Journal of Mathematics Instruction, Social Research and Opinion
Vol. 4, No. 2, June 2025, pp. 529 – 542, https://doi.org/10.58421/misro.v4i2.524
ISSN 2962-7842

Who Reaches the Top? Investigating Gender-Based Performance and Maximum Score Gaps in University Entrance Exams

Hermina Disnawati[1], Siti Irene Astuti Dwiningrum[2], Yohanis Ndapa Deda[3], Salman Rashi[4], Mariano Dos Santos[5]
[1,2]Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[1,3] Universitas Timor, Kefamenanu, Indonesia
[4]University of Turbat, Turbat, Pakistan
[5]Instituto Católico para a Formação de Professores (ICFP) Baucau, Baucau, Timor - Leste

Article Info

Article history:

Received 2025-05-19
Revised 2025-06-25
Accepted 2025-06-26

Keywords:

Academic readiness
English proficiency
Gender differences
Quantitative Knowledge
University Entrance Exam

ABSTRACT

Disparities in the academic readiness of first-year university students remain evident, continuing to challenge higher education systems, particularly when entrance assessments fail to reflect foundational competencies accurately. This study investigates whether gender-based differences exist in entrance exam performance and analyzes students' maximum scores to assess peak academic achievement in two core domains: English proficiency and Quantitative Knowledge. The research was conducted at a public university in East Nusa Tenggara, Indonesia, involving 1,186 students admitted through the independent selection pathway. A comparative quantitative approach was applied using descriptive statistics, Welch's t-test, and boxplot visualizations in JASP 0.18.3 software. The results showed no statistically significant gender differences in either domain (English: $p = 0.504$, $d = 0.040$; Quantitative: $p = 0.543$, $d = –0.036$), with nearly identical mean scores and score variability. However, the maximum scores of 17/20 in English and 13/20 in Quantitative Knowledge were far below the full mark, indicating systemic gaps in academic preparedness and possible ceiling effects in test design. These findings support the need to redesign entrance assessments using Item Response Theory (IRT) and to implement bridging or matriculation programs. These interventions can help universities and policymakers ensure fairer and inclusive pathways to promote equitable access and success in higher education.

Corresponding Author:

Hermina Disnawati
Program Studi Pendidikan Matematika, Universitas Timor, Kefamenanu, Indonesia
Email: disnawati@unimor.ac.id

1. INTRODUCTION

Higher education in Indonesia plays a pivotal role in preparing a competent workforce to face global challenges [1], [2]. One critical component of this preparation is the university entrance examination, which is designed to assess the basic academic abilities

of prospective students to ensure their readiness for higher education [3]. However, analysis of entrance exam results often reveals an imbalance in the initial abilities of new students, which can be caused by differences in access or the quality of their previous education. This inequality suggests that the quality of primary and secondary education, including access to learning resources [4], significantly contributes to the academic readiness of prospective students [5].

Gender is one of the important factors that contribute to the inequality of students' early academic abilities [6], [7]. Cross-national studies suggest that this disparity is not merely cognitive but also shaped by psychological and sociocultural dynamics. For instance, research in Finland shows that female students are more likely to avoid risks when answering entrance exam questions, leaving more items blank than their male counterparts, a pattern that reduces their chances of admission despite comparable academic abilities [8].

A similar pattern was observed in Turkey, where entrance exams that emphasize linguistic and mathematical logic tend to favor male students, although female students exhibit more diverse multiple intelligence profiles [9]. However, the extent to which risk aversion affects total performance is considered modest [10]. Complementing these findings, a study in Spain revealed that male students reported higher levels of self-confidence when tackling math-based questions, potentially giving them an advantage in tasks requiring logical reasoning and quick decision-making [11].

Gender is one of the important factors that contribute to the inequality of students' early abilities [6], [7]. Cross-national studies suggest that this disparity is not merely cognitive but also shaped by psychological and sociocultural dynamics. For instance, research in Finland shows that female students are more likely to avoid risks when answering entrance exam questions, leaving more items blank than their male counterparts, a pattern that reduces their chances of admission despite comparable academic abilities [8]. A similar pattern was observed in Turkey, where entrance exams that emphasize linguistic and mathematical logic tend to favor male students, although female students exhibit more diverse multiple intelligence profiles [9]. However, the impact of this risk aversion on overall scores is relatively small [10]. Complementing these findings, a study in Spain revealed that male students reported higher levels of self-confidence when tackling math-based questions, potentially giving them an advantage in tasks requiring logical reasoning and quick decision-making [11].

Previous research has shown that academic ability inequality can be influenced by various factors, including gender, socioeconomic background, and the distribution of educational resources [8]. Gender often affects academic performance in complex ways; for example, men are more often advantaged in math and science exams in Western countries, while women tend to excel in reading skills in different countries [12]. Studies also indicate that women perform better in low-stress situations, but underperform in high-stakes exams, reflecting distinct gendered responses to test anxiety [13]. A Nigerian study further emphasized that sociocultural expectations encourage women to pursue non-technical disciplines, which in turn influences their academic outcomes in STEM-related exams [7].

