Journal of Measurement & Educational Evaluation Studies
Vol. 6, No. 15, Autumn 2016

فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال ششم، شمارهٔ ۱۵، پائیز ۱۳۹۵، صص۳۷- ۱۱

تأثیر حجم نمونه و طول آزمون بر نمرات همتراز شده و خطای همترازسازی: مورد مطالعه آزمون‌های ملی ایران

جلیل یونسی*

تاریخ دریافت مقاله: ۹۵/۰۱/۲۶

تاریخ پذیرش مقاله: ۹۵/۰۸/۰۴

چکیده

هدف از انجام پژوهش حاضر ارزیابی تأثیر حجم نمونه و طول آزمون بر نمره‌های همتراز شده و خطای همترازسازی روش کرنل (KE) (با شیوه‌های مختلف هموارسازی رشته‌ای[1] و PSE[2]) و همچنین مزایا و معایب این روش در مقایسه با تکنیک‌های همترازسازی کلاسیک بوده است. جامعه آماری و گروه نمونه پژوهش حاضر، داده‌های آزمون‌های ملی ایران (آزمون تولیمو و آزمون‌های جامع کنکورهای آزمایشی شرکت تعاونی سازمان سنجش آموزش کشور در سال ۹۲-۹۱) بوده است. آزمون تولیمو دارای ۱۷ سؤال لنگر در هر فرم و ۱۲۳ سؤال بود. در آزمون های جامع کنکورهای آزمایشی شرکت تعاونی سازمان سنجش آموزش کشور صرفاً از سؤال‌های مشترک درس‌های عمومی رشته‌های ریاضی- فیزیک، علوم تجربی و علوم انسانی استفاده شد. به‌منظور بررسی تأثیر حجم نمونه بر دقت نتایج همترازسازی، از مجموعه داده‌های مورد نظر به طور کاملاً تصادفی سه نمونه ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری انتخاب و تحلیل شدند. برای بررسی تأثیر طول آزمون بر دقت نتایج همترازسازی از درس‌های عمومی آزمون‌های جامع کنکورهای آزمایشی سنجش نمونه‌ای ٤٠ تایی از سؤال‌ها (از هر درس ۱۰ سؤال) به‌طور کاملاً تصادفی انتخاب شد. بدین ترتیب در آزمون‌های جامع دو آزمون ۱۰۰ و ٤٠ سؤالی در حجم‌های نمونه مختلف مورد تحلیل قرار گرفته است. طرح همترازسازی مناسب در آزمون تولیمو طرح گروه‌های غیر همتا با آزمون لنگر (NEAT[3]) و در آزمون‌های جامع طرح گروه‌های همسان بوده است. روش همترازسازی در آزمون‌های مورد نظر، روش میانگین، روش خطی، روش همصدک، روش قوس دایره‌ای (Circle arc) و روش کرنل (KE) بوده است. به‌طور کلی هرچه حجم نمونه آزمون‌شوندگانی که نمراتشان در تحلیل همترازسازی وارد می‌شود بیشتر باشد، خطای استاندارد همترازسازی کوچک‌تر خواهد بود. نتایج تحلیل‌ها به‌طور کلی نشان داد که همچنان که حجم نمونه افزایش یافته، برازش مربوط به هموارسازی کرنل نیز بهبود یافته است و بهبود هموارسازی کرنل با افزایش طول آزمون همراه بوده است. به‌طور کلی زمانی که حجم نمونه کوچک باشد، روش کرنل بزرگ‌ترین مزیت‌ها را بر سایر روش‌های همترازسازی کلاسیک دارد.

واژگان کلیدی: همترازسازی، روش کرنل (KE)، آزمون تولیمو، سؤالات لنگر، نظریه کلاسیک آزمون (CTT)، روش قوس دایره‌ای، خطای همترازی

---

* عضو هیئت علمی گروه سنجش و اندازه‌گیری دانشگاه علامه طباطبائی (نویسنده مسئول:
jalilyounesi@gmail.com)

1. chain equating (CE)
2. poststratification equating (PSE)
3. The non-equivalent groups with anchor test (NEAT) design

مقدمه

امروزه آزمون‌هایی از قبیل آزمون‌های ورودی دانشگاه‌ها، آزمون‌های استخدام ادواری و آزمون‌های دوره‌ای زبان انگلیسی از مهم‌ترین آزمون‌های مطرح در کشور هستند. این آزمون‌ها از آن جهت مهم هستند که وسیله اتخاذ تصمیم‌های سرنوشت‌ساز در مورد افراد هستند. از طرف دیگر، با گسترش روزافزون کاربرد آزمون‌ها موضوع رعایت عدل و انصاف در تمامی مراحل آزمونگری از اهمیت قابل ملاحظه‌ای برخوردار می‌شود. به‌طور مثال، در شرایطی که فرم‌های متعددی از یک آزمون وجود دارد باید شرایطی به وجود آید که دشواری فرم‌های مختلف تعدیل شود و نمره‌های حاصل از فرم‌های مختلف قابل مقایسه با یکدیگر باشند. استفاده از فرم‌های مختلف هر آزمون ایجاب می‌کند که فرم‌ها دارای محتوا و ویژگی‌های آماری یکسان باشند تا صفت خاصی را به‌طور یکسان اندازه‌گیری کنند. بر این اساس، سازندگان و پرورش دهندگان آزمون‌ها از ویژگی‌های آزمون به‌عنوان خط راهنما استفاده می‌کنند تا فرم‌های آزمون از نظر محتوا و ویژگی‌های آماری تا حد ممکن شبیه هم باشند؛ این فرم‌ها را فرم‌های موازی یا معادل گویند. درمجموع می‌توان گفت که معادل‌سازی آزمون‌ها از آن جهت شایان اهمیت است که موجب تصمیم‌گیری‌هایی دقیق‌تر و عادلانه‌تر می‌شود (برنان[1]، ۲۰۰۶؛ لرد، ۱۹۸۰؛ ترجمه دلاور و یونسی، ۱۳۹۱).

اگر نتوان گفت همترازسازی در همه برنامه‌های سنجش، می‌توان گفت در اغلب برنامه‌های سنجش امری ضروری است و همترازسازی نمره‌های آزمون‌ها از شیوه‌های افزایش عدالت[2] در سنجش است. چون راهی برای تضمین این موضوع نیست که فرم‌های آزمون دارای دشواری یکسانی هستند، در برنامه‌های سنجش باید بتوان این فرم‌ها را به‌گونه‌ای همتراز کرد تا اطمینان حاصل شود که نمره‌های آزمون‌ها قابل مقایسه‌اند؛ بدین معنا که فرقی نکند که آزمودنی به کدام فرم آزمون پاسخ دهد (ون داویر، هالند و تایر[3]، ۲۰۰٤).

همترازسازی دربردارندهٔ عوامل متعددی است که می‌تواند بر نتایج نهایی تأثیرگذار باشد. مقایسه روش‌های همترازسازی اغلب به پارامترها و عوامل مختلفی بستگی دارد که از آن بین می‌توان به ۱) حجم نمونه، ۲) تفاوت در توانایی گروه‌هایی که دو فرم آزمون را دریافت کرده‌اند، ۳) طول آزمون و ٤) ویژگی‌های آزمون لنگر اشاره کرد. به

---

1. Brennan
2. fairness
3. von Davier, Holland & Thayer

همین خاطر بیشتر تحقیقات همترازسازی بر این نکته معطوف شده‌اند که آیا عملکرد روش‌های مورد نظر در شرایط معین متفاوت است یا خیر.

بر اساس نتایج پژوهش‌های سینارای و هالند[1] (۲۰۰۷) به‌طور کلی طول آزمون لنگر، طول آزمون و حجم نمونه رابطه مثبتی با عملکرد همترازسازی دارند، در حالی که تفاوت‌های موجود در توانایی گروه و متغیر بودن دشواری آزمون لنگر به‌طور منفی بر نتایج همترازسازی اثر می‌گذارند.

گادفری[2] (۲۰۰۷) نیز عملکرد همترازسازی چندگانه را با توجه به عوامل تفاوت‌های موجود در توانایی گروه، حجم نمونه‌ها، طول آزمون لنگر و طول آزمون بررسی کرده است. سه نمونه با اندازه‌های متفاوت ۱۰۰۰، ۱۰۰۰۰، ۱۰۰۰۰۰ در این مطالعه شبیه سازی شد. طول آزمون مورد بررسی شامل ۲۰ سؤال، ٦۰ سؤال و ۱۰۰ سؤال بود. سؤال‌های لنگر مورد استفاده ۲۰ درصد، ۳۵ درصد و ۵۰ درصد طول کل آزمون را تشکیل می‌دادند. میزان توانایی گروه‌ها مورد مقایسه قرار گرفت و میزان توانایی گروه‌ها متفاوت بود. نتایج نشان داد که در شرایط واقعی مختلف روش همترازسازی بر نمره‌های آزمودنی‌ها به‌طور متفاوتی تأثیر می‌گذارد. برای نمونه هنگامی که حجم نمونه زیاد باشد شیوه‌های همترازسازی تمایل دارند نتایج کاملاً مشابهی را ارائه کنند.

هانسون و بگوئین[3] (۲۰۰۲) برای بررسی عملکرد شیوه‌های همترازسازی نظریه سؤال- پاسخ در دو نمونه دارای سؤال‌های مشترک، یک مطالعه شبیه‌سازی انجام دادند. سه عاملی که در مطالعه آنها بررسی شد، حجم نمونه (۳۰۰۰ یا ۱۰۰۰)، تعداد سؤال‌های مشترک (۱۰ یا ۲۰) و اینکه توانایی گروه‌ها در این آزمون برابر است (میانگین هر دو ۰) یا نابرابر (میانگین یک گروه = ۰ و میانگین گروه دیگر = ۱) بود. با بررسی هر یک از شیوه‌های همترازسازی نظریه سؤال- پاسخ، مشخص شد زمانی که حجم نمونه زیادتر باشد خطای مجذور میانگین (MSE) پایین‌تر خواهد بود. آنها همچنین دریافتند که خطای مجذور میانگین همترازسازی در گروه‌های برابر کمتر از گروه‌های نابرابر است (هانسون و بگوئین، ۲۰۰۲).

