15(57), Autumn 2024, 33-53
eISSN: 2470-6240 ISSN: 2252-004x
Article Type: Reserch Article
Received March 30, 2024 Received in revised form August 25, 2024
Published Online October 06, 2024 Accepted September 19, 2024

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

ATU PRESS

The Equating Requirements of Scores in Alternative Forms of High-Stakes Tests: The Case Study of the National Entrance Exam of the Foreign Language Applicants

Abdolkarim Shadmehr* *Corresponding Author*, Assistant Professor, National Organization for Assessment and Evaluation of Education System. Email: shadmehrabdolkarim@yahoo.com

Enayatollah Zamanpour Assistant Professor, Department of Educational Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran.

Shoeayb Qasemi PhD in Assessment and Measurement, National Organization for Assessment and Evaluation of Education System, Tehran, Iran.

Abstract

When administering two nationwide tests in a single year, it becomes necessary to create alternate forms of the exams and equate their scores. This equating process involves specific requirements. In the context of high-stakes tests, due to security concerns, pilot testing of items is often not feasible, further highlighting the need for careful equating procedures. The primary objectives of this study were to delve into the requirements for equating the alternate forms of the first and second 2023 English language foreign nationwide examinations. The research in question is of an applied and psychometric nature, relying on secondary data analysis as its methodological foundation. A total of 13,091 candidates from the first test and 19,915 candidates from the second test of the English language 2023 nationwide examination were chosen randomly for analysis. The data from the tests administered in January 2023 and July 2023 were scrutinized as part of this study. Both tests exhibited similar levels of difficulty and discrimination, as indicated by their comparable mean difficulty coefficients and discrimination coefficients. The construct equivalence of the two tests was confirmed through single-factor confirmatory factor analysis, as well as by comparing the performance of male and female candidates and examining the correlation with high school GPA. Additionally, the study examined the equal consistency requirement, finding a value of 0.95 in both test forms. To equate alternate forms, it's crucial to meet certain requirements, and test developers may employ judgmental estimation of item difficulty, as well as utilize the test specification table, to create alternate forms meeting those requirements. Guidelines have been outlined to fulfill the prerequisites for equating alternate forms and crafting items with comparable difficulty levels.

Keywords: National Exam, High-Stakes Test, Alternate Forms; Equating

Cite this Article: Shadmehr, A., Zamanpour, E., & Qasemi, S. (2024). The Equating Requirements of Scores in Alternative Forms of High-Stakes Tests: The Case Study of the National Entrance Exam of the Foreign Language Applicants. *Educational Measurement*, *15*(57), 33-53. https://doi.org/10.22054/jem.2024.81723.3563

Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2024.81723.3563

Extended Abstract

Introduction

It's worth noting that the National Organization of Educational Testing takes into account both the judgmental difficulty of items, as well as the cognitive skills and content areas to be examined, when constructing alternative test forms. Due to strict security measures, it has not been possible to administer and psychometrically assess the test items from the nationwide examination.

In line with the established scientific practices for creating alternate test forms, the organization has prepared their item developers to accurately estimate item difficulty and utilize the judgmental difficulty values stated in the test specification table. This approach aids in minimizing the difference in difficulty between the two test forms and helps to satisfy the requirements for equating high-stakes tests, while adhering to the security measures that prevent direct testing of test items for psychometric evaluation.

It is worth noting that, due to the unique circumstances surrounding nationwide tests, there is currently a lack of existing research focusing on the establishment of equating requirements for these types of examinations. This study has focused on evaluating the success of the establishment of equating requirements through comparing the validity, reliability, and difficulty of the items from the first and second nationwide examinations in 2023 within the English language group (foreign languages).

Literature Review

The process of equating is used to compensate for slight differences in difficulty between alternate test forms. In order to create alternative forms for high-stakes tests, in addition to maintaining the necessary security measures, it is crucial that the items used are designed and developed within specific standards. In this situation, the estimation of the difficulty of the items by a group of judges can be a useful. Test developers or judges sometimes categorize items as easy, moderate, and difficult, or directly estimate difficulty values. Judges generally tend to underestimate the difficulty of new items. According to Goodwin, the disparity between the judges' opinion and the respondents' actual performance is likely due to the fact that judges are often experts in the field, possessing high levels of experience, knowledge, and education.

This expertise may lead to the judges having excessively high expectations of the respondents' abilities and performance. Various suggestions have been proposed to improve the accuracy of difficult judgments, such as ongoing training for developers and judges, as well as the use of established anchor items with known psychometric properties.

Methodology

This research is an applied psychometric study conducted using secondary data analysis. The study population involves all candidates from the English group who participated in the 2023 nationwide foreign language (English) examination rounds held in January and July 2023 respectively. A total of 13,091 respondents participated in the first round of the English language exam, while 19,915 volunteers participated in the second round. Each test consisted of 70 multiple-choice questions, each worth 1 point for a correct answer and 0 points for an incorrect answer or no answer. The tests were divided into 6 sections, comprising: a grammar section with 15 items, a vocabulary portion with 20 items, a section on sentence structure with 5 items, a language functions section with 5 items, a cloze test with 10 items, and a reading comprehension segment with 15 items. Various statistical techniques, including the difficulty index, modified point-biserial correlation, composite reliability coefficient, independent t-test, Cohen's d and r2 effect sizes, confirmatory factor analysis, and Spearman's correlation coefficient, were applied to analyze the data. Statistical analysis was performed using R 4.3.3 software along with the psych, lavaan, and ggplot2 packages.

Results

A comparison of the difficulty indices between the two tests revealed a difference of 0.04 and an effect size d of 0.32, with r2 being 0.026. The average point-biserial correlation for the first test was 0.43, while for the second test it was 0.42. Multiple lines of evidence were used to assess the need to measure the same construct through two distinct tests. One-dimensional factor analysis revealed that both English tests exhibited a one-dimensional structure. The fit indices for the one-dimensional model were comparable and similar in both tests.

The distribution of factor loadings also showed similarities between the two tests. In both tests, men scored higher on average compared to

women, and the effect sizes related to group differences were small and similar in both cases. The correlation coefficients between both tests and GPA were positive. The requirement for equal reliability between the tests was verified using the composite reliability coefficient. The reliability coefficient for both tests was found to be 0.95.

Conclusion

The results of this study provide stronger support for the idea that equating requirements were met in the first and second nationwide English language group tests in 2023. The difficulty difference between the two test forms was found to be relatively small. This study also indicates that the two requirements of measuring the same construct and maintaining equal reliability have been achieved. This helps to establish fairness and population invariance in the testing process.

These findings suggest that scores from both tests can be equated and used for making decisions about university applicants in Iran. Future research could focus on other test groups, especially the science test group, in order to further explore the validity and reliability of test equating in different subject areas. In addition, comparing the performance of different subgroups, such as place of residence (Tehran vs. provincial capitals), parents' education level, and school type, can provide further evidence for the requirement of testing the same construct across different subpopulations.

