Predictor Validity as an Overview of Criterion Validity: A Literature Review

Ermawati[1, a)], Achmad Ridwan[2, b)] and Dinny Devi Triana [2, c)]

Author Affiliations

[1]*Graduate Program of Educational Research and Evaluation, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta Timur 13220, Indonesia.*
[2]*Fakultas Pascasarjana, Universitas Negeri Jakarta, Jl. Rawamangun Muka, Jakarta Timur 13220, Indonesia.*
Author Emails
[a)] ermawati2171@gmail.com
[b)] achmadridwan@unj.ac.id
[c)] dinnydevi@unj.ac.id

ABSTRACT

This study emphasizes the importance of validity in measurement, particularly predictive and criterion validity, to ensure accurate decision-making in educational selection and non-cognitive assessment. A literature review was conducted examining 1,013 articles from the Scopus, ScienceDirect, Google Scholar, and ERIC databases. After PRISMA screening, 34 relevant studies were identified. The analysis revealed that no single predictor consistently demonstrated the highest predictive and criterion validity. Cognitive tests such as the Scholastic Aptitude Test (SAT) or the Medical College Admission Test (MCAT) tended to be more accurate in predicting initial academic performance, while prior academic achievement and non-cognitive factors made significant additional contributions. In the context of health and psychosocial outcomes, physiological and behavioral indicators also demonstrated variable validity, with multi-predictor models consistently outperforming single-test approaches. In conclusion, implementing a multi-predictor approach ensures greater accuracy for selection and assessment, strengthening evidence-based practices in education, health, and workforce development. These findings contribute not only to theoretical advancements but also to impactful outcomes that align with the Sustainable Development Goals (SDGs), particularly SDG 4 (Quality Education), SDG 3 (Good Health and Well-Being), and SDG 8 (Decent Work and Economic Growth). Strengthening the multi-predictor approach supports the development of more inclusive, equitable, and impactful systems that benefit individuals and society.

Keywords: predictive validity, criterion validity, educational selection, scholastic, SDGs

1. INTRODUCTION

Validity is very important in a research. It is important because validity will result in the accuracy of the results and the right Decisions. The accuracy of the results shows that the results of the study using valid measuring instruments will be more accurate and reliable. Valid research results can be used to make better decisions, such as in the selection of new students, employee

A. D. Utami et al. (eds.), *Proceedings of the 3rd International Conference on Environmental Learning and Educational Technologies (3rd ICELET 2025)*, Advances in Social Science, Education and Humanities Research 1007,
https://doi.org/10.2991/978-2-38476-567-6_25

selection, development of training programs or policy making, and others. Better decisions will result in the right decisions and impact the Sustainable Development Goals (SDGs).

A test will be able to use the results if the extent to which a measuring instrument can measure what you want to measure (Azwar, 2022)[1] (Naga, 2022)[2] . One of the validities related to measurement and can be used to assess the quality of a measuring instrument is the validity of prediction and the validity of criteria. The relationship between prediction validity and criterion validity is that the two validities complement each other (Diamantopoulos et al., 2012)[3] (Siswanto, 2008) [4]. Prediction validity shows the ability of measuring instruments to predict the future, while criterion validity shows the relationship of measuring instruments with existing criteria.(Cunningham et al., 2019)[5]

One of the uses of prediction validity and the validity of criteria to measure is the student acceptance test. Student admission tests are needed to see the picture of a student's success at the advanced level. In addition to the success of preparation at the beginning of the school year, the estimated success of graduating on time from the two validity measurements is also estimated. The picture of student success is measured by test standards that are considered the same (Armstrong, 2000)[6] (Goldhaber & Özek, 2019)[7] which applies universally throughout the world. Test standards are certainly not exactly the same between countries, but the criteria are the same in the world, for example the standard Scholastic Aptitude Test (SAT) or American College Testing (ACT) test which is widely used in America in the process of accepting new students. SAT is one of the tests for college entrance that tests students' ability to understand and analyze reading texts, numerical questions, questions that test critical thinking logic and problem-solving (Liu et al., 2014) [8]. The ACT is a test that uses computers with tested material, such as academic potential tests (verbal, numerical, logic), English, competency tests in certain fields, and aptitude interest tests (Olszewski-Kubilius & Corwith, 2018)[9]. The test for student admissions in each country is indeed different but the estuary of the test leads to the aptitude test (TBS), SAT or ACT model, average school exam scores, rankings, and academic potential tests (TPA).

The university entrance selection test in each country has a different selection system. Deep futon (Asrijanty, 2014)[10] said that in general, the variation of the selection system comes from three things, namely 1) the quality of the candidate who is the selection criterion, whether the talent (Aptitude) or performance (Achievement); 2) Assessment references, whether based on criteria (Criterion-based assessment) or comparison in groups (Norm-based assessment); and 3) the context of the assessment, namely whether the basis for selection is the assessment carried out in secondary school (Internal Assessment) or assessment from outside school, external assessment.

The college entrance test as a pre-university measure uses various measures that are not the same in each country. One of them is academic achievement which is identified as the most significant predictor of university performance (Van Herpen et al., 2017) [11]. The entrance tests used are those that use scholastic aptitude tests and some use achievement tests as a measure. Some use three measures, namely high school average scores, general aptitude tests, and academic achievement acceptance tests (Vista & Alkhadim,2022) [12], some only use academic potential or achievement test scores with the SAT or ACT, or some are a combination of both. In Indonesia, the history of student admissions in universities starting from SKALU, Sipenmaru, UMPTN, SPMB, SNMPTN, SBMPTN, and SNBT has undergone several changes in the selection test. Selection for university entrance in Indonesia uses scholastic aptitude tests or academic potential as a selection tool after 2001 (Asrijanti, 2014). Previously, the selection test only used basic achievement and ability tests. The use of achievement tests and academic potential tests after the regulation of the Minister of National Education that regulates the authority of universities in the

university selection criteria. Previously, many public and private universities used scholastic aptitude tests and basic ability tests for independent university entrance exams.

Now in 2025 the Ministry of Primary and Secondary Education (Kemendikdasmen, 2025)[13] uses the Academic Ability Test (TKA) which aims to provide a fairer assessment for all students, considering that the standards in each school are different. TKA is a national standard assessment designed to measure students' academic achievement in certain subjects in accordance with the applicable curriculum. TKA is needed to enter university achievement pathways (SNBP) and other needs to continue school. TKA is not mandatory and does not determine graduation, intended for those who feel ready and need it to complete the current assessment system and does not replace assessment by educational units.

The use of predictive validity and criterion validity is not only for student admission tests, but also to determine certain ability tests, such as reading and other mental abilities that are not related to academics, such as physical health and mental health, fitness, physical endurance, employee acceptance or employee performance, and so on. This is also related to the Sustainable Development Goals (SDGs), especially SDG 3 and SDG 8. Non-academic tests are also carried out to see the picture of future predictions of the test results associated with the specified criteria. The researcher divided the literature discussion into three parts. The first part or table 1 is about Overview of prediction validity and criterion validity based on college entrance exams. The second part of the college entrance exam is based on tests but the test is different from the first part (table 1), and finally table 3 contains an overview of the validity of the predictions and criteria of mental and action.

The study of the predictive validity and validity of criteria in university entrance tests and non-cognitive instruments has direct implications for the achievement of the Sustainable Development Goals or SDGs. By proving the reliability of the selection instruments, this study supports SDG 4 (Quality Education) which emphasizes the importance of access to equitable, inclusive, and quality-based higher education. The selection process using valid measurement tools contributes to increasing the accuracy of predicting students' academic success, while helping to reduce the dropout rate and improve the efficiency of the education system (UNESCO, 2020) [14]. The student selection system and the estimated success in education as a process of realizing quality and successful education in accordance with SDG 4.

Furthermore, in table 3 the focus on non-cognitive instruments that include aspects of mental health, physical endurance, and work behavior also strengthens the contribution to SDG 3 (Good Health and Well-being) and SDG 8 (Decent Work and Economic Growth). The validity of the instrument in this domain allows for early detection of health risk factors, so that preventive interventions can be designed more effectively (WHO, 2021) [15]. In addition, in the context of employment, proven instruments support fair, transparent, and productive human resource recruitment and development practices, in line with the vision of decent work and sustainable economic growth (ILO, 2019) [16]. Thus, this research not only makes a theoretical contribution to the measurement literature, but also supports the global agenda of sustainable development through improving the quality of education, health, and employment.

2. LITERATURE REVIEW

This section will discuss the definition and explanation of predictive validity and criterion validity. How the relationship between predictive validity and criterion validity will be briefly explained.

Definition of Predictive validity and Validity of criteria

Cronbach & Meehl (1955) say that predictive validity refers to the extent to which scores on an exam or measurement scale predict achievement on a particular criterion factor. (Rothstein, 2004)[17] ; (Agostino & Powers, 2009)[18] , (Atkinson & Geiser, 2009)[19] according to Fulcher and Davidson in (Razmi et al., 2021)[20] Predictive validity is a term used when test scores are used to predict several future criteria, such as academic success, endurance success, health, adaptability, mental resilience, response speed, etc. Predictive validity is a subcategory of criteria-oriented validation criteria that evaluates the relationship between a particular test and the criteria we want to predict (Messick, 1990)[21] ,(Mishler, 1990)[22] The validity of a prediction predicts the future or the validity of a prediction. From the results of the prediction validity score, it can be associated with other validities and can estimate or predict future results.

Dali S. Naga in his book said that the measurement variable is designed in relation to other variables. Measurement variables generate predictors, while other variables generate criteria. There is a correlation between predictors and criteria. For example, the new student admission exam, the purpose of the exam is to obtain students who will get high learning outcomes in the future. Conceptually, there is a connection between the results of the student admission exam (predictor) and the learning outcomes later (criteria) if there is a connection, then the new student admission exam is valid, if there is no connection, then the new student admission exam is invalid. Another example of a new employee admission exam is the purpose of the exam to obtain employees who perform well in the future. Conceptually, there is a relationship between the results of the employee admission exam (predictor) and their future performance (criteria). If there is a connection, the new employee admission exam is valid. Meanwhile, if there is no connection, the new employee admission exam is invalid (Naga, 2022) [2].

Criterion validity is the match between the predictor (predictor score) and the criterion (criterion score). The validity of the criteria is aimed at whether or not the predictor (predictor score) is good. If the validity of the criteria is good, then the predictor measuring tool (predictor score) can be used for various similar purposes. The validity of the criteria is to associate the measuring instrument with other measuring instruments as a criterion, whether the measuring instrument can be explained by the results of its correlation with the criteria based on the existing theory (Flynn et al., 1994)[23]; (Kimberlin & Winterstein, 2008) [24]. The validity of the criteria will be a prediction picture if the predictive results of the test are significant with the results of the test being compared.

