‌گیری شرطی نمره‌ها و نوسان کمتر در میانه توزیع، مجذور میانگین خطای کمتری نسبت به روش‌های دیگر داشته است.

بحث و نتیجه‌گیری

این پژوهش به‌منظور بررسی نقش هموارسازی در خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس نرمال انجام گرفته 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 و۱۰۰۰۰ نمونه واقعی آزمون سراسری سال ۱۳۹۵ در گروه آزمایشی ریاضی و فنی هستند به سه روش تبدیل مقیاس نرمال بدون هموارسازی، تبدیل مقیاس نرمال با پیش‌هموارسازی فراوانی نمره‌ها و تبدیل مقیاس نرمال با پس‌هموارسازی نمره‌ها 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. برای پیش‌هموارسازی فراوانی نمره‌ها از روش دوجمله‌ای کرنل و برای پس‌هموارسازی نمره‌های مقیاس از توابع اسپلاین بهره گرفته شد. از نظریه نمره حقیقی قوی برای بررسی خطای استاندارد اندازه‌گیری شرطی نمره‌های مقیاس و همچنین محاسبه پایایی نمره‌ها استفاده شد. برای تحلیل هموار بودن نمره‌ها از گشتاورهای اول تا چهارم در کنار وارسی مقطعی اعتبار بهره برده و با استفاده از مجذور میانگین خطای استاندارد اندازه‌گیری شرطی دقت و کارآمدی نمره‌های مقیاس ساخته شده، بررسی شد.

با توجه به هدف از اجرای این پژوهش که بررسی تأثیر هموارسازی بر خطای استاندارد اندازه‌گیری شرطی نمره‌های ساخته شده به روش تبدیل مقیاس نرمال بود به پرسش‌های پژوهش پاسخ داده شد.

پرسش اول: استفاده از روش‌های هموارسازی (پیش و پس هموارسازی) چه تأثیری بر دقت نمره‌های مقیاس نرمال خواهد داشت؟ با مراجعه به مقادیر ضریب پایایی ارائه شده در جدول‌های (۳ تا ۸) مشاهده می‌شود که هیچ‌کدام از روش‌های نرمال‌سازی نتوانسته‌اند تغییر فاحشی در مقدار ضریب پایایی ایجاد کنند. به‌گونه‌ای که در برخی از موارد یا مقادیر پایایی تغییری نکرده یا در صورت تغییر به اندازه چندهزارم تغییر در ضریب پایایی مشاهده شده است که نشان می‌دهد روش‌های پیش‌هموارسازی و یا پس‌هموارسازی نتوانستند پایایی نمره‌های مقیاس را به میزان قابل توجهی افزایش دهند. این نتیجه هم در داده‌های شبیه‌سازی و هم در داده‌های واقعی دیده می‌شود.

پرسش دوم: کارآمدترین تبدیل مقیاس نرمال در میان روش‌های تبدیل مقیاس نرمال، با توجه به میزان خطای استاندارد اندازه‌گیری شرطی هر روش کدام است؟ در پاسخ به این پرسش باید گفت که با استفاده از روش پیش‌هموارسازی، می‌توان تا حد مطلوبی خطای استاندارد اندازه‌گیری شرطی را کاهش داد؛ به این معنی که در میان روش‌های:

