Vol 16, Issue 60, 2025, 83-108
eISSN: 2470-6240 ISSN: 2252-004x
Received April 14, 2024 Received in revised form January 15, 2025
Published Online June 25, 2025 Accepted May 31, 2025
Article Type: Reserch Article

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

Utilizing Simple Random Sampling (SRS) in Classical Test Theory for Large-Scale Assessments: A Time-Efficiency Approach

Behrooz Kavehie *

Corresponding Author, Department of Statistics, National Organization for Educational Testing, Tehran, Iran. E-mail: kavehiebehrooz@Yahoo.com

Abstract

This study proposes a novel method to significantly reduce data processing time for calculating facility (P) and discrimination indices in large-scale assessments. By integrating Simple Random Sampling (SRS) with Classical Test Theory (CTT), we developed a computational approach that minimizes processing time while maintaining statistical accuracy. For illustration, a test question requiring 7,393 minutes (5.5 days) for full analysis yielded a P-index of 0.50. Our SRS-CTT method achieved comparable results (P-index = 0.503) in just 129.2 minutes (mean across 10 iterations), representing a 98.3% reduction in processing time. The accompanying computer program enables efficient analysis of national-level datasets, making this method particularly valuable for high-stakes testing organizations. We strongly recommend its adoption for nationwide educational assessments.

Keywords: Classical Test Theory; Large-Scale Dataset; Process Time; Simple Random Sampling

Cite this Article: Kavehie, B. (2025). Utilizing Simple Random Sampling (SRS) in Classical Test Theory for Large-Scale Assessments: A Time-Efficiency Approach. *Educational Measurement*, *16*(60), 83-108. https://doi.org/10.22054/jem.2025.79211.3533

Publisher: Allameh Tabataba'i University Press

Introduction

Within Iran's higher education system, undergraduate admissions are regulated by parliamentary legislation and administered through a centralized national entrance examination. The National Organization for Educational Testing (NOET) conducts this high-stakes assessment annually at predetermined dates, with university placements determined solely by examination performance. The current system distributes approximately 205,869 available seats across diverse academic disciplines annually.

The 2023 administration of Iran's national university entrance exam highlights the scale of this high-stakes assessment. Within just the experimental sciences group (medical fields), 484,576 candidates competed for 60,968 available seats. With each examination comprising 150 test items, the resulting dataset presents significant analytical challenges.

While calculating standard psychometric indices (e.g., Facility and Discrimination indices) remains computationally simple for small datasets, national-level analysis introduces substantial complexity. Processing approximately 500,000 examinees' responses (representing 75 million data points) demand:

1. Advanced computational methods
2. Optimized statistical procedures
3. Efficient data processing frameworks

This scale of analysis necessitates innovative approaches to maintain both accuracy and timeliness in test evaluation.

Research Question(s)

Can Simple Random Sampling (SRS) significantly accelerate Classical Test Theory (CTT) analyses for large-scale testing datasets while maintaining statistical fidelity?

Literature Review

Classical Test Theory (CTT) is built upon several fundamental psychometric developments, including:

1. The classical true score model (Lord & Novick, 1968)
2. Observed score (X), true score (T), and error score (E) random variables, where $X = T + E$
3. Reliability coefficients (e.g., Cronbach's α, split-half methods)
4. Standard error of measurement (SEM) and its role in score interpretation

5. Test length effects on score precision and reliability
6. Single-administration reliability estimators (e.g., Kuder-Richardson formulas)
7. Models of score equivalence (e.g., parallel, tau-equivalent tests)

These constructs form the theoretical backbone of CTT and remain widely applied in large-scale assessments (Algina & Swaminathan, 2015; Feldt & Brennan, 1989).

In both criterion-referenced and norm-referenced testing, post-administration item analysis represents a critical phase of test evaluation. This process systematically examines key psychometric properties-including item difficulty (P-index), discrimination power (D-index), and distractor effectiveness—to identify potential flaws in question design such as ambiguous wording or poor discriminatory capability. Through such analysis, testing organizations gain empirical insights that not only reveal the scientific robustness of individual items but also inform iterative improvements to test quality. The resulting data enables evidence-based refinements to item construction, test blueprints, and scoring protocols, particularly crucial for high-stakes assessments where validity and reliability directly impact score interpretations. This cyclical evaluation process ensures continuous enhancement of measurement precision and fairness in subsequent test administrations (Haladyna & Rodriguez, 2013).

Given that test publishers possess the financial and technical resources to conduct comprehensive item analyses, this process has become integral to the development of large-scale standardized assessments, where rigorous psychometric evaluation is essential. Beyond its role in test refinement, item analysis serves as a valuable diagnostic tool for educators by revealing patterns of student misconceptions—when a significant proportion of examinees select an incorrect response, it may indicate either flaws in item design or gaps in instructional delivery. This dual function makes item analysis critical for both improving assessment quality and identifying areas needing pedagogical intervention, particularly in standardized testing contexts where the stakes are high and the impact of measurement errors is substantial.

The discriminant index method offers a practical solution for question analysis, particularly for educators lacking access to automated scoring systems or computational tools. This technique generates a numerical discriminant index (D) for each question, ranging

Homepage: https://jem.atu.ac.ir/

from -1 to +1, which serves as a straightforward metric for evaluating item effectiveness. A higher positive D-value indicates stronger alignment between the individual question and the overall test's measurement objectives, with values approaching +1 suggesting that high-performing test-takers answered correctly while low-performing test-takers answered incorrectly. This simple yet informative measure enables teachers to identify questions that effectively discriminate between students of differing ability levels without requiring sophisticated analytical resources.

The discriminant analysis method operates on the premise that students who answer one item correctly are more likely to answer other items correctly, reflecting overall test consistency. The procedure begins by ranking all students according to their total test scores. The top 27% of performers (high-scoring group) and bottom 27% (low-scoring group) are then selected for comparison. This 27% threshold has been empirically shown to optimally balance two competing needs: (1) maintaining sufficiently large group sizes to ensure stable estimates, while (2) maximizing performance differences between groups to enhance discriminant sensitivity. For practical classroom applications with 25-35 students, using 10 students per group (approximately 28-40% of the class) provides a reasonable compromise between computational simplicity and statistical reliability of the resulting discriminant indices (Cunningham, 1998). This approach remains robust even without sophisticated analytical tools, making it particularly valuable for teacher-led assessments.

The difficulty level of a test question, commonly referred to as the P-value (or Facility Index), is simply the proportion of examinees who answer it correctly. This metric ranges from 0 to 1, where a higher value indicates an easier question. For example, if 60% of test-takers answer a question correctly, its P-value would be 0.60. Conversely, a P-value of 0.30 would indicate a relatively difficult question that only 30% of respondents answered properly. This straightforward measure provides educators with immediate insight into question performance, where values around 0.50 typically represent questions of moderate difficulty that effectively discriminate between different ability levels.

Question difficulty, expressed as the P-value (Facility Index), ranges from 0 to 1, where lower values indicate more challenging items. For instance:

- An extremely difficult question (P = 0.10) would be answered correctly by only 10% of examinees
- An exceptionally easy question (P = 0.90) would be answered correctly by 90% of examinees

The discriminant power (D-index) of a question reflects its ability to differentiate between high-performing and low-performing test-takers. Importantly, this conceptual framework assumes that the underlying construct being measured exists on a continuum rather than as a dichotomy. For example, when assessing psychological traits like depression, we recognize a spectrum of severity rather than simply categorizing individuals as depressed or non-depressed. This continuous-variable assumption is crucial for proper interpretation of both difficulty and discrimination indices in psychological and educational measurement.

Multiple methods exist for calculating a question's discriminant power (D-index), with most following similar computational logic (Oosterhof, 1976). The fundamental approach involves comparing response patterns between high-performing and low-performing examinees for each test item. Consider a mathematics assessment administered to 100 students, where each student's performance is recorded both as a total score and per-item responses.

The critical methodological decision involves defining the "high" and "low" performance groups. Kelley's (1939) seminal work demonstrated that selecting the top and bottom 27% of performers optimizes discriminant sensitivity, though marginal variations (25-30%) yield negligible differences in practice. This approach aligns with Rehfisch's (1958) findings using 25% groupings in strength scale analyses.

Note that both facility (P) and discriminant (D) indices are conventionally expressed as decimal values ranging from 0 to 1, despite sometimes being discussed in percentage terms (e.g., P = 0.75 described as "75%").

