Vol 16, Issue 59, 2025, 7-35 ISSN: 2252-004x
eISSN: 2470-6240
Article Type: Reserch Article
Received in revised form October 17, 2023
Accepted May 07, 2024
Received October 02, 2022
Published Online March 25, 2025

Educational Measurement
https://jem.atu.ac.ir/

ATU PRESS

Identification and Selection of the Optimal Approach in the Equating of Small Samples

Azam Ghafouri — Ph.D. Student in Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: b.ghafuori@gmail.com

Noorali Farrokhi* — *Corresponding Author*, Associate Professor, Department of Educational Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: farrokhinoorali@gmail.com

Jalil Younesi — Associate Professor, Department of Educational Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: jalilyounesi@gmail.com

Abstract

The present study aimed to explore and evaluate Equating methods based on the classical test theory, with a focus on quantifying errors, as well as identifying factors impacting the accuracy of these methods in small sample sizes. The primary aim of this research was to compare different approaches for Equating and evaluate the accuracy of each method, culminating in the determination of the optimal Equating method. Consequently, two types of studies were planned, which employed two distinct research approaches. The first study drew upon real data derived from the California Critical Thinking Questionnaire (Forms A and B) to compare Equating methods. The statistical population for this study comprised all employees working in Qazvin Municipality during the year 2019. For the simulation study, two tests were created using simulated data in various scenarios, following the average discrimination parameter obtained from the actual implementation of the questionnaire on a sample of 200 individuals. Additionally, a random group sampling method was employed to ensure a fair and independent selection of participants, resulting in two samples of 200 people each. The findings indicated that the mean Equating method yielded the lowest error in a sample size of 50 people, while the percentile Equating method demonstrated the least error rate in terms of the BE index for the same sample size of 50 participants. The analysis of various indicators related to Equating error revealed that as the sample size grew, the amount of error decreased. Moreover, an increase in the difficulty difference between the two tests tended to elevate the error rate in all four Equating methods.

Keywords: Equating; Mean Equating; Linear Equating; Equipercentile Equating; Circle – Arc Equating

Cite this Article: Ghafouri, A., Farrokhi, N., & Younesi, J. (2025). Identification and Selection of the Optimal Approach in the Equating of Small Samples. *Educational Measurement*, *16*(59), 7-35. https://doi.org/10.22054/jem.2024.64044.3405

Publisher: Allameh Tabataba'i University Press

Introduction

One crucial aspect to consider in the use of measurement tools and decision-making processes is fairness and justice (Aşiret et al., 2016). According to Kolen et al. (2014), test equating is defined as a statistical process aimed at adjusting scores derived from different forms of the same test, to ensure that they can be interchanged and used interchangeably. The effectiveness of test equating hinges on the fulfillment of certain assumptions related to data collection and equating characteristics. Hambleton et al. (1985) classified these conditions as symmetry, same specification, equity, and group linking (Kolen et al., 2014).

Within the realm of classical test theory-based Equating methods, there are four primary approaches utilized: mean equating, linear equating, equipercentile equating, and Circle – Arc equating (Mechael, 2008). Comparable Equating methods are often evaluated and compared based on various parameters and factors, such as the sample size, ability differences between recipient groups for two test forms, test length, a guessability factor, the difference in difficulty between forms, data collection patterns, and the characteristics of anchor tests.

Researchers have found that the sample size has a direct impact on the standard error of equating (Kim et al., 2008). This implies that the larger the sample size, the smaller the standard error and vice versa. Assessing the random error, verifying the compatibility of equating results with previous outcomes, and examining the satisfaction of equating assumptions (like symmetry, item equivalence, and others) can function as bases to define an equating evaluation criterion (Kolen et al., 2014).

Research Question(s)

Which equating method in classical test theory is considered the most optimal?

Which of the equating methods is more likely to yield accurate results when equating test forms?

Literature Review

There is a significant discrepancy between the available empirical researches on equating in small sample sizes. According to Iriyadi et al.'s (2018) study titled "Equating Methods for Small Samples," the weighted nominal mean method is deemed more suitable than the linear equating method. Their study revealed that the root mean square error

in the weighted nominal mean method was found to be smaller compared to the linear equating method. Livingston and Kim (2010) performed a comparative analysis of Equating methods, including equipercentile, linear, mean, and Circle – Arc equating methodologies (both symmetrical and simple). In summary, the findings demonstrated the efficacy of the Circle – Arc equating method in terms of its high equating accuracy, outperforming other methods within various sample size groups.

In a study similar to Skaggs (2005), the accuracy of Equating methods based on random groups in very small samples was compared. The findings suggested that, in scores lower than the mean, the mean equating method demonstrated the highest accuracy. In scores close to the mean, all Equating methods operated similarly. Lastly, for scores higher than the mean, the equipercentile equating method showcased the greatest accuracy.

In the study conducted by Parshall et al. (1995), the effect of sample size (ranging from 15 to 100 individuals) on the stability and bias of linear equating, when applied to two parallel test forms based on equivalent group patterns with shared items (anchor) was investigated. Their findings indicated that, as the sample size diminishes, a minimal bias is observed in the equating results. Nevertheless, when it comes to the standard error of equating, reducing the sample size correspondingly leads to an amplification of this error.

Methodology

This research employed two different approaches, resulting in two distinct categories of data analysis. The data was divided into two sections: real data and data obtained through simulation, following a pattern of random group sampling. The mentioned approach was applied in both categories of data. The first part of this research utilized real data derived from administering Forms A and B of the California Critical Thinking Skills Test to assess the accuracy of Equating methods based on classical test theory. Additionally, the results obtained from each method were compared to ascertain which technique proved to be the most accurate.

This study's research methodology adopts a descriptive approach. The target population in this research includes all employees across various occupational categories who were employed in Qazvin Municipality in the year 2020. A total of 2048 individuals were selected

as the sample size. Furthermore, to implement the random group equating pattern, two independent and randomly selected samples consisting of 200 individuals each were chosen from the population mentioned above, employing the Cochran sample size determination formula and a margin of error of 0.01.

For the second part of the study, two simulated tests were developed based on various scenarios (considering sample size and difficulty coefficients) while taking into account the mean parameter derived from the real administration of the test for 200 participants. Subsequently, to further analyze the accuracy of Equating methods in small sample sizes, two extra sample groups consisting of 50 and 100 individuals were chosen and scrutinized as well.

Conclusion

This research scrutinized the performance of four Equating methods (linear equating, mean equating, equipercentile equating, and Circle – Arc equating) within a random group pattern, taking into account two independent variables: sample size (50, 100, and 200 individuals) and difficulty coefficients (0.1, 0.4, and 0.7). The results suggest a correlation between expanding the sample size and diminishing the error in equating methods. Consequently, the standard error of equating was found to decrease as the sample size increased.

The most noticeable effect of sample size on bias error reduction was observed in the equipercentile and Circle – Arc equating methods. Moreover, increasing the difference in difficulty resulted in an amplification of the bias error. Among the mentioned Equating methods, the least impact was registered in the equipercentile method and the most significant impact was evident in the Circle – Arc equating method. With regard to the standard error of equating (SEE), all four equating methods demonstrated a decline in error with an expansion of the sample size, wherein the equipercentile method yielded the lowest error. Furthermore, an increase in the difficulty difference caused a corresponding increase in the standard error across all four equating methods.

اندازه‌گیری تربیتی
https://jem.atu.ac.ir/

شناسایی و انتخاب رویکرد بهینه در همترازسازی نمونه‌های با حجم کم

اعظم غفوری دانشجوی دکتری رشته سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: b.ghafuori@gmail.com

نورعلی فرخی* *نویسنده مسئول*، دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: farrokhinoorali@gmail.com

جلیل یونسی دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: jalilyounesi@gmail.com

چکیده

همترازسازی فرم‌های جدید یک آزمون با فرم‌های قبلی یکی از موضوعات مهم و پرکاربرد در امر سنجش و اندازه‌گیری است. پژوهش حاضر با هدف مطالعه و ارزیابی روش‌های همترازسازی بر مبنای نظریه کلاسیک آزمون برحسب خطاها و همچنین بررسی عوامل مؤثر بر افزایش دقت روش‌های همترازسازی در نمونه‌های با حجم کم انجام شده است. نظر به اینکه در این پژوهش محقق در پی مقایسه رویکردهای مختلف در همترازسازی و دقت اندازه‌گیری و درنهایت تعیین روش بهینه همترازسازی بوده، لذا دو نوع مطالعه بر اساس دو روش پژوهش طرح‌ریزی شد. روش پژوهش در این مطالعه به‌طور عام جزء پژوهش‌های توصیفی است. با توجه به دو نوع داده مورداستفاده در پژوهش حاضر، دو نوع جامعه نیز مطرح است. در مطالعه اول به‌منظور بررسی مقایسه‌ای روش‌های همترازسازی و با توجه به اجرای پرسش‌نامه تفکر انتقادی کالیفرنیا (فرم الف و ب)، از داده‌های واقعی استفاده شد. جامعه آماری در این مطالعه شامل کلیه کارکنان شاغل در شهرداری قزوین در سال ۱۳۹۹ است. همچنین به جهت استفاده از الگوی گروه‌های تصادفی، دو نمونه به‌صورت تصادفی و مستقل با حجم ۲۰۰ نفر از جامعه مذکور انتخاب شد. در خصوص مطالعه شبیه‌سازی نیز در حالات مختلف (حجم نمونه و ضرایب دشواری) و با توجه به میانگین پارامتر تمیز به‌دست‌آمده از اجرای واقعی برای ۲۰۰ نفر آزمودنی، دو آزمون با داده‌های شبیه‌سازی ایجاد شد. با توجه به شاخص‌های محاسبه‌شده در خصوص میزان خطای همترازسازی، نتایج نشانگر کاهش میزان خطا با افزایش حجم نمونه بود. همچنین با افزایش مقدار ضرایب دشواری میزان خطا نیز در هر چهار روش همترازسازی افزایش یافت.

