14(56), Summer 2024, 7-41
eISSN: 2470-6240 ISSN: 2252-004x
Article Type: Reserch Article
Received in revised form September 08, 2020
Accepted September 19, 2020
Received January 13, 2020
Published Online July 05, 2024

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

ATU PRESS

The Role of Test Unidimensionality Violation in Equating Errors of IRT and Classical Theory Models

Somayeh Bahmanabadi — Ph.D. in Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: S.bahmanabady@yahoo.com

Mohammadreza Falsafinejad * — *Corresponding Author*, Associate Professor, Department of Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: falsafinejad@yahoo.co.uk

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

Asghar Minaei — Associate Professor, Department of Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran. E-mail: asghar.minaei@yahoo.com

Abstract

A critical aspect of the field of measurement and assessment is identifying the consequences of any violations of the assumptions inherent in measurement models. Understanding these implications is important for ensuring the validity and reliability of assessment measures being used. The primary objective of the current study was to investigate the impact of test dimensionality violations on equating errors in both item response theory (IRT) and classical theory models. The research design employed in the study was experimental, utilizing a 3 × 3 factorial design to uncover the effects of various conditions. The study's population encompassed all Mathematical and Technical Sciences Entrance Examiners from 2017 and 2018. To ensure a representative sample, 5000 examiners were randomly selected and included in the study group. A Mathematics test consisting of 55 items was utilized for equating purposes. Through this test, three distinct data structures were generated, namely one-dimensional, two-dimensional, and three-dimensional data. Data was equated using a combination of three methods: the equipercentile method, true score equating method, and observed score equating technique. Utilizing metrics such as equating standard errors, bias, and root mean square error, the impact of independent variables was assessed and evaluated. Data analysis revealed that the violation of dimensionality has a significant effect on equating standard error, biasing equating results, and also increases root-mean-square error. To quantify the extent of this effect, the equating results were repeated 100 times in different samples for all three datasets and three equating methods. The findings indicate that violating dimensionality has a similar impact on equating errors, regardless of the equating method being employed. Moreover, there is no appreciable difference in the effect on equating errors between classical theory and IRT models.

Keywords: Violations of unidimensionality; True score equating; Observed score equating; Equipercentile equating; Equating standard error

Cite this Article: 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

Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2024.49153.1991

1. Introduction

Given the significant implications that fateful tests can have on an individual's future career, it is crucial to prioritize the elimination of bias as a top concern in the field of educational assessment and measurement. Ensuring fairness in testing procedures and outcomes is essential for maintaining the integrity and validity of these assessments and the decisions they influence.

It is crucial to emphasize that one of the primary indicators of fairness and impartiality in testing situations is the maintenance of test question secrecy until they are administered, as highlighted by Arıkan & Gelbal (2008) and Kim (2000). To safeguard the confidentiality of the test questions, it is necessary to administer different test forms at different times, ensuring comparability between diverse candidates through these varied forms.

Equating techniques comprise statistical and psychometric approaches that are utilized to standardize scores obtained from various test forms and to make them comparable (Dorans & Holland, 2014; Meng, 2012; Zhang, 2012). Complying with a specific set of assumptions regarding the data collection plan and equating characteristics is critical for appropriately applying equating methods (Kim, 2018). One of the prerequisites for equating is the unidimensionality of the test, but this feature is frequently violated due to factors such as fatigue, test conditions, cheating, guessing, and other aspects. Although the Item Response Theory (IRT) model is resilient to violating the unidimensionality assumption of the test, disregarding this assumption entirely cannot justify the use of unidimensional equating methods. Such violations can introduce bias and errors in the equating results.

2. Literature Review

While research has examined the influence of dimensionality and various equating methods on equating performance (Hirsch, 1989; Cook, Dorans, Eignor, & Petersen, 1985; Camilli, Wang & Fesq, 1995; Ricker, 2007; Brossman, 2010), the results of these studies have been inconclusive and lacking in clarity. Moreover, the nature and focus of these previous studies differ fundamentally from the aims and methods of the current research.

Given that in Iran, the performance of equating multidimensional tests through the application of unidimensional equating methods has

not been thoroughly investigated, and traditional studies have solely employed unidimensional methods, the current study aims to assess the equating errors of different equating methods in both unidimensional and multidimensional test scenarios. This research aims to shed light on the accuracy and validity of equating methods in various test contexts.

3. Methodology

In terms of the research methodology, the study employed an experimental approach with a 3 x 3 factorial design. The research population comprised candidates from the departments of mathematics and technical sciences who participated in the national entrance exams in 2017 (totaling 148,429 individuals) and 2018 (totaling 144,437 individuals). The utilization of this research design allows for a comprehensive evaluation of the equating errors in both multidimensional and unidimensional test scenarios.

The analysis encompassed the responses of 5000 candidates from both years. The mathematics test contained a total of 55 questions. Three datasets were derived for the purpose of the study, each consisting of 20 questions representing a one-factor, two-factor, and three-factor scenario, respectively. SPSS software was utilized for performing Exploratory Factor Analysis (EFA) using the Principal Component method. NOHARM software was then employed to confirm the results obtained from EFA.

For conducting equating using the percentile method, the Equate package in the R software library was utilized. For equating using the true score and observed IRT score methods, the ltm package in R software (Rizopoulos, 2006) was first used to estimate the question and ability parameters. The estimated parameters were subsequently entered into PIE software (Hanson and Zeng, 2004). Microsoft Excel software was then employed to calculate the equating error. The equating performance was assessed through the utilization of three statistics, namely SEE, BIAS, and RMSE. To calculate equating errors, previous studies have indicated that a repetition of 100 repetitions is a reasonable consideration.

4. Results

A two-factor analysis of variance (ANOVA) with a 3 x 3 design was utilized to investigate the significance of the influence of dimensionality and equating methods on the Standard Error of Equating

(SEE). The analysis yielded a significant impact of dimensionality alone ($p < 0.0001$, $F = 11.58$, $\eta^2 = 0.114$). The subsequent Tukey's follow-up test revealed a significant difference between the types of unidimensional and two-dimensional tests ($p < 0.05$) as well as between the unidimensional and three-dimensional test types ($P < 0.0001$) in terms of the standard error of equating.

In terms of the percentile method, the highest level of bias was observed when equating three-dimensional data, while the least bias was present when equating unidimensional data. A two-factor analysis of variance indicated a significant effect of dimensionality ($P < 0.0001$, $F = 20.18$, $\eta^2 = 0.183$). The outcomes of the Tukey's follow-up test indicated that there was a statistically significant divergence between the two types of unidimensional and three-dimensional tests ($P < 0.01$), as well as between the two types of two-dimensional and three-dimensional tests ($P < 0.0001$) in terms of the degree of biasedness.

The magnitude of RMSE in all equating methods was found to be higher in three-dimensional data compared to two-dimensional data, and in two-dimensional data it was higher than in unidimensional data. A two-factor analysis of variance indicated a significant impact of dimensionality on the RMSE ($P < 0.0001$, $F = 15.38$, $\eta^{\wedge 2} = 0.146$).

The findings of the Tukey's follow-up test revealed a statistically significant divergence between the two types of unidimensional and two-dimensional tests ($P < 0.01$), as well as between the two one-dimensional and three-dimensional tests ($P < 0.0001$) and the two-dimensional and three-dimensional tests ($P < 0.01$), indicating that the RMSE values differed significantly across different test dimensions and equating methods.

5. Conclusion

The study demonstrated a significant influence of dimension on SEE, BIAS, and RMSE. However, the interaction between equating method and test dimension was not found to be statistically significant. This implies that the effect of test dimension on equating error metrics remains consistent regardless of the equating method employed. In conclusion, when the assumption of unidimensionality is not met, the performance of equating is negatively impacted, as evidenced by significant differences observed in SEE, BIAS, and RMSE between different test dimensions and equating methods. This indicates that

increasing the number of dimensions in a test causes a corresponding increase in SEE, BIAS, and RMSE.

The current study's findings align with those from previous research by Spence (1996), Patterson (2014), Ricker (2007), and Lim and Lee (2016). Within Item Response Theory, if unidimensionality is violated, the conversion of scores from one test form to another becomes dependent on the population tested, thus disrupting population invariance. As a consequence, equating functions derived from tests measuring different latent traits differ among subgroups of participants. This violation of unidimensionality leads to an increase in equating error.

The violation of the assumption of unidimensionality has far-reaching implications for equating results, primarily due to its disruptive effect on the assumption of local independence. This compromise leads to several issues, including inaccuracies in maximum likelihood estimation, imprecise estimation of Item Response Theory, inaccuracies in estimating the item characteristic curve (ICC), errors in scale conversion, and incorrect placement of parameters on a common scale. As a result, there is a risk of misestimating the test characteristic curve (TCC) and question fit in true score equating, as reported in the research of Ricker (2007), Zhang and Stone (2008).

In the study conducted by Kim, Lim, and Lee (2019), they found that the unidimensional bifactor model within the Item Response Theory framework was found to be a suitable method in situations where the degree of local dependence among test questions was low. This suggests that the bifactor model is better suited for handling local dependence in situations where unidimensionality is violated.

The research indicates that the structure of the test plays a crucial role in the quality and performance of equating. If a test has more than one-dimension, traditional unidimensional equating methods may not effectively handle the data. In cases where unidimensionality is breached, the study suggests exploring alternative equating methods to ensure accurate score comparison and equivalence across different test forms.

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

دوره ۱۴، شماره ۵۶، تابستان ۱۴۰۳، ۷-۴۱
ISSN: 2252-004x
eISSN: 2470-6240

نوع مقاله: مقاله پژوهشی

تاریخ دریافت: ۱۳۹۸/۱۰/۲۳؛ تاریخ بازنگری: ۱۳۹۹/۰۶/۱۸
تاریخ پذیرش: ۱۳۹۹/۰۶/۲۹؛ تاریخ انتشار: ۱۴۰۳/۰۴/۱۵

نقش تخطی از تک‌بعدی بودن آزمون در خطاهای همترازسازی مدل‌های نظریه سؤال پاسخ و کلاسیک

سمیه بهمن‌آبادی دکتری رشته سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: s.bahmanabady@yahoo.com

محمدرضا فلسفی‌نژاد* *نویسنده مسئول*، دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: falsafinejad@yahoo.co.uk

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

اصغر مینایی دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: asghar.minaei@yahoo.com

چکیده

شناسایی عواقب تخطی از مفروضه‌های مدل‌های اندازه‌گیری از دغدغه‌های اصلی در حوزه سنجش و اندازه‌گیری است، هدف پژوهش حاضر، مطالعه نقش نقض مفروضه تک‌بعدی بودن در خطای همترازسازی در نظریه کلاسیک و نظریه سؤال پاسخ بود. روش پژوهش آزمایشی و طرح آن طرح عاملی ۳×۳ بود. جامعه آماری مشتمل بر کلیه داوطلبان گروه ریاضی و فیزیک کنکور سراسری سال‌های ۱۳۹۶ و ۱۳۹۷ بود. با استفاده از نمرات ۵۰۰۰ نفر از داوطلبان در آزمون ریاضی (۵۵ سؤالی) کنکور سراسری، سه مجموعه داده ۲۰ سؤالی با ساختار متفاوت تک‌بعدی، دوبعدی و سه‌بعدی تشکیل شد. هر سه مجموعه داده با استفاده از روش‌های همترازسازی هم‌صدک، همترازسازی نمره واقعی نظریه سؤال پاسخ و روش همترازسازی نمره مشاهده‌شده نظریه سؤال پاسخ همتراز شدند. برای ارزیابی اثرات بعدیت و روش‌های همترازسازی از آماره‌های خطای استاندارد همترازسازی، سوگیری و مجذور میانگین مربع خطا استفاده شد و برای تعیین آماره‌های خطا، نتایج همترازسازی در هر سه مجموعه داده و سه روش همترازسازی در نمونه‌های مختلف ۱۰۰ بار تکرار شد. نتایج تجزیه‌وتحلیل داده‌ها نشان داد که تخطی از بعدیت، خطای استاندارد همترازسازی، سوگیری نتایج همترازسازی و میزان مجذور میانگین مربع خطا را افزایش می‌دهد. تأثیر تخطی از بعدیت در خطاهای همترازسازی بین روش‌های مختلف همترازسازی و در مدل‌های نظریه کلاسیک و نظریه سؤال پاسخ تفاوتی نداشت.

کلیدواژه‌ها: نقض تک‌بعدی، همترازسازی نمره واقعی نظریه سؤال پاسخ، همترازسازی نمره مشاهده‌شده، همترازسازی هم‌صدک، خطای استاندارد همترازسازی

استناد به این مقاله: بهمن‌آبادی، سمیه، فلسفی نژاد، محمدرضا، فرخی، نورعلی، و مینایی، اصغر. (۱۴۰۳). نقش تخطی از تک‌بعدی بودن آزمون در خطاهای همترازسازی مدل‌های نظریه سؤال پاسخ و کلاسیک. *فصلنامه اندازه‌گیری تربیتی*، ۱۴(۵۶)، ۷-۴۱. https://doi.org/10.22054/jem.2024.49153.1991

ناشر: دانشگاه علامه طباطبائی

مقدمه

تصمیم‌سازی بر مبنای آزمون‌های سرنوشت‌ساز، زندگی و آینده شغلی افراد را به شدت تحت تأثیر قرار می‌دهد، ازاین‌رو اطمینان از عدم سوگیری و اجرای منصفانه آن‌ها یکی از اولویت‌های اصلی مطالعات حوزه سنجش و ارزیابی آموزشی است. یکی از مصادیق اصلی رعایت انصاف و عدالت آزمون‌ها، محرمانه بودن آزمون‌ها تا قبل از اجرا و جلوگیری از فاش شدن سؤالات آن است (Gelbal & Arıkan, 2018; Kim, 2000). جهت محرمانه ماندن سؤالات لازم است فرم‌های مختلف آزمون در زمان‌های مختلف اجرا شود و این فرم‌ها باید قابلیت مقایسه داوطلبان مختلف را فراهم نمایند. با وجود این، اغلب غیرممکن است که فرم‌های آزمون دقیقاً از دشواری یکسانی برخوردار باشند و در این صورت ممکن است یک آزمودنی به فرم آسان و آزمودنی دیگر به فرم دشوار پاسخ دهد. تکنیک‌های همترازسازی روش‌هایی آماری و روان‌سنجی هستند که برای تعدیل نمرات به‌دست‌آمده از فرم‌های مختلف آزمون و قابل‌مقایسه کردن آزمون‌هایی با محتوای یکسان مورداستفاده قرار می گیرند (Dorans & Holland, 2000; Meng, 2012; Zhang, 2013; Chen, 2014).

