Iranian Journal of Psychoeducational Assessment
University of Zanjan
June 2026. Vol 2. Issue 1

Research Paper

# More Data, Less Validity: The Paradox of Incorporating Academic Records into Student Selection Equations

**Ebrahim Khodaie[1*], Mahdi Karvandi Renani[1], Mojtaba Jahanifar[2]**

1. Department of Teaching Methods and Educational and Curriculum Planning, Faculty of Psychology and Education, University of Tehran, Tehran, Iran.
2. Department of Educational Sciences, Faculty of Education and Psychology, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

**Article info:**
Received: 02.06.2026
Revised: 04.06.2026
Accepted: 10.06.2026

**Keywords:**
Validity, National Entrance Examination, Ceiling Effect, Simulation, Final Exams



Publisher: University of Zanjan

## Abstract

**Background:** Iran's high-stakes national university entrance examination calculates candidates' final scores by combining the Konkur (a norm-referenced test) and academic records (a criterion-referenced test). This study critically evaluates the claim that combining these scores provides a more comprehensive assessment of candidates' true abilities.

**Methods:** Using a two-stage analytical-simulation design, we first critiqued the theoretical assumptions of this policy using Kane's argument-based approach to validation, focusing on score interpretation and decision consequences. Second, a statistical simulation quantified the impact of the ceiling effect on candidates' rankings.

**Results:** Results revealed that the ceiling effect severely diminishes the discriminating power of academic records among top-tier candidates (the top 10%). As mean scores approach the scale's upper limit, the rank correlation between true ability and observed scores nears zero, rendering rankings highly susceptible to random error.

**Conclusion:** Findings indicate that the fundamental premise of combining these assessments is untenable from a validity perspective. The ceiling effect in final exams reduces this score component into an inefficient, error-prone source for differentiating elite candidates. This undermines the ranking system's credibility in the most competitive segment and compromises selection fairness by introducing systematic error. Consequently, combining Konkur scores and academic records is methodologically questionable, necessitating a serious revision of admission policies to safeguard deserving candidates' rights.

Use your device to scan and read the article online

Citation: Khodaei, E., & Karvandi Renani, M., & Jahanifar, M. (2026). More Data, Less Validity: The Paradox of Incorporating Academic Records into Student Selection Equations. *Iranian Journal of Psychoeducational Assessment*, 2 (1), 1-20.
https://doi.org/10.30470/ijpa.2026.736668

***Corresponding Author:** Ebrahim Khodaie
**Address:** Faculty of Psychology and Education, University of Tehran, Tehran, Iran.
**Email:** khodaie@ut.ac.ir

# Extended Abstract


## Introduction

Validity represents the foundational paradigm in developing and evaluating educational and psychological assessments (AERA, APA, NCME, 2014; Messick, 1989). As testing instruments scale from individual diagnostics to macro-level, high-stakes societal selection tools, the empirical validation of their interpretation and use becomes critically paramount. Public claims regarding selection equity and accuracy demand rigorous public justification (Kane, 2013). In Iran's higher education matrix, the national university entrance examination (the Konkur) has undergone a contentious structural paradigm shift. The current policy mandates combining the norm-referenced Konkur scores with criterion-referenced high school academic records. The institutional claim undergirding this integration posits that combining these disparate metrics yields a more comprehensive, multi-dimensional, and fair assessment of a candidate's true scholastic capability. However, this policy introduces an acute psychometric contradiction. Norm-referenced instruments are purposely engineered to maximize inter-individual score variance and rank-order candidates across an expansive ability continuum. Conversely, criterion-referenced final exams are structurally designed to assess curriculum mastery against fixed standards, independent of variance maximization goals (Crocker & Algina, 1986). This structural misalignment generates a severe psychometric artifact known as the ceiling effect. As candidate scores saturate the upper scale boundary, the ability of academic records to accurately differentiate among elite applicants deteriorates. This study utilizes Kane's validation framework and advanced statistical simulations to critically evaluate how the mathematical limitations of the ceiling effect undermine the validity, fairness, and ranking stability of this combined selection policy, focusing specifically on high-achieving cohorts.

## Method

To examine this policy claim, the study used a two-stage analytical–simulation framework. First, it critically analyzed the Interpretation and Use Argument (IUA) of the integrated selection system based on Kane's argument-based validation approach, with particular emphasis on the scoring inference. Second, a large-scale simulation was conducted in R using a synthetic population of 100,000 candidates to estimate the effect of scale truncation on candidate rankings. True scores were generated from a normal distribution under three mean ability conditions (μ = 13, 14, and 15), with random measurement error added according to Classical Test Theory assumptions. Three scoring conditions were compared: an unconstrained baseline model, a post-error truncation model, and a true score truncation model with secondary clipping. Rank outcomes were then evaluated using Spearman rank correlations and Mean Absolute Rank Differences for the full sample, the top 10%, and the lower 90%.

## Results

The analytical critique demonstrates that the primary psychometric vulnerability lies within the scoring inference, where scale boundaries induce non-linear distortions. Under Classical Test Theory (CTT) assumptions, an individual's observed score is conventionally expressed as:

$$X_i = T_i + E_i$$

where $X_i$ is the observed score, $T_i$ is the true latent ability, and $E_i$ is the random error. When an assessment is bound by a rigid upper scale ceiling, S, the observed score formula undergoes a non-linear transformation:

$$X_i^{(capped)} = min(X_i, S)$$

For high-achieving sub-populations whose latent true abilities exceed or equal the scale ceiling ($T_i >= S$), the observed score collapses into a fixed value:

$$X_i^{(capped)} = S$$

Consequently, the observed score variance within this elite segment approaches zero, severely attenuating the true underlying ability variance (Nunnally & Bernstein, 1994):

$$Var\left(X_i^{(capped)}\right) << Var(T_i)$$

When true score variance is extinguished, the relative weight of the random error component expands exponentially, creating an extreme ranking hyper-sensitivity to random noise:

$$\left.\frac{\Delta Rank(X_i^{capped})}{\Delta E_i}\right|_{T_i>S} \gg \left.\frac{\Delta Rank(X_i)}{\Delta E_i}\right|_{T_i<S}$$

The quantitative simulation showed that the ceiling effect is largely hidden at the overall population level, with the Spearman correlation between true ability and observed scores remaining very high (ρ ≥ 0.99), since most candidates are still below the constrained upper range. However, among the top 10% elite group, the ceiling effect severely disrupts ranking accuracy. The Spearman rank correlation declines sharply from 0.764 at a mean of 13, to 0.331 at 14, and to -0.005 at 15, making selection nearly equivalent to a random lottery. The Mean Absolute Rank Difference also confirms this instability, increasing from 1,939 at a mean of 13, to 3,461 at 14, and 5,376 at 15, indicating substantial and arbitrary reshuffling of top candidates' ranks.

## Conclusion

The quantitative simulation supports these theoretical assertions. At the aggregate population level, the ceiling effect remains masked; the overall Spearman rank correlation between true ability and observed scores stays exceptionally stable (ρ>=0.99) across all simulated means because most candidates fall within the unconstrained range of the distribution. However, within the competitive top 10% elite cohort, the ceiling effect causes a sharp decline in ranking integrity. When the population mean is 13, the Spearman correlation between true ability and observed ranks under Ceiling Model 2 is 0.764. As the mean rises to 14 and more elite candidates approach the scale ceiling, the rank correlation drops to 0.331. Critically, when the mean reaches 15, the Spearman correlation falls to -0.005, rendering selection nearly equivalent to a random lottery. The Mean Absolute Rank Difference further supports this pattern. Within the top 10% elite cohort, the average rank displacement increases from 1,939 positions at a mean of 13 to 3,461 at a mean of 14, and then to 5,376 at a mean of 15, revealing substantial and arbitrary rank shifts among top candidates.

## Ethical considerations

This article was conducted based on the analysis of data related to the National University Entrance Examination, and no independent contact with participants was made for research purposes during the course of the study.

## Funding

This study was financially supported by the National Organization for Educational Testing of Iran.

## Author contributions

The first author conceptualized and supervised the study. The second author drafted the manuscript, conducted the statistical analyses, and revised the text. The third author performed the simulations and followed up on the proposed revisions. All authors contributed to and approved the final version of the manuscript.

## Conflict of interest

The authors also declare that there was no conflict of interest at any stage of the present study.

مقاله پژوهشی

# داده‌های بیشتر، روایی کمتر:
# پارادوکس ورود سوابق تحصیلی به معادلات پذیرش دانشجو

ابراهیم خدایی[۱*]، مهدی کروندی رنانی[۱]، مجتبی جهانی‌فر[۲]
۱. گروه روش ها و برنامه ریزی آموزشی و درسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران.
۲. گروه علوم تربیتی، دانشکده علوم تربیتی و روانشناسی، دانشگاه شهید چمران اهواز، اهواز، ایران.

**اطلاعات مقاله:**
تاریخ دریافت: ۱۴۰۵/۰۳/۱۲
تاریخ داوری: ۱۴۰۵/۰۳/۱۴
تاریخ پذیرش: ۱۴۰۵/۰۳/۲۰

**واژه‌های کلیدی:**
روایی، کنکور، اثر سقف، شبیه‌سازی، امتحانات نهایی.



ناشر: دانشگاه زنجان

## چکیده

**زمینه:** آزمون سراسری ورود به دانشگاه در ایران به‌عنوان آزمونی سرنوشت‌ساز، نمره داوطلبان را از ترکیب کنکور (آزمون هنجار – مرجع) و سوابق تحصیلی (آزمون ملاک – مرجع) محاسبه می‌کند. هدف این پژوهش، ارزیابی انتقادی روایی این ادعا است که ترکیب این دو نمره، سنجشی جامع‌تر و معتبرتر از توانایی حقیقی داوطلبان فراهم می‌آورد.

**روش:** روش پژوهش مبتنی بر طرحی تحلیلی – شبیه‌سازی دو مرحله‌ای است. در فاز اول، با استفاده از چارچوب روایی استدلال‌محور کین، پیش‌فرض‌های نظری این ادعا نقد شد. در فاز دوم، شبیه‌سازی آماری برای نمایش کمّی تأثیر اثر سقف بر رتبه‌بندی داوطلبان انجام گرفت.

**یافته‌ها:** نتایج شبیه‌سازی نشان داد که اثر سقف، قدرت تمایزگذاری سوابق تحصیلی را در داوطلبان ممتاز (۱۰ درصد برتر) به‌شدت تضعیف می‌کند. هم‌زمان با نزدیک شدن میانگین نمرات جامعه به سقف مقیاس، همبستگی رتبه‌ای میان توانایی واقعی و نمره مشاهده‌شده در این گروه به سمت صفر میل می‌کند و رتبه‌بندی به‌شدت تحت تأثیر خطای تصادفی قرار می‌گیرد.

**نتیجه‌گیری:** نتیجه‌گیری این پژوهش نشان می‌دهد که ادعای بنیادین سیاست ترکیب دو آزمون، از نظر روایی قابل دفاع نیست. اثر سقف در امتحانات نهایی، این بخش از نمره کل را به منبعی اطلاعاتی ناکارآمد و پرنوسان (خطا) برای تمایزگذاری میان داوطلبان برتر تبدیل می‌کند. این امر نه‌تنها اعتبار کل سیستم رتبه‌بندی را در حساس‌ترین بخش رقابت تضعیف می‌کند، بلکه با وارد کردن خطای سیستماتیک، به کاهش عدالت در فرایند پذیرش دانشجو منجر می‌شود؛ بنابراین، ترکیب فعلی نمرات کنکور و سوابق تحصیلی عملی سؤال‌برانگیز در روش‌شناختی است که نیازمند بازنگری جدی در سیاست‌گذاری‌های نظام سنجش و پذیرش است تا از تضییع حقوق داوطلبان شایسته جلوگیری شود.