کلیدواژه‌ها: همترازسازی، همترازسازی میانگین، همترازسازی خطی، همترازسازی همصدک، همترازسازی قوس دایره‌ای

استناد به این مقاله: غفوری، اعظم، فرخی، نورعلی، و یونسی، جلیل. (۱۴۰۴). شناسایی و انتخاب رویکرد بهینه در همترازسازی نمونه‌های با حجم کم مقاله. *فصلنامه اندازه‌گیری تربیتی*، ۱۶(۵۹)، ۷-۳۵.
https://doi.org/10.22054/jem.2024.64044.3405

ناشر: دانشگاه علامه طباطبائی

دوره ۱۶، شماره ۵۹، ۱۴۰۴، ۷-۳۵
eISSN: 2470-6240 ISSN: 2252-004x
نوع مقاله: مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۰۷/۱۰؛ تاریخ بازنگری: ۱۴۰۲/۰۷/۲۵
تاریخ پذیرش: ۱۴۰۳/۰۲/۱۸؛ تاریخ انتشار: ۱۴۰۴/۰۱/۰۵

مقدمه

در حوزه‌های آموزشی، تربیتی، و به‌ویژه روان‌شناسی روش‌های اندازه‌گیری یا آزمون‌ها به‌صورت گسترده‌ای مورداستفاده قرار می‌گیرند؛ ازجمله موضوعات مهم در خصوص نحوه استفاده از ابزار اندازه‌گیری و اخذ تصمیمات رعایت انصاف و عدالت است (Aşiret et al., 2016). علی‌رغم تلاش آزمون‌سازان و استفاده‌کنندگان از آزمون‌های استاندارد مبنی بر ارزیابی دقیق و باثبات و روا از موضوعات مختلف، موضوع امنیت و فاش نشدن سؤالات آزمون‌ها در هر بار اجرا بسیار موردتوجه بوده و تهدیدی جدی است. ازاین‌رو به دست آوردن نتایج دقیق و منصفانه با رعایت کامل عدالت در فرایند اجرای آزمون عاملی برای بررسی راه‌کارهای حذف و رفع تهدیدهایی از این نوع شده است (Caglak, 2016). همترازسازی[1] آزمون برای رفع این مشکلات مورداستفاده قرار می‌گیرد و می‌تواند تفسیر نمرات به‌دست‌آمده از فرم‌های آزمون را قابل‌معاوضه[2] کند (VonDavier et al., 2004). همچنین در شرایطی که آزمون‌ها دارای فرم‌های چندگانه هستند و این فرم‌ها ویژگی‌های یکسان اما دشواری متفاوتی دارند همترازسازی آزمون‌ها به دلیل کنترل کردن اثر متغیر مخدوش‌کننده فرم آزمون، تعدیل دشواری فرم‌های مختلف آزمون و قابل‌مقایسه کردن نمرات آن‌ها، یک ضرورت در اندازه‌گیری به شمار می‌آید (Dorans et al., 2010). همترازسازی دقیق پیش‌نیازی برای تفسیر معتبر نمرات فرم‌های چندگانه آزمون است. چنانچه همترازسازی به‌طور دقیق انجام شود، نمرات از یک فرم آزمون با نمرات فرم‌های دیگر آزمون قابل‌معاوضه و مقایسه خواهد بود و عدالت و انصاف در ارزیابی امکان‌پذیر می‌شود (بهمن‌آبادی و همکاران، ۱۴۰۳).

همترازسازی آزمون را به‌عنوان فرایند آماری مورداستفاده جهت تعدیل نمرات حاصل از اجرای فرم‌های متفاوت یک آزمون تعریف‌شده به‌طوری‌که نمرات فرم‌های مختلف به‌صورت قابل‌معاوضه‌ای مورداستفاده قرار می‌گیرند (Kolen et al., 2014).

هدف همترازسازی تولید نمرات قابل‌معاوضه از نسخه‌های مختلف یک آزمون است که معمولاً بر اساس مجموعه استانداری از شرایط اندازه‌گیری گردآوری شده‌اند (Dorans et al., 2023). موفقیت همترازسازی مستلزم رعایت مجموعه‌ای از مفروضات درباره نحوه گردآوری داده‌ها و ویژگی‌های همترازسازی است و عدم توجه به این مفروضات موجب

1. equating
2. interchangeably

کاهش دقت همترازسازی می‌شود. برخی از پژوهشگران شرایط متعددی را برای همترازسازی آزمون‌ها ذکر کرده‌اند. Hambleton و همکاران (1985) این شرایط را تحت عناوین ویژگی تقارن[1]، یکسانی ویژگی‌های سؤال[2]، برابری (بی‌طرفی)[3] و گروه فهرست‌بندی کرده‌اند (Kolen et al., 2014).

معمولاً اولین گام در همترازسازی، انتخاب یک روش مناسب است. روش‌های همترازسازی خانواده‌ای از مدل‌ها و روش‌های آماری هستند که برای تعدیل نمرات به دست آمده از فرم‌های مختلف آزمون و مقایسه‌پذیر کردن آزمون‌هایی با محتوای یکسان مورداستفاده قرار می‌گیرند (Dorans et al., 2000). چنان چه فرم‌های مختلف آزمون دشواری متفاوتی داشته باشند مقایسه نمرات افرادی که این فرم‌ها را می‌گیرند به سبب آسان بودن برخی فرم‌ها و دشواری برخی دیگر، عادلانه نخواهد بود، ازاین‌رو همترازسازی آزمون‌ها رویکردی است که عدالت و انصاف در ارزیابی را تضمین کرده و ضمن تأمین امنیت آزمون‌ها نمرات فرم‌های مختلف آزمون را در مقیاس یکسان قرار می‌دهد (Dorans et al., 2010).

روش‌های همترازسازی بر اساس نظریه کلاسیک آزمون (همترازسازی نمره مشاهده‌شده) و نظریه سؤال – پاسخ (همترازسازی نمره واقعی) طبقه‌بندی می‌شوند. روش‌های همترازسازی کلاسیک آزمون به دو دسته روش‌های غیرخطی (همترازسازی میانگین، خطی، قوس دایره‌ای) و خطی (همترازسازی همصدک) تقسیم می‌شوند (Mechael, 2008). با توجه به این‌که در پژوهش حاضر روش‌های مبتنی بر نظریه کلاسیک آزمون مورداستفاده قرار خواهد گرفت، ازاین‌رو هر یک از این روش‌ها شامل روش همترازسازی میانگین[4]، روش همترازسازی خطی[5]، روش همترازسازی همصدک[6]، روش همترازسازی قوس دایره‌ای[7] به‌اختصار توضیح داده می‌شود. در روش همترازسازی میانگین فرض بر این است که فرم X و فرم Y ازنظر دشواری به‌واسطه یک مقدار ثابت در طول مقیاس نمرات متفاوت هستند (Kolen et al., 2014).

---

1. symmetry
2. same specifications
3. equity
4. mean equating method
5. linear equating
6 .equipercentile equating
7. circle – arc method

روش همترازسازی خطی مبتنی بر این مفروضه است که شکل توزیع نمرات فرم X و فرم Y یکسان، ولی میانگین و انحراف استاندارد آن‌ها متفاوت هستند. همترازسازی خطی زمانی استفاده می‌شود که نمرات استاندارد به‌دست‌آمده از این فرم‌ها هم‌تراز فرض گرفته می‌شد. Donlon (1984) اظهار می‌کند که اگر گروه‌های آزمون‌شوندگانی که فرم‌های مختلفی از یک آزمون را دریافت کردند سطوح توانایی برابری داشته باشند، می‌توان همترازسازی خطی را انجام داد (Aşiret et al., 2016).

همترازسازی همصدک: در همترازسازی همصدک، از یک منحنی برای توصیف تفاوت در دشواری فرم‌ها استفاده می‌شود که باعث می‌شود همترازسازی همصدک رایج‌تر از همترازسازی خطی باشد. این روش زمانی که شکل توزیع نمره فرم‌ها متفاوت هستند پیشنهاد شده است. وقتی دو آزمون توزیع نمرات متفاوتی دارند، ارتباط بین نمرات آن‌ها غیرخطی است. در این موقعیت همترازسازی همصدک ترجیح داده می‌شود (Heh, 2007). در همترازسازی همصدک، فرم X امکان دارد از فرم Y برای نمرات بالا و پایین دشوارتر باشد، درحالی‌که برای نمرات متوسط دشواری کمتری داشته باشد. اگر توزیع نمرات فرم X که به نمرات فرم Y تبدیل شده با توزیع نمرات فرم Y برابر باشد، تابع همترازسازی بین دو فرم تابع همترازسازی همصدک نامیده می‌شود (Kolen et al., 2014).

در روش همترازسازی همصدک به دلیل انتخاب آزمون‌شوندگان از یک یا چند جامعه، به هنگام رسم توزیع نمره‌های خام، برخی بی‌نظمی‌ها می‌توانند به‌عنوان نتیجه‌ای از خطاهای نمونه‌گیری ظاهر شوند (Kolen et al., 2014). با افزایش حجم نمونه خطاهای نمونه‌گیری کاهش می‌یابد. با توجه به پیشنهاد Kolen and Brennan (1995) در خصوص کفایت نمونه ۱۵۰۰ نفری در روش همترازسازی همصدک و با توجه به عدم دسترسی به این حجم نمونه در هر موقعیتی، روش‌های یکنواخت سازی[1] به کار برده می‌شوند (Cui et al., 2009).

Livingston و همکاران (2009b) یک روش جدید برای همترازسازی فرم‌های آزمون در نمونه‌های کوچک پیشنهاد کرده‌اند. مفروضه این روش خطی نبودن رابطه همترازسازی است. همچنین آنان دریافتند که روش قوس دایره‌ای در مقایسه با روش همترازسازی میانگین، به‌ویژه در انتهای توزیع نمرات، سوگیری و مجذور میانگین مربع خطای کمتری دارد (Babcock et al., 2019).

1. smoothing

دو نسخه از روش همترازسازی قوس دایره‌ای توسط Livingston و همکاران (2008) تحت عناوین متقارن[1] و ساده[2] معرفی شده است که تابعی بین فرم‌های آزمون به‌ویژه وقتی حجم نمونه در فرم جدید کمتر از ۳۰ نفر است ایجاد می‌کند. تفاوت بین دو نسخه مربوط به نحوه ترسیم تابع همترازسازی است. برای به دست آوردن تابع همترازسازی در هر دو نسخه احتمالاً به محاسبات ریاضی ساده‌ای نیاز است، اما برآورد تصویر هندسی منحنی در محور X در نسخه ساده فرایند تقریباً پیچیده‌تری دارد درحالی‌که قوس منحنی باید با نقاط در نسخه متقارن برازش داشته باشد. Livingston و همکاران (2011) خاطرنشان کردند هر دو نسخه همترازسازی قوس دایره‌ای نتایج مشابهی به همراه دارند (Caglak, 2016).