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

دوره ۱۵، شماره ۵۷، پاییز ۱۴۰۳، ۳۳-۵۳
ISSN: 2252-004x
eISSN: 2470-6240

نوع مقاله: مقاله پژوهشی
تاریخ دریافت: ۱۴۰۳/۰۱/۱۱؛ تاریخ بازنگری: ۱۴۰۳/۰۶/۰۴
تاریخ پذیرش: ۱۴۰۳/۰۶/۲۹؛ تاریخ انتشار: ۱۴۰۳/۰۷/۱۵

الزامات همترازسازی نمرات در نسخه‌های جایگزین آزمون‌های سرنوشت‌ساز: موردمطالعه آزمون سراسری داوطلبان زبان‌های خارجی

عبدالکریم شادمهر* *نویسنده مسئول*، استادیار، سازمان ملی سنجش و ارزشیابی نظام آموزش کشور، تهران، ایران. رایانامه: shadmehrabdolkarim@yahoo.com

عنایت‌اله زمانپور استادیار سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: zamanpour@atu.ac.ir

شعیب قاسمی دکتری سنجش و اندازه‌گیری، سازمان ملی سنجش و ارزشیابی نظام آموزش کشور، تهران، ایران. q.shoeayb@gmail.com

چکیده

برگزاری دو نوبت آزمون سراسری در یک سال مستلزم ساخت نسخه‌های جایگزین و همترازسازی نمرات آن‌ها است. همترازسازی نسخه‌های جایگزین نیازمند برقراری الزامات خاصی است. در آزمون‌های سرنوشت‌ساز به علت مسائل امنیتی امکان اجرای آزمایشی سؤال‌ها وجود ندارد. این مطالعه باهدف بررسی برقراری الزامات همترازسازی نسخه‌های جایگزین نوبت‌های اول و دوم آزمون سراسری گروه زبان خارجی انگلیسی سال ۱۴۰۲ انجام شد. پژوهش حاضر به لحاظ هدف کاربردی و از نوع روان‌سنجی که بر اساس تحلیل ثانویه داده‌ها انجام شده است. بدین منظور از بین کلیه داوطلبان گروه آزمایشی زبان انگلیسی سال ۱۴۰۲، از نوبت اول ۱۳۰۹۱ و از نوبت دوم ۱۹۹۱۵ داوطلب به‌صورت تصادفی انتخاب و داده‌های مربوط به آزمون‌های اختصاصی گروه آزمایشی زبان‌های خارجی (انگلیسی) اجراشده در دی‌ماه ۱۴۰۱ و تیرماه ۱۴۰۲ تحلیل شده است. میانگین ضرایب دشواری و تشخیص دو نوبت آزمون نزدیک به هم گزارش شده است. برقراری الزام سازهٔ یکسان دو نوبت آزمون، با نتایج تحلیل عاملی تک‌بعدی، مقایسه عملکرد زنان و مردان و رابطه نمره آزمون با معدل دبیرستان مورد تائید قرار گرفت. همچنین الزام پایایی یکسان نیز بررسی و میزان آن در هر دو نوبت آزمون ۰/۹۵ بود. هم‌ترازی نسخه‌های جایگزین مستلزم رعایت الزاماتی است که برآورد قضاوتی دشواری سؤال‌ها و استفاده از جدول مشخصات آزمون می‌تواند به طراحان در ساخت نسخه‌های جایگزین کمک کند. به‌منظور برقراری الزامات و همچنین طراحی سؤال‌هایی با دشواری مشابه رهنمودهایی ارائه شده است.

کلیدواژه‌ها: آزمون سراسری، آزمون سرنوشت‌ساز، نسخه‌های جایگزین، همترازسازی

استناد به این مقاله: شادمهر، عبدالکریم، زمانپور، عنایت‌اله، و قاسمی، شعیب. (۱۴۰۳). الزامات همترازسازی نمرات در نسخه‌های جایگزین آزمون‌های سرنوشت‌ساز: موردمطالعه آزمون سراسری داوطلبان زبان‌های خارجی. اندازه‌گیری تربیتی، ۱۵(۵۷)، ۳۳-۵۳. https://doi.org/10.22054/jem.2024.81723.3563

ناشر: دانشگاه علامه طباطبائی

مقدمه

مصوبه سیاست‌ها و ضوابط ساماندهی سنجش و پذیرش متقاضیان ورود به آموزش عالی (پس از پایان متوسطه) که در جلسه ۸۴۳ مورخ ۱۴۰۰/۰۴/۱۵ شورای عالی انقلاب فرهنگی به تصویب رسید، سازمان ملی سنجش و ارزشیابی نظام آموزش کشور را مکلف به برگزاری دو نوبت آزمون سراسری و یک مرحله پذیرش برای ورودی در سال ۱۴۰۲ به نظام آموزش عالی کشور کرده است. سازمان با توجه به الزام قانونی به برگزاری دو نوبت آزمون سراسری در سال، اقدام به ساخت نسخه‌های جایگزین[1] برای دو نوبت آزمون سراسری کرده است. در شرایطی که ایمنی و حفاظت آزمون دارای اهمیت باشد و وقتی‌که پرسش‌های آزمون و کلید آن باید به‌موجب قانون بعد از اجرای آزمون منتشر شود، لازم است هر زمان که آزمون به‌کار برده می‌شود نسخه جدیدی از آن تهیه شود. هر آزمودنی حق دارد انتظار داشته باشد نسخه خاصی از آزمون که او پاسخ داده است، تأثیر زیادی بر نمره او نداشته باشد. در این مورد خطاهای اندازه‌گیری که ناشی از تفاوت در محتوای نسخه‌های آزمون هستند دغدغه اصلی آزمودنی‌ها است (Crocker & Algina, 2008).

نسخه‌های مختلف از یک آزمون باید یک خصیصه یکسان را اندازه‌گیری کنند. سازندگان آزمون تمام تلاش خود را انجام می‌دهند تا نسخه‌های آزمون تا حد ممکن موازی[2] باشند. چنانچه نسخه‌های دو آزمون موازی باشند نیاز به همترازسازی نیست، اما تفاوت در دشواری نسخه‌ها اجتناب‌ناپذیر است (Livingston, 2014; Gonzalez & Wiberg, 2017؛ مقدم‌زاده، ۱۳۹۵). افزایش تفاوت دشواری نسخه‌های جایگزین، خطای روش‌های مختلف همترازسازی را بیشتر می‌کند (Sun & Kim, 2023). نسخه‌های جایگزین، دو نسخه یک آزمون هستند که یک سازه یکسان را اندازه می‌گیرند و در ساخت آن‌ها تلاش شده که موازی باشند و ممکن است میانگین، واریانس و همبستگی آن‌ها با سایر آزمون‌ها، یکسان (یا بسیار مشابه) باشد؛ علاوه بر این نسخه‌های جایگزین باید بر اساس محتوا و مشخصات روان‌سنجی یکسان ساخته شوند (Allen & Yen, 1979; Kolen & Brennan, 2014).