**استناد:** خدایی، ا.، و کروندی رنانی، م، و جهانی‌فر، م. (۱۴۰۵). داده‌های بیشتر، روایی کمتر: پارادوکس ورود سوابق تحصیلی به معادلات پذیرش دانشجو. *سنجش روانی تربیتی*، ۲(۱)، ۱–۲۰.
https://doi.org/10.30470/ijpa.2026.736668

* **نویسنده مسئول:** ابراهیم خدایی
**نشانی:** دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران.
**پست الکترونیکی:** khodaie@ut.ac.ir

## مقدمه

«روایی[1] بنیادی‌ترین ملاحظه در فرایند ساخت و ارزیابی آزمون‌ها است» (انجمن تحقیقات آموزشی[2]، انجمن روان‌شناسی آمریکا[3] و مجمع ملی اندازه‌گیری در آموزش[4]، ۲۰۱۴: ۱۱). این جمله یا مشابه آن را می‌توان پرتکرارترین گزاره در ادبیات آزمون‌های روانی و آموزشی دانست (برای چند مثال به مارکوس[5] و برسبوم[6]، ۲۰۱۳؛ نیوتون[7] و شاو[8] ۲۰۱۴؛ مسیک[9]، ۱۹۸۹؛ کرونباخ[10]، ۱۹۸۸ و کلی[11]، ۱۹۲۷، نگاه کنید)؛ علاوه بر این، باید تأکید داشت که هرچه آزمون‌ها از سطح کاربردهای فردی فراتر بروند و به ابزار تصمیم‌گیری‌های کلان آموزشی و اجتماعی بدل شوند، توجه به روایی آن‌ها اهمیتی دوچندان می‌یابد. همان‌طور که کِلی بیان کرده است:

> «تا زمانی که فردی آزمون یا امتحانی را صرفاً برای مقاصد شخصی خود طراحی و به کار گیرد، پرسش از روایی اساساً مطرح نمی‌شود؛ اما هنگامی که معلمان از آزمون‌ها برای اهداف آموزشی استفاده می‌کنند، به دلیل پیوند عمیق این اهداف با آینده و رفاه دانش‌آموزان، نمی‌توان روایی آزمون را بدون بررسی رها کرد. در واقع، خود دانش‌آموز، به‌ویژه دانش‌آموز امروزی با رویکرد خودانگیخته و فعال در سنت فکری دیویی، نیروی محرک اصلی پشت شکل‌گیری و گسترش جنبش توجه به روایی است» (کلی، ۱۹۲۷: ۱۳).

بنابراین، این حق ذاتی ذی‌نفعان در هر آزمونی است تا از شواهد روایی آن اطلاع پیدا کنند و یا به بیان کین[12]: «ادعاهای عمومی نیازمند توجیه عمومی هستند» (کین، ۲۰۱۳: ۱). همچنین باید تأکید داشت که ارائه شواهد روایی هر آزمون بر عهده توسعه‌دهندگان و کاربران (به‌کاربرندگان) نهایی آزمون است (انجمن تحقیقات آموزشی، انجمن روان‌شناسی آمریکا و مجمع ملی اندازه‌گیری در آموزش، ۲۰۱۴).

آزمون سراسری ورود به دانشگاه (در مقطع کارشناسی) در ایران که از ترکیب نمره کنکور و نمره امتحانات نهایی تشکیل می‌شود، به‌عنوان سرنوشت‌سازترین آزمون در نظام آموزشی کشور شناخته می‌شود. این آزمون نه‌تنها مسیر و کیفیت ادامه تحصیل دانش‌آموزان را تعیین می‌کند، بلکه فرصت‌های شغلی و موقعیت‌های اجتماعی آینده آن‌ها را نیز تحت تأثیر قرار می‌دهد. اهمیت این آزمون (آزمون‌ها) از این جهت است که نتایج آن به تصمیم‌گیری‌های کلان آموزشی و پذیرش در دانشگاه‌ها منتهی می‌شود؛ بنابراین، با توجه به ابعاد آن، این آزمون (آزمون‌ها) نقش تعیین‌کننده‌ای در زندگی دانش‌آموزان دارد؛ لذا رواسازی[13] این آزمون (آزمون‌ها) برای حفظ اعتبار سیستم پذیرش و همچنین اعتماد ذی‌نفعان به آن اهمیتی حیاتی دارد.

در زمینه رواسازی آزمون‌ها، به‌ویژه آزمون‌های ترکیبی و کلان، مانند آزمون سراسری، هنوز همگرایی روش‌شناختی کاملی وجود ندارد (مارکوس و برسبوم، ۲۰۱۳ و نیوتون و شاو، ۲۰۱۵). این امر ناشی از پیچیدگی‌های ذاتی مفهوم روایی و همچنین سنجش روانی است. تعریف‌های متعددی را می‌توان برای روایی ذکر کرد. قدیمی‌ترین تعریف روایی به این شرح است که: مسئله روایی این است که آیا یک آزمون واقعاً آنچه ادعای اندازه‌گیری آن را دارد، اندازه‌گیری می‌کند یا خیر؟ (کلی، ۱۹۲۷: ۱۴). این تعریف به‌صورت ضمنی اشاره به بحث حقیقت[14] در ذات مسئله روایی دارد (برسبوم و ویسن[15]، ۲۰۱۶). با این حال، تعریف‌های بعدی (برای مثال: مسیک، ۱۹۸۹)، روایی را بر اساس شواهد[16] معرفی کرده‌اند: آیا برای تفسیر یک نمره معیّن، دلایل تجربی و نظری کافی در دست داریم؟ این دو پرسش که در اصل ممکن است گاهی ناسازگار باشند، با پرسش سوم نیز همراه شده‌اند: آیا استفاده از آزمون، با توجه به نتایج و پیامدهای آن، به‌طور کلی قابل توجیه است یا خیر؟ (کین، ۲۰۱۳). این تعریف سوم فراتر از شواهد و یا حقیقت است و به اهداف، ایدئولوژی و اخلاقیات مربوط می‌شود (برسبوم و ویسن، ۲۰۱۶).

بخشی از اختلافات موجود در تعریف روایی را می‌توان به تفاوت در فهم و تلقی از سنجش روانی نسبت داد. برخی پژوهشگران آزمون‌ها را ابزارهایی صرفاً برای اندازه‌گیری ویژگی‌ها و توانایی‌ها می‌دانند. به بیان دقیق‌تر، روایی مفهومی مرتبط با حقیقت اندازه‌گیری است و بنابراین، تنها درمورد آزمون‌هایی قابل طرح است که به‌طور مشخص با هدف اندازه‌گیری یک ویژگی یا توانایی طراحی شده‌اند (برسبوم و همکاران، ۲۰۰۴). از این منظر، روایی معادل صحت اندازه‌گیری تلقی می‌شود. در مقابل، دیدگاه عمل‌گرایانه[17] معتقد است آزمون‌ها می‌توانند هر نوع ابزاری را شامل شوند که کارایی و فایده آن برای اهداف مشخص قابل اثبات باشد. از این رو، مفهوم روایی فراتر از اندازه‌گیری صرف و در ارتباط با توجیه کاربرد و تفسیر نمرات حاصل از آزمون معنا پیدا می‌کند (کین، ۲۰۰۶ و مسیک، ۱۹۸۹).

---

[1] validity
[2] American Educational Research Association
[3] American Psychological Association
[4] National Council on Measurement in Education
[5] Markus
[6] Borsboom
[7] Newton
[8] Shaw
[9] Messick
[10] Cronbach
[11] Kelley
[12] Kane
[13] validation
[14] truth
[15] Wijsen
[16] evidence
[17] pragmatic

با توجه به اینکه آزمون سراسری ماهیتاً آزمون انتخاب و گزینش است و نه ابزاری با ادعای مستقیم در خصوص اندازه‌گیری (کروندی رنانی، در حال داوری) و نیز با در نظر گرفتن پیامدهای فردی و اجتماعی گسترده‌ای که به همراه دارد، به‌کارگیری چارچوبی عمل‌گرایانه برای رواسازی (باید دقت شود که سنجش روایی نیز بخشی از فرایند رواسازی در نظر گرفته می‌شود) مناسب‌تر به نظر می‌رسد. به همین دلایل، در ادامه چارچوب روایی کین (۲۰۰۶، ۲۰۱۳ و ۲۰۱۶) به‌صورت مختصر معرفی شده و سپس روایی آزمون سراسری در این چارچوب مورد بحث قرار گرفته است.

چارچوب روایی کین که به نام رویکرد استدلال‌محور به روایی[1] شناخته می‌شود، یکی از منسجم‌ترین و پرنفوذترین چارچوب‌ها در سنجش و اندازه‌گیری آموزشی و روان‌شناختی است (برای مثال، این رویکرد مبنای آخرین سری کتاب استانداردهای سنجش آموزشی و روان‌شناختی بوده است و یا فصل روایی کتاب اندازه‌گیری آموزشی که به‌عنوان کتاب مقدس این حوزه شناخته می‌شود، توسط کین نگاشته شده است). هسته اصلی این رویکرد، تغییر کانون توجه از رواسازی آزمون به رواسازی تفسیرها و کاربردهای نمرات حاصل از آزمون است. بر این اساس، ادعاهای مطرح‌شده باید به‌صورت استدلالی منسجم بیان شوند که در آن، تمامی استنتاج‌ها و مفروضات لازم برای رسیدن از پاسخ‌های خام آزمون به تفاسیر و کاربردهای نهایی مشخص شده باشد. در نهایت، رواسازی در این چارچوب، ارزیابی انسجام، کامل بودن و معقول بودن این استدلالِ یکپارچه است (کین، ۲۰۰۶ و ۲۰۱۳).

کین (۲۰۱۲) ریشه‌های رویکرد استدلال‌محور به روایی را به مقاله پراستناد کرونباخ و میهل[2] (۱۹۵۵) با نام روایی سازه در آزمون‌های روان‌شناختی[3] نسبت داده است:

> ۱) کرونباخ و میهل (۱۹۵۵) فرض می‌کنند که این تفسیری است که باید رواسازی شود و تفسیر پیشنهادی به‌تفصیل شرح داده خواهد شد.
> ۲) کرونباخ و میهل (۱۹۵۵) بدیهی می‌دانند که ارزیابی تفسیر پیشنهادی به‌جای مطالعه‌ای واحد، شامل یک برنامه پژوهشی گسترده خواهد بود.
> ۳) تمرکز کرونباخ و میهل (۱۹۵۵) بر ارزیابی نظریه‌ای که یک سازه را تعریف می‌کند، این ضرورت را مطرح ساخت که تفسیر پیشنهادی باید در برابر تفسیرهای رقیب نیز مورد ارزیابی قرار گیرد.
>
> رویکرد مبتنی بر استدلال از این سه اصل توسعه یافته است (کین، ۲۰۱۲: ۶۸).