There are limitations to the validity of the criteria related to the calculation of the validity coefficient. The criterion validity coefficient is the linear correlation coefficient between the predictor sector (X) and the criterion score (Y). There are several limitations in calculating the validity coefficient, including the clarity of the criteria data (Naga, Dali S., 2022). If the predictor score is clear, for example the results of the selection of new students or the selection of new employees, then often the score of criteria is not always certain, for example in the admission of new employees, what is the score of the criteria. So it requires clear criteria and clear measurement results. According to Powers (2004) Investigations of validity are particularly susceptible to inconsistencies due to reliance on limited sample sizes, unreasonable criteria setting, excessive generalizations based on limited sources of unreliable information, ignoring the effects of compensatory selection, and lack of consideration for the estimated effects of range limitation in predictive measures and criteria (Razmil et al., 2021) [20]. So the clarity of the criteria determines whether the validity of the prediction can describe the validity of the criteria.

A test can be a snapshot of another test and the results are close to the real thing if errors, biases, etc. that make an inaccurate result are minimized (C.R. Reynolds, et al, 2021)[25]. The picture of predictive validity to the validity of the criteria will be in line and can be predicted if the test work is pure and can measure what must be measured. The picture of test results can be wrong to estimate the future if test takers in filling out the test think that the answers are not the result of thinking. Likewise, the picture of the future can be wrong if the measurement is not correct, for example, if you want to measure the understanding and reasoning of logic given by the History test, this is certainly not correct. Because it is prone to errors, biases and others, the determination of prediction tests and criteria tests from the beginning has been confirmed and tested.

Hubert, Cohen, & Staub (2022)[26] state a conceptual argument about how predictive validity can be hidden or misinterpreted due to the effects of compensatory selection. Predictors, according to Hubert et al., are standardized test scores and other predictors in the application (e.g. school grades, letters of recommendation, personal statements) are used compensatory in selection. Meanwhile, the criteria are academic success (degree completion) in their simulation. In the article it was stated that failure to find a positive correlation between test scores and success among admitted college students does not automatically mean that the test is invalid as a predictor. Conversely, the absence of a neutral correlation or correlation can be consistent with predictive validity if the selection uses a mechanism between predictors. So the aspect of predictive validity according to Hubert et al. From the results of the simulation used, a valid standard test can lose correlation among accepted students due to the compensatory effect of the selection. A test is considered a valid criterion if the test score is related to success in the simulation, although this relationship can be masked by compensatory selection. The absence of a positive correlation among accepted students is not proof that the test is invalid, even if the correlation is zero, it could be that the test remains valid in the full applicant population.

Alarcón-Bustamante, Eduardo. et al (2025)[27] improved the way to assess predictive validity When the criteria (GPA) are not fully observed, by providing valid identification boundaries under transparent selection beliefs or assumptions. This can be determined as proof of the validity of the (predictive) criteria, but in a bounded/sensitivity-analysis format it is not a single point estimate. So to assess predictions related to criteria that are not always visible or fully observed, Eduardo et al (2025)[27] provide a solution by providing identification limits with sensitivity analysis so that the assessment is not only one way.

3. METHODOLOGY

This study is an analytical review of the literature review on the predictive validity and validity of the criteria of each article. The researcher looks for predictive validity, the validity of the criteria, and looks at the results. This research was conducted to answer the research question:

a. Is the article reviewed related and significant to the validity of the prediction with the validity of the criteria?
b. Which predictor gives the closest picture to the predetermined criteria?
c. Which predictors give a high and significant prediction value to the criteria?

The process of searching for articles goes through the following stages:

1. Stage 1 searches Scopus, google scholar, and Eric with a search for the keywords Predictive validity, Criterion validity, Scholasticism, Educational Selection, admission test, SDGs with a

combination of adding the word and and or if one of the keywords is removed if it is not found or if there is a search error.

2. The initial search was not limited to the type of publication and year, google scholar as many as 868 documents scopus 91 documents and ERIC 54 documents.
3. Searches by title that correspond to the relevance of prediction validity and criterion validity from 1013 articles to 959 were discarded 54 articles.
4. Reading the appropriate titles and keywords also sees similarities and there are no duplicate sources from Scopus, google scholar, and Eric which are only based on the source of the article (books and other sources are not included) so it becomes 456 articles, which 503 documents are discarded
5. Viewing and reading titles, keywords, and abstracts is taken into 200 articles. Articles Deleted 256 articles
6. Re-reading the abstract and connecting with the research objectives, then it becomes 150 articles that are discarded 50 articles.
7. Reading the title and abstract carefully that is appropriate to see the validity of the prediction and the validity of the criteria (qualitative), then it becomes 134 articles taken, articles that are discarded 16 articles.
8. The search is only based on English articles, easy to access and can be searched for predictions and criteria (quantitative) so that it becomes 86 articles that 52 articles are discarded.
9. Manual search based on title, abstract, and overall content of the article that fits the purpose to 34 articles.

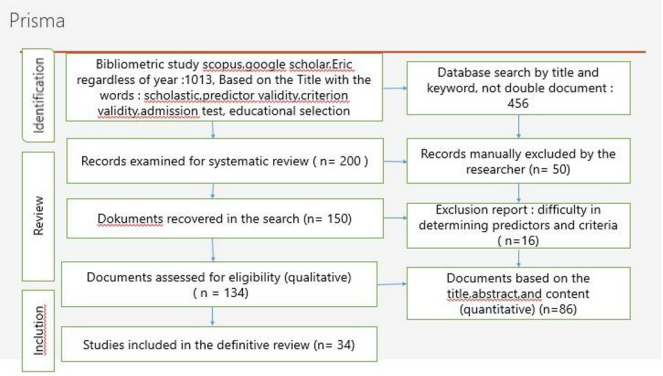

FIGURE 1. Investigation Process

4. RESULT

The researcher limited the research to the prediction of the university entrance exam, the prediction of achievement exam scores with the cumulative score of the school, compared to the value of the specified criteria (tables 1 and 2). Then in table 3 is the validity of the prediction and the validity of the criteria of the test outside the college and school entrance exam. So the division of the table:

1. Table 1 contains the prediction validity and validity of the criteria of the article on the college entrance exam with research limitations Scholastic Aptitude Test (TBS), UN, achievement test, Academic Potential Test (TPA), Basic ability test, school average score, school ranking, and college performance or GPA
2. Table 2 contains the predictive validity and criterion validity of the article on college entrance exams with the research limitation that the test used for college entrance is different from table 1
3. Table 3 contains the predictive validity and criterion validity of test articles outside of college and school entrance exams .

TABLE 1. The Prediction Validity and Validity of the Criteria of The Article

No	Title,Author, Year,Journal,Country	Validity Predictions	Validity Criterion	Result
1.	Assessing The Predictive validity of pre-admission criteria on dental students Academic Performance : a cross-sectional study,Almalki,Sultan A. et al., [28] 2024,BMC Oral Health,Saudi Arabia	High school grades, achievement and aptitude test scores (entrance exams) and demographic/psychosocial factors (age, marital status, accommodation, suspension history, English proficiency)	Current GPA-Dental Students	Pearson's correlation and regression showed a meaningful positive relationship between high-school grades, achievement, and aptitude tests and GPA; age is negatively correlated; a number of other covariates are also significant. This shows that pre-entry indicators have predictive power on actual academic performance (predictive validity as well as criterion-related).
2.	Predictive validity of Admission tests and Educational	TMS and HAM-Nat entrance exam scores and high-school GPA.	Pre-clinical academic performance (preclinical	TMS, HAM-Nat, and GPA were significantly correlated with pre-clinical performance (correlation range ~0.23–0.48 for entrance tests; 0.19–0.35

	Attainment on Preclinical Academic Performance-a Multisite Study, author: Jaehn, Malvin. et al. ,2025,[29]BMC Medical Education Journal,German		GPA/PCGPA) and M1 national exam scores	for GPA). This indicates criterion-related validity because the predictor relates to real criteria. Incremental validity: both entrance tests provide additional predictive power above the GPA, while the additional contribution of GPA after controlling for test scores is small. The author's conclusion: both the entrance test and GPA are worthy of being used as selection criteria; however, amid concerns over GPA, entrance tests are emerging as a more predictively robust alternative.
3	Predictive Validity of Scholastic Talent and Learning Achievement as Criteria for Higher Education Entrance Selection. Author : Asrijanti, 2014 Journal : Journal of Education and Culture, Indonesia	evaluate the predictive validity of the Scholastic Aptitude Test (TBS) and National Examination (UN) scores in predicting academic success in higher education	comparing the results of FFB and UN scores with the Cumulative Grade Point Average (GPA) in the first four semesters of higher education	The results showed that the TBS subtests (verbal, quantitative, and reasoning) had significant predictive validity for the Economics major, while the UN score was insignificant.For Engineering majors, verbal and quantitative TBS subtests and UN scores in English and Mathematics are significant predictors. The results showed that the TBS subtest and some UN scores had a significant correlation with GPA, indicating the validity of the criteria
4	The Quality of the Academic Potential (TPA) 07 A Author : [30],Journal of Research and	The TPA test was analyzed based on the answer data of 80 students accepted into the UGM Psychology Master's Program in 2007.	The validity of the TPA 07A criteria was tested using the first semester GPA as a criterion.	The results of the study showed that TPA scores were less able to predict the Cumulative Grade Point Average (GPA) in the first semester significantly. The correlation between TPA scores and first-semester GPA was very low, suggesting that TPA 07A did not have satisfactory predictive validity for predicting academic success in the first semester The first-semester GPA was very homogeneous and not varied enough as a criterion, resulting in low predictive validity estimates. Thus, the

	Evaluation,Indonesia			validity of the TPA 07A criteria is also considered unsatisfactory because it cannot accurately predict academic success based on the first-semester GPA
5	Predictions of scholastic aptitude tests on the learning achievement of high school students. Author: ([31],Indonesian Journal of Educational Assessment,Indonesia	Scholastic Aptitude Test (TBS) scores with students' academic achievement in high school as a prediction overview	The validity of the FFB criteria was tested using the semester GPA score in high school as a criterion.	The results of the study show that the FFB prediction model is in accordance with empirical data. The variant explained by the potential of scholastic talent on learning achievement is 14%. This means that TBS can predict 14% of the variance in students' academic achievement in high school. FFB tends to be a better predictor for social science (IPS) programs compared to natural sciences (IPA). In the social studies department, the contribution of scholastic talent to achievement is 37%, while in the science department it is only 7% TBS has a fairly good validity criterion in predicting students' academic achievement in high school, especially in the social studies department. The benchmarking process results in four benches on each TBS subtest, each of which has different potential characteristics. These characteristics help in describing students' abilities based on FFB results
6.	Aptitude Tests and Successful College Students: The Predictive Validity of the General Aptitude Test (GAT) in Saudi Arabia. Author :	General aptitude Test (GAT), National Achievement Test (NAT), High School Grade Point Average (HSGPA), all three tests are used to predict students'	Correlation between GAT and cumulative GPA, correlation between NAT and GAT, and correlation between HSGPA and cumulative GPA at university.	1. Predictive of Academic Success: The combination of high school grade point average (HSGPA) and the National Achievement Test (NAT) is the best predictor for academic success in university, measured by a cumulative GPA or a freshman GPA. However, GAT is the best predictor of graduation as a criterion for success. **2.** Correlations: The correlation between GAT and cumulative GPA is 0.33, which means that about 9% of the variation in GPA can be explained by GAT. The highest correlation was between NAT and GAT

