Innovation in Language Testing and Teaching (ILTT)
Volume 1 Issue 1 Year 2025 Pages 21 - 38
E–ISSN : 3110-5483 | DOI: 10.61227
E-mail : iltt.pusmedia@gmail.com
https://journal.pusmedia.com/index.php/ILTT/index

How University Entrance Examinations Can Threaten Validity in Higher Education Admissions

Hossein Salarian [1]*

[1] University of Tehran, Iran

Correspondence e-mail * : hos_salarian@ut.ac.ir

Abstract: Although university entrance examinations are designed to promote fairness and objectivity in admissions, they can at times compromise the validity of the process by prioritizing test-taking strategies over genuine academic ability or potential. To this end, this study explor the validity, fairness, and predictive value of university entrance examinations among 400 students via a mixed-methods research design. A cross-sectional survey was carried out to collect quantitative data analyzed through statistical software (SPSS), while semi-structured interviews supplied qualitative depth to explore the nuances behind the statistical trends. Also, ANOVA tests assessed differences across regions. Interview transcriptions were coded through NVIVO 8 and analyzed using thematic analysis. Moreover, academic records (GPA and entrance exam scores) obtained to investigate predictive validity. Quantitative results showed moderate acceptance of entrance exams' predictive power, but significant concerns around fairness, regional equity, and psychological stress. Semi-structured interviews revealed that many students felt the exams forced rote learning and unfairly favored urban, well-resourced applicants. Students reported stress, anxiety, and health impacts from the exam's high-stakes nature. Prior research supports these findings, noting GPA as a stronger long-term predictor of performance. Overall, the study highlights that while entrance exams can motivate, they risk undermining fairness and mental health. Reform efforts, including multi-dimensional admissions models combining GPA, interviews, and personal statements, are recommended to achieve greater equity and validity in higher education admissions. This study has some theoretical and practical implications along with providing some suggestions for further studies to improve it.

Keyword : University entrance exam; High-stakes test; Positive/negative washback; Stakeholders

Article info: Submitted : 2025-04-04 | Accepted : 2025-06-10 | Published : 2025-07-15

INTRODUCTION

A fundamental step in test design is the enactment of the test's validity. When a test involves in validity, it has a means of measuring what it asserts to measure, and this addresses connecting test scores to achievement in other attempts. In this view, learners' assessment submission displays their actual capability. Otherwise, when there is no specific standards-based system of measurement for passing the test legitimately, test results are just self-referential (Boud & Bearman, 2024). According to Maxwell (2004), validity is a goal and deals with the question of the connection between conclusions and reality. Furthermore, in a validity analysis (VA) we

determine probable threats and discuss and identify how to consider them. A validity threat (VT) is a particular manner in which we might be wrong. A particular selection or action applied to develop validity through dealing with a specific threat is known as a mitigation strategy. These threats consider determinants that can debilitate the accuracy and reliability of study findings, potentially resulting in incorrect interpretations or conclusions. They can impact various facets of validity, composing of internal, external, construct, and statistical conclusion validity. Other threatening variables can be maturation, history, testing, instrumentation interaction of testing and maturation, interaction of testing and the experimental factor (Zheng, et al. 2024).

On the other hand, Alderson and Wall (1993) believed that high-stakes tests can not merely exert direct and influential impacts on teaching and learning tasks, but they can also have powerful domination on instructors' and learners' attitudes, behaviors, and motivation as Shohamy (1997) maintained it. Likewise, Moses and Nanna (2007) asserted that high-stakes tests are the ones in which their scores have a direct influence on a testee's life options and chances. In fact, it is on such foundation that evaluation systems are conceived as high-stakes while they involve in critical consequences for students (Thomas, 2005; Ysseldyke et al., 2004). Thus, the determining component for pondering a test as high-stakes is the ascription of consequences to the outcomes. Such consequences have been growingly at the focal point of language testing scholars' concentration ever since Messick's (1989) validity matrix introduction, based on which Alderson and Wall (1993) paradoxically asked "does washback exist?" and Brennan (2006) undeniably maintained that maybe the most controversial matter in validity has been the function of consequences.