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 (Simon, 2008؛ مقدم‌زاده، ۱۳۹۱) و عدالت و انصاف در ارزیابی امکان‌پذیر می‌شود. مقایسه‌پذیری و همترازسازی آزمون‌ها از چنان اهمیتی برخوردار است که استانداردهای انجمن روان‌شناسی آمریکا[1]، انجمن تحقیقات آموزشی آمریکا[2] و شورای ملی اندازه‌گیری در آموزش‌وپرورش[3] (۱۹۹۹) تأکید کرده‌اند که «هنگامی که نمرات از فرم‌های مختلف آزمون به دست می‌آید و ادعا می‌شود که می‌توان این فرم‌ها را به‌جای هم استفاده کرد، باید منطق و شواهد حمایتی روشن برای این ادعا فراهم شود» (Kolen & Brennan, 2004).

به‌کارگیری روش‌های همترازسازی مستلزم رعایت مجموعه‌ای از مفروضات در خصوص طرح گردآوری داده‌ها و ویژگی‌های همترازسازی است. بدیهی است شکست در رعایت این مفروضات، دقت همترازسازی را با مشکل مواجه می‌کند (Kim, 2018). یکی

1. American Psychological Association (APA)
2. American Educational Research Association (AERA)
3. National Council on Measurement in Education (NCME)

از این پیش‌شرط‌ها، رعایت مفروضه تک‌بعدی بودن آزمون است، اما این ویژگی همواره تحت تأثیر عوامل مختلفی همچون خستگی، شرایط برگزاری آزمون، تقلب، حدس و سایر عوامل نقض می‌شود (Kim et al., 2018; Li et al., 2012). در حوزه روان‌شناسی و تعلیم و تربیت نیز، برخی از آزمون‌ها ماهیتاً چندبعدی محسوب شده و نمی‌توان آن‌ها را به‌عنوان آزمون‌های تک‌بعدی در نظر گرفت (Kim et al., 2018). برای مثال یک آزمون ریاضی درعین‌حال که توانایی محاسبه را موردسنجش قرار می‌دهد، به سنجش توان جبر، هندسه، مثلثات و حتی توانایی خواندن نیز می‌پردازد. همچنین در یک آزمون زبانِ انگلیسی، علاوه بر سنجش توانایی گرامر یا ساختار فرد، توانایی واژگان، درک مطلب و سایر مهارت‌های زبانی نیز سنجش می‌شود، ازاین‌رو، نمی‌توان در خصوص تک‌بعدی بودن این آزمون‌ها با قطعیت قضاوت نمود.

به نظر می‌رسد مدل‌های نظریه سؤال پاسخ تا حدودی نسبت به نقض تک‌بعدی بودن آزمون مقاوم هستند اما چنانچه این مفروضه به‌طور کامل نقض گردد، استفاده از روش‌های تک‌بعدی همترازسازی قابل توجیه نبوده و منجر به سوگیری و خطا در نتایج می‌شود (Brossman & Lee, 2013). نقض مفروضه تک‌بعدی اثرات منفی بر بسیاری از کاربردهای نظریه سؤال پاسخ همچون برازش سؤال، دقت برآورد سؤال و توانایی، آگاهی و پایایی آزمون دارد (Chen, 2014) و روایی کاربرد مدل‌های تک‌بعدی در همترازسازی، بررسی کنش افتراقی، نمره‌گذاری و سنجش انطباقی در موقعیتی که فضای چندبعدی معقول و منطقی است، موردتردید واقع شده است (Seo & Weiss, 2015). با وجود این، برخی پژوهشگران به این نتیجه رسیده‌اند که حتی زمانی که آزمون چندبعدی است، روش IRT تک‌بعدی منجر به‌دقت در نتایج همترازسازی می‌شود (Lee et al., 2014؛ Lee, 2013).

Kim (2018) دقت بیشتر همترازسازی نمره واقعی چندبعدی با ساختار ساده را در مقایسه با روش‌های معمول تک‌بعدی برای آزمون‌های چندبعدی نشان داده است. در مطالعه Ricker (2007) همترازسازی هنگامی‌که فقط سؤالات مشترک بعد دوم را اندازه می‌گرفتند، در برابر نقض تک‌بعدی بودن مقاومت نشان داد. Spence (1996) نشان داد همچنان که تعداد سؤالات چندبعدی در آزمون بالا می‌رود، اثر چندبعدیتی بر همترازسازی بیشتر می‌شود. در پژوهش Peterson (2014) و نیز مدل‌های چندبعدی برای داده‌های چندبعدی بهتر عمل کردند درحالی‌که مدل‌های تک‌بعدی برای داده‌های تک‌بعدی عملکرد بهتری داشتند و

Andrews (2011) نشان داد که با افزایش تعداد ابعاد آزمون، روش همترازسازی نمره مشاهده‌شده نظریه سؤال پاسخ چندبعدی نسبت به روش‌های تک‌بعدی عملکرد بهتری دارد. نتایج برخی پژوهش‌ها شواهدی از تأثیرپذیری همترازسازی نمره واقعی و نمره مشاهده‌شده از مقدار بار عاملی و تغییرپذیری عامل‌های مشترک و اختصاصی را نشان داد (Lim & lee, 2016)، اما در مطالعه Fesq al و همکاران (1995) پس از تأیید چندبعدی بودن داده‌ها از طریق تحلیل عاملی مرتبه اول و دوم، نقض مفروضه تک‌بعدی اثر چندانی بر همترازسازی نمره واقعی نداشت.

در زمینه مقاومت مدل‌های IRT در مقابل نقض مفروضه تک‌بعدی بودن، Dorans and Kingston (1985) به این نتیجه رسیدند که در مدل IRT سه پارامتری لوجستیک، روش همترازسازی نمره واقعی مقاومت بیشتری نسبت به نقض تک‌بعدی بودن دارد (به نقل از Fesq et al., 1995). علاوه بر این، سطح وابستگی موضعی سؤالات نیز بر همترازسازی نمره واقعی و نمره مشاهده‌شده نظریه سؤال پاسخ اثر معنی‌دار نشان داد (Chen, 2014). در بررسی تفاوت و شباهت‌های همترازسازی نمره مشاهده‌شده و نمره واقعی نظریه سؤال پاسخ با همترازسازی همصدک در آزمون‌های ریاضی و زبان پذیرش دانشگاه‌های آمریکا، نتایج حاکی از آن بود که همترازسازی نمره واقعی و نمره مشاهده‌شده نظریه سؤال پاسخ از نتایج پایاتری نسبت به همترازسازی همصدک برخوردار است (Han et al., 1997).

علی‌رغم انجام پژوهش‌هایی در بررسی اثر تخطی از بعدیت و اثر روش‌های مختلف همترازسازی نظریه سؤال پاسخ و نظریه کلاسیک در خصوص عملکرد همترازسازی (Fesq et al., 1995; Brossman, 2010 ؛Hirsch, 1989; Cook et al., 1985)، نکته حائز توجه این است که چنین پژوهش‌هایی نتایج روشن و مشخصی را نشان نداده‌اند، همچنین ماهیت این پژوهش‌ها با پژوهش حاضر متفاوت است، چراکه به‌طور هم‌زمان تفاوت داده‌های مختلف با ساختار تک‌بعدی و چندبعدی را مدنظر قرار نداده‌اند. با توجه به عدم همسویی نتایج پژوهش‌ها در سطح جهان و با توجه به اینکه در کشور ایران، عملکرد همترازسازی آزمون‌های چندبعدی در شرایط استفاده از روش‌های تک‌بعدی مشخص نشده است و تاکنون در تمام پژوهش‌های انجام‌شده صرفاً از روش‌های تک‌بعدی استفاده شده است (مقدم زاده، ۱۳۹۲؛ رضوانی فر، ۱۳۹۱؛ شاطریان محمدی، ۱۳۸۱؛ واشقانی فراهانی، ۱۳۸۰)، پژوهش حاضر به بررسی و مقایسه خطاهای همترازسازی روش‌های مختلف همترازسازی در

آزمون‌های تک‌بعدی و چندبعدی می‌پردازد. به‌عبارت‌دیگر، پژوهش حاضر به دنبال این است که بررسی کند که نقض تک‌بعدی بودن تا چه حد قابل‌چشم‌پوشی است، وجه تفاوت این پژوهش استفاده از سه نوع داده مختلف بر اساس بعدیت و استفاده از داده‌های واقعی به‌جای داده‌های شبیه‌سازی است.

روش

روش پژوهش آزمایشی و طرح پژوهش، طرح عاملی ۳ × ۳ بود که در آن اثر دو متغیر مستقل بعدیت آزمون (با سه سطح تک‌بعدی، دوبعدی و سه‌بعدی) و روش همترازسازی (با سه سطح روش همترازسازی نمره واقعی IRT، همترازسازی نمره مشاهده‌شده IRT و روش همصدک) بر انواع خطای همترازسازی بررسی شد. کلیه داوطلبان گروه علوم ریاضی و فنی کنکور سراسری سال‌های ۱۳۹۶ (برابر با ۱۴۸۴۲۹ نفر) و ۱۳۹۷ (برابر با ۱۴۴۴۳۷ نفر) جامعه پژوهش را تشکیل دادند و داده‌های پاسخ ۵۰۰۰ نفر از داوطلبان هر دو سال به آزمون ریاضی کنکور سراسری موردمطالعه قرار گرفت. آزمون ریاضی دارای ۵۵ سؤال ۴ گزینه‌ای است که از آن سه مجموعه داده ۲۰ سؤالی با توجه به هدف پژوهش استخراج گردید. برای استخراج سه مجموعه داده، ابتدا برای کل سؤالات آزمون، تحلیل عاملی اکتشافی گرفته شد و ۲۰ سؤال که یک بعد غالب را بر اساس ملاک‌های مختلف همچون مقدار ویژه و درصد واریانس تبیینی موردسنجش قرار می‌دادند، به‌عنوان آزمون تک‌بعدی در نظر گرفته شد، در مرحله بعد ۲۰ سؤال به‌گونه‌ای انتخاب شد که ۲ عامل غالب را موردسنجش قرار دهد و در مرحله سوم ۲۰ سؤال به‌نحوی‌که سه عامل غالب در داده‌ها موردسنجش قرار گیرد، انتخاب شد. جهت اطمینان بیشتر از تعیین ابعاد داده‌ها از نرم‌افزار NOHRAM (Fraser & McDonald, 2012) به دلیل مناسب بودن این نرم‌افزار در تعیین ابعاد داده‌های چندبعدی و مفید بودن آن برای تحلیل داده‌های دو ارزشی استفاده شد. لازم به ذکر است که در آزمون ریاضی سال ۱۳۹۷ تعداد ۵ سؤال و در سال ۱۳۹۶ تعداد ۶ سؤال در هر سه مجموعه داده مشترک بود. سه مجموعه داده انتخاب‌شده (تک‌بعدی، دوبعدی و سه‌بعدی) در آزمون ریاضی کنکور سراسری دو سال ۹۶ و ۹۷ با استفاده از طرح گروه‌های معادل[1] و در سه روش همترازسازی همصدک، همترازسازی نمره مشاهده‌شده و همترازسازی نمره واقعی نظریه سؤال پاسخ همتراز شدند. در طرح گروه‌های معادل، دو گروه از ارزیابی‌شوندگان مستقل از

1. equivalence group

جامعه نمونه‌گیری می‌شوند و هر گروه یک فرم آزمون را می‌گیرد، در این پژوهش منظور از دو گروه از مشارکت‌کنندگان، داوطلبانی است که در سال‌های ۱۳۹۶ و ۱۳۹۷ فرم‌های مختلف آزمون ریاضی را پاسخ داده‌اند (González & Wiberg, 2017: 11). در طرح گروه‌های معادل با مدل‌های تک‌بعدی نظریه سؤال پاسخ (UIRT) نیازی به فرایند هموار کردن مقیاس و در مقیاس مشترک قرار دادن نمرات نیست، زیرا دو گروه از جمعیت یکسانی استخراج شده‌اند و در این صورت نیاز به تنظیم عرض از مبدأ و واحد مقیاس از بین می‌رود (Kim, 2018).