While this discussion focuses on Classical Test Theory (CTT) applications, it is important to acknowledge its methodological limitations. Recent scholarship (Adekunle et al., 2022) has specifically recommended that the Examination Council of Lesotho (ECOL) transition from CTT to modern test theory frameworks like Item Response Theory (IRT) for test development and item analysis. Such contemporary approaches offer distinct advantages, including sample-

independent item parameters and more precise ability estimation, particularly for high-stakes assessments. These considerations, however, extend beyond the scope of the current examination of CTT-based methods.

Methodology

In the CTT method, analysis typically uses a subset of data comprising the top 27% and bottom 27% of students based on test scores. This paper proposes an alternative approach: instead of simple random sampling, extracting just 1% from each of these high and low performance groups for calculations. This method reduces the data processing volume to a manageable size, achieving a 99% decrease in computational operations - while maintaining necessary measurement accuracy.

Conclusion

The results show that for calculating the Facility index, using a simple random sampling method with 1% of the 27% data yields the same results as standard calculations. Similarly, the Discriminant index calculations show identical results to standard methods. Therefore, to save processing time for extracting classical psychometric indices, results from just 1% of candidates can be used for question analysis and decision-making. These findings demonstrate approximately 99% savings in data processing time.

Acknowledgments

The authors extend their sincere gratitude to the esteemed leadership and dedicated staff of the National Organization for Educational Testing (NOET) for their generous provision of test data and invaluable support throughout this research endeavor. Their contribution was essential to the successful completion of this study.

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

دوره ۱۶، شماره ۶۰، ۱۴۰۴، ۸۳-۱۰۸
eISSN: 2470-6240 ISSN: 2252-004x
نوع مقاله: مقاله پژوهشی
تاریخ دریافت: ۱۴۰۳/۰۱/۲۶؛ تاریخ بازنگری: ۱۴۰۳/۱۰/۲۶
تاریخ پذیرش: ۱۴۰۴/۰۳/۱۰؛ تاریخ انتشار: ۱۴۰۴/۰۴/۰۱

استفاده از نمونه‌گیری تصادفی ساده در تحلیل کلاسیک آزمون برای داده‌های بزرگ با هدف کاهش زمان پردازش

بهروز کاوه‌ئی*

نویسنده مسئول، گروه آمار، سازمان سنجش آموزش کشور، تهران، ایران. رایانامه: KavehieBehrooz@Yahoo.com

چکیده

هدف این پژوهش یافتن یک روش جدید برای کاهش زمان پردازش داده‌ها و محاسبه شاخص آسانی و شاخص تمیز سؤالات در آزمون‌های بزرگ است. برای این منظور از روش نمونه‌گیری تصادفی ساده در محاسبه شاخص آسانی و شاخص تمیز در روش تحلیل کلاسیک آزمون و تولید برنامه کامپیوتری برای اجرای نمونه آزمایشی با داده‌های بزرگ استفاده شده است. نتایج نشان داد که استفاده از این روش باعث کاهش زمان برآورد پارامترها می‌شود. برای مثال، سؤالی که با صرف ۷۳۹۳ دقیقه (بیش از ۱۳۳ ساعت – ۵ شبانه‌روز) مقدار شاخص P را برابر ۰/۵۰ به دست آورده بود، با ۱۰ بار تکرار و صرف زمان متوسط ۱۲۹/۲ دقیقه، مقدار متوسط این شاخص را برابر ۰/۵۰۳ محاسبه کرد. درنتیجه ترکیب روش نمونه‌گیری تصادفی ساده با روش تحلیل کلاسیک آزمون، باعث کاهش چشمگیر زمان پردازش داده‌ها شده و لذا استفاده از این روش به سازمان‌های بزرگ مانند سازمان‌هایی که آزمون‌هایی را در سطح ملی برگزار می‌کنند قویاً توصیه می‌شود.

کلیدواژه‌ها: تحلیل کلاسیک آزمون، داده‌های بزرگ، زمان پردازش، نمونه‌گیری تصادفی ساده

استناد به این مقاله: کاوه‌ئی، بهروز. (۱۴۰۴). استفاده از نمونه‌گیری تصادفی ساده در تحلیل کلاسیک آزمون برای داده‌های بزرگ با هدف کاهش زمان پردازش. *فصلنامه اندازه‌گیری تربیتی*، ۱۶(۶۰)، ۸۳-۱۰۸.
https://doi.org/10.22054/jem.2025.79211.3533

ناشر: دانشگاه علامه طباطبائی

مقدمه

در نظام آموزش عالی ایران، بر اساس قانون مجلس شورای اسلامی، آزمون ورودی دانشگاه‌ها و موسسه‌های آموزش عالی برای ورود به مقطع کارشناسی، در یک‌زمان مشخص از سال و به‌صورت متمرکز توسط سازمان سنجش آموزش کشور[1] (NOET) برگزار می‌شود و برگزیدگان این آزمون ملی برای ورود به دانشگاه‌ها معرفی می‌شوند. درمجموع و در همه رشته‌های امتحانی حدود ۲۰۵۸۶۹ جایگاه (صندلی) برای تصرف داوطلبان وجود دارد. تعداد داوطلبان این آزمون تنها در گروه آزمایشی علوم تجربی در سال ۱۴۰۰ برابر ۴۸۴۵۷۶ نفر بوده است[2] که این تعداد برای تصاحب ۶۰۹۶۸ صندلی با هم رقابت کرده‌اند. به‌علاوه تعداد سؤالات پرسش شده از هر داوطلب در این آزمون ۱۵۰ سؤال است.

انجام محاسبات مربوط به شاخص دشواری و شاخص تمیز سؤالات، زمانی که پردازش داده‌ها برای تعداد محدودی از شرکت‌کنندگان در آزمون موردنظر باشد کار دشوار و زمان‌بری نیست. مشکل آنجا ظاهر می‌شود که یک آزمون بزرگ در سطح ملی اجرا شده و باید تحلیل آزمون برای ۵۰۰ هزار شرکت‌کننده انجام شود.

با توجه به فرمول‌های محاسبه شاخص‌های P و D مربوط به روش تحلیل آزمون کلاسیک داریم:

- حداقل تعداد داوطلب لازم، ۲۷ درصد از داوطلبان قوی (۱۳۰۸۳۵ نفر) و ۲۷ درصد از داوطلبان ضعیف (۱۳۰۸۳۵ نفر). درمجموع می‌شود ۲۶۱۶۷۰ داوطلب.
- مرتب‌سازی نزولی (یا صعودی) همه داوطلبان ۴۸۴۵۷۶ نفر بر اساس نمره کل.
- جداسازی ۱۳۰۸۳۵ نفر از بالای فهرست (داوطلبان قوی).
- جداسازی ۱۳۰۸۳۵ نفر از پایین فهرست (داوطلبان ضعیف).
- محاسبه میانگین نمره ۱۳۰۸۳۵ نفر داوطلبان قوی.
- محاسبه میانگین نمره ۱۳۰۸۳۵ نفر داوطلبان ضعیف.
- محاسبه واریانس (و انحراف معیار) نمره ۱۳۰۸۳۵ نفر داوطلبان قوی.
- محاسبه واریانس (و انحراف معیار) نمره ۱۳۰۸۳۵ نفر داوطلبان ضعیف.
- محاسبات شاخص‌های P و D با ایجاد حلقه‌های محاسباتی تودرتو به شکل و کران‌های زیر:

1. National Organization for Educational Testing (NOET).

۲. گزارشات رسمی سایت سازمان سنجش آموزش کشور در سال ۱۴۰۰ خورشیدی.

حلقه داوطلبان :For I = ۱ to ۲۶۱,۶۷۰

حلقه پاسخ‌های داوطلبان :For J = ۱ to ۳۹,۲۵۰,۵۰۰ (۱۵۰*۲۶۱۶۷۰)

کل حلقه‌های اجرایی (حاصل‌ضرب دو حلقه): ۱۰,۲۷۰,۶۷۸,۳۳۰,۰۰۰ چرخش!

- تولید گزارش خروجی.

محاسبه ردیف‌های بالا با یک کامپیوتر پرقدرت با مشخصات زیر برای هر سؤال حدود ۱۲۳:۱۳ ساعت (۷۳۹۳ دقیقه) زمان مصرف کرده است:

CPU: Intel® Core(TM) I7-7700 @4.2 GHz, 4.2 GHz.
Ram: 16 GB DDR2.
HDD: SSD 1TB (High Speed).

درصورتی‌که زمان اجرای همه ۱۵۰ سؤال آزمون محاسبه شود باید عددی حدود ۱۸۴۵۰ = ۱۲۳* ۱۵۰ ساعت وقت صرف شود.