مقایسه روش‌های همترازسازی اغلب به پارامترها و عوامل مختلفی بستگی دارد که از این میان می‌توان به حجم نمونه، تفاوت در توانایی گروه‌های دریافت‌کننده دو فرم، طول آزمون، عامل حدس‌پذیری و تفاوت دشواری بین فرم‌ها، الگوی جمع‌آوری داده، و ویژگی‌های آزمون لنگر اشاره کرد. بر این مبنا پژوهش‌های متعددی در خصوص عملکرد روش‌های مختلف همترازسازی در شرایط معین انجام شده است. حجم نمونه در مطالعات همترازسازی یک متغیر مشترک محسوب می‌شود. بر اساس یافته‌های پژوهشگران حجم نمونه بر میزان خطای استاندارد همترازسازی تأثیر مستقیمی دارد (Kim et al., 2008).

همچنین در انجام همترازسازی آزمون لازم است برای جمع‌آوری داده‌ها، یا دو آزمون روی آزمون‌شوندگان مشترک اجرا شود یا اینکه سؤال‌های مشترک در دو آزمون قرار داده شوند؛ به این دلیل، الگوی جمع‌آوری داده‌ها[3] توسعه پیدا کرده است. الگوی جمع‌آوری داده‌ها طرحی برای جمع‌آوری داده‌های موردنیاز همترازسازی است. الگوی جمع‌آوری داده‌ها ممکن است به‌عنوان یک الگوی آزمودنی‌های مشترک[4] یا به‌عنوان الگوی سؤال‌های مشترک[5] طبقه‌بندی شود. الگوهای رایج مورداستفاده برای همترازسازی شامل الگوی تک گروهی[6]، الگوی تک گروهی متوازن (با موازنه)[7]، الگوی گروه‌های تصادفی[8] یا گروه‌های

1. symmetric
2. simplified
3. data collection design
4. common – examinees design
5. common – items design
6. single – group design
7. counterbalanced single – group design
8. random – groups design

معادل[1] (Kolen et al., 2014)، و الگوی گروه نامعادل با آزمون لنگر[2] که به الگوی گروه نامعادل با سؤالات مشترک[3] نیز شناخته می‌شود (Devdass, 2011) است.

در همترازسازی آزمون‌ها، علاوه بر انتخاب الگوی همترازسازی، داشتن یک تعریف عملیاتی از همترازسازی و تعیین ملاکی جهت ارزشیابی نتایج همترازسازی ضروری است؛ بنابراین پس از انجام فرایند همترازسازی نتایج باید مورد ارزیابی قرار بگیرند. چنین ارزیابی نیازمند تعیین یک ملاک برای همترازسازی است. برآورد خطای تصادفی، سازگاری نتایج همترازسازی با نتایج قبلی، بررسی تحقق ویژگی‌های (مفروضه‌های) همترازسازی (نظیر ویژگی تقارن، ویژگی یکسانی خصوصیات آماری، و ...) می‌تواند به‌عنوان پایه‌هایی برای تعریف معیار ارزشیابی همترازسازی مورداستفاده قرار گیرد (Kolen et al., 2014). ازجمله معیارهای رایج جهت بررسی و ارزیابی نتایج همترازسازی، بررسی میزان دقت و خطای ناشی از همترازسازی است. مفهوم خطا در نظریه کلاسیک آزمون به‌عنوان تفاوت بین نمره واقعی و نمره مشاهده تعریف شده است؛ اما در اغلب کاربردهای آماری، خطا به‌عنوان تفاوت بین ارزش واقعی یک پارامتر در جامعه و مقدار برآورد شده آن پارامتر تعریف می‌شود (Jaeger, 1981). خطاهای همترازسازی به دو منبع تقسیم می‌شوند: خطای تصادفی و خطای منظم. این دو نوع منبع خطا در همترازسازی غیرقابل‌اجتناب است (Devdass, 2011).

خطای تصادفی همترازسازی (خطای نمونه‌گیری) به هنگام برآورد پارامترهایی نظیر میانگین، انحراف استاندارد، و رتبه صدکی، یک نمونه که از کل جامعه بیرون کشیده شده‌اند، رخ می‌دهد (Kolen et al., 2014)؛ به‌عبارت‌دیگر خطای تصادفی به‌عنوان تفاوت بین رابطه هم‌تراز برآورد شده برای نمونه‌ها و کل جامعه در نظر گرفته می‌شود. خطاهای منظم زمانی رخ می‌دهد که از مفروضه‌های آماری یا شرایط روش‌های همترازسازی تخطی صورت گیرد (Kolen et al., 2014)؛ بنابراین اگر ملاکی قادر باشد این خطاها را به‌خوبی ارزیابی نماید، ملاک کارآمدی خواهد بود. در زیر برخی از این ملاک‌ها موردبحث قرار گرفته است.

1. equivalent groups design
2. non – equivalent group anchor
3. common items non – equivalent group design

خطای استاندارد همترازسازی[1] شاخص مفیدی در تعیین مقدار خطای تصادفی همترازسازی است. خطای استاندارد همترازسازی به‌عنوان انحراف استاندارد اختلاف بین نمرات هم‌تراز شده در یک روش همترازسازی و ملاک همترازساری است (Heh, 2007).

خطای منظم همترازسازی بر اساس میزان خطای سوگیری همترازسازی[2] نتایج همترازسازی محاسبه می‌شود که این خطا در اثر نقض مفروضه‌های روش‌های همترازسازی یا استفاده از روش‌های نامناسب در همترازسازی نمرات فرم‌های آزمون ایجاد می‌شود (Caglak, 2016).

جهت تعیین خطای کلی همترازسازی از مجذور میانگین مربع خطای[3] همترازسازی استفاده می‌شود که برابر است با مجذور خطای استاندارد همترازسازی و سوگیری همترازسازی (Aşiret et al., 2016).

با توجه با مسائل مطرح‌شده پژوهش در خصوص روش‌های مختلف همترازسازی در نظریه کلاسیک آزمون، در رفع و حل مشکلات و مسائل موجود در آزمون‌سازی و همترازسازی کمک‌کننده خواهد بود. بر این اساس در این پژوهش در پی پاسخ به این سؤال هستیم که بهینه‌ترین روش همترازسازی در نظریه کلاسیک آزمون کدام است و با کدام‌یک از روش‌های همترازسازی مورداستفاده می‌توان به نتایج دقیق‌تری در همترازسازی فرم‌های آزمون دست یافت؟

پیشینه پژوهش

در خصوص همترازسازی در نمونه‌های با حجم کم مطالعات تجربی اندکی وجود دارد. ازجمله این مطالعات می‌توان به پژوهش Iriyadi و همکاران (2018)، Kim و همکاران (2010)، Livingston و همکاران (2010)، Skaggs (2005)، Asiret و همکاران (2016)، Babcock و همکاران (2012)، Parshall و همکاران (1995)، Puhan و همکاران (2009)، Kim و همکاران (2006)، Caglak (2016) اشاره کرد.

Iriyadi و همکاران (2018) در پژوهش خود تحت عنوان روش‌های همترازسازی برای نمونه کوچک (پژوهش مقایسه‌ای در روش میانگین وزنی اسمی و روش خطی)، با هدف تعیین دقت روش‌های میانگین وزنی اسمی و روش خطی در همترازسازی با استفاده از نمونه

1 - standard error of equating
2. equating bias
3. root mean square error

کوچک به‌عنوان ابزاری برای معلمان در همترازسازی نمرات دانش‌آموزان در کلاس، تعداد ۳۰ نفر موردبررسی قرار گرفتند. تعداد سؤال‌های آزمون ۳۰ سؤال و تعداد سؤال‌های لنگر ۶ سؤال در نظر گرفته شد. نتایج به‌دست‌آمده حاکی از ارجحیت روش میانگین وزنی اسمی[1] نسبت به روش خطی بود. به‌نحوی‌که مقدار مجذور میانگین مربع خطاها[2] در روش میانگین وزنی اسمی کوچک‌تر از روش همترازسازی خطی بود.

همچنین در مطالعه Kim و همکاران (2010) با عنوان مقایسه روش‌های همترازسازی در نمونه کوچک با الگوی سؤال مشترک، روش‌های همترازسازی شامل همترازسازی همصدک دنباله‌ای[3] از توزیع‌های هموار یا یکنواخت، همترازسازی خطی دنباله‌ای، همترازسازی میانگین دنباله‌ای، روش قوس دایره‌ای متقارن، روش قوس دایره‌ای آسان شده با حجم نمونه‌های ۱۰، ۲۵، ۵۰ و ۱۰۰ موردبررسی و مقایسه قرار گرفت. نتایج نشان داد در توزیع نمرات پایین روش همترازسازی میانگین دنباله‌ای و در توزیع نمرات بالاتر از متوسط دو روش قوس دایره‌ای نتایج دقیق‌تری به دست دادند.

Livingston و همکاران (2010) در مطالعه دیگری با عنوان همترازسازی گروه‌های تصادفی در نمونه‌هایی از ۵۰ تا ۴۰۰ آزمودنی، روش‌های همترازسازی شامل همترازسازی همصدک، خطی، میانگین، و قوس دایره‌ای (متقارن و ساده) را مورد مقایسه قرار دادند. بر اساس نتایج به‌دست‌آمده در صدک ۵۰ همه روش‌های همترازسازی در میزان دقت همترازسازی (به‌عبارت‌دیگر در مقدار خطای همترازسازی) یکسان بودند، درحالی‌که با افزایش نمرات به سمت صدک ۷۵ در میزان عملکرد و دقت همترازسازی بین روش قوس دایره‌ای و سایر روش‌های همترازسازی اختلاف بیشتر می‌شد. در صدک ۲۵ و پایین‌تر روش همترازسازی قوس دایره‌ای و روش همترازسازی میانگین عملکرد بهتری داشته و نتایج دقیق‌تری را نسبت به سایر روش‌های همترازسازی به دست دادند. درمجموع نتایج برتری روش قوس دایره‌ای را بر اساس بیشترین دقت همترازسازی در بین سایر روش‌های همترازسازی در همه گروه‌های با حجم نمونه‌های مختلف را نشان داد.