Moreover, despite the widely applications of university entrance examinations for evaluation and selecting applicants for higher education, these tests are intended to supply a standardized, fair and objective method of assessing student readiness. Nevertheless, their implementation can inadvertently threaten and undermine the validity of this process in the cases of content relevance, equity, predictive accuracy and variables beyond academic abilities and hence result in unfair or inaccurate admissions decisions. As a result, validity-types framework should be applied to arrange and contextualize threats to validity in instructional applications of supervised learning (Chapelle and Voss, 2021). Meanwhile, test score pollution which is any impact that influences the accuracy of achievement test scores as contaminants in Messick's (1984) term, can threaten the validity. In other word, this practice of teaching to the test which raises test scores, though not associated with the construct being measured, is one sort of test score pollution creating construct-irrelevant test sore variance (Gipps, 1995). This is what Linn (1981) regarded it as improvement test score without actually mastering the skills or constructs being assessed (cited in Gipps, 1995). To this reason, achievement in this kind of entrance exams is a sophisticated constellation of knowledge and skills and no current test appears to be

comprehensive toward the end of assessing the complete domain (Haladnya, 1991). In the same vein, Kane (2013) contended that the assessment of test score uses involves in a measurement and consideration into the consequences of the supposed uses, and that negative consequences can cause a score use to become unreasonable and unpleasant So, this study considers the following questions:

1. To what extent do university entrance examination scores predict academic performance in higher education?
2. How do students perceive the fairness and relevance of university entrance examinations?
3. What are the regional disparities, if any, in access to resources that support preparation for entrance exams?
4. How do students experience psychological stress related to university entrance examinations, and what factors contribute to this stress?

Review of Literature

1. Validity

Haertel and Herman (2005) argued that concept of validity involve in a great history with various points of view, frameworks, and terminology. In other word, validity was historically regarded as a statistical property of tests. Nevertheless, in recent years, most researchers perceive validity not as an inherent characteristic of the tests themselves but rather as an assessment of the suitability and importance of the results drawn from test scores (Chapelle & Voss, 2021). It's often shown as having a gatekeeping function in scientific research (Johnson et al., 2008). In the setting of modern validity theory and the Standards for Educational and Psychological Testing (AERA, 2014; Kane, 2013), the main emphasis or concentration is on the process of validation and not merely on the tool being validated. Based on Messick's unified framework (1989), construct validity presumes a focal function through combining different validity facets. In cooperation with Chapelle and Voss (2021), Kane (2013) supported applying Toulmin's (2003) informal logic principles in data gathering and analysis for validation composing of evidence, backing, warrants, counterevidence, and qualifiers.

2. Validity Threats and Entrance Examinations

Scholars propose two usual methods for dealing with these threats: 1) They can supply a comprehensive description of the construct of interest. For instance, in the case of applying supervised learning to assess "authentic questioning" (Kelly et al., 2018). 2) They can confirm any argument or challenges in putting into operation the construct. For instance, in foreseeing graduation through clarifying the problem of explaining "drop-out" given that learners usually quitting college for periods of time while wanting to return.

On the other hand, Wohlin et al. (2000) considered four key kinds of validity threats: internal, external, conclusion and construct. Internal validity emphasizes the amount of confidence about the treatment that really caused the results. There can be other components that have caused the results, components that we do not have considered or have not assessed them. Conclusion validity deals with the amount of confidence about the treatment we applied in an experiment. It is actually associated with the real result we observed. Generally, this concerns whether there is a statistically significant impact on the results. Threats to external validity involve in components that confine the generalizability of the results to other contexts, or occasions or populations. These threats can decrease the proportion of the results of a research can be certainly utilized to situations beyond the particular context of the research. In other word, this kind of validity deals with whether we can generalize the outcomes outside the scope of the study. Construct validity considers the association between the theory relating to the experiment and the observation(s). Even when we have realized that there is a casual connection between the treatment of our experiment and the observed result, the treatment may not related to the cause we consider we have dealt with and changed (Fulcher, 2014).

In the same vein, Bailey (1999) argued that many entrance exams focus narrowly on academic knowledge and specific subject areas, which may not reflect the full range of skills required for success in university. For instance, critical thinking, creativity, and communication skills are often underrepresented, despite being crucial in higher education settings. This disconnection can reduce the content validity of the test. Construct validity is compromised when exams fail to measure the constructs they claim to assess. For example, a test intended to measure mathematical reasoning may instead assess rote memorization or test-taking strategies. If students are rewarded for strategies unrelated to true understanding, the test results become less meaningful (Bennett et al., 2011). Entrance exams are often justified by their ability to predict future academic success to consider the predictive validity. However, studies have shown that high school GPA is often a better predictor of university performance than standardized test scores. Relying too heavily on entrance exams can result in overlooking students who may succeed in a more holistic academic environment (Bird et al., 2021)

Furthermore, Borsboom et al. (2004) asserted a realist, causality-based explanation in which variation in the quality under examination creates variation in test scores. This formalization of a concept is related to (reflective) latent facet modeling because responses to items are seemed to be an influence of the respective latent variable(s). In addition, a validation crisis is a forerunner to a replication crisis, causing even replicable outcomes uninformative (Schimmack, 2021). Similarly, questionable assessment practices deal with the degrees of freedom that scholars purposefully or unintentionally achieve as a desired outcome (Flake & Fried, 2020).