While international literature has broadly explored gender disparities in STEM education and university performance globally [14], [15], most of these studies focus on

long-term academic trajectories or macro-structural determinants. In contrast, Indonesian research has predominantly examined mean score differences, general academic readiness, or sociocultural barriers to higher education access [14], [16]. For instance, Sriyati [17] conducted a descriptive analysis of the performance of grade 12 students on the Scholastic Potential Test, finding that the average scores across five components remained below 50%, indicating low readiness for national entrance exams, such as UTBK. Similarly, Anjar [18] reported that high school students in Metro, Lampung, achieved only 26.84% on SNMPTN-related readiness indicators, emphasizing the limited academic preparedness of university applicants. Their study also highlighted the critical role of school counselors in providing psychological and academic support to students facing entrance examinations.

In addition to performance-based studies, Nainggolan et al. [19] analyzed the cognitive demands of SBMPTN and SIMAK UI exam questions, revealing a strong emphasis on higher-order thinking skills such as logical reasoning and complex comprehension. While these studies offer valuable insights into the challenges faced by prospective students, they often focus on descriptive patterns or test content, rather than addressing the underlying issues. However, limited attention has been given to gender-based performance patterns in standardized university entrance exams, particularly within the independent admission pathway, where each institution determines selection criteria. Existing studies rarely conduct diagnostic evaluations that analyze maximum score distributions or subject-specific gender parity, particularly in core areas such as mathematics and English, at the point of entry, especially under the independent admission track.

This study addresses that gap by conducting a diagnostic analysis of gender-based performance disparities and score extremes in a public university's independent admission exam, with a specific focus on mathematics and English. Unlike previous research that primarily examines general readiness or post-admission academic outcomes, this study provides fine-grained, entry-level insights into performance equity. By focusing on maximum scores and gap severity, the study offers a sharper lens for understanding gender fairness in high-stakes admission contexts.

The analysis is framed through the lens of structural inequality in higher education [20], which posits that academic disparities are often rooted in systemic educational barriers rather than individual deficits. Additionally, the study applies fairness and diagnostic validity frameworks [21] to critically assess whether entrance exams equitably measure academic readiness across gender groups. These conceptual foundations support the study's dual objectives: (1) to uncover gender-based performance patterns and (2) to evaluate maximum scores as diagnostic indicators of test alignment and readiness gaps.

The analysis in this study is expected to provide deeper insights into the pattern of inequality in students' early abilities and the factors that influence it, as well as serve as the basis for a more inclusive education policy. The resulting recommendations can also be used to evaluate the effectiveness of the self-paced test design and ensure that the selection process accommodates the diverse backgrounds of participants [13].

In this context, the analysis of entrance exam scores serves not only as a selection tool but also as a reflection of the broader education system. One of the key findings in the study was that the maximum score was well below the full score, suggesting challenges in

the design of the exam questions. This phenomenon sparks discussions about how entrance exams are designed and whether they accurately reflect the fundamental abilities of new students, or instead reflect limitations in the previous education system [22].

2. METHOD

2.1 Research Design

This study employed a comparative quantitative design to examine gender-based differences in average scores for two core skill domains, English and Quantitative Knowledge (Mathematics), among new students admitted through the independent selection track at a public university in Indonesia. In addition, the analysis emphasized the maximum scores in both domains as indicators of the distribution and ceiling of students' initial academic abilities.

The research population comprises new students admitted during the 2024/2025 academic year through the *jalur mandiri* (independent admission track), with a final sample of 1,186 students consisting of 481 males and 705 females. Initially, 1,189 student records were collected; however, three records were excluded due to missing gender information.

Although the entrance examination assessed five subject areas, this study focused specifically on English and Quantitative Knowledge, as they are strongly associated with core competencies required for academic success in both STEM and non-STEM fields. Each subtest had a maximum score of 20, contributing to a total test score of 100. English proficiency was operationalized as performance on a 20-item sub-test measuring vocabulary and reading comprehension. Quantitative Knowledge was measured using a 20-item sub-test covering basic arithmetic, numerical series, and logical reasoning problems. The independent variable was gender (male or female). The dependent variables included English score, Quantitative Knowledge score, and the respective maximum scores as indicators of peak performance.

Since the researchers did not participate in the test design or validation process, it is assumed that the test items had undergone internal quality assurance by the university’s selection committee. Due to the nature of the institutional data, internal reliability coefficients such as KR-20 or Cronbach’s alpha were not available for analysis.

All data used in this study were obtained in an anonymized form. The researchers did not collect or access personally identifiable information, and the study involved no direct interaction with human participants. Based on institutional guidelines, the study was exempt from formal ethical review. Nevertheless, all ethical principles related to confidentiality and responsible data use were strictly upheld.