با وجود اینکه رویکرد کین شباهت‌هایی به چارچوب‌های روایی پیشین (ازجمله چارچوب کرونباخ و میهل، ۱۹۵۵ و مسیک، ۱۹۸۹) دارد، تفاوت‌های کلیدی و مهمی نیز بین آن‌ها دیده می‌شود؛ برای مثال، مسیک تمام فرایندهای رواسازی را زیر چتر روایی سازه[4] جای داده بود و هرگونه شواهدی را با توجه به تفسیر سازه آزمون معنا می‌داد. از سوی دیگر، کین (۲۰۱۲) اساساً در چارچوب خود به‌صورت عمدی از به‌کارگیری اصطلاح سازه خودداری کرده است و استفاده از آن را غیرسازنده می‌داند؛ چراکه اصطلاح سازه چنان معنا و کاربردهای گسترده‌ای یافته است که نمی‌توان درک درستی از سازه مورد بحث به دست آورد (کین، ۲۰۱۲: ۶۷).

این رویکرد بر دو استدلال مکمل استوار است: استدلال تفسیر و کاربرد[5] (IUA) که در این مرحله ادعاها مشخصاً بیان می‌شوند و استدلال روایی[6] که در آن این ادعاها ارزیابی می‌شوند (کین، ۲۰۱۳).

گام نخست، تدوین استدلال تفسیر و کاربرد (IUA) است؛ یعنی ترسیم شبکه‌ای از استنتاج‌ها و پیش‌فرض‌هایی که آزمونگر را از عملکرد مشاهده‌شده فرد در موقعیت آزمون به تفسیرهای گسترده‌تر و در نهایت تصمیم‌های عملی هدایت می‌کند؛ برای نمونه، از مشاهده پاسخ‌ها به نمره خام می‌رسیم (استنتاج نمره‌گذاری) و سپس از نمره خام به برآورد توانایی در دامنه بزرگ‌تری از تکالیف تعمیم می‌دهیم (استنتاج تعمیم). ممکن است این برآورد به پیش‌بینی عملکرد در آینده یا موقعیت‌های دیگر بسط یابد (استنتاج برون‌یابی) یا به شکل سازه‌های نظری تفسیر شود (استنتاج علّی) و در نهایت، نمره به تصمیمی آموزشی، شغلی یا بالینی منجر شود (استنتاج تصمیم). هریک از این مراحل بر مفروضاتی استوار است که باید شناسایی و در فرایند رواسازی مورد آزمون قرار گیرند.

گام دوم، ساخت استدلال روایی است که بر پایه IUA شکل می‌گیرد. استدلال روایی شامل ارزیابی انسجام منطقی و شواهد تجربی‌ای است که از استنتاج‌ها و مفروضات IUA پشتیبانی می‌کنند. در اینجا اصل مهم این است که هرچه ادعاها مبتنی بر نمره بلندپروازانه‌تر باشند، نیازمند شواهد بیشتر و قوی‌تری خواهند بود؛ برای مثال، ادعای محدود به توانایی اجرای تکالیف مشابه با مقوله‌های آزمون به شواهد اندکی نیاز دارد، در حالی که تفسیر در سطح سازه‌های نظری یا پیش‌بینی موفقیت تحصیلی آینده مستلزم شواهد نظری و تجربی گسترده است. همچنین کین تأکید می‌کند که پیامدهای استفاده از

[1] argument-based approach to validity
[2] Meehl
[3] construct validity in psychological tests
[4] construct validity
[5] Interpretation/Use Argument
[6] Validity Argument

نمره‌ها بخشی جدایی‌ناپذیر از رواسازی هستند؛ به این معنا که اگر کاربرد نمره‌ها پیامدهای منفی جدی به همراه داشته باشد، حتی با وجود شواهد تجربی قوی، آن کاربرد ناموجه خواهد بود.

نکته‌ای که در بحث روایی همواره مطرح بوده، این است که فرایند رواسازی ذاتاً می‌تواند بی‌پایان باشد. چرا؟ به این دلیل که هر استنتاج در IUA بر مجموعه‌ای از مفروضات تکیه دارد و هر مفروضه می‌تواند نیازمند شواهد و بررسی‌های تازه باشد و چرخه‌ای بی‌نهایت ایجاد کند (به دوهیم[1] (۲۰۲۱) نگاه کنید). همان‌گونه که کرونباخ و میهل نیز اشاره کرده بودند، اگر بخواهیم برای همه استنتاج‌ها و همه مفروضات، پشتوانه تجربی فراهم کنیم، هیچ‌گاه به پایان نخواهیم رسید؛ چراکه زنجیره‌ای نامحدود از پرسش‌ها و آزمون‌ها به وجود می‌آید. در پاسخ به این مشکل، کین پیشنهاد کرده است که در مرحله استدلال روایی باید بر بحث‌برانگیزترین ادعاها و مفروضه‌ها تمرکز شود. این بدین معنا نیست که بررسی دیگر تفاسیر و کاربردها به حمایت از رواسازی آزمون کمک نمی‌کند، بلکه به عدم تقارن در شواهد حمایت‌کننده و ردکننده روایی اشاره دارد؛ به بیان دیگر، هرچند شواهد همگرا با ادعاهای مطرح‌شده در خصوص آزمون استدلال روایی، آن را تقویت (نه تأیید کامل) می‌کنند، در مقابل، یک مورد از نقض پیش‌فرض‌ها و یا ادعاها می‌تواند به‌صورت جدی و کامل استدلال روایی را زیر سؤال ببرد. تمرکز بر ادعاهای بحث‌برانگیز از آن جهت حائز اهمیت است که یک مفروضه و یا ادعای توجیه‌ناپذیر باعث از دست رفتن کل روایی یک IUA می‌شود. به بیان کین:

> «قیود احتمال[2] به‌طور میانگین خنثی نمی‌شوند و وجود ضعفی جدی در هریک از استنتاج‌های اصلی می‌تواند کل استدلال را تضعیف کند؛ حتی اگر سایر استنتاج‌ها به‌خوبی پشتیبانی شده باشند. استنتاج‌ها را می‌توان به‌سان دهانه‌های پل در نظر گرفت که از عملکردهای آزمون آغاز و به نتایج و تصمیم‌های مستتر در تفسیر و کاربرد پیشنهادی منتهی می‌شوند. اگر یکی از دهانه‌ها فروبریزد، کل پل از کار می‌افتد؛ حتی اگر سایر دهانه‌ها به‌خوبی استوار باشند» (کین، ۲۰۱۳: ۱۳).

## رواسازی آزمون سراسری

پیش از پرداختن به رواسازی آزمون سراسری، لازم است تا با کارکرد و اهداف این آزمون (آزمون‌ها) دقیق‌تر آشنا شویم. آزمون سراسری در ایران از دو بخش معدل نمرات کتبی متوسطه دوم و نمرات کنکور سراسری ایجاد می‌شود. در حال حاضر، مراحل زیر برای ساخت نمره کل داوطلبان آزمون سراسری طی می‌شود:

**۱. محاسبه نمره خام آزمون کنکور سراسری:** هر پاسخ درست ۳ امتیاز و هر پاسخ غلط ۱- امتیاز دارد. درصد خام با فرمول ۱۰۰ × [(۳ × درست) – غلط]/ (۳ × کل سؤالات) برای هر درس به‌صورت مستقل محاسبه می‌شود.

**۲. تبدیل نمرات خام به نمرات تراز مقیاس‌شده:** در این مرحله نمرات خام (حاصل از مرحله قبل) در ابتدا به نمرات مقیاس نرمال تبدیل و سپس تراز می‌شوند. در این مرحله در ابتدا توزیع فراوانی نسبی نمره‌ها محاسبه و سپس در گام دوم، با استفاده از فراوانی نسبی و توزیع تراکمی نمره‌ها، رتبه درصدی متناظر با هر نمره محاسبه می‌شود. در گام سوم، نمره استاندارد متناظر با رتبه درصدی به دست آمده در توزیع نرمال، با استفاده از معکوس تابع تراکمی زیر محاسبه می‌شود:

$$\Phi(z) = \frac{\hat{Q}(y)}{100} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\omega^2/2}\, d\omega \qquad (1)$$

در آن $\Phi(z)$ تابع تراکمی نرمال استاندارد، $\omega$ متغیر انتگرال‌گیری با دامنه $-\infty$ تا $z$ است و $\hat{Q}(y)$ نشان‌دهنده رتبه درصدی است (جهانی‌فر و همکاران، ۲۰۱۸). در نهایت، نمره‌های مقیاس‌شده با استفاده از فرمول ۲، به نمرات تراز آزمون سراسری با میانگین ۵۰۰۰ و انحراف استاندارد ۲۲۵۰ تبدل می‌شوند:

$$sc = \sigma(sc)z + \mu(sc) \qquad (2)$$

در آن $SC$ نمره‌های مقیاس‌شده هستند.

**۳. وزن‌دهی دروس:** هر درس با ضریبی مشخص و از پیش تعیین‌شده‌ای وزن‌دهی می‌شود؛ مثلاً ریاضی در گروه آزمایشی ریاضی، ضریب ۱۲ دارد (ضرایب در جدول ۱ نمایش داده شده‌اند).

**۴. انتخاب بالاترین تراز از بین نمرات آزمون‌های دو ساله:** در حال حاضر، هرساله آزمون سراسری در دو مقطع برگزار می‌شود تا برای داوطلبان امکان جبران نتایج را فراهم کند و همچنین به سازمان سنجش اجازه دهد تا برآوردی دقیق‌تر از توانایی‌های داوطلبان داشته باشد؛ علاوه بر این، نتایج هر آزمون تا دو سال معتبر خواهد بود؛ این بدان معنا است که برای مثال، نتایج حاصل‌شده در آزمون سراسری ۱۴۰۳ در سال ۱۴۰۴ نیز معتبر بوده است؛

[1] Duhem

[2] qualifiers

اما نکته مهم‌تر این است که از بین چهار آزمون سراسری معتبر برای پذیرش در سال تحصیلی پیش رو، بالاترین نمره تراز هر داوطلب به‌عنوان مبنای تصمیم‌گیری و رتبه‌بندی در نظر گرفته می‌شود:

$$NS_{final} = MAX\begin{bmatrix} NS_{11} & NS_{21} \\ NS_{12} & NS_{22} \end{bmatrix} \qquad (3)$$

در فرمول بالا، $NS$ نشان‌دهنده نمره کل آزمون سراسری است که میانگین وزنی نمرات تراز هر درس با ضریب مشخص آن درس (جدول ۱) ایجاد می‌شود. زیرنویس‌ها نیز اشاره به سال و نوبت آزمون دارد؛ برای مثال، $NS_{12}$ اشاره به نمره کل آزمون سراسری سال اول در نوبت دوم دارد.

**۵. ترکیب با سوابق تحصیلی:** برای آزمون ۱۴۰۴، نمره نهایی با فرمول (نمره آزمون × ۰٫٤) + (نمره سوابق تحصیلی × ۰٫٦) محاسبه می‌شود و بدین ترتیب، نمره کل نهایی برای رتبه‌بندی داوطلبان ایجاد می‌شود که این عدد به نزدیک‌ترین عدد طبیعی گرد می‌شود و مبنای رتبه‌بندی داوطلبان را فراهم می‌کند. شایان ذکر است که نمرات سوابق تحصیلی از معدل نمرات آزمون‌های نهایی دوره متوسطه دوم با وزن‌های مشخص هر درس ایجاد می‌شود.