در پژوهشی مشابه Skaggs (2005) میزان دقت روش‌های همترازسازی با الگوی گروه‌های تصادفی در نمونه‌های خیلی کوچک را مورد مقایسه قرار داد. هدف این پژوهش

---

1. nominal weight mean
2. root mean square error
3. chained

بررسی اثربخشی روش‌های همترازسازی با نمونه‌های خیلی کوچک با استفاده از الگوی گروه‌های تصادفی بود. روش‌های همترازسازی مورداستفاده شامل همترازسازی خطی، همترازسازی میانگین، همترازسازی همصدک ناهموار، و همترازسازی همصدک در نمونه‌های ۲۵، ۵۰، ۷۵، ۱۰۰، ۱۵۰ و ۲۰۰ نفر بود. همان‌گونه که انتظار می‌رفت مقدار خطای استاندارد همترازسازی با افزایش حجم نمونه کاهش یافت؛ هرچند میزان سوگیری همترازسازی به‌عنوان تابعی از حجم نمونه تغییر اندکی نشان داد. علی‌رغم این موضوع در خصوص نمونه با حجم ۲۰۰ نفر نیز در بخش‌های کمی از مقیاس نمرات خام، میزانی از خطای استاندارد مشاهده شد. درمجموع نتایج به‌دست‌آمده حاکی از این بود که در نمرات پایین‌تر از میانگین، روش همترازسازی میانگین، و در نمرات نزدیک میانگین همه روش‌های همترازسازی و در نمرات بالاتر از میانگین روش همترازسازی همصدک بیشترین دقت را داشتند. همچنین نتایج حاکی از کاهش خطای استاندارد روش‌های همترازسازی در نمونه‌های با دامنه ۲۵ تا ۵۰ نفر بوده است.

در پژوهشی تحت عنوان بررسی روش‌های همترازسازی در نمونه‌های با حجم پایین با توجه به عوامل متعدد، Asiret و همکاران (2016) به مقایسه روش‌های همترازسازی هویت[1]، میانگین، خطی، قوس دایره‌ای، روش همصدک با پیش‌هموارسازی، در نمونه‌هایی با اندازه مختلف (۱۰، ۲۵، ۵۰، ۷۵، ۱۰۰، ۱۵۰، ۲۰۰) با الگوی گروه‌های تصادفی پرداختند. در این مطالعه تأثیر عوامل مختلفی نظیر سطوح دشواری متفاوت (۰,۱، ۰,۴، و ۰,۷) و عامل حدس بر میزان دقت و عملکرد روش‌های همترازسازی موردبررسی قرار گرفت. نتایج نشان داد که در سطح دشواری ۰,۴ با حجم نمونه ۵۰ و بالاتر روش‌های همترازسازی مقدار خطای مجذور ریشه میانگین کمتری داشتند و درمجموع نیز روش‌های همترازسازی قوس دایره‌ای و میانگین خطای همترازسازی کمتری نسبت به سایر روش‌ها به دست دادند.

Babcock و همکاران (2012) در پژوهش خود با موضوع همترازسازی میانگین وزنی اسمی: یک روش برای نمونه‌های خیلی کوچک به مقایسه روش‌های همترازسازی همصدک، تاکر، میانگین، میانگین وزنی اسمی، هویت، ترکیبی[2]، هویت، و قوس دایره‌ای با الگوی گروه‌های ناهمسان با سؤالات مشترک پرداختند. حجم نمونه موردبررسی در این پژوهش ۲۰، ۲۵، و ۸۰ آزمودنی بود. نتایج پژوهش نشان داد روش میانگین وزنی اسمی

1. identity
2. synthetic

تحت هر شرایط (سطوح دشواری مختلف، سطوح توانایی مختلف، حجم نمونه) کاراترین روش است. روش همترازسازی هویت صرفاً درصورتی‌که دو فرم آزمون تفاوتی در سطوح دشواری نداشتند بیشترین دقت را داشت. در صورت تفاوت در دشواری و همسانی در سطوح توانایی روش‌های همترازسازی میانگین و میانگین وزنی اسمی مقدار خطای کمتری را به دست می‌دادند. در گروه‌های با سطوح دشواری و سطوح توانایی متفاوت روش‌های همترازسازی میانگین وزنی اسمی و قوس دایره‌ای بهترین عملکرد را داشتند.

در پژوهش Parshall و همکاران (1995) تأثیر حجم نمونه (۱۵، ۲۵، ۵۰ و ۱۰۰) نفر در میزان ثبات و سوگیری همترازسازی خطی با دو فرم موازی مبتنی بر الگوی گروه‌های معادل با سؤالات مشترک (لنگر) بررسی شد. نتایج آن‌ها نشان داد با کاهش حجم نمونه مقدار بسیار کمی از سوگیری نتایج همترازسازی مشاهده می‌شود ولی در خصوص خطای استاندارد همترازسازی کاهش حجم نمونه منجر به افزایش خطای مذکور می‌شود. همچنین در خصوص نمراتی که در اطراف میانگین نمره خام قرار داشتند خطای استاندارد همترازسازی در کمترین حالت بود و درعین‌حال با افزایش فاصله نمرات از میانگین (انحراف بالای نمرات از میانگین) در مقدار خطای استاندارد همترازسازی به‌طور یکنواختی افزایش مشاهده می‌شد. این یافته‌ها نشان دادند که خطای استاندارد همترازسازی در نمونه‌های کوچک با افزایش فاصله نمرات از میانگین ارتباط داشته و مقدار آن حالت افزایشی دارد.

Puhan و همکاران (2009) در پژوهش خود که با هدف بررسی همترازسازی نمونه‌های کوچک با استفاده از الگوی آزمون تقریباً هم‌ارز تک گروهی[1] در نمونه‌هایی با حجم ۱۰، ۱۵، ۲۵ و ۵۰ آزمودنی انجام شده، نشان دادند میانگین خطای همترازسازی با توجه به هر سه شاخص محاسبه‌شده (مجذور میانگین مربع خطاها، خطای استاندارد همترازسازی و خطای سوگیری همترازسازی) در الگوی آزمون تقریباً هم‌ارز تک گروهی در نمونه‌های با حجم پایین کمتر از الگوی همترازسازی گروهای ناهمسان با سؤالات مشترک (لنگر) است.

نتایج پژوهش Kim و همکاران (2006) در خصوص مقایسه روش‌های همترازسازی با استفاده از الگوی گروه‌های نامعادل با آزمون لنگر، نمونه‌هایی با حجم ۱۵، ۲۰، ۵۰ و ۱۰۰ نفر را موردبررسی قرار دادند. نتایج حاکی از عملکرد بهتر روش‌های همترازسازی نظیر هویت، خطی و ترکیبی (ترکیب روش خطی و هویت) نسبت به روش‌های همترازسازی سنتی در نمونه‌های حجم کم بود.

1. single – group nearly equivalent test

در پژوهشی با عنوان مقایسه چندین روش همترازسازی در نمونه‌های با حجم پایین تحت الگوی گروه‌های نامعادل با سؤالات مشترک (لنگر)، Caglak (2016) تأثیر حجم نمونه‌هایی با اندازه ۱۰، ۲۰، ۵۰، و ۱۰۰ آزمودنی را در عملکرد روش‌های همترازسازی هویت، میانگین وزنی اسمی، و قوس دایره‌ای، مورد ارزیابی قرار داد. در این پژوهش روش تابع ترکیبی[1] (سیستم وزن‌دهی یکسان) معرفی و تأثیر آن بر سایر روش‌های همترازسازی موردبررسی قرار گرفت. این روش به‌عنوان عاملی در کنترل خطای همترازسازی مطرح شده است. نتایج این مطالعه حاکی از این بود که با استفاده از روش تابع ترکیبی روش همترازسازی میانگین وزنی خطای همترازسازی کمتری از سایر روش‌ها ایجاد می‌کند. همچنین یافته‌ها نشان دادند کاربرد روش تابع ترکیبی در روش‌های همترازسازی به استفاده صرف این روش‌ها ارجحیت دارد و نتایج دقیق‌تری به دست می‌دهد.

با هدف همترازسازی نمونه‌های کوچک با استفاده از تابع پیونددهی ترکیبی، Kim و همکاران (2006) پژوهشی با استفاده از الگوی گروه‌های نامعادل با آزمون لنگر انجام دادند. در این پژوهش روش‌های همترازسازی ترکیبی، هویت، و خطی دنباله‌ای در گروه‌هایی با حجم نمونه ۱۵، ۲۵، ۵۰، و ۱۰۰ نفری مورد مقایسه قرار گرفتند. نتایج نشان داد روش همترازسازی تابع ترکیبی زمانی که حجم نمونه خیلی بزرگ نیست نسبت به دو روش هویت و خطی دنباله‌ای بهترین انتخاب است و مقدار خطای سوگیری و خطای همترازسازی کمتری را ایجاد می‌کند.

Kim و همکاران (2011) در پژوهشی با عنوان اطلاعات جانبی در همترازسازی در نمونه‌های با حجم کم، به بررسی تأثیر حجم نمونه (۱۰، ۲۵، ۵۰، ۷۵، ۱۰۰، و ۲۰۰ نفر) در میزان دقت همترازسازی با استفاده از روش‌های خطی دنباله‌ای و میانگین دنباله‌ای پرداختند. با استفاده از فرم اصلی دو فرم جدید طراحی شد. حجم نمونه در مورد فرم جدید از ۱۰ تا ۲۰۰ متغیر و در مورد فرم اصلی تعداد ۲۰۰ نفر ثابت بود. یکی از دو فرم جدید ازنظر دشواری تفاوت زیادی با فرم اصلی نداشت، درحالی‌که فرم دیگر از فرم اصلی دشوارتر بود. نتایج نشان داد در مورد فرمی که ازنظر دشواری تفاوتی با فرم اصلی نداشت دقت همترازسازی به‌ویژه وقتی حجم نمونه کمتر می‌شد بیشتر بود و در مورد فرم دوم دقت همترازسازی کمتر بود. علت این تفاوت در تفاوت دشواری فرم‌ها بود.