2.2 Data Analysis Procedure

The dataset was organized in Excel with three key columns: Gender, English score, and Quantitative Knowledge score. The data were then analyzed in JASP version 0.18.3, which supports both descriptive and inferential statistical procedures, as well as data visualization.

The analysis consisted of four stages:

a. Descriptive Statistics

Descriptive statistics were calculated to provide an overview of score distributions by gender. These included mean, standard deviation, minimum, maximum, and maximum score. Maximum score analysis was conducted to understand achievement ceilings and the extent to which students approached the full marks.

b. Assumption Testing and Hypothesis Testing

Prior to conducting inferential analysis, the assumptions of normality and homogeneity of variance were tested using the Shapiro-Wilk and Levene's tests, respectively. As assumptions were not fully met, Welch's t-test was used to compare group means in English and Quantitative Knowledge domains. This test is robust under conditions of unequal variances and non-normal distributions in large samples.

i. Null hypothesis (H_0): There is no significant difference in mean scores between male and female students.
ii. Alternative hypothesis (H_1): There is a significant difference in mean scores between male and female students.

In addition to the t-value, degrees of freedom (df), and p-value, Cohen's d was calculated to determine the magnitude of the difference. Effect sizes were interpreted according to Cohen's guidelines: small (d = 0.2), medium (d = 0.5), and large (d = 0.8) [23], [24].

c. Maximum Score Analysis

Maximum score data were analyzed by gender to identify potential patterns of achievement disparity or ceiling effects. This helped determine whether certain groups approached or failed to reach full performance levels.

d. Data Visualization

Boxplots were generated for each domain to visually depict the distributions of scores, medians, interquartile ranges (IQRs), outliers, and score extremes by gender.

This method is designed to provide comprehensive insight into score distribution patterns, the achievement of maximum scores, and potential significant differences based on gender, thereby supporting the research objectives holistically.

3. RESULTS AND DISCUSSION

3.1 Results

The findings are organized into subsections that reflect key research focuses, namely score distribution, gender comparison, and analysis of maximum achievement.

a. Descriptive Statistics on English and Quantitative Knowledge

Descriptive analysis was conducted on the entrance exam results of 1,186 students (481 males and 705 females) admitted through the independent admission track. The study focused on two of the five exam domains: English proficiency and Quantitative Knowledge, each with a maximum score of 20.

Table 1. Descriptive Statistics by Domain and Gender

	English Proficiency		Quantitative Knowledge (Mathematics)	
	Male	Female	Male	Female
Valid (N)	481	705	481	705

Missing	0	0	0	0
Mean	5.310	5.216	5.210	5.284
Std. Deviation	2.389	2.374	2.025	2.087
Minimum	0.000	0.000	1.000	1.000
Maximum	17.000	16.000	12.000	13.000

The result presented in Table 1 reveals that male students achieved an average English proficiency score of 5.310 (SD = 2.389), while their female counterparts scored slightly lower, with a mean of 5.216 (SD = 2.374). The minimum score for both genders was 0, indicating the presence of students with extremely limited proficiency. In contrast, the maximum score in English was marginally higher for males (17) than for females (16), though neither group approached the full mark of 20.

In the domain of Quantitative Knowledge, the pattern was reversed. Female students attained a slightly higher average score of 5.284 (SD = 2.087), compared to 5.210 (SD = 2.025) among males. As with English, the score distribution reflects a considerable variation in individual performance, suggesting heterogeneous levels of preparedness within each gender group.

b. Gender-Based Score Comparison

This section aims to examine whether there are statistically significant differences in English proficiency and Quantitative Knowledge scores between male and female students. The hypothesis testing was conducted using Welch's t-test, a robust method suitable for samples with unequal variances and non-normal distributions. Figure 1 presents the summary of Welch's t-test results for both domains.

Independent Samples T-Test

	Test	Statistic	df	p	Cohen's d	SE Cohen's d
English	Student	0.669	1184.000	0.504	0.040	0.059
	Welch	0.668	1027.078	0.504	0.040	0.059
Math (Quantitative)	Student	-0.605	1184.000	0.546	-0.036	0.059
	Welch	-0.608	1051.302	0.543	-0.036	0.059

Figure 1. The results of the Independent Samples T-Test of two domains and gender

The results indicate that for English proficiency, the p-value was 0.504 (> 0.05), and for Quantitative Knowledge, the p-value was 0.543 (> 0.05). As both p-values exceed the standard significance threshold of 0.05, H_0 is retained in both cases. This means that there is no statistically significant difference in mean scores between male and female students in either domain. Additionally, the effect sizes (Cohen's *d*) were very small, 0.040 for English and 0.036 for Quantitative, suggesting negligible practical differences between the two groups. According to Cohen's benchmarks, effect sizes can be interpreted as small ($d \approx 0.2$), medium ($\approx d = 0.5$), and large (≥ 0.8) [23].