جدول ۱

ضرایب دروس به تفکیک هر گروه آزمایشی در کنکور سراسری

| رشته | درس | ضریب |
|---|---|---|
| ریاضی – فیزیک | ریاضی | ۱۲ |
| | فیزیک | ۹ |
| | شیمی | ۷ |
| علوم تجربی | زمین‌شناسی | ۱ |
| | ریاضی | ۷ |
| | زیست‌شناسی | ۱۲ |
| | فیزیک | ۷ |
| | شیمی | ۹ |
| علوم انسانی | ریاضی | ۶ |
| | اقتصاد | ۲ |
| | ادبیات فارسی (تخصصی) | ۸ |
| | عربی (تخصصی) | ۵ |
| | تاریخ و جغرافیا | ۵ |
| | علوم اجتماعی | ۵ |
| | فلسفه و منطق | ۵ |
| | روان‌شناسی | ۲ |
| هنر | درک عمومی هنر | ۱۲ |
| | درک عمومی ریاضی – فیزیک | ۵ |
| | خلاقیت بصری و تخیلی | ۳ |
| زبان‌های خارجی | زبان تخصصی | ۱ |

## استدلال تفسیر و کاربرد آزمون سراسری

همان‌طور که پیش‌تر بیان شد، اولین گام در رواسازی آزمون‌ها، مشخص کردن ادعاها و پیش‌فرض‌های مرتبط با آن‌ها است. معمولاً این ادعاها باید توسط توسعه‌دهندگان آزمون صراحتاً بیان و ارزیابی شوند. با این حال، این نکته در خصوص آزمون سراسری صدق نمی‌کند. به همین دلیل، ما در این بخش و با توجه به اطلاعات بیان‌شده در خصوص آزمون سراسری، به بیان برخی از این ادعاها خواهیم پرداخت.

### استنتاج نمره‌گذاری[1]

این استنتاج، عملکرد خام داوطلبان را به یک نمره کل نهایی برای رتبه‌بندی تبدیل می‌کند. این فرایند شامل چندین مرحله است که تنها چند مورد به‌صورت نمونه مطرح شده است:

۱. ابتدا پاسخ‌های صحیح و غلط هر داوطلب در هر درس (کنکور) به یک نمره درصد خام تبدیل می‌شود. این مرحله بر این فرض (ادعا) استوار است که فرمول محاسبه درصد خام (با نمره منفی برای پاسخ غلط) اثر حدس‌پذیری سؤالات را خنثی می‌کند و به‌درستی نمایانگر دانش داوطلب در آن درس است.

۲. سپس نمرات خام هر درس با استفاده از روش‌های آماری (محاسبه رتبه درصدی و استفاده از توزیع نرمال) به نمرات تراز با میانگین و انحراف استاندارد ثابت تبدیل می‌شوند. این مرحله بر این فرض استوار است که نرمال‌سازی، نمرات دروس مختلف و گروه‌های آزمایشی متفاوت را قابل مقایسه و اثر دشواری متفاوت آزمون‌ها را خنثی می‌کند.

۳. از بین تمام نمرات کل کسب‌شده توسط داوطلب در آزمون‌های معتبر طی دو سال، بالاترین نمره به‌عنوان نمره نهایی کنکور انتخاب می‌شود. ادعا: بالاترین نمره، دقیق‌ترین برآورد از توانایی داوطلب است و این روش به کاهش خطای اندازه‌گیری و افزایش عدالت کمک می‌کند.

۴. نمره نهایی کنکور با نمره سوابق تحصیلی (معدل امتحانات نهایی) با وزن‌های مشخص (مثلاً ۶۰٪ سوابق و ۴۰٪ کنکور برای آزمون ۱۴۰۴) ترکیب و نمره کل نهایی برای رتبه‌بندی داوطلبان ایجاد می‌شود. ادعا: ترکیب این دو منبع اطلاعاتی (آزمون هنجار – مرجع و ملاک – مرجع) ارزیابی جامع‌تر و معتبرتری از شایستگی تحصیلی داوطلب ارائه می‌دهد و وزن‌های تخصیص داده‌شده بهینه هستند.

فراتر از ادعاهای مطرح‌شده در خصوص نمره‌گذاری، ادعاهای مرتبط با استنتاج تعمیم[2]، برون‌یابی[3] و تصمیم‌گیری[4] نیز قابل شناسایی هستند؛ اما از آنجا که پیش‌تر به این موضوع در مقالات دیگر پرداخته شده است و به هدف این پژوهش نیز مربوط نمی‌شود، از بیان مجدد آن‌ها خودداری کردیم.

در نهایت، هدف ما در این پژوهش بررسی ادعای چهارم (ترکیب نمرات کنکور سراسری و سوابق تحصیلی (آزمون هنجار – مرجع و ملاک – مرجع) ارزیابی جامع‌تر و معتبرتری از شایستگی تحصیلی داوطلب ارائه می‌دهد و وزن‌های تخصیص داده‌شده بهینه هستند) در خصوص آزمون سراسری و بررسی استدلال روایی آن است. همان‌طور که در چارچوب روایی کین مطرح شد، اولویت در مرحله استدلال روایی با ادعاهایی است که بیش از سایر ادعاها بحث‌برانگیز باشند. انتخاب این بخش از IUA برای رواسازی نیز با توجه به همین دیدگاه بوده است.

## روش

این پژوهش با هدف ارزیابی انتقادی – روایی ترکیب نمرات کنکور سراسری و سوابق تحصیلی در نظام پذیرش دانشجوی ایران، از یک طرح پژوهشی تحلیلی – شبیه‌سازی در دو فاز متوالی بهره می‌برد. مبنای نظری این تحقیق، چارچوب روایی استدلال محور کین است که در آن روایی آزمون از طریق ارزیابی انسجام و استحکام ادعاهای مطرح‌شده در خصوص تفسیر و کاربرد نمرات آن سنجیده می‌شود.

### فاز اول: بررسی تحلیلی و واکاوی استدلال روایی

از روش تحلیلی برای واکاوی استدلال تفسیر و کاربرد (IUA) آزمون سراسری استفاده شد. تمرکز این بخش بر یکی از بحث‌برانگیزترین ادعاهای این سیستم بود: ترکیب نمرات کنکور (آزمون هنجار – مرجع) و سوابق تحصیلی (آزمون ملاک – مرجع)، ارزیابی جامع‌تر و معتبرتری از شایستگی تحصیلی داوطلب ارائه می‌دهد و وزن‌های تخصیص داده‌شده بهینه هستند. این انتخاب، مطابق با توصیه کین مبنی بر اولویت دادن به ارزیابی مناقشه‌برانگیزترین مفروضه‌ها در یک استدلال روایی صورت گرفت. در این مرحله، با استناد به مبانی نظری روان‌سنجی، به‌ویژه نظریه کلاسیک آزمون، پیش‌فرض‌های این ادعا مورد نقد قرار گرفت.

### فاز دوم: شبیه‌سازی آماری

به‌منظور ملموس‌سازی نتایج تحلیل نظری و نمایش کمّی تأثیرات اثر سقف، در فاز دوم، مطالعه شبیه‌سازی طراحی و اجرا شد. جزئیات این شبیه‌سازی به شرح زیر است:

**۱. تولید داده‌ها:** جامعه آماری‌ای متشکل از ۱۰۰,۰۰۰ فرد شبیه‌سازی شد. نمرات واقعی این افراد از توزیعی نرمال با انحراف استاندارد ثابت (۴) و سه میانگین متفاوت (۱۳، ۱۴ و ۱۵) برای مدل‌سازی سطوح مختلف میانگین توانایی جامعه، تولید گردید. به هر نمره واقعی، یک خطای اندازه‌گیری تصادفی (با توزیع نرمال، میانگین صفر و انحراف استاندارد ۰.۵) افزوده شد.

[1] scoring
[2] generalization
[3] extrapolation
[4] decision

**۲. مدل‌سازی اثر سقف:** سه نوع نمره مشاهده‌شده ایجاد شد: ۱) نمره بدون اعمال سقف، ۲) مدل سقف اول که در آن نمرات در بازه ۰ تا ۲۰ پس از افزودن خطا محدود شدند و ۳) مدل سقف دوم که در آن نمره حقیقی ابتدا محدود، سپس خطا به آن افزوده شده و در نهایت، نتیجه مجدداً محدود گردیده است.

**۳. تحلیل داده‌ها:** برای تحلیل تأثیر اثر سقف، رتبه‌های افراد در هر چهار حالت (نمره حقیقی و سه مدل مشاهده‌شده) محاسبه گردید. تحلیل‌ها به‌صورت جداگانه برای کل نمونه و دو زیرگروه «بالا» (۱۰ درصد برتر بر اساس نمره حقیقی) و «سایر» (۹۰ درصد باقی‌مانده) انجام شد. برای مقایسه، از دو شاخص آماری استفاده شد: ۱) همبستگی رتبه‌ای اسپیرمن برای سنجش میزان هم‌راستایی رتبه‌بندی‌ها و ۲) میانگین قدر مطلق تفاوت رتبه‌ها به‌عنوان معیاری برای شدت جابه‌جایی در رتبه‌بندی.

## یافته‌ها

### استدلال روایی مربوط به ادعای چهارم

ادعای چهارم در استدلال تفسیر و کاربرد آزمون سراسری چنین بیان می‌کند که ترکیب نمرات کنکور سراسری (به‌عنوان آزمون هنجار – مرجع) و سوابق تحصیلی (به‌عنوان آزمون ملاک – مرجع)، تصویری جامع‌تر و معتبرتر از شایستگی تحصیلی داوطلب فراهم می‌آورد و وزن‌های تخصیص‌یافته به این دو منبع نیز بهینه هستند. ظاهر این ادعا قانع‌کننده به نظر می‌رسد؛ چراکه منطقی است تصور کنیم ترکیب دو منبع مستقل اطلاعات، دیدی کامل‌تر از توانایی‌های تحصیلی ایجاد کند؛ اما در چارچوب استدلال روایی کین، اعتبار این ادعا تنها زمانی پذیرفتنی است که پیش‌فرض‌های آن به‌دقت واکاوی شوند.

یکی از مهم‌ترین این مفروضات آن است که سوابق تحصیلی واقعاً قدرت افتراقی لازم برای تمایز میان داوطلبان را دارند و می‌توانند مکملی معتبر برای کنکور در امر رتبه‌بندی داوطلبان باشند؛ اما آیا با توجه به ماهیت ملاک – مرجع امتحانات نهایی، این مفروضه قابل حمایت است؟ ما استدلال می‌کنیم که چنین نیست.