On the hand, Standardized entrance exams can disadvantage students from underprivileged backgrounds who may lack access to test preparation resources, tutoring, or high-quality schooling. Cultural and linguistic biases embedded in test items can also skew results, further disadvantaging minority students. Therefore, fairness and equity are important issues in entrance examinations (Moses & Nanna, 2007). Meanwhile, the high stakes associated with entrance exams can place immense psychological pressure on students, leading to stress, anxiety, and even burnout. These pressures can distort performance and do not necessarily reflect a student's true potential (Kane, 2013). Besides, theory of action (TOA) deals with consequential studies at different levels, from individual scores to aggregate outcomes along with the social and educational settings of examinations. TOA studies presents more prominence to the impacts of the evaluation system on learners and institutions as well as to the underlying procedures behind those impacts. This theory addresses a social point of view relating research into consequences to engage in not merely intended but also unintended impacts (Bennett et al., 2011).

3. Washback Effect EFL Testing for University Entrance on EFL Teaching

The term 'washback' came to be used in the field to involve in the power that high-stakes tests could have on language teaching and learning although the terms 'impact' or 'consequences' are more commonly used in the field of education (Scott, 2007). It is believed that high-stakes tests (e.g. university entrance examinations) can have significant impact not only on individuals but also on practices and policies in the classroom, the school, the education system and in society as a whole. High-stakes tests usually involve in a set of determining functions in testees' lives ranging from employment and promotion to placement and achievement (Wall, 2005). The outcomes of the examinations, as Wall (2000) regarded them as "differentiating rituals" can often be so significant in the testees' future lives that they involve the test-takers to take any possible measures to results of the tests. There is a clear tendency for teachers to adopt certain skills, techniques, and tasks in order to meet the test demands and satisfy the students' needs.

Similarly, one consequence of high-stakes testing is the danger of test-driven learning and teaching. When students and other test-takers know that one single measure of performance will determine their lives, they are less likely to take a positive attitude toward learning. The motives in such a context are almost exclusively extrinsic, with little likelihood of stirring intrinsic interests. Teachers also get caught up in the wave of this world-wide issue. The effect of this policy was undue pressure on teachers to make sure their students excelled in the exam, possibly at the risk of ignoring other objectives in their curriculum (Douglas Brown, 2004).

Frederiksen in his paper called 'The real test bias' in 1984 suggested that because test information is important in attempting to hold schools accountable, the

influence of tests on what is taught is potentially great. Also, Madaus (1988) on the impact of testing on the curriculum and teaching states the power of tests and exams to affect individuals, institutions, curriculum or instruction is a perceptual phenomenon. If teachers, students or administers believe that the results of an examination are important, it matters very little whether this is really true or false.

The effect is produced by what individuals perceive to be the case, and if important decisions are presumed to be related to test results, then teachers will teach to the test (p.92).

Method

Participants

A sample of 400 male and female participants of different cities in Iran took part in this study based on the random sampling. They were part of a larger population and were chosen from applicants of University Entrance Examinations form various cities. Their ages ranged from 17 to 25 years. All the participants spoke Persian as their first language.

Research Design

This study employed a mixed-methods research design, integrating both quantitative and qualitative approaches and made use of triangulation. Dornyei (2007) believes that the value of triangulation is that it reduces the researcher bias and enhances the validity and reliability of the information. A cross-sectional survey was conducted to collect quantitative data, while interviews provided qualitative depth to explore the nuances behind the statistical trends. The study included 400 participants from various regions across the country, selected through stratified random sampling to ensure regional representation.

Materials and Instruments

The primary instruments used in this study were:

1. A structured questionnaire consisting of both closed-ended and Likert-scale items designed to measure students' perceptions of entrance exam relevance, fairness, and stress levels.
2. Semi-structured interview guides for in-depth interviews with selected participants to gather qualitative insights.
3. Academic records (GPA and entrance exam scores) obtained with participant consent to assess predictive validity.