همین محاسبات با روش پیشنهادی در این نوشتار (استفاده از روش‌های نمونه‌گیری آماری) برای یک سؤال (که در حالت عادی ۱۲۳:۱۳ ساعت زمان لازم داشت) به‌طور متوسط حدود ۱۲۹ دقیقه (حدود ۲ ساعت و ۹ دقیقه) زمان برده است. شایان‌ذکر است که روش پیشنهادی مبتنی بر روش‌های نمونه‌گیری آماری بوده است که خود مبتنی بر احتمال است و معمولاً دقت روش‌های آماری را حدود ۹۵ درصد در نظر می‌گیرند. بااین‌حال اگر به مسئله تحلیل هزینه-فایده[1] توجه کنیم، به نظر می‌رسد استفاده از روش پیشنهادی بسیار مقرون‌به‌صرفه و قابل‌قبول است.

در چنین محاسباتی برای کاهش زمان باید یکی از روش‌های زیر را بکار بست:

افزایش تعداد کامپیوترها و پردازش موازی روی چند دستگاه،

استفاده از روش نمونه‌گیری پیشنهادی در این نوشتار.

برنامه‌نویسان کامپیوتر که تجربه پردازش داده‌های بزرگ را دارند به‌خوبی می‌دانند که از مشکلات دیگر پردازش داده‌های بزرگ در کامپیوترها این است که زمان لازم برای پردازش دقیقاً از حالت ضرب استفاده نمی‌کند و به‌صورت تصاعدی افزایش پیدا می‌کند؛ یعنی لزوماً برای ۱۵۰ سؤال (هر سؤال ۱۲۳ ساعت)، عددی برابر ۱۸۴۵۰ ساعت رخ نمی‌دهد و احتمالاً زمان شگفت‌انگیز بیشتری لازم است.

البته این حجم از داده‌ها تنها مربوط به بخش مجموعه رشته علوم تجربی بود. گروه‌های آزمونی دیگری نیز در این آزمون وجود دارند. درمجموع در این آزمون داوطلبان در

1. Cost-Benefit Analysis.

گروه‌های پنج‌گانه، گروه علوم ریاضی و فنی، گروه علوم تجربی (علوم پزشکی)، گروه علوم انسانی، گروه زبان‌های خارجی و گروه هنر شرکت می‌کنند. توضیح اینکه برخی داوطلبان در دو یا سه گروه آزمون می‌دهند. طبق اعلام رسمی سازمان برگزارکننده این آزمون بزرگ، این آزمون درمجموع برای ۵ گروه آزمونی حدود ۱,۳۶۷,۹۳۱ نفر داوطلب دارد. البته برای محاسبات تحلیل کلاسیک آزمون، تعداد داوطلبان غایب باید از تعداد اعلام‌شده کم شود.

این سازمان برای آزمون‌های ورودی به دوره‌های تحصیلات تکمیلی و آزمون‌های استخدامی نیز بیش از ۲ میلیون نفر را آزمون می‌کند. اکنون لزوم این پژوهش بیش‌ازپیش مشخص شد. از اصلی‌ترین علت‌های اجرای فرآیندهای نمونه‌گیری می‌توان به مواردی چون کاهش زمان انجام مطالعه، کاهش هزینه انجام مطالعه، جلوگیری از تخریب آزمودنی‌ها و غیره اشاره کرد. البته موارد بیشتری وجود دارند که اشاره به همه آن‌ها از حوصله این بحث خارج است. با توجه به اینکه سازمان‌های مجری آزمون در طول سال چند بار آزمون‌های متفاوت و بزرگ‌مقیاس را طراحی و اجرا می‌کنند، تحلیل روان‌سنجی تک‌تک سؤالات هر آزمون برای بهبود آزمون‌های بعدی امری ضروری است. اگر قرار باشد که زمان تحلیل هر آزمون بیش‌ازحد به طول بینجامد، برای بهینه‌سازی طراحی سؤالات آزمون‌های بعدی، فعالیت بهنگام و به‌موقع صورت نمی‌پذیرد و به‌تبع آن بهینه‌سازی آزمون‌های بعدی امکان‌پذیر نخواهد بود. با توجه به توضیحات بالا مشخص شد که دست‌اندرکاران آزمون‌های بزرگ در سطح ملی و بین‌المللی با مسئله بسیار جدیِ زمان و هزینه تحلیل آزمون مواجه هستند. مسئله اصلی پژوهش این است که آیا می‌توان مستقیماً از ۰/۰۱ داوطلبان قوی، و ۰/۰۱ داوطلبان ضعیف استفاده شود؟

با انجام عمل مرتب‌سازی داوطلبان بر اساس نمره کل‌های آزمون، قوی‌ترین افراد در بالا یا پایین فهرست و ضعیف‌ترین آن‌ها در سمت دیگر جایابی می‌شوند (بسته به صعودی یا نزولی بودن دستور مرتب‌سازی). همین امر سبب ایجاد فاصله علمی بیشتر بین گروه‌های انتخاب‌شده (یک درصد بالا و یک درصد پایین) می‌شود. با توجه به اینکه بهترین‌ها در بالای فهرست (یا پایین فهرست) قرار می‌گیرند، علت این جمله این است که در مرتب‌سازی نمره‌های داوطلبان، (برای مثال نزولی) بهترین نمره‌ها در ۱٪ بالا قرار می‌گیرند. این استدلال در مورد افراد گروه ضعیف نیز در خصوص ضعیف‌تر بودن آن‌ها صادق است. همین امر

استاندارد روش CTT[1] را زیر سؤال می‌برد. برای رفع این مشکل، باید یک درصد از افراد به‌صورت تصادفی از میان ۲۷ درصد بالا انتخاب شوند. (به همین صورت برای افراد گروه پایین).

پیشینه پژوهش

ازجمله تحولات اساسی کلاسیک ارائه‌شده در نظریه آزمون عبارت‌اند از: مدل نمره واقعی کلاسیک، متغیر تصادفی امتیاز مشاهده‌شده، امتیاز واقعی، متغیر تصادفی نمره خطا، ضریب قابلیت اطمینان، خطای استاندارد اندازه‌گیری، اثر طول آزمون بر خواص آماری نمره‌های آزمون، ضرایب پایایی تک مدیریته و مدل‌هایی برای درجات هم ارزی بین نمرات آزمون (Algina & Swaminathan, 2015).

در هر آزمونی (ملاکی یا هنجاری)، پس از برگزاری آزمون، تحلیل سؤال‌های آزمون یکی از مهم‌ترین و ضروری فرآیندهای پس از آزمون به شمار می‌آید. علت این است که با تحلیل آزمون و سؤال‌ها آن، قابلیت‌ها یا مشکلات احتمالی ناشی از طراحی سؤال‌ها از قبیل میزان سختی یا آسانی سؤال‌ها، توانا بودن سؤال‌ها در ایجاد تفکیک بین آزمون دهندگان، روایی و سایر موارد علمی روشن‌شده و با استفاده از این اطلاعات طراحان سؤال تلاش خواهند کرد که برای آزمون‌های بعد کیفیت آزمون و سؤال‌های آن را بالا ببرند.

ازآنجاکه ناشران آزمایشی، منابع مالی لازم برای تجزیه و تحلیل دقیق سؤال را دارند روند تحلیل سؤال معمولاً با ساخت آزمون‌های استانداردشده در مقیاس بزرگ درهم‌آمیخته است. در آن شرایط، تحلیل دقیق و کامل سؤال ضروری است. تحلیل سؤال‌ها به دو دلیل توسط دبیران آن‌طور که باید استفاده نمی‌شود.

۱) بسیاری از دبیران هرگز فرایند تحلیل سؤال را نیاموخته‌اند و همچنین تخصص لازم برای تفسیرِ تحلیل شخص دیگری را ندارند.

۲) درصورتی‌که دبیر مهارت و دانش لازم برای تحلیل را داشته باشد، ممکن است پشتکار لازم برای عملی کردن این کار را نداشته باشد.

افراد نباید از حجم کار بترسند زیرا برخی از روش‌های تحلیل سؤال را می‌توان بدون تلاش گزاف و غیرمنطقی اجرایی کرد (Popham, 1999).