1. synthetic function

روش

نظر به اینکه در این پژوهش محقق در پی مقایسه رویکردهای مختلف در همترازسازی و به دست آوردن برآوردهای باثبات‌تر پارامترها (اعم از پارامترهای سؤال و توانایی) و افزایش اعتبار تصمیم‌ها و دقت اندازه‌گیری و درنهایت تعیین روش بهینه همترازسازی بوده، لذا دو نوع مطالعه بر اساس دو روش پژوهش طرح‌ریزی شد. ازاین‌رو متناسب با نوع مطالعه، داده‌های مورداستفاده جهت تجزیه‌وتحلیل در دو بخش داده‌های واقعی و داده‌های تجربی حاصل از شبیه‌سازی مورداستفاده قرار گرفت.

در مطالعه اول داده‌های واقعی به‌دست‌آمده از اجرای فرم‌های الف و ب آزمون مهارت تفکر انتقادی کالیفرنیا با روش‌های همترازسازی مبتنی بر نظریه کلاسیک آزمون بررسی و ارزیابی شد و درنهایت دقت هر یک از روش‌ها با یکدیگر مورد مقایسه قرار گرفته است. روش پژوهش در این مطالعه به‌طور عام جزء پژوهش‌های توصیفی است. همچنین در مطالعه دوم عمدتاً به دلیل وقت‌گیر و هزینه‌بر بودن روش‌های تجربی در دسترسی به داده‌ها و عدم امکان اجرای پژوهش در شرایط واقعی، از داده‌های شبیه‌سازی‌شده استفاده شد. ازاین‌رو با استفاده از داده‌های شبیه‌سازی‌شده با دست‌کاری و مشاهده عواملی نظیر حجم نمونه و ضریب دشواری به‌عنوان متغیرهای مستقل و نحوه تأثیر این عوامل بر میزان دقت روش‌های همترازسازی خطای استاندارد اندازه‌گیری، سوگیری نتایج همترازسازی، و مجذور میانگین مربع خطاها به‌عنوان متغیر وابسته، موردبررسی قرار گرفت؛ که از حیث نوع تجزیه‌وتحلیل آماری روش این بخش از مطالعه جزء پژوهش‌های آزمایشی است؛ بنابراین در این مطالعه با استفاده از ویژگی‌های داده‌های واقعی آزمون تفکر انتقادی کالیفرنیا[1] (فرم الف و ب)، به ایجاد شرایط و تولید داده‌ها در موقعیت‌هایی با حجم نمونه ۵۰، ۱۰۰، و ۲۰۰ و اختلاف در سطوح دشواری سؤالات (۰/۱، ۰/۴، و ۰/۷) پرداخته‌شده و عوامل مذکور در چهار روش همترازسازی خطی، میانگین، همصدک و قوس دایره‌ای مورد مقایسه قرار گرفت.

با توجه به دو نوع داده مورداستفاده در پژوهش حاضر (داده‌های واقعی و داده‌های شبیه‌سازی‌شده)، و متناسب با آن دو نوع مطالعه، دو نوع جامعه نیز مطرح است. در موردمطالعه اول به‌منظور بررسی مقایسه‌ای روش‌های همترازسازی و با توجه به اجرای پرسش‌نامه تفکر انتقادی کالیفرنیا (فرم الف وب)، از داده‌های واقعی استفاده شد. جامعه

1. California Critical Thinking Skills Test

آماری در این مطالعه شامل کلیه کارکنان در تمامی رده‌های شغلی شاغل در شهرداری قزوین در سال ۱۳۹۹ با حجم ۲۰۴۸ نفر است. همچنین به جهت استفاده از الگوی گروه‌های تصادفی همترازسازی، دو نمونه به‌صورت کاملاً تصادفی و مستقل با حجم ۲۰۰ نفر با استفاده از فرمول تعیین حجم نمونه کوکران با مقدار خطای ۰,۰۱ از جامعه مذکور انتخاب شد. در خصوص مطالعه شبیه‌سازی نیز در حالات مختلف (حجم نمونه و ضرایب دشواری) و با توجه به میانگین پارامتر تمیز به‌دست‌آمده از اجرای واقعی برای ۲۰۰ نفر آزمودنی، دو آزمون با داده‌های شبیه‌سازی ایجاد شد. همچنین به‌منظور تعیین میزان دقت روش‌های همترازسازی در نمونه‌های حجم کم، ۲ گروه نمونه دیگر با حجم ۵۰ و ۱۰۰ نفر نیز انتخاب و موردبررسی قرار گرفت.

داده‌های واقعی با استفاده از دو فرم الف و ب پرسش‌نامه تفکر انتقادی کالیفرنیا جمع‌آوری شد. این پرسش‌نامه توسط Facion و همکاران (1990) به‌منظور سنجش مهارت‌های اساسی تفکر انتقادی بزرگ‌سالان در دو فرم موازی "الف" و "ب" طراحی و ساخته شده است. هر دو فرم این پرسش‌نامه شامل ۳۴ سؤال چندگزینه‌ای (۲۰ سؤال ۴ گزینه‌ای و ۱۴ سؤال ۵ گزینه‌ای) با یک پاسخ صحیح در پنج حوزه مهارت‌های تفکر انتقادی شامل ارزشیابی، استنباط، تحلیل، استدلال قیاسی و استدلال استقرایی است. امتیاز نهایی بین ۰ و ۳۴ است. پایایی پرسش‌نامه در پژوهش Facion و همکاران (1997) برای فرم "الف" ۰/۷۰ و برای فرم "ب" ۰/۷۱ گزارش شده است. همچنین در ایران نیز مهری نژاد (۱۳۸۶) پایایی فرم "الف" را به روش دونیمه کردن ۰/۷۸، با روش اسپیرمن براون ۰/۸۸ و با روش آلفای کرونباخ ۰/۸۳؛ عسگری و همکاران (۱۳۸۹) پایایی فرم "ب" را به روش کودرریچاردسون ۰/۶۸، به روش دو نیمه کردن ۰/۵۵، و با روش بازآزمایی ۰/۶۸، دهقانی و همکاران (۱۳۸۹) ۰/۷۸، خلیلی و همکاران (۱۳۸۲) ۰/۶۲ گزارش کرده‌اند. همچنین پایایی فرم "ب" در سال‌های اخیر در پژوهش اصغری (۱۳۹۶) ۰/۸۲ به دست آمده است.

پس از اجرا و تحلیل سؤالات مربوط به پرسش‌نامه تفکر انتقادی کالیفرنیا، ابتدا شاخص تمیز سؤال‌های هر یک از دو فرم الف و ب آزمون تفکر انتقادی کالیفرنیا برآورد و میانگین و انحراف استاندارد شاخص تمیز داده‌های واقعی تعیین شد و با استفاده از آن داده‌های شبیه‌سازی تولید شدند. طرح تولید داده‌ها یک طرح متقاطع ۳ (حجم نمونه) در ۳ (ضریب دشواری) برای ۴ روش همترازسازی بود.

در این پژوهش محاسبه خطای همترازسازی به‌عنوان ملاک ارزیابی میزان دقت نتایج همترازسازی است که کاهش آن حاکی از دقت نتایج همترازسازی است. با توجه به اینکه خطاهای ایجادشده در فرایند همترازسازی ناشی از دو منبع خطاهای تصادفی و خطاهای منظم هستند (Kolen et al., 2014)، به‌منظور دستیابی به اهداف پژوهش در خصوص ارزیابی روش‌های همترازسازی در قالب طرح گروه‌های تصادفی بر مبنای نظریه کلاسیک آزمون برحسب خطاها و همچنین بررسی عوامل مؤثر بر افزایش دقت روش‌های همترازسازی و نهایتاً انتخاب بهینه‌ترین و مطلوب‌ترین روش همترازسازی در نمونه‌های با حجم کم، و با توجه به‌ضرورت انتخاب ملاک‌هایی جهت ارزیابی نتایج همترازسازی با توجه به دو نوع خطای تصادفی و منظم ملاک‌های ارزیابی عملکرد روش‌های همترازسازی در خصوص داده‌های واقعی، جهت ارزیابی روش‌های همترازسازی میانگین، خطی، همصدک و قوس دایره‌ای از روش همترازسازی همصدک و قوس دایره‌ای به‌عنوان معیار مقایسه استفاده شده است. همچنین در خصوص داده‌های شبیه‌سازی‌شده از ملاک‌هایی همچون خطای استاندارد همترازسازی برای ارزیابی خطای تصادفی همترازسازی، ملاک سوگیری همترازسازی برای ارزیابی خطای منظم همترازسازی، و مجذور میانگین مربع خطاها برای ارزیابی دقت کلی همترازسازی استفاده شده است.

الگوی مورداستفاده در مطالعه حاضر در هر دو دسته از داده‌ها یعنی داده‌های واقعی و داده‌های شبیه‌سازی، الگوی گروه‌های تصادفی است. در الگوی گروه‌های تصادفی هر یک از فرم‌های آزمون به‌طور تصادفی صرفاً در مورد یک گروه اجرا می‌شود. به این صورت که پس از گروه‌بندی تصادفی اعضای نمونه در دو گروه اول و دوم، فرم‌های الف و ب به‌طور تصادفی در اختیار یک گروه قرار داده می‌شود. در این حالت به دلیل انتخاب و گمارش تصادفی نمونه در گروه‌ها فرض معادل بودن گروه‌ها نیز مدنظر قرار می‌گیرد. روش‌های هم‌تراز سازی مورداستفاده در پژوهش حاضر شامل روش‌های همترازسازی مبتنی بر نظریه کلاسیک آزمون شامل همترازسازی میانگین، خطی، همصدک و قوس دایره‌ای است. قبل از انجام همترازسازی با روش‌های مذکور سؤالات آزمون و نمرات بر اساس نظریه کلاسیک مورد تجزیه‌وتحلیل قرار گرفت. به‌این‌ترتیب در ابتدا شاخص‌های اولیه نظیر ضرایب تمیز، دشواری و پایایی و ... بررسی و سپس با روش‌های مختلف همترازسازی، هم‌تراز و با استفاده از ملاک‌های ارزیابی شامل ریشه میانگین مجذور تفاوت، میانگین خطا و ریشه میانگین

مجذور خطا مورد مقایسه قرار خواهند گرفت. جهت تجزیه‌وتحلیل داده‌های جمع‌آوری‌شده از نرم‌افزارهای R packages: equate, mirt, SNSequate, ggplot2 استفاده شده است.