Further descriptive statistics, as shown in Figure 2, reinforce this pattern of similarity in Table 1. The mean English score for males was 5.310 (SD = 2.389), while females scored 5.216 (SD = 2.374). In Quantitative Knowledge, females slightly outperformed males (5.284

vs. 5.210), although this difference was not statistically significant. The standard errors indicate that the group means are estimated with high precision, and the similar CV values suggest homogeneous score distributions across genders. Therefore, even though there are slight numerical differences in scores, they do not reflect meaningful or statistically supported gender-based disparities. These results contribute to the growing evidence that gender alone is not a determining factor in students' performance in core academic domains, reinforcing the importance of focusing on broader systemic factors when addressing learning outcomes.

Group Descriptives

	Group	N	Mean	SD	SE	Coefficient of variation
English	L	481	5.310	2.389	0.109	0.450
	P	705	5.216	2.374	0.089	0.455
Math (Quantitative)	L	481	5.210	2.025	0.092	0.389
	P	705	5.284	2.087	0.079	0.395

Figure 2. Standard Error and Coefficient of Variation by Gender and Domain

Although these statistical results demonstrate gender equity in performance, they do not mask the systemic issue of low academic readiness. As previously discussed in Table 1, average scores hovering around 5-6 out of 20, along with substantial floor effects, suggest that a large portion of students from both genders enter university with foundational academic gaps.

c. Median Score Analysis and Distribution Patterns

To explore the extent of students' highest academic achievement in each domain, Figure 3 illustrates the score distribution for English and Quantitative Knowledge using box plots. This visualization offers insights into central tendency (median), score variability, and the presence of outliers across genders.

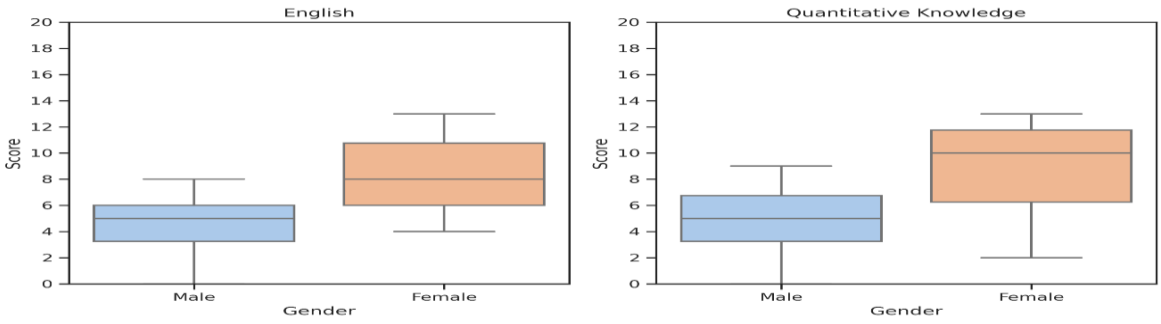

Figure 3. Boxplot of Score Distribution by gender and domain

As shown in Figure 3, the median scores for English proficiency were nearly identical between male and female students, indicating comparable central tendencies. The interquartile ranges for both groups were also comparable, indicating a consistent spread of scores within the central 50% of the distribution. However, male students had a slightly wider range, with the maximum score reaching 17, while the maximum score for female

students was 16. The minimum score for both groups was 0, indicating the presence of students with the lowest possible proficiency.

In contrast, the Quantitative Knowledge domain showed a somewhat different pattern. While the medians between genders were still very close, female students achieved a slightly higher maximum score (13) compared to males (12). This difference, although small, may reflect a marginal advantage among high-performing female students in this domain. Additionally, the presence of low-end outliers in both genders signals a minority of students with significantly lower-than-average abilities.

3.2 Discussion

This study employed a comparative quantitative method to examine gender-based performance differences and academic ceilings in English proficiency and Quantitative Knowledge. Descriptive statistics, Welch's t-test, and boxplot visualizations were performed using JASP software. Welch's t-test was selected due to its robustness under unequal variances and sample sizes. Boxplots were used to observe score distributions and identify score extremes across gender groups.

The results in Table 1 show no statistically significant gender differences in either domain, as indicated by similar means and standard deviations. These findings align with international benchmarks such as TIMSS and PISA, and meta-analyses that consistently report minimal or context-dependent gender differences in mathematics and language achievement [25], [26], [27].

This outcome reflects a selection system that appears gender-neutral on the surface; however, it may not be entirely so. However, when examined through the lens of structural inequality theory [20], such systems may inadvertently reproduce disparities that originate earlier in the education pipeline. Institutional cultures, admission norms, and the design of assessments often reflect the values and experiences of dominant groups, thereby overlooking latent disadvantages experienced by underrepresented or marginalized student populations.