جهت تعیین ابعاد آزمون ریاضی از تحلیل عاملی اکتشافی به روش مؤلفه‌های اصلی در نرم‌افزار SPSS استفاده شد و با استفاده از ملاک‌های مقدار ویژه (بزرگ‌تر از یک) و مقدار تبیین واریانس کل آزمون (بیشتر از ۵ درصد واریانس) در خصوص تعداد ابعاد آزمون تصمیم گرفته شد و نتایج با نرم‌افزار NOHARM موردبررسی و تأیید قرار گرفت. همترازسازی داده‌ها به روش همصدک با استفاده از بسته equate در نرم‌افزار R انجام شد و جهت همترازسازی داده‌ها در دو روش نمره واقعی و نمره مشاهده‌شده IRT، ابتدا پارامترهای سؤال و توانایی با استفاده از بسته ltm (Rizopouos, 2006) در نرم‌افزار R برآورد شده و سپس این پارامترها در نرم‌افزار PIE (Hanson & Zeng, 2004) برای انجام همترازسازی وارد شدند. برآورد آماره‌های خطا با استفاده از فرمول‌های خطای همترازسازی در نرم‌افزار excel فرمول‌نویسی شد.

جهت ارزیابی دقت همترازسازی و تعیین میزان خطای آن، چندین آماره مدنظر پژوهشگران قرار گرفته است، خطاهای ایجادشده در فرایند همترازسازی شامل خطاهای تصادفی و خطاهای منظم هستند (Kolen & Brennan, 2004; Zhang, 2012)، ازاین‌رو، ضروری است که ملاک‌هایی برای ارزیابی انتخاب شوند که بتوانند روش‌های مختلف همترازسازی را در این دو نوع خطا ارزیابی کنند، مهم‌ترین آن‌ها آماره‌های خطای استاندارد همترازسازی (SEE[1]) برای ارزیابی خطای تصادفی، سوگیری (BIAS) همترازسازی برای ارزیابی خطای منظم و مجذور میانگین مربع خطای همترازسازی (RSME[2]) برای اندازه‌گیری دقت کلی همترازسازی است (Zhang, 2012؛ Lee, 2013؛ Kolen & Brennan, 2004؛ Kim, 2018). خطای استاندارد همترازسازی به‌عنوان انحراف استاندارد

1. Equating Standard Er
2. Root mean square error statistic

نمرات همترازشده در طی تکرارهای فرضی یک فرایند همترازسازی در نمونه‌های یک جامعه یا جامعه‌های آزمودنی تعریف می‌شود. در یک تکرار فرضی، تعداد مشخصی از آزمودنی‌ها به‌طور تصادفی از جامعه انتخاب می‌شوند، سپس نمرات همتراز شده فرم Y با فرم X در سطوح مختلف نمرات با استفاده از یک روش همترازسازی همتراز می‌شوند. خطای استاندارد همترازسازی در هر سطح، انحراف استاندارد نمرات همترازساز شده فرم Y و X در تکرارهای مختلف است (Kolen & Brennan, 2004)، سوگیری به معنای تفاوت مطلق میانگین نمرات همترازشده در تکرارهای مختلف نتایج همترازسازی در نمونه‌های مختلف از نمره همترازسازی به‌دست‌آمده از جامعه است و مجذور میانگین مربع خطای همترازسازی نیز ریشه دوم مجموع مجذور سوگیری و خطای استاندارد همترازسازی است.

در پژوهش حاضر، جهت ارزیابی عملکرد همترازسازی و تعیین خطای آن از سه آماره SEE، BIAS و RMSE استفاده شد. استفاده از آماره‌های خطا مستلزم تکرار نتایج همترازسازی در نمونه‌های مختلف است، ازاین‌رو باید تعداد تکرارها به‌طور منطقی انتخاب گردد تا سوگیری و خطای نمونه‌گیری کاهش یابد، در پژوهش حاضر جهت تعیین تعداد خطا از مطالعات قبلی و پیشینه پژوهشی استفاده شد. در پیشینه مطالعات همترازسازی از ۱۰ تا ۵۰۰ تکرار مدنظر قرار گرفته است. در پژوهش ژانگ (۲۰۱۲) با توجه به میانگین تعداد تکرارها در پژوهش‌های قبلی ۲۰ بار تکرار برای شبیه‌سازی در نظر گرفته شد. در پژوهش Lee (2013) تعداد تکرارها ۵۰ و در پژوهش‌های Lu & Guo (2018)، Kim (2018)، Lim (2016) و Shin (2015) تعداد ۱۰۰ بار تکرار جهت تولید داده‌های چندبعدی در نظر گرفته شد. در پژوهش حاضر نیز تعداد ۱۰۰ بار تکرار بر اساس مطالعات قبلی جهت تعیین خطاهای همترازسازی در نظر گرفته شده است .جهت قابل‌مقایسه بودن نتایج همترازسازی علاوه بر استفاده از آماره‌های استانداردشده، سعی شد فرایند تولید داده‌ها، فرایند نمونه‌گیری از داده‌ها، تعداد تکرارها، روش برآورد پارامترها، قرار دادن برآوردها در PIE جهت انجام همترازسازی و استفاده از بسته‌های آماری در همه روش‌ها و در سه مجموعه داده یکسان باشد.

یافته‌ها

میانگین نمرات درس ریاضی در دو گروه داوطلبان سال ۱۳۹۶ و سال ۱۳۹۷ در جدول زیر آمده است.

جدول ۱. شاخص‌های توصیفی نمرات آزمودنی‌ها در دو فرم آزمون ریاضی

سؤالات	داوطلبان	میانگین	انحراف استاندارد	کمینه	بیشینه	کجی	کشیدگی
آزمون تک‌بعدی	سال ۱۳۹٦	۳/۳۹	۳/٦٤	0	۲۰	۱/۵۱	۱/۹٤
	سال ۱۳۹۷	۳/۰۲	۳/٤٤	0	۱۹	۱/۵۱	۱/۹۸
آزمون دوبعدی	سال ۱۳۹٦	۲/۰۱	۲/۵٤	0	۱۹	۲/۰٦	۵/۳۷
	سال ۱۳۹۷	۲/۰۳	۲/۸٦	0	۱۹	۲/۱۲	۵/۰۹
آزمون سه‌بعدی	سال ۱۳۹٦	۲/۱۱	۲/۵۰	0	۱۷	۱/۹۳	٤/۵۸
	سال ۱۳۹۷	۲/۲۰	۲/۸۸	0	۲۰	۲/۲۲	٦/۰٤

همان‌طور که در جدول فوق مشاهده می‌شود در آزمون تک‌بعدی، در سال ۱۳۹۶، میانگین و انحراف استاندارد نمرات به ترتیب برابر با ۳/۳۹ و ۳/۶۴ و در سال ۱۳۹۷ برابر با ۳/۰۲ و ۳/۴۴ است. در آزمون دوبعدی نیز میانگین و انحراف استاندارد نمرات در سال ۱۳۹۶ به ترتیب برابر با ۲/۰۱ و ۲/۵۴ و در سال ۱۳۹۷ برابر با ۲/۰۳ و ۲/۸۶ بود و در آزمون سه‌بعدی میانگین و انحراف استاندارد نمرات در سال ۱۳۹۶ به ترتیب برابر با ۲/۱۱ و ۲/۵۰ و در سال ۱۳۹۷ برابر با ۲/۲۰ و ۲/۸۸ بود.

ابعاد آزمون و داده‌های موردمطالعه. جهت تعیین ابعاد سه مجموعه داده موردمطالعه از روش تحلیل عاملی اکتشافی به روش مؤلفه‌های اصلی در نرم‌افزار SPSS استفاده شده است، همچنین از نرم‌افزار NOHARM جهت تعیین بارهای عاملی و تأیید بیشتر ابعاد تعیین‌شده استفاده شد، نتایج مربوط به تعیین بعد در مجموعه اول داده‌ها در دو سال ۱۳۹۶ و ۱۳۹۷ در زیر آمده است.

جدول ۲. تعداد ابعاد زیربنایی داده‌ها و مقدار واریانس تبیین شده توسط هر بعد در داده‌های تک‌بعدی

عامل‌ها	سال ۱۳۹٦			سال ۱۳۹۷		
	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی
۱	۵/۳۵	۲٦/۷۹	۲٦/۷۹	۵/۰۲	۲۵/۱۰	۲۵/۱۰
۲	۱/۳۷	٦/۸٤	۳۳/٦٤	۱/۲۵	٦/۲۷	۳۱/۳۷
۳	–	–	–	۱/۰۲	۵/۱۳	۳٦/۵۱

همان‌طور که در جدول فوق مشاهده می‌شود داده‌ها و سؤالات آزمون در مجموعه اول داده‌ها به نحوی انتخاب شده است که یک عامل غالب (که حداقل دو برابر عامل‌های دیگر واریانس کل را تبیین نماید و دارای مقدار ویژه بزرگ‌تر از ۱ باشد) بر داده‌ها حاکم است.

نتایج مربوط به نرم‌افزار NOHARM نیز تک‌بعدی بودن این داده‌ها را تأیید نمود. در نرم‌افزار NOHARM مقدار مجموع مجذورات باقی‌مانده‌ها در آزمون سال ۱۳۹۷ برابر با ۰/۰۰۳ و ریشه دوم میانگین مجذورات باقی‌مانده‌ها[1] (RMSR) برابر با ۰/۰۰۲ است و در سال ۱۳۹۶ این مقادیر به ترتیب برابر با ۰/۰۰۲ و ۰/۰۰۳ است که هر چه مقدار آن کوچک‌تر باشد، نشان‌دهنده برازش بهتر مدل است، همچنین می‌توان ملاک RMSR را با چهار برابر ریشه دوم معکوس حجم نمونه مقایسه نمود که در این پژوهش مقدار RMSR از ملاک ذکرشده کمتر است، شاخص نیکویی برازش تاناکا[2] در داده‌های سال ۹۷ برای مدل تک‌بعدی برابر با ۰/۹۹۳ و در سال ۱۳۹۶ برابر با ۰/۹۹۴ و بیشتر از ۰/۹۵ است که نشان‌دهنده برازش مناسب مدل تک‌بعدی با داده‌ها است (Kline, 2011). بارهای عاملی در مدل تک‌بعدی در داده‌های سال ۱۳۹۶ و ۱۳۹۷ در جدول زیر آمده است.

جدول ۳. بارهای عاملی مربوط به بعد کنکور ریاضی در دو سال ۱۳۹۶ و ۱۳۹۷ در داده‌های تک‌بعدی

سال ۱۳۹۶		سال ۱۳۹۷	
سؤال	عامل اول	سؤال	عامل اول
۱	۰/۶۸	۱	۰/۲۴
۲	۰/۶۹	۲	۰/۴۳
۳	۰/۷۲	۳	۰/۴۹
۴	۰/۵۵	۴	۰/۶۲
۵	۰/۷۶	۵	۰/۷۰
۶	۰/۸۴	۶	۰/۵۶
۷	۰/۴۱	۷	۰/۶۰
۸	۰/۷۳	۸	۰/۶۸
۹	۰/۷۱	۹	۰/۲۵
۱۰	۰/۷۷	۱۰	۰/۷۴
۱۱	۰/۶۸	۱۱	۰/۷۶
۱۲	۰/۸۸	۱۲	۰/۸۰

1. Root mean square of residuals
2. Tanaka index of goodness of fit

سال ۱۳۹۶		سال ۱۳۹۷	
سؤال	عامل اول	سؤال	عامل اول
۱۳	۰/۸۳	۱۳	۰/۷۱
۱٤	۰/۸۲	۱٤	۰/۷۵
۱۵	۰/۵۹	۱۵	۰/٦۷
۱٦	۰/۳۷	۱٦	۰/۷۳
۱۷	۰/۳۲	۱۷	۰/٦٦
۱۸	۰/٤٤	۱۸	۰/۸۰
۱۹	۰/۳۰	۱۹	۰/۸۱
۲۰	۰/۲۹	۲۰	۰/٦٦

همان‌طور که در جدول فوق مشاهده می‌شود کلیه سؤالات آزمون به‌جز سؤال ۲۰ در آزمون سال ۱۳۹۶ و سؤال ۹ در آزمون سال ۱۳۹۷ دارای بار عاملی بزرگ‌تر از ۰/۳ در عامل اول هستند. در جدول (۴) تعداد ابعاد زیربنایی داده‌ها و مقدار واریانس در داده‌های دوبعدی ارائه شده است.

جدول ٤. تعداد ابعاد زیربنایی داده‌ها و مقدار واریانس تبیین شده توسط هر بعد در داده‌های دوبعدی

عامل‌ها	سال ۱۳۹٦			سال ۱۳۹۷		
	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی
۱	۲/۸٦	۱٤/۳٤	۱٤/۳٤	۳/۹۹	۱۹/۹۵	۱۹/۹۵
۲	۲/٤٦	۱۲/۳۱	۲٦/٦٦	۲/۲۷	۱۱/۳۸	۳۱/۳٤

در داده‌های مجموعه دوم دو بعد اصلی داده‌ها را پوشش می‌دهد که هر دو بعد به میزان تقریباً یکسانی میزان واریانس کل آزمون را تبیین می‌کنند. مقادیر ویژه هر دو بعد بزرگ‌تر از ۱ است. نتایج مربوط به نرم‌افزار NOHARM نیز دوبعدی بودن این داده‌ها را تأیید نمود. در نرم‌افزار NOHARM مقدار مجموع مجذورات باقی‌مانده‌ها در آزمون سال ۱۳۹۷ برابر با ۰/۰۰۰۸ و ریشه دوم میانگین مجذورات باقی‌مانده‌ها (RMSR) برابر با ۰/۰۰۲ است و در سال ۱۳۹۶ این مقادیر به ترتیب برابر با ۰/۰۰۰۴ و ۰/۰۰۱ است که هر چه مقدار آن کوچک‌تر باشد، نشان‌دهنده برازش بهتر مدل است، همچنین شاخص نیکویی برازش تاناکا در داده‌های سال ۹۷ برای مدل تک‌بعدی برابر با ۰/۹۹۵ و در سال ۱۳۹۶ برابر با ۰/۹۹۷ و بیشتر از ۰/۹۵

است که نشان‌دهنده برازش مناسب مدل دوبعدی با داده‌ها است. بارهای عاملی در مدل دوبعدی بعد از چرخش واریماکس در جدول زیر آمده است.