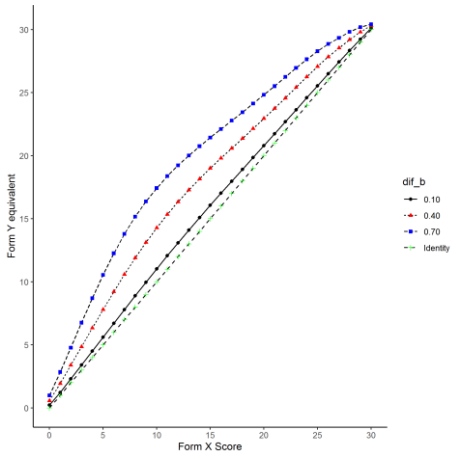
یافته‌ها

یافته‌ها در دو بخش داده‌های شبیه‌سازی‌شده و داده‌های واقعی ارائه شده‌اند. در قسمت شبیه‌سازی عملکرد چهار روش همترازسازی در حالت‌های مختلف شبیه‌سازی با سه شاخص SEE، BE، و RMSE مقایسه شد. تأثیر حجم نمونه و تفاوت دشواری فرم‌ها بر این شاخص‌ها با استفاده از نمودارها بررسی شده است. در بخش داده واقعی، نخست ویژگی‌های جمعیت شناختی نمونه در قالب جدول فراوانی و نمودار ستونی ارائه شده‌اند. سپس نمرات فرم B با نمرات فرم A هم‌تراز شده‌اند و عملکرد روش‌های همترازسازی همصدک، خطی، میانگین و قوس دایره‌ای با سه شاخص SEE، BE، و RMSE که به روش بوت استرپینگ به دست آمده‌اند مقایسه شده است.

بخش شبیه‌سازی

در فرم y میانگین ضرایب دشواری ۰/۰۳ با انحراف معیار ۰/۸٦ است. میانگین و انحراف معیار ضرایب تمیز به ترتیب ۱/۳۳ و ۰/٤۷ است. ضرایب تمیز فرم x با فرم y یکسان بود. پارامترهای دشواری فرم x با اضافه کردن مقادیر ۰/۱۰، ۰/٤۰، و ۰/۷۰ به ضرایب دشواری فرم y به دست آمد.

نمودار ۱. رابطه هم‌تراز سازی واقعی به تفکیک تفاوت ضرایب دشواری فرم‌ها



نمودار ۱ رابطه همترازسازی واقعی که به روش نمره مشاهده‌شده مبتنی بر IRT محاسبه‌شده را نشان می‌دهند. در نمودار خط سبزرنگ رابطه این‌همانی (identity) را نشان می‌دهد که هم‌تراز هر نمره خود آن نمره است. رابطه هم‌تراز سازی برای تفاوت دشواری ۰/۱۰، ۰/۴۰، و ۰/۷۰ به ترتیب با خطوط سیاه، قرمز و آبی مشخص شده است. با افزایش تفاوت دشواری فرم‌های x و y همترازهای برآورده در مقایسه با ملاک این‌همانی، افزایش بیشتری می‌یابند. بیشترین افزایش در بازه نمره ۵ تا ۱۵ اتفاق می‌افتد. برای مثال در حالت‌های تفاوت دشواری ۰/۱۰، ۰/۴۰، ۰/۷۰، همترازهای واقعی نمره ۱۰ به ترتیب برابر ۱۱/۰۲، ۱۴/۲۸، ۱۷/۴۳ هستند.

شاخص خطای سوگیری همترازسازی (EB)

جدول ۱. سوگیری همترازسازی در حالت‌های مختلف شبیه‌سازی#

حالت	حجم نمونه	تفاوت دشواری	همصدک	خطی	میانگین	قوس دایره‌ای
۱	۵۰	۰/۱	۰/۳۷	۰/۳۸	۰/۳۸	۰/٤٥
۲	۱۰۰	۰/۱	۰/۲۸	۰/۳۲	۰/۳۰	۰/۳۷
۳	۲۰۰	۰/۱	۰/۱۵	۰/۲٤	۰/۲٤	۰/۲۸
٤	۵۰	۰/٤	۰/۹۵	۱/۱۵	۱/۱۵	۱/۳۹
۵	۱۰۰	۰/٤	۰/۸۲	۱/۰۷	۱/۰۸	۱/۳۲
٦	۲۰۰	۰/٤	۰/۸۲	۱/۰۷	۱/۰۸	۱/۳۲
۷	۵۰	۰/۷	۱/٦۱	۲/۰۱	۲/۰۱	۲/٤۲
۸	۱۰۰	۰/۷	۱/۵٤	۱/۹٦	۱/۹٦	۲/۳٦
۹	۲۰۰	۰/۷	۱/۵۲	۱/۹۳	۱/۹٤	۲/۳۵

#در هر حالت زیر کمترین مقدار خط کشیده شده است.

همان‌طور که در جدول ۱ مشاهده می‌شود، در همه ۹ حالت شبیه‌سازی روش همترازسازی همصدک کمترین میزان سوگیری (bias) را در بین ۴ روش دارد. در بین ۹ حالت، کمترین میزان سوگیری ۰/۱۵ است که مربوط به حالت ۳ (حجم نمونه ۲۰۰، تفاوت دشواری ۰/۱۰) است. روش قوس دایره‌ای در همه حالت‌ها بیشترین میزان سوگیری را دارد. در بین ۹ حالت بیشترین سوگیری ۲/۴۲ است، که مربوط به حالت ۷ (حجم نمونه ۵۰، تفاوت

دشواری ۰/۷۰) است. نکته دیگر شباهت عملکرد دو روش همترازسازی میانگین، و خطی است، که خطوط آن‌ها بسیار نزدیک به هم هستند.

جدول ۲. سوگیری همترازسازی به تفکیک حجم نمونه و تفاوت دشواری

مقادیر		همصدک	خطی	میانگین	قوس دایره‌ای
حجم نمونه	۵۰	۰/۹۸	۱/۱۸	۱/۱۸	۱/٤۲
	۱۰۰	۰/۸۸	۱/۱۲	۱/۱۱	۱/۳۵
	۲۰۰	۰/۸۳	۱/۰۸	۱/۰۹	۱/۳۲
تفاوت دشواری	۰/۱۰	۰/۲۷	۰/۳۱	۰/۳۱	۰/۳۷
	۰/٤۰	۰/۸٦	۱/۱۰	۱/۱۰	۱/۳۵
	۰/۷۰	۱/٥٦	۱/۹۷	۱/۹۷	۲/۳۸

جدول ۲ سوگیری (bias) را به تفکیک حجم نمونه و تفاوت دشواری نشان می‌دهد. برای هر چهار روش همترازسازی با افزایش حجم نمونه میزان سوگیری کاهش می‌یابد. حجم نمونه بیشترین تأثیر را در کاهش سوگیری برای روش همترازسازی همصدک داشته است. میزان سوگیری از ۰/۹۸ در حجم نمونه ۵۰ نفر به ۰/۸۳ در حجم نمونه ۲۰۰ نفر کاهش یافته است. تأثیر حجم نمونه برای دو روش خطی و میانگین مشابه بوده است. در روش قوس دایره‌ای میزان سوگیری از ۱/۴۲ به ۱/۳۲ رسیده است، که بیشترین کاهش پس از روش همصدک است. در هر ۴ روش با افزایش تفاوت دشواری دو فرم میزان سوگیری روش‌های همترازسازی افزایش می‌یابد. روش همترازسازی دایره‌ای بیشترین تأثیر را از تفاوت دشواری فرم‌ها می‌پذیرد. کمترین تأثیر تفاوت دشواری فرم‌ها بر سوگیری، مربوط به روش همصدک است. تأثیر تفاوت دشواری بر سوگیری همترازسازی روش‌های خطی و میانگین مشابه است.

خطای استاندارد همترازسازی (SEE)

جدول ۳. خطای استاندارد همترازسازی در حالت‌های مختلف شبیه‌سازی[#]

حالت	حجم نمونه	تفاوت دشواری	هم صدک	خطی	میانگین	قوس دایره‌ای
۱	۵۰	۰/۱	۱/٥٦	۱/۲۷	۱/۰۸	۰/۸۹
۲	۱۰۰	۰/۱	۱/۲۰	۱/۰۱	۰/۸٦	۰/۷۲
۳	۲۰۰	۰/۱	۰/۸۸	۰/۷۲	۰/٦۲	۰/٥٥
٤	۵۰	۰/٤	۱/٦۳	۱/۳۷	۱/۱۷	۱/۱۹
۵	۱۰۰	۰/٤	۱/۲٦	۱/۰۸	۰/۹٥	۱/۰٥

حالت	حجم نمونه	تفاوت دشواری	هم صدک	خطی	میانگین	قوس دایره‌ای
٦	۲۰۰	۰/٤	۰/۹۱	۰/۷٤	۰/٦٥	۰/۸٥
۷	٥۰	۰/۷	۱/٦٦	۱/٤۲	۱/۱۷	۱/٥۰
۸	۱۰۰	۰/۷	۱/۳٥	۱/۱٦	۰/۹٦	۱/۳۷
۹	۲۰۰	۰/۷	۰/۸۹	۰/۷٦	۰/٦٤	۱/۱۷

#در هر حالت زیر کمترین مقدار خط کشیده شده است.

همان‌طور که در جدول ۳ مشاهده می‌شود، در ۳ حالت با تفاوت دشواری ۰/۱۰ روش قوس دایره‌ای کمترین خطای استاندارد همترازسازی (SEE) را دارد. در ۶ حالت با تفاوت دشواری‌های ۰/۴۰ و ۰/۷۰ روش همترازسازی میانگین کمترین میزان SEE را دارد. کمترین SEE=۰/۵۵ است که مربوط روش قوس دایره‌ای در حالت ۳ (حجم نمونه ۲۰۰، تفاوت دشواری ۰/۱۰) است. روش همصدک در هفت حالت بیشترین SEE را دارد. در دو حالت (تفاوت دشواری ۰/۷۰ با حجم نمونه ۱۰۰ و ۲۰۰) روش قوس دایره‌ای بیشترین خطای استاندارد را دارد. در بین حالت‌های مختلف بیشترین SEE=۱/۶۶ است، که مربوط به روش همصدک در حالت ۷ (حجم نمونه ۵۰، تفاوت دشواری ۰/۷۰) است.