آزمون‌های ملاک – مرجع به‌صورت کلی و امتحانات نهایی به شکل خاص با هدف سنجش میزان دستیابی دانش‌آموزان به معیارهای آموزشی مصوب طراحی می‌شوند و در نتیجه، پوشش مناسب دامنه مطالب مورد آموزش مهم‌ترین ملاک در بررسی کیفیت این آزمون‌ها است؛ بنابراین، سازنده آزمون ملاک – مرجع نباید نگران گزینش گویه‌ها با هدف بیشینه‌سازی واریانس باشد (کروکر[1] و آلجینا[2]، ۱۹۸۶: ۳۲۹). باید توجه داشت که همین بیشینه‌سازی واریانس هدف اساسی در ساخت آزمون‌های هنجار – مرجع به‌صورت کلی و کنکور به شکل خاص است. در آزمون‌های هنجار – مرجع، هدف اصلی تمایزگذاری میان افراد است و به همین دلیل، طراحان آزمون می‌کوشند گویه‌هایی را انتخاب کنند که بیشترین واریانس پاسخ‌ها را ایجاد و امکان تفکیک بهتر میان شرکت‌کنندگان را فراهم می‌کند. همچنین مجموع گویه‌های گنجانده‌شده در آزمون به نحوی است که طیف وسیعی از توانایی مورد سنجش را پوشش دهد؛ برای مثال، در آزمون ورودی دانشگاه، سؤال‌ها باید چنان طراحی شوند که بتوانند داوطلبان ضعیف، متوسط و قوی را از یکدیگر بازشناسی کنند. در مقابل، آزمون‌های ملاک – مرجع، مانند امتحانات نهایی، بر این امر متمرکز هستند که آیا دانش‌آموزان به سطح معیار تعیین‌شده در برنامه درسی دست یافته‌اند یا خیر؛ به بیان دیگر، اگر تمام دانش‌آموزان به یک سؤال (حتی کل سؤالات) درست پاسخ دهند و در نتیجه، واریانس آن صفر شود، این امر در آزمون ملاک – مرجع نه‌تنها ضعف محسوب نمی‌شود، بلکه نشان می‌دهد هدف آموزشی محقق شده است. از همین رو، طراحان آزمون ملاک – مرجع نباید دغدغه‌ای درباره بیشینه‌سازی واریانس داشته باشند؛ چراکه کارکرد بنیادین چنین آزمون‌هایی سنجش میزان دستیابی به معیارهای از پیش تعریف‌شده است، نه تمایزگذاری آماری بین دانش‌آموزان (نانالی[3] و برنشتاین[4]، ۱۹۹۴).

این عدم تمایزگذاری بین داوطلبان در آزمون‌های ملاک – مرجع (ازجمله امتحانات نهایی) به شکل اثر سقف[5] نمایان شده است. اثر سقف به حالتی گفته می‌شود که نمره‌ها یا مشاهدات در حد بالایی متوقف می‌شوند و امکان فراتر رفتن از آن وجود ندارد (نانالی و برنشتاین، ۱۹۹۴: ۵۷۰). باید دقت شود که اثر سقف زمانی ایجاد می‌شود که توانایی حقیقی بخشی از داوطلبان بیشتر از سقف در نظر گرفته‌شده در آزمون باشد. در امتحانات نهایی، اثر سقف باعث می‌شود تمام داوطلبانی که توانایی واقعی‌شان از سقف بالاتر است، نمره کامل و یا نزدیک به سقف دریافت کنند. در این وضعیت، نمرات فشرده‌شده در سقف، حساسیت رتبه‌بندی به خطای تصادفی را افزایش می‌دهند؛ به بیان دیگر، هرگونه تغییر کوچک ناشی از خطای تصادفی تنها می‌تواند نمره‌ها را پایین‌تر ببرد و همین ریزتفاوت‌ها به‌طور نامتناسب بر رتبه‌ها تأثیر می‌گذارند (به‌ویژه به دلیل متراکم شدن نمرات در این بازه)؛ به‌گونه‌ای که تفاوت‌های مشاهده‌شده دیگر بازتاب توانایی واقعی نیستند.

برای درک بهتر این مشکل می‌توان آن را از نگاه نظریه کلاسیک آزمون[6] (CTT) مورد توجه قرار داد. فرض کنیم نمره مشاهده‌شده $X_i$ از داوطلب $i$ از رابطه کلاسیک CTT حاصل شود:

[1] Crocker
[2] Algina
[3] Nunnally
[4] Bernstein
[5] ceiling effect
[6] Classical Test Theory

$$X_i = T_i + E_i \qquad (4)$$

در آن $T_i$ نمره حقیقی و $E_i$ خطای تصادفی با میانگین صفر و واریانس $\sigma_E^2$ است. در امتحانات نهایی اگر سقف آزمون برابر با $S$ باشد، نمره مشاهده‌شده محدود خواهد شد:

$$X_i^{capped} = \min(X_i, S) \qquad (5)$$

برای داوطلبانی که $T_i > S$، تمام نمرات برابر با سقف $S$ خواهد بود؛ بنابراین:

$$X_i^{capped} = S \qquad (6)$$

در این شرایط، واریانس مشاهده‌شده در نزدیکی سقف کاهش می‌یابد و پراکندگی نمرات تقریباً صفر می‌شود:

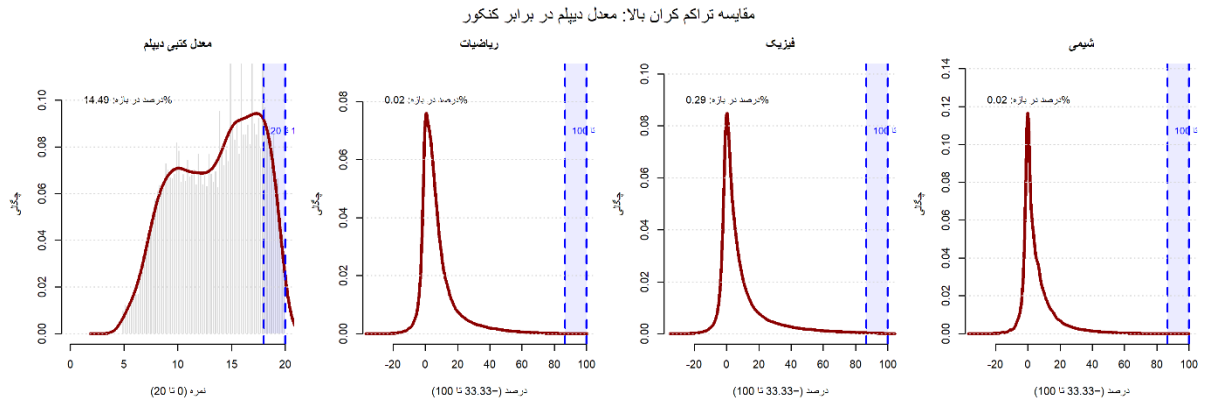
$$Var\left(X_i^{capped}\right) \ll Var(T_i) \qquad (7)$$

همچنین اگر خطای تصادفی $E_i$ باعث کاهش نمره برخی داوطلبان از سقف شود، حتی تغییرات کوچک $\Delta X_i = -E_i$ می‌تواند نسبتاً بزرگ جلوه کند؛ زیرا این تغییرات روی رتبه‌بندی در بالاترین سطح تمرکز دارند و هرگونه اختلاف کوچک که در واریانس طبیعی غیرسقف اهمیت کمتری داشت، اکنون اثر نامتناسبی بر رتبه‌ها ایجاد می‌کند؛ به بیان دیگر، حساسیت رتبه‌بندی به خطای تصادفی در این حالت افزایش می‌یابد:

$$\left.\frac{\Delta Rank(X_i^{capped})}{\Delta E_i}\right|_{T_i>S} \gg \left.\frac{\Delta Rank(X_i)}{\Delta E_i}\right|_{T_i<S} \qquad (8)$$

در ادامه، به‌منظور تقویت ادعای مطرح‌شده درباره تراکم نمرات در کران بالای توزیع نمرات امتحانات نهایی (معدل کتبی دیپلم)، تراکم نمرات معدل داوطلبان آزمون سراسری سال ۱۴۰۳ در ۱۰٪ بالاییِ بازه نمرات ممکن (بازه ۱۸ تا ۲۰) با بازه مشابه در نمرات دروس کنکورِ گروه آزمایشی ریاضی و فنی (معادل ۸۶.۶۷ تا ۱۰۰ درصد) مقایسه و بررسی شد. توزیع فراوانی این نمرات در شکل ۱ ارائه شده است.

نتایج نشان داد که در معدل کتبی دیپلم، حدود ۱۴.۴۹٪ از داوطلبان در بازه ۱۸ تا ۲۰ قرار گرفته‌اند، در حالی که سهم داوطلبان در بازه متناظر ۸۶.۶۷ تا ۱۰۰ برای دروس کنکور گروه ریاضی و فنی به‌مراتب کمتر بوده است (به‌ترتیب، ریاضیات ≈ ۰.۰۲٪، فیزیک ≈ ۰.۰۲۹٪ و شیمی ≈ ۰.۰۲٪). این اختلاف چشمگیر بیانگر انباشت قابل توجه نمرات در کران بالای توزیع معدل دیپلم و در نتیجه، کاهش قدرت تمایزدهی این شاخص در سطوح بالا (اثر سقف) است.




شکل ۱

مقایسه تراکم نمرات در کرانِ بالایی توزیع: معدل کتبی دیپلم در برابر دروس کنکور گروه ریاضی و فنی (ریاضیات، فیزیک و شیمی)

شکل توزیع نمرات دروس کنکور در گروه ریاضی و فنی در شکل ۱، الگوی متعارفی را نشان می‌دهد که معمولاً در آزمون‌های نسبتاً دشوار و تمایزبخش مشاهده می‌شود. در این آزمون‌ها بخش عمده فراوانی در نواحی میانی و پایین‌تر متمرکز است و در عین حال، دُم بالایی توزیع (کران بالا) کم‌تراکم و کشیده باقی می‌ماند. این ویژگی به این معنا است که آزمون، به‌ویژه در سطوح بالای توانایی، دچار انباشت نمره نمی‌شود و بنابراین، سقف نمره به‌عنوان محدودکننده تمایز عمل نمی‌کند.

در مقابل، توزیع معدل کتبی دیپلم در کران بالای دامنه، رفتاری متفاوت نشان می‌دهد. در این توزیع، پس از تمرکز قابل توجه در نمرات بالا، در بازه‌های انتهایی نزدیک به سقف نمره، کاهش نسبتاً ناگهانی مشاهده می‌شود. چنین الگویی (افت تند فراوانی درست در مجاورت کران بالا) لزوماً به معنای افزایش تمایزگذاری ذاتی میان داوطلبان قوی نیست؛ زیرا اگر ابزار اندازه‌گیری به‌طور ساختاری در ناحیه بالای توانایی تمایزبخش باشد، انتظار می‌رود کاهش فراوانی در کران بالا به‌صورت تدریجی و پیوسته و متناسب با کاهش تعداد افراد بسیار قوی رخ دهد، نه به شکل شکست یا افت جهشی در انتهای توزیع.

این پدیده بیش از آنکه قدرت تمایزگذاری امتحانات نهایی را نشان دهد، احتمال وجود چند سؤال کشنده را مطرح می‌کند (کیم[1]، ۲۰۲۵).

سؤالات کشنده در بستر آزمون‌هایی مطرح می‌شوند که در کران بالای توزیع نمره با اثر سقف مواجه هستند. در چنین شرایطی، طراحان گاه با افزودن تعداد محدودی سؤال بسیار دشوار که حتی برای داوطلبان قوی نیز چالش‌برانگیز است، می‌کوشند تا فاصله‌ای بین نمرات در ناحیه سقف ایجاد و از هم‌نمره شدن گسترده در بالاترین سطوح جلوگیری کنند. پیشینه مربوط نشان می‌دهد این سؤالات معمولاً با بار شناختی بالا و اتکای بیشتر به استدلال سطح بالا (فراتر از سطح قابل انتظار برای آزمون) طراحی می‌شوند و کارکردشان بیش از آنکه افزایش اطلاعات سنجش در کل دامنه باشد، ایجاد تمایز مصنوعی و موضعی در نزدیکی سقف نمره است (کیم، ۲۰۲۵؛ لی[2]، ۲۰۲۳؛ یئونگ[3] و سئو[4]، ۲۰۲۳ و کیم، ۲۰۲۳).