This corresponds with prior evidence from Fang [15], which examined first-generation engineering undergraduate (FGEU) students in the United States. Fang found that the correlation between college entrance examination scores (ACT) and cumulative grade point average (GPA) was statistically significant for male students but not for their female counterparts. This suggests that standardized test scores, such as the ACT, may have limited predictive power for the academic performance of female students, particularly those from underrepresented or vulnerable backgrounds, like FGEU. In the Indonesian context, this highlights a cautionary insight: although entrance exam outcomes may appear gender-neutral, disparities in academic readiness, social support, or learning experiences may only become apparent after students enter university.

Therefore, equity analysis in university admissions should not be limited to comparing average or maximum scores; it should also examine the relationship between entrance performance and long-term academic outcomes. This approach is essential for designing a selection system that is not only objective but also responsive to the sociocultural and gender dynamics that shape higher education trajectories.

Figure 1 illustrates the descriptive statistics and maximum scores across groups, while Figure 2 confirms the lack of significant differences through Welch's t-test results and negligible effect sizes. These statistical indicators suggest that gender does not significantly influence entrance exam performance. However, deeper concerns arise from the low average scores (5–6 out of 20) and presence of floor-level scores (0–1), suggesting that many students answered only 25–30% of the questions correctly. This reflects foundational learning gaps, consistent with frameworks of college readiness that emphasize mastery of core competencies as crucial for postsecondary success [28], [29].

The limited maximum scores of 17 in English and 13 in Quantitative Knowledge further imply a misalignment between the entrance exam content and students' prior educational experiences [29], [30]. Research warns that overly difficult or poorly calibrated test items may penalize students [31], especially those from under-resourced academic backgrounds. Supporting literature [32], [33] similarly points to the role of teaching quality and curriculum relevance in shaping student performance.

The low performance in Quantitative Knowledge, as visualized in Figure 3, is particularly concerning. Prior studies have highlighted frequent errors in mathematical reasoning among first-year students, including misunderstandings of problem statements and incorrect procedural steps [34]. More than 70% of students misunderstand basic concepts such as set theory [35], while others struggle with advanced reasoning due to insufficient grounding in trigonometry and related skills [36]. These findings point to systemic gaps in instructional delivery at the secondary level. Moreover, the efficacy of teachers plays a vital role in addressing such challenges. As emphasized by Yoong and Hoe [37], strengthening teacher preparedness and diagnostic capabilities can improve student learning outcomes and reduce readiness gaps.

The presence of low-end outliers and the failure to reach full-score ceilings highlight inconsistencies between exam expectations and student capacity. From a diagnostic perspective, the low maximum scores in both domains, 17 and 16 for English, and 13 and 12 for Quantitative Knowledge, are noteworthy. Considering that the full score was 20, the results suggest that no student in the sample demonstrated full mastery of the tested material. This implies a ceiling effect, which may indicate that the exam items were too difficult or that students lacked adequate preparation at the secondary level.

These findings suggest the need for reform in the development of diagnostic testing. Item Response Theory (IRT) offers a sophisticated approach to ensure test validity, fairness, and the identification of Differential Item Functioning (DIF) [38], [39], [40], [41]. In diverse and educationally stratified contexts, such as Indonesia's 3T regions, or especially for students from marginalized backgrounds [42], IRT-based tools are essential for detecting subtle biases and improving equity [43].

As gender is not a significant determinant of performance gaps, the findings reinforce the need for system-wide readiness interventions. Moreover, the consistent presence of low-performing outliers suggests the need for matriculation programs or structured foundational training, as recommended by Black [44] and Maxwell and J. Gleason [45], to ensure equitable student success at the beginning of university life. Studies have consistently demonstrated that high-impact matriculation and bridge programs significantly enhance

student readiness and retention in higher education. A study by Black [44] reports that bridging programmes reinforce academic persistence by providing scaffolded support and smoother transitions to college-level coursework. Similarly, a comprehensive 2024 guide from the Education Strategy Group [46] highlights the multifaceted benefits, both academic and socio-emotional, of summer bridge initiatives, particularly for students from under-resourced backgrounds, such as Indonesia's 3T regions, where educational disparities are prevalent.

Therefore, these results warrant the implementation of integrated interventions, including redesigned entrance assessments, strengthened secondary education instruction, and targeted post-admission support. Such measures are crucial not only to elevate academic preparedness but also to create equitable conditions for student success at the onset of higher education.

In summary, although no gender-based inequalities were found, the consistently low performance patterns reveal a deeper systemic issue. Entrance exams must evolve to reflect realistic academic competencies better better and be accompanied by robust support systems that enable all students to thrive in higher education.

4. CONCLUSION

This study found that gender-based differences in entrance exam scores for English proficiency and Quantitative Knowledge were statistically negligible; however, the persistently low average and maximum scores across all students signal a deeper issue of academic underpreparedness. These findings are significant because they suggest that entrance assessments may not accurately reflect students' foundational competencies, especially in contexts with unequal access to quality secondary education.