1. Classical test theory.

علاوه بر این، مشکلات لجستیکی را می‌توان با رایانه‌های شخصی و اسکنرهای آزمونی برطرف کرد. فناوری جدید در آینده این امکان را برای دبیران فراهم خواهد کرد که تحلیل سؤال را با تلاش کمتری نسبت به حال حاضر انجام دهد. تحلیل کردن سؤالات علاوه بر مشخص کردن سؤال‌هایی که نیازمند بهبود یافتن و یا حذف شدن هستند، بازخوردی از مهارت‌های طراحی سؤال نیز به دبیر می‌دهد. این فرآیند دبیر را قادر می‌سازد تا در آزمون‌سازی بهتر عمل کند و درنتیجه تعداد سؤالات ضعیف کاهش خواهند یافت. همچنین فرآیند تحلیل سؤال می‌تواند دبیر را از اطلاعات نادرست کسب‌شده توسط دانش‌آموزان مطلع سازد. اگر تعداد زیادی از دانش‌آموزان یک پاسخ اشتباه را ارائه دهند، قطعاً نقصی در آموزش وجود داشته است. روش تمیز سؤال برای انجام تحلیل سؤال، این مزیت را دارد که تکنیک مناسبی برای معلمانی است که به امتیازدهی ماشینی یا تحلیل کامپیوتری دسترسی ندارند. این روش یک مقدار عددی (شاخص تمیزیا D برای یک سؤال) را در اختیار معلم قرار می‌دهد که می‌تواند از منفی یک تا مثبت یک متغیر باشد. هرچه قدر D برای یک سؤال بیشتر باشد، این احتمال بیشتر است که سؤال با چیزی که کل آزمون اندازه‌گیری می‌کند سازگار باشد. دانش‌آموزی که به یک مورد به‌درستی پاسخ می‌دهد، احتمالاً باید پاسخ صحیح را برای همه موارد در آزمون به دست آورد. وقتی یک تست قابل‌اعتماد باشد، همه سؤالات یا بیشتر آن‌ها هدف را به‌صورت همانند اندازه‌گیری می‌کنند. با تغییر یا حذف سؤال‌هایی که مقدار D آن‌ها کوچک است، قابلیت اطمینان یک آزمون را می‌توان افزایش داد. اولین گام در انجام تحلیل سؤال‌ها با استفاده از روش تمیز سؤال، مرتب‌سازی دانش‌آموزان بر اساس تعداد سؤال‌هایی است که به‌درستی پاسخ داده‌اند. دو گروه مشخص می‌شوند، ۲۷ درصد از دانش‌آموزان در بالا (نمره‌های بهتر) و تعداد مشابه در پایین فهرست (نمره‌های پایین). برای افزایش ثبات، مطلوب است که در هر گروه تا حد امکان دانش‌آموز وجود داشته باشد. درعین‌حال این دو گروه باید تا حد امکان متفاوت باشند تا تفاوت‌ها واضح‌تر شود. به نظر می‌رسد ۲۷ درصد، اختلاف میان این دو گروه را به حداکثر می‌رساند. در عمل، اگر بین بیست‌وپنج تا سی‌وپنج دانش‌آموز در کلاس خود دارید، احتمالاً بهتر است ده دانش‌آموز را در هر گروه قرار دهید. این امر محاسبات ریاضی را بدون هیچ اثر منفی قابل‌توجهی بر روی ثبات شاخص تمیز ساده می‌کند (Cunningham, 1998).

شاخص تمیز (D) با کم کردن تعداد دانش‌آموزانی که یک سؤال را در گروه بالا (RU) به‌درستی پاسخ داده‌اند از تعداد افرادی که آن را در گروه پایین (RL) به‌درستی پاسخ داده‌اند و تقسیم حاصل تفریق بر تعداد افراد یک گروه (N) محاسبه می‌شود.

فرمول به شرح زیر است:

$$\text{شاخص تمیز}(D) = \frac{RU - RL}{N}$$

به زبانی ساده، دشواری یک سؤال به درصد افرادی است که به آن سؤال پاسخ صحیح می‌دهند. توجه داشته باشید که هر چه این درصد بیشتر باشد، سؤال موردنظر راحت‌تر است. سؤالی که ۶۰ درصد پاسخ‌دهندگان به آن پاسخ صحیح داده‌اند دارای P برابر ۰,۶۰ (برحسب درصد) است. یک سؤال سخت که تنها ۱۰ درصد به آن پاسخ صحیح داده‌اند یک P برابر ۰,۱۰ دارد و یک سؤال آسان که ۹۰ درصد دانش‌آموزان به‌درستی به آن پاسخ داده‌اند دارای یک P برابر ۰,۹۰ است. همه سؤال‌های آزمون دارای پاسخ صحیح هستند. یکی از دلایل لزوم بررسی سطح دشواری سؤال‌ها این است که می‌توان با انتخاب درست، آزمون‌هایی در سطوح دشواری مختلف ایجاد کنیم. به‌طورکلی، از دیدگاه روان‌سنجی، آزمون‌ها باید دارای سختی متوسط باشند و میانگین درجه سختی آن‌ها (P) برابر ۰/۵۰ تعریف شود. توجه داشته باشید که این امر منجر به یک میانگین امتیاز نزدیک به ۵۰ درصد می‌شود که ممکن است استانداردی کاملاً دشوار به نظر برسد. دلیل این امر این است که شاخص درجه سختی ۰,۵۰ متمایزترین سؤال‌ها را ایجاد می‌کند، سؤال‌هایی که تفاوت‌های فردی را مشخص می‌کنند. برای مثال، سؤالاتی را در نظر بگیرید که یا بسیار دشوارند، P= ۰/۰۰ یا بسیار آسان، P=۱/۰۰. ازنظر روان‌سنجی، چنین سؤالاتی مفید نیستند زیرا تفاوت‌های موجود بین افراد را مشخص نمی‌کنند. میزان مفید بودن سؤالات با میزان و نوع پاسخ‌دهی افراد مختلف به آن‌ها مرتبط است؛ بنابراین به‌طورکلی، مفیدترین سؤال‌ها دارای p نزدیک به ۰,۵۰ هستند. البته این مسئله تا حدودی پیچیده‌تر است. به‌طوری‌که فقط دو نمره خام، صفر یا ۱۰۰ وجود دارد که وضعیت بسیار نامطلوبی است. یکی از راه‌های دور زدن این موضوع این است که سؤال‌هایی را انتخاب کنید که میانگین میزان سختی آن‌ها ۰/۵۰ است، اما درواقع ممکن است دامنه وسیعی داشته باشد، برای مثال از ۰/۳۰ تا ۰/۷۰ یا مقادیر مشابه. عامل پیچیده دیگر مربوط به مخاطبان، کسانی که آزمون برای آن‌ها طراحی می‌شود، است. قدرت تمیز سؤال در آزمون درس حساب، آن است که هر سؤال در آن آزمون به‌طور ایده‌آل باید بین کسانی که موضوع

را می‌دانند و آن‌هایی که نمی‌دانند تفاوت قائل شود. اگر آزمون افسردگی داشته باشیم، هر سؤال باید به‌طور ایده‌آل بین افراد افسرده و غیر افسرده تمایز قائل شود. قدرت تمیز سؤال به توانایی یک سؤال برای «ایجاد تفکیک صحیح» بین افرادی که در متغیر موردنظر بالاتر هستند و کسانی که پایین‌تر هستند اشاره دارد. توجه داشته باشید که برای اکثر متغیرها معمولاً فرض بر این نیست که پاسخ دوحالتی است، یک متغیر پیوسته را در نظر می‌گیریم - یعنی معتقد نیستیم که جهان توسط دو نوع آدم افسرده و غیر افسرده پر شده است، بلکه افراد مختلف می‌توانند درجات متفاوت افسردگی را نشان دهند. روش‌های مختلفی برای محاسبه شاخص‌های قدرت تمیز سؤال‌ها وجود دارد، اما اکثر آن‌ها کاملاً مشابه هستند (Oosterhof, 1976) و اساساً شامل مقایسه امتیازات عملکرد گروه قوی با امتیازات گروه ضعیف برای هر سؤال است. برای مثال فرض کنید یک آزمون درس حساب داریم که توسط ۱۰۰ کودک انجام شده است. برای هر کودک، یک نمره خام کل در آزمون، و یک خط از عملکرد آن‌ها در هر سؤال داریم. برای محاسبه شاخص قدرت تمیز سؤال برای هر سؤال، ابتدا باید تصمیم بگیریم که چگونه «امتیاز بالا» و «امتیاز پایین» را تعریف کنیم. سال‌ها پیش، (Kelley, 1939) نشان داد که بهترین استراتژی، انتخاب ۲۷ درصد بالا و ۲۷ درصد پایین است، اگرچه انحرافات جزئی از این، مانند ۲۵ درصد یا ۳۰ درصد، چندان اهمیتی ندارد. توجه داشته باشید که (Rehfisch, 1958) در رابطه با مثال مقیاس استحکام، ۲۵ درصد بالا و پایین را ازنظر استحکام تحلیل کرد. توجه داشته باشید که شاخص قدرت تمیز به‌صورت درصد بیان می‌شود. ما می‌توانیم همان محاسبات را روی نمرات خام انجام دهیم، در این موارد تعداد پاسخ‌های صحیح از ۲۷ پاسخ است، اما در صورت تغییر اندازه نمونه، نتایج ممکن است از آزمونی به آزمون دیگر متفاوت باشد. از اطلاعات به‌دست‌آمده از چنین تحلیلی می‌توان برای ایجاد تغییرات در سؤالات و بهبود آزمون استفاده کرد. به‌عنوان‌مثال، توجه داشته باشید که به نظر می‌رسد که سؤال ۱ به‌خوبی تمایز قائل است. اکثر افراد دارای امتیاز بالا (۸۵٪) به سؤالات پاسخ صحیح دادند، درحالی‌که تعداد بسیار کمتری از افراد که دارای امتیاز پایین (۲۲٪) هستند به سؤال پاسخ صحیح دادند. ازنظر تئوری، یک سؤال با قدرت تمیز کامل دارای مقدار D برابر ۱۰۰٪ خواهد بود. سؤال‌های ۲ و ۳ به‌خوبی تمایز قائل نمی‌شوند، سؤال ۲ بسیار آسان و سؤال ۳ بسیار دشوار است. سؤال ۴ کار می‌کند اما برعکس، تعداد کمی از افراد دارای امتیاز بالا سؤال را به‌درستی پاسخ داده‌اند. اگر این سؤالی است که در آن پاسخ