Data Gathering Procedure

Ethical clearance was obtained from relevant institutional review boards. Participants were recruited through schools and universities, with informed consent collected prior to participation. The questionnaire was administered online and in-

person, depending on the region's accessibility. Semi-structured interviews were conducted with a subset of 120 participants (30% of the sample), either face-to-face or via video conferencing. Academic records were collected and anonymized to protect participant privacy.

Data Analysis

Quantitative data were analyzed using statistical software (SPSS):

1. Descriptive statistics (means, standard deviations) summarized the responses.
2. Inferential statistics, including correlation and regression analysis, assessed the relationship between entrance exam scores and academic performance.
3. ANOVA tests evaluated differences across regions.

Qualitative data from interviews were transcribed and analyzed using thematic analysis. Emerging themes related to perceptions of fairness, stress, and educational equity were coded and interpreted to complement the quantitative findings.

Results and Discussion

Results

The first research question aim to examine to what extent university entrance examination scores predict academic performance in higher education. The results of this study shows that most of the students generally view university entrance examination scores as an efficient instrument for academic performance in higher education. The results are summarized in Table 1, with mean scores and standard deviations for each statement related to the effectiveness of university entrance examination:

Table 1.
Effectiveness of University Entrance Examination as a Tool For Academic Performance in Higher Education

No	Statements	Mean	Std. Deviation
1.	The entrance exam accurately reflects my academic abilities.	40.38	5.063
2.	My entrance exam score is a good predictor of my university performance.	33.24	5.184
3.	I believe my high school GPA better represents my abilities than my entrance exam score.	29.75	5.355
4.	I feel that the exam measures important skills for university success.	34.73	5.283
5.	I believe my high school GPA better represents my abilities than my entrance exam score.	35.69	5.394

6. The entrance exam motivates students to study harder.	38.65	5.520
7. University entrance examination can be a good instrument for academic performance in higher education.	41.61	5.076

Mean Average = 35.39

Table 1 summarizes participants' responses concerning the effectiveness of the university entrance examination as a predictor of academic performance in higher education. The results indicate a generally moderate level of agreement, with a mean average of 35.39 across the seven statements. Notably, the statement "University entrance examination can be a good instrument for academic performance in higher education" received the highest mean score (M = 41.61, SD = 5.076), suggesting that some respondents still perceive a role for entrance exams in predicting academic readiness. Conversely, the statement "I believe my high school GPA better represents my abilities than my entrance exam score" scored relatively lower (M = 29.75, SD = 5.355), pointing to a perception that high school GPA might more accurately reflect sustained performance than a one-time test. Additionally, a significant proportion of participants felt that entrance exams motivated students to study harder (M = 38.65, SD = 5.520), but only modest confidence was shown in the predictive validity of entrance exam scores themselves (M = 33.24, SD = 5.184). Overall, these data suggest that while entrance exams are viewed as somewhat useful tools, their ability to predict long-term academic achievement is perceived to be limited, aligning with the findings highlighted in the discussion of Research Question 1.

The second research question explore how students perceive the fairness and relevance of university entrance examinations. In other word, this study explored the manner students perceive the fairness and relevance of university entrance examinations. By analyzing students' responses, it is evident that university entrance examinations are fair and relevant for academic performance in higher education but they are much greater than many students' proficiency. This result is provided in Table 2.

Table 2.
Fairness and Relevance of University Entrance Examinations

No	Statements	Mean	Std. Deviation
1.	The entrance exam content is relevant to what I studied in high school.	40.78	5.061
2.	The entrance exam is fair to students from all backgrounds.	35.54	1.174
3.	My entrance exam score truly reflects my potential for university success.	32.71	1.265
4.	Entrance exams promote equal opportunities for all students.	33.64	1.173
5.	Tutoring resources were equally available to all students in my community.	32.39	5.294
6.	The entrance exam was fair to students from diverse backgrounds.	34.67	5.122
7.	The entrance exam score reflects our true potential.	35.61	1.276

Mean Average = 33.68

Table 2 presents students' perceptions of the fairness and relevance of university entrance examinations. The overall mean responses suggest moderate agreement with statements about fairness and relevance, with the highest mean (M = 40.78, SD = 5.061) reported for the relevance of entrance exam content to high school studies. However, students were less convinced that their entrance exam score fully reflected their potential for university success (M = 32.71, SD = 5.265), revealing concerns about the exam's validity as a measure of true academic ability. Statements addressing fairness to students from diverse backgrounds (M = 34.67, SD = 1.122) and the promotion of equal opportunities (M = 33.6, SD = 5.173) also scored moderately, indicating that while some students saw entrance exams as reasonably fair, others remained skeptical, especially about equal access to tutoring resources (M = 32.39, SD = 5.294). Altogether, these results highlight a tension between perceiving entrance exams as broadly fair while still questioning their capacity to capture diverse students' actual academic readiness, which supports earlier literature on equity challenges in standardized testing.