From a theoretical standpoint, the results reinforce ongoing debates in educational assessment fairness, supporting the need for diagnostic-oriented approaches and alignment with readiness theory. The mismatch between test difficulty and student ability indicates a misalignment between assessment design and actual educational experience, an issue central to equity in access and selection.

To address these concerns, higher education institutions and national testing authorities should collaborate to redesign entrance exams using psychometric models. We recommend piloting IRT-based diagnostic tests that adjust item difficulty in real-time, along with bridging programs or matriculation that focus on foundational math and reading skills. Such interventions not only improve fairness but also support inclusive academic transition.

While the current study offers important insights, its scope was limited to two out of five exam domains and did not incorporate variables such as socioeconomic status or regional disparities. These results should be interpreted with caution, as the study did not include a control group and was conducted at a single institution. Future research should encompass a broader range of academic competencies, employ longitudinal tracking, and utilize Cognitive Diagnostic Models (CDMs) to identify latent learning gaps.

These conclusions should be considered in light of the study's limitations. The analysis was confined to a single institution and focused solely on two skill domains, without incorporating socioeconomic variables or control groups. Future research should expand the

diagnostic framework to include affective and contextual factors, as well as longitudinal tracking and validation across diverse educational settings, to inform more equitable and effective selection policies.

ACKNOWLEDGEMENTS

This research was supported by the Indonesia Endowment Fund for Education (LPDP). We sincerely thank the university for granting access to the entrance exam data and the anonymous reviewers whose valuable feedback significantly improved the clarity, rigor, and overall quality of this article.

REFERENCES

[1] O. D. M. Tuegeh, M. Harangi-Rákos, and A. S. Nagy, "Industry 4.0 and human resource in Indonesia: A systematic literature review," *Economic Annals-XXI*, vol. 190, no. 5, pp. 171–180, 2021, doi: 10.21003/EA.V190-16.

[2] E. Hermawan and A. L. Arifin, "What Expert Say about Empowering Human Resources in Supporting Leadership Function in Higher Education in the 21st Century," *Jurnal Iqra' : Kajian Ilmu Pendidikan*, vol. 6, no. 2, pp. 27–38, 2021, doi: 10.25217/ji.v6i2.1493.

[3] M. Nagy and R. Molontay, "Comprehensive analysis of the predictive validity of the university entrance score in Hungary," *Assessment and Evaluation in Higher Education*, vol. 46, no. 8, pp. 1235–1253, 2021, doi: 10.1080/02602938.2021.1871725.

[4] A. Aditomo and N. F. Faridz, "Ketimpangan Mutu dan Akses Pendidikan di Indonesia: Potret Berdasarkan Survei PISA 2015," Jan. 02, 2019. doi: 10.31227/osf.io/k76g3.

[5] E. James-Brabham, E; Loveridge, T; Sella, F; Wakeling, P.; Carroll, D. J.; & Blakey, "How do socioeconomic attainment gaps in early mathematical ability arise.pdf," pp. 1550–1565, 2023.

[6] R. I. Saptatiningsih and P. Jana, "Gender Perspective Concerning the Basic Mathematical Abilities," *PONTE International Scientific Researchs Journal*, vol. 75, no. 12, pp. 182–187, 2019, doi: 10.21506/j.ponte.2019.12.12.

[7] Adaeze Regis-Onuoha and K. U. Chukwu, "Gender Differences and the Writing Achievement of University Fresh Students: A study of Federal University of Technology Owerri," *Journal of Gender, Culture and Society*, vol. 2, no. 1, pp. 11–16, 2022, doi: 10.32996/jgcs.2022.2.1.2.

[8] T. Pekkarinen, "Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations," *Journal of Economic Behavior and Organization*, vol. 115, pp. 94–110, 2015, doi: 10.1016/j.jebo.2014.08.007.

[9] A. Kan and O. Bulut, "Examining the Relationship Between Gender DIF and Language Complexity in Mathematics Assessments," *International Journal of Testing*, vol. 14, no. 3, 2014, doi: 10.1080/15305058.2013.877911.

[10] P. Akyol, J. Key, and K. Krishna, "HIT OR MISS ? TEST TAKING BEHAVIOR IN MULTIPLE CHOICE EXAMS," *Annals of Economics and Statistics*, vol. 147, no. 147, 2022.

[11] F. Reinhold *et al.*, "The role of spatial, verbal, numerical, and general reasoning abilities in complex word problem solving for young female and male adults," *Mathematics Education Research Journal*, vol. 32, no. 2, pp. 189–211, 2020, doi: 10.1007/s13394-020-00331-0.

[12] S. L. Tsai, M. L. Smith, and R. M. Hauser, *Gender gaps in student academic achievement and inequality*, vol. 20. 2018. doi: 10.1108/S1479-353920180000020008.