صحیح وجود دارد، یک D منفی به ما هشدار می‌دهد که مشکلی در سؤال وجود دارد و باید بازبینی شود. اگر سؤالی از یک آزمون شخصیت داشتیم که در آن پاسخ صحیحی وجود نداشت، به دست آمدن D منفی درواقع به ما می‌گفت که باید امتیاز را معکوس کنیم. ما انتخاب کرده‌ایم که امتیاز بالا و امتیاز پایین را بر اساس خود نمره کل آزمون تعریف کنیم. این ممکن است کمی حلقوی به نظر برسد، اما درواقع کاملاً درست است. اگر آزمون «دانش حساب» را اندازه‌گیری کند، نمره بالا در حساب درواقع از آن فردی است که نمره بالایی در دانش حساب می‌گیرد.

در نظریه احتمالات، قانون اعداد بزرگ قضیه‌ای است که نتیجه انجام یک آزمایش مشابه را برای چندین بار توصیف می‌کند. طبق این قانون، میانگین نتایج به‌دست‌آمده از تعداد زیادی آزمایش، باید به مقدار مورد انتظار (امید ریاضی[1]) نزدیک باشد و با انجام آزمایش‌های بیشتر به مقدار مورد انتظار نزدیک‌تر می‌شود (Dekking, 2005).

نکته مهم درباره قانون اعداد بزرگ این است که این قانون، همان‌طور که از نامش پیداست، تنها زمانی اعمال می‌شود که تعداد زیادی مشاهده در نظر گرفته شوند. هیچ اصلی وجود ندارد که تعداد کمی از مشاهدات با مقدار مورد انتظار منطبق شوند. همچنین مهم است که توجه داشته باشید که قانون اعداد بزرگ فقط برای میانگین اعمال می‌شود.

صورت ریاضی آن بدین شکل است:

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{X_i}{n} - \bar{X} = 0$$

فرمول‌های دیگری که مشابه به نظر می‌رسند قابل‌قبول نیستند؛ مانند انحراف معیارِ «نتایج نظری»:

$$\sum_{i=1}^{n} X_i - n \times \bar{X}$$

این فرمول نه‌تنها با افزایش n به سمت صفر همگرا نمی‌شود، بلکه با افزایش n به یک مقدار ثابت میل خواهد کرد (Dekking, 2005؛ Yao & Gao, 2016). در پژوهش دیگری که به مقایسه روش کلاسیک با روش نظریه پرسش و پاسخ پرداخته است آمده است که بر طبق تحلیل‌های نظریه کلاسیک آزمون، میانگین ضریب پایایی کل آزمون‌ها با

1. Expected Value.

استفاده از فرمول آلفای کرونباخ ۶۶/۰ به دست آمد و این مطلب حاکی از آن است که آزمون‌های مذکور از این نظر در سطح نسبتاً مطلوبی قرار دارند. تحلیل گزینه‌های انحرافی نشان داد که گزینه‌های انحرافی تمامی سؤالات در همهٔ آزمون‌ها هم احتمال نیستند و میزان جذابیت این گزینه‌ها برای آزمودنی‌ها یکسان نیست و گزینه‌های انحرافی از این حیث، عملکرد معیوبی داشته‌اند. تحلیل ویژگی‌های روان‌سنجی کل سؤالات آزمون‌ها طبق IRT[1] نشان داد که آزمون‌ها تک‌بعدی بوده و استقلال موضعی در آن‌ها برقرار است. در بررسی پارامتر دشواری سؤالات مشخص گردید که سؤالات دارای پارامتر دشواری منظمی نیستند و از ساده به مشکل مرتب نشده‌اند (جلیلی و همکاران، ۱۳۸۹). جلیلی در پژوهش دیگری با تمرکز بر روش بیزی و خطای اندازه‌گیری به این نتیجه رسید که در نظر گرفتن خطای اندازه‌گیری تک‌تک سؤالات بر تحلیل‌ها در چارچوب ینزی و با استفاده از نمونه‌گیری گیبس به‌طور چشمگیری توانمندی تحلیل‌های چندسطحی را بهبود می‌بخشد و افزایش معنادار نسبت واریانس تبیین شده را در پی خواهد داشت (جلیلی و همکاران، ۱۳۹۳). تمام سؤالات امتحانات نهایی رشته علوم تجربی سال سوم مقطع متوسطه در خردادماه ۱۳۹۰ برای تعیین ویژگی‌های روان‌سنجی آن‌ها از عملکرد ۶۰۰ دانش‌آموز ناحیه یک شهرستان خرم‌آباد در دو درس که به‌صورت تصادفی انتخاب شده بودند، موردبررسی قرار گرفت. ضریب اعتبار آزمون‌های زیست‌شناسی و ادبیات فارسی با استفاده از آلفای کرونباخ ۰/۹۷ و ۰/۹۶ تعیین گردید. بر اساس CTT، میانگین ضرایب دشواری و تمیز سؤالات در آزمون‌های ادبیات فارسی ۰/۶۵ و ۰/۵۷ و در درس زیست‌شناسی ۰/۵۰ و ۰/۶۵ به دست آمد (فلسفی نژاد، فرخی و بهرامی، ۱۳۹۵).

البته انتقادهایی هم به روش CTT وارد است که خارج از بحث این مقاله است. در همین راستا (Adekunle et al, 2022) با توجه به انتقادها توصیه می‌کنند که شورای امتحانات لسوتو (ECOL[2]) برای توسعه آزمون و تحلیل سؤال، از CTT به نظریه آزمون مدرن (مانند IRT) روی آورد که مزایای متعددی را ارائه می‌دهد.

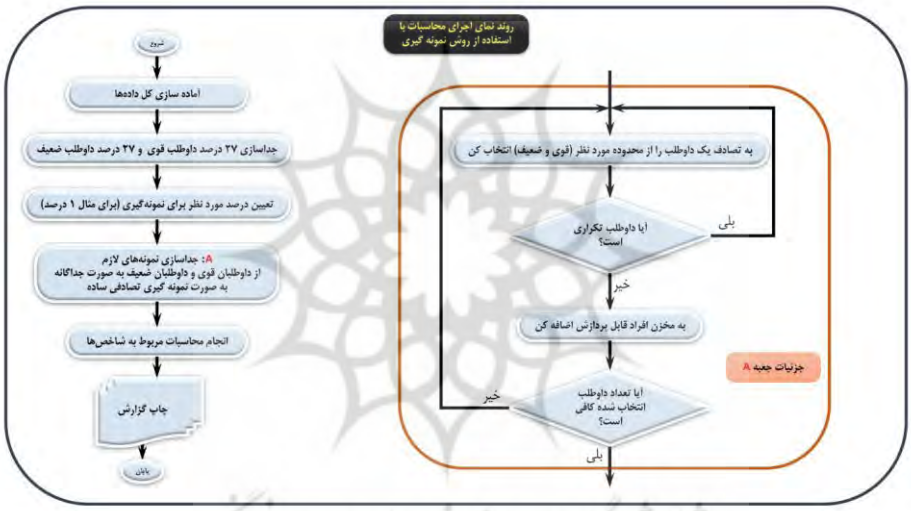
1. Item-response Theory.
2. Examination Council of Lesotho (Country in Southern Africa)

روش

این پژوهش از نوع توسعه‌ای است و تصمیم دارد با به‌کارگیری روش‌های احتمالی آماری با کاهش زمان محاسبات، برآوردی از پارامترها به دست آورد که با دقت خیلی بالا نزدیک به پارامترهای محاسبه‌شده با روش‌های مرسوم و استاندارد می‌باشند.