جدول ٤ خطای استاندارد همترازسازی به تفکیک حجم نمونه و تفاوت دشواری

	مقادیر	همصدک	خطی	میانگین	قوس دایره‌ای
	٥۰	۱/٦۱	۱/۳٥	۱/۱٤	۱/۱۹
حجم نمونه	۱۰۰	۱/۲۷	۱/۰۹	۰/۹۲	۱/۰٥
	۲۰۰	۰/۸۹	۰/۷٤	۰/٦٤	۰/۸٦
	۰/۱۰	۱/۲۱	۱	۰/۸٥	۰/۷۲
تفاوت دشواری	۰/٤۰	۱/۲۷	۱/۰۷	۰/۹۲	۱/۰۳
	۰/۷۰	۱/۳۰	۱/۱۱	۰/۹۲	۱/۳٥

جدول ۴ خطای استاندارد همترازسازی (SEE) را به تفکیک حجم نمونه و تفاوت دشواری نشان می‌دهد. برای هر چهار روش همترازسازی با افزایش حجم نمونه میزان خطای استاندارد کاهش می‌یابد. حجم نمونه بیشترین تأثیر را در کاهش خطای استاندارد برای روش همترازسازی همصدک داشته است، که میزان خطای استاندارد از ۱/۶۲ در حجم نمونه ۵۰ نفر به ۰/۹۰ در حجم نمونه ۲۰۰ نفر کاهش یافته است. در روش قوس دایره‌ای میزان خطای استاندارد از ۱/۲۰ به ۰/۸۶ رسیده است، که کمترین کاهش است. خطای استاندارد دو روش

همترازسازی خطی و میانگین هم با افزایش حجم نمونه کاهش یافته است که کاهش روش خطی بیشتر بوده است. در روش میانگین با افزایش تفاوت دشواری دو فرم از ۰/۱۰ به ۰/۴۰ میزان خطای استاندارد همترازسازی افزایش می‌یابد؛ اما خطای استاندارد برای ۰/۴۰ و ۰/۷۰ یکسان است. در سه روش همترازسازی دیگر، با افزایش تفاوت دشواری دو فرم، خطای استاندارد همترازسازی نیز افزایش می‌یابد. بیشترین تأثیر تفاوت دشواری بر خطای استاندارد در روش قوس دایره‌ای مشاهده شد که خطای استاندارد از ۰/۷۲ به ۱/۳۵ افزایش یافت.

مجذور میانگین مربع خطاها (RMSE)

جدول ۵. شاخص RMSE در حالت‌های مختلف شبیه‌سازی[#]

حالت	حجم نمونه	تفاوت دشواری	همصدک	خطی	میانگین	قوس دایره‌ای
۱	۵۰	۰/۱	۱/۶۱	۱/۳٤	۱/۱٦	۱
۲	۱۰۰	۰/۱	۱/۲٤	۱/۰۸	۰/۹۲	۰/۸۲
۳	۲۰۰	۰/۱	۰/۹۰	۰/۷۷	۰/٦۸	۰/٦۲
٤	۵۰	۰/٤	۱/۹۲	۱/۸۹	۱/۷۰	۱/۸۷
۵	۱۰۰	۰/٤	۱/۵۳	۱/٦۱	۱/۵۰	۱/۷٤
٦	۲۰۰	۰/٤	۱/۲۵	۱/۳۸	۱/۳۱	۱/٦٤
۷	۵۰	۰/۷	۲/۳۸	۲/٦۰	۲/٤۰	۲/۹٤
۸	۱۰۰	۰/۷	۲/۱۱	۲/٤۰	۲/۲۵	۲/۸۵
۹	۲۰۰	۰/۷	۱/۸۱	۲/۱٦	۲/۱	۲/۷۸

#در هر حالت زیر کمترین مقدار خط کشیده شده است.

در جدول ۵ شاخص RMSE ارائه شده است. در سه حالت با تفاوت دشواری ۰/۱۰، روش همترازسازی قوس دایره‌ای کمترین میزان RMSE را در بین ۴ روش دارد. در تفاوت دشواری ۰/۴۰ دو حالت با حجم نمونه ۵۰ و ۱۰۰ نفر، روش میانگین کمترین RMSE را دارد. در حالت دشواری ۰/۴۰ با حجم نمونه ۲۰۰ نفر و سه حالت با تفاوت دشواری ۰/۷۰، روش همصدک کمترین RMSE را دارد. کمترین مقدار RMSE=۰/۶۳ است که مربوط به روش قوس دایره‌ای در حجم نمونه ۲۰۰ نفر و تفاوت دشواری ۰/۱۰ است. بیشترین مقدار RMSE=۲/۹۵ است، که مربوط به روش قوس دایره‌ای در حجم نمونه ۵۰ نفر و تفاوت دشواری ۰/۷۰ است.

جدول ٦. شاخص RMSE همترازسازی به تفکیک حجم نمونه و تفاوت دشواری

	مقادیر	همصدک	خطی	میانگین	قوس دایره‌ای
حجم نمونه	۵۰	۱/۹۷	۱/۹٤	۱/۷٦	۱/۹٤
	۱۰۰	۱/٦۳	۱/۷۰	۱/۵٦	۱/۸۰
	۲۰۰	۱/۳۲	۱/٤٤	۱/۳٦	۱/٦۸
تفاوت دشواری	۰/۱۰	۱/۲۵	۱/۰۷	۰/۹۲	۰/۸۱
	۰/٤۰	۱/۵۷	۱/٦۳	۱/۵۱	۱/۷۵
	۰/۷۰	۲/۱۰	۲/۳۹	۲/۲۵	۲/۸٦

جدول ۶ شاخص RMSE را به تفکیک حجم نمونه و تفاوت دشواری نشان می‌دهد. برای هر چهار روش همترازسازی با افزایش حجم نمونه میزان RMSE کاهش می‌یابد. حجم نمونه بیشترین تأثیر را در کاهش RMSE برای روش همترازسازی همصدک داشته است، که میزان RMSE از ۱/۹۷ در حجم نمونه ۵۰ نفر به ۱/۳۲ در حجم نمونه ۲۰۰ نفر کاهش یافته است. روش همترازسازی خطی پس از روش همصدک بیشترین کاهش RMSE را درنتیجه افزایش حجم نمونه داشته است. در روش قوس دایره‌ای میزان RMSE از ۱/۹۴ به ۱/۶۸ رسیده است، که کمترین کاهش در بین ۴ روش است. در هر ۴ روش با افزایش تفاوت دشواری دو فرم میزان RMSE روش‌های همترازسازی افزایش می‌یابد. روش همترازسازی قوس دایره‌ای بیشترین تأثیر را از تفاوت دشواری فرم‌ها می‌پذیرد که از ۰/۸۲ به ۲/۸۶ افزایش می‌یابد. کمترین تأثیر تفاوت دشواری فرم‌ها بر RMSE، مربوط به روش همصدک است که از ۱/۲۵ به ۲/۱۰ افزایش می‌یابد. تأثیر تفاوت دشواری بر RMSE دو روش خطی و میانگین نزدیک به هم است، البته میزان RMSE روش خطی بیشتر از روش میانگین است.

بحث و نتیجه‌گیری

در این پژوهش عملکرد چهار روش همترازسازی خطی، میانگین، همصدک و قوس دایره‌ای در یک الگوی گروه‌های تصادفی با توجه به متغیرهای مستقل شامل حجم نمونه (۵۰، ۱۰۰، ۲۰۰ نفر) و سطوح دشواری (۰,۱، ۰,۴، ۰,۷) موردبررسی قرار گرفت.

نتایج به‌دست‌آمده در پژوهش حاضر با نتایج سایر پژوهش‌ها ازجمله مطالعه Parshall و همکاران (1995)، Asiret and Sünbül (2016)، Babcock و همکاران (2012)، Livingston و همکاران (2010) و Skaggs (2005)، همسو و حاکی از تأثیر افزایش حجم

نمونه در کاهش مقدار خطا در روش‌های هم‌تراز سازی همترازسازی بوده است. چنان چه در پژوهش Skaggs (2005) همان‌گونه که انتظار می‌رفت مقدار خطای استاندارد همترازسازی با افزایش حجم نمونه کاهش یافت؛ هرچند میزان سوگیری همترازسازی به‌عنوان تابعی از حجم نمونه تغییر اندکی نشان داد. علی‌رغم این موضوع در خصوص نمونه با حجم ۲۰۰ نفر نیز در بخش‌های کمی از مقیاس نمرات خام، میزانی از خطای استاندارد مشاهده شد. درمجموع نتایج به‌دست‌آمده حاکی از این بود که در نمرات پایین‌تر از میانگین، روش همترازسازی میانگین، و در نمرات نزدیک میانگین همه روش‌های همترازسازی و در نمرات بالاتر از میانگین روش همترازسازی همصدک بیشترین دقت را داشتند. همچنین نتایج حاکی از کاهش خطای استاندارد روش‌های همترازسازی در نمونه‌های با دامنه ۲۵ تا ۵۰ نفر بوده است. در پژوهش Asiret and Sünbül (2016) نیز نتایج نشان داد که در سطح دشواری ۰,۴ با حجم نمونه ۵۰ و بالاتر روش‌های همترازسازی مقدار خطای مجذور ریشه میانگین کمتری داشتند و درمجموع نیز روش‌های همترازسازی قوس دایره‌ای و میانگین خطای همترازسازی کمتری نسبت به سایر روش‌ها به دست دادند. در مطالعه Babcock و همکاران (2012) نیز نتایج پژوهش نشان داد روش میانگین وزنی اسمی تحت هر شرایط (سطوح دشواری مختلف، سطوح توانایی مختلف، حجم نمونه) کاراترین روش است. روش همترازسازی هویت صرفاً درصورتی‌که دو فرم آزمون تفاوتی در سطوح دشواری نداشتند بیشترین دقت را داشت. در صورت تفاوت در دشواری و همسانی در سطوح توانایی روش‌های همترازسازی میانگین و میانگین وزنی اسمی مقدار خطای کمتری را به دست می دادند. در گروه‌های با سطوح دشواری و سطوح توانایی متفاوت روش‌های همترازسازی میانگین وزنی اسمی و قوس دایره‌ای بهترین عملکرد را داشتند.