حجم نمونه در مطالعات همترازسازی یک متغیر مشترک محسوب می‌شود. مرور ادبیات مرتبط با این موضوع آشکار کرد که پژوهشگران انتخاب‌های بسیار متفاوتی داشته‌اند. به‌طور کلی زمانی که حجم نمونه کوچک باشد روش کرنل بزرگ‌ترین

---

1. Sinharay & Holland
2. Godfrey
3. Hanson & Béguin

مزیت‌ها را نسبت به سایر روش‌های همترازسازی کلاسیک دارد. گرانت و همکاران[1] (۲۰۰٦) از طرح NEAT برای مطالعه همترازسازی نمونه کوچک در KE استفاده کردند تا تأثیر آن را بر خطای استاندارد همتراز سازی بررسی کنند.

عملکرد روش کرنل در نمونه‌هایی با حجم ۱۰۰۰، ۵۰۰، ۲۵۰، ۱۲۵ و ۷۵ مورد مقایسه قرار گرفت. زمانی که حجم نمونه کوچک می‌شد، ناهمواری بیشتری در توزیع نمره‌ها مشاهده شد. همچنان‌که انتظار می‌رفت نتایج نشان داد که با کاهش حجم نمونه دقت همترازسازی کاهش می‌یابد. افزایش حجم یک نمونه کوچک‌تر، نتایج همترازسازی را بیش از افزایش حجم یک نمونه بزرگ‌تر بهبود می‌بخشید؛ اما حجم نمونه ۷۵ و ۱۲۵ خیلی کوچک بودند و در عمل به‌ندرت مورد استفاده قرار می‌گرفتند. واقعیت آن است که حجم نمونه بزرگ معمولاً اطلاعات بیشتری را دربارهٔ برآورد توزیع نمره‌ها و برازش مدل‌های نظریه سؤال- پاسخ فراهم می‌کند؛ چراکه میزان خطای تصادفی در همترازسازی را کاهش می‌دهد و متعاقب آن نتایج همترازسازی بهبود می‌یابد (کولن و برنان[2]، ۲۰۰٤). به همین نحو افزایش حجم نمونه می‌تواند به دقت بیشتر نتایج همترازسازی منجر شود. در مقابل، کافی نبودن حجم نمونه می‌تواند با کاهش کیفیت برآورد کردن پارامتر و کاهش سودمند بودن برآوردهای لنگر بر روی همترازسازی تأثیر منفی بگذارد (پترسن و کوک[3]، ۱۹۸۹).

در نمونه‌هایی که حجم نمونه نیز پایین است همترازسازی دقیقی باید صورت گیرد. کولن و برنان (۲۰۰٤) پیشنهاد می‌کنند هنگام استفاده از شیوه‌های همترازسازی خطی کمترین حجم نمونه در مورد هر فرم ٤۰۰ نفر و هنگام استفاده از همترازسازی معادل درصدی هر فرم ۱۵۰۰ نفر باشد. در همترازسازی نظریه سؤال- پاسخ در مدل راش ٤۰۰ نفر و در مدل سه‌پارامتری ۱۵۰۰ نفر نیاز است. علاوه بر این آنها نشان دادند که برای به دست آوردن دقتی در همین سطح، طرح گروه‌های معادل[4] نسبت به طرح گروه منفرد[5] و طرح گروه‌های غیر همتا با آزمون لنگر (NEAT) به حجم نمونه بزرگ‌تری نیاز دارد (کولن و برنان، ۲۰۰٤). لی و ون داویر[6] (۲۰۱۰) در پژوهشی به شیوهٔ شبیه‌سازی، تأثیر تعداد سؤال‌ها و حجم نمونه را بر سودمندی هموارسازی کرنل

---

1. Grant et al
2. Kolen, & Brennan
3. Peterson, & Cook
4. equivalent group (EG)
5. single group (SG)
6. Lee & von Davier

به‌منظور برآورد منحنی ویژگی سؤال بررسی کرد. با استفاده از داده‌های شبیه‌سازی شده بر مبنای الگوی سؤال- پاسخ دو پارامتری، سه نمونه با حجم متفاوت (۱۰۰،۲۰۰، ۱۰۰۰) و طول آزمون (۲۰، ٤۰، ۸۰) بررسی شد. نتایج نشان داد در صورتی که طول آزمون و حجم نمونه افزایش یابد مقادیر خطا کاهش و دقت برآوردها افزایش خواهد یافت. با توجه به مطالب بیان شده و پیشینه تجربی پژوهش مهم‌ترین هدف کاربردی این پژوهش، استفاده از نتایج آن در همترازسازی نمره‌های آزمون‌های ملی ایران بوده است. لذا هدف اصلی این پژوهش بررسی اثر حجم نمونه و تعداد سؤال‌ها (طول آزمون) بر روی نمره‌های همتراز شده و خطای همترازسازی بوده است.

روش پژوهش

در این پژوهش از داده‌های واقعی مربوط به اجرای آزمون‌های تولیمو و آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش استفاده شده است. با این اطلاعات دربارهٔ رتبه و وضعیت داوطلبان شرکت‌کننده در آزمون زبان تولیمو و منتخبی از دروس آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش تصمیماتی اتخاذ شده است. از این لحاظ می‌توان با قدرت نتایج پژوهش را تعمیم داد و استواری نتایج این پژوهش را بیشینه ساخت. علاوه بر این، می‌توان با استفاده از داده‌های واقعی حاصل از عملکرد داوطلبان در آزمون‌های مورد نظر، تصنعی بودن شرایط و اثرات منفی را که این عامل به‌تنهایی و یا در تعامل با سایر عوامل بر قابلیت تعمیم یافته‌ها می‌گذارد به حداقل رساند. در این پژوهش محقق در پی مقایسه رویکردهای مختلف در همترازسازی نتایج حاصل از اجرای فرم‌های مختلف آزمون، به‌دست آوردن برآوردهای باثبات‌تر پارامترها (اعم از پارامترهای سؤال و توانایی آزمودنی‌ها) به‌منظور افزایش اعتبار تصمیم‌ها و دقت اندازه‌گیری بوده است؛ لذا روش این پژوهش به‌طور عام جزو پژوهش‌های توصیفی است. پژوهش‌های توصیفی شامل مجموعه روش‌هایی است که هدف آنها توصیف کردن شرایط یا پدیده‌های مورد بررسی است؛ اجرای پژوهش توصیفی می‌تواند صرفاً برای شناخت بیشتر شرایط موجود یا یاری دادن به فرایند تصمیم‌گیری باشد (سرمد، بازرگان و حجازی، ۱۳۸٤).

جامعه آماری و گروه نمونه

به‌منظور دستیابی به اهداف پژوهش، از اطلاعات واقعی به دست آمده از اجرای آزمون تولیمو (۱۳۹۲-۱۳۹۱) و داده‌های آزمون‌های جامع کنکور آزمایشی (درس‌های عمومی مشترک رشته‌های ریاضی، تجربی و علوم انسانی) استفاده شده است. بر این اساس، جامعه آماری و گروه نمونه پژوهش حاضر، جامعه آماری و گروه نمونه شرکت‌کننده در آزمون‌های مذکور می‌شود که در سال تحصیلی ۱۳۹۲-۱۳۹۱ به اجرا درآمده است. البته با توجه به اینکه حجم نمونه از متغیرهای مهم در دقت نتایج همترازسازی است، در این پژوهش نمونه‌های ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری به‌طور کاملاً تصادفی از مجموعه داده‌های آزمون‌های تولیمو و آزمون‌های جامع کنکور آزمایشی سازمان سنجش انتخاب و سپس پارامترها برآورد شده است.

ابزار اندازه‌گیری

داده‌های این پژوهش از طریق آزمون‌های سراسری به‌دست آمده‌اند که توسط سازمان سنجش آموزش کشور در سال تحصیلی ۹۲-۹۱ اجرا شده‌اند. به‌طور دقیق از داده‌های حاصل از آزمون تولیمو در سال ۹۲-۹۱ و داده‌های آزمون‌های جامع کنکور آزمایشی در سال ۹۲-۹۱ استفاده شده است.

الف) آزمون تولیمو

هدف از برگزاری آزمون تولیمو سنجش توانش زبانی افرادی است که زبان مادری آنان زبان انگلیسی نیست. نمره‌های حاصل از این آزمون برای مقاصد زیر کاربرد دارد: ۱- پذیرش دانشجو در مقاطع مختلف تحصیلی به‌ویژه کارشناسی ارشد و دکتری در رشته‌های مختلف دانشگاهی؛ ۲- اعطای بورس تحصیلی و یا شرط استخدام قابل استفاده توسط مؤسسات، سازمان‌ها و ارگان‌های دولتی و غیردولتی. آزمون تولیمو هرساله چندین نوبت به‌طور منظم در سطح کشور برگزار می‌شود و همه افراد بدون توجه به سن، جنسیت، قومیت یا ملیت می‌توانند در آن شرکت کنند. این آزمون که توسط کارشناسان در دفتر آزمون‌سازی سازمان سنجش تهیه می‌شود شامل سه بخش است که در جدول (۱) ساختار کلی آن نشان داده شده است.