در روش پیشنهادی انتخاب تصادفی یک درصد از گروه بالا و یک درصد از گروه پایین جامعه تحت مطالعه به‌جای انتخاب همه ۲۷ درصد از گروه‌های بالا و پایین که به‌صورت دقیق در حالت‌های مختلف اجرا شده است. شکل ۱ روند اجرا را نمایش می‌دهد.

شکل ۱. روندنمای انجام محاسبات



توضیحات مربوط به روندنما و الگوریتم اجرای محاسبات:

– داده‌های موردنظر برای محیط برنامه‌نویس آماده می‌شوند. در مورد این مثال داده‌ها در محیط بانک اطلاعاتی MS SQL Server 2016 آماده و شاخص‌گذاری شده‌اند.

– زبان برنامه‌نویسی استفاده شده برای تولید نرم‌افزار کاربردی Embarcadero Delphi XE Rio 3.10 است.

– داده‌ها باید شامل اطلاعاتی مانند، شماره داوطلبان، نمره کل آزمون کسب‌شده توسط داوطلبان (برای مرتب‌سازی)، همه پاسخ‌های داوطلبان، کلید همه سؤالات و سایر اطلاعات

موردنیاز باشند (به‌منظور جلوگیری از مشکل افزونگی داده‌ها[1] که باعث کندی سیستم می‌شود، توصیه می‌شود مانند این مثال، داده‌ها در جدول‌های مختلفی نگهداری شوند)

– ۲۷ درصد از داده‌های بالای فهرست (افراد قوی) و ۲۷ درصد از داده‌های پایین فهرست (افراد ضعیف) جدا شوند (در برنامه مورداستفاده توسط نویسنده مقاله این افراد داخل یک جدول نگهداری شده‌اند).

– ضریب درصد نمونه‌گیری را تعیین کنید (برای مثال ۱ درصد). تعداد ۱ درصد از ۲۷ درصد را محاسبه کنید.

در مورد مثال ما:

۴۸۴۵۷۶ * ۰/۲۷ * ۰/۰۱ = ۱۳۰۸/۳۵ ≅ ۱۳۰۸

بنابراین تعداد کل افراد قوی و ضعیف که قرار است پردازش شوند عبارت است از:

۱۳۰۸*۲=۲۶۱۶

– اکنون به‌صورت جداگانه از هر گروه تعداد ۱۳۰۸ نفر را طبق نمودار بالا جدا کنید. دقت کنید که افراد تکراری انتخاب نشوند. در علم آمار به این روش «نمونه‌گیری بدون جایگزینی[2]» گفته می‌شود.

– محاسبات مربوط به شاخص‌های P، D و پاسخ‌های غلط را انجام دهید.

– گزارش نهایی را چاپ کنید.

بزرگ‌ترین عدد مربوط به حجم نمونه طبق جدول تعیین حجم نمونه Krejcie and Morgan (1980) مقدار ۳۸۴ مورد است که اطمینان ۹۵ درصدی را در محاسبات فراهم می‌کند. پس با توجه به اینکه حجم نمونه ما بزرگ‌تر است نگران میزان اطمینان برآوردگرها نخواهیم بود و حداقل ۹۵ درصد اطمینان خواهیم داشت.

جدول محاسبات شبیه‌سازی‌شده به قرار زیر است:

آمارهای توصیفی و مشخصات آزمون:

- نوع آزمون: تستی (چهارگزینه‌ای)،
- تعداد داوطلب: ۴۸۴۵۷۶،
- تعداد ۲۷ درصد افراد جداشده قوی: ۱۳۰۸۳۵ نفر،
- تعداد ۲۷ درصد افراد جداشده ضعیف: ۱۳۰۸۳۵ نفر،

1. Data Redundancy.
2. Sampling without replacement.

- درصد جداسازی نمونه‌گیری: ۱ درصد (۰/۰۱) برای گروه قوی برابر ۱۳۰۸ نفر،
- درصد جداسازی نمونه‌گیری: ۱ درصد (۰/۰۱) برای گروه ضعیف برابر ۱۳۰۸ نفر،
- تعداد سؤال: ۱۵۰ (که البته در این مقاله تعداد ۳ سؤال به‌عنوان نمونه پردازش شده است. سؤال ۱: اولین سؤال درس ادبیات، سؤال ۲: اولین سؤال درس زبان انگلیسی و سؤال ۳: اولین سؤال درس ریاضی).
- تعداد شبیه‌سازی اجرا شده برای هر سؤال: ۱۰ مرتبه.

جدول ۱. توزیع فراوانی سؤال ۱

	اولین سؤال ادبیات فارسی		
	کل جامعه	۱ درصد بالا	۱ درصد پائین
پاسخ صحیح	۲۳۱۸۵۱	۹۷۲	۳۱۷
پاسخ غلط	۲۵۲۷۲۵	۳۳۶	۹۹۱
جمع	٤٨٤٥٧٦	۱۳۰۸	۱۳۰۸

جدول ۲. توزیع فراوانی سؤال ۷٦

	اولین سؤال زبان انگلیسی		
	کل جامعه	۱ درصد بالا	۱ درصد پائین
پاسخ صحیح	۱۰٤۱۵۳	۵۸۲	۲۵۸
پاسخ غلط	۳۸۰٤۲۳	۷۲٦	۱۰۵۰
جمع	٤٨٤٥٧٦	۱۳۰۸	۱۳۰۸

جدول ۳. توزیع فراوانی سؤال ۱۲٦

	اولین سؤال ریاضی		
	کل جامعه	۱ درصد بالا	۱ درصد پائین
پاسخ صحیح	۳۲۵۱٤	۱۰۹	۷۳
پاسخ غلط	٤٥۲۰٦۲	۱۱۹۹	۱۲۳۵
جمع	٤٨٤٥٧٦	۱۳۰۸	۱۳۰۸

یافته‌ها

به‌منظور بررسی اینکه آیا برآوردهای به‌دست‌آمده مقدار واقعی پارامتر را در جامعه درست برآورد می‌کنند یا خیر، از آزمون‌های آماری متناسب استفاده شده است.

جدول ٤. محاسبات شاخص‌ها با استفاده از روش استاندارد (۲۷٪) و روش نمونه‌گیری تصادفی ساده (۱٪ از ۲۷٪)

نوبت شبیه‌سازی	سؤال ۱			سؤال ٧٦			سؤال ١٢٦		
	P	D	زمان*	P	D	زمان*	P	D	زمان*
پردازش استاندارد**	۰/۵۰	۰/۵۱	۷۳۹۳	۰/۱۵	۰/۱۱	۷۰۹۱	۰/۰۸	۰/۰٤	۷٤۹۰
نمونه‌گیری ۱	۰/۵۱	۰/۵۱	۱۲۰	۰/۱۵	۰/۱۱	۱۱۰	۰/۰۸	۰/۰٦	۲٤٥
نمونه‌گیری ۲	۰/۵۰	۰/٤۸	۱۰۹	۰/۱۵	۰/۱۱	۳۱	۰/۰۸	۰/۰٥	۱۱۳
نمونه‌گیری ۳	۰/۵۰	۰/۵۰	٦۱	۰/۱۵	۰/۱۱	۱۲٤	۰/۰۷	۰/۰٤	۱۰٦
نمونه‌گیری ٤	۰/٤۹	۰/۵۰	۱۲۷	۰/۱٤	۰/۱۰	۱۱۳	۰/۰۹	۰/۰۳	۱۰۸
نمونه‌گیری ٥	۰/۵۱	۰/۵۱	۱۳۳	۰/۱۵	۰/۱۱	۱۰۸	۰/۰۸	۰/۰٤	۱۱۷
نمونه‌گیری ٦	۰/۵۰	۰/۵۰	۱۱۸	۰/۱۵	۰/۱٤	۱۰٦	۰/۰۷	۰/۰۳	۱۰۹
نمونه‌گیری ۷	۰/۵۱	۰/۵۰	۱۱۱	۰/۱٤	۰/۱۰	۱۱۷	۰/۰۸	۰/۰٥	۱٦۷
نمونه‌گیری ۸	۰,۵۰	۰/۵۳	۱۰۹	۰/۱۵	۰/۱۱	۱۳۷	۰/۰۷	۰/۰٥	۱۱۱
نمونه‌گیری ۹	۰/۵۱	۰/۵۰	۲۲٥	۰/۱٤	۰/۱۰	۱۱۰	۰/۰۸	۰/۰٥	۱۰۹
نمونه‌گیری ۱۰	۰/۵۰	۰/٤۸	۱۷۹	۰/۱٤	۰/۰۸	۱۳٤	۰/۰۸	۰/۰٤	۱۲۲
میانگین	۰/۵۰۳	۰/۱٤٦	۱۲۹,۲	۰/۱٤٦	۰/۱۰۷	۱۰۹	۰/۰۷۸	۰/۰٤٤	۱۳۰,۷

* زمان برحسب دقیقه.