	[32]International Education Studies, Saudi Arabia	academic success at universities as measured by cumulative GPA, freshman GPA, and graduation status.		(R=0.547), which showed that 29% of the variation in GAT was explained by variation in NAT. Overall, the findings of this study support the importance of the combination of HSGPA and NAT in predicting academic success at university, while GAT is the best predictor for graduation. The multipredictor model (HSGPA + NAT -/+ GAT) is more accurate than a single test
7.	The Relationship between Cumulative Grade Point Average (GPA) and National Final Exam (UAN) Scores: A Case Study at FMIPA Unsyiah,[33],Journal of Statistics,Indonesia	National final exam scores predict students' cumulative grade point average (GPA)	National exam scores are positively correlated with the GPA of students at FMIPA Unsyiah, although limited. GPA as a criterion.	The results of the study showed that the UAN score was positively correlated with the GPA of students at FMIPA Unsyiah. However, when all UAN factors were used simultaneously to see their relationship with GPA, the resulting determination factor was quite small, ranging from 29.9% to 58.1%. This suggests that UAN scores have limited predictive validity in predicting a student's GPA. The most influential factor on the GPA of FMIPA Unsyiah students is the Indonesian language score and district origin. This shows that Indonesian and district origin scores have a stronger validity criteria compared to other UAN factors in predicting student GPA.
8.	Pre-university Measures and University Performance Author: Alvin Vista, Ghadah S, Alkhadim, 2022, Journal of Frontiers in Education, Saudi Arabia.	HSGPA (High School GPA), GAT (Verbal & Quantitative), SAAT, Latent construct (combined pre university measures). Pre-university measures (high school averages, General Aptitude Tests, and academic	Average GPA, prep period GPA, university performance, graduation status. Pre-university sizes are valid for use in various subgroups of students. The scores generated from this measure make a	1. The results of the Friedman test showed no significant difference: $\Delta\chi 2(df = 3) = 0.353$, $p = 0.9498$. This suggests that these rankings do not differ from each other, and that universities that rank high in their students' high school GPA also tend to rank high in other pre-university measures. 2.These overall results look at the rankings of the entire size of pre-universities for 23 universities. To check if the same applies to measurements that are based only on standardized tests (i.e., GAT-Q, GAT-V, and SAAT), a follow-up analysis is performed focusing on the ratings for all three measurements. The result was again insignificant,

		merit acceptance tests) are consistent in predicting university performance. This includes performance in the preparatory year and whether the student remains regular and eventually completes university studies.	statistically significant contribution to the prediction of bachelor's degree completion	$\Delta\chi 2(df = 2) = 0.609$, $p = 0.7376$. This shows that universities are consistently ranked based on how they attract students based on their standard results**.** **3.** The test that has the most effect on the performance of the university is SAAT. HSGPA is one of the important indicators of latent pre-university performance, GAT-V also contributes significantly to the variance of pre-university constructs. Predictive Consistency: All pre-university measures (high school average scores, General Aptitude Tests, and academic merit acceptance tests) are consistent in predicting university performance. This supports previous results that pre-university measure performance predicts final performance at university, including performance in the preparatory year and whether students remain regular and eventually complete university studies Influence on University Performance: The size of pre-university has a significant influence on university performance. Multilevel structural equation modeling analysis showed that pre-university performance affects the performance of the university's preparatory year, and this effect remains significant throughout the university's journey until students graduate
9.	The Predictive Validity of Entrance Scores and Short-term Performance for Long-term Success in Engineering Education. Author	1. High school coursework score 2.General Ability Test Achievement Test 3.Preparatory year GPA score/student's initial performance at university (Year-1, Year-2, Year-3 GPA).	Cumulative GPA at the end of the program / Graduation GPA in an engineering program The correlation between the entrance score and the	Predicted Academic Performance: Admission scores (high school grades, General Ability Test, and Achievement Test) have good predictive validity for the preparatory year GPA, but not for the cumulative GPA. First- to third-year GPA is a better predictor of graduation GPA. Female Student Performance: Female students perform better compared to male students in entrance scores and GPA. Therefore, female students tend to have higher cumulative GPAs.

	: [34],Journal of Applied Research in Higher Education, Saudi Arabia.	Admission scores (high school grades, General Ability Test, and Achievement Test) have a significant correlation with the GPA of the preparatory year, but not with the cumulative GPA. The correlation was also significant between the GPA of the first to third year and the GPA of graduation. The GPA of the adjacent years is a better predictor of future GPA. GPA at the time of graduation can be well predicted by GPA during the years of study in an engineering program after controlling for entrance scores	preparatory year GPA shows that the entrance score can predict the preparatory year GPA before the student enters the engineering program. However, the correlation between entry scores and GPAs of subsequent years during engineering studies tends to be lower. A stronger correlation was found when linking a particular GPA with an adjacent previous year's GPA.	Practical Implications: These findings have practical implications that are useful for admission and monitoring of student progress in engineering education programs. These results can help develop program curriculum and committees in designing admission criteria. Overall, the findings of this study support the importance of academic performance during the study in predicting long-term success in engineering education
10.	Sats, Achievement Tests, and High-School Class Rank As Predictors of	SAT scores (Scholastic Aptitude Test) – verbal & math. Achievement Tests – achievement tests in a	College GPA (cumulative GPA, especially the first year and beyond).	1. High School Class Ranking (CLR): CLR has higher predictive validity compared to the SAT and achievement test (ACH) in predicting college performance. CLR significantly adds to the overall prediction of student academic performance.

	College Performance Author : [35],Educational and Psychological Measurement,Pennsylvania	specific academic field. High School Class Rank (HSCR) – the ranking of students in the class/secondary school. - This study found that high school class rankings (CLR) and achievement tests (ACH) had higher predictive validity compared to the SAT in predicting college performance. CLR and ACH significantly add to the overall prediction, while the SAT does not. - The SAT does add predictive scores in some individual subjects, but its contribution is not as large as the CLR and ACH	Indicators of academic success in higher education. CLR is the only variable that significantly increases the prediction of dropout rates or the number of courses not completed. This suggests that CLR has a stronger criterion validity in predicting long-term academic success compared to the SAT and ACH.	2. Achievement Test (ACH): ACH also has higher predictive validity compared to the SAT in predicting college performance. ACH significantly adds to the overall prediction of student academic performance. 3.SAT: Although the SAT adds to score prediction in some individual subjects, its contribution is not as large as the CLR and ACH. The SAT does not significantly improve the overall prediction of a student's academic performance. 4. Validity of Criteria: CLR is the only variable that significantly increases the prediction of dropout rates or the number of courses not completed. This suggests that CLR has a stronger criterion validity in predicting long-term academic success compared to the SAT and ACH. 5.Overall, the results show that high school grade rankings and achievement tests are better predictors of college performance compared to the SAT.
11.	Group Data on High School Grade Point Averages and	1.High School GPA (HSGPA) — high school average grade.	The correlation between group percentage and criteria (pass rate) was much	1. Graduation Prediction: Persistence to graduation as a characteristic of the student population in college can be predicted much better at the group level compared to the individual level within the college. The

	Scores on Academic Aptitude Tests as Predictors of Institutional Graduation Rates" by [36],Educational and Psychological Measurement, US	2. Academic Aptitude Test Scores — academic aptitude test scores (similar to the SAT/ACT, depending on the context). The results showed that SAT I Verbal and Math scores as well as ACT scores in the 25th and 75th percentiles were good predictors for the percentage of students who graduated from the institution. The correlation between this score and the graduation rate ranges from 0.62 to 0.73 The percentage of freshmen who have a high school GPA of at least 3.00 is also a good predictor of graduation rates, with a correlation of 0.49	higher compared to the predictive validity coefficients of SAT I and ACT scores for individual graduation reported in the literature Regression analysis showed that ACT scores in the 25th percentile and SAT I Math scores in the 25th percentile made the greatest contribution to predicting the percentage of students who graduated within 6 years of entering college	greatest contribution to high multiple correlations at the group level is given by ACT scores in the 25th percentile within colleges. 2. SAT I and ACT scores: A high SAT I Math score in the 25th percentile and a high SAT I Verbal score help with multiple predictions, but a high SAT I Math score does not. A low Verbal SAT I score or high high school GPA (3.00 or higher) also does not contribute significantly to graduation prediction 3. Critical Variables: A critical variable in student persistence at the college level is an average score in the 25th percentile on the SAT I Math and/or ACT. Colleges with student populations that have low scores on the SAT I Math or ACT in the 25th percentile tend to have higher dropout rates. 4. Quantitative Ability: A score in the 25th percentile on SAT I Math is a slightly better predictor for persistence compared to a score in the 75th percentile. This may be due to the fact that most college students do not take majors that require very high quantitative ability. Overall, the results show that SAT I and ACT scores and high school GPA are good predictors of an institution's graduation rate, with scores in the 25th percentile making the greatest contribution to predicting graduation
12.	The Validity of Tradisonal Cognitive	the predictive validity of various tests such as CTBS, HSGPA,	The validity of the criteria in this article was evaluated by	The results showed that the CTBS reading vocabulary test had a high validity criterion as a predictor of academic success.