تهیه طرح کلی[3] آزمون یکی از مراحل اولیه در ساخت آزمون است. در آزمون‌های تحصیلی، طرح کلی یا جدول مشخصات آزمون یک جدول دوبعدی است که در یکی از

1. alternate forms
2. parallel
3. blueprint

دو بعد آن حوزه‌های اصلی محتوای موردِاندازه‌گیری و در بعد دیگر هدف‌های یادگیری یا فرایندهای شناختی قرار می‌گیرد (Crocker & Algina, 2008). ایجاد جدول مشخصات آزمون و وفاداری به آن کمک شایانی به همتایی[1] محتوای نسخه‌های جایگزین می‌کند. مشخصات آزمون عملکرد مهمی در اطمینان‌یافتن از این امر دارد که همه نسخه‌های یک آزمون همتا هستند. به این طریق همه آزمودنی‌ها، صرف‌نظر از این‌که به کدام نسخه از آزمون پاسخ دهند، پوشش مناسبی از محتوا را دریافت می‌کنند که تفسیرهای منصفانه و مناسب از نمرات آن‌ها امکان‌پذیر می‌شود. علاوه بر تناسب محتوایی آزمون، مشخصات روان‌سنجی مطلوب هم باید برقرار باشد. حداقل باید دشواری کلی آزمون و توزیع موردنظر برای دشواری‌ها و ضرایب تشخیص سؤال‌ها متناسب با طرح کلی مدنظر باشد. نسخه‌های تقریباً همتا با ساخت نسخه‌هایی بر اساس مشخصات محتوا و تطبیق ویژگی‌های روان‌سنجی به بهترین شکل قابل‌دستیابی است. در اصل اگر مشخصات محتوایی و روان‌سنجی به‌خوبی تعریف شده باشند و نیز بانک سؤال‌های مناسب در دسترس باشد، فرایند ساخت نسخه‌های جایگزین نسبتاً سرراست است. در اینجا باید هزینه-فایده مدنظر قرار گیرد به‌نحوی‌که ساده‌ترین راه برای دستیابی به نسخه‌های جایگزین این است که آزمون‌ها را بر اساس شرایط بازگو شده دقیقاً یکسان طراحی شوند. بسیاری از طراحان اعتقاد دارند که ایده‌آل این است که همه نسخه‌های آزمون تا حد ممکن شبیه ساخته شوند، تا این‌که اجرای هرکدام از نسخه‌های آزمون روی آزمودنی‌ها تفاوتی نداشته باشد؛ اما اگر آزمودنی‌هایی که آزمون اول را پاسخ داده‌اند، سؤال‌هایی را با سایر آزمودنی‌ها به اشتراک بگذارند، ممکن است نسخه‌هایی که بسیار به هم شبیه باشند موجب ایجاد مشکلات امنیتی شوند؛ بنابراین نسخه‌هایی که سؤال‌های یکسان یا بسیار شبیه به یکدیگر دارند، کمتر ایده‌آل خواهند بود. نسخه‌های مختلف یک آزمون باید شامل سؤال‌های متفاوتی باشند که مفاهیم یکسانی را اندازه می‌گیرند و مشخصات آزمون باید اجازه این‌گونه تمایزها را در بین نسخه‌ها بدهند (Wendler & Walker, 2015).

دو مرحله اساسی در فرایند ساخت آزمون، اجرای میدانی سؤال‌ها بر روی یک نمونه بزرگ که معرف جامعه آزمودنی‌های موردنظر آزمون و تعیین ویژگی‌های روان‌سنجی سؤال‌ها و حذف سؤال‌های است که ملاک‌های موردنظر را برآورد نمی‌کنند. به‌منظور لحاظ

1. equivalency

انصاف[1] درباره داوطلبانی که به نسخه‌های مختلف آزمون پاسخ می‌دهند لازم است دشواری نسخه‌ها یکسان باشد. چنانچه پارامترهای دشواری و قدرت تشخیص سؤال‌ها از طریق اجرای میدانی در دسترس باشند، می‌توان سؤال‌های نسخه‌های جایگزین را به‌گونه‌ای انتخاب کرد که کمترین تفاوت در دشواری آن‌ها وجود داشته باشد (Crocker & Algina, 2008; Schmeiser & Welch, 2006; Liang et al., 2021).

با توجه به اجتناب‌ناپذیری تفاوت دشواری نسخه‌های جایگزین، برای حذف اثرات تفاوت دشواری نسخه‌ها و رعایت انصاف و عدالت در تصمیمات حساس بر اساس نتایج آزمون، نیاز به همترازسازی[2] نمرات نسخه‌های مختلف وجود دارد. همترازسازی دربرگیرنده‌ی تعدیل‌های آماری کوچک است تا تفاوت‌های جزئی[3] در دشواری نسخه‌های جایگزین را جبران کند. پس از همترازسازی، نمرات نسخه‌های جایگزین را می‌توان به‌جای یکدیگر به کار برد (Aera, 2014; Kolen & Brennan, 2014؛ لرد، ۱۳۹۱/۱۹۸۰). این در حالی است که پنج الزام ضروری برای موفقیت در همترازسازی وجود دارد. این الزامات به ترتیب تاریخ پیدایش آن‌ها در ادبیات همترازسازی به‌صورت زیر هستند (Holland & Dorans, 2006):

الزام سازه یکسان[4]: آزمون‌ها باید سازه‌های یکسانی را اندازه بگیرند.

الزام پایایی یکسان: آزمون‌ها باید پایایی یکسانی داشته باشند.

الزام تقارن[5]: تابع همترازسازی برای همترازکردن نمرات Y با X باید معکوس تابع همترازکردن نمرات X با Y باشد.

الزام انصاف[6]: برای آزمودنی نباید تفاوتی وجود داشته باشد که با کدام‌یک از دو آزمونی که همتراز می‌شوند، آزمون شود.

الزام تغییرناپذیری جامعه[7]: انتخاب جامعه یا زیرجامعه‌ای که در برآورد تابع همترازسازی نمرات X و Y استفاده می‌شود نباید اهمیت داشته باشد، یعنی تابع همترازسازی استفاده‌شده برای پیونددادن نمرات X و Y باید نسبت به جامعه تغییرناپذیر باشد.

---

1. equity
2. equating
3. minor
4. equal construct requirement
5. symmetry
6. equity
7. population invariance

الزامات ۱ و ۲ به این معنی هستند که آزمون‌ها باید با مشخصات[1] یکسان ساخته شوند. این دو الزام برای اینکه نمرات نسخه‌های جایگزین قابل تعویض باشند، ضروری هستند (Kolen & Brennan, 2014). تشخیص سازه موردِاندازه‌گیری توسط یک آزمون به سادگی‌ای که به نظر می‌رسد نیست. بااین‌وجود رایج‌ترین راه برای بررسی کردن الزام اندازه‌گیری سازه یکسان توسط نسخه‌های مختلف، تعیین محتوا و جمله‌بندی[2] سؤال‌های آزمون است؛ به‌عبارت‌دیگر، این‌که آزمون‌ها ازنظر داوران ماهر[3] یک چیز را اندازه می‌گیرند. طبق تعریف، این موقعیت الزام سازه یکسان را برآورده می‌کند. یک روش تجربی ساده برای اثبات الزام سازه یکسان این است که دو آزمون، زیرگروه‌های مختلف از آزمودنی‌ها را ازنظر میانگین نمرات آن‌ها دقیقاً به یک شیوه مرتب کند. به‌علاوه می‌توان از روش‌های تحلیل عاملی برای تعیین صفات موردِاندازه‌گیری توسط دو آزمون استفاده کرد (Dorans & Holland, 2000). اگر دو آزمون سازه‌های متفاوتی را اندازه بگیرند، علاوه بر نقض الزام ۱، الزام ۴ نیز نقض می‌شود زیرا داوطلبان، آزمونی را انتخاب خواهند کرد که فکر می‌کنند در آن نمره بالاتری می‌گیرند. اگر دو آزمون یک سازه را اندازه بگیرد اما پایایی متفاوتی داشته باشند، آزمودنی‌های توانمندتر، آزمون با پایایی بالاتر را ترجیح می‌دهند، درحالی‌که آزمودنی ضعیف‌تر آزمونی را ترجیح خواهند داد که پایایی کمتری دارد. نقض الزام‌های ۱ و ۲ موجب می‌شود که روش‌های همترازسازی نتایجی به دست بدهند که الزام تغییرناپذیری در زیرگروه‌های خاصی از جامعه نیز برآورده نشود (Holland & Dorans, 2006).