[13] D. Montolio and P. A. Taberner, "Gender differences under test pressure and their impact on academic performance: A quasi-experimental design," *Journal of Economic Behavior and Organization*, vol. 191, pp. 1065–1090, 2021, doi: 10.1016/j.jebo.2021.09.021.

[14] H. Wen, Y. Liu, and N. Zhao, "Longitudinal Cognitive Diagnostic Assessment Based on the HMM/ANN Model," *Frontiers in Psychology*, vol. 11, 2020, doi: 10.3389/fpsyg.2020.02145.

[15] N. Fang, "The gender effect on academic success among first-generation engineering undergraduate (FGEU) students," *Global Journal of Engineering Education*, vol. 27, no. 1, pp. 16–21, 2025.

[16] A. Ocvitasari, D. E. Widiasi, and M. Firmansyah, "Analisis Faktor Kesiapan Akademik Terkait Orientasi Motivasi Berprestasi, M-SCORE dan Kegiatan Non-Akademik Terhadap Prestasi Akademik Mahasiswa," *Jurnal Kedokteran Komunitas*, vol. 8, no. 2, pp. 74–82, 2020.

[17] S. Sriyati, "Analisis Hasil Tes Potensi Skolastik sebagai Indikator Kesiapan Siswa Menghadapi Tes UTBK 2022," *Tarbiyah Wa Ta'lim: Jurnal Penelitian Pendidikan dan Pembelajaran*, vol. 9, no. 2, pp. 74–83, 2022, doi: 10.21093/twt.v9i1.4210.

[18] T. Anjar, "Student Readiness High School Entrance Exams Higher Education and the Role of School Counselors," *GUIDENA: Jurnal Ilmu Pendidikan, Psikologi, Bimbingan dan Konseling*, vol. 5, no. 1, p. 50, 2015, doi: 10.24127/gdn.v5i1.62.

[19] B. R. Nainggolan, P. Purnawarman, and D. Sukyadi, "Comprehension Processes and Critical Thinking Skills in Two University Entrance Exams in Indonesia: a Content Analysis of the Sbmptn and the Simak Ui 2008-2019," *Polyglot: Jurnal Ilmiah*, vol. 18, no. 1, p. 1, 2022, doi: 10.19166/pji.v18i1.4908.

[20] R. Naylor and N. Mifsud, "Structural inequality in higher education: Creating institutional cultures that enable all students," 2019.

[21] M. S. Nitko, Anthony J..,Brookhart, *Educational Assessment of Students: Pearson New International Edition. Always learning. Pearson custom library.* PEARSON, 1987.

[22] A. Sudheesh and R. M. Golden, "Exploring Temporal Functional Dependencies Between Latent Skills in Cognitive Diagnostic Models," in *Springer Proceedings in Mathematics and Statistics*, 2021, pp. 253–262. doi: 10.1007/978-3-030-74772-5_23.

[23] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Second Edi. USA: Lawrence Erlbaum Associates, 1988.

[24] G. M. Sullivan and R. Feinn, " Using Effect Size—or Why the P Value Is Not Enough ," *Journal of Graduate Medical Education*, vol. 4, no. 3, pp. 279–282, 2012, doi: 10.4300/jgme-d-12-00156.1.

[25] E. Ghasemi and H. Burley, "Gender, affect, and math: a cross-national meta-analysis of Trends in International Mathematics and Science Study 2015 outcomes," *Large-Scale Assessments in Education*, vol. 7, no. 1, pp. 1–25, 2019, doi: 10.1186/s40536-019-0078-1.

[26] D. Giofrè, C. Cornoldi, A. Martini, and E. Toffalini, "A population level analysis of the gender gap in mathematics: Results on over 13 million children using the INVALSI dataset," *Intelligence*, vol. 81, no. June, p. 101467, Jul. 2020, doi: 10.1016/j.intell.2020.101467.

[27] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, "Cross-National Patterns of Gender Differences in Mathematics: A Meta-Analysis," *Psychological Bulletin*, vol. 136, no. 1, pp. 103–127, 2010, doi: 10.1037/a0018053.

[28] D. T. Conley, "College Readines," pp. 1–11, 2007.

[29] R. J. Shavelson *et al.*, "Gap analysis and data visualization: Creating a more equitable assessment culture in higher education," *Thinking Skills and Creativity*, vol. 47, no. 8, pp. 1274–1288, 2022, doi: 10.1080/02602938.2022.2067265.

[30] M. Fine and K. Pryiomka, "Assessing College Readiness through Authentic Student Work: How the City University of New York and the New York Performance Standards Consortium Are Collaborating toward Equity," *Learning Policy Institute*, no. July, 2020.

[31] R. Liu and K. D. Bradley, "Differential Item Functioning Among English Language Learners on a Large-Scale Mathematics Assessment," *Frontiers in Psychology*, vol. 12, no. July, pp. 1–12, 2021, doi: 10.3389/fpsyg.2021.657335.