جدول ۵. بارهای عاملی مربوط به بعد کنکور ریاضی در دو سال ۱۳۹۶ و ۱۳۹۷ در داده‌های دوبعدی

سال ۱۳۹۶			سال ۱۳۹۷		
سؤال	عامل اول	عامل دوم	سؤال	عامل اول	عامل دوم
۱	۰/۵۶	۰/۲۷	۱	۰/۶۳	−۰/۰۳
۲	۰/۵۰	۰/۵۳	۲	۰/۴۴	۰/۲۶
۳	۰/۵۸	۰/۳۸	۳	۰/۴۷	۰/۳۲
٤	۰/۶۲	۰/۲٤	٤	۰/٤٤	۰/٤۸
۵	۰/۵۳	۰/۱۵	۵	۰/٤٦	۰/۵٦
٦	۰/٦۰	۰/۰۳	٦	۰/٤٤	۰/٤۲
۷	۰/۵٤	۰/٤۸	۷	۰/۵٦	۰/۳۹
۸	۰/۵۲	۰/۳٤	۸	۰/۵۲	۰/٤۹
۹	۰/۵۰	۰/۲۲	۹	۰/۵۵	۰/۰۲٦
۱۰	۰/۵٦	۰/۲۹	۱۰	۰/۳٤	۰/٦٦
۱۱	۰/۱۲	۰/۸۳	۱۱	۰/۲۵	۰/۷۲
۱۲	۰/۱۱	۰/۷۷	۱۲	۰/٤۲	۰/٦۸
۱۳	۰/۳۸	۰/۵۹	۱۳	۰/٤۷	۰/۵۵
۱٤	۰/۲۳	۰/٦۹	۱٤	۰/٤۹	۰/٦۰
۱۵	۰/٦۱	۰/۳۳	۱۵	۰/۲۳	۰/٦۳
۱٦	۰/۳۳	۰/٤۷	۱٦	۰/۳۹	۰/٦۳
۱۷	۰/۲۳	۰/۳۳	۱۷	−۰/۰۳	۰/۷۵
۱۸	۰/۵٤	۰/۱۵	۱۸	۰/۱۵	۰/۸۱
۱۹	۰/٤٦	۰/۱۹	۱۹	۰/۱۷	۰/۸۲
۲۰	۰/۳٦	۰/٤۹	۲۰	۰/۵۷	۰/٤٦

مقادیر بار عاملی در جدول شماره (۵) برای داده‌های دوبعدی نشان می‌دهد که در داده‌های آزمون سال ۱۳۹۶ تعداد ۱۳ سؤال دارای بار عاملی بالا در عامل اول و تعداد ۹ سؤال دارای بار عاملی بالا در عامل دوم هستند، در آزمون سال ۱۳۹۷ نیز تعداد ۷ سؤال در عامل

اول و تعداد ۱۳ سؤال در عامل دوم دارای بار عاملی مناسب هستند. در جدول (۶) تعداد ابعاد و میزان واریانس تبیینی هر بعد در داده‌های سه‌بعدی ارائه شده است.

جدول ٦. تعداد ابعاد زیربنایی داده‌ها و مقدار واریانس تبیین شده توسط هر بعد در داده‌های سه‌بعدی

عامل‌ها	سال ۱۳۹٦			سال ۱۳۹۷		
	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی	مقدار ویژه	درصد تبیین واریانس	واریانس تجمعی
۱	۲/۵۳	۱۲/٦۹	۱۲/۹٦	۲/۷۲	۱۳/٦۲	۱۳/٦۲
۲	۱/۹۹	۹/۹۸	۲۲/٦۸	۲/۳۹	۱۱/۹۸	۲۵/٦۰
۳	۱/۸٤	۹/۱۹	۳۱/۸۷	۲/۱۸	۱۰/۹۰	۳٦/۵۰

در داده‌های مجموعه سوم، سه بعد اصلی داده‌ها را پوشش می‌دهد که هر سه بعد به میزان تقریباً یکسانی میزان واریانس کل آزمون را تبیین می‌کنند و دارای بار عاملی بزرگ‌تر از ۱ هستند. نتایج مربوط به تعیین بعدیت در نرم‌افزار NOHARM نیز سه‌بعدی بودن این داده‌ها را تأیید نمود. در نرم‌افزار NOHARM مقدار مجموع مجذورات باقی‌مانده‌ها در آزمون سال ۱۳۹۷ برابر با ۰/۰۰۰۵ و ریشه دوم میانگین مجذورات باقی‌مانده‌ها (RMSR) برابر با ۰/۰۰۱ است و در سال ۱۳۹۶ این مقادیر به ترتیب برابر با ۰/۰۰۰۴ و ۰/۰۰۱ است که هر چه مقدار آن کوچک‌تر باشد، نشان‌دهنده برازش بهتر مدل است، همچنین شاخص نیکویی برازش تاناکا[1] در داده‌های سال ۹۷ برای مدل تک‌بعدی برابر با ۰/۹۹٦ و در سال ۱۳۹۶ برابر با ۰/۹۹۸ و بیشتر از ۰/۹۵ است که نشان‌دهنده برازش مناسب مدل دوبعدی با داده‌ها است. مقادیر بار عاملی در هر سه بعد، در جدول زیر آمده است.

جدول ۷. بارهای عاملی مربوط به بعد کنکور ریاضی در دو سال ۱۳۹٦ و ۱۳۹۷ در داده‌های سه‌بعدی

	سال ۱۳۹٦				سال ۱۳۹۷		
سؤال	عامل اول	عامل دوم	عامل سوم	سؤال	عامل اول	عامل دوم	عامل سوم
۱	۰/٤۸	۰/۲۹	۰/۱٤	۱	۰/٤۷	۰/۲۱	۰/۱۸
۲	۰/٦٦	۰/۲٤	۰/۲۰	۲	۰/۵۸	۰/٤۵	۰/۲۷
۳	۰/۷۱	۰/۰۵	۰/۲۷	۳	۰/۵۰	۰/۳۳	۰/۳۰
٤	۰/۸۳	۰/۱۵	۰/۲۸	٤	۰/٦۰	۰/۱۰	۰/۲۵

1. Tanaka index of goodness of fit

سال ۱۳۹۶				سال ۱۳۹۷			
سؤال	عامل اول	عامل دوم	عامل سوم	سؤال	عامل اول	عامل دوم	عامل سوم
۵	**۰/۷۲**	۰/۲۳	۰/۳۳	۵	**۰/۵۳**	۰/۰۹	۰/۲۱
٦	**۰/۷۸**	۰/۰۸	۰/۳۰	٦	**۰/۵۹**	۰/۳٤	۰/۱۹
۷	۰/۳۱	**۰/٦۷**	۰/۱۵	۷	**۰/٤۹**	۰/۳٤	۰/۳٤
۸	-۰/۰٦	**۰/۵۳**	۰/۰۰۹	۸	۰/۲۳	**۰/٦٤**	۰/۱۰
۹	۰/۳۵	**۰/۵٤**	۰/۲۸	۹	۰/۱۲	**۰/۷۳**	۰/۰٦
۱۰	۰/۰۲	**۰/۵۲**	۰/۰۲	۱۰	۰/۳۷	**۰/٦۹**	۰/۲۹
۱۱	۰/۱۵	**۰/۵۵**	۰/۱۸	۱۱	۰/٤۵	**۰/٦۰**	۰/۲۷
۱۲	۰/۳۳	۰/۰۸	**۰/۷٦**	۱۲	۰/۲۳	**۰/۵۲**	۰/۳۵
۱۳	۰/۲۰	۰/۲۳	**۰/۷۳**	۱۳	۰/۰۸	**۰/٦۰**	۰/۱۸
۱٤	۰/٤۰	۰/۲۹	**۰/٤۸**	۱٤	۰/۴۷	۰/۲٦	**۰/۵۲**
۱۵	۰/۳۳	۰/۱٦	**۰/٦۳**	۱۵	۰/٤۱	۰/۱۸	**۰/٦۱**
۱٦	۰/۲۷	**۰/۵٤**	۰/۲۰	۱٦	**۰/٤۸**	۰/۱۹	۰/٤۲
۱۷	۰/۳٤	**۰/۵٤**	۰/۲۵	۱۷	۰/۳۷	۰/۱۷	**۰/۵۲**
۱۸	۰/۱۷	**۰/۳۵**	۰/۲۵	۱۸	۰/۱٦	۰/۰۹	**۰/۷۳**
۱۹	۰/۳۷	**۰/٤٤**	۰/۳۲	۱۹	۰/۲۷	۰/۳۰	**۰/۷٤**
۲۰	۰/۳۳	۰/۲٦	**۰/٤۲**	۲۰	۰/۳۳	۰/۲۵	**۰/۷۳**

مقادیر بارهای عاملی در مجموعه داده‌های سه‌بعدی نشان می‌دهد که در آزمون سه‌بعدی سال ۱۳۹۶ تعداد ۶ سؤال در عامل اول، تعداد ۹ سؤال در عامل دوم و تعداد ۵ سؤال در عامل سوم دارای بار عاملی مناسبی هستند، همچنین در آزمون سه‌بعدی سال ۱۳۹۷ تعداد ۸ سؤال در عامل اول،۶ سؤال در عامل دوم و تعداد ۶ سؤال در عامل سوم بارهای عاملی قابل قبولی دارند.

- نمرات همتراز شده.

سه مجموعه داده مربوط به نمرات آزمون ریاضی کنکور سراسری گروه ریاضی و فنی در دو سال ۱۳۹۶ و ۱۳۹۷ از طریق سه روش همترازسازی نمره واقعی نظریه سؤال پاسخ تک‌بعدی، همترازسازی نمره مشاهده‌شده نظریه سؤال پاسخ تک‌بعدی و همترازسازی

همصدک همتراز شده‌اند، نتایج مربوط به همترازسازی نمرات در جدول شماره (۸) آمده است.

جدول ۸. نمرات همتراز شده در سه روش همترازسازی در سه مجموعه داده تک‌بعدی، دوبعدی و سه‌بعدی در آزمون ریاضی سال ۹۶ و ۹۷

نمرات خام	آزمون تک‌بعدی			آزمون دوبعدی			آزمون سه‌بعدی		
	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده
۰	۰/۲۴۱	۰	−۰/۱۹۰	۰/۰۹۴	۰	−۰/۰۴۱	−۰/۲۴۲	۰	−۰/۱۶۳
۱	۱/۲۵۶	۰/۶۲۶	۰/۵۰۵	۱/۱۸۸	۰/۹۰۵	۰/۸۷۰	۰/۳۷۵	۰/۴۳۰	۰/۳۷۱
۲	۲/۲۷۵	۱/۵۸۰	۱/۴۸۵	۲/۱۴۹	۲/۱۳۵	۲/۱۱۵	۱/۱۶۰	۰/۹۵۵	۰/۹۹۳
۳	۳/۳۰۰	۲/۴۹۲	۲/۴۳۸	۳/۰۱۱	۳/۴۶۸	۳/۴۳۹	۱/۹۰۷	۱/۵۲۹	۱/۵۰۴
۴	۴/۳۹۴	۳/۳۳۵	۳/۳۱۰	۳/۸۱۹	۴/۸۲۱	۴/۸۰۸	۲/۶۱۲	۲/۱۶۲	۲/۱۹۳
۵	۵/۵۲۰	۴/۱۳۷	۴/۱۲۶	۴/۶۴۲	۶/۱۵۴	۶/۱۵۳	۳/۳۸۴	۲/۸۵۷	۲/۸۸۰
۶	۶/۵۳۶	۴/۹۳۲	۴/۹۱۴	۵/۴۷۰	۷/۴۴۵	۷/۴۳۹	۴/۲۴۳	۳/۶۱۲	۳/۶۲۰
۷	۷/۴۵۱	۵/۷۴۲	۵/۷۲۵	۶/۲۹۳	۸/۶۸۵	۸/۶۸۲	۵/۰۳۶	۴/۴۱۸	۴/۴۲۵
۸	۸/۵۷۳	۶/۵۸۵	۶/۵۸۱	۷/۱۷۰	۹/۸۷۵	۹/۸۷۷	۵/۸۶۳	۵/۲۷۱	۵/۲۸۲
۹	۹/۷۰۳	۷/۴۷۴	۷/۴۸۱	۸/۰۱۸	۱۱/۰۲۳	۱۱/۰۲۴	۶/۶۶۴	۶/۱۷۳	۶/۱۹۴
۱۰	۱۰/۹۲۲	۸/۴۱۸	۸/۴۳۳	۸/۷۷۶	۱۲/۱۳۶	۱۲/۱۳۷	۷/۳۴۸	۷/۱۳۸	۷/۱۷۲
۱۱	۱۲/۰۹۲	۹/۴۳۱	۹/۴۵۶	۹/۴۵۲	۱۳/۲۰۸	۱۳/۲۰۴	۸/۲۱۶	۸/۱۹۰	۸/۲۳۳
۱۲	۱۳/۰۳۹	۱۰/۵۲۷	۱۰/۵۶۹	۱۰/۴۲۷	۱۴/۲۱۷	۱۴/۲۰۱	۹/۳۲۵	۹/۳۶۱	۹/۴۱۲
۱۳	۱۳/۸۴۰	۱۱/۷۳۰	۱۱/۷۹۵	۱۱/۶۰۵	۱۵/۱۳۵	۱۵/۱۰۸	۱۰/۵۱۵	۱۰/۶۸۲	۱۰/۷۴۹
۱۴	۱۴/۸۹۵	۱۳/۰۹۴	۱۳/۲۱۶	۱۲/۸۲۲	۱۵/۹۳۸	۱۵/۹۴۲	۱۱/۹۶	۱۲/۱۷۶	۱۲/۲۴۷
۱۵	۱۶/۱۳۸	۱۴/۷۳۹	۱۴/۹۵۸	۱۳/۷۸۴	۱۶/۶۲۳	۱۶/۷۱۹	۱۳/۸۸۲	۱۳/۸۳۹	۱۳/۹۳۰
۱۶	۱۷/۰۶۲	۱۶/۶۷۷	۱۶/۸۲۸	۱۵/۷۴۹	۱۷/۲۰۱	۱۷/۴۱۰	۱۵/۳۱۵	۱۵/۶۴۱	۱۵/۷۳۳
۱۷	۱۷/۹۲۷	۱۸/۲۷۴	۱۸/۲۱۴	۱۷/۱۶۶	۱۷/۶۹۴	۱۸/۱۹۲	۱۶/۴۰	۱۷/۴۸۶	۱۷/۴۳۴
۱۸	۱۸/۵	۱۹/۲۰۱	۱۹/۲۴۹	۲۰/۵	۱۸/۱۴۱	۱۸/۸۸۰	۱۷/۸۰	۱۹/۰۸۶	۱۸/۷۸۷
۱۹	۱۹/۵	۱۹/۷۲۰	۲۰/۱۶	۲۰/۵	۱۸/۶۳۰	۱۹/۴۵۹	۱۹/۵	۱۹/۹۲۹	۱۹/۸۶۲
۲۰	۲۰/۵	۲۰	۲۰/۴۷۳	۲۰/۵	۲۰	۲۰/۳۹۲	۲۰/۵	۲۰	۲۰/۴۳۸