همان‌طور که بیان شد همترازسازی برای جبران تفاوت جزئی دشواری سؤال‌های نسخه‌های جایگزین است. به‌منظور ساخت نسخه‌های جایگزین در آزمون‌های سرنوشت‌ساز که اهمیت بالایی داشته و بر روی تمام جنبه‌های زندگی داوطلبان تأثیر بسزایی دارد (بهمن‌آبادی و همکاران، ۱۴۰۳)، علاوه بر حفظ امنیت باید سؤال‌ها مطابق با استانداردها طراحی شود. در این شرایط برآورد دشواری سؤال‌ها توسط یک گروه از قضاوت‌کننده‌ها می‌تواند رهگشا باشد. طراحان آزمون یا داوران، در برخی موارد سؤال‌ها را به‌صورت آسان، متوسط و دشوار دسته‌بندی می‌کنند، یا مستقیماً مقادیر دشواری کلاسیک را برآورد می‌کنند (Thorndike, 1982/1996; Liang et al., 2021; Berenbon, 2023). داوران عمدتاً تمایل

1. specifications
2. wording
3. competent judge

به زیربرآورد دشواری سؤال‌های جدید دارند (Rezigalla, 2024). بر اساس نظر گودوین (Goodwin, 1996) تفاوت بین نظر قضاوت‌کنندگان با عملکرد واقعی آزمودنی‌ها، احتمالاً ناشی از این باشد که قضاوت‌کنندگان اغلب متخصصان آن حوزه هستند که تجربه، دانش و تحصیلات بالایی دارند، یا اینکه توقعات بسیار بالایی از آزمودنی‌ها دارند. پیشنهادهای مختلفی ازجمله آموزش مستمر طراحان و داوران (Kiessling et al., 2018 Ferrara et al., 2011) و نیز استفاده از سؤال‌هایی که قبلاً ویژگی‌های روان‌سنجی آن‌ها مشخص‌شده به‌عنوان سؤال‌های لنگر (Hambleton & jirka, 2006) به‌منظور افزایش دقت قضاوت دشواری بیان شده است.

سازمان سنجش در ساخت فرم‌های جایگزین، سطح دشواری قضاوتی سؤال‌ها را همراه با سطوح مهارت‌های شناختی و محتوای مورداندازه‌گیری لحاظ کرده است. به دلیل الزامات امنیتی، اجرای آزمایشی سؤال‌های آزمون سراسری و برآورد ویژگی‌های روان‌سنجی آن‌ها امکان‌پذیر نبوده است. سازمان با تکیه بر یافته‌های علمی در زمینه ساخت نسخه‌های جایگزین در آزمون‌های سرنوشت‌ساز، طراحان سؤال را در زمینه برآورد دشواری سؤال‌ها، آموزش داده و از دشواری‌های قضاوتی طراحان سؤال در جدول مشخصات آزمون استفاده کرده تا تفاوت دشواری نسخه‌ها به حداقل ممکن برسد و از برقراری الزامات همترازسازی اطمینان حاصل شود. به دلیل جدید بودن شرایط، تاکنون هیچ پژوهشی بر روی برقراری الزامات همترازسازی آزمون سراسری انجام نشده است. این پژوهش از طریق مقایسه شواهد روایی، پایایی و ضرایب دشواری سؤال‌های آزمون سراسری نوبت اول و دوم سال ۱۴۰۲ در گروه آزمایشی زبان‌های خارجی (زبان انگلیسی)، میزان موفقیت در برقراری الزامات همترازسازی نسخه‌های جایگزین را بررسی کرده است.

روش

این پژوهش یک تحقیق کاربردی از نوع روان‌سنجی است که بر اساس تحلیل داده‌های ثانویه انجام شده است. جامعه آماری این پژوهش شامل کلیه داوطلبان گروه آزمایشی زبان‌های خارجی (انگلیسی) آزمون سراسری سال ۱۴۰۲ در نوبت‌های اول (دی‌ماه ۱۴۰۱) و دوم (تیرماه ۱۴۰۲) است. نمونه تحلیل‌شده در نوبت اول ۱۳۰۹۱ داوطلب و در نوبت دوم ۱۹۹۱۵ داوطلب بود. ابزار اندازه‌گیری، آزمون اختصاصی گروه آزمایشی زبان انگلیسی نوبت اول و دوم بود. هر آزمون متشکل از ۷۰ سؤال چهارگزینه‌ای است که برای هر پاسخ

صحیح نمره ۱ و برای پاسخ نادرست و عدم پاسخ نمره صفر در نظر گرفته شده است. سؤال‌های آزمون دارای ۶ بخش دستور زبان (۱۵ سؤال)، واژگان (۲۰ سؤال)، ساختار جمله (۵ سؤال)، عملکردهای زبان (۵ سؤال)، کلوز تست (۱۰ سؤال) و درک خواندن (۱۵ سؤال) هستند. از روش‌های آماری ضریب دشواری کلاسیک، همبستگی دورشته‌ای نقطه‌ای اصلاح‌شده، ضریب پایایی ترکیبی، آزمون t مستقل، اندازه اثرهای d کوهن و r^2، تحلیل عاملی تأییدی و ضریب همبستگی اسپیرمن برای تحلیل داده‌ها استفاده شد. تحلیل‌های آماری با استفاده از نرم‌افزار R 4.3.3 و بسته‌های psych، lavaan و ggplot2 انجام شدند.