در پژوهش پارشال و همکاران (۱۹۹۵) نتایج نشان داد با کاهش حجم نمونه مقدار بسیار کمی از سوگیری نتایج همترازسازی مشاهده می‌شود ولی در خصوص خطای استاندارد همترازسازی کاهش حجم نمونه منجر به افزایش خطای مذکور می‌شود. همچنین در خصوص نمراتی که در اطراف میانگین نمره خام قرار داشتند خطای استاندارد همترازسازی در کمترین حالت بود و درعین‌حال با افزایش فاصله نمرات از میانگین (انحراف بالای نمرات از میانگین) در مقدار خطای استاندارد همترازسازی به‌طور یکنواختی افزایش مشاهده می‌شد.

این یافته‌ها نشان دادند که خطای استاندارد همترازسازی در نمونه‌های کوچک با افزایش فاصله نمرات از میانگین ارتباط داشته و مقدار آن حالت افزایشی دارد.

درمجموع همان‌گونه که نتایج نشان داد هر چهار روش همترازسازی در حجم نمونه کم خطای بالاتری را به‌تناسب حجم نمونه بالا ایجاد کردند. بیشترین تأثیر حجم نمونه در کاهش مقدار خطای سوگیری (EB) به ترتیب در روش همصدک و قوس دایره‌ای به دست آمد. همچنین افزایش تفاوت دشواری منجر به افزایش خطای سوگیری شده و کمترین میزان تأثیر در روش همصدک و بیشترین میزان تأثیر در روش قوس دایره‌ای مشاهده شد.

در خصوص خطای استاندارد همترازسازی (SEE) کمترین خطای استاندارد همترازسازی در هر سه حجم نمونه ۵۰، ۱۰۰ و ۲۰۰ نفر در تفاوت دشواری ۰٫۱ در روش همترازسازی قوس دایره‌ای مشاهده شد. در تفاوت دشواری ۰٫۴ و ۰٫۷ در هر سه حجم نمونه کمترین خطای استاندارد در روش هم‌تراز سازی میانگین بده دست آمد. همچنین روش همصدک در حالت حجم نمونه ۵۰ نفری و تفاوت دشواری ۰٫۷ بیشترین خطای استاندارد را ایجاد کرد. در هر چهار روش همترازسازی با افزایش حجم نمونه خطای استاندارد کمتری مشاهده شد و کمترین خطا در روش همصدک مشاهده شد. در هر چهار روش همترازسازی افزایش تفاوت دشواری منجر به افزایش خطای استاندارد گردید. بیشترین تأثیر در روش قوس دایره‌ای بود. در بررسی شاخص مجذور میانگین مربع خطاها (RMSE) در هر سه حجم نمونه ۵۰، ۱۰۰ و ۲۰۰ نفری در تفاوت دشواری ۰٫۱ کمترین خطا مربوط به روش قوس دایره‌ای بود و در تفاوت دشواری ۰٫۴ در حجم نمونه ۵۰ و ۱۰۰ روش همترازسازی میانگین و در حجم نمونه ۲۰۰ نفری روش همصدک کمترین خطا را نشان دادند. در هر چهار روش همترازسازی افزایش تفاوت دشواری منجر به افزایش خطا شد.

تعارض منافع

نویسندگان هیچ‌گونه تعارض منافعی ندارند.

سپاسگزاری

مقاله حاضر برگرفته از رساله دکتری رشته سنجش و اندازه‌گیری دانشگاه علامه طباطبایی است.

منابع

اصغری، سیما. (۱۳۹٦). *ویژگی‌های روان‌سنجی آزمون تفکر انتقادی کالیفرنیا در مدیران مدارس شهر تهران* (پایان‌نامه کارشناسی ارشد). دانشکده روان‌شناسی و علوم تربیتی. دانشگاه علامه طباطبایی. تهران.

بهمن‌آبادی، سمیه، فلسفی نژاد، محمدرضا، فرخی، نورعلی، و مینایی، اصغر. (۱٤۰۳). نقش تخطی از تک‌بعدی بودن آزمون در خطاهای همترازسازی مدل‌های نظریه سؤال پاسخ و کلاسیک. *اندازه‌گیری تربیتی*، ۱٤(٥٦)، ۷-٤۱. https://doi.org/10.22054/jem.2024.49153.1991

ثرندایک، رابرت (۱۹۸۲). *روان‌سنجی کاربردی*. ترجمه حیدرعلی هومن. (۱۳۷٥). تهران: انتشارات دانشگاه.

خلیلی، حسین، و سلیمانی، محسن. (۱۳۸۲). تعیین اعتماد، اعتبار، و هنجار نمرات آزمون مهارت‌های تفکر انتقادی کالیفرنیا فرم ب (CCTST-B). *دانشگاه علوم پزشکی بابل*، ۲، ۸٤-۹۰.

عسگری، محمد، و ملکی، سامان. (۱۳۸۹). اعتبار، رواسازی و هنجاریابی آزمون مهارت‌های تفکر انتقادی کالیفرنیا برای دانشجویان. اندازه‌گیری تربیتی، ۱٦۹ - ۱٤۷.

مقدم زاده، علی. (۱۳۹٥). روش بهینه همترازسازی با توجه به ویژگی‌های بومی آزمون‌های ملی ایران: موردمطالعه آزمون تولیمو و آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش آموزش کشور. *اندازه‌گیری تربیتی*، ٦ (۲۲)، ۲٦۱-۲۸۷.

مهری نژاد، سید ابوالقاسم. (۱۳۸٦). انطباق و هنجاریابی آزمون مهارت‌های تفکر انتقادی کالیفرنیا. *تازه‌های علوم شناختی*، ۹ (۳)، ۷۲-٦۳.

References

Asghari, S. (1017). *Psychometric characteristics of California's critical thinking test in school principals in Tehran*, Master Dissertation, Allameh Tabatabai University. [In Persian]

Asgari, M., & Maleki, S. (2010). Validation, validation, and standardization of the California Test of Critical Thinking Skills for College Students. *Educational measurement*, 169-147. [In Persian]

Asiret, S., & Sünbül, S. Ö. (2016). Investigating Test Equating Methods in Small Samples through Various Factors. *Educational Sciences: Theory and Practice*, *16*(2), 647-668.

Babcock, B., Albano, A; & Raymond, M. (2012). *Nominal weights mean equating: A method for very small samples*. *Educational and Psychological Measurement*, 72 (4), 608 – 628.

Babcock, B. & Hodge, K. J. (2019) *Rasch Versus Classical Equating in Context of Small Sample Sizes. Educational and Psychological Measurement*. 1-23.

Bahmanabadi S., Falsafinejad, M. R., Farrokhi, N., Minaei, A. (2024). The Role of Test Unidimensionality Violation in Equating Errors of IRT and Classical Theory Models. *Educational Measurement, 14* (56), 7-41. https://doi.org/10.22054/jem.2024.49153.1991 [In Persian]

Caglak, S. (2016). *Comparison of several small sample equating methods under the NEAT design*. *Turkish Journal of Education, 5* (3), 96-118.

Devdass, S. (2011). *Conditions affecting the accuracy of classical equating methods for small samples under the NEAT design: A simulation study (Doctoral dissertation).* University of North Carolina, NC.

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (ETS Research Report NO. RR – 10-29). Princeton, NJ: ETS.

Dorans, N. J., & Holland, P. W. (2000). *Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case*. *Journal of Educational Measurement, 37*(4), 281-306.

Dorans, N. J., & Moses, T. (2023). Score equating: an aspirational form of score linking. *International Encyclopedia if Education (*Fourth Edition). 236-248. DOI:10.1016/B978-0-12-818630-5.10034-x

Heh, V K. (2007). *Equating accuracy using Small Samples in the random group design (Doctoral dissertation)*. Patton College of Education at Ohio University, Athens, OH.

Jaeger, R. M. (1981). *Some Exploratory Indices for Selection of a Test Equating Method*. *Journal of Educational Measurement, 18* (1), 23-38.

Khalili, H., & Soleimani, M. (2003). Determining reliability, validity, and norm scores of the California Critical Thinking Skills Test Form B (CCTST-B). Journal of Babol University of Medical Sciences. 2. 90-84. [In Persian]

Kim, S., Von Davier, A. A., & Haberman, S. (2008). *Small – Sample Equating Using a Synthetic Linking Function*. Journal of Educational Measurement, 45 (4), 325-342.

Kim, S., & Livingston, S. A. (2010). Comparisons among Small Sample Equating Methods in a Common–Item Design. *Journal of Educational Measurement, 47* (3), 286-298.

Kim, S., Livingston, S. A., & Lewis, C. (2011). Collateral Information for Equating in Small Sample: A Preliminary Investigation. *Applied Measurment in Education, 24*, 302-323.

Kim, S., von Davier, A. A., & Haberman, S. (2006). *An alternative to equating with small samples in the non-equivalent groups anchor test design*. Paper presented at the annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Method and practice (3nd ed)*. New York, NY: Spiringer.

Livingston, S. A; Kim, S. (2009b). The circle – arc method for equating in small samples. *Journal of Educational Measurement, 46* (3), 330 – 343.

Mechael, M. (2008). *The Impact of the errors of Equating and errors of measurement on the reported scores*. (Doctoral dissertation). Department of Psychology at Fordhanm University, New York.

Mehrinejad, A. (2007). Adaptation and standardization of the California Test of Critical Thinking Skills. *Cognitive science news, 9* (3), 72-63.[In Persian]

Moghadamzadeh, A. (2013). Optimal Smoothing Method of Data in Test Equating: The Case of TOLIMO and Comprehensive Trial Tests of Iran Educational Testing Organization. [In Persian]

Parshall, C. G., Houghton, P. D., & Kromrey, J. D. (1995*).* Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32* (1), 37 – 54.

Puhan, G., Moses, T. P., Grant, M. C., & McHale, F. (2009). Small – Sample Equating Using a Single – Group Nearly Equivalent Test (SiGNET) Design. *Journal of Educational Measurement, 46* (3), 344 – 362.

Skaggs, G. (2005). Accuracy of Random Groups Equating with Very Small Samples. *Journal of Educational Measurement, 42* (4), 309-330.

Thorndike, R. (1982). *Applied psychometrics*. Translated by Heydar Ali Homan. Tehran University Publications. [In Persian]

VonDavier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of test equating.* New York, NY: Springer-Verlag.

Iriyadi, D; Rahayu,W; & Naga, D. (2018). Equating Method for Small Sample: Comparative research on nominal weight mean and linear method. *Advances in Social Science, Education and Humanities Research*, 295, 178-182.