The third research question investigated the regional disparities, if any, in access to resources that support preparation for entrance exams. Regarding this research question, which examines how university entrance exams influence students' motivation and attitudes, the data suggests a nuanced picture.

Table 3.
Regional Disparities in Access to Resources for Entrance Exams

No	Statements	Mean	Std. Deviation
1.	The entrance exam was regionally and culturally unbiased.	42.68	4.061
2.	My school provided enough support to help me prepare for the entrance exam.	38.44	4.194
3.	My entrance exam score truly reflects my potential for university success.	28.81	4.255
4.	I think high school GPA should weigh more than entrance exam scores in admissions decisions.	37.83	4.183
5.	There were no regional disparities in access to resources of entrance exams.	36.69	4.194
6.	I believe the entrance exam disadvantages students from rural areas.	35.48	4.120
7.	All of the students in everywhere have the same resources for this exam.	34.51	4.176
	Mean Average = 34.52		

As shown in Table 3, the data reveals clear inequities. This Table shows that while some students perceived adequate support in their schools (M = 38.44), many disagreed with the notion that all students have equal resources (M = 34.51), and concerns about regional or rural disadvantages were evident (M = 35.48). These findings confirm prior studies (Reay, 2018; Yang & Gustafsson, 2004) showing that urban students benefit from better preparation resources compared to rural students, raising questions about the fairness and inclusiveness of a system heavily reliant on standardized tests.

The last research question dealt with the manner students experienced psychological stress related to university entrance examinations, and what factors contributed to this stress. For Research Question 4, which focuses on comparing high school GPA and entrance exam scores in predicting academic achievement, evidence points to a clear skepticism toward the predictive validity of entrance exams.

Table 4.
Students' Experience for Psychological Stress Related to University Entrance Exam Examinations

No	Statements	Mean	Std. Deviation
1.	Preparing for the entrance exam caused me significant stress.	42.78	5.062
2.	The pressure of the entrance exam negatively affected my health.	41.54	5.184
3.	Entrance exams should be replaced by other forms of assessment.	39.71	5.155
4.	I feel confident in my performance during the entrance exam.	35.62	5.173
5.	I believe interviews and personal statements should be part of the admission process.	36.49	5.174
6.	I believe the entrance exam disadvantages students from rural areas.	37.68	5.160
7.	I feel emotionally before, during, and after taking the exam.	38.61	5.156
	Mean Average = 36.59		

For Research Question 4, related to students' psychological stress experiences, the results in Table 4 highlight substantial concerns. Students reported high stress levels preparing for the exam (M = 42.78), health impacts (M = 41.54), and supported alternative forms of assessment (M = 39.71). This supports previous research (Putwain, 2008; von der Embse et al., 2018) indicating that high-stakes tests can generate significant anxiety, potentially reducing students' well-being and performance, and pointing to a need for reform in admissions systems.

In addition, for investigation of the differences across regions ANOVA was used through the latest version (v26.0) of SPSS. It was used to determine whether there are statistically significant differences in the perceptions of fairness. For this reason the average Likert-scale score for "fairness" for the questionnaire was considered as the dependent variable and the various regions of the country were considered as the independent variables (Table 5).

Table 5.
ANOVA Analysis Results for Mean Scores of the Two Study Groups Based on Fairness

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2458.056	2	1384.064	64.74	.164
Within Groups	3594.139	396	62.25		
Total	5692.146	398			

Note. The Sig. value for ANOVA is more than the alpha level ($P > 0.05$); the assumption of homogeneity of variances is not violated.