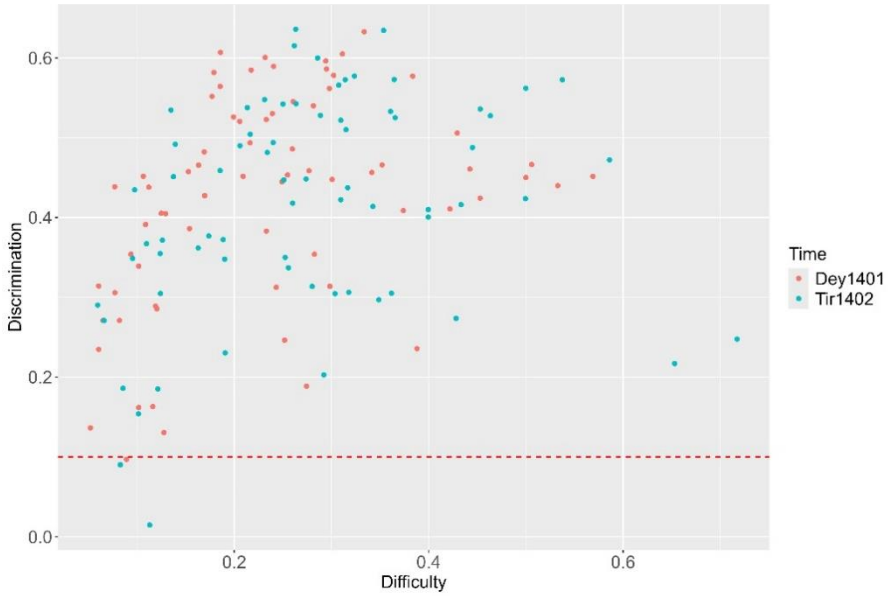
یافته‌ها

در مرحله اول شاخص‌های مربوط به تحلیل سؤال در نظریه کلاسیک شامل دشواری و همبستگی دورشته‌ای نقطه‌ای اصلاح‌شده محاسبه شدند (جدول ۱). ضرایب دشواری سؤال‌های نوبت اول در دامنه ۰/۰۵ و ۰/۵۷ و میانگین آن‌ها ۰/۲۴ بود. ضرایب دشواری سؤال‌های نوبت دوم در دامنه ۰/۰۶ و ۰/۷۲ و میانگین آن‌ها ۰/۲۸ بود. تفاوت میانگین ضرایب دشواری سؤال‌ها دو نوبت ۰/۰۴ با اندازه اثر d=۰/۳۲ و r^2=۰/۰۲۶ بود. مقادیر d کوهن ۰/۲، ۰/۵۰، و ۰/۸۰ و r^2، ۰/۰۴، ۰/۲۵ و ۰/۶۴ به ترتیب اثرهای کوچک، متوسط و بزرگ محسوب می‌شوند (Ferguson, 2009). میانگین قدرت تشخیص سؤال‌های نوبت اول ۰/۴۳ و نوبت دوم ۰/۴۲ بود.

جدول ۱. دشواری و تشخیص سؤال‌های آزمون زبان انگلیسی نوبت اول و دوم آزمون سراسری ۱۴۰۲

	نوبت اول		نوبت دوم	
	ضریب دشواری	قدرت تشخیص	ضریب دشواری	قدرت تشخیص
میانگین	۰/۲٤	۰/٤۳	۰/۲۸	۰/٤۲
انحراف معیار	۰/۱۳	۰/۱۳	۰/۱٤	۰/۱٤
حداقل	۰/۰۵	۰/۱۰	۰/۰٦	۰/۰۱
حداکثر	۰/۵۷	۰/٦۳	۰/۷۲	۰/٦٤

شکل ۱. نمودار پراکندگی دشواری و قدرت تمیز سؤال‌های گروه آزمون زبان انگلیسی در نوبت اول و دوم آزمون سراسری ۱۴۰۲



به‌منظور بررسی بیشتر توزیع ضرایب دشواری و قدرت تشخیص سؤال‌های دو نوبت آزمون، نمودار پراکندگی آن‌ها ترسیم شد که در شکل ۱ مشاهده می‌شود. در آزمون نوبت اول کمترین ضریب تشخیص ۰/۱۰ بود، اما در آزمون نوبت دوم قدرت تشخیص ۲ سؤال کمتر از ۰/۱۰ بود که در سطح نامطلوب است (Penfield, 2013). در هر دو نوبت آزمون، سؤال‌های با دشواری بین ۰/۲۰ تا ۰/۴۰ بیشترین فراوانی را دارند و سپس مقادیر دشواری کمتر از ۰/۲۰ قرار دارند. توزیع سؤال‌های دو نوبت آزمون در طول پیوستار دشواری و قدرت تشخیص مشابه است، به‌گونه‌ای که تقریباً در ازای هر سؤال در آزمون نوبت اول، یک سؤال مشابه در آزمون نوبت دوم وجود دارد. یک مورد استثنا در این خصوص، دو سؤال آزمون نوبت دوم با دشواری بیشتر از ۰/۶۰ هستند، درحالی‌که هیچ‌کدام از سؤال‌های آزمون نوبت اول دشواری بیشتر از ۰/۶۰ ندارد.

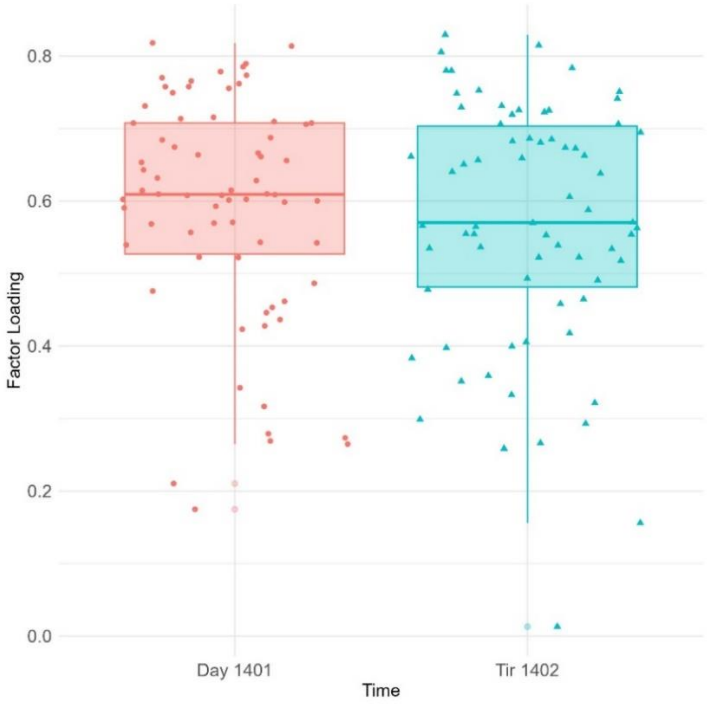
با توجه به نمره‌گذاری سؤال‌ها که به‌صورت صفر و یک است و اندازه نمونه، تحلیل عاملی تأییدی بر روی ماتریس همبستگی تتراکوریک و با روش برآورد WLSMV انجام شد (Viladrich et al., 2017). برای ارزیابی برازش مدل از چند شاخص برازش استفاده

شد. به دلیل این‌که آماره کای اسکوئر نسبت به اندازه نمونه بسیار حساس است و برازش دقیق را آزمون می‌کند که یک ملاک بسیار سخت‌گیرانه است، از شاخص برازش تطبیقی[1] (CFA)، شاخص توکر لوئیس[2] (TLI)، ریشه دوم میانگین مجذور خطای تقریب[3] (RMSEA)، و ریشه دوم میانگین مجذور باقیمانده‌های استانداردشده[4] (SRMR) استفاده شد. عموماً زمانی که CFI≥۰/۹۰، TLI≥۰/۹۰، RMSEA≤ ۰/۰۸ و SRMR≤۰/۰۸ باشند، برازش مدل مناسب تلقی می‌شود (Hu & Bentler, 1999).