همان‌طور که در جدول فوق مشاهده می‌شود در سه مجموعه داده، نمرات همتراز شده در دو روش نمره واقعی IRT و نمره مشاهده‌شده IRT مشابه با یکدیگر هستند و تنها در چند صدم نمره با هم متفاوت هستند، همچنین در داده‌های تک‌بعدی و دوبعدی تفاوت بین دو

روش همترازسازی تحت نظریه سؤال پاسخ با روش همصدک زیاد است اما در داده‌های سه‌بعدی تفاوت بین سه روش همترازسازی اندک است.

- اثر بعدیت آزمون و روش همترازسازی بر خطای استاندارد همترازسازی (SEE).

جهت بررسی اثر بعدیت آزمون و روش همترازسازی در خطای استاندارد همترازسازی از تحلیل واریانس عاملی استفاده شد. قبل از ارائه نتایج تحلیل واریانس، میزان خطای استاندارد همترازسازی در سه روش همترازسازی (روش همترازسازی همصدک، نمره واقعی و نمره مشاهده‌شده) و سه مجموعه آزمون (تک‌بعدی، دوبعدی و سه‌بعدی) در جدول زیر ارائه شده است.

جدول ۹. خطای استاندارد همترازسازی (SEE) در سه روش همترازسازی در مجموعه داده‌های تک‌بعدی، دوبعدی و سه‌بعدی در سال ۱۳۹۶ و ۱۳۹۷

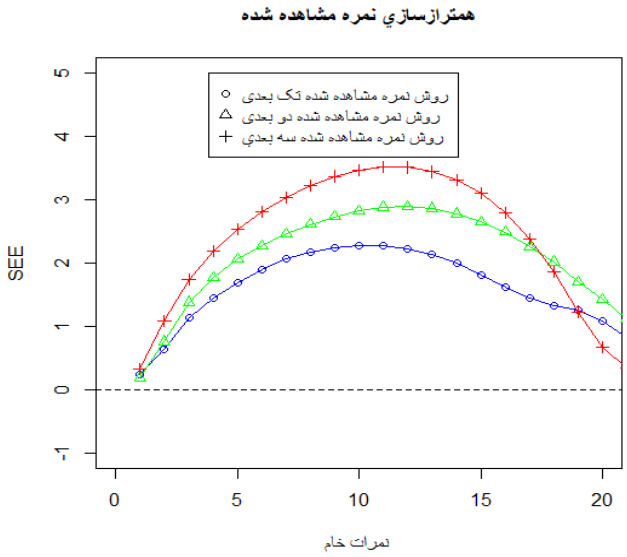
نمرات خام	آزمون تک‌بعدی (SEE)			آزمون دوبعدی (SEE)			آزمون سه‌بعدی (SEE)		
	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده
۰	۰/۲۰	۰	۰/۲٤	۰/۲۹	۰	۰/۱۹	۰/۳٦	۰	۰/۳۳
۱	۰/٤۲	۰/٦٦	۰/٦٤	۰/٤٤	۰/۸۱	۰/۷٦	۰/٦۱	۱/۰۹	۱/۰۸
۲	۰/۸۱	۱/۱۷	۱/۱۳	۰/۵۰	۱/٤۳	۱/۳۸	۰/۹٦	۱/۷۹	۱/۷٤
۳	۰/۹۷	۱/۵۱	۱/٤۵	۰/٦۲	۱/۸۱	۱/۷۷	۱/۱٦	۲/۲۵	۲/۱۹
٤	۱/۲۱	۱/۷۷	۱/٦۹	۰/۷۵	۲/۰۹	۲/۰٦	۱/٤۲	۲/٦۰	۲/۵٤
۵	۱/٤۷	۱/۹۷	۱/۹۰	۰/۸۷	۲/۳۰	۲/۲۷	۱/۷۱	۲/۸۸	۲/۸۱
٦	۱/٦٦	۲/۱۳	۲/۰٦	۰/۹۳	۲/٤۸	۲/٤٦	۱/۹۰	۳/۱۱	۳/۰٤
۷	۱/۸۳	۲/۲۵	۲/۱۷	۰/۸۳	۲/٦۲	۲/٦۱	۲/۰۱	۳/۳۰	۳/۲۳
۸	۱/۹۷	۲/۳۳	۲/۲٤	۰/۹٤	۲/۷۳	۲/۷۳	۲/۱۹	۳/٤٤	۳/۳۷
۹	۱/۹۳	۲/۳۵	۲/۲۸	۱/۱٤	۲/۸۲	۲/۸۲	۲/۲٤	۳/۵٤	۳/٤۷
۱۰	۱/۹٤	۲/۳۳	۲/۲۸	۱/۳۳	۲/۸٦	۲/۸۸	۲/۳۵	۳/۵۸	۳/۵۲
۱۱	۱/۹۳	۲/۲٦	۲/۲۳	۱/٦۵	۲/۸۷	۲/۸۹	۲/۵۳	۳/۵۷	۳/۵۱
۱۲	۱/۸٤	۲/۱٤	۲/۱۳	۱/۷٦	۲/۸۲	۲/۸٦	۲/۵۵	۳/۵۰	۳/٤٤
۱۳	۱/۵۹	۱/۹۷	۱/۹۹	۱/٤۰	۲/۷۳	۲/۷۸	۲/۱۲	۳/۳۷	۳/۳۱
۱٤	۱/٦۱	۱/۷٦	۱/۸۱	۱/۱٤	۲/۵۹	۲/٦۵	۱/۹۸	۳/۱۵	۳/۱۰
۱۵	۱/٤۳	۱/۵٤	۱/٦۱	۱/۲۵	۲/۳۹	۲/٤۹	۱/۹۰	۲/۸۵	۲/۷۹
۱٦	۱/٤٤	۱/۳۸	۱/٤٤	۱/۳۰	۲/۱٤	۲/۲۵	۱/۹۵	۲/٤٤	۲/۳۸
۱۷	۱/۳۳	۱/۲۱	۱/۳۲	۱/۲۸	۱/۸٤	۲/۰۲	۱/۸۵	۱/۹۲	۱/۸٦

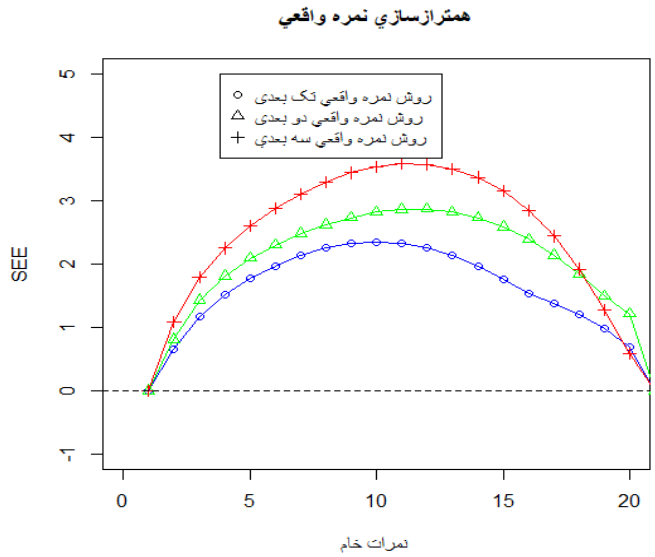
نمرات خام	آزمون تک‌بعدی (SEE) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون دوبعدی (SEE) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون سه‌بعدی (SEE) همصدک	نمره واقعی	نمره مشاهده‌شده
۱۸	۱/۳٤	۰/۹۸	۱/۲۵	۰/۸۹	۱/٤۹	۱/۷۰	۱/٦۱	۱/۲۷	۱/۲۳
۱۹	۱/۲۵	۰/٦۸	۱/۰۹	۰/۸٤	۱/۲۱	۱/٤۲	۱/۵۱	۰/۵۸	۰/٦۷
۲۰	۱/۰۳	۰	۰/۸۳	۰/۵٤	۰	۱/۱۲	۱/۱٦	۰	۰/۳٤

نتایج مربوط به آماره خطای استاندارد همترازسازی در جدول فوق نشان می‌دهد که در سه روش همترازسازی با افزایش تعداد ابعاد آزمون، میزان خطای همترازسازی افزایش می‌یابد.

در نمودارهای ۱ میزان خطای استاندارد در سه روش همترازسازی در سه مجموعه داده ارائه شده است.

نمودار ۱. میزان خطای استاندارد همترازسازی در روش همصدک، نمره مشاهده‌شده و نمره واقعی در داده‌های تک‌بعدی، دوبعدی و سه‌بعدی





همان‌طور که نمودارهای فوق نشان می‌دهد در هر سه روش همترازسازی، هر چه تعداد ابعاد بیشتر می‌شود، میزان خطای استاندارد همترازسازی افزایش‌یافته و کمترین خطا در هر سه روش مربوط به همترازسازی داده‌های تک‌بعدی و بیشترین خطای استاندارد در همه روش‌ها مربوط به داده‌های سه‌بعدی است.

برای بررسی معنی‌داری اثر بعدیت و اثر روش‌های همترازسازی بر خطای استاندارد همترازسازی از آزمون تحلیل واریانس دوعاملی ۳ ×۳ استفاده شد. نتایج مربوط به تحلیل واریانس دوعاملی در جدول زیر آمده است.

جدول ۱۰. نتایج تحلیل واریانس دوعاملی جهت تعیین اثر بعدیت و روش همترازسازی بر خطای همترازسازی

منابع تغییرات	مجموع مجذورات	درجه آزادی	میانگین مجذورات	F	معنی‌داری	مجذور اتای تفکیکی
بعدیت آزمون	۱۷/۷٦	۲	۸/۸۸	۱۱/۵۸	۰/۰۰۰۱	۰/۱۱٤
روش همترازسازی	٤/۵۸	۲	۲/۲۹	۲/۹۹	۰/۰۵۳	۰/۰۳۲
بعدیت *روش	۰/۵٤	٤	۰/۱۳	۰/۱۷	۰/۹۵	۰/۰۰٤
خطا	۱۳۸/۰۰۱	۱۸۰	۰/۷٦			
کل	۸۳۹/٦٤	۱۸۹				

نتایج مربوط به تحلیل واریانس دوعاملی نشان می‌دهد که فقط اثر بعدیت بر خطای استاندارد همترازسازی معنی‌دار است (P<۰/۰۰۰۱، F=۱۱/۵۸، η^2=۰/۱۱۴). مجذور اتای تفکیکی نشان‌دهنده این است که حدود ۱۱ درصد از تغییرات خطای استاندارد همترازسازی مربوط به اثر بعدیت آزمون است. اما اثرات روش همترازسازی و اثر متقابل بعدیت و روش همترازسازی بر خطای استاندارد همترازسازی معنی‌دار نیست (P>۰/۰۵). نتایج مربوط به آزمون تعقیبی توکی نشان داد که بین دو نوع آزمون تک‌بعدی و دوبعدی (P<۰/۰۵) و بین دو نوع آزمون تک‌بعدی و سه‌بعدی (P<۰/۰۰۰۱) ازلحاظ خطای استاندارد همترازسازی تفاوت معنی‌داری وجود دارد و با توجه به میانگین بیشتر خطای همترازسازی در آزمون دوبعدی و سه‌بعدی نسبت به آزمون تک‌بعدی می‌توان گفت با افزایش بعدیت آزمون، خطای همترازسازی افزایش می‌یابد.