۱- نرمال‌سازی هموار نشده، ۲- نرمال‌سازی با پیش‌هموارسازی و ۳- نرمال‌سازی با پس‌هموارسازی، روش نرمال به همراه پیش‌هموارسازی به دلیل کمتر بودن میانگین خطای استاندارد اندازه‌گیری شرطی (طبق جدول ۹) کارآمدتر بوده و توانسته میانگین خطای کمتری را به خود اختصاص دهد. نتایج نشان‌دهنده برتری روش پیش‌هموارسازی نمره‌ها در کاهش خطای استاندارد اندازه‌گیری شرطی برای توزیع نمره‌ها بود؛ به‌طوری که با وجود خطای استاندارد اندازه‌گیری شرطی بالایی که هر سه روش برای ابتدا و انتهای توزیع نمره‌ها از خود نشان دادند، روش پیش‌هموارسازی توانسته ضمن کاهش میزان خطای استاندارد اندازه‌گیری شرطی برای میانه توزیع موجب کاهش نوسان‌های خطا شده و میانگین خطای استاندارد اندازه‌گیری شرطی این روش از دو روش دیگر کمتر بوده است. جدول (۹) کاهش خطای استاندارد اندازه‌گیری شرطی به این روش را برای داده‌های شبیه‌سازی شده و واقعی نشان می‌دهد و به نظر می‌رسد استفاده از پیش‌هموارسازی فراوانی نمره‌ها برای تبدیل نمره‌های خام به نمره‌های مقیاس نرمال می‌تواند باعث کاهش خطای استاندارد اندازه‌گیری برای همه سطوح نمره‌ها شده و تا حدودی به هم‌ترازی خطا برای همه سطوح نمره‌ها کمک کند.

نکته دیگر اینکه، هرچند روش پس‌هموارسازی نتوانست میانگین مناسبی برای خطای استاندارد اندازه‌گیری شرطی به دست بدهد ولی با افزایش واریانس نمره‌های مقیاس توانسته تا حدودی تفکیک مناسب‌تری بین سطوح نمره‌ها ایجاد کند و کمی از فشردگی نمره‌ها در میانه توزیع بکاهد. به‌طور کلی، تبدیل نمره‌های خام به نمره‌های مقیاس به روش نرمال‌سازی باعث افزایش واریانس نمره‌ها و همچنین میزان دقت آنها می‌شود، ولی همچنان همسان نبودن میزان خطا برای همه سطوح نمره‌ها یکی از مشکلات جدی این روش تبدیل مقیاس است. هرچند که پیش‌هموارسازی هم در داده‌های شبیه‌سازی و هم واقعی تا حدودی توانسته این همسانی را به وجود آورد ولی هنوز برای همه سطوح نمره‌ها به یک اندازه خطا نشده و افراد با نمره‌های متفاوت دارای خطای استاندارد اندازه‌گیری شرطی یکسانی نیستند. این موضوع در ابتدای توزیع نمره‌ها و در انتهای آن به روشنی دیده می‌شود.

در این پژوهش تنها از یکی از روش‌های تبدیل نمره‌های خام به نمره‌های مقیاس استفاده شد. ضمن اینکه رویکرد استفاده از روش‌های هموارسازی در آن تنها محدود به دو روش دوجمله‌ای کرنل و توابع اسپلاین بوده که می‌توان از سایر رویکردهای

هموارسازی بهره برد. یکی از این روش‌ها، روش هموارسازی توزیع بتای چهارپارامتری (لرد، ١٩٦٥) است. نکته آخر اینکه در این پژوهش از توزیع دوجمله‌ای نمره حقیقی قوی برای توزیع شرطی نمره‌های مشاهده شده و حقیقی بهره گرفته شد که با بسط این توزیع می‌توان از توزیع دوجمله‌ای گسترش یافته (کولن، ۱۹۹۲) نیز برای بررسی خطای استاندارد اندازه‌گیری شرطی استفاده کرد.

منابع

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

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

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.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorridike (Ed.). *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education. (Reprinted as 'W. A. Angoff, Scales, norms, and equivalent scores'. Princeton, NJ: Educational Testing Service, 1984.).

Brennan, Robert L. & Lee, Won-Chan (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.

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

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.

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*, Retrieved: www.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. (2014). *Test Equating, Scaling and Linking*, 3rd 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.

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. (1955). Estimating Test Reliability. *ETS Research Bulletin Series*, 1955, 1–17.

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

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

Liu, C. (2011). A comparison of statistics for selecting smoothing parameters for log-linear pre-smoothing and cubic spline post-smoothing under a random groups design (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3461186).

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

Moses, T. & Holland, P. W. (2009). Selection strategies for univariate log-linear smoothing models and their effect on equating function accuracy. *Journal of Educational Measurement*, 46, 159–176.

SAT (2015). *SAT technical manual*. Retrieved from collegereadiness.collegeboard.org.

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.