جدول ۲. پایایی ترکیبی و برازش مدل یک‌بعدی با سؤال‌های نوبت اول و دوم آزمون سراسری گروه زبان انگلیسی ۱۴۰۲

نوبت آزمون	χ^2(df)	CFI	TLI	RMSEA	SRMR	پایایی ترکیبی
دی ۱۴۰۱	۲۹۲۰۱/۳۶(۲۳۴۵)	۰/۹۲	۰/۹۲	۰/۰۳۰	۰/۰۵۳	۰/۹۵
تیر ۱۴۰۲	۵۶۲۰۸/۴۳(۲۳۴۵)	۰/۹۱	۰/۹۱	۰/۰۳۴	۰/۰۵۶	۰/۹۵

شکل ۲. نمودار جعبه‌ای بارهای عاملی آزمون سراسری گروه زبان انگلیسی در نوبت‌های اول و دوم ۱۴۰۲



1. comparative fit index
2. Tucker Lewis index
3. Root mean square error of approximation
4. standardized root mean square residual

شاخص‌های برازش برای مدل تک‌بعدی نوبت اول و دوم در جدول ۲ ارائه شده است. آماره کای اسکوئر برای هر دو مدل در سطح ۰/۰۰۱ معنادار بود. با توجه به اندازه نمونه بسیار بزرگ، سایر شاخص‌ها موردتوجه قرار گرفتند که همه شاخص‌ها در سطح مطلوب بودند و مدل یک‌بعدی با داده‌های نوبت اول و نوبت دوم برازش مناسبی داشت. بارهای عاملی سؤال‌ها در نوبت‌های اول و دوم در نمودار جعبه‌ای شکل ۲ نشان داده شده است. در تحلیل عاملی، بارهای عاملی بیشتر از ۰/۳۰ قابل‌قبول در نظر گرفته می‌شوند (Hair et al., 2010). در هر دو نوبت ۶ سؤال بار عاملی کمتر از ۰/۳۰ داشتند. همان‌گونه که در نمودار جعبه‌ای مشخص است توزیع بارهای عاملی در دو نوبت مشابه یکدیگر است. میانه بارهای عاملی نوبت اول ۰/۶۱ و میانه بارهای عاملی نوبت دوم ۰/۵۷ است و دامنه میان‌چارکی نمودارهای جعبه‌ای دو نوبت نیز با هم همپوشانی بالایی دارد.

میانگین نمرات زنان و مردان در دو نوبت با آزمون t مقایسه شد (جدول ۳). به دلیل اندازه نمونه بسیار بزرگ، علاوه بر آزمون t مستقل، اندازه اثر d کوهن و r^2 نیز محاسبه شد. در نوبت اول میانگین‌های نمرات مردان (۱۹/۱۵) و زنان (۱۵/۰۲) تفاوت معناداری داشت (t=۱۶/۵۳، p<۰/۰۰۱). در آزمون نوبت دوم نیز میانگین مردان (۱۷/۸۱) بالاتر از میانگین زنان (۲۲/۵۸) بود (t=۲۲/۹۹، p<۰/۰۰۱). اندازه اثر تفاوت میانگین‌ها در نوبت اول d=۰/۳۲ و در نوبت دوم d=۰/۳۷ بود.

جدول ۳. مقایسه میانگین نمره کل زنان و مردان در نوبت اول و دوم آزمون گروه زبان انگلیسی ۱۴۰۲

نوبت آزمون	جنسیت	تعداد	میانگین	انحراف معیار	t	df	P-Value	Cohen's d
دی ۱۴۰۱	زن	۸۵۶۶	۱۵/۰۲	۱۱/۷۳	۱۶/۵۳	۷۷۲۱	۰/۰۰۱	۰/۳۲
	مرد	۴۵۲۵	۱۹/۱۵	۱۴/۴۶				
تیر ۱۴۰۲	زن	۱۳۰۴۸	۱۷/۸۱	۱۱/۹۹				
	مرد	۶۹۲۹	۲۲/۵۸	۱۴/۹۳	۲۲/۹۹	۱۱۷۵۶	۰/۰۰۱	۰/۳۷

یکی دیگر از شواهد روایی سازه که بررسی شد ارتباط نمره کل آزمون با معدل اعلامی داوطلبان بود. معدل اعلامی توسط داوطلبان در زمان ثبت‌نام در سیستم ثبت می‌شود، ازاین‌رو برخی از مقادیر معدل صحیح نیستند. در محاسبه همبستگی، داوطلبانی که معدل اعلامی آنان

کمتر از ۱۰ و بیشتر از ۲۰ بود از تحلیل کنار گذاشته شدند. ضریب همبستگی بین نمره آزمون زبان انگلیسی و معدل در هر دو نوبت اول و دوم ۰/۱۶ بود (p<۰/۰۰۱).

پایایی دو نوبت آزمون به روش پایایی ترکیبی[1] یا ضریب اومگا[2] محاسبه شد. پایایی ترکیبی بر اساس نسبت واریانس واقعی به‌دست‌آمده از پارامترهای مدل تحلیل عاملی تأییدی و مجموع واریانس‌ها و کوواریانس‌های سؤالات ضمنی مدل[3] محاسبه می‌شود. در مواردی که شرایط تائو-معادل[4] برقرار باشد، ضریب پایایی ترکیبی برابر با ضریب آلفای کرونباخ است. در شرایط آزمون متجانس[5]، آلفای ترکیبی در مقایسه با ضریب آلفای کرونباخ برآورد بهتری از پایایی واقعی است، زیرا محدودیت مفروضه یکسان بودن بارهای عاملی را ندارد (Raykov, 2001). ضریب پایایی ترکیبی آزمون‌های نوبت اول و نوبت دوم با هم برابر و ۰/۹۵ بود (جدول ۲).

بحث و نتیجه‌گیری

سازمان ملی سنجش و ارزشیابی نظام آموزش کشور، به‌منظور افزایش عدالت آموزشی بر اساس قانون موظف به برگزاری دو نوبت آزمون سراسری در سال شده است. زمانی که دو نسخه از یک آزمون ساخته می‌شوند همترازسازی نمرات آن‌ها ضرورت می‌یابد. اجرای همترازسازی نیازمند آن است که بین دو نسخه تفاوت دشواری اندکی وجود داشته باشد. بعلاوه سازه یکسان و پایایی یکسان هم جزو الزامات همترازسازی هستند. این دو الزام با برقراری الزام انصاف و الزام تغییرناپذیری جامعه مرتبط هستند. با توجه به این‌که آزمون سراسری یک آزمون سرنوشت‌ساز با حساسیت بسیار بالا است، این پژوهش به بررسی برقراری شرایط و الزامات همترازسازی در آزمون‌های نوبت اول و دوم سال ۱۴۰۲ در گروه آزمایشی زبان انگلیسی پرداخت.

اندازه اثر تفاوت میانگین دشواری دو نوبت آزمون یک تفاوت کوچک بود که سؤال‌های نوبت دوم اندکی ساده‌تر از نوبت اول بودند. این تفاوت کوچک، همترازسازی دو نوبت آزمون زبان انگلیسی را امکان‌پذیر می‌کند و خطای همترازسازی را کاهش می‌دهد. ساده‌تر بودن سؤال‌های نوبت دوم ممکن است تا حدودی از انگیزه بالاتر و افزایش توانایی

1. composite reliability
2. omega
3. model implied
4. tau-equivalent
5. congeneric

داوطلبان در نوبت دوم ناشی شود که موجب شده ضرایب دشواری کلاسیک به‌دست‌آمده در نوبت دوم بالاتر از نوبت اول باشد. میانگین قدرت تشخیص سؤال‌های دو نوبت نیز تفاوت بسیار اندکی داشت. پراکندگی سؤال‌ها دو نوبت آزمون ازنظر دشواری و قدرت تشخیص به‌گونه‌ای بود که تا حدود زیادی با استفاده از روش زیرمجموعه‌های تصادفی جورشده (Allen & Yen, 1979)، از معادل بودن دو نوبت آزمون اطمینان حاصل کرد.