- اثر بعدیت آزمون و روش همترازسازی بر سوگیری نتایج همترازسازی (BIAS)

در جدول (۱۱) میزان سوگیری نتایج همترازسازی در سه روش همترازسازی موردمطالعه و به تفکیک سه مجموعه داده تک‌بعدی، دوبعدی و چندبعدی ارائه شده است.

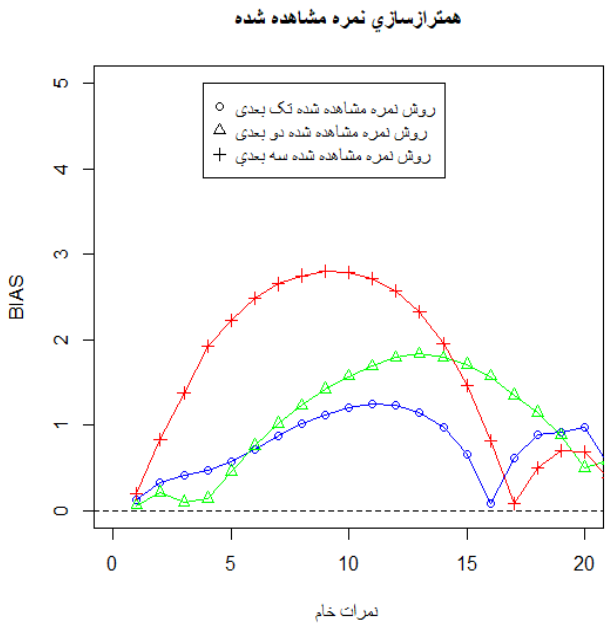
جدول ۱۱. میزان سوگیری نتایج همترازسازی (BIAS) در سه روش همترازسازی در مجموعه داده‌های تک‌بعدی، دوبعدی و سه‌بعدی در سال ۱۳۹۶ و ۱۳۹۷

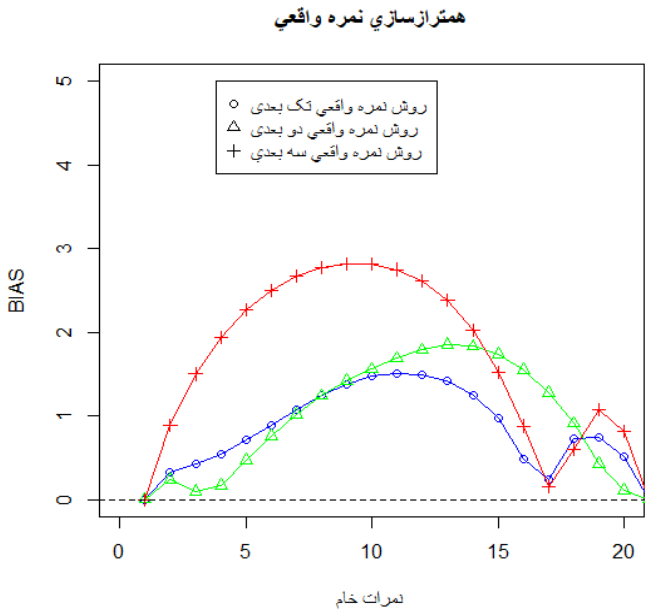
نمرات خام	آزمون تک‌بعدی (BIAS)			آزمون دوبعدی (BIAS)			آزمون سه‌بعدی (BIAS)		
	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده
0	0	0	۰/۱۳	۰/۱۱	0	۰/۰۶	۰/۲۹	0	۰/۲۰
۱	۰/۷۸	۰/۳۲	۰/۳۳	۰/۲۶	۰/۲٤	۰/۲۱	۰/٦٦	۰/۸۹	۰/۸۳
۲	۰/۹۷	۰/٤۳	۰/٤۱	۰/۲۱	۰/۱۰	۱/۱۰	۰/۹۹	۱/۵۰	۱/۳۸
۳	۱/۰۷	۰/۵٤	۰/٤۷	۰/۰٦	۰/۱۷	۰/۱٤	۱/۳۱	۱/۹٤	۱/۹۲
٤	۱/۱۹	۰/۷۱	۰/۵۷	۰/۲۰	۰/٤۷	۰/٤٦	۱/٦۳	۲/۲۷	۲/۲۲
۵	۱/۳۰	۰/۸۹	۰/۷۱	۰/۳٤	۰/۷٦	۰/۷٦	۲/۰۳	۲/۵۰	۲/٤۸
٦	۱/۳۷	۱/۰۸	۰/۸۷	۰/٤۲	۱/۰۲	۱/۰۲	۲/۳۰	۲/٦۷	۲/٦۵
۷	۱/٤۱	۱/۲٤	۱/۰۲	۰/٤۷	۱/۲٤	۱/۲۳	۲/۵۷	۲/۷۷	۲/۷۵
۸+	۱/٤۷	۱/۳۷	۱/۱۲	۰/٦۰	۱/٤۲	۱/٤۲	۲/۷۷	۲/۸۱	۲/۸۰
۹	۱/٤۹	۱/٤۷	۱/۲۰	۰/۷۰	۱/۵٦	۱/۵۷	۲/۹۱	۲/۸۱	۲/۷۹
۱۰	۱/۵۳	۱/۵۱	۱/۲٤	۱/۰٤	۱/٦۹	۱/٦۹	۳/۰۱	۲/۷۵	۲/۷۱
۱۱	۱/۵۹	۱/٤۹	۱/۲۳	۱/۵۸	۱/۷۹	۱/۷۹	۲/۸٦	۲/٦۱	۲/۵۷

نمرات خام	آزمون تک‌بعدی (BIAS) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون دوبعدی (BIAS) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون سه‌بعدی (BIAS) همصدک	نمره واقعی	نمره مشاهده‌شده
۱۲	۱/۵۹	۱/۴۲	۱/۱۵	۱/۵۷	۱/۸۵	۱/۸۳	۲/۴۱	۲/۳۸	۲/۳۳
۱۳	۱/۴۵	۱/۲۵	۰/۹۸	۱/۵۰	۱/۸۳	۱/۷۹	۱/۸۵	۲/۰۲	۱/۹۵
۱۴	۱/۴۰	۰/۹۷	۰/۶۶	۱/۳۵	۱/۷۴	۱/۷۱	۱/۲۸	۱/۵۲	۱/۴۶
۱۵	۱/۳۷	۰/۴۸	۰/۰۸	۱/۲۴	۱/۵۵	۱/۵۷	۰/۶	۰/۸۸	۰/۸۱
۱۶	۱/۳۹	۰/۲۴	۰/۶۲	۰/۲۶	۱/۲۸	۱/۳۵	۰/۵۹	۰/۱۵	۰/۰۸
۱۷	۱/۳۶	۰/۷۳	۰/۸۹	۰/۱۰	۰/۹۱	۱/۱۵	۰/۶۴	۰/۶۰	۰/۵۰
۱۸	۱/۲۶	۰/۷۵	۰/۹۱	۲/۱۳	۰/۴۳	۰/۸۸	۰/۲۵	۱/۰۸	۰/۷۰
۱۹	۱/۲۲	۰/۵۲	۰/۹۷	۱/۷۱	۰/۱۱	۰/۵۰	۰/۳۲	۰/۸۱	۰/۶۹
۲۰	۱/۰۷	۰	۰/۵۴	۰	۰	۰/۵۸	۰/۴۶	۰	۰/۳۹

نتایج مربوط به سوگیری در جدول (۱۱) نشان می‌دهد که بیشترین میزان سوگیری نتایج مربوط به روش‌های همترازسازی در داده‌های سه‌بعدی است. نتایج مربوط به سوگیری همترازسازی به‌صورت دیداری در نمودار زیر ارائه شده است.

نمودار ۲. میزان سوگیری نتایج مربوط به سه روش همترازسازی در سه مجموعه داده تک‌بعدی، دوبعدی و سه‌بعدی





نمودار (۲) نشان می‌دهد که در روش همصدک، بیشترین میزان سوگیری مربوط به همترازسازی داده‌های سه‌بعدی و کمترین میزان سوگیری مربوط به داده‌های تک‌بعدی است. برای بررسی معنی‌داری اثر بعدیت و اثر روش‌های همترازسازی بر سوگیری نتایج همترازسازی از آزمون تحلیل واریانس دوعاملی ۳ ×۳ استفاده شد. نتایج مربوط به تحلیل واریانس دوعاملی در جدول زیر آمده است.

جدول ۱۲. نتایج تحلیل واریانس دوعاملی جهت تعیین اثر بعدیت و روش همترازسازی بر سوگیری همترازسازی

منابع تغییرات	مجموع مجذورات	درجه آزادی	میانگین مجذورات	F	معنی‌داری	ضریب اتای تفکیکی
بعدیت آزمون	۲۱/۳۷	۲	۱۰/٦۸	۲۰/۱۸	۰/۰۰۰۱	۰/۱۸۳
روش همترازسازی	۰/۱۷	۲	۰/۰۸	۰/۱٦	۰/۸٤۷	۰/۰۰۲
بعدیت *روش	۱/۵۳	٤	۰/۳۸	۰/۷۲	۰/۵۷۵	۰/۰۱٦
خطا	۹۵/۳۲	۱۸۰	۰/۵۳			
کل	۳۵۸/۳۰	۱۸۹				

نتایج مربوط به تحلیل واریانس دوعاملی نشان می‌دهد که اثر بعدیت بر سوگیری نتایج همترازسازی معنی‌دار است ($P<۰/۰۰۰۱$، $F=۲۰/۱۸$، $\eta^2=۰/۱۸۳$)، ضریب اتای تفکیکی نیز

نشان می‌دهد که حدود ۱۸ درصد از تغییرات سوگیری همترازسازی مربوط به اثر بعدیت است. اما اثر روش همترازسازی و اثر متقابل بعدیت و روش همترازسازی بر سوگیری نتایج همترازسازی معنی‌دار نیست (P>۰/۰۵). نتایج مربوط به آزمون تعقیبی توکی نشان داد که بین دو نوع آزمون تک‌بعدی و سه‌بعدی (P<۰/۰۱) و بین دو آزمون دوبعدی و سه‌بعدی (P<۰۰۰۱) ازلحاظ سوگیری نتایج همترازسازی تفاوت معنی‌داری وجود دارد و با توجه به میانگین بیشتر سوگیری در آزمون سه‌بعدی نسبت به آزمون تک‌بعدی و دوبعدی می‌توان گفت میزان سوگیری نتایج همترازسازی در آزمون سه‌بعدی نسبت به آزمون تک‌بعدی و دوبعدی بیشتر است.

- اثر بعدیت آزمون و روش همترازسازی بر مجذور میانگین مربع خطای همترازسازی (RMSE)

در جدول (۱۳) مجذور میانگین مربع خطای همترازسازی[1] (RMSE) به تفکیک روش‌های مختلف همترازسازی در سه مجموعه داده تک‌بعدی، دوبعدی و سه‌بعدی ارائه شده است.

جدول ۱۳. مجذور میانگین مربع خطا (RMSE) در سه روش همترازسازی در مجموعه داده‌های تک‌بعدی، دوبعدی و سه‌بعدی در سال ۱۳۹۶ و ۱۳۹۷

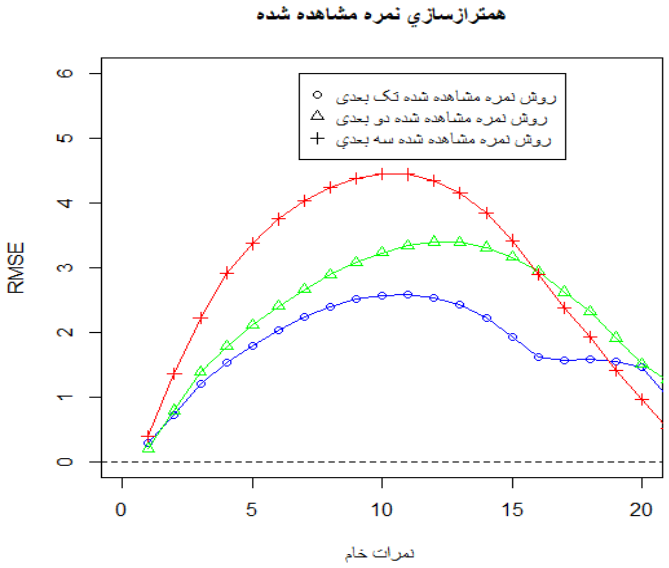
نمرات خام	آزمون تک‌بعدی (RMSE)			آزمون دوبعدی (RMSE)			آزمون سه‌بعدی (RMSE)		
	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده	همصدک	نمره واقعی	نمره مشاهده‌شده
۰	۰/۳۶	۰	۰/۲۸	۰/۲۰	۰	۰/۲۰	۰/٤٦	۰	۰/۳۹
۱	۰/٦۱	۰/۷٤	۰/۷۲	۰/۵۳	۰/۸۵	۰/۷۹	۰/۹٦	۱/٤۱	۱/۳٦
۲	۰/۹٦	۱/۲۵	۱/۲۱	۰/۸۰	۱/٤۳	۱/۳۹	۱/٤۵	۲/۳۳	۲/۲۲
۳	۱/۱٦	۱/٦۱	۱/۵۳	۱/۱۱	۱/۸۲	۱/۷۸	۱/۹۲	۲/۹۷	۲/۹۲
٤	۱/٤۲	۱/۹۰	۱/۷۹	۱/۵٦	۲/۱٤	۲/۱۱	۲/۳۲	۳/٤۵	۳/۳۷
۵	۱/۷۱	۲/۱۷	۲/۰۳	۱/۸۸	۲/٤۲	۲/٤۰	۲/۹۰	۳/۸۲	۳/۷۵
٦	۱/۹۰	۲/۳۹	۲/۲٤	۲/۲٤	۲/٦۸	۲/٦٦	۳/۳۸	٤/۱۰	٤/۰٤
۷	۲/۰۱	۲/۵۷	۲/٤۰	۲/٤۵	۲/۹۰	۲/۸۹	۳/۸۵	٤/۳۱	٤/۲٤
۸	۲/۱۹	۲/۷۰	۲/۵۱	۲/۵۷	۳/۰۸	۳/۰۸	٤/۱۹	٤/٤۵	٤/۳۸
۹	۲/۲٤	۲/۷۷	۲/۵۷	۲/٦۱	۳/۲۲	۳/۲۳	٤/٤٤	٤/۵۲	٤/٤۵
۱۰	۲/۳۵	۲/۷۷	۲/۵۹	۲/٦۱	۳/۳۲	۳/۳٤	٤/۵۸	٤/۵۱	٤/٤٤
۱۱	۲/۵۳	۲/۷۱	۲/۵٤	۲/۷۸	۳/۳۸	۳/٤۰	٤/۵٤	٤/٤۲	٤/۳۵
۱۲	۲/۵۵	۲/۵٦	۲/٤۲	۲/٦۸	۳/۳۷	۳/۳۹	٤/۲۱	٤/۲۳	٤/۱۵