[32] L. Stoffelsma and W. Spooren, "The Relationship Between English Reading Proficiency and Academic Achievement of First-Year Science and Mathematics Students in a Multilingual Context," *International Journal of Science and Mathematics Education*, vol. 17, no. 5, pp. 905–922, 2019, doi: 10.1007/s10763-018-9905-z.

[33] A. Tunde, "Educational Resources Availability and Utilization As Determinant of Students Academic Performance in South West Nigeria," *African Journal of Education and Practice*, vol. 8, no. 5, pp. 45–63, 2022, doi: 10.47604/ajep.1636.

[34] H. Disnawati and Y. N. Deda, "Miskonsepsi Mahasiswa Pada Materi Himpunan: Analisis Menggunakan Kriteria Certainty of Response Index," *Jurnal Pendidikan Matematika (Jupitek)*, vol. 4, no. 2, pp. 95–102, 2022, doi: 10.30598/jupitekvol4iss2pp95-102.

[35] H. Disnawati, C. N. Salsinha, N. Deda, and N. L. Bees, "Student s ' Mathematical Reasoning Errors in Solving Computer Based Written Exam Model Questions Based on Newman ' s Theory," vol. 7, no. 2, pp. 490–501, 2023.

[36] Nisa Triyatul Fitri, Arif Muchyidin, and Reza Oktiana Akbar, "Mastery of Trigonometry Material and Its Effect on Students' Ability to Solving Three Dimensional Problems," *Journal of Mathematics Instruction, Social Research and Opinion*, vol. 1, no. 2, pp. 97–106, 2022, doi: 10.58421/misro.v1i2.23.

[37] S. M. Yoong and Fu Sai Hoe, "Teachers' Efficacy to Teach Students with Mathematics Learning Difficulties," *Journal of Mathematics Instruction, Social Research and Opinion*, vol. 1, no. 3, pp. 117–130, 2022, doi: 10.58421/misro.v1i3.29.

[38] E. Istiyono, W. S. B. Dwandaru, Y. A. Lede, F. Rahayu, and A. Nadapdap, "Developing IRT-based physics critical thinking skill test: A CAT to answer 21st century challenge," *International Journal of Instruction*, vol. 12, no. 4, pp. 267–280, 2019, doi: 10.29333/iji.2019.12417a.

[39] H. Ravand, "Assessing Measurement Invariance in a University Entrance Exam: A Comparison of Multigroup Confirmatory Factor Analysis Alignment Method vs. Multigroup Item Response Theory," *Educational Methods and Psychometrics*, vol. 2, no. 1, pp. 21–30, 2024, doi: 10.61186/emp.2024.4.

[40] E. Apino, H. Retnawati, W. Purbani, and K. Hidayati, "The Statistical Literacy of Mathematics Education Students: An Investigation on Understanding the Margin of Error," *TEM Journal*, vol. 13, no. 1, pp. 293–302, 2024, doi: 10.18421/TEM131-31.

[41] J. Arlinwibowo, H. Retnawati, and B. Kartowagiran, "Item Response Theory Utilization for Developing the Student Collaboration Ability Assessment Scale in STEM Classes," *Ingenierie des Systemes d'Information*, vol. 26, no. 4, pp. 409–415, 2021, doi: 10.18280/ISI.260409.

[42] S. Ketabi, Dr. S. M. Alavi, and Dr. H. Ravand, "Diagnostic Test Construction: Insights from Cognitive Diagnostic Modeling," *International Journal of Language Testing*, vol. 11, no. 1, 2021, [Online]. Available: https://api.semanticscholar.org/CorpusID:276481758

[43] S. M. Alavi, H. Karami, and M. H. Kouhpaeenejad, "Examining the Fairness of the University Entrance Exam: A Latent Class Analysis Approach to Differential Item Functioning," *Atu Press Issues in Language Teaching (Ilt)*, vol. 10, no. 1, pp. 147–170, 2021.

[44] A. M. Black, "The role of bridging programmes in supporting student persistence and prevention of attrition: a UK case study," *Studies in Higher Education*, vol. 49, no. 9, pp. 1519–1531, 2024, doi: 10.1080/03075079.2023.2269246.

[45] M. Maxwell and J. Gleason, "Item efficiency: an item response theory parameter with applications for improving the reliability of mathematics assessment," *International Journal of Mathematical Education in Science and Technology*, vol. 50, no. 2, pp. 216–243, 2019, doi: 10.1080/0020739X.2018.1492038.

[46] Learn & Work Ecosystem Library, "Bridging the Gap: A State Leader's Guide to Implementing Summer Bridge Programs (Education Strategy Group)." Accessed: Jun. 25, 2025. [Online]. Available: https://learnworkecosystemlibrary.com/initiatives/bridging-the-gap-a-state-leaders-guide-to-implementing-summer-bridge-programs-education-strategy-group/