Table 5 displays the outcomes of an ANOVA analysis utilized to make sure where the observed mean difference was statistically significant. After that, a Scheffé's test was employed to reveal this issue. As can be shown in this table, the two groups conducted quite well on the tests; nonetheless, the mean scores were diverse to each other, suggesting that the two conditions have variously impacted on the achievement scores of the students. The mean difference is in favor of the "north" and "central" regions in which students have prepared themselves in various distinctive ways to achieve better scores in this exam. To further explore the meaningfulness of the mean difference across these groups across regional disparities in the perceptions entrance exam a one-way ANOVA was conducted using region as an independent variable. Then, a post-hoc analysis was done to identify accurately which regions differ from each other.

Table 6.
Results of the Post Hoc Comparison of the Mean Scores

(I) Exp. Groups	(J) Exp. Groups	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
A	B	18.43*	2.235	.026	.58	15.94
	C	19.052*	2.235	.000	7.43	25.74
B	A	9.027*	2.235	.023	4.59	4.25
	C	6.437*	2.235	.005	3.81	8.51
C	A	8.868*	2.235	.000	2.98	4.62
	B	4.631*	2.235	.004	1.34	2.46

Note. The Sig. value for each comparison set is smaller than the alpha value (($P > 0.05$), and no confidence interval contains zero; all the three means are significantly different. A: South, B: North, C: Rural

As displayed in Table 6, all the means are significantly different, as the mean difference is statistically significant for each comparison set ($P > 0.05$), showing that the students' test performance scores were diversely impacted by all the facilities of the specific regions. Based on this table, the means of urban group are greater than that of rural group and the mean of the central region group and capital student group are more than those of the others. Therefore, the north and capital student group performed better than the other groups.

Semi-Structured Interviews

Semi-structured interviews intended to complement the quantitative data obtained from students' questionnaires. As a qualitative research approach, this sort of interview engaged in asking eight open-ended questions composing of prompt

discussion and background questions with diverse factors. Through coding, this sort of interview was applied to investigate these students' attitudes towards the fairness of the exam, their experiences on preparing for the university entrance examination and their reactions towards the scores in this exam reflecting their true potential.

Likewise, their answers and recommendations were divided into codes. Lastly, the codes were integrated and thus, four classifications were constructed. By means of thematic analysis, the obtained data were investigated based on Schmidt's (2004) five phases. In the initial phase, analytical classifications were constructed via all the transcripts to investigate the themes and the associated facets aspects focused on the fairness and equity of this exam. In the second phase, the data entailed were formulated and gathered. In the third phase, based on coding guide the analytical procedure was identified. In the fourth phase, the coding outcomes were quantified so as to supply a common summary of the distributions in the transcripts relating to frequencies in the analytical classifications. In the last phase, specific considerations of each interviewee's transcript were created to propose new suggestions. These categories were: overall fairness and equity (f=33), idea development (f=20), teaching to test (f=29), reflection of true potential (f=37). What's more, for analyzing the transcripts NVIVO 8 was applied analysis.

Furthermore, the semi-structured interviews complemented the surveys by providing in-depth views on exam fairness, stress, and true potential. Many students described how the entrance exam pushed them to memorize content without true understanding, leading to surface-level learning. Others shared that unequal tutoring opportunities distorted fairness, while repeated references to anxiety and fear supported the quantitative data on stress. Participants also valued GPA as a fairer indicator of their real abilities, proposing more comprehensive evaluation methods beyond a single test. These qualitative findings strengthen the argument that a combination of GPA, personal statements, and interviews could offer a more balanced and valid admissions framework.

Discussion

Overall, the study confirms that while entrance exams moderately predict academic performance (RQ1), their fairness, equity, and predictive validity are subject to skepticism. Students expressed doubts that entrance exams fully captured their skills, suggesting a preference for a more holistic evaluation. This aligns with studies questioning the sole use of standardized tests in higher education admissions. Besides, while entrance examinations were perceived as moderately effective in predicting academic performance, students questioned their accuracy, as evidenced by the modest mean score (M = 33.24). This supports studies by Atkinson and Geiser (2009), who argued that standardized exams are limited in measuring academic potential comprehensively.

In terms of fairness and relevance (RQ2), students appreciated that the exams covered relevant content but worried about unequal opportunities and cultural bias, echoing Alderson and Wall (1993) on test washback effects. They feared that these exams perpetuate social inequalities rather than level the playing field. Although some students felt that entrance exams covered relevant content (M = 3.78), concerns about fairness, equal opportunities, and availability of resources persisted. This resonates with previous findings by Madaus and Clarke (2001), who noted that entrance tests often perpetuate educational inequities.