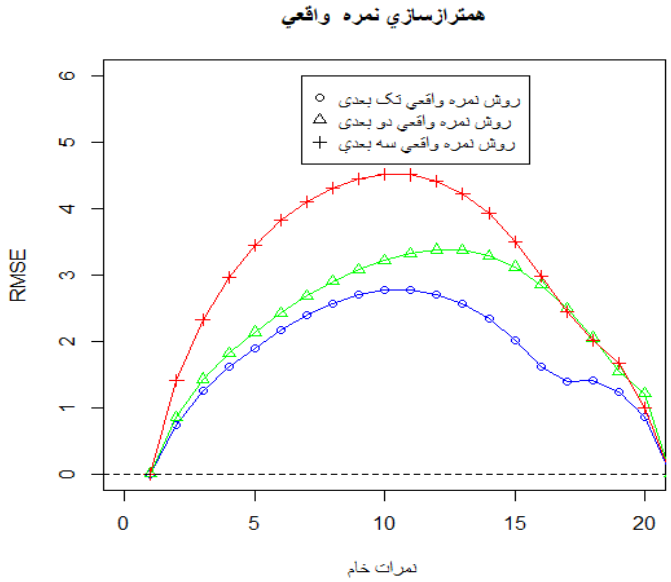
1. Root mean square error statistic

نمرات خام	آزمون تک‌بعدی (RMSE) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون دوبعدی (RMSE) همصدک	نمره واقعی	نمره مشاهده‌شده	آزمون سه‌بعدی (RMSE) همصدک	نمره واقعی	نمره مشاهده‌شده
۱۳	۲/۱۲	۲/۳٤	۲/۲۲	۲/۵۰	۳/۲۹	۳/۳۱	۳/۷۱	۳/۹۳	۳/۸٤
۱٤	۱/۹۸	۲/۰۱	۱/۹۳	۲/۳۱	۳/۱۲	۳/۱٦	۳/۱۷	۳/۵۰	۳/٤۲
۱۵	۱/۹۰	۱/٦۲	۱/٦۱	۲/۰٤	۲/۸۵	۲/۹٤	۲/۲۷	۲/۹۸	۲/۹۰
۱٦	۱/۹۵	۱/٤۰	۱/۵۷	۱/٤۱	۲/٤۹	۲/٦۲	۱/۵۰	۲/٤۵	۲/۳۸
۱۷	۱/۸۵	۱/٤۱	۱/۵۹	۱/٤٤	۲/۰۵	۲/۳۲	۱/۱٦	۲/۰۱	۱/۹۳
۱۸	۱/٦۱	۱/۲٤	۱/۵۵	۲/۷۳	۱/۵۵	۱/۹۱	۰/۸۵	۱/٦۷	۱/٤۱
۱۹	۱/۵۱	۰/۸٦	۱/٤۹	۲/۲٦	۱/۲۱	۱/۵۱	۰/۸۷	۱	۰/۹٦
۲۰	۱/۱٦	۰	۰/۹۹	۰	۰	۱/۲٦	۰/۸۷	۰	۰/۵۱

همان‌طور که در جدول فوق مشاهده می‌شود، میزان RMSE در همه روش‌های همترازسازی در داده‌های سه‌بعدی بیشتر از داده‌های دوبعدی و در داده‌های دوبعدی بیشتر از یک‌بعدی است. در نمودار (۳) تفاوت میزان RSME در سه روش موردمطالعه در سه مجموعه داده تک‌بعدی، دوبعدی و سه‌بعدی به‌صورت دیداری ارائه شده است.

نمودار ۳: میزان RMSE نتایج مربوط به سه روش همترازسازی در سه مجموعه داده تک‌بعدی، دوبعدی و سه‌بعدی





همان‌طور که در نمودار فوق مشاهده می‌شود در تمامی روش‌های همصدک، نمره مشاهده‌شده و روش نمره واقعی، مقدار RMSE در داده‌های سه‌بعدی بیشتر از داده‌های دوبعدی و تک‌بعدی است. برای بررسی معنی‌داری اثر بعدیت و اثر روش‌های همترازسازی بر مجذور میانگین مربع خطای همترازسازی از آزمون تحلیل واریانس دوعاملی ۳×۳استفاده شد. نتایج مربوط به تحلیل واریانس دوعاملی در جدول زیر آمده است.

جدول ١٤: نتایج تحلیل واریانس دوعاملی جهت تعیین اثر بعدیت و روش همترازسازی بر مجذور میانگین مربع خطای همترازسازی

منابع تغییرات	مجموع مجذورات	درجه آزادی	میانگین مجذورات	F	معنی‌داری	مجذور اتای سهمی
بعدیت آزمون	٣٥/٦٠	٢	١٧/٨٠	١٥/٣٨	٠/٠٠٠١	٠/١٤٦
روش همترازسازی	٣/٩٥	٢	١/٩٧	١/٧٠	٠/١٨٤	٠/٠١٩
بعدیت *روش	١/٣٨	٤	٠/٣٤	٠/٢٩	٠/٨٧٨	٠/٠٠٧
خطا	٢٠٨/٣٠	١٨٠	١/١٥			
کل	١١٩٧/٩٨	١٨٩				

نتایج مربوط به تحلیل واریانس دوعاملی نشان می‌دهد که اثر بعدیت بر مجذور میانگین مربع خطای همترازسازی معنی‌دار است (P<۰/۰۰۰۱، F=۱۵/۳۸، η^2=۰/۱۴۶) و میزان ۱۴ درصد از تغییرات مجذور میانگین مربع خطای همترازسازی مربوط به اثر بعدیت آزمون است. اما اثر روش همترازسازی و اثر متقابل بعدیت و روش همترازسازی بر مجذور میانگین مربع خطای همترازسازی معنی‌دار نیست (P>۰/۰۵). نتایج مربوط به آزمون تعقیبی توکی نشان داد که بین دو نوع آزمون تک‌بعدی و دوبعدی (P<۰/۰۱) و بین دو آزمون تک‌بعدی و سه‌بعدی (P<۰/۰۰۰۱) و دو آزمون دوبعدی و سه‌بعدی (p<۰/۰۱) ازلحاظ مجذور میانگین مربع خطای همترازسازی تفاوت معنی‌داری وجود دارد و با توجه به میانگین بیشتر مجذور میانگین مربع خطای همترازسازی در آزمون سه‌بعدی نسبت به آزمون تک‌بعدی و دوبعدی می‌توان گفت میزان مجذور میانگین مربع خطای همترازسازی در آزمون سه‌بعدی نسبت به آزمون تک‌بعدی و دوبعدی بیشتر است. همچنین میزان خطای آزمون دوبعدی به‌طور معنی‌داری از خطای آزمون تک‌بعدی بیشتر است.

بحث و نتیجه‌گیری

همترازسازی روشی آماری برای اطمینان از عادلانه بودن نتایج آزمون‌ها است، در پیشینه نظریه سؤال پاسخ، شرط استفاده از این مدل‌ها برقراری مفروضه تک‌بعدی بودن آزمون‌ها است، اما این مفروضه در اکثر مواقع نقض می‌شود و نمی‌توان از تک‌بعدی بودن آزمون‌ها اطمینان حاصل کرد. هدف پژوهش، مطالعه اثر نقض تک‌بعدی بودن و اثر روش‌های همترازسازی بر خطاهای همترازسازی بود. در این پژوهش، اثر بعدیت بر خطای استاندارد همترازسازی معنی‌دار بود، اما اثر روش همترازسازی و اثر متقابل بعدیت و روش همترازسازی بر خطای استاندارد همترازسازی معنی‌دار نبود. این یافته تأثیر نقض مفروضه تک‌بعدی بودن را بر عملکرد همترازسازی نشان می‌دهد، همان‌طور که یافته‌ها نشان می‌دهند بین دو نوع آزمون تک‌بعدی و دوبعدی و بین دو نوع آزمون تک‌بعدی و سه‌بعدی ازلحاظ خطای استاندارد همترازسازی تفاوت معنی‌داری وجود داشت و با افزایش بعدیت آزمون، خطای همترازسازی افزایش یافت. این یافته با پژوهش Spence (1996) و Peterson (2014) که نشان دادند با افزایش تعداد سؤالات چندبعدی در آزمون، اثر چندبعدیتی بر همترازسازی بیشتر می‌شود و همچنین مدل‌های چندبعدی برای داده‌های چندبعدی بهتر عمل می‌کنند، همسو است. در پژوهش Ricker (2007) همترازسازی هنگامی‌که فقط سؤالات مشترک

بعد دوم را اندازه می‌گرفتند، در برابر نقض تک‌بعدی بودن مقاوم بود. همچنین این یافته در راستای پژوهش Lim and Lee (2016) است که در پژوهش آن‌ها، نتایج همترازسازی برای روش همترازسازی نمره واقعی و نمره مشاهده‌شده صفر مرتبه با مقدار بار عاملی و تغییرپذیری عامل‌های مشترک و اختصاصی تحت تأثیر قرار گرفت. در نظریه سؤال پاسخ با نقض تک‌بعدی بودن، تبدیل نمرات از یک فرم به فرم دیگر مستقل از آزمودنی‌هایی که آزمون بر روی آن‌ها اجرا شده است، نخواهد بود، ازاین‌رو ویژگی نامتغیر بودن آزمودنی‌ها در این شرایط نقض شده و تابع همترازسازی به‌دست‌آمده از آزمون‌هایی که صفات پنهان مختلف را اندازه‌گیری می‌کنند، برای زیرگروه‌های آزمودنی‌ها متفاوت خواهد بود (Champlain, 1996)، لذا نقض تک‌بعدی بودن سبب نقض ویژگی نامتغیر بودن می‌شود و از این طریق خطا را افزایش می‌دهد.

در خصوص عدم معنی‌داری اثر روش‌های مختلف همترازسازی بر خطای استاندارد همترازسازی، این یافته همسو با پژوهش Chen (2012) است که در آن تفاوتی بین همترازسازی نمره مشاهده‌شده IRT و نمره مشاهده‌شده لوین تفاوتی مشاهده نشد، در توجیه این یافته می‌توان گفت ازآنجاکه سه روش مورداستفاده، روش‌های تک‌بعدی نظریه سؤال پاسخ و نظریه کلاسیک بودند، لذا ماهیت تقریباً یکسان این روش‌ها به لحاظ تک‌بعدی بودن تفاوتی در همترازسازی داده‌های تک‌بعدی و چندبعدی ایجاد نکرده است و می‌توان گفت تفاوت در عملکرد همترازسازی نه به فرایند همترازسازی بلکه به ساختار آزمون وابسته است و چنانچه ساختار آزمون از چندین بعد تشکیل شده باشد، استفاده از روش‌های تک‌بعدی همترازسازی توجیه منطقی نداشته و باید از روش‌های مناسب برای همترازسازی چنین داده‌هایی استفاده گردد. در این زمینه پیشنهاد می‌شود روش‌های نوین همترازسازی چندبعدی جایگزین روش‌های سنتی و معمول شود.

اثر بعدیت بر سوگیری نتایج همترازسازی معنی‌دار بود و میزان سوگیری نتایج همترازسازی در آزمون سه‌بعدی نسبت به آزمون تک‌بعدی و دوبعدی بیشتر بود. به نظر می‌رسد در شرایطی که آزمون بیش از یک صفت پنهان را اندازه بگیرد اما پاسخ‌ها با استفاده از مدل سؤال پاسخ تک‌بعدی تحلیل شود، برآورد توانایی و برآورد پارامترهای سؤال به میزان زیادی سوگیرانه خواهد بود. درنتیجه استفاده از مدل‌ها و روش‌های تک‌بعدی برای همترازسازی آزمون‌هایی که بیش از یک صفت پنهان را اندازه می‌گیرند، منجر به روابط

همترازی سودار می‌شود (Brossman & Lee, 2013). همچنین با نقض مفروضه تک‌بعدی بودن، مفروضه استقلال موضعی به خطر افتاده و نقض این مفروضه آثار سوئی بر نتایج همترازسازی ازجمله خطا در برآورد بیشینه درستنمایی، برآورد نادقیق پارامترهای سؤال، عدم دقت در برآورد منحنی ویژگی‌های سؤال (ICC[1])، خطا در تبدیل مقیاس و قرار دادن پارامترها در مقیاس مشترک، برآورد نادرست منحنی ویژگی‌های آزمون (TCC[2]) و برازش سؤال در همترازسازی نمره واقعی خواهد شد (Ricker, 2007; Zhang & Stone, 2008; Chen, 2014). در پژوهش Kim et al. (2019) مدل تک‌بعدی دو ارزشی نظریه سؤال پاسخ در شرایطی که میزان وابستگی موضعی سؤال پایین بود روش مناسبی بود.