	Measures and of Scales of The study Attitudes and Methods Survey in The Prediction of The Academic Success of Educational Opportunity Program Students. Author : [37],Journal of Educational and Psychological Measurement, Los Angeles	and entrance exam scores (ACT/SAT) in predicting the academic success of EOP students. Traditional cognitive tests SAT-Verbal SAT-Mathematics High School GPA (HSGPA) Non-cognitive (Study Attitudes and Methods Survey / SAMS) Scales: Study Habits, Motivation, Attitudes, and other learning strategies.	comparing test results (CTBS, HSGPA, entrance exam scores) with the academic performance of EOP students. Freshman GPA (first-year GPA) Educational *Opportunity Program (EOP) student* at the University of Los Angeles. GPA is considered a key indicator of "academic success."	The SAMS scale makes a significant contribution to some regression analyses and the SAMS scale shows potential as an additional indicator. Best Predictor: Reading vocabulary measurements are the best predictors for EOP students' success in their first year of college after a summer prep program . SAMS Validity: Two SAMS (Academic Drive-Conformity and Alienation Toward Authority) scales show potential as additional indicators for the academic success of minority students. Influence of Preparatory Programs: Summer prep programs may have influenced the validity of predictor coefficients, especially for students who are weakest in the attributes measured.
13.	Predictability of Engineering Students' Performance at the University of Engineering and Technology, Peshawar from Admission Test Conducted by Educational Testing and Evaluation	Using three tests, namely FSc, Entry Test Scores, and overall Merit as predictors to measure the academic performance of Engineering students at the University of Engineering and Technology (UET), Peshawar	Students' academic performance at engineering colleges (Engineering program performance) in the form of *CGPA* or cumulative academic scores during studies. Overall, these findings suggest that ETEA entrance tests have good predictive validity for	**1.** Significant Relationship: There is a significant relationship between FSc scores (equivalent to high school grades), ETEA entrance test scores, and overall merit with the academic achievement of engineering students at UET Peshawar. However, this relationship is stronger for male students compared to female students. **2.** Gender Differences: The results of the study show that there is a negative relationship between predictor scores and academic achievement for female students and some engineering disciplines. This is due to the smaller sample size for female college students. **3.** Predictive Validity: ETEA entrance tests have good predictive validity for the academic performance of engineering students, but there is

	Agency (ETEA), NWFP, Pakistan.Author : [38],Journal of Social and Behavioral Sciences, Pakistan		engineering students' academic performance, but there is a need to develop additional predictors, such as interviews or non-cognitive domain tests, to improve the accuracy of admissions decisions	a need to develop additional predictors, such as interviews or non-cognitive domain tests, to improve the accuracy of admissions decisions. 4. Overall, the findings of this study support the importance of the ETEA entrance test in predicting the academic performance of engineering students at UET Peshawar, although there are some differences based on gender and engineering disciplines
14.	The Effect of National Exam Scores on the Cumulative Achievement Index of FMIPA Students at Cendrawasih University Author : [39],Electronic Journal of Cendrawasih University, Papua,Indonesia	UN score has a significant influence on the GPA of FMIPA Uncen students	Although UN scores have a significant influence on GPA, the low coefficient of determination suggests that there are still many other factors that affect a student's GPA. Thus, UN scores are not the only factor that determines the academic success of students at FMIPA Uncen.	Positive Correlations: There is a significant positive correlation between the scores of the National Examination (UN) and the Cumulative Achievement Index (GPA) of FMIPA Uncen students. This means that UN scores can affect a student's GPA, although the effect is not very great. Coefficient of Determination: The coefficient of determination of 9.048% indicates that the UN score can only explain about 9.048% of the variance in the student's GPA. This means that there are still many other factors that affect a student's GPA that are not explained by the UN score. Influence of Entry Path: The analysis shows that student entry pathways (SLSB, SNMPTN, SBMPTN) also have an influence on the relationship between UN scores and GPA. The SNMPTN entrance pathway shows significant results, which means that UN scores can be used as a better predictor for students entering through this pathway. Regression Model: The regression model obtained from the analysis shows that the UN value has a significant influence on GPA, but the effect is relatively small. The resulting regression model was GPA = -

				0.969 + 0.418UN, which showed that every one-unit increase in the UN value would increase the GPA by 0.418. Overall, the results of the analysis show that UN scores have an influence on students' GPA, but the influence is not dominant and there are still many other factors that need to be considered in predicting students' academic success.
15.	Predictive Validity of Admission Criteria in Predicting Academic Performance of Medical Students: A Retrospective Cohort Study. By [40],Frontiers in Medicine,Bahrain	1.Science test scores (science test of admission criteria) 2.High School GPA (HSGPA) 3. English test scores. there was a statistically significant positive correlation between the admission criteria and academic performance of medical students in Year 1, Year 4, the B.Sc. (Bachelor of Medical Science) exam, and the MD (Medical Doctor) exam during the 6 years of the medical program.	1. Academic grades in Year 1 (Year 1 GPA) 2. Grade in Year 4 (Year 4 GPA). 3. Exam Score B.Sc. (Bachelor of Science) 4. Final exam score MD exam (final medical exam). Not all admission criteria used for student selection are good indicators of, for example, the English language test does not show a significant correlation with the average score of the MD exam for students who enroll directly in Year 1.	Significant Positive Correlations: There is a statistically significant positive correlation between admission criteria and academic performance of medical students in Year 1, Year 4, the B.Sc. (Bachelor of Medical Sciences) exam, and the MD (Medical Doctor) exam during the 6 years of the medical program. However, the English Test did not show a significant correlation with the average MD exam score for students who enrolled directly in Year 1. Primary Student Performance: Results related to students admitted to the Primary Programme show that there is no significant correlation between their Secondary School Grade Point Average (HSGPA) and their academic performance in Year 1, Year 4, B.Sc., and MD exams. Predictor variables: All predictive variables correlated significantly with all outcome variables (academic performance). Science test scores show a greater contribution in predicting academic performance compared to other predictors. Science test scores showed 27.7%, 15.0%, 19.7%, and 12.6% variations in student performance in Year 1, Year 4, B.Sc., and MD exams, respectively. Regression Analysis: Regression analysis shows that there is a statistically significant linear relationship between student performance and admission

		Science test scores show a greater contribution in predicting academic performance compared to other predictors. Science test scores show 27.7%, 15.0%, 19.7%, and 12.6% variations in student performance in Year 1, Year 4, B.Sc., and MD exams.	Science test scores were found to be more predictive of academic performance compared to other predictors.	scores. Science test scores were found to be more predictive of academic performance compared to other predictors.
16.	Quality Improvement: A Practical Nursing Program's Admission Test. By [41],Journal of Nurse Education Today, Canada	The results showed that biology entrance test scores were a significant predictor for student success in the Canadian Practical Nurse Registration Examination (CPNRE). Biology entrance test scores had a significant positive correlation with CPNRE results. Chemistry, English, and math entrance test scores, as well as the	Academic factors such as program entry GPAs, English entrance test scores, math, biology, and chemistry are just a few of the characteristics to consider when determining the predictability of a practical nursing student's success. Therefore, it is important for the program admission process to identify and	**1.** Academic Predictors: Of the four components required on the program entrance test, three components (biology, chemistry, and English) are statistically and clinically significant for passing on the Canadian Practical Nurse Registration Examination (CPNRE). Biology entrance test scores are the strongest predictors for student success in CPNRE. **2.** Biology Entrance Test Scores: The average score of the biology entrance test (M = 74.96) is a predictor of student success in CPNRE. Chemistry entrance test scores and English language entrance test results also showed a positive association with CPNRE passing. **3.** Program Entry GPA: The analysis shows that a higher program entry GPA correlates with a higher graduation rate on the CPNRE. Students who pass CPNRE have a higher program GPA compared to those who fail. **4.** Explained Variance: The regression model shows that 9% of the variance in CPNRE results can be explained by biology,

		program's grade point average (GPA) had no significant effect beyond the model's predictors. This model shows that 9% of the variance is explainable, and 91% is unexplained.	measure non-academic entry criteria, as academic criteria have limited predictability.	chemistry, English, and program entry test scores. However, 91% of the variance is not explained by this model, suggesting that non-academic factors may also play an important role in student success.

Table 1 Discussion

Table 1 summarizes the evidence on the predictive validity and criterion validity of various pre-university measures—ranging from high school grades, grade ranks, merit/aptitude tests, to a combination of both—with academic performance outputs in college such as first-year GPA, cumulative GPA, retention, and graduation. A common pattern that immediately appears is the absence of a single indicator that excels in all contexts; In fact, the multipredictor model is consistently more stable than relying on a single test. Report card/school scores (HSGPA) and academic achievement-based indicators often provide a strong basis for prediction, while general aptitude tests add value primarily to project initial achievement after admission.

In many studies, HSGPA and class rankings show a solid association with GPA and even with graduation. These findings make substantive sense: The HSGPA aggregates cross-subject performance, long study periods, and repetitive assessment practices—thus containing information about study habits, persistence, and performance consistency. However, the superiority of the HSGPA is not a reason to deny standardized tests. In some contexts (e.g., health and engineering programs), entrance test scores—both achievement and general ability—contribute additionally above the HSGPA, especially when predicting a prep year or first-year GPA that is strongly influenced by specific cognitive readiness.

Another recurring trend is the decline in predictive power as the timing of the outcome gets farther away. Entrance scores are informative enough to project initial performance, but sequential annual GPAs tend to be the best predictors of graduation GPA. This implies that good selection needs to be followed by careful management of study progress: early indicators help "capture the starting point", while GPA trajectories from year to year capture the dynamics of learning, adaptation, and academic support that students experience.

The results also showed that the quality of the criteria affected the validity estimate. An overly homogeneous GPA results in a low coefficient, so it is best to use a more varied measure (retention, graduation time, or annual GPA). Differences across study programs, gender, and entry paths also need to be considered because they moderate the power of prediction.

Practically, Table 1 supports a selection strategy based on an indicator package: a combination of HSGPA/class ranking with a minimum of one achievement test and one general ability test. To map initial readiness, weight can be given to standard tests, while for long-term predictions, monitoring the progress of studies becomes more important.

Finally, there are two prerequisites for selection to be more fair and accurate: (1) the right criteria—sufficiently varied and relevant to the program, and (2) cross-group fairness audits—so that selection decisions can be contextually adapted.

Analysis Division Table 1

1. General View
   Based on table 1, it shows that there is a positive correlation between college entrance test scores (predictors) and student academic performance (criteria). The measurement is through the initial semester GPA, cumulative GPA, on-time graduation, or other academic achievements. The correlation rate varied from low ($r = 0.10$-0.30) to moderate ($r = 0.40$-0.60) while some cases reached high at the group level ($r > 0.70$ as in the Stumpf & Stanley study, 2002). cases of low predictive validity (such as TPA 07A-Azwar, 2008) due to the homogeneity of the criteria scores, so that the variation of the data is small and the correlation coefficient is low.
2. Country Context and Differences in Selection System

In Indonesia, FFB and UN (before 2023) were tested jointly whose results showed that the contribution of FFB tends to be more consistent than the UN, especially in non-science departments. In Saudi Arabia and other Middle Eastern countries (Alnahdi, 2015; Abdelfattah, 2022; Almalki, Sultan A. et al., 2024), a combination of several Pre-university indicators (HSGPA, GAT, NAT, SAAT) provides a more stable prediction of university performance. In the United States, HSGPA and Class Rank are more powerful predictors of graduation than the SAT alone (Baron & Norman, 1992; Stumpf & Stanley, 2002). In Germany, both entrance tests and GPAs are eligible as selection criteria; however, amid concerns over GPA, entrance tests emerged as a more predictively robust alternative (Jaehn, Malvin. et al., 2025). In Pakistan, ETEA entrance tests have good predictive validity, but researchers recommend additional non-cognitive predictors to improve accuracy (Ali & Ali, 2010).

3. High Validity Predictors
   First, the Scholastic Aptitude Test (TBS) shows that predictive validity is significant in several fields of study, especially social sciences (Asrijanti, 2014; Krisna, 2018). Second, high school grade point average (HSGPA) is a strong predictor of academic performance in college, sometimes stronger than standardized tests such as the SAT (Baron & Norman, 1992). Third, the National Achievement Test (NAT) and the National Exam (UN) scores have varying validity—in several significant studies (Ferdhiana, 2015; Dani & Sroyer, 2016), but the contribution to GPA variation is often small (<30%). Fourth, Combined Tests (e.g., a combination of HSGPA, general aptitude tests, and academic achievement tests—Vista & Alkhadim, 2022) tend to increase predictive power compared to a single test.
4. Factors Affecting the Power of Validity
   The first factor is the type of test, tests that measure general cognitive ability (aptitude) tend to be more stable in predicting early performance (first year), while long-term predictions (cumulative GPA) are often more influenced by non-test factors (motivation, learning support, social adaptation). The second factor is the field of study, Predictive validity differs between study programs. For example, TBS is more accurate in predicting achievement in the social studies department than science (Krisna, 2018) The third factor is the homogeneity of the criterion score, If the criterion GPA has a narrow distribution (homogeneous), then the predictive validity coefficient tends to be low (Azwar, 2008). The fourth factor is gender, several studies (Abdelfattah, 2022; Ali & Ali, 2010) found differences in the power of prediction based on gender.
5. Synthesis of Predictive Validity-Criterion Validity Relationship
   There is a positive relationship between predictors (entrance tests) and criteria (GPA/academic performance). There is no single single test that is a perfect predictor for all contexts. The combination of multiple predictors results in stronger criterion validity. Short-term predictions (first year) tend to be more accurate than long-term predictions (graduation GPA). On non-achievement tests (e.g., general ability or scholastic aptitude), predictive validity is often more noticeable in certain areas.

Core Findings of Cross-Study

1. There is no single predictor that "wins in all contexts". Consistent results show the need for a combination of indicators (HSGPA + merit tests) rather than a single test; the combined model provides a more stable explanation of GPA/graduation.
2. HSGPA/class rankings are often very strong (even surpassing the SAT in many contexts) for predicting GPA/graduation; This power is also evident at the institutional level.

3. The aptitude test remains important: a significant contribution, especially for predicting initial performance (Year-1/Year-prep) and as incremental validity above the HSGPA (multisite/country findings).
4. Short-term > long-term predictions: the coefficients for the initial GPA are higher; Adjacent year-to-year GPA is the best predictor of graduation GPA.
5. Important level of analysis: at the group/institution level, the correlation with graduation rates can be very high (r≈0.62–0.73), exceeding individual studies.

Patterns per Predictor (Based on Table 1 Items)

1. Local FFB/aptitude: significant for some study programs (strong in social studies; variation in science).
2. UN/national achievement tests: tend to be positive but limited (variants are described as small; influenced by subject composition & school/region context).
3. HSGPA/class: consistently strong for GPA/graduation; Very useful when used in conjunction with standard tests.
4. Combined tests (GAT/NAT/SAAT + HSGPA): improve selection accuracy and prediction of graduation/completion of studies; Inter-university rankings over pre-university sizes are relatively consistent.
5. The case of TPA 07A (UGM): low validity due to the homogeneity of the criteria (narrow 1st semester GPA), a classic example of how the range restriction / small variation lowers the coefficient.

Moderator & Causes of Coefficient Variation

1. Field of study: the predictive power differs between study programs (e.g. FFB is more accurate in social studies than in science).
2. Characteristics of the criteria: narrow distribution of GPA → down correlation.
3. Gender & selection pathways: some studies show differences in the power of predictions by gender/pathway of entry.
4. Level of measurement: institution (group) vs individual; at the institutional level, the lower quantile indicator (p25 SAT/ACT) is very informative for 6-year graduation.

.

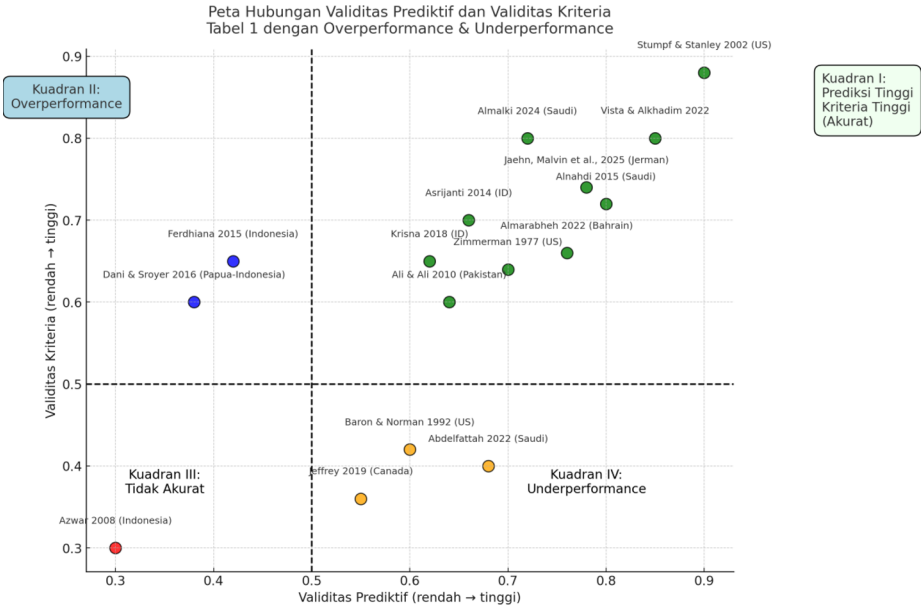

FIGURE 2. Predictive Validity and Criterion Validity Relationship Map (Table 1). A picture of the distribution of the articles in Table 1 into four quadrants (Accurate, Overperformance, Inaccurate, Underperformance). Quadrant 1 is accurate, quadrant 2 is overperformance, quadrant 3 is inaccurate, quadrant 4 is underperforming.

Quadrant I – Accurate (High Prediction, High Criteria)

The majority of articles fall in this quadrant (e.g. Almalki 2024; Jaehn, Malvin. et al.2025 ; Asrijanti 2014; Krisna 2018; Alnahdi 2015; Vista & Alkhadim 2022; Stumpf & Stanley 2002; Zimmerman 1977; Ali & Ali 2010; Almarabheh 2022). Their results consistently show that high school GPA, achievement tests, and aptitude tests have significant predictive power on GPA, early-year GPA, or graduation, so the predictive validity and criteria are equally high.

Quadrant II – Overperformance (Low Prediction, High Criteria)

Articles such as Ferdhiana 2015 and Dani & Sroyer 2016 show that National Examination (UN) scores explain only a small fraction of the variation in GPA. However, the actual performance of students turned out to be better than the test prediction results, indicating that other factors (motivation, entry path, academic support) played a role.

Quadrant III – Inaccurate (Low Prediction, Low Criteria)

The case of Azwar 2008 (TPA UGM) confirms how tests with limited reliability and homogeneous criteria (first-semester GPA) produce very low validity. Both the predictors and the criteria are weak so they are unable to give an accurate picture.

Quadrant IV – Underperformance (High Prediction, Low Criteria)

Several articles, for example Abdelfattah 2022; Baron & Norman 1992; Jeffrey 2019, shows the phenomenon of overprediction. Entrance tests or initial academic scores are indeed high in predictive validity, but they fail to correlate with long-term performance (cumulative GPA, graduation, or final exams). This means that the instrument sets expectations too high compared to actual performance.

Quadrants and Article Interpretation

Article	Quadrant	Interpretation
Almalki 2024 (Saudi)	I (Accurate)	High school grades, achievement & aptitude tests significantly predict a dental student's GPA.
Jaehn, Malvin. Et al., 2025 (German)	I (Accurate)	Entrance test scores and school GPA are predictive of pre-clinical GPA; there is *incremental validity*.
Asrijanti 2014 (Indonesia)	I (Accurate)	The FFB subtest is significant in the Economics major; some significant UN values in Engineering.
Krisna 2018 (Indonesia)	I (Accurate)	TBS predicts that high school achievement will contribute more to social studies than science.
Alnahdi 2015 (Saudi)	I (Accurate)	A combination of HSGPA & predictive NAT for GPA; GAT is the best predictor for graduation.
Vista & Alkhadim 2022 (Saudi)	I (Accurate)	HSGPA, GAT, SAAT consistently predicts university performance and graduation.
Stumpf & Stanley 2002 (US)	I (Accurate)	A school's SAT/ACT & GPA scores are strong predictive for an institution's graduation rate.
Zimmerman 1977 (US)	I (Accurate)	Traditional cognitive tests & predictive learning attitude scale first-year GPA EOP.
Ali & Ali 2010 (Pakistan)	I (Accurate)	The ETEA entrance test is valid for engineering performance, despite gender differences.
Almarabheh 2022 (Bahrain)	I (Accurate)	Science test scores are highly predictive of the short & long term performance of medical students.
Ferdhiana 2015 (Indonesia)	II (Overperformance)	The UN only partially explains GPA; students have a GPA higher than the test prediction.
Dani & Sroyer 2016 (Papua, Indonesia)	II (Overperformance)	The UN value has little effect; The entry pathway significantly affects the GPA.
Azwar 2008 (Indonesia)	III (Inaccurate)	TPA 07A failed to predict first-semester GPA; low validity due to the homogeneity of the criteria.
Abdelfattah 2022 (Saudi)	IV (Underperformance)	The entrance test is predictive for an initial GPA, but not for a long-term cumulative GPA.
Baron & Norman 1992 (US)	IV (Underperformance)	Class rank & achievement test is stronger; Weak SAT → indication of *overprediction*.

Article	Quadrant	Interpretation
Jeffrey 2019 (Canada)	IV (Underperformance)	Biology tests are significant for passing the CPNRE exam, but other predictors are weak.

Practical Implications for Selection & Monitoring Design

1. Use indicator packs: HSGPA/class + one achievement test + one general ability test (aptitude/TBS/GAT). This gives incremental validity to the initial year GPA and graduation chances.
2. Separate prediction objectives:– Short-term goals (Year-1/Year-prep): emphasize test scores (aptitude/achievement).– Long-term (graduation GPA): monitor the trajectory of the annual GPA as a superior predictor.
3. Pay attention to the variants of the criteria: avoid using criteria that are too homogeneous (e.g. an initial semester GPA that "piled up" in a narrow range); Consider alternative/composite criteria (e.g. year GPA, retention, graduation).
4. Fairness audit across study programs & gender: predictor×study program/gender interaction test; adjust the weight of the study program-specific if necessary.
5. Use institutional information: lower quantile indicators (p25 SAT/ACT or equivalent) in cohorts are useful for institution-level dropout risk management

In addition, the findings in Table 1 have direct relevance to the Sustainable Development Goals (SDGs) agenda, especially SDG 4: Quality Education. The validity of predictions and the validity of criteria in university entrance selection are an important part of efforts to provide inclusive, fair, and quality higher education. When selection indicators rely only on one measure—for example, aptitude tests or national test scores—the potential for bias increases and can hinder access for students from diverse socioeconomic backgrounds. In contrast, a more comprehensive selection strategy based on a package of indicators, as shown in the studies in Table 1, reinforces the principle of equity in higher education.

Furthermore, success in predicting academic achievement (e.g. continuous GPA until graduation) contributes to SDG 8: Decent Work and Economic Growth, because effective and equitable higher education produces graduates who are ready to enter the job market with relevant competencies. Thus, a selection policy based on predictive validity and criteria not only improves academic accuracy, but also serves as a mechanism to accelerate the achievement of sustainable development targets in the fields of education and employment.

TABLE 2. Result

No.	Title,Author, year,journal,country	Prediction Validity	Validity of Criteria	Result
1.	Selection and academic success of medical students in Hamburg,German ? Author : [42],BMC Medical Education Journal,German	The tests used to predict are the Hamburg Natural science Test (Ham-Nat), Pre-University Education Attainment (PEA), and Multiple mini-Interview (HAM-Int). The three tests were to predict the academic success of medical students at the University of Hamburg	Students admitted through entrance tests and pre-university education achievement quotas perform better in the first three semesters. The validity of the criteria of HAM-Nat compared to the achievement of pre-university education shows that HAM-Nat provides little additional validity compared to the achievement of pre-university education	**1.** HAM-Nat Predictive Validity: The HAM-Nat test had low predictive validity (r = 0.31) after correction for range restrictions. This score may be due to the effects of self-selection, high selection ratios, and test difficulties that exceed the demands of the medical curriculum. **2.** Performance Based on Admission Pathway: Students admitted through entrance tests or pre-university educational achievement quotas show better performance compared to those admitted through waitlists or foreign student quotas. **3.** Sociodemographic factors: Factors such as the parent's age, gender, and educational background have little influence on the performance of the entrance test. **4.** State Influence: Differences in education policies between German states affect the relationship between pre-university educational attainment and study outcomes.
2.	An Examination of the Predictive Validity of a Measure of College Admissions Applicants' Attributions of Success and Failure By [43]. Journal of Postsecondary Student Success, US	Attribution scores as predictors were tested against future outcomes (4-year graduation, 5-year graduation, UGPA) using logistic & linear regression. This article found that the attribution scores resulting from the measure made a	It explicitly does not use the term "validity criteria," but outcomes such as GPA and passing are tangible criteria that are relevant. This size is valid for use in various subgroups of students. Nonetheless, the	Validity of Attribution Measures: The attribution measures used in this study are valid for use in various subgroups of students. The attribution scores resulting from this measure make a statistically significant but nominal contribution to the prediction of four-year bachelor's degree completion. Predicted Cumulative GPA: Attribution scores make no statistically or practically significant contribution to the prediction of undergraduate cumulative GPA. .

		statistically significant but nominal contribution to the prediction of four-year bachelor's degree completion. However, attribution scores make no statistically or practically significant contribution to predicting undergraduate cumulative GPA and five-year degree completion	contribution of this measure to the prediction of important educational outcomes such as cumulative GPA and completion of a bachelor's degree remains minimal. Outcomes (especially 4-year graduation) can be seen as evidence that the attribution construct has a criterion relevance to academic success	Five-Year Degree Completion: Attribution scores also do not make a statistically or practically significant contribution to the prediction of five-year degree completion.
3.	Integrating Self-Regulated Learning and Individual Differences in the Prediction of University Academic Achievement Across a Three-Year-Long Degree.[44],Metacognition and Learning,Spanish	1.Previous academic performance (combination of high school grades + entrance test scores)Cognitive SRL strategy (regulation of cognition & metacognition) 2. Non-cognitive factors: such as conscientiousness, academic engagement, self-efficacy, regulation of behavior and context 3. Cognitive /	1. **Annual GPA score** for 3 years of undergraduate study. 2. **Construct** latent GPA factor in *structural equation modeling (SEM).* The two structural equation models tested explain about 27-29% variance in latent GPA factors exclusively from the global variable proxy of non-cognitive factors as latent predictors, which is	1. GPA correlation: The results showed that GPA correlated with previous academic performance (a combination of high school GPA and college entrance test scores), academic self-efficacy, academic engagement, independent learning strategies (SRL), and conscientiousness. 2. Hierarchical Regression Analysis: Non-cognitive factors (academic engagement, academic self-efficacy, behavioral and context regulation, and rigor) alone explain 17-25% of the variance in GPA over three years, and previous academic performance explains an additional 25% of variance, so that the total variance described reaches up to 50% in GPA. 3. Additional Validity: Academic engagement and regulation of behavior and context show additional validity over cognitive predictors such as previous

		intelligence ability (inductive reasoning). Predictions of non-cognitive factors significantly explain the 17-25% variance in GPA over three years. Previous academic performance also explains an additional 25% variance, so that the total variance described reaches up to 50% in GPA.	new and promising evidence of strong criterion validity.	academic performance, inductive reasoning, and regulation of cognition and metacognition. 4. Structural Equation Model (SEM): Two structural equation models explain approximately 27-29% variance in latent GPA factors exclusively from the global variable proxy of non-cognitive factors as latent predictors, which is new and promising evidence of the strong validity of their criteria.
4.	The predictive validity of the MCAT for medical school performance and medical board licensing examinations: a meta-analysis of the published research. By [45],Canada.	The MCAT has significant predictive validity to performance in medical school and medical licensing exams. This means that MCAT scores can be used to predict how well a student will perform in medical programs and licensure exams. The MCAT includes subtests:	The validity of the criteria in this context refers to the extent to which the MCAT score correlates with actual performance in medical school and medical licensing exams. This article shows that while there is a significant correlation, the validity of MCAT criteria varies from small to moderate, suggesting that other	1. MCAT Predictive Validity: 1. Preclinical Years: The predictive validity coefficient for the MCAT in preclinical years was (r = 0.39) (95% confidence interval [CI], 0.21-0.54). 2. USMLE Step 1: The predictive validity coefficient for USMLE Step 1 is (r = 0.60) (95% CI, 0.50-0.67). 2. Biological Sciences Subtest: 1. The biological sciences subtest was found to be the best predictor of performance in the preclinical years with (r = 0.32) (95% CI, 0.21-0.42). 2. For USMLE Step 1, this subtest has a coefficient of validity (r = 0.48) (95% CI, 0.41-0.54).

		Biological Sciences (BS) Physical Sciences (PS) Verbal Reasoning (VR) Writing Sample (WS)	factors also play a role in determining student performance. Medical school performance: pre-clinical GPA, clinical GPA, total GPA. Medical licensing examinations: USMLE Step 1 and Step 2 (medical licensing exam).	The MCAT consistently has a moderate–strong predictive validity of academic performance in medical school. Predictive validity was higher for medical licensure exams ($r \approx 0.60$) than for medical school GPAs. Different subtests predict different outcomes: BS/PS for Step 1 (basic science), VR for Step 2 and clinical performance. The results of the meta-analysis confirm the MCAT as a test with good criterion validity for the selection of medical students
5.	Dear admission committee. . .: Which moves in application essays predict student master grades?,by [46], Journal PLoS ONE,Netherlands	This article shows that an application essay can be a valid prediction tool for study success. Using the Latent Dirichlet Allocation (LDA) method, the authors identify seven "moves" or units of text in the essay that can predict student grades in the master's program OLS (linear regression) models that include *moves* as predictors show that the addition of *moves* adds to the described variance (R^2	Of the seven moves identified, five of them had a significant effect on student grades. Moves related to intrinsic motivation and specific knowledge, such as "master specific" and "interest to learn", have a positive effect. On the other hand, moves such as "research skills", "societal impact", and "city and university" have a negative effect. This article provides evidence that an	Conclusion: This study shows that application essays can be used as a valid admission tool. Topics related to intrinsic motivation and specific knowledge have a positive effect on student grades, while topics such as research skills and social impact have a negative effect. Practical Implications: These findings may assist universities in evaluating application essays more effectively, with a focus on intrinsic motivation and specific knowledge as the primary criteria. Limitations: This research is limited to one faculty at one university, so the results may not be generalized to other contexts.

		adjusted > 0.2). In other words, *the moves* in the application essay have the ability to predict the master's academic performance.	application essay can be used as a valid admission criterion, especially if it focuses on intrinsic motivation and specific knowledge	
6.	Considering vocational training as selection criterion for Medical students: evidence for predictive validity. By [47], Advances in Health Sciences Education,Germany	This study shows that relevant vocational training has significant predictive validity in predicting the academic performance of medical students during the first two years of study. Vocational training showed a positive predictive association with study outcomes beyond GPA (Grade Point Average) and TMS (Test für Medizinische Studiengänge) cognitive test results.	Vocational training makes an additional contribution to the prediction of study outcomes beyond cognitive selection criteria such as GPA and TMS2. The multi-level model used in the study showed that vocational training, along with GPA and TMS, explained 14.5% of the variance in study outcomes Overall, the results of this study support the use of vocational training as an additional selection criterion for medical studies, which may add	Positive Correlations: Relevant vocational training had a significant positive correlation with the study outcomes of medical students during the first two years of study. Vocational training showed a positive predictive association with study outcomes beyond GPA (Grade Point Average) and TMS (Test für Medizinische Studiengänge) cognitive test results. Multi-Level Model: A multi-level model with GPA, TMS, vocational training, and gender variables as predictors explains 14.5% of the variance in study outcomes. Vocational training showed a positive predictive association with study outcomes (\u03B2 = 0.33, p = 0.008) outside of GPA (\u03B2 = 0.38, p < 0.001) and TMS (\u03B2 = 0.26, p < 0.001). Location Differences: The analysis showed that there were significant differences in the study results between the two study sites, namely the University of Heidelberg and the University of Göttingen. However, vocational training still showed significant predictive validity in both locations. Overall, the results of the analysis support the use of vocational training as an additional selection criterion for medical studies, which can add predictive validity to cognitive criteria such as GPA and TMS.

			predictive validity to cognitive criteria.	

Table 2 Discussion

If we notice, Table 2 contains a selection instrument that differs from the general pattern in Table 1. If Table 1 is dominated by classic cognitive tests (such as TBS, UN, SAT, HSGPA), then the articles in Table 2 actually introduce a more diverse form of selection: ranging from domain-specific tests (HAM-Nat), multiple mini-interviews (MMI), to non-cognitive factors such as attribution, self-regulated learning, and even vocational experience. There are also interesting examples such as the application essay and the standardized professional test (MCAT).

First, for the context of medicine in Germany (Meyer et al.), the HAM-Nat test was shown to provide only a small additional prediction compared to school achievement. Here it can be seen that specific science tests are indeed relevant to the medical curriculum, but their effect is diminished because the very strict selection makes the student variance too narrow. In contrast, admissions pathways based on school achievement are more consistent in predicting early success. This shows that the "fit" between the test and the curriculum needs to be balanced with the reality of selection in the field.

Second, the Paris & Heiser (2022) study tried to include psychological factors in the form of attribution of success or failure. The results were somewhat surprising: attribution scores were indeed associated with four-year graduation, but they were not significant at all for a five-year GPA or graduation. This means that psychological factors like this may be more suitable to be viewed as a risk indicator (e.g. persistence) than a direct predictor of academic performance.

Third, strong evidence comes from the research of Pérez-González et al. (2022) in Spain. They showed that non-cognitive variables such as academic engagement, self-efficacy, and self-regulation could explain up to a quarter of the variance in GPA, even close to 50% in total when combined with previous achievements. Such a large number is quite convincing. I see this as a signal that the world of higher education cannot rely solely on cognitive tests; Motivational factors and student learning strategies have an important role.

Fourth, in the realm of international medicine, the MCAT has long been debated. The meta-analysis of Donnon et al. (2007) gives a fairly clear answer that the MCAT validly predicts both academic performance and medical licensing exams. Interestingly, the strongest validity can be seen in Step 1 which is based on basic science. This makes sense, because the content of the test and the outcome criteria are completely in line (content alignment).

Fifth, there is a more qualitative approach: application essays. Research by Boer et al. (2024) shows that the content of essays can be analyzed quantitatively and turns out to have predictive power. Moves that emphasize intrinsic motivation and interest in learning are actually positive, while moves such as "research skills" or "societal impact" are negatively related to grades. This is a bit counter-intuitive, but it could be because the statements are cliché and don't really describe a student's readiness to learn.

Sixth, additional evidence comes from Germany again (Amelung et al., 2022) that examined vocational training. The results are clear: relevant work experience makes an additional contribution to predicting the performance of medical studies, even after being controlled with GPAs and cognitive tests. From this we can draw a lesson that practical experience should not be underestimated.

Reflection & Red Thread

In general, Table 2 shows a new direction in college entrance selection. There is no longer a complete reliance on academic test scores. Instead, the combination of various factors—from

school grades, special tests, non-cognitive factors, to practical experience—provides a more complete picture of a student's potential.

Of course, each approach has its limitations. Domain-specific tests are prone to range restriction, non-cognitive constructs are sometimes difficult to measure objectively, while essays are prone to assessor bias. But with more advanced methodologies (e.g. SEM or text mining), the evidence for its validity is getting stronger.

Relevance to SDG 4 (Quality Education)

The analysis of Table 2 has direct implications for SDG 4. Target 4.3 emphasizes equal access to higher education. By expanding the selection instruments, students from diverse backgrounds have more entrances—not just those who excel on cognitive tests.

In addition, targets 4.1 and 4.4 emphasize the quality of learning outcomes and job skills. Tests such as the MCAT or HAM-Nat ensure academic readiness that matches the demands of the program, while non-cognitive variables and vocational experience increase students' chances of survival and success until the end.

Thus, the holistic selection approach as illustrated in Table 2 is not only about "choosing who is accepted," but also about supporting the sustainability and success of the study. That said, this table provides empirical evidence that more inclusive and varied selection practices are an important part of achieving quality education for all.

Overall, the Table 2 map confirms that the multipredictor approach—combining cognitive, non-cognitive, and practical experience—provides more accurate predictive results. These findings support the principles of fair selection and improving the quality of higher education as outlined in SDG 4 (Quality Education). In addition, the relationship with SDG 3 (Good Health and Well-being) can be seen in the validity of the selection of health workers, while the contribution to SDG 8 (Decent Work and Economic Growth) is present through the relevance of vocational selection to job readiness.

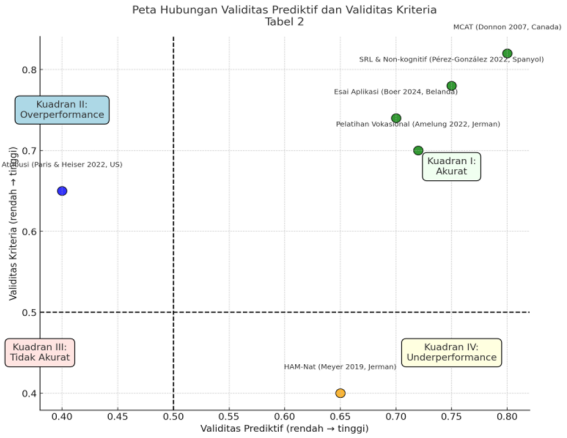

FIGURE 3. Predictive Validity and Criterion Validity Relationship Map (Table 2). The following figure shows the distribution of Table 2 articles into four quadrants (Accurate, Overperform, Inaccurate, Underperformance). Most of the studies fell into Quadrant I, while there were 1 study in Quadrant II, 1 study in Quadrant IV, and no studies in Quadrant III.

Summary Table of Articles Table 2

Article	Quadrant	Brief Interpretation

Meyer et al., 2019 (Germany) – HAM-Nat	IV (Underperformance)	Low predictive validity (r≈0.31), small contribution compared to pre-university achievement; influenced by self-selection and policy.
Paris & Heiser, 2022 (US) – Attribution	II (Overperformance)	Significant for 4-year graduation, but weak for cumulative GPA & 5-year graduation; Practical contributions are limited.
Pérez-González et al., 2022 (Spain) – SRL & Non-cognitive	I (Accurate)	Non-cognitive factors + previous academic performance explain up to 50% of GPA variances; The validity of the additional is strong.
Donnon et al., 2007 (Canada) – MCAT Meta-analysis	I (Accurate)	Predictive validity MCAT (r≈0.39–0.60); subtests are relevant to different outcomes; The validity of the criteria is consistent.
Boer et al., 2024 (Netherlands) – Application Essay	I (Accurate)	The application essay predicts the master's grade (R^2>0.20); positive motivation & specific knowledge; other aspects are negative.
Amelung et al., 2022 (Germany) – Vocational Training	I (Accurate)	Vocational experience was positively associated with the initial results of medical studies; additional contributions on top of GPA & TMS.

Practical implications

1. Multi-instrumental-based selection needs to be prioritized: the combination of cognitive tests, non-cognitive factors, and experiential indicators has proven to be more accurate.
2. Standard cognitive tests (MCAT, TMS, HAM-Nat) are still relevant, but they should not be the only measure.
3. Non-cognitive & application essays provide an overview of motivation, engagement, and self-regulation, thus enriching the selection process.
4. Vocational experience can be considered as an additional criterion, especially for fields that demand practical readiness.
5. It is important to conduct a context audit: educational policies, study locations, and types of outcomes must be considered so that the validity of the predictors–criteria is more accurate and fair.

Table 3 is different from tables 1 and 2 because it measures mental or non-cognitive, such as physical endurance, intention to quit work, leadership behavior, health forecast, etc. Non-cognitive measurement tables are differentiated from cognitive measurements to see the picture of prediction validity and criterion validity, whether prediction validity and criterion validity are related to subsequent actions. In Health, the description of the relationship between the validity of the prediction and the validity of the criteria is seen from the initial measurement and then compared with the measurement of the final outcome whether the estimated action at the beginning can predict the outcome of the action at the end. Meanwhile, those outside of health research see the validity of the results of the treatment carried out. The description of the relationship can be read in table 3:

TABLE 3. Summary of Research Articles on Predictive and Criterion Validity

No.	Title,Author, year,journal,country	Prediction Validity	Validity of Criteria	Result
1.	Development of Questionnaire on the Sense of Workplace Involution for Newly Recruited Employees and Its Relationship with Turnover Intention [48], International Journal of Environmental Research and Public Health, China.	This article shows that "sense of workplace involution" has a significant predictive effect on "turnover intention" or intention to quit a job. This means that high scores on questionnaires about involution in the workplace can predict new hires' intentions to leave their jobs in the future.	The validity of the criteria in this article was tested by measuring the correlation between "sense of workplace involution" and "turnover intention." The results showed a significant positive correlation, which means that the questionnaire developed had good criterion validity because the results corresponded to the expected criteria.	1. Questionnaire Development: This article develops a Questionnaire on a Sense of Involution in the Workplace for New Employees. The questionnaire was tested on 571 new hires and showed good reliability and validity. 2. Factor Analysis Results: Exploratory and confirmatory factor analysis identified four main dimensions of a sense of involution in the workplace: resource scarcity, forced commitment, futility of effort, and negative experiences. 3. Relationship with Intention to Quit: A sense of involution in the workplace is significantly and positively correlated with an intention to quit work. This means that the higher the sense of involution, the higher the employee's intention to quit. 4. Practical Implications: This study provides a valid tool for measuring feelings of involution in the workplace and demonstrates the importance of addressing feelings of involution to reduce employee intent to quit and retain employees.
2.	A validity study of a work sample test of leadership behavior using supervisor and subordinate ratings as criteria by [49], Scandinavian Journal of Psychology, Sweden.	The work sample test used in this study aimed to predict leadership behavior more than 6 months later.The results showed that the test scores of the work sample for the "Consideration" dimension had a	Assessments from subordinates and superiors are used as criteria to measure leadership dimensions such as "Influencing others", "Consideration", and "Planning". The results of the study show that the work sample test can	Predictor Validity: 1. The test score of the work sample for the "Consideration" dimension has a significant predictive validity to the assessment of subordinates with a correlation of $\rho = 0.33$. 2. Test scores also showed a positive association with assessment from superiors, although the correlation was lower ($\rho = 0.22$) and the confidence interval included zero. Validity of Criteria: 1. Assessments from subordinates and superiors are used as criteria to measure leadership dimensions such as "Influencing others", "Consideration", and "Planning".

		significant predictive validity to the assessment from the subordinate (ρ = 0.33) and also showed a positive relationship with the assessment from the superior (ρ = 0.22), although the confidence interval for the assessment from the superior was zero	be used mainly to predict the dimension of "Consideration"	2. The work sample test proved to be effective especially in predicting the dimension of "Consideration". Statistical Analysis: 1. The correlation between test scores and assessments from superiors and subordinates shows that the work sample test can predict leadership behavior quite well. 2.Corrections for range limitation and criterion unreliability help improve the accuracy of results. Overall, the results of this study support the use of the work sample test as a valid tool for assessing leadership behavior, especially in the "Consideration" dimension.
3.	Role of diaphragmatic ultrasound in predicting weaning success from mechanical ventilation in pediatric intensive care unit by [50]-Egyptian Journal of Radiology and Nuclear Medicine, Egypt	Diaphragmatic Excursion (DE): DE was found to be a predictor of weaning success. The cutoff values for the right DE are >4.1 mm and the left DE >5.5 mm, with the area under the curve (AUC) of 0.737 and 0.831 respectively. Diaphragmatic Thickening Fraction (DTF) and Time to Peak Inspiratory Amplitude (TPIA): No significant correlation was found between DTF or TPIA and weaning results	This study used a prospective cohort design with 30 pediatric patients undergoing mechanical ventilation.Ultrasound measurements were performed during a spontaneous breathing test (SBT) and were compared between the successful and failed weaning groups. **Results:** The duration of mechanical ventilation was longer in the group that failed weaning. The DE showed significant	The cutoff values for the right DE are >4.1 mm and the left DE >5.5 mm, with the area under the curve (AUC) of 0.737 and 0.831 respectively. Diaphragmatic Thickening Fraction (DTF) and Time to Peak Inspiratory Amplitude (TPIA): No significant correlation was found between DTF or TPIA and weaning results. **1.** Weaning Success: Out of 30 patients, 19 patients (63.3%) were successfully weaned from mechanical ventilation. **2.** Ventilation Duration: The duration of mechanical ventilation was longer in the group that failed to weaning (P=0.017). **Sec. 3**. Diaphragmatic Excursion (DE): There was a significant difference between the successful and failed groups in the right (P=0.032) and left (P=0.022) DE with cut-off values of >4.1 mm and >5.5 mm. 4.Diaphragmatic Thickening Fraction (DTF) and TPIA: There is no significant difference between the two groups in DTF or TPIA.

			differences between the successful and failed groups, while DTF and TPIA showed no significant differences	
4.	Criterion-Related Validation of a Music-Based Attention Assessment for Individuals with Traumatic Brain Injury. By [51],Journal of Environmental Public Health	The validity of the predictors in this article was tested by linking MAA results with various standard neuropsychological tests, such as the Wechsler Adult Intelligence Scale (DST), the Delis-Kaplan Executive Functioning Scale (color-word interference test), and Conner's Continuous Performance Test.The results show a significant correlation between MAA and these neurocognitive batteries, suggesting that MAA can be a valid tool for measuring different types of attention deficits[1]	The validity of the criteria was also confirmed through multiple regression analysis which showed that MAA was a significant factor in predicting attention ability. This means that the MAA can be used as a reliable measure to assess attention deficits in patients with TBI	1. Validity of MAA Criteria: Music-based assessment (MAA) showed a significant correlation with several neuropsychological tests, demonstrating the validity of a good criterion for measuring attention ability in individuals with traumatic brain injury (TBI). 2. Predictors of Cognitive Capacity: MAA performance and education level were found to be significant predictors of general cognitive capacity, demonstrating the potential of MAA as a predictive tool in neurocognitive rehabilitation. 3. Auditory Attention Measurement: MAA can measure different types of auditory attention and attention incapacitation, which is essential for assessment in a realistic auditory environment.

5.	Demographic predictors of bedtime procrastination in the Japanese population by [52],Sleep and Biological Rhythms,Japan.	Age was found to be a strong predictor for sleep procrastination (BP), with younger age groups (≤40 years) having a higher risk of BP. No significant effect of employment status	Criterion Validity: CPM shows good internal consistency (Cronbach's α = .90–.91) and test-retest reliability (ICC = .86). BPS also had a moderate positive correlation with insomnia and a negative correlation with sleep quality and sleep duration on weekdays	1. Validity and Reliability of BPS: The Japanese version of the Bedtime Procrastination Scale (BPS) shows good validity and reliability. 2. Demographic Predictors: Age is a strong predictor of sleep procrastination (BP), with younger individuals having a higher risk of BP than older ones. 3. No Influence of Gender and Employment Status: No significant differences were found in BP scores based on gender or employment status. 4. Relationship with Insomnia: BP has a significant positive correlation with insomnia symptoms.
6.	Development and assessment of algorithms for predicting brain amyloid positivity in a population without dementia by [53], Alzheimer's Research & Therapy, France.	The most effective predictive models combine demographic data, cognitive assessment, ApoE status, and Alzheimer's-related blood biomarkers (Aβ42/40 and P-tau181). The model achieved an Area Under the Curve (AUC) of 0.82 in the MEMENTO sample and 0.90 in the external validation sample[1]. This model is significantly better	The validity of the criteria was tested through internal validation using cross-validation and external validation using data from the Amsterdam Dementia Cohort. The model was also evaluated in sub-samples that met the key criteria of the Appropriate Use Recommendations (AUR) for lecanemab, demonstrating comparable performance.	**1.** Model Accuracy: + The prediction model achieved an Area Under the Curve (AUC) of 0.82 in the MEMENTO cohort and 0.90 in the external validation cohort. + This model shows a significant improvement in accuracy compared to reference models that use only demographic and cognitive data. **2.** External Validation: + External validation using data from the Amsterdam Dementia Cohort shows that this model is reliable and applicable to different populations. **3.** Performance Evaluation: + Brier Score: This model has a low Brier score, indicating an accurate probability prediction. + Calibration Curve: The calibration curve indicates that the probability prediction of the model corresponds to the observed results, indicating a good calibration. **4.** Sensitivity Analysis: The model remains stable and accurate despite variations in data or model assumptions, indicating high robustness.

		than the reference model that uses only demographic and cognitive data, with an AUC difference of 0.10	This model suggests that demographic data, cognitive assessments, and blood biomarkers can be used to predict amyloid status with high accuracy without the need for ApoE genotype or brain MRI data	Overall, the results of this study show that the predictive model developed can be used to detect brain amyloid positivity with high accuracy, which can aid in early detection and Alzheimer's intervention
7.	Criterion validity of functional movement screen as a predictor of sports injury risk in Chinese Police staff. By [54], Environmental Research and Public Health.China	This article shows that the FMS total score can be used to predict the risk of sports injuries. The study found that lower FMS scores correlated with a history of previous injuries. The identified threshold score was 13.5, with an AUC (Area Under Curve) value of 0.701, indicating an acceptable level of prediction validity.	The validity of the criteria in this context refers to how well the FMS can predict injuries based on established standards. This article uses injury history as a criterion to evaluate the validity of FMS.The results showed that the FMS had good criteria validity in identifying the risk of injury to police personnel	**1.** FMS Score Distribution: The total FMS score of the participant indicates a normal distribution. The average score for police staff with a history of sports injuries was 10.6 ± 2.28, lower compared to those with no history of injury (12.4 ± 2.26). **2.Threshold** Values: The total FMS score used as a warning of sports injury risk is 13.5, with an acceptable AUC value (0.701). **3.** Age Influence: FMS results show a clear downward trend with age. **4.** Prediction Validity: FMS can be considered by coaches and clinicians as a pre-workout physical examination to avoid sports injuries in Chinese police staff.
8.	canonical babbling ratio-concurrent and predictive evaluation of the 0.15 criterion. By [55], Journal of	Predictive validity refers to the ability of the Canonical Babbling Ratio (CBR) to predict future language development.	Criterion validity refers to how well the CBR with criterion 0.15 reflects the actual canonical babbling stage.The study found	**1**. Concurrent Validity: The Canonical Babbling Ratio (CBR) with the criterion of 0.15 was found to be a valid measure to determine whether a child has reached the canonical babbling stage at 10 months of age. The Under the Curve Area (AUC) of the Receiver-Operator Characteristics (ROC) is 0.87, indicating high validity.

	Communication Disorders, Sweden	In this study, CBR with a criterion of 0.15 was used to predict children's language skills at an older age.The results showed that children who reached or exceeded a ratio of 0.15 at a given age tended to have better language development in the future.	that the 0.15 criterion is a good indicator to determine whether a child has reached the canonical babbling stage, which is an important stage in early language development	**Sec. 2**. Predictive Validity: CBR with criterion 0.15 also shows good predictive ability for future language development. Children who reach or exceed a ratio of 0.15 at 10 months of age tend to have better language development at 30-36 months of age. **3.** Optimal Criteria: The analysis showed that the optimal criterion for CBR was 0.14, which provided the best combination of sensitivity (0.96) and specificity (0.70). **4.** Prediction of Speech/Language Difficulties: CBR with criteria 0.14 or 0.15 showed similar sensitivity (0.71) in predicting speech/language difficulties at 30-36 months of age, but criterion 0.14 had slightly better specificity (0.52 vs 0.42). These results suggest that CBR is a valid and reliable measure for evaluating the stages of canonical babbling and predicting future language development[l].
9.	Constructs Derived from the Addiction Cycle: Predict Alcohol Use Disorder Treatment Outcomes and Recovery: Three Years Following Treatment. By [56], American Psychological Association, US	This study shows that addiction cycle domains, such as reward/incentive salience, relief/negative emotionality, and loss of control/executive functioning, have significant predictive validity in predicting AUD treatment outcomes.	The addiction cycle domain is more strongly associated with treatment outcomes compared to other measures that clinicians may use to predict outcomes (e.g., AUD symptoms). The domains of relief/negative emotionality and reward/incentive salience showed a significant association	**1.** Addiction Cycle Domain Validity: Addiction cycle domains such as reward/incentive salience, relief/negative emotionality, and loss of control/executive functioning have significant predictive validity in predicting the outcome of alcohol use disorder (AUD) treatment and recovery three years after treatment. **2.** Prediction of Treatment Outcomes: Relief/negative emotionality and reward/incentive salience were significantly associated with treatment outcomes at one and three years after treatment, while executive functioning also predicted non-abstinent recovery at three years. **3.** Clinical Utility: The addiction cycle domain is more strongly associated with treatment outcomes compared to other measures that clinicians may use to predict outcomes (e.g., AUD symptoms). This suggests that this

		Relief/negative emotionality and reward/incentive salience were significantly associated with treatment outcomes at one and three years after treatment, while executive functioning also predicted non-abstinent recovery at three years	with treatment and recovery outcomes, suggesting that these domains could be used to predict AUD treatment outcomes and long-term recovery. Overall, the results of this study support the utility of domains relevant to the addiction cycle in predicting AUD treatment outcomes and recovery among individuals seeking treatment for AUD.	domain can be used to predict AUD treatment outcomes and long-term recovery.
10.	childhood language skills as predictors of social, adaptive and behavioral outcomes of adolescents with autism spectrum disorder. By [57], Journal of Research in Autism Spectrum Disorders, Spain	Predictor validity in this context refers to the extent to which childhood language skills can predict social, adaptive, and behavioral outcomes in adolescence This study found that structural and pragmatic language skills are significant predictors for outcomes social and	Criterion validity refers to the extent to which the outcomes predicted by childhood language skills correspond to the actual outcomes measured. In this study, the validity of the criteria was tested by comparing predictions of language skills with actual social, adaptive, and behavioral	**1**. Pragmatic Language: Pragmatic language significantly predicts all domains of development, especially stereotypical language and contextual use. **2.Structural** Language: Structural language (especially semantics) only affects prosocial behavior and socialization skills. **3.Adaptive** Skills: Pragmatic language skills as a whole explain 36% of variance in daily life skills. **4.** Behavioral Difficulties: Pragmatic language, particularly stereotyped language and the use of context, predicts behavioral problems.

		behavioral in adolescents with ASD	outcomes in adolescents with ASD	
11.	Derivation and Validation of a Clinical Prediction Score to Identify the Isolation of Pseudomonas in Pneumonia.By Yana Maskov, et al., 2022[58]Journal of the American Society for Microbiology, USA	The clinical prediction score developed has significant predictive validity in predicting isolation of Pseudomonas aeruginosa