برای درک بهتر تأثیر سؤالات کشنده، بار دیگر می‌توان از نظریه کلاسیک آزمون و بسط آن (نظریه تعمیم‌پذیری) استفاده کرد که در آن واریانس خطا تفکیک می‌شود:

$$\sigma_X^2 = \sigma_T^2 + \underbrace{\sigma_{E'}^2 + \sigma_I^2}_{\sigma_E^2} \qquad (9)$$

در آن $\sigma_I^2$ واریانس نامرتبط با سازه یا همان دشواری مصنوعی است. هدف طراحان آزمون از طرح سؤالات کشنده، افزایش واریانس کل و کاهش احتمال وقوع اثر سقف است؛ اما این تعدیل (در ظاهر) مشکل اساسی اثر سقف بر رتبه‌بندی، یعنی تأثیرپذیری غیرطبیعی رتبه‌ها از خطای تصادفی را رفع نمی‌کند. اگر دو داوطلب $\mathrm{i}$ و $\mathrm{j}$ را در نظر بگیریم که داوطلب $\mathrm{i}$ توانایی واقعی بالاتری دارد ($\Delta T = T_i - T_j > 0$)، احتمال اینکه در آزمون، داوطلب ضعیف‌تر نمره بهتری بگیرد (خطای رتبه‌بندی)، تابعی از نویز کل سیستم ($\sigma_Z = \sqrt{\sigma_{E'}^2 + \sigma_I^2}$) خواهد بود:

$$P(X_i < X_j) = \Phi\left(\frac{-\Delta T}{\sqrt{2\left(\sigma_{E'}^2 + \sigma_I^2\right)}}\right) \qquad (10)$$

همان‌طور که در معادله بالا مشخص است، با ورود سؤالات کشنده و افزایش واریانس نامرتبط، مخرج کسر بزرگ‌تر و مقدار داخل تابع نرمال ($\Phi$) به صفر نزدیک‌تر می‌شود. این امر به معنای افزایش احتمال جابه‌جایی رتبه‌ها و نزدیک شدن نتایج آزمون به حالت تصادفی است (معادله ۱۰ بر این فرض بنا شده است که خطاهای اندازه‌گیری برای داوطلبان مختلف حول صفر، توزیع نرمال دارند. این خطاها بین افراد مستقل هستند و خطای هر فرد از توانایی یا نمره حقیقی مستقل است)؛

علاوه بر این، معمولاً سؤالات کشنده اثر سقف را خنثی نمی‌کنند، بلکه تنها باعث جابه‌جایی دامنه آن می‌شوند. همان‌طور که پیش‌تر بیان شد، در آزمونی استاندارد که دچار اثر سقف است، بخشی از جامعه آماری (داوطلبان ممتاز) دارای توانایی واقعی $T_i$ بالاتر از سقف آزمون ($S$) هستند. در نتیجه، تابع چگالی احتمال نمرات مشاهده‌شده در نقطه $S$ دارای جهشی جرمی است:

$$P(X = S) = \int_S^{\infty} f_T(t)\, dt \gg 0 \qquad (11)$$

زمانی که طراح آزمون یک یا چند سؤال کشنده با دشواری بالا و یا نامرتبط با سازه ($\delta$) طرح می‌کند، عملاً نمره مورد انتظار این داوطلبان ممتاز را به اندازه $\delta$ کاهش می‌دهد. اگر فرض کنیم سؤال کشنده چنان دشوار است که قسمت قابل توجهی از داوطلبان قادر به پاسخگویی به آن نیستند، معادله نمره برای این گروه به‌صورت زیر تغییر می‌کند:

[1] Kim
[2] Lee
[3] Yeung
[4] Seo

$$X_i^{\text{new}} \approx T_i^{\text{capped}} - \delta \qquad (12)$$

در این حالت، آن بخش از جمعیت که پیش‌تر در نمرات نزدیک به ۲۰ (سقف) متراکم شده بودند، اکنون به نمره $S - \delta$ (مثلاً ۱۸) منتقل می‌شوند. به زبان ساده، جرم احتمالی که در نقطه $S$ متمرکز بود، ناپدید نمی‌شود، بلکه به بازه پایین‌دست منتقل می‌شود و یک تراکم موضعی جدید ایجاد می‌کند (در شکل ۱ نمرات معدل در بازه قبل از ۱۸ بیشترین تراکم را نشان می‌دهند که مؤید این نتیجه‌گیری است)؛ بنابراین، نتایج حاصل از معادله ۸ همچنان پابرجا و تنها بازه آن تغییر کرده است.

مسئله زمانی جدی‌تر می‌شود که به وزن‌دهی دو منبع ایجادکننده نمره کل در آزمون سراسری (امتحانات نهایی و کنکور) توجه کنیم. هرچه وزن امتحانات نهایی در نمره کل بیشتر باشد، اثر سقف و پیامدهای ناشی از آن (منظور، اثر خطای تصادفی در رتبه‌بندی است) برجسته‌تر خواهد شد؛ به بیان دیگر، وقتی سوابق تحصیلی سهم بالاتری در ترکیب دارند، محدودیت‌های ناشی از فشردگی نمرات در سقف بیش از پیش بر رتبه‌بندی نهایی تأثیر می‌گذارند. در چنین شرایطی، تفاوت‌های ناچیز در سوابق که ممکن است بازتابی از عوامل تصادفی مانند تفاوت در نمره‌گذاری معلمان بوده باشد، می‌تواند سرنوشت رتبه داوطلبان را به شکل غیرمنصفانه‌ای تغییر دهد. در نتیجه، هم جامعیت ادعا شده زیر سؤال می‌رود و هم بهینه بودن وزن‌ها از منظر روایی قابل دفاع نخواهد بود.

برای ملموس‌تر شدن این نتیجه، در ادامه با یک طرح شبیه‌سازی، تأثیر اثر سقف بر رتبه‌بندی نشان داده شده است.

**شبیه‌سازی اثر سقف در امتحانات نهایی**

در این شبیه‌سازی، جامعه‌ای بزرگ شامل ۱۰۰،۰۰۰ فرد در نظر گرفته شد تا دقت برآوردها افزایش یابد. نمرات واقعی هر فرد از توزیعی نرمال با میانگین‌های مختلف (۱۳، ۱۴ و ۱۵) و انحراف معیار ثابت برابر با ۴ تولید شد. سپس به هر نمره یک جزء خطای نرمال با میانگین صفر و انحراف معیار ۰.۵ افزوده شد تا فرایند اندازه‌گیری بازنمایی شود. از ترکیب این دو منبع، سه نوع نمره مشاهده‌شده ساخته شد: ۱) نمره بدون سقف که صرفاً حاصل جمع نمره واقعی و خطا است، ۲) مدل سقف اول که در آن ابتدا خطا به نمره افزوده و سپس حاصل در بازه صفر تا ۲۰ محدود می‌شود و ۳) مدل سقف دوم که در آن ابتدا نمره واقعی در محدوده صفر تا ۲۰ بریده می‌شود، سپس خطا افزوده و در نهایت مجدداً در همان بازه محدود می‌گردد. این دو مدل سقف بیانگر دو رویه متفاوت از نحوه تأثیر محدودیت سقف بر نمرات هستند.

برای تحلیل اثر سقف، رتبه‌های افراد در چهار حالت (نمره واقعی، بدون سقف، سقف ۱ و سقف ۲) محاسبه شدند؛ همچنین افراد به دو زیرگروه تقسیم شدند: ده درصد بالاترین نمرات که به‌عنوان گروه ممتاز (Top) در نظر گرفته شدند و بقیه افراد که به‌عنوان گروه سایر (Rest) تعریف شدند. مقایسه‌ها در دو سطح انجام شد: ۱) همبستگی رتبه‌ای (اسپیرمن) بین مدل‌های مختلف و ۲) میانگین اختلاف رتبه‌ها به‌عنوان شاخص تغییر مطلق در جایگاه افراد. این محاسبات به‌صورت جداگانه برای کل نمونه، گروه ممتاز و گروه سایر گزارش شد (جدول‌های ۲ و ۳). در ادامه، کدهای مربوط به شبیه‌سازی (در محیط R) گزارش شده است.

## Ceiling Effect Simulation

```r
set.seed(1377)

### --- Parameters ---
n <- 100000
top_prop <- 0.10
ceiling_min <- 0
ceiling_max <- 20

### --- Helper functions ---
compute_ranks <- function(scores) {
  rank(-scores, ties.method = "min")
}

avg_rank_diff <- function(r1, r2) {
  mean(abs(r1 - r2))
}
```

```
split_ranks <- function(r, top_idx, rest_idx) {
  list(top = r[top_idx], rest = r[rest_idx])
}

run_simulation <- function(mu) {
  ### --- Data generation ---
  true_final <- rnorm(n, mean = mu, sd = 4)
  error_final <- rnorm(n, mean = 0, sd = 0.5)

  # No ceiling
  observed_nc <- true_final + error_final

  # Ceiling Model 1: add error first, then ceiling
  observed_c1 <- pmin(ceiling_max, pmax(ceiling_min, observed_nc))

  # Ceiling Model 2: ceiling true first, add error, ceiling again
  true_ceiled <- pmin(ceiling_max, pmax(ceiling_min, true_final))
  observed_c2 <- true_ceiled + error_final
  observed_c2 <- pmin(ceiling_max, pmax(ceiling_min, observed_c2))

  ### --- Ranks ---
  r_true <- compute_ranks(true_final)
  r_nc  <- compute_ranks(observed_nc)
  r_c1  <- compute_ranks(observed_c1)
  r_c2  <- compute_ranks(observed_c2)

  ### --- Top vs. rest split ---
  top_n <- round(n * top_prop)
  top_idx <- order(true_final, decreasing = TRUE)[1:top_n]
  rest_idx <- setdiff(seq_len(n), top_idx)

  r_true_s <- split_ranks(r_true, top_idx, rest_idx)
  r_nc_s  <- split_ranks(r_nc,  top_idx, rest_idx)
  r_c1_s  <- split_ranks(r_c1,  top_idx, rest_idx)
  r_c2_s  <- split_ranks(r_c2,  top_idx, rest_idx)

  ### --- Correlations ---
  cor_results <- tibble(
    Comparison = c("nc_vs_c1","nc_vs_c2","c1_vs_c2","true_vs_nc","true_vs_c1","true_vs_c
2"),
    Overall = c(cor(r_nc, r_c1, method="spearman"),
            cor(r_nc, r_c2, method="spearman"),
            cor(r_c1, r_c2, method="spearman"),
            cor(r_true, r_nc, method="spearman"),
            cor(r_true, r_c1, method="spearman"),
            cor(r_true, r_c2, method="spearman")),
    Top = c(cor(r_nc_s$top, r_c1_s$top, method="spearman"),
```

```
        cor(r_nc_s$top, r_c2_s$top, method="spearman"),
        cor(r_c1_s$top, r_c2_s$top, method="spearman"),
        cor(r_true_s$top, r_nc_s$top, method="spearman"),
        cor(r_true_s$top, r_c1_s$top, method="spearman"),
        cor(r_true_s$top, r_c2_s$top, method="spearman")),
  Rest = c(cor(r_nc_s$rest, r_c1_s$rest, method="spearman"),
        cor(r_nc_s$rest, r_c2_s$rest, method="spearman"),
        cor(r_c1_s$rest, r_c2_s$rest, method="spearman"),
        cor(r_true_s$rest, r_nc_s$rest, method="spearman"),
        cor(r_true_s$rest, r_c1_s$rest, method="spearman"),
        cor(r_true_s$rest, r_c2_s$rest, method="spearman"))
 )

 ### --- Average rank differences ---
 avg_results <- tibble(
  Comparison = c("nc_vs_c1","nc_vs_c2","c1_vs_c2","true_vs_nc","true_vs_c1","true_vs_c
2"),
  Overall = c(avg_rank_diff(r_nc, r_c1),
           avg_rank_diff(r_nc, r_c2),
           avg_rank_diff(r_c1, r_c2),
           avg_rank_diff(r_true, r_nc),
           avg_rank_diff(r_true, r_c1),
           avg_rank_diff(r_true, r_c2)),
  Top = c(avg_rank_diff(r_nc_s$top, r_c1_s$top),
        avg_rank_diff(r_nc_s$top, r_c2_s$top),
        avg_rank_diff(r_c1_s$top, r_c2_s$top),
        avg_rank_diff(r_true_s$top, r_nc_s$top),
        avg_rank_diff(r_true_s$top, r_c1_s$top),
        avg_rank_diff(r_true_s$top, r_c2_s$top)),
  Rest = c(avg_rank_diff(r_nc_s$rest, r_c1_s$rest),
        avg_rank_diff(r_nc_s$rest, r_c2_s$rest),
        avg_rank_diff(r_c1_s$rest, r_c2_s$rest),
        avg_rank_diff(r_true_s$rest, r_nc_s$rest),
        avg_rank_diff(r_true_s$rest, r_c1_s$rest),
        avg_rank_diff(r_true_s$rest, r_c2_s$rest))
 )

 list(cor=cor_results, avg=avg_results)
}
```