عدم تفاوت سوگیری بین روش‌های مختلف همترازسازی در این پژوهش همسو با پژوهش Lee (2013)، Brossman (2010) و Peterson (2014) است که نشان دادند که روش‌های همترازسازی تک‌بعدی نظریه سؤال پاسخ مشابه هم عمل می‌کنند.

بررسی اثر بعدیت بر مجذور میانگین مربع خطای همترازسازی نشان‌دهنده معنی‌دار بودن این اثر بود و با افزایش بعدیت آزمون مجذور میانگین مربع خطا افزایش یافت. این یافته با مطالعه Fesq و همکاران (1995) که نشان داد نقض تک‌بعدیتی اثر قابل‌توجهی بر همترازسازی نمره واقعی ندارد، ناهم‌سو است، دلیل عدم همسویی نتایج دو مطالعه ممکن است مربوط به نوع آزمون باشد، زیرا در پژوهش Fesq و همکاران (1995) از داده‌های آزمون پذیرش دانشکده حقوق که نوعی آزمون استعداد است، استفاده شده است، اما در پژوهش حاضر از آزمون ریاضی استفاده شده است. نتایج پژوهش حاضر با نتایج Topczewski (2013) که نشان داد هنگامی‌که مفروضه تک‌بعدی بودن IRT در همترازسازی عمودی نقض می‌شود، برآورد نمرات خطای استاندارد و سوگیری خواهند داشت، همسو است. Peterson (2014) و Lee And Lissitz (2000) پیشنهاد کرده‌اند که در شرایطی که مفروضه تک‌بعدی بودن نقض می‌شود، استفاده از روش‌های همترازسازی چندبعدی کامل، رویکرد قابل‌قبول‌تری نسبت به روش‌های تک‌بعدی است.

با توجه به نتایج پژوهش می‌توان گفت یکی از عوامل اثرگذار بر کیفیت و عملکرد همترازسازی، ساختار آزمون است، لذا نتایج این پژوهش مبین این است که درصورتی‌که ساختار آزمون دارای بیش از یک بعد است، نمی‌توان از روش‌های تک‌بعدی معمول جهت

1. Item characteristic curve
2. Test characteristic curve

همترازسازی داده‌ها استفاده کرد و در صورت نقض تک‌بعدی بودن باید به فکر روش‌های جایگزین همترازسازی برای آزمون‌های کشور افتاد. نکته دیگری که با توجه به نتایج پژوهش حائز اهمیت است، توجه به این نکته است که جهت افزایش دقت و کاهش خطای همترازسازی لازم است قبل از انتخاب روش همترازسازی مناسب برای آزمون‌ها باید ابتدا شرایط هر آزمون و ساختار آن مشخص شود و سپس روش همترازسازی مناسب انتخاب گردد.

نتایج این پژوهش می‌تواند جهت همترازسازی آزمون‌ها در تمامی حوزه‌های روان‌شناسی و تربیتی که آزمون‌ها دارای چندین بعد هستند، مورداستفاده قرار گیرد. پیشنهاد می‌شود که در پژوهش‌های آینده در همترازسازی آزمون‌های چندبعدی با مقیاس بزرگ و سرنوشت‌ساز که با تحصیل و آینده شغلی افراد مرتبط است، از روش‌های معتبرتر و جدیدتر همترازسازی همچون روش‌های همترازسازی چندبعدی که متناسب با ساختار چندبعدی داده‌ها است استفاده نموده و کارایی این روش‌ها را در همترازسازی داده‌های چندبعدی و تک‌بعدی موردمطالعه قرار دهند. همچنین پیشنهاد می‌گردد تأثیر شرایط مختلف همچون همبستگی بین ابعاد و تفاوت فرم‌های مختلف آزمون نیز بررسی شود تا مشخص گردد کدام روش در چه شرایطی نتایج معتبرتری فراهم می‌کند.

تعارض منافع

نویسندگان هیچ‌گونه تعارض منافعی ندارند.

منابع

شاطریان محمدی، فاطمه. (۱۳۸۲). *مقایسه سه روش همتراز سازی همصدک هموار نشده نمره مشاهده شده IRT و نمره واقعی IRT در طرح گروه های نامعادل با سوالات لنکر (پایان نامه کارشناسی ارشد). دانشکده روانشناسی و علوم تربیتی*. دانشگاه علامه طباطبایی. تهران.

مقدم زاده، علی. (۱۳۹۱). *روش بهینه همترازسازی با توجه به ویژگی های بومی آزمونهای ملی ایران: مورد مطالعه آزمون تولیمو و آزمون‌های جامع کنکورهای آزمایشی سازمان سنجش آموزش کشور* (رساله دکترا). دانشکده روانشناسی و علوم تربیتی. دانشگاه علامه طباطبایی. تهران.

رضوانی فر، شیرین. (۱۳۹۱). *همترازسازی نمرات دروس ریاضی و فیزیک رشته علوم تجربی آزمون کنکور سراسری سال‌های ۱۳۸۸ و ۱۳۸۹ براساس نظریه‌های کلاسیک و جدید اندازه‌گیری* (پایان نامه کارشناسی ارشد). دانشکده روانشناسی و علوم تربیتی. دانشگاه علامه طباطبایی. تهران.

واشقانی فراهانی، مریم. (۱۳۸۰). *کاربرد روش همتراز سازی همصدک در معادل سازی نمرات آزمون های ورودی دانشگاهها (کنکور ورودی سال ۱۳۸۷)* (پایان نامه ارشناسی ارشد). دانشکده روانشناسی و علوم تربیتی. دانشگاه علامه طباطبایی، تهران.

References

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Andrews, B. J. (2011). *Assessing first-and second-order equity for the common-item nonequivalent groups design using multidimensional IRT* [Doctoral dissertation, University of Iowa].

Arıkan, Ç. A., & Gelbal, S. (2018). A Comparison of Traditional and Kernel Equating Methods. *International Journal of Assessment Tools in Education*, *5*(3), 417-427. DOI:10.21449/ijate.409826

Brossman, B. G. (2010).*Observed score and true score equating procedures for multidimensional item response theory* [Doctoral dissertation, University of Iowa].

Brossman, B. G., & Lee, W. C. (2013). Observed score and true score equating procedures for multidimensional item response theory. *Applied Psychological Measurement*, *37*(6), 460-481. DOI:10.1177/0146621613484083

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement*, 32(1), 79-96.

Champlain, A. F. (1996). The Effect of Multidimensionality on IRT True-Score Equating for Subgroups of Examinees. *Journal of Educational Measurement*, *33*(2), 181-201. https://doi.org/10.1111/j.1745-3984.1996.tb00488.x

Chen, H. (2012). A Comparison Between Linear IRT Observed-Score Equating and Levine Observed-Score Equating Under the Generalized Kernel Equating Framework. *Journal of Educational Measurement*, *49*(3), 269-284. https://doi.org/10.1111/j.1745-3984.2012.00175.x

Chen, J. (2014). *Model selection for IRT equating of testlet-based tests in the random groups design* [Doctoral dissertation, University of Iowa].

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). An Assessment of the Relationship Between the Assumption of Unidimensionality and the Quality of IRT True-Score Equating 1, 2, 3. *ETS Research Report Series*, 1985(2), i-68. http://dx.doi.org/10.1002/j.2330-8516.1985.tb00115.x

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational*

Measurement, *22*(4), 249-262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Dorans, N.J. & Holland, P.W. (2000). Population invariance and equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37, 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

González, J., & Wiberg, M. (2017). *Applying test equating methods*. New York: Springer. doi, 10, 978-3.

Han, T., Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true-and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education*, *10*(2), 105-121. DOI: 10.1207/s15324818ame1002_1

Hanson, B., & Zeng, L. (2004). PIE: A computer program for IRT equating. (Windows *Console Version, Revised by Z. Cui, May 20, 2004*) [Manual]. Unpublished manuscript, College of Education, University of Iowa, Iowa City, Iowa.

Hirsch, T. M. (1989). Multidimensional equating. *Journal of Educational Measurement*, *26*(4), 337-349. https://doi.org/10.1111/j.1745-3984.1989.tb00338.x

Kim, K. Y., Lim, E., & Lee, W. C. (2019). A Comparison of the Relative Performance of Four IRT Models on Equating Passage-Based Tests. *International Journal of Testing*, 19(3), 248-269. https://doi.org/10.1080/15305058.2018.1530239

Kim, S. Y. (2018). *Simple structure MIRT equating for multidimensional tests* [Doctoral dissertation, University of Lowa].

Kim, S., Cole, K. L., & Mwavita, M. (2018). FIPC Linking Across Multidimensional Test Forms: Effects of Confounding Difficulty within Dimensions. *International Journal of Testing*, *18*(4), 323-345. https://doi.org/10.1504/IJQRE.2019.100168

Kline, R. B. (2011). *Principles and practice of structural equation modeling*, (3rd Ed). New York, NY: Guilford.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer-Verlag.

Lee, E, Lee, W.C., Brennan, R. L. (2014). *Equating Multidimensional Tests under a Random Groups Design: A Comparison of Various Equating Procedures, Center for Advanced Studies in Measurement and Assessment*, CASMA Research Report.

Lee, E. (2013). *Equating multidimensional tests under a random groups design: a comparison of various equating procedures* [Doctoral dissertation, The University of Iowa].

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*(2), 115-138. https://doi.org/10.1177/01466216000242002

Li, Y., Jiao, H., & Lissitz, R. W. (2012). Applying multidimensional item response theory models in validating test dimensionality: An example of K–12 large-scale science assessment. *Journal of Applied Testing Technology*, *13*(2).

Lim, E. (2016). *Subscore equating with the random groups design* [Doctoral dissertation, University of Iowa].

Lim, E; Lee, w. c. (2016). *Subscore Equating and Reporting. Center for Advanced Studies in Measurement and Assessment*, CASMA Research Report.

Lu, R., & Guo, H. (2018). A Simulation Study to Compare Nonequivalent Groups With Anchor Test Equating and Pseudo-Equivalent Group Linking. *ETS Research Report Series*, 2018(1), 1-16. https://doi.org/10.1002/ets2.12196

Meng, Y. (2012). *Comparison of Kernel Equating and Item Response Theory Equating Methods*. ProQuest LLC. 789 East Eisenhower Parkway, PO Box 1346, Ann Arbor, MI 48106.

Moghadamzadeh, A. (2013). *Optimal Smoothing Method of Data in Test Equating: The Case of TOLIMO and Comprehensive Trial Tests of Iran Educational Testing Organization.* [Doctoral Dissertation, Allameh Tabataba'i University]. [In Persian]

Oshima, T. C., Davey, T. C., & Lee, K. (2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37, 357-373.* https://doi.org/10.1111/j.1745-3984.2000.tb01092.x

Peterson, J. L. (2014). *Multidimensional item response theory observed score equating methods for mixed-format tests* [Doctoral dissertation, University of Iowa].

Ricker, K. L. (2007). *The Consequence of Multidimensionality IRT Equating Outcomes Using a Common-Items Nonequivalent Groups Design* [Doctoral dissertation, university of Alberta].

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of statistical software*, *17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Rizvanifar, Shirin. (2012). Equating of the scores of mathematics and physics courses in the field of experimental sciences in the national entrance exams based on classical and new measurement theories. [Master Dissertation, Allameh Tabataba'i University]. [In Persian]

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement*, *75*(6), 954-978. https://doi.org/10.1177/0013164415575147

Shaterian Mohammadi, F. *Comparison of three unsmoothed percentile equating of the observed and True score IRT methods in unequal group design with anchor questions.* [Master Dissertation, Allameh Tabataba'i University]. [In Persian]

Shin, M. (2015). *An Investigation of Subtest Score Equating Methods under Classical Test Theory and Item Response Theory Frameworks* [Doctoral dissertation, University of Massachusetts].

Simon, M. K. (2008). *Comparison of concurrent and separate multidimensional IRT linking of item parameters* [Doctoral Dissertation, University of Minnesota].

Spence, P. D. (1996). *The effect of multidimensionality on unidimensional equating with item response theory* [Doctoral dissertation, University of Florida].

Topczewski, A. M. (2013). *Effect of violating unidimensional item response theory vertical scaling assumptions on developmental score scales* [Doctoral dissertation, University of Iowa].

Vashghani Farahani, M. (2010). *The application of the equipercentile method in equating university entrance exam scores (2017 entrance exam).* [Master Dissertation, Allameh Tabataba'i University]. [In Persian]

Zhang, B., Stone, C. A. (2008). Evaluating Item Fit for Multidimensional Item Response Models, *Educational and Psychological Measurement*, 68(2), 181-196 .DOI:10.1177/0013164407301547

Zhang, O. (2012). *Observed score and true score equating for multidimensional item response theory under nonequivalent group anchor test design* [Doctoral Dissertation, University of Florida].