Regarding motivation and regional disparities (RQ3), the exams did motivate students to study harder, but unfair differences in regional preparation resources risk undermining this motivational benefit. Students from rural or under-resourced areas were particularly disadvantaged, reflecting the ongoing challenge of achieving fairness across diverse student populations. Students acknowledged that these tests motivated them to study harder (M = 38.65), a pattern similar to findings by Messick (1989) emphasizing the positive washback effect of high-stakes testing. Nonetheless, when fairness is compromised, motivation may suffer, consistent with contemporary discussions around test anxiety and inequity.

Finally, for psychological stress (RQ4), the study shows how the pressure and consequences of these exams produce anxiety, stress, and health problems, consistent with prior literature on test-induced anxiety. This underscores the urgent need for admissions reforms, integrating multiple measures of ability to reduce negative impacts on student mental health and well-being. Furthermore, students perceived GPA as a more valid indicator of long-term performance, reinforcing the literature that views GPA as a better predictor of academic persistence and success (Geiser & Santelices, 2007). Together, these results suggest the need to balance entrance exams with other assessments to improve fairness and predictive accuracy.

Moreover, the findings of this study reveal significant concerns regarding the validity of university entrance examinations. Quantitative data showed a weak correlation between entrance exam scores and university GPA, supporting the argument that these exams may not effectively predict academic success. Additionally, participants across different regions expressed concerns over the fairness and accessibility of the exams, especially in under-resourced areas. Similarly, the qualitative data echoed these findings. Participants reported feelings of stress and anxiety linked to the high-stakes nature of the exams. Many felt that the exams did not reflect their true abilities or potential. Thematic analysis highlighted themes of systemic inequality, lack of preparation resources, and cultural bias embedded in test content.

These results suggest that current entrance exam systems may inadvertently reinforce social inequalities and fail to serve their intended purpose as equitable and reliable assessment tools.

Conclusion, Implications dan Limitation

Conclusion

This study has highlighted the various ways in which university entrance examinations can undermine the validity of higher education admissions. With evidence of limited predictive validity, perceived unfairness, and psychological strain, it becomes clear that relying solely on standardized testing is both inadequate and inequitable. Moving forward, it is crucial for educational stakeholders to develop more inclusive and comprehensive admission frameworks that recognize diverse forms of student achievement and potential.

While university entrance examinations offer a standardized method for evaluating candidates, they can pose significant threats to validity when used as the primary admissions criterion. To foster a fairer and more accurate admissions process, universities should consider broader measures that reflect the diverse skills and backgrounds of prospective students. Furthermore, given the rapid adoption of machine learning methods by education researchers, and the growing acknowledgment of their inherent risks, there is an urgent need for tailored methodological guidance on how to improve and evaluate the validity of inferences drawn from these methods.

To address these threats to validity, many educators and policymakers advocate for a more holistic admissions process. This might include a combination of high school GPA, letters of recommendation, personal statements, and interviews. Some institutions have moved toward test-optional policies, recognizing that a single exam may not fully capture a student's abilities or potential.

Implications

Also, the study's outcomes have several important implications:

1. For Policy Reform: There is a need for policymakers to reconsider the structure and function of entrance exams. Reforms should aim to create more holistic and equitable admissions criteria.
2. For Educational Equity: Institutions should address disparities in access to preparatory resources. Outreach programs and support systems could help bridge the gap for disadvantaged students.
3. For Alternative Assessments: Universities should consider integrating alternative metrics, such as high school GPA, portfolio assessments, and interviews, to complement or replace standardized tests.
4. For Mental Health Considerations: Given the reported psychological impact, there should be support mechanisms in place to assist students in managing exam-related stress.

Limitations

This study acknowledges several limitations that may affect the generalizability and comprehensiveness of its findings. First, the research was confined to Iranian participants who spoke Persian as their first language, which limits the cross-cultural applicability of the results to other educational systems and cultural contexts. Second, the cross-sectional design captured only a snapshot of student perceptions at one point in time, potentially missing longitudinal changes in attitudes and experiences related to university entrance examinations. Third, while the sample size of 400 participants was substantial, the stratified random sampling may not have fully captured the diversity of socioeconomic backgrounds and educational experiences across all regions of Iran. Fourth, the reliance on self-reported data through questionnaires and interviews may introduce social desirability bias, where participants might have provided responses they perceived as more acceptable rather than their true feelings. Additionally, the study's focus on student perceptions, while valuable, did not include perspectives from university admissions officers, policymakers, or educators, which could have provided a more comprehensive understanding of the validity threats in entrance examination systems. Finally, the academic records used to assess predictive validity were limited to GPA scores and entrance exam results, without considering other important indicators of academic success such as retention rates, degree completion times, or post-graduation outcomes.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*, 115–129.