**results**

```
### --- Run for mean = 13, 14, 15 ---
results <- lapply(c(13,14,15), run_simulation)
names(results) <- paste0("Mean_", c(13,14,15))

# Combine into long tables
cor_table <- bind_rows(lapply(names(results), function(nm) {
```

```
 results[[nm]]$cor %>% mutate(Mean = nm)
}))

avg_table <- bind_rows(lapply(names(results), function(nm) {
 results[[nm]]$avg %>% mutate(Mean = nm)
}))
```

جدول ۲

همبستگی اسپیرمن برای میانگین‌های مختلف

| مقایسه‌ها | کلی | ۱۰% بالا | سایر | میانگین نمره حقیقی |
|---|---|---|---|---|
| nc_vs_c۱ | ۱/۰۰۰ | ۰/۹۶۴ | ۱/۰۰۰ | میانگین ۱۳ |
| nc_vs_c۲ | ۱/۰۰۰ | ۰/۸۹۳ | ۱/۰۰۰ | |
| c۱_vs_c۲ | ۱/۰۰۰ | ۰/۹۲۴ | ۱/۰۰۰ | |
| true_vs_nc | ۰/۹۹۲ | ۰/۹۲۵ | ۰/۹۸۹ | |
| true_vs_c۱ | ۰/۹۹۲ | ۰/۸۸۳ | ۰/۹۸۹ | |
| true_vs_c۲ | ۰/۹۹۱ | ۰/۷۶۴ | ۰/۹۸۹ | |
| nc_vs_c۱ | ۱/۰۰۰ | ۰/۸۲۴ | ۱/۰۰۰ | میانگین ۱۴ |
| nc_vs_c۲ | ۱/۰۰۰ | ۰/۵۷۸ | ۱/۰۰۰ | |
| c۱_vs_c۲ | ۱/۰۰۰ | ۰/۶۸۶ | ۱/۰۰۰ | |
| true_vs_nc | ۰/۹۹۲ | ۰/۹۲۵ | ۰/۹۸۹ | |
| true_vs_c۱ | ۰/۹۹۱ | ۰/۷۲۴ | ۰/۹۸۹ | |
| true_vs_c۲ | ۰/۹۹۱ | ۰/۳۳۱ | ۰/۹۸۹ | |
| nc_vs_c۱ | ۰/۹۹۹ | ۰/۳۹۱ | ۱/۰۰۰ | میانگین ۱۵ |
| nc_vs_c۲ | ۰/۹۹۸ | ۰/۳۰۴ | ۱/۰۰۰ | |
| c۱_vs_c۲ | ۰/۹۹۹ | ۰/۳۴۱ | ۱/۰۰۰ | |
| true_vs_nc | ۰/۹۹۲ | ۰/۹۲۶ | ۰/۹۸۹ | |
| true_vs_c۱ | ۰/۹۹۱ | ۰/۳۲۵ | ۰/۹۸۹ | |
| true_vs_c۲ | ۰/۹۸۹ | -۰/۰۰۵ | ۰/۹۸۹ | |

نکته: nc= نمره بدون سقف، ۱c= نمره با اثر سقف (اعمال خطا پیش از اثر سقف)، ۲c= نمره با اثر سقف (اعمال خطا پس از اثر سقف)، true= نمره حقیقی.

جدول ۳

میانگین تفاوت‌های رتبه‌بندی (رند شده)

| مقایسه‌ها | کلی | ۱۰% بالا | سایر | میانگین نمره حقیقی |
|---|---|---|---|---|
| nc_vs_c۱ | ۸۵ | ۸۵۳ | ۰ | میانگین ۱۳ |
| nc_vs_c۲ | ۱۰۲ | ۱۰۱۸ | ۰ | |
| c۱_vs_c۲ | ۸۳ | ۸۲۹ | ۰ | |
| true_vs_nc | ۲۷۸۴ | ۹۹۹ | ۲۹۸۳ | |
| true_vs_c۱ | ۲۸۵۵ | ۱۷۰۳ | ۲۹۸۳ | |
| true_vs_c۲ | ۲۸۷۸ | ۱۹۳۹ | ۲۹۸۳ | |
| nc_vs_c۱ | ۲۳۸ | ۲۳۵۱ | ۳ | میانگین ۱۴ |
| nc_vs_c۲ | ۲۷۷ | ۲۶۴۲ | ۱۴ | |
| c۱_vs_c۲ | ۲۲۷ | ۲۱۶۹ | ۱۱ | |

| مقایسه‌ها | کلی | ۱۰٪ بالا | سایر | میانگین نمره حقیقی |
|---|---|---|---|---|
| true_vs_nc | ۲۷۹۷ | ۱۰۰۰ | ۲۹۹۷ | |
| true_vs_c۱ | ۲۹۹۶ | ۲۹۶۸ | ۳۰۰۰ | |
| true_vs_c۲ | ۳۰۵۵ | ۳۴۶۱ | ۳۰۱۰ | |
| nc_vs_c۱ | ۵۸۶ | ۴۵۷۷ | ۱۴۲ | |
| nc_vs_c۲ | ۶۴۹ | ۴۴۷۱ | ۲۲۵ | |
| c۱_vs_c۲ | ۵۴۴ | ۴۶۴۸ | ۸۸ | میانگین ۱۵ |
| true_vs_nc | ۲۷۸۶ | ۹۹۹ | ۲۹۸۵ | |
| true_vs_c۱ | ۳۲۸۴ | ۴۶۹۵ | ۳۱۲۷ | |
| true_vs_c۲ | ۳۳۹۷ | ۵۳۷۶ | ۳۱۷۸ | |

نکته: nc= نمره بدون سقف، ۱c= نمره با اثر سقف (اعمال خطا پیش از اثر سقف)، ۲c= نمره با اثر سقف (اعمال خطا پس از اثر سقف)، true= نمره حقیقی.

نتایج جدول ۲ نشان می‌دهد که در سطح کل نمونه، همبستگی بین انواع نمرات بسیار بالا و نزدیک به یک باقی می‌ماند. این امر بدین معنا است که اثر سقف در جمعیت کلی چندان آشکار نمی‌شود؛ زیرا بخش عمده‌ای از افراد در دامنه میانی توزیع نمره قرار دارند.

بررسی گروه ممتاز (۱۰٪ بالا) الگوی کاملاً متفاوتی را نشان می‌دهد. در میانگین ۱۳، همبستگی‌ها هنوز بالا هستند (برای مثال، همبستگی بین نمره واقعی و مدل سقف ۲، برابر با ۰.۷۶۴ است). با این حال، با افزایش میانگین به ۱۴، همبستگی‌ها به‌طور چشمگیری کاهش می‌یابند (برای مثال، میزان همبستگی رتبه‌ای اسپیرمن میان نمره حقیقی و نمره اعمال‌شده با مدل سقف دوم برای گروه ممتاز برابر با ۰٫۳۳۱ است). در میانگین ۱۵، همبستگی بین نمره واقعی و مدل سقف ۲ عملاً از بین می‌رود (میزان همبستگی رتبه‌ای اسپیرمن میان نمره حقیقی و نمره اعمال‌شده با مدل سقف دوم برای گروه ممتاز برابر با ۰-٫۰۰۵ است). این کاهش شدید نشان می‌دهد که هرچه میانگین نمره به سقف نزدیک‌تر شود، رتبه‌بندی افراد ممتاز به‌شدت تحریف می‌شود و اطلاعات تمایزبخش آن‌ها تقریباً به‌طور کامل از بین می‌رود.

در گروه سایر، همبستگی‌ها تقریباً کامل باقی می‌مانند (حدود ۰٫۹۸۹ یا بالاتر). این الگو بیانگر آن است که اثر سقف اساساً پدیده‌ای است که بیش از همه بر گروه‌های نزدیک به سقف اثر می‌گذارد، در حالی که رتبه‌بندی بقیه جمعیت کم و بیش دست‌نخورده باقی می‌ماند. یافته‌های جدول ۳ تصویری مکمل از این الگو ارائه داده‌اند. در سطح کل نمونه، میانگین اختلاف رتبه‌ها با افزایش میانگین نمره بیشتر می‌شود (برای مثال، true_vs_c2 از ۲۸۷۸ در میانگین ۱۳ به ۳۳۹۷ در میانگین ۱۵ می‌رسد). این افزایش نشان‌دهنده اثر تجمعی سقف بر ساختار کلی رتبه‌بندی است. در گروه ممتاز، اختلاف رتبه‌ها به‌مراتب شدیدتر بود؛ برای مثال، اختلاف میانگین رتبه‌ها بین نمره واقعی و مدل سقف ۲ در میانگین ۱۳ برابر با ۱۹۳۹ است؛ اما این مقدار در میانگین ۱۴ به ۳۴۶۱ و در میانگین ۱۵ به ۵۳۷۶ می‌رسد. چنین مقادیری به‌وضوح بیانگر این نکته هستند که افراد برتر به‌طور جدی جایگاه خود را در رتبه‌بندی از دست می‌دهند و امکان تفکیک بین آنان تقریباً محو می‌شود.

## بحث و نتیجه‌گیری

این پژوهش با هدف ارزیابی اعتبار یکی از کلیدی‌ترین ادعاهای زیربنایی آزمون سراسری ورود به دانشگاه در ایران، یعنی «ارائه ارزیابی جامع‌تر و معتبرتر از طریق ترکیب نمرات کنکور و سوابق تحصیلی»، انجام شد. با به‌کارگیری چارچوب روایی استدلال‌محور کین (۲۰۰۶ و ۲۰۱۳)، نشان داده شد که این ادعا با وجود ظاهر منطقی، بر پیش‌فرض‌های متزلزلی استوار است که روایی کل سیستم را به‌ویژه برای داوطلبان ممتاز، به‌طور جدی به چالش می‌کشد.