** استفاده از ۲۷٪ بالا (داوطلبان قوی) و ۲۷٪ پایین (داوطلبان ضعیف).

در خصوص سؤال ۱، مقدار پارامتر جامعه P=۰/۵۰ محاسبه شده است. میانگین مقدارهای P به‌دست‌آمده از محاسبات به روش نمونه‌گیری ۰/۵۰۳ است. با توجه به اینکه مشاهدات دارای توزیع نرمال نبودند ($Shapiro - Wilk\ P_{Value} = ۰/۰۱۵$) از آزمون رتبه علامت دار یک نمونه‌ای ویلکاکسون[1] برای تساوی میانگین نمونه با ۰/۰۵ استفاده شده است. نتایج به‌دست‌آمده نشان داد که میانگین نمونه به‌دست‌آمده با مقدار پارامتر برابر است ($P_{Value} =$

1. Wilcoxon Signed Rank Test.

۰/۱۸۰). بنابراین برای صرفه‌جویی در زمان پردازش، استفاده از روش پیشنهادی با حجم نمونه مناسب (۱٪ از ۲۷٪) توصیه می‌شود.

محاسبات برای مقایسه ضریب تمیز و دشواری هر سه سؤال در جدول ۵ آمده است:

جدول ۵. آزمون‌های آماری مربوط به شاخص P با استفاده از روش CTT (۲۷٪) و روش نمونه‌گیری تصادفی ساده (۱٪ از ۲۷٪)

	P استاندارد	میانگین نمونه	نام آزمون	P_{Value}	*تصمیم
سؤال ۱	۰/۵۰	۰/۵۰۳	One sample Wilcoxon signed rank test	۰/۱۸۰	قبول: H_0
سؤال ۷۶	۰/۱۵	۰/۱۴۶	One sample Wilcoxon signed rank test	۰/۰۴۶	قبول: H_0
سؤال ۱۲۶	۰/۰۸	۰/۰۷۸	One sample Wilcoxon signed rank test	۰/۳۱۸	قبول: H_0

* $\alpha = ۰/۰۱$

نتایج جدول ۵ نشان می‌دهند که بین مقادیر به‌دست‌آمده میانگین شاخص دشواری در میانگین‌های نمونه و پارامتر جامعه در سه سؤال مذکور، تفاوت معنی‌داری مشاهده نشده است؛ بنابراین با مشاهده مقادیر به‌دست‌آمده ضرایب دشواری (شاخص P)، به‌طور خلاصه می‌توان گفت استفاده از روش نمونه‌گیری تصادفی ساده یک‌درصدی از ۲۷ درصد مشاهدات (۲٪ از ۵۴٪)، همان نتایجی را به دست می‌دهد که محاسبات استاندارد خواهد داد. مطلب قابل‌ذکر این است که محاسبات مربوط به شاخص P با استفاده از همه مشاهدات انجام می‌شود؛ اما نتایج نشان داد که حتی برای این مورد هم می‌توان از روش نمونه‌گیری استفاده کرد.

جدول ۶. آزمون‌های آماری مربوط به شاخص D با استفاده از روش CTT (۲۷٪) و روش نمونه‌گیری تصادفی ساده (۱٪ از ۲۷٪)

	P استاندارد	میانگین نمونه	نام آزمون	P_{Value}	*تصمیم
سؤال ۱	۰/۵۱	۰/۵۰۲	One sample Wilcoxon signed rank test	۰/۰۹۸	قبول: H_0
سؤال ۲	۰/۱۱	۰/۱۰۷	One sample Wilcoxon signed rank test	۰/۴۰۸	قبول: H_0
سؤال ۳	۰/۰۴	۰/۰۴۴	One sample Wilcoxon signed rank test	۰/۲۰۶	قبول: H_0

* $\alpha = ۰/۰۱$

نتایج جدول ۶ نشان می‌دهند که مقادیر به‌دست‌آمده میانگین شاخص تمیز در میانگین‌های نمونه و پارامتر جامعه در سه سؤال مذکور، تفاوت معنی‌داری مشاهده نشده است؛ بنابراین با مشاهده مقادیر به‌دست‌آمده ضرایب تمیز و دشواری، به‌طور خلاصه می‌توانند گفت استفاده از روش نمونه‌گیری تصادفی ساده یک‌درصدی از ۲۷ درصد داوطلبان قوی و ضعیف (درمجموع ۲٪ از ۵۴٪)، همان نتایجی را به دست می‌دهد که محاسبات استاندارد خواهد داد.

بحث و نتیجه‌گیری

نتایج نشان می‌دهند برای محاسبه شاخص دشواری، استفاده از روش نمونه‌گیری تصادفی ساده یک درصد از ۲۷ درصد مشاهدات (۲٪ از ۵۴٪)، همان نتایج را به دست می‌دهد که محاسبات استاندارد خواهد داد. همچنین نتایج دیگر نشان می‌دهند برای محاسبه شاخص تمیز، استفاده از روش نمونه‌گیری تصادفی ساده یک‌درصدی از ۲۷ درصد داوطلبان قوی و ضعیف (درمجموع ۲٪ از ۵۴٪)، همان نتایج را به دست می‌دهد که محاسبات استاندارد خواهد داد. بنابراین می‌توان برای صرفه‌جویی در زمان پردازش لازم برای استخراج شاخص‌های روان‌سنجی کلاسیک از اطلاعات مربوط به ۱٪ داوطلبان برای تصمیم‌گیری و تحلیل سؤالات استفاده کرد. همچنین یافته‌ها نشان می‌دهند چنین کاری می‌تواند باعث صرفه‌جویی به میزان حدود ۹۹ درصد در زمان پردازش داده‌ها شود.

بر اساس «قانون ضعیف اعداد بزرگ»، میانگین نتایج به‌دست‌آمده از تعداد زیادی آزمایش باید نزدیک به مقدار مورد انتظار (امید ریاضی) و نزدیک‌تر به آزمایش‌های مورد انتظار باشد (Dekking, 2005).

به لحاظ نظری، نتایج به‌دست‌آمده با روش پیشنهادی از پشتیبانی نظریه قانون ضعیف اعداد بزرگ نیز برخوردارند. در خصوص دقت برآوردگرها هم می‌توان گفت بزرگ‌ترین عدد مربوط به حجم نمونه طبق جدول تعیین حجم نمونه کرجسی و مورگان مقدار ۳۸۴ است که اطمینان ۹۵ درصدی را در محاسبات فراهم می‌کند. پس با توجه به اینکه حجم نمونه ما بزرگ‌تر از این عدد است، نگران میزان اطمینان برآوردگرها هم نخواهیم بود و حداقل ۹۵ درصد اطمینان خواهیم داشت.