Alderson, J. C., & Hamp-Lyons, L. (1996). TOEFL Preparation Courses: A Study of Washback. *Language Testing, 13*(3), 280-297.

AERA, A. NCME, American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education

(NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bachman, L. (1995). *Fundamental consideration in language testing*. Oxford: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Bailey, K. M. (1999). Washback in Language Testing. TOEFL Monograph Series. Report Number: RM-99-04. TOEFL-MS-15. Princeton. NJ: Educational Testing Service.

Bennett, R. E., Kane, M. T., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment*. Paper presented at invitational Research Symposium on Through Course Summative Assessment, Atlanta, GA.

Bird K. A., Castleman B. L., Mabel, Z., & Song Y. (2021). Bringing transparency to predictive analytics: A systematic comparison of predictive modeling methods in higher education. *AERA Open*, 7, 1–19

Boud, D., & Bearman, M. (2024). The assessment challenge of social and collaborative learning in higher education. *Educational Philosophy and Theory,* 56 (5), 459–468.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement.

In R. L. Brennan (Ed.), Educational measurement (4th ed.). Praeger.

Dornyei, Z. (2007). *Research methods in applied linguistics.* Oxford University Press. Thousand Oaks, California: Sage Publications, Inc.

Chapelle, C. A., & Voss, E. (2021). *Validity argument in language testing: Case studies of validation research*. Cambridge University Press.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*, 456–465.

Frederiksen, J., & Collins, A. (1989). A systems Approach to Educational Testing. *Educational Researcher. 18*(4), 22-32.

Fulcher, G. (2014). *Philosophy and language testing*. In A. Kunnan (Ed.), The companion to language assessment. John Wiley & Sons.

Gipps, C. V. (1995). *Beyond testing:* Toward a theory of educational assessment. The Falmer Press.

Haertel, E. H., & Herman, J. L. (2005). *A historical perspective on validity arguments for accountability testing.* Yearbook of the National Society for the Study of Education, Cambridge University Press.

Haladyna, T., Nolen, S. and Hass, N. (1991). Raising standardized achievement test score pollution. *Educational Researcher.20* (5), 2-7.

Hughes, A. (1989). *Testing for Language Teachers,* Cambridge: CUP

Hymes, D. H. (1972). *On communicative competence in Pride.* University of Chicago press.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–13.

Kelly S., Olney A. M., Donnelly P., Nystrand, M., D. & Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464.

Mackey, A. & Gass, S. M. (2005). *Second language research: methodology and design*. Lawrence Elbaum Associates.

Madaus, G. (1988). *The influence of Testing on the Curriculum in Tanner (ED) Critical Issues in Curriculum.* Yearbook of NSSE, part 1, Chicago, IL, University of Chicago press.

Maxwell, J. (2004). *Qualitative research design: An interactive approach*. Sage Publications Inc.

Moses, M. S., & Nanna, M. J. (2007). The testing culture and the persistence of high stakes testing reforms. *Education and Culture, 23*(1), 55–72.

Schimmack, U. (2021). The validation crisis in psychology. Meta-Psychology, 5, Article 1645. https://doi.org/10.15626/MP.2019.1645

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? *Language Testing,14*, 34–49.

Thomas, R. M. (2005). *High-stakes testing: Coping with collateral damage*. Lawrence Erlbaum Associates Publishers.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted? *Language Testing, 14*(2), 197–221.

Wall, D. (2005). *The Impact of high-stakes examinations on classroom teaching*: A case study using insights from testing and innovation theory. Cambridge: CUP.

Wohlin, M., Host, P., Runeson, M., Ohlsson, B., Regnell, & Wessl´en, A. (2000). *in software engineering*: an introduction. Kluwer Academic Publishers.

Ysseldyke, J., Nelson, J., Christenson, S., Johnson, D., Dennis, A., Triezenberg, H., & Hawes, M. (2004). What we know and need to know about the consequences of high- stakes testing for students with disabilities. *Council for Exceptional Children, 71*(1), 75–94.

Zheng Y., Nydick S., Huang S., Zhang S. (2024). MxML (exploring the relationship between measurement and machine learning): Current state of the field. *Educational Measurement: Issues and Practice*, 43(1), 19–38.