تحلیل نظری و نتایج شبیه‌سازی به‌روشنی نشان داد که مشکل اصلی نه در نفس ترکیب دو منبع اطلاعاتی، بلکه در ماهیت متفاوت و اهداف متعارض این دو آزمون نهفته است. کنکور سراسری به‌عنوان آزمونی هنجار – مرجع با هدف اصلی بیشینه‌سازی واریانس و تمایزگذاری دقیق میان داوطلبان طراحی می‌شود. در مقابل، امتحانات نهایی به‌عنوان آزمونی ملاک – مرجع صرفاً به دنبال سنجش میزان دستیابی دانش‌آموزان به اهداف آموزشی از پیش تعیین‌شده است و دغدغه‌ای برای تمایزگذاری میان آن‌ها ندارد.

این تفاوت بنیادین منجر به بروز پدیده‌ای به نام اثر سقف در نمرات امتحانات نهایی می‌شود؛ به این معنا که تعداد زیادی از داوطلبان، به‌ویژه آن‌هایی که توانایی بالاتری دارند، به نمره کامل یا نزدیک به آن دست می‌یابند. نتایج شبیه‌سازی نشان داد که هرچه میانگین توانایی جامعه به سقف نمرات (نمره ۲۰) نزدیک‌تر شود، قدرت تمایزگذاری سوابق تحصیلی در گروه ممتاز (۱۰ درصد برتر) به‌شدت افت می‌کند و همبستگی رتبه‌ها با توانایی حقیقی آن‌ها عملاً از بین می‌رود. این یافته مؤید آن است که در بازه رقابتی برای کسب رتبه‌های برتر، سوابق تحصیلی به شاخصی با خطای بالا و اتکاناپذیر تبدیل می‌شود که در رتبه‌بندی به‌جای توانایی حقیقی، به خطاهای تصادفی حساس است.

بر اساس چارچوب کین (۲۰۱۳)، روایی آزمون به قوت و انسجام زنجیره‌ای از استدلال‌ها بستگی دارد که از نمره خام به تصمیم نهایی منجر می‌شود. اگر حلقه‌ای از این زنجیره سست باشد، کل استدلال فرومی‌ریزد. در مدل فعلی آزمون سراسری، این حلقه سست همان فرضِ معتبر بودن سوابق تحصیلی برای تمایزگذاری میان داوطلبان برتر است. وقتی جزئی از نمره نهایی (در اینجا سوابق تحصیلی) در حساس‌ترین بخش رقابت، کارایی خود را از دست می‌دهد، ادعای جامعیت و روایی کل سیستم زیر سؤال می‌رود؛ علاوه بر این، ادعای «بهینه بودن وزن‌ها» نیز بی‌معنا می‌شود؛ زیرا تخصیص هرگونه وزنی به متغیری که در بخش مهمی از جامعه هدف فاقد واریانس معنادار است نه‌تنها به بهبود دقت سنجش کمکی نمی‌کند، بلکه به دلیل وارد کردن خطا، به‌طور بالقوه باعث کاهش عدالت و روایی می‌شود.

در نتیجه می‌توان گفت که سیاست ترکیب کنکور سراسری و سوابق تحصیلی، حداقل در شکل کنونی آن، از منظر روایی قابل دفاع نیست. این ترکیب به‌جای آنکه تصویری کامل‌تر ارائه دهد، با تکیه بر یک منبع اطلاعاتی ناکارآمد (در بخش بالایی توزیع)، موجب تحریف رتبه‌بندی و تضعیف هدف اصلی آزمونی سرنوشت‌ساز، یعنی انتخاب عادلانه و دقیق شایسته‌ترین افراد، می‌شود. برای آینده، سیاست‌گذاران باید به طور جدی درمورد هدف از لحاظ کردن سوابق تحصیلی بازنگری کنند. اگر هدف، تمایزگذاری است، ساختار امتحانات نهایی باید به‌گونه‌ای تغییر کند که قادر به ایجاد واریانس در تمام سطوح توانایی و به‌ویژه سطوح بالای آن (به شکلی سازنده) باشد. در غیر این صورت، تکیه بر آن برای رتبه‌بندی داوطلبان، به‌ویژه با وزن‌های بالا، اشتباهی روش‌شناختی است که پیامدهای فردی و اجتماعی گسترده‌ای به همراه دارد (برای نگرشی جایگزین در خصوص آزمون‌های ورودی دانشگاه، به کروندی رنانی و همکاران (۲۰۲۵) نگاه کنید).

البته در تفسیر نتایج این پژوهش، ضروری است که شواهد دال بر وجود اثر سقف (به‌ویژه شکل ۱) به‌طور مجزا و منفرد ارزیابی نشوند. در مقابل، اتخاذ رویکردی استنتاجی مبتنی بر منطق بیزی[1] می‌تواند راهگشا باشد؛ رویکردی که در آن احتمال صحت یک فرضیه با تکیه بر شواهد موجود و در پرتو احتمالات و انتظارات پیشین بررسی می‌گردد. به بیان دقیق‌تر، شواهد ارائه‌شده پیرامون اثر سقف را باید در چارچوب شناخت علمی ما از مکانیزم عمل آزمون‌های ملاک – مرجع و انتظار پیشین وقوع این اثر، تحلیل و تفسیر کرد.

علاوه بر این، در تفسیر نهایی یافته‌های این پژوهش، ضروری است رویکرد و نتایج کار را در چارچوب ارزیابی و تحلیل سیاست‌گذاری‌ها درک کنیم. تحلیل سیاست‌گذاری‌های آموزشی لزوماً اثباتی قطعی یا حل مسئله با پاسخ‌های مطلق نیست، بلکه ماهیتی استدلالی[2] دارد (مانند مباحثه، در آن گروه‌های متفاوت شواهد خود را در تأیید یا رد ادعاها ارائه می‌کنند). بر همین اساس، خروجی مدل‌ها و شبیه‌سازی‌های به کار رفته در این مطالعه نباید به‌عنوان حقایق عینی[3] و خدشه‌ناپذیر پنداشته شوند. در عوض، نتایج این شبیه‌سازی‌ها صرفاً شواهدی[4] هستند که در کنار سایر مبانی نظری قرار می‌گیرند تا ادعایی مشخص را در فرایند بحث تقویت کنند (به ماجون[5] (۱۹۸۹) نگاه کنید).

در نهایت، مؤثرترین پیشنهاد کاربردی، تغییر کارکرد سوابق تحصیلی و البته آزمون سراسری، از متغیری رتبه‌ساز به یک حد نصاب کیفی است. در این مدل، ورود به رقابت برای صندلی‌های دانشگاه‌های برتر منوط به کسب حداقل نمره‌ای مشخص در آزمون‌ها خواهد بود و پس از این فیلتر، از شاخص‌های شخصی‌سازی‌شده به‌عنوان معیار ورود به آموزش عالی استفاده می‌شود (کروندی رنانی و همکاران، ۲۰۲۵)؛ اما فارغ از تغییر یا عدم تغییر کارکرد آزمون‌ها، اهمیت روایی و رواسازی در فرایند آزمون‌سازی همچنان به قوت خود باقی است و باید دغدغه اصلی نهادهای توسعه‌دهنده آزمون‌ها باشد (امری که در کشور ما مغفول مانده است).

## سپاس‌گزاری

این مقاله برگرفته از طرح‌های پژوهشی مصوب سازمان سنجش آموزش کشور است و نویسندگان از همکاری و حمایت آن سازمان، به‌ویژه در فراهم‌سازی داده‌های آزمون سراسری و سایر مساعدت‌های انجام‌شده، صمیمانه قدردانی می‌کنند.

## ملاحظات اخلاقی

این مقاله بر پایه تحلیل داده‌های مربوط به آزمون سراسری انجام شده است و در فرایند اجرای آن، مراجعه مستقلی به شرکت‌کنندگان با هدف پژوهش صورت نگرفته است

## حامی مالی

این مطالعه با حمایت مالی سازمان سنجش آموزش کشور انجام شده است.

[1] Bayesian
[2] argumentative
[3] fact
[4] evidence
[5] Majone

## مشارکت نویسندگان

ایده‌پردازی اصلی پژوهش و نظارت بر حسن اجرای تمامی مراحل آن بر عهده نویسنده اول بوده است. تهیه پیش‌نویس مقاله، انجام تحلیل‌های آماری و اعمال کلیه ویرایش‌های متنی بر عهده نویسنده دوم بوده است. همچنین، انجام شبیه‌سازی‌ها و پیگیری اصلاحات پیشنهادی بر عهده نویسنده سوم بوده است. در نهایت، هر سه نویسنده در تهیه و تأیید نسخه نهایی مقاله مشارکت داشته‌اند.

## تعارض منافع

نویسندگان اعلام می‌دارند که در هیچ‌یک از مراحل پژوهش حاضر هیچ‌گونه تضاد منافعی وجود ندارد.

## منابع


جهانی‌فر، م؛ خدایی، ا؛ یونسی، ج و موسوی، س. (۱۳۹۶). نقش هموارسازی در تبدیل غیرخطی نمره‌های خام به نمره‌های مقیاس نرمال. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*. ۷(۲۰). ۱۰۳-۱۳۳.

کروندی رنانی، م؛ صالحی، ک و خدایی، ا. (۱۴۰۵). فراسوی عینیت مکانیکی: چالش‌های عینیت و راهکارهای نوین در پذیرش دانشجو. *فصلنامه پژوهش و برنامه‌ریزی در آموزش عالی*. e729749.


## References


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Borsboom, D., & Wijsen, L. D. (2016). Frankenstein's validity monster: The value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 281-283.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061-1071.

Crocker, L. & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. Harcourt, New York, 527.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In *Test validity.* (pp. 3-17). Lawrence Erlbaum Associates, Inc.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Duhem, P. (2021). *The Aim and Structure of Physical Theory*. Princeton: Princeton University Press.

Jahanifar,M, Khodaie,E, Yunesi,J and Musavi,S A . (2018). Smoothing methods role in raw scores Non-linear transformation to Normalized scale scores. *Educational Measurement and Evaluation Studies*, *7*(20), 103-133. (Persian)

Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), 66-70.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (Fourth Edition ed., pp. 17-64).

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kane, M. T. (2016). Validity as the evaluation of the claims based on test scores. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 309-311.

Karvandi Renani,M , Salehi,K and Khodaie,E . (2026). Moving Beyond Mechanical Objectivity: Addressing the Challenges and Exploring Novel Solutions in Iranian University Admissions. *Quarterly Journal of Research and Planning in Higher Education*, *32*(2), 165-187.

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Co.

Kim, J.-E. (2025). What makes a killer question killer? A text mining analysis of high-difficulty questions in the Korean CSAT English section. *English Teaching, 80*(3), 3–22.

Kim, Y. (2023, December 17). South Korea to reform education system by removing “killer questions” from Suneung. The Herald Insight.

Lee, H.-J. (2023, June 26). 'Killer' CSAT questions require obscure knowledge. Korea JoongAng Daily. https://koreajoongangdaily.joins.com/2023/06/26/national/socialAffairs/CSAT-college-exam-killer-questions/20230626191501015.html

Majone, G. (1989). *Evidence, Argument, and Persuasion in the Policy Process*. New Haven, CT: Yale University Press.

Markus, K. A., Borsboom, D. (2013). *Frontiers in test validity theory : measurement, causation, and meaning* (1st edition ed.). Routledge.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement (3rd ed.)* (pp. pp. 13–100). American Council on Education.

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational & Psychological Assessment*.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory (3rd ed.)*. New York: McGraw-Hill.

Yeung, J., & Seo, Y. (2023, July 1). South Korea is cutting ‘killer questions’ from an 8-hour exam some blame for a fertility rate crisis. CNN. https://edition.cnn.com/2023/07/01/asia/south-korea-college-exam-fertility-pressure-intl-hnk-dst