هنگامی‌که تحلیل کل یک جمعیت غیرممکن است، آماردانان روش‌های نمونه‌گیری را با انتخاب نمونه‌ای کوچک از عناصر جامعه هدف به‌عنوان الگویی از کل جامعه به کار

می‌گیرند؛ بنابراین، یکی از بنیادی‌ترین وظایف علم آمار، ارائه حدومرز برای حجم نمونه است که برآورد مناسبی برای پارامترهای جامعه ارائه دهد، و از روش‌های احتمالی برای تضمین این هدف، تحت شرایط مختلف استفاده می‌شود. تقریباً همه این تضمین‌ها مبتنی بر مدل‌های احتمالی کلاسیک هستند که فرض می‌کنند جامعه هدف ثابت است و به نمونه جمع‌آوری‌شده در طول فرآیند بستگی ندارد. درواقع چنین فرضی همیشه واقع‌بینانه نیست. در این کار یک چارچوب انتزاعی را بررسی می‌کنیم که این فرض را حذف می‌کند، و ثابت می‌کنیم که فرآیندهای نمونه‌گیری طبیعی و کارآمد نمونه‌هایی را تولید می‌کنند که به‌خوبی نشان‌دهنده جمعیت هدف هستند. موقعیت‌هایی که فرآیند نمونه‌گیری به‌طور صریح یا ضمنی بر جمعیت هدف تأثیر می‌گذارد، در تحلیل داده‌های مدرن فراوان است (Alon et al., 2021). همان‌طور که از اطلاعات به‌دست‌آمده از جدول ۴ پیداست، میانگین نتایج به هدف همگرا بوده و استفاده از روش نمونه‌گیری در پردازش داده‌های بزرگ‌مقیاس را در روش تحلیل کلاسیک آزمون، مناسب ارزیابی می‌کند. البته پیشنهاد پژوهشگر این است که درصد نمونه‌گیری (۱٪) طوری انتخاب شود که حجم نمونه به‌دست‌آمده کمتر از ۳۸۴ داوطلب نشود تا به استناد «فرمول کاکران برای تعیین حجم نمونه برای نسبت‌ها» حداقل ۹۵ درصد اطمینان به پارامترهای برآورد شده داشته باشیم.

یادآور می‌شود که در بررسی‌ها مشخص شد که هیچ پژوهشی با این سبک (استفاده از روش‌های نمونه‌گیری آماری) و در این ابعاد (آزمون‌های ملی با این حجم از داوطلب) گزارش نشده است

با توجه به نتایج محاسبات به‌دست‌آمده در بالا، به نظر می‌رسد که روش نمونه‌گیری تصادفی ساده که محاسبات این مقاله بر مبنای آن انجام شده است نتایج مطلوبی را ارائه کرده است. کاهش زمان محاسبه از ۷۳۹۳ دقیقه به متوسط ۱۲۹ دقیقه یک نتیجه عالی به نظر می‌رسد. درواقع این صرفه‌جویی در زمان پردازش تا حدود کمتر از ۲٪ در برخی موارد می تواند از انصراف از تحلیل یا کاهش دقت در تحلیل جلوگیری کند.

همچنین درصورتی‌که محدودیت‌های زمانی برای ارائه فوری گزارش بعد از آزمون را نیز در نظر بگیریم به نظر می‌رسد استفاده ازاین‌گونه روش‌ها اجتناب‌ناپذیر است. با توجه به مزایای مطرح‌شده در بالا بی‌شک هزینه انجام تحلیل‌های موردبحث هم به‌شدت کاهش می‌یابد.

شایان‌ذکر است که روش‌های نمونه‌گیری دیگری نیز در علم آمار وجود دارند که خواننده علاقه‌مند می‌تواند این محاسبات را با استفاده از آن‌ها انجام دهد. برای مثال نمونه‌گیری آماری با استفاده از روش طبقه‌بندی.

مشارکت نویسندگان

ارائه ایده، تهیه داده‌ها، مطالعات اولیه، برنامه‌نویسی و نوشتن گزارش نهایی توسط بهروز کاوه ئی انجام شده است.

تعارض منافع

نویسنده هیچ‌گونه تعارض منافعی ندارد.

سپاسگزاری

از مدیران و کارکنان محترم سازمان سنجش آموزش کشور که با گشاده‌رویی داده‌های آزمون را در اختیار قرار دادند سپاسگزاری کند.

منابع

فلسفی‌نژاد، محمدرضا، فرخی، نورعلی، و بهرامی، لیلا (۱۳۹۵). ویژگی‌های روان‌سنجی امتحانات نهایی سال سوم متوسطه و قابلیت آن‌ها در گزینش داوطلبان ورود به دوره‌های کارشناسی. *فصلنامه اندازه‌گیری تربیتی*، ۷(۲۳)، ۴۵-۷۶. https://doi.org/10.22054/jem.2017.6387.1196

یونسی، جلیل، دلاور، علی، و فلسفی‌نژاد، محمدرضا. (۱۳۸۹). بررسی ویژگی‌های روان‌سنجی سؤالات تخصصی آزمون‌های فراگیر رشته روان‌شناسی دانشگاه پیام نور در سال ۱۳۸۵. *فصلنامه اندازه‌گیری تربیتی*، ۱ (۲)، ۱۳۹-۱۶۹. https://jem.atu.ac.ir/article_5636.html

یونسی، جلیل، اسکندری، فرزاد، دلاور، علی، فلسفی‌نژاد، محمدرضا، و فرخی، نورعلی. (۱۳۹۳). مقایسه توانمندی رویکرد بیزی مدل IRT چندسطحی و مدل کلاسیک چندسطحی: تحلیل داده‌های آزمون فیزیک تیمز پیشرفته ۲۰۰۸. *فصلنامه اندازه‌گیری تربیتی*، ۵ (۱۵)، ۱۶۶-۱۸۶. https://jem.atu.ac.ir/article_277.html

References

Adekunle Ayanwale, M., Chere-Masopha, J., & Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the Examination Council of Lesotho. *International Journal of Learning, Teaching and Educational Research, 21*(8), 409-426. https://doi.org/10.26803/ijlter.21.8.22

Algina, J., & Swaminathan, H. (2015). Psychometrics: Classical test theory. In J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., Vol. 19, pp. 423-430). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.44049-4

Alon, N., Ben-Eliezer, O., Dagan, Y., Moran, S., Naor, M., & Yogev, E. (2021). Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing* (pp. 447-455). https://doi.org/10.1145/3406325.3451029

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101-133. https://doi.org/10.1214/ss/1009213286

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley. ISBN-13: 978-0471162407

Cunningham, G. K. (1998). *Assessment in the classroom: Constructing and interpreting texts*. Falmer Press.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2005). *A modern introduction to probability and statistics*. Springer. https://doi.org/10.1007/1-84628-168-7

Domino, G., & Domino, M. L. (2006). *Psychological testing: An introduction* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511813757

Falsafi-Nejad, M. R., Farrokhi, N., & Bahrami, L. (2016). [Psychometric characteristics of third-grade high school final examinations and their

capability in selecting candidates for bachelor's degree programs]. *Educational Measurement Quarterly*, 7(23), 45-76. https://doi.org/10.22054/jem.2017.6387.1196 [In Persian]

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4125-4130. https://doi.org/10.21037/jtd.2017.09.14

Illowsky, B., Dean, S., & Community College Open Educational Resources Group. (2018). *Introductory statistics*. OpenStax. https://openstax.org/details/books/introductory-statistics

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*(1), 17-24. https://doi.org/10.1037/h0057123

Khare, V., Nema, S., & Baredar, P. (2020). *Ocean energy modeling and simulation with big data: Computational intelligence for system optimization and grid integration*. Elsevier. https://doi.org/10.1016/C2018-0-05113-1

Oosterhof, A. C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement, 13*(2), 145-150. https://doi.org/10.1111/j.1745-3984.1976.tb00005.x

Popham, W. J. (1999). *Classroom assessment: What teachers need to know* (2nd ed.). Allyn & Bacon.

Rehfisch, J. M. (1958). Some scale and test correlates of a personality rigidity scale. *Journal of Consulting Psychology, 22*(5), 372-374. https://doi.org/10.1037/h0048746

Wallis, S. A. (2013). Binomial confidence intervals and contingency tests: Mathematical fundamentals and the evaluation of alternative methods. *Journal of Quantitative Linguistics, 20*(3), 178-208. https://doi.org/10.1080/09296174.2013.799918

Yao, K., & Gao, J. (2016). Law of large numbers for uncertain random variables. *IEEE Transactions on Fuzzy Systems, 24*(3), 615-621. https://doi.org/10.1109/TFUZZ.2015.2466080

Younesi, J., Delavar, A., & Falsafi-Nejad, M. R. (2010). [Investigating psychometric characteristics of specialized questions in comprehensive exams of psychology at Payame Noor University in 2006]. *Educational Measurement Quarterly*, 1(2), 139-169. https://jem.atu.ac.ir/article_5636.html [In Persian]

Younesi, J., Eskandari, F., Delavar, A., Falsafi-Nejad, M. R., & Farrokhi, N. (2014). [Comparing the capability of Bayesian approach of multilevel IRT model and classical multilevel model: Analysis of TIMSS advanced 2008 physics test data]. *Educational Measurement Quarterly*, 5(15), 166-186. https://jem.atu.ac.ir/article_277.html [In Persian]

Zar, J. H. (1999). *Biostatistical analysis* (4th ed.). Prentice Hall.