الزام اندازه‌گیری سازه یکسان توسط دو نوبت آزمون با استفاده از شواهد چندگانه بررسی شد. در وهله نخست تعداد سؤال‌های دو نوبت آزمون و ساختار آن‌ها یکسان بود. تحلیل عاملی تک‌بعدی نشان داد که هر دو نوبت آزمون زبان انگلیسی دارای ساختار یک‌بعدی هستند. شاخص‌های برازش مدل تک‌بعدی با هر دو نوبت آزمون نزدیک به هم و مشابه بودند. توزیع بارهای عاملی دو نوبت آزمون نیز مشابه بود. در هر دو نوبت آزمون، میانگین نمره کل مردان بالاتر از زنان بود و اندازه اثرهای مربوط به تفاوت گروهی نیز نزدیک به هم و در سطح کوچک قرار داشتند. ضریب همبستگی هر دو نوبت آزمون با معدل یک همبستگی مثبت بود که با پژوهش‌های پیشین همخوان است (صادقی و همکاران، ۱۳۹۸). مقدار همبستگی برای دو نوبت آزمون یکسان بود. درمجموع، نتایج نشان داد الزام اندازه‌گیری سازه یکسان توسط دو نوبت آزمون سراسری زبان انگلیسی، بر اساس شواهد پیشنهادشده توسط Dorans and Holland (2000) برقرار است.

الزام پایایی یکسان دو نوبت آزمون با استفاده از ضریب پایایی ترکیبی بررسی شد. برای آزمون‌های سرنوشت‌ساز که در آن‌ها در مورد افراد تصمیم‌گیری انجام می‌شود ضرایب پایایی بالاتر از ۰/۹۰ پیشنهاد شده است (Nunnally, 1978). برخی از پژوهشگران برای اجتناب از حشو[1] و فدا نشدن روایی آزمون، حداکثر مقدار قابل‌قبول برای پایایی ترکیبی را ۰/۹۵ پیشنهاد کرده‌اند (Hair et al, 2021). میزان پایایی هر دو نوبت آزمون ۰/۹۵ با هم برابر بود که در سطح بالا و قابل‌قبول قرار دارد.

یافته‌های این پژوهش تا حدود زیادی از برقراری الزامات همترازسازی در آزمون‌های نوبت اول و دوم گروه آزمایشی زبان انگلیسی سال ۱۴۰۲ حمایت می‌کند. تفاوت دشواری دو نسخه در سطح کوچک قرار دارد. دو الزام اندازه‌گیری سازه یکسان و پایایی یکسان نیز برقرار هستند و برقراری این دو به برقراری الزام‌های انصاف و تغییرناپذیری جامعه کمک می‌کند. ازاین‌رو نمرات به‌دست‌آمده از دو فرم آزمون می‌توانند پس از همترازسازی به‌جای

1. redudency

یکدیگر به کار برده شوند. تصمیم‌گیری برای ورود داوطلبان به دانشگاه بر اساس نتایج این آزمون در ضمن برآورد کرده امنیت، عادلانه بودن را هم تأمین می‌کند. بر اساس نتایج این تحقیق و بررسی ادبیات تحقیق در زمینه برآورد ویژگی‌های سؤال‌های آزمون‌های سرنوشت‌ساز به روش قضاوتی، پیشنهادهای زیر برای کاهش تفاوت نسخه‌های جایگزین آزمون مطرح می‌شود:

۱- برخی از سؤال‌های آزمون گروه زبان انگلیسی ازنظر قدرت تشخیص در سطح مطلوب قرار نداشتند. تناسب نوع سؤال چندگزینه‌ای با محتوای موضوعی پیش‌بین‌کننده قدرت تشخیص سؤال است (Berenbon, 2023). به‌علاوه ایرادات رایج در سؤالات چندگزینه‌ای به سؤالاتی منتهی می‌شود که قدرت تشخیص کمتری دارند (Tarrent & Ware, 2008). بررسی کردن این موارد در جلساتی که توسط طراحان و متخصصان برای برآورد دشواری سؤال‌ها برگزار می‌شود، در بهبود قدرت تشخیص آن‌ها مفید خواهد بود.

۲- جلسات ارزیابی و برآورد سؤال‌ها به‌صورت گروهی برگزار شود (Hambleton & jirka, 2006/2014).

۲۳ در صورت امکان یک گروه قضاوت‌کننده متشکل از دانشجویان مقطع کارشناسی که مسلط به حیطه موضوعی هستند تشکیل شود. ترکیب برآوردهای این گروه با برآوردهای گروه معلمان و طراحان، به کاهش سوگیری برآورد دشواری کمک می‌کند (van de Watering & van der Rijt, 2006).

۴- آموزش مستمر قضاوت‌کنندگان در خصوص عوامل مؤثر بر سطح دشواری سؤال می‌تواند به بهبود عملکرد آنان منجر شود. برخی از این عوامل عبارت‌اند از: لغات و عبارات منفی، ارجاعات، واژگان، طول جمله یا پاراگراف، انتزاعی بودن متن، جایگاه متن سؤال، سطوح و تعداد مهارت‌های شناختی لازم برای پاسخگویی به سؤال، و نزدیکی گزینه‌های انحرافی به گزینه صحیح (Kiessling et al., 2018; Ferrara et al., 2011).

۵- استفاده از سؤالات لنگر[1]. یک دسته سؤال لنگر پنهان که ضریب دشواری آن‌ها از طریق اجرای آزمایشی به دست آمده است همراه با سؤال‌ها مورد ارزیابی، به اعضای پنل داده شود. ضرایب دشواری این سؤال‌ها لنگر در اختیار اعضای پنل قرار نمی‌گیرد و پس از برآورد دشواری سؤال‌ها توسط اعضای پنل، از اطلاعات این سؤال‌ها برای اصلاح سوگیری ارزیابی اعضای پنل استفاده شود. سؤال‌های لنگر آشکار نیز می‌توانند استفاده شوند. این‌ها

1. anchor

سؤال‌هایی هستند که ضریب دشواری آن‌ها از قبل مشخص است و در زمان برآورد دشواری سؤال‌ها توسط اعضای پنل در اختیار آن‌ها قرار داده می‌شوند (Hambleton & jirka, 2006/2014).