جدول (۱) ساختار آزمون تولیمو

بخش‌های آزمون	هدف	بخش	تعداد سؤال
۱- دستور و نگارش، تعداد سؤال‌ها: ۴۰ زمان: ۳۰ دقیقه	سنجش توانایی داوطلبان در انتخاب ساختارهای درست زبان انگلیسی	قسمت A ـ یافتن پاسخ صحیح	۱۵
		قسمت B ـ یافتن گزینه غلط	۲۵
۲- خواندن و درک مطلب زمان: ۵۵ دقیقه	سنجش توانایی داوطلبان در خواندن و درک صحیح متون علمی در موضوعات گوناگون. سؤال‌های مربوط به هر متن دقیقاً اجزای مختلف مهارت خواندن ازجمله یافتن موضوع اصلی مورد بحث متن، استنتاج، حدس معانی کلمات مشکل به کمک بافت کلامی موجود و غیره را مورد سنجش قرار می‌دهد.	۵ یا ۶ متن آکادمیک	۵۰
۳ـ درک شنیداری تعداد سؤال‌ها: ۵۰ زمان: ۳۵ دقیقه	سنجش توانایی داوطلبان در زمینهٔ درک شفاهی مکالمات روزمره و آکادمیک و نیز سخنرانی‌های کلاسی و دانشگاهی	قسمت A ـ مکالمات کوتاه	۳۰
		قسمت B ـ مکالمات طولانی‌تر	۷ یا ۸
		قسمت C ـ سخنرانی‌ها	۱۳ یا ۱۲

بعد از آنکه نمرهٔ خام هر داوطلب در هر یک از بخش‌های آزمون مشخص شد هریک از آنها به نمره تراز که برای تمامی بخش‌های آزمون بین ۲۸ تا ۶۸ است تبدیل می‌شود. سپس سه نمرهٔ تراز با یکدیگر جمع و حاصل جمع در عدد ۱۰ ضرب و مجموعِ حاصل، تقسیم بر عدد ۳ می‌شود. نمرهٔ به دست آمده نمرهٔ کل هر داوطلب محسوب می‌شود. نمره کل هر داوطلب بین ۲۰۰ تا ۶۷۷ متغیر است. نمرهٔ مطلوب در آزمون تولیمو از طرف دفتر آزمون‌سازی سازمان سنجش آموزش کشور تعریف نمی‌شود بلکه دانشگاه‌ها، ادارات، مؤسسات و غیره هستند که برحسب انتظارات خود حد نمره مطلوب را مشخص می‌کنند. ولی به‌طور کلی نمرهٔ زیر ۴۰۰ ضعیف ارزیابی می‌شود و بیشتر دانشگاه‌های انگلیسی‌زبان برای پذیرش دانشجو در مقطع کارشناسی حداقل نمره ۵۵۰ را ملاک قرار می‌دهند. این دانشگاه‌ها برای پذیرش دانشجو در مقطع کارشناسی ارشد و دکتری عموماً نمره‌ای برابر یا بالاتر از ۶۰۰ می‌خواهند. برای هر داوطلب با توجه به نمرات اکتسابی در هر بخش آزمون و نمرهٔ کل، رتبهٔ صدکی خاصی گزارش می‌شود. رتبه صدکی جایگاه نمرهٔ دریافتی هر داوطلب را در مقایسه با نمرات اخذ شدهٔ سایر داوطلبان مشخص می‌کند.

ب) آزمون‌های آزمایشی جامع شرکت تعاونی کارکنان سازمان سنجش آموزش کشور

شرکت تعاونی خدمات آموزشی کارکنان سازمان سنجش آموزش کشور با بررسی‌های کارشناسی مجموعهٔ آزمون‌های آزمایشی را در نُه نوبت طراحی کرده است. شش نوبت از آزمون‌های آزمایشی به‌صورت مرحله‌ای و سه نوبت آزمون جامع برگزار خواهد شد تا داوطلبان شرکت‌کننده در آزمون‌های مرحله‌ای و جامع، با ارزیابی کاملاً علمی و استاندارد، از وضعیت علمی و تحصیلی خود شناخت پیدا کرده و در هر مرحله از آزمون‌ها نسبت به رفع مشکلات تحصیلی خود اقدام کنند. دانش‌آموزان پس از اتمام آزمون‌های مرحله‌ای، با شرکت در سه نوبت آزمون‌های آزمایشی جامع و حضور در جلسات مشابه آزمون سراسری و پاسخگویی به سؤالاتی شبیه به آزمون سراسری، آمادگی خود را روز به روز افزایش داده و موقعیت علمی و تحصیلی خود را در سه نوبت، به‌طور جدی محک می‌زنند.

یافته‌ها

برای تجزیه و تحلیل داده‌ها از نرم‌افزارهای EXCEL، SPSS و R، KE (چن و همکاران[1]، ۲۰۰۷) استفاده شده است. سؤال‌ها و آزمون‌های این پژوهش به‌صورت دوارزشی نمره‌گذاری شده و روش‌های همترازسازی کرنل و همترازسازی نمرهٔ مشاهده شده در آن مورد توجه بوده است. برای انجام روش همترازسازی کرنل از نرم افزار KE 3.0 (چن و همکاران، ۲۰۰۷) استفاده شده است. درنهایت، برآوردهای پارامترهای سؤال‌های دو فرم آزمون و همچنین برآوردهای توانایی با استفاده از برنامه‌ای در نرم‌افزارR (تیم مرکزی گسترش نرم‌افزار[2]R، ۲۰۱۰) همتراز شده‌اند. در این مطالعه فرم‌های مختلف آزمون با روش‌های مختلفی که پیش‌تر بیان شد همتراز شدند و نتایج با استفاده از ملاک‌هایی ارزیابی و مقایسه شدند. از دو طول آزمون (دو سطح)، حجم نمونه (سه سطح)، برای تولید ٦ موقعیت برای هر روش همترازسازی استفاده شد.

[1]. Chen et al
[2]. R Development Core Team

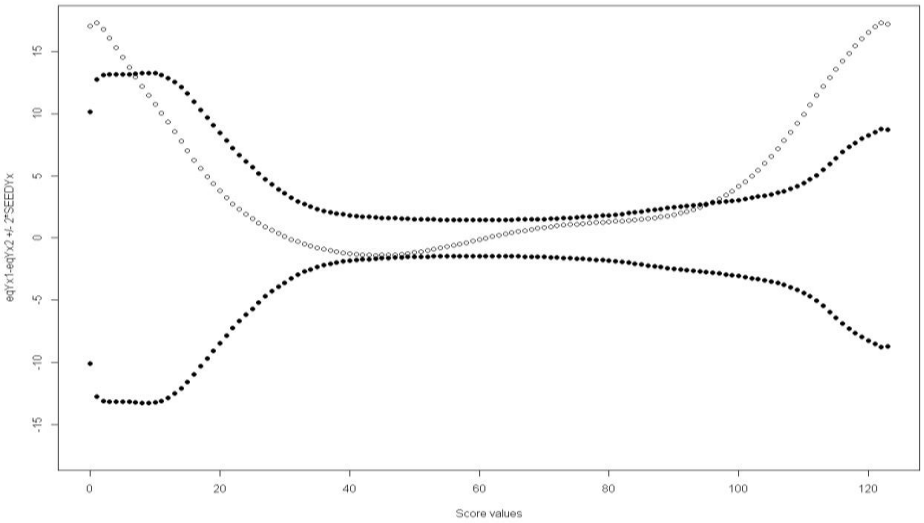
الف) آزمون تولیمو

در جدول زیر این موضوع بررسی شده که در یک روش به‌خصوص همترازسازی (روش میانگین، روش خطی، روش همصدک، روش قوس دایره‌ای، روش کرنل به شیوهٔ هموارسازی رشته‌ای (CS) و روش کرنل به شیوهٔ هموارسازی PostStratification) در داده‌های آزمون‌های تولیمو، حجم نمونه چه تأثیری بر شاخص‌های آماری مرتبط (میانگین، انحراف استاندارد، کجی، کشیدگی و دقت همترازسازی) داشته است.

جدول (۲) اثر حجم نمونه بر چهار گشتاور اول (میانگین و...) نمرات مشاهده شده در روش همترازسازی کرنل به شیوه هموارسازی رشته‌ای

حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندارد کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
۲۰۰	۵۳/۹۹	۹/۲۰	-۰/۱۰۶	۰/۱۷۲	-۰/۳۱۳	۰/۳۴۲	۴/۱۲۷
۵۰۰	۵۳/۰۵۵	۱۰/۱۶۱	۰/۱۹۱	۰/۱۰۹	۰/۵۰۵	۰/۲۱۸	۳/۳۰۸
۱۰۰۰	۵۴/۳۳	۹/۵۷	۰/۱۸۳	۰/۰۷۷	۰/۲۱۶	۰/۱۵۵	۲/۰۴۸

بر اساس اطلاعات جدول (۲)، در روش همترازسازی کرنل به شیوهٔ هموارسازی رشته‌ای (CS) تأثیر حجم نمونه بر شاخص میانگین در نمونه ۲۰۰ نفری برابر ۵۳/۹۹، در نمونه ۵۰۰ نفری برابر ۵۳/۰۵ و در نمونه ۱۰۰۰ نفری برابر ۵٤/۳۳ بوده است. به همین ترتیب تأثیر حجم نمونه بر سایر گشتاورها به‌خوبی در جدول (۲) قابل مشاهده است. بر این اساس، در این روش همترازسازی در نمونه ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است. در شکل (۱) تأثیر حجم نمونه ۲۰۰ نفری و ۱۰۰۰ نفری در روش کرنل به شیوهٔ هموارسازی رشته‌ای (CS) به‌خوبی قابل مشاهده است. توجه به این نکته مهم است که شکل فوق بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۲۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر به دست آمده و بر آن اساس، فاصله اطمینان ترسیم شده است.



شکل (۱) تأثیر حجم نمونه بر روش همترازی کرنل به شیوه Chain Smoothing بین دو نمونه ۱۰۰۰ نفری و ۲۰۰ نفری

در نمودار بالا، نقاط توخالی (خاکستری رنگ) نشان‌دهندهٔ تفاوت پارامتر همترازسازی برای یک نمرهٔ خاص است و نقاط سیاه حدود فاصله اطمینان برای تفاوت است. با توجه به شکل به‌خوبی می‌توان مشاهده کرد که اولاً اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۵۰–۳۰ نقاط زیر این خط قرار می‌گیرد و این مطلب بدان معناست که مقادیر عددی نمره‌های همتراز شده نمونه ۲۰۰ تایی در این بازه از مقادیر نمونه ۱۰۰۰ نفری بیشتر بوده است؛ ثانیاً می‌توان مشاهده کرد که در بازهٔ نمرهٔ ۱۰۰–۱۰ نقاط در داخل فاصله اطمینان قرار گرفته‌اند و در اول و آخر بازه (نمره‌های ۱۰–۰ و ۱۲۰–۱۰۰)، نقاط خارج از این فاصله اطمینان قرار گرفته‌اند و این امر یعنی اینکه خطای تفاوت همترازسازی روش کرنل به شیوهٔ هموارسازی رشته‌ای (CS) در دو نمونه ۲۰۰ و ۱۰۰۰ نفری عدد بزرگی است و درواقع تفاوت معنادار است؛ سوم، بازهٔ اطمینان در دامنهٔ ۱۰۰–۳۰ بسته‌تر و در سایر محدوده‌ها بازتر است. باید این نکته را خاطرنشان کرد که در نمودار مذکور، محور عمودی فاصله اطمینان تفاوت بین دو برآورد را نشان می‌دهد (فرمول روی نمودار است) و این درست مثل آزمون تفاوت دو نمونه همبسته است. در اینجا پارامتر همترازسازی Y به X از طریق دو نمونه X1 و X2 انجام شده است و مانند آن است که پارامتر را دو بار اندازه‌گیری کرده باشیم. چون برای هر پارامتر خطای برآورد پارامتر (SE) را هم داریم می‌توان

معناداری تفاوت این دو پارامتر (eqYx1 و eqYx2) را بررسی کرد. حالا این کار را می‌شود به شیوه برآورد نقطه‌ای (همان آزمون t معمولی) و یا با استفاده از فاصله اطمینان انجام داد.

a. با استفاده از برآورد نقطه‌ای

در نظریه همترازسازی، فرض می‌شود که هر پارامتر همترازسازی دارای یک توزیع زیربنایی است (جامعه آماری) و مقدار واقعی آن از نمونه باید برآورد شود. به همین دلیل برای هر پارامتر- درست مثل میانگین یک نمونه- یک خطای استاندارد برآورد هم خواهیم داشت. طبق این تعریف، مقدار eqYx1 درواقع برآورد حداقل مجذورات (OLS) یا درست نمایی بیشینه (ML) پارامتر موردنظر است با خطای SEyx1. به همین طریق برای برآورد دوم آزمون t به‌صورت زیر خواهد بود:

$$t = \frac{eqYx1 - eqYx2}{SE\ of\ difference}$$

در همترازسازی مقدار مخرج با استفاده از خطای استاندارد تفاوت همترازسازی (standard error of equating difference or SEED) به‌دست می‌آید. در آزمون t همبسته، معناداری تفاوت را با استفاده از فاصله اطمینان هم می‌توان بررسی کرد. به این شکل که اگر فاصله اطمینان شامل ۰ (صفر) باشد می‌شود گفت که تفاوت معنادار نیست. برای محاسبه فاصله اطمینان هم باید از مقادیر صورت و مخرج کسر آزمون t استفاده کرد. مثلاً (با فرض آزمون یک دامنه):

$$t = \frac{\bar{x}1 - \bar{x}2}{SE\ of\ difference}$$

$$CI_{95\%}: (\bar{x}1 - \bar{x}2) \pm t_{95\%} \times SE\ of\ difference$$

همین چارچوب را می‌شود به بررسی تفاوت دو پارامتر همترازسازی تعمیم داد. در چارچوب همترازسازی خواهیم داشت:

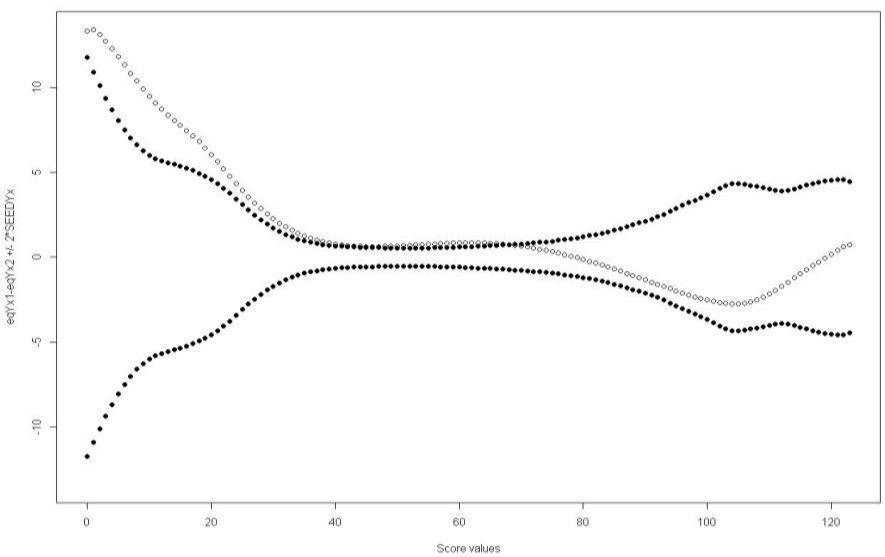
$$CI_{95\%}: (eqYx1 - eqYx2) \pm t_{95\%} \times SEED$$

و چون در همترازسازی با نمونه‌های بزرگ سروکار داریم به‌جای توزیع t از توزیع نرمال استفاده می‌کنیم و درنتیجه بجای $t_{۹۵٪}$ داریم ۱/۹۶ که معادل $Z_{۹۵٪}$ هست. معمولاً در محاسبات این عدد به ۲ گرد می‌شود؛ بنابراین داریم:

$$CI_{۹۵٪}: (eqYx۱ - eqYx۲) \pm ۲ \times SEED$$

این همان چیزی است که روی محور عمودی نمودار داریم. در آن نمودار نقاط توخالی نشان‌دهندهٔ تفاوت پارامتر همترازسازی برای یک نمرهٔ خاص است و نقاط سیاه حدود فاصله اطمینان برای تفاوت است. اگر این تفاوت مشاهده شده در بازه فاصله اطمینان قرار بگیرد به این معنا است که تفاوت مشاهده شده اختلاف معناداری با صفر ندارد. برای هر نمره در آزمون، یک پارامتر همترازسازی برآورد می‌شود. بنابراین برای هر نمره یک خطای استاندارد برآورد متفاوت و متعاقباً یک فاصله اطمینان متفاوت خواهیم داشت.

در شکل زیر تأثیر حجم نمونه ۵۰۰ نفری و ۱۰۰۰ نفری در روش کرنل به شیوهٔ هموارسازی رشته‌ای (CS) به‌خوبی قابل مشاهده است:



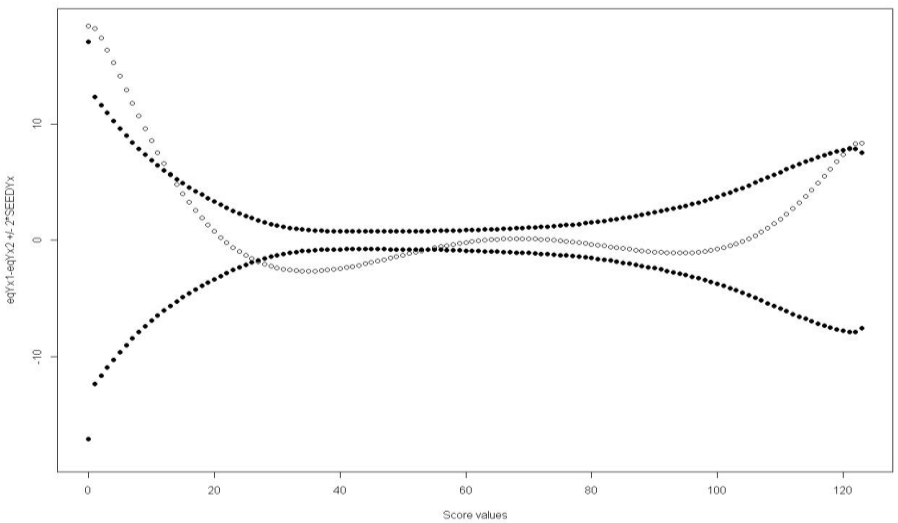
شکل (۲) تأثیر حجم نمونه بر روش همترازی کرنل به شیوه هموارسازی رشته‌ای (CS) بین دو نمونه ۱۰۰۰ نفری و ۵۰۰ نفری

با عنایت به این نکته که شکل فوق بر اساس تفاضل مقادیر عددی نمره‌های همترازشده حجم نمونه ۵۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر به‌دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، به‌خوبی می‌توان مشاهده کرد که اولاً اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۱۱۰–۷۰ نقاط زیر این خط قرار می‌گیرد؛ به این معنا که مقادیر همترازسازی نمونه ۵۰۰ تایی در این بازه از مقادیر نمونه ۱۰۰۰ نفری بیشتر بوده است؛ ثانیاً می‌توان مشاهده کرد که در دامنه ۱۲۰–۴۰، همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند ولی در ابتدای بازه نمرات، همه نمرات تفاوت همترازسازی خارج از این فاصله اطمینان قرار گرفته باشد و این امر بدان معناست که خطای تفاوت همترازسازی روش کرنل به شیوه هموارسازی رشته‌ای (CS) در دو نمونه ۵۰۰ و ۱۰۰۰ نفری در این محدوده عددی بزرگ است و درواقع تفاوت در این بخش معنادار است؛ سوم، بازهٔ اطمینان در دامنهٔ ۸۰–۴۰ بسته‌تر و در سایر محدوده‌ها بازتر است. در جدول (۳) تأثیر حجم نمونه بر چهار گشتاور اول و خطای همترازسازی در روش کرنل به شیوه Post Stratification Smoothing در داده‌های آزمون‌های تولیمو، گزارش شده است.

جدول (۳) اثر حجم نمونه بر چهار گشتاور اول (میانگین و...) نمره‌های مشاهده شده در روش همترازسازی کرنل به شیوه PSE

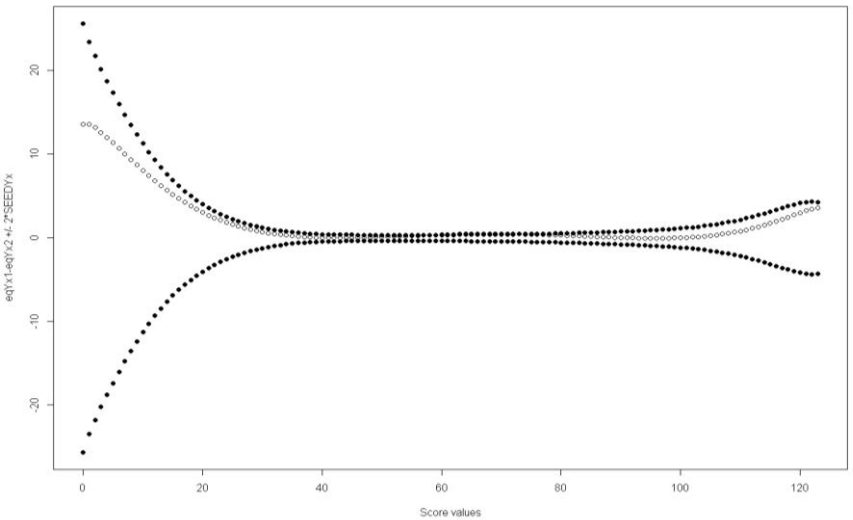
حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندارد کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
۲۰۰	۵۸/۵۷	۶/۴۳	۰/۰۶۵	۰/۱۷۲	۰/۰۹۴	۰/۳۴۲	۲/۷۵۵
۵۰۰	۵۷/۲۳۸	۷/۲۶۴	-۰/۰۲۹	۰/۱۰۹	۰/۱۲۰	۰/۲۱۸	۲/۴۹۴
۱۰۰۰	۵۷/۸۲	۷/۱۸	۰/۰۲۲	۰/۰۷۷	۰/۰۰۹	۰/۱۵۵	۱/۱۷۲

بر اساس اطلاعات جدول (۳)، در روش همترازسازی کرنل به شیوه PSE Smoothing در نمونه ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است. در شکل زیر تأثیر حجم نمونه ۲۰۰ نفری و ۱۰۰۰ نفری در روش کرنل به شیوه PSE Smoothing قابل مشاهده است:



شکل (۳) تأثیر حجم نمونه بر روش همترازی کرنل به شیوه PSE Smoothing بین دو نمونه ۱۰۰۰ نفری و ۲۰۰ نفری

با توجه به شکل (۳) (که بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۲۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر به‌دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است) به‌خوبی می‌توان مشاهده کرد که اولاً اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ٦٠–٢٠ و ١٠٠–٨٠ نقاط زیر این خط قرار می‌گیرد، به این معنا که مقادیر همترازسازی نمونه ۲۰۰ تایی در این بازه از مقادیر نمونه ۱۰۰۰ نفری بیشتر بوده است؛ ثانیاً می‌توان مشاهده کرد که در بازهٔ نمرهٔ ۱۰۰–۱۰ نقاط در داخل فاصله اطمینان قرار گرفته‌اند و در اول و آخر بازه (نمره‌های ۱۰–۰ و ۵۰–۳۰)، نقاط خارج از این فاصله اطمینان قرار گرفته‌اند و این امر یعنی اینکه خطای تفاوت همترازسازی روش کرنل به شیوه PSE Smoothing در دو نمونه ۲۰۰ و ۱۰۰۰ نفری عدد بزرگی است و درواقع تفاوت معنادار است؛ سوم، بازه اطمینان در دامنه ۱۰۰–۲۰ بسته‌تر و در سایر محدوده‌ها بازتر است. در شکل (٤) تأثیر حجم نمونه ۵۰۰ نفری و ۱۰۰۰ نفری در روش کرنل به شیوهٔ PSE Smoothing به‌خوبی قابل مشاهده است.



شکل (٤) تأثیر حجم نمونه بر روش همترازی کرنل به شیوه PSE Smoothing بین دو نمونه ۱۰۰۰ نفری و ۵۰۰ نفری

با عنایت به این نکته که شکل (٤) بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۵۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر به‌دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، به‌خوبی می‌توان مشاهده کرد که اولاً اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۱۱۰–۲۰ (بخش اعظم پیوستار نقاط) نقاط دقیقاً روی این خط قرار می‌گیرد، به این معنا که در این محدوده مقادیر عددی نمره‌های همترازشده نمونه ۵۰۰ تایی از مقادیر نمونه ۱۰۰۰ نفری تفاوت معنادار ندارد؛ ثانیاً می‌توان مشاهده کرد که در تمام طول پیوستار نمره‌های ممکن، همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند و هیچ نقطه‌ای را نمی‌توان یافت که خارج از این فاصله اطمینان قرار گرفته باشد و این امر بدان معناست که خطای تفاوت همترازسازی روش کرنل به شیوهٔ PSE Smoothing در دو نمونه ۵۰۰ و ۱۰۰۰ نفری در این محدوده عددی بسیار کوچک است و درواقع تفاوت در این محدوده معنادار نیست؛ ۳) نکته قابل توجه آنکه بازهٔ اطمینان در دامنه ۱۱۰–۲۰ بسته و فقط در محدوده ۲۰–۰ باز است. در جدول (٤) تأثیر حجم نمونه بر چهار گشتاور اول و خطای همترازسازی در روش‌های مختلف همترازسازی میانگین، همترازسازی خطی، همترازسازی همصدک (equipercentile) و همترازسازی قوس دایره‌ای در داده‌های آزمون‌های تولیمو، گزارش شده است.

جدول (٤) اثر حجم نمونه بر چهار گشتاور اول (میانگین و...) نمره‌های مشاهده شده در روش های مختلف همترازسازی

روش همترازسازی	حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندارد کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
میانگین	۲۰۰	۵۸/۸۵	۱٤/۰۸	۰/۸۰٤	۰/۱۷۲	۰/۸۹٦	۰/۳٤۲	۱/۰٦٦
	۵۰۰	۵۷/۲۸۰	۱٤/۲۱۲	۰/۸۵۰	۰/۱۰۹	۱/۰۹۳	۰/۲۱۸	۰/٦۲۷
	۱۰۰۰	۵۷/۷۷	۱٤/۲٦	۰/۸۰٦	۰/۰۷۷	۱/۲۹۷	۰/۱۵۵	۰/۵۵۳
همترازسازی خطی	۲۰۰	۵۸/۸۵	٦/۷۱	۰/۸۰٤	۰/۱۷۲	۰/۸۹٦	۰/۳٤۲	۱/۵۹٤
	۵۰۰	۵۷/۲۸۰	۷/۱۹۹	۰/۸۵۰	۰/۱۰۹	۱/۰۹۳	۰/۲۱۸	۱/۰۵۰
	۱۰۰۰	۵۷/۷۷	۷/۱٦	۰/۸۰٦	۰/۰۷۷	۱/۲۹۷	۰/۱۵۵	۰/۷۹۳
همصدک	۲۰۰	۵۲/٤۵	۱٤/۸۷	-۲/۲۷۵	۰/۱۷۲	٦/٤٤۱	۰/۳٤۲	٤/٤۸۰
	۵۰۰	۵۳/۵۵۹	۱۱/۳٦۹	۰/۹٦۸	۰/۱۰۹	۱۰/۱۲۵	۰/۲۱۸	٦/۸۱٤
	۱۰۰۰	۵٤/۸٦	۱۰/٤۹	۱/۵۵٦	۰/۰۷۷	۱۰/۹۸۰	۰/۱۵۵	٦/۲٦۳
قوس دایره‌ای	۲۰۰	۵۸/۳۳	۱٤/۵۳	۰/۵۸۷	۰/۱۷۲	۰/۳۱٦	۰/۳٤۲	۰/۹۵۸
	۵۰۰	۵۷/۰۳۲	۱٤/۵۱٦	۰/٦٦۲	۰/۱۰۹	۰/٦۳۲	۰/۲۱۸	۰/٤۳۸
	۱۰۰۰	۵۷/٤۰	۱٤/۵٤	۰/٦۰۰	۰/۰۷۷	۰/۸۷۳	۰/۱۵۵	۰/۳۹٤

ب) آزمون عمومی سنجش جامع

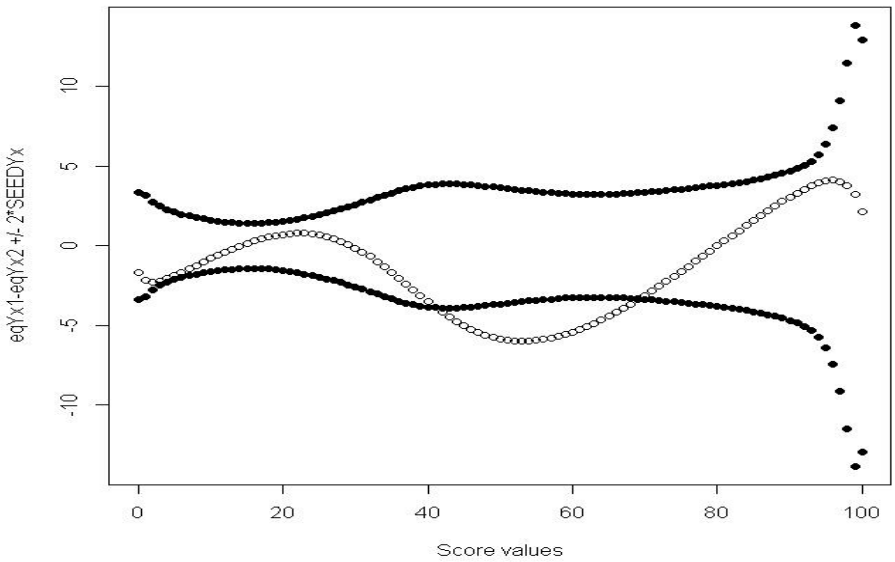
در ادامه این موضوع بررسی شده که در روش به‌خصوص همترازسازی (روش میانگین، روش خطی، روش همصدک، روش قوس دایره‌ای، روش کرنل) در داده‌های آزمون‌های سنجش جامع، طول آزمون و حجم نمونه چه تأثیری بر شاخص‌های آماری مرتبط (میانگین، انحراف استاندارد، کجی، کشیدگی و دقت همترازسازی) داشته است. در جدول (۵) تأثیر طول آزمون و حجم نمونه بر چهار گشتاور اول و خطای همترازسازی در روش همترازسازی کرنل در داده‌های آزمون‌های جامع سنجش، گزارش شده است.

جدول (۵) اثر طول آزمون و حجم نمونه بر چهار گشتاور اول (میانگین و...) نمره‌های مشاهده شده در روش همترازسازی کرنل

طول آزمون	حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندار د کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
٤۰ سؤالی	۲۰۰	۱٦/۰۲	۷/۰۳	۰/۲۷	۰/۱۷	-۰/٦٦	۰/۳٤	۱/۱۰۹
	۵۰۰	۱۵/۹۵	٦/۵۷	۰/۲۷	۰/۱۱	-۰/٤٤	۰/۲۲	۰/۸۵۹
	۱۰۰۰	۱٦/٦۳	۷/۰۱	۰/۱۷	۰/۰۸	-۰/۵۸	۰/۱۵	۰/۵۷۹

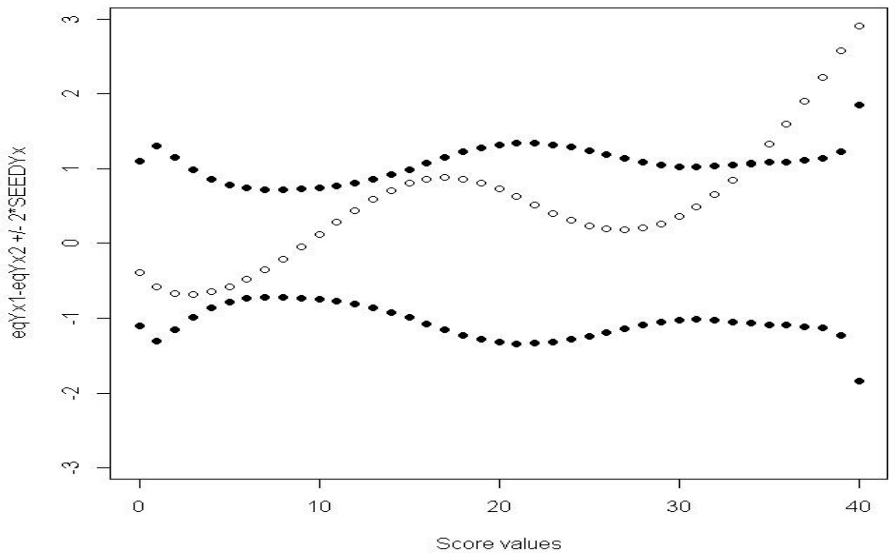
طول آزمون	حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندارد کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
۱۰۰ سؤالی	۲۰۰	۳۷/۹٤	۱۷/۲۸	۰/۵۱٤	۰/۱۷۲	-۰/۵۰۳	۰/۳٤۲	۳/٦۷۰
	۵۰۰	۳۵/۷٦	۱٦/٦۰	۰/۵۸۱	۰/۱۰۹	-۰/۱٤۱	۰/۲۱۸	۲/۲۲۱
	۱۰۰۰	۳۵/۸۵	۱۵/۷۸	۰/٤۱۰	۰/۰۷۷	-۰/۳۵٦	۰/۱۵۵	۲/۰۵۳

براساس اطلاعات جدول (۵)، در روش همترازسازی کرنل به شیوهٔ هموارسازی رشته‌ای (CS) تأثیر حجم نمونه بر شاخص میانگین در نمونه‌های ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری در دو آزمون درس‌های عمومی با طول ٤۰ و ۱۰۰ سؤال، تفاوت معنادار ندارد. به همین ترتیب تأثیر حجم نمونه بر سایر گشتاورها به‌خوبی از روی جدول قابل مشاهده است. همچنین با توجه به مندرجات جدول بالا و توجه به خطای همترازسازی هم در آزمون ٤۰ سؤالی و هم در آزمون ۱۰۰ سؤالی گویای آن است که در این روش همترازسازی در نمونه ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است. در شکل (۵) تأثیر حجم نمونه ۲۰۰ نفری و ۱۰۰۰ نفری در روش کرنل به شیوه هموارسازی رشته‌ای (CS) برای آزمون ۱۰۰ سؤالی سنجش به‌خوبی قابل مشاهده است.



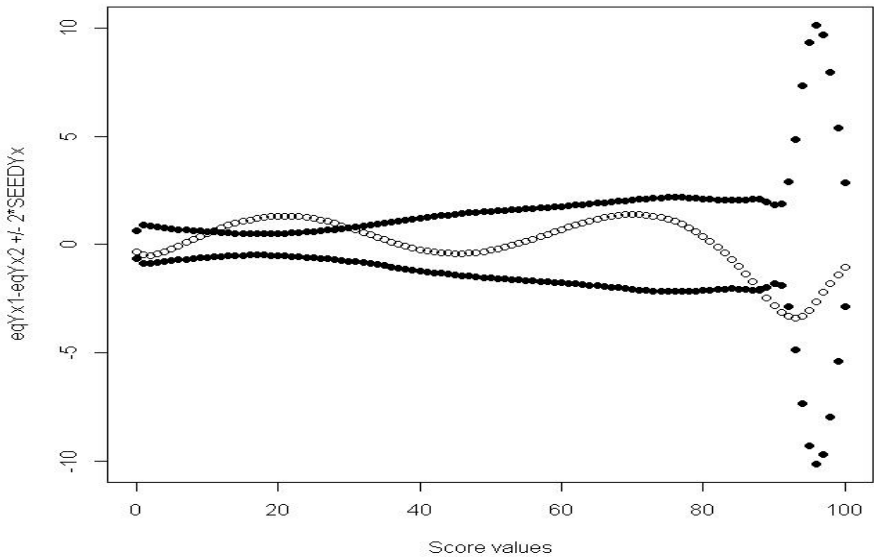
شکل (۵) تأثیر حجم نمونه بر روش همترازی کرنل بین دو نمونه ۱۰۰۰ نفری و ۲۰۰ نفری در آزمون ۱۰۰ سؤالی سنجش

با در نظر گرفتن این نکته که شکل فوق بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۲۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر به‌دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، با توجه به شکل بالا به‌خوبی می‌توان مشاهده کرد که ۱) اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۸۰–۳۰، نقاط زیر این خط قرار می‌گیرد و این مطلب بدان معناست که در این محدوده مقادیر همترازسازی نمونه ۲۰۰ تایی از مقادیر نمونه ۱۰۰۰ نفری بیشتر است؛ ۲) همچنین می‌توان مشاهده کرد که به‌جز دامنه ۸۰–۳۰ همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند؛ ۳) نکته قابل توجه آنکه بازهٔ اطمینان تقریباً در تمام دامنه نمره‌های بسته و فقط در محدوده انتهایی بالا (۱۰۰–۹۰) باز است. در شکل (٦) تأثیر حجم نمونه ۲۰۰ نفری و ۱۰۰۰ نفری در روش کرنل برای آزمون ٤۰ سؤالی عمومی سنجش به‌خوبی قابل مشاهده است.



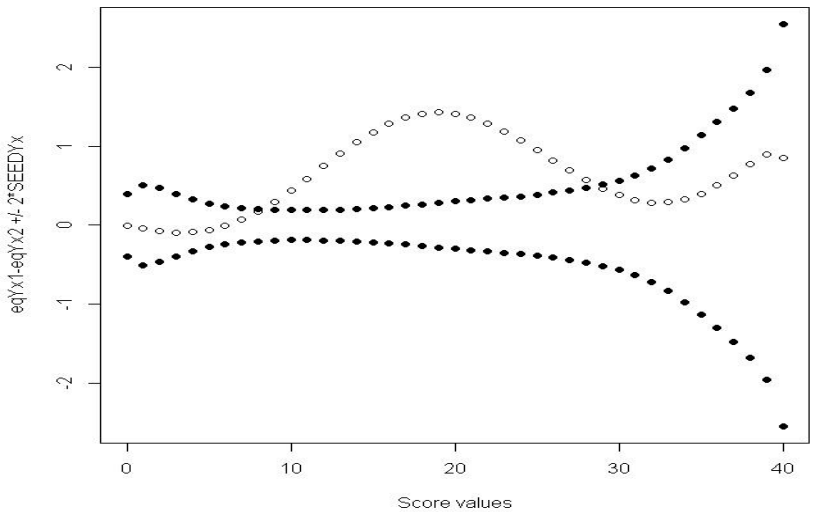
شکل (٦) تأثیر حجم نمونه بر روش همترازی کرنل بین دو نمونه ۱۰۰۰ نفری و ۲۰۰ نفری در آزمون ٤٠ سؤالی سنجش

با عنایت به این نکته که شکل فوق بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۲۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر در آزمون ٤٠ سؤالی به دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، با توجه به شکل بالا به‌خوبی می‌توان مشاهده کرد که ۱) اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم کرد حدوداً در بازهٔ نمرهٔ ۱۰-۰، نقاط زیر این خط قرار می‌گیرد و این مطلب بدان معناست که به‌جز این محدوده مقادیر عددی نمره‌های همتراز شده نمونه ۲۰۰ تایی از مقادیر نمونه ۱۰۰۰ نفری کمتر است؛ ۲) همچنین می‌توان مشاهده کرد که به‌جز دامنه ٤٠-۳۰ همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند؛ ۳) نکته قابل توجه آنکه بازهٔ اطمینان تقریباً در تمام دامنه نمره‌های بسته است. در شکل (۷) تأثیر حجم نمونه ۵۰۰ نفری و ۱۰۰۰ نفری در روش کرنل برای آزمون ۱۰۰ سؤالی عمومی سنجش به‌خوبی قابل مشاهده است.



شکل (۷) تأثیر حجم نمونه بر روش همترازی کرنل بین دو نمونه ۱۰۰۰ نفری و ۵۰۰ نفری در آزمون ۱۰۰ سؤالی سنجش

با در نظر گرفتن این نکته که شکل فوق بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۵۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر در آزمون ۱۰۰ سؤالی سنجش به دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، با توجه به شکل بالا به‌خوبی می‌توان مشاهده کرد که ۱) اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۱۰۰–۸۰، نقاط زیر این خط قرار می‌گیرد، به این معنا که در این محدوده مقادیر عددی نمره‌های همتراز شده نمونه ۵۰۰ تایی از مقادیر نمونه ۱۰۰۰ نفری بیشتر است؛ ۲) همچنین می‌توان مشاهده کرد که به‌جز دامنه ۱۰۰–۹۰ همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند؛ ۳) نکته قابل توجه آنکه بازهٔ اطمینان به‌جز بخش خیلی کوچکی از بالای پیوستار نمره‌ها، تقریباً در تمام دامنه نمره‌ها خیلی بسته است. در شکل (۸) تأثیر حجم نمونه ۵۰۰ نفری و ۱۰۰۰ نفری در روش کرنل برای آزمون ۴۰ سؤالی عمومی سنجش به‌خوبی قابل مشاهده است.



شکل (۸) تأثیر حجم نمونه بر روش همترازی کرنل بین دو نمونه ۱۰۰۰ نفری و ۵۰۰ نفری در آزمون ٤٠ سؤالی سنجش

با ملاحظه این نکته که شکل فوق بر اساس تفاضل مقادیر همترازسازی حجم نمونه ۵۰۰ از مقادیر همترازسازی حجم ۱۰۰۰ نفر در آزمون ٤٠ سؤالی سنجش به‌دست آمده و بر آن اساس فاصله اطمینان ترسیم شده است، با توجه به شکل بالا به‌خوبی می‌توان مشاهده کرد که ۱) اگر از نقطه صفر خطی به‌موازات محور طول‌ها ترسیم کنیم حدوداً در بازهٔ نمرهٔ ۱۰-۰، نقاط روی این خط قرار می‌گیرد، به این معنا که در این محدوده مقادیر عددی نمره‌های همتراز شده نمونه ۵۰۰ تایی با مقادیر نمونه ۱۰۰۰ نفری تفاوت معناداری ندارد؛ ۲) همچنین می‌توان مشاهده کرد که در دامنه ۳۰-۱۰ همه نقاط در خارج از فاصله اطمینان قرار گرفته‌اند؛ ۳) نکته قابل توجه آنکه بازهٔ اطمینان تقریباً در تمام دامنه نمره‌ها بسته و فقط در محدوده کران بالا (٤٠-۳۰) به‌طور معناداری باز است. در جدول (٦) تأثیر طول آزمون و حجم نمونه بر چهار گشتاور اول و خطای همترازسازی در روش‌های مختلف همترازسازی (میانگین، خطی، همصدک و قوس دایره‌ای) در داده‌های آزمون‌های جامع سنجش، گزارش شده است.

جدول (٦) اثر طول آزمون و حجم نمونه بر چهار گشتاور اول (میانگین و...) نمره‌های مشاهده شده در روش‌های مختلف همترازسازی

روش همترازسازی	طول آزمون	حجم نمونه	میانگین	انحراف استاندارد	کجی	خطای استاندارد کجی	کشیدگی	خطای استاندارد کشیدگی	خطای همترازسازی
میانگین	٤٠	۲۰۰	۱٦/۰۳	۷/٤۷	۰/۲۹	۰/۱۷	-۰/٥٦	۰/۳٤	۰/۷۷۰
		٥۰۰	۱٥/۹٥	٦/۹۸	۰/۲۱	۰/۱۱	-۰/٥٦	۰/۲۲	۰/٤۱۳
		۱۰۰۰	۱٦/٦٤	۷/۲۰	۰/۳۱	۰/۰۸	-۰/٥۱	۰/۱٥	۰/۳٥٥
	۱۰۰	۲۰۰	۳۷/۹٤	۱٥/٥٥	۰/۷۰٦	۰/۱۷۲	۰/۲٥٦	۰/۳٤۲	۱/٤۱۷
		٥۰۰	۳٥/۷٥	۱٦/٤۸	۰/۷۷۲	۰/۱۰۹	۰/۱۲٦	۰/۲۱۸	۰/۹٦۹
		۱۰۰۰	۳٥/۸٥	۱٥/۸۱	۰/٥۷۱	۰/۰۷۷	-۰/۲۸٥	۰/۱٥٥	۰/۷٦۸
خطی	٤٠ سؤال	۲۰۰	۱٦/۰۳	۷/۰٤	۰/۲۹	۰/۱۷	-۰/٥٦	۰/۳٤	۰/۹٥۲
		٥۰۰	۱٥/۹٥	٦/٥۸	۰/۲۱	۰/۱۱	-۰/٥٦	۰/۲۲	۰/٦۳۷
		۱۰۰۰	۱٦/٦۳	۷/۰۲	۰/۳۱	۰/۰۸	-۰/٥۱	۰/۱٥	۰/٤۹٦
	۱۰۰ سؤال	۲۰۰	۳۷/۹٤	۱۷/۲۹	۰/۷۰٦	۰/۱۷۲	۰/۲٥٦	۰/۳٤۲	۳/٤٦٥
		٥۰۰	۳٥/۷٥	۱٦/٦۰	۰/۷۷۲	۰/۱۰۹	۰/۱۲٦	۰/۲۱۸	۱/۹۰۷
		۱۰۰۰	۳٥/۸٥	۱٥/۷۸	۰/٥۷۱	۰/۰۷۷	-۰/۲۸٥	۰/۱٥٥	۱/۲۹٥
همصدک	٤٠	۲۰۰	۱٦/۰٤	۷/۰۳	۰/۲۹	۰/۱۷	-۰/٦۹	۰/۳٤	۱/۱۸٥
		٥۰۰	۱٥/۹٥	٦/٥۸	۰/۲۷	۰/۱۱	-۰/٤٥	۰/۲۲	۰/۸٤٦
		۱۰۰۰	۱٦/٦٤	۷/۰۱	۰/۱۷	۰/۰۸	-۰/٥۸	۰/۱٥	۰/٥۹٤
	۱۰۰	۲۰۰	۳۷/۹۳	۱۷/۲۹	۰/٥۱۹	۰/۱۷۲	-۰/٤۸۹	۰/۳٤۲	۲/۸۷٦
		٥۰۰	۳٥/۷٥	۱٦/٥۹	۰/٥۸۸	۰/۱۰۹	-۰/۱۳٦	۰/۲۱۸	۲/٥۸۷
		۱۰۰۰	۳٥/۸٦	۱٥/۷۹	۰/٤۱۰	۰/۰۷۷	-۰/۳٥۷	۰/۱٥٥	۱/٦۱۳
Circle Arc	٤٠	۲۰۰	۱٦/۰٤	۷/٤٦	۰/۳۰	۰/۱۷	-۰/٥٥	۰/۳٤	۰/٥۰۷
		٥۰۰	۱٦/۰۲	٦/۹۲	۰/۲٦	۰/۱۱	-۰/٥۲	۰/۲۲	۰/۳۷۲
		۱۰۰۰	۱٦/٦۰	۷/۲۲	۰/۲۹	۰/۰۸	-۰/٥۳	۰/۱٥	۰/۲۲٦
	۱۰۰	۲۰۰	۳۷/۳۸	۱٦/٤۲	۰/٥۳۱	۰/۱۷۲	-۰/۰۸٥	۰/۳٤۲	۱/۲۹٦
		٥۰۰	۳٥/٤۰	۱٦/۹٥	۰/٦۷٦	۰/۱۰۹	-۰/۰٤۹	۰/۲۱۸	۰/۷۳۸
		۱۰۰۰	۳٥/٤٥	۱٦/٤۹	۰/٤۷۰	۰/۰۷۷	-۰/٤۳۳	۰/۱٥٥	۰/٥٤۳

بحث و نتیجه‌گیری

در آزمون تولیمو مقایسه دو نمونه ۲۰۰ نفر و ۱۰۰۰ نفر در روش کرنل به شیوهٔ CE نشان داد که خطای تفاوت همترازسازی روش کرنل به شیوهٔ هموارسازی رشته‌ای (CS) در دو نمونه ۲۰۰ و ۱۰۰۰ نفری عدد بزرگی است و تفاوت معنادار است؛ درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران بالا و پایینی توزیع نمره‌ها است. بنابراین تأثیر حجم نمونه در وسط پیوستار (نمره‌های ۸۰-

۴۰) کمتر و در انتهای پیوستار (نمره‌های ۴۰-۰ و ۱۲۰-۸۰) یا بازهٔ اطمینان تأثیر حجم نمونه بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران بالا و پایینی توزیع نمره‌ها است. درنهایت مشاهده شد که تفاوت مقادیر همترازسازی بین دو نمونه ۵۰۰ و ۱۰۰۰ نفری در مقایسه با دو نمونه ۲۰۰ و ۱۰۰۰ نفری خیلی ناچیزتر است. مقایسه دو نمونه ۲۰۰ نفر و ۱۰۰۰ نفر در روش کرنل به شیوهٔ PSE نشان داد که تأثیر حجم نمونه در وسط پیوستار نمره‌ها (۱۰۰-۲۰) کمتر و در انتهای پیوستار یا بازهٔ اطمینان تأثیر حجم نمونه بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران بالا و پایینی توزیع نمره‌ها است. مقایسه با روش‌های همترازسازی کرنل متفاوت نشان می‌دهد که در روش همترازسازی کرنل با شیوهٔPSE Smoothing تفاوت دو نمونه ۲۰۰ و ۱۰۰۰ نفری کمتر است و مقادیر تفاوت خطای همترازسازی بیشتر در دامنه اطمینان قرار گرفته‌اند.

مقایسه دو نمونه ۵۰۰ نفر و ۱۰۰۰ نفر در روش کرنل به شیوهٔ PSE نشان داد که تأثیر حجم نمونه بر دقت همترازسازی در این شیوهٔ همترازسازی فقط در دامنه نمره ۲۰-۰ که دامنه خیلی کوچکی است بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران پایینی توزیع نمره‌ها است. درنهایت می‌توان به‌خوبی مشاهده کرد که تفاوت مقادیر عددی نمره‌های همتراز شده بین دو نمونه ۵۰۰ و ۱۰۰۰ نفری در مقایسه با دو نمونه ۲۰۰ و ۱۰۰۰ نفری خیلی ناچیزتر است. نتیجه نهایی اینکه مقایسه این چهار نمودار با روش‌های همترازسازی کرنل متفاوت نشان می‌دهد که در روش همترازسازی کرنل با شیوه PSE Smoothing تفاوت دو نمونه ۵۰۰ و ۱۰۰۰ نفری کمتر است و مقادیر تفاوت خطای همترازسازی کاملاً در دامنه اطمینان قرار گرفته‌اند. به‌طور کلی، در روش همترازسازی کرنل در نمونهٔ ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است. به‌طور کلی هرچه اندازه نمونه آزمون‌شوندگانی که نمراتشان در تحلیل همترازسازی وارد می‌شود بیشتر باشد، خطای استاندارد همترازسازی کوچک‌تر خواهد بود. به‌طور دقیق‌تر نام این خطا "خطای استاندارد همترازسازی مشروط[1]" است، چون خطای استاندارد همترازسازی برای نمره‌های خام مختلف متفاوت است. در میانه توزیع نمره‌ها که بیشتر نمره‌های آزمون‌شوندگان در آنجا قرار

[1]. conditional standard error of equating

دارند، خطای استاندارد همترازسازی احتمالاً کوچک است. در دو کران بالایی و پایینی توزیع نمره‌ها که در آنجا داده‌ها کم تعداد هستند، خطای استاندارد همترازسازی احتمالاً بیشتر است. بر اساس نتایج، در روش‌های مختلف همترازسازی (میانگین، خطی، همصدک و قوس دایره‌ای) تأثیر حجم نمونه بر شاخص میانگین در نمونه‌های ۲۰۰، ۵۰۰ و ۱۰۰۰ دارای تفاوت معنادار نیست. بر این اساس، در این روش همترازسازی در نمونه ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است.

در آزمون ۱۰۰ سؤالی آزمایشی جامع سنجش، مقایسه دو نمونه ۲۰۰ نفر و ۱۰۰۰ نفر در روش کرنل نشان داد که تأثیر حجم نمونه در این شیوه همترازسازی فقط در دامنه نمره ۱۰۰–۹۰ که دامنه خیلی کوچکی است بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران بالای توزیع نمره‌ها است. در آزمون ۱۰۰ سؤالی سنجش، مقایسه دو نمونه ۵۰۰ نفر و ۱۰۰۰ نفر در روش کرنل نشان داد که به‌جز دامنه ۱۰۰–۹۰ همه نقاط در داخل فاصله اطمینان قرار گرفته‌اند و بازهٔ اطمینان به‌جز بخش خیلی کوچکی از بالای پیوستار نمره‌ها، تقریباً در تمام دامنه نمره‌ها خیلی بسته است. بنابراین تأثیر حجم نمونه در این شیوهٔ همترازسازی فقط در دامنه نمره ۱۰۰–۸۰ که دامنه خیلی کوچکی است، بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های کران بالای توزیع نمره‌ها است.

در آزمون ٤۰ سؤالی آزمایشی جامع سنجش، مقایسه دو نمونه ۲۰۰ نفر و ۱۰۰۰ نفر در روش کرنل نشان داد که بازهٔ اطمینان تقریباً در تمام دامنه نمره‌ها بسته است. بنابراین تأثیر حجم نمونه در این شیوهٔ همترازسازی فقط در دامنه نمره ٤۰–۳۰ که دامنه خیلی کوچکی است، بیشتر و بارزتر است. درواقع، در این روش همترازسازی حجم نمونه بیشترین تأثیر را برای نمره‌های کران بالای توزیع نمره‌ها داشته است. در آزمون ٤۰ سؤالی سنجش، مقایسه دو نمونه ۵۰۰ نفر و ۱۰۰۰ نفر در روش کرنل نشان داد که در دامنه ۳۰–۱۰ همه نقاط در خارج از فاصله اطمینان قرار گرفته‌اند و بازهٔ اطمینان تقریباً در تمام دامنه نمره‌ها بسته و فقط در محدوده کران بالا (٤۰–۳۰) به‌طور معناداری باز است. بنابراین تأثیر حجم نمونه در این شیوهٔ همترازسازی فقط در دامنه نمره ۳۰–۱۰ که دامنه بزرگی است بیشتر و بارزتر است. درواقع، در این روش همترازسازی نقش حجم نمونه بیشتر برای نمره‌های میانی توزیع نمره‌ها بوده است.

بر اساس نتایج، در روش‌های مختلف همترازسازی (میانگین، خطی، همصدک و قوس دایره‌ای) تأثیر حجم نمونه بر شاخص میانگین در نمونه‌های ۲۰۰، ۵۰۰ و ۱۰۰۰ نفری در دو آزمون درس‌های عمومی با طول ٤٠ و ۱۰۰ سؤال، تفاوت معنادار ندارد. همچنین با توجه به مندرجات جدول (٦) و توجه به خطای همترازسازی هم در آزمون ٤٠ سؤالی و هم در آزمون ۱۰۰ سؤالی گویای آن است که در همه روش‌های همترازسازی در نمونه ۲۰۰ نفری میزان خطای همترازسازی (SEE) بیشتر و در نمونه ۱۰۰۰ نفری کمتر گزارش شده است.

متغیرهای مشترک تأثیرگذار بر شرایط همترازسازی که در دیگر پژوهش‌ها و این پژوهش به‌طور اخص، بررسی شده‌اند عبارتند از: مجموعه داده‌ها، حجم نمونه و طول آزمون. در مطالعات همترازسازی انتخاب حجم نمونه یکی از مسائل مهم است. مرور پیشینه مرتبط با این موضوع آشکار کرد که پژوهشگران انتخاب‌های بسیار متفاوتی داشته‌اند. به‌طور کلی زمانی که حجم نمونه کوچک باشد، روش کرنل بزرگ‌ترین مزیت‌ها را نسبت به سایر روش‌های همترازسازی کلاسیک دارد. گرانت و همکاران (۲۰۰٦) طرح NEAT را برای مطالعه همترازسازی نمونه کوچک در روش کرنل به کار بردند تا تأثیر آن را بر خطای استاندارد همترازسازی بررسی کنند. عملکرد روش کرنل در نمونه‌هایی با حجم ۱۰۰۰، ۵۰۰، ۲۵۰، ۱۲۵ و ۷۵ مورد مقایسه قرار گرفت. زمانی که حجم نمونه کوچک می‌شد، ناهمواری بیشتری در توزیع نمره‌ها مشاهده شد. همچنان‌که انتظار می‌رفت نتایج نشان داد که با کاهش حجم نمونه دقت همترازسازی کاهش می‌یافت. افزایش یافتن حجم یک نمونه کوچک‌تر نتایج همترازسازی را بیش از افزایش حجم یک نمونه بزرگ‌تر بهبود می‌بخشید، اما حجم نمونه ۷۵ و ۱۲۵ خیلی کوچک بودند و در عمل به‌ندرت مورد استفاده قرار می‌گرفتند. واقعیت آن است که حجم نمونه بزرگ معمولاً اطلاعات بیشتری را دربارهٔ برآورد توزیع نمرات فراهم می‌کند؛ چراکه میزان خطای تصادفی در همترازسازی را کاهش می‌دهد و متعاقب آن نتایج همترازسازی بهبود می‌یابد (کولن و برنان، ۲۰۰٤).

افزایش حجم نمونه می‌تواند دقت نتایج همترازسازی را افزایش دهد، ولی کافی نبودن حجم نمونه می‌تواند کیفیت برآورد پارامترها و سودمندی برآوردهای سؤال‌های لنگر را کاهش دهد و بر روی همترازسازی تأثیر منفی بگذارد (پترسن و کوک، ۱۹۸۹). پژوهشگران علاقه‌مند هستند که در نمونه‌هایی با کمترین حجم نیز همترازسازی دقیقی را انجام دهند. کولن و برنان (۲۰۰٤) پیشنهاد می‌کنند هنگام استفاده از شیوه های همترازسازی خطی کمترین حجم نمونه در مورد هر فرم ٤۰۰ نفر و هنگام

استفاده از همترازسازی معادل درصدی هر فرم ۱۵۰۰ نفر باشد. همچنین آنها نشان دادند که برای به‌دست آوردن دقتی در همین سطح، طرح گروه‌های معادل نسبت به طرح گروه منفرد و طرح گروه‌های غیر همتا با آزمون لنگر (NEAT) به حجم نمونه بزرگ‌تری نیاز دارد (لیوینگستون، دورانز و رایت[1]، ۱۹۹۰؛ کولن و برنان، ۲۰۰٤). به طور کلی طول آزمون و حجم نمونه رابطه مثبتی با عملکرد همترازسازی دارند، در حالی که تفاوت‌های موجود در توانایی گروه و متغیر بودن دشواری سؤال‌های لنگر به طور منفی بر نتایج همترازسازی اثر می‌گذارند (سینارای و هالند، ۲۰۰۷). هماهنگ با نتایج پژوهش‌های پیشین (لی[2]، ۲۰۰۷) در این پژوهش نیز مشخص شد که همچنان که حجم نمونه افزایش یافته است برازش مربوط به هموارسازی کرنل نیز بهبود یافته است و بهبود هموارسازی کرنل با افزایش طول آزمون همراه بوده است. پیشنهاد می شود علاوه بر بررسی تأثیر حجم نمونه و روش‌های مختلف همترازسازی، تأثیر عوامل دیگر همچون «تفاوت توانایی گروه‌ها» و «تفاوت تعداد ابعاد آزمون‌های مورد همترازسازی» مطالعه و بررسی شود.

1. Livingston, Dorans & Wright
2. Lee

منابع

سرمد، زهره؛ بازرگان، عباس و حجازی، الهه (۱۳۸٤). *روش‌های تحقیق در علوم رفتاری*. تهران: نشر آگاه.

لرد، فردریک (۱۹۸۰). *کاربردهای نظریه سؤال- پاسخ*؛ ترجمه علی دلاور و جلیل یونسی. تهران: انتشارات رشد.

Brennan, R. L (2006). (Ed.). *Educational measurement* (4th ed.). Westport, CT: Praeger.

Godfrey, K. E. (2007). *A comparison of Kernel equating and IRT true score equating methods.* Unpublished doctoral dissertation, University of North Carolina, Greensboro. Retrieved from ProQuest. (AAT 3273329).

Grant, M. C.; Zhang, L.; Damiano, M. & Lonstein, L. (2006). An evaluation of the kernel equating method: Small sample equating in non-equivalent groups. *Paper presented at the national conference of AERA/NCME, 2006.*

Hanson, B. A. & Béguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common item nonequivalent groups equating design. *Applied Psychological Measurement*, 26 (1), 3-24.

Kolen, M. J. & Brennan, R. L. (2004). *Test Equating Methods and Practices*. New York: Springer-Verlag.

Lee, Y., & von Davier, A. A. (2010). Equating through alternative kernels. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 159-173). New York: Springer.

Lee, Y. H. (2007). Contributions to the statistical analysis of item response time in educational testing. *Unpublished doctoral dissertation*, Columbia University, New York.

Livingston, S. A., Dorans, N. J. & Wright, N. K. (1990). What combination of sampling and equating methods work best? *Applied Measurement in Education, 3*, 73-95.

Peterson, N. S and Cook L.L (1989). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied psychological measurement*. 11, 225- 244.

R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Sinharay, S. & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44* (3), 249-275.

Von Davier, A. A., Holland, P. W., Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York: Springer-Verlag.

استناد به این مقاله:
یونسی، جلیل (۱۳۹۵). تأثیر حجم نمونه و طول آزمون بر نمرات همتراز شده و خطای همترازسازی: مورد مطالعه آزمون‌های ملی ایران، *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ٦ (۱۵)، ۱۱ – ۳۷.