بر اساس قانون نتایج آزمون سراسری برای دو سال دارای اعتبار است. این امر به آن معنا است که ۴ نسخه مورداستفاده در هر دو سال باید دارای الزامات همترازسازی باشند. انجام پژوهش بر روی ۴ نوبت آزمون که برای پذیرش داوطلبان در یک سال دارای اعتبار هستند، در اطمینان از عادلانه بودن و صحت فرایند همترازسازی مفید خواهد بود. در اختیار داشتن نتایج چهار نوبت آزمون، این امکان را فراهم می‌کند تا با استفاده از روش‌های مبتنی بر تحلیل عاملی، تائو-معادل یا متجانس بودن چهار نوبت آزمون را بررسی کرد (Graham, 2006). در این پژوهش تنها یکی از گروه‌های آموزشی پنج‌گانه موردبررسی قرار گرفت. محدودیت دیگر پژوهش این بود که در مقایسه عملکرد زیرگروه‌های مختلف برای اطمینان از اندازه‌گیری سازه یکسان توسط دو نوبت آزمون، فقط تفاوت ازنظر جنسیت بررسی شد. پژوهش‌های آینده می‌توانند بر روی سایر گروه‌های آزمایشی، به‌ویژه گروه آزمایش تجربی متمرکز شوند. همچنین مقایسه عملکرد زیرگروه‌های مختلف مانند محل سکونت (تهران، مرکز استان و شهرستان)، تحصیلات والدین و نوع مدرسه می‌تواند شواهد بیشتری برای الزام سازه یکسان فراهم کند.

تعارض منافع

نویسندگان هیچ‌گونه تعارض منافعی ندارند.

سپاسگزاری

از کلیه کارکنان سازمان ملی سنجش و ارزشیابی نظام آموزش کشور که ما را در این پژوهش یاری کردند سپاسگزاری می‌شود.

منابع

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

صادقی، میثم، فلسفی نژاد، محمدرضا، دلاور، علی، فرخی، نورعلی و جمالی، احسان. (۱۳۹۷). تأثیر مدل وزن دهی و نمره کل سازی سوابق تحصیلی بر کارایی گزینش داوطلبان ورود به دانشگاه‌ها و مراکز آموزش عالی کشور. *پژوهش در نظام‌های آموزشی*، ۱۲(ویژه‌نامه), ۲۷-۴۳.

لرد، فردریک. ام. (۱۳۹۱). *کاربردهای نظریه سؤال-پاسخ* (دلاور، علی و یونسی، جلیل، مترجمان). تهران، انتشارات رشد. (انتشار نسخه اصلی، ۱۹۸۰)

مقدم‌زاده، علی. (۱۳۹۵). روش بهینه هموارسازی داده‌ها در همترازسازی: موردمطالعه آزمون تولیمو و آزمون‌های جامع آزمون‌های آزمایشی سازمان سنجش آموزش کشور. *فصلنامه اندازه‌گیری تربیتی*، ٦ (۲۲)، ۲٦۱-۲۸۷.

References

Aera, A. P. A. (2014). Standards for educational and psychological testing. *New York: American Educational Research Association*.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, California: Brooks/Cole Publishing Company.

Bahmanabadi, S., Falsafinejad, M., 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. [in Persian]

Berenbon, R. F., & McHugh, B. C. (2023). Do Subject Matter Experts' Judgments of Multiple-Choice Format Suitability Predict Item Quality?. *Educational Measurement: Issues and Practice*, *42*(3), 13-21.

Crocker, L.M., Algina, J. (2008). Introduction to Classical and Modern Test Theory. Cengage Learning.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of educational measurement*, *37*(4), 281-306.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test development with performance standards and achievement growth in mind. Educational Measurement: Issues and Practice, 30(4), 3–15. https://doi.org/10.1111/j.1745-3992.2011.00218.x

González Burgos, J. A., & Wiberg, M. (2017). Applying test equating methods, using R.

Goodwin, L. D. (1996). Focus on quantitative methods: Determining cut-off scores. Research in Nursing & Health, 19, 249–256.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. Educational and Psychological Measurement, 66(6), 930-944.

Hair J.F., Jr., Black W.C., Babin B.J., Anderson R.E. *Multivariate Data Analysis.* 7th ed.

Hair Jr, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook* (p. 197). Springer Nature.

Hambleton, R. K., & Jirka, S. J. (2014). Anchor-based methods for judgmentally estimating item statistics. In *Handbook of test development* (pp. 413-434). Routledge.

Holland, P. W., & Dorans, N. J. (2006).Linking and equating. *Educational measurement*, *4*, 187-220.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1-55. https://doi.org/10.1080/10705519909540118

Kiessling, C., Lahner, F.-M., Winkelmann, A., & Bauer, D. (2018). When predicting item difficulty, is it better to ask authors or reviewers? Medical Education, 52(5), 571–572. https://doi.org/10.1111/medu.13570

Kolen, Michael & Brennan, Robert. (2014). Test equating, scaling, and linking. Methods and practices. 3rd revised ed. 10.1007/978-1-4939-0317-7.

Liang, Z., Zhang, M., Huang, F., Kang, D., & Xu, L. (2021). Application innovation of educational measurement theory, method, and technology in China's New College Entrance Examination Reform. *Chinese/English Journal of Educational Measurement and Evaluation*, *2*(1), 3.

Livingston, S. A. (2014). Equating test scores (without IRT). *Educational testing service*.

Lord, F. M. (2012). Applications of Item Response Theory to Practical Testing Problems (Delavar, A., Younesi, J, Trans). Tehran, Roshd publication. (Original work published 1980). [in Persian]

MoghadamZade, A. (2015). Optimal Smoothing Method of Data in Test Equating: The Case of TOLIMO and Comprehensive Trial Tests of Iran Educational Testing Organization. Quarterly of Educational Measurement, 6(21), 261-287. [in Persian]

Nunnally, J. C. (1978). Psychometric theory. New York, NY: McGrawHill.

Penfield, R. D. (2013). Item analysis. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 121–138). American Psychological Association. https://doi.org/10.1037/14047-007

Raykov, T. (2001). Bias of coefficient α fixed congeneric measures with correlated errors. *Applied psychological measurement*, *25*(1), 69-76.

Rezigalla, A. A. (2024). AI in medical education: uses of AI in construction type A MCQs. *BMC medical education*, *24*(1), 247.

Sadeghi, M., Falsafinezhad, M., Delavar, A., Farrokhi, N; & Jamali, E (2018). The effect of the weighting model and the composite score of academic records on the efficiency of selecting candidates to enter the universities and higher education centers of the country. *Journal of Research in Educational Systems*, 12, (Special Issue), 27-43. [in Persian]

Schmeiser, C. B., & Welch, C. J. (2006). Test development. *Educational measurement*, *4*, 307-353.

Sun, T., & Kim, S. Y. (2023). Evaluating Equating Methods for Varying Levels of Form Difference. *Educational and Psychological Measurement*, 00131644231176989.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple choice questions on student achievement in high-stakes nursing assessments: Item-writing flaws and student achievement. Medical Education, 42(2), 198–206. https://doi.org/10.1111/j.1365-2923. 2007.02957.

Thorndike, R, L. (1996). Apllied psychometrics (Hooman, H,A, Trans). Houghton Mifflin School. (Original work published 1982)

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. Educational Research Review, 1(2), 133–147. https://doi.org/10.1016/j.edurev.2006.05.001

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de psicología*, *33*(3), 755-782.

Wendler, C. L., & Walker, M. E. (2015). Practical issues in designing and maintaining multiple test forms. In *Handbook of test development* (pp. 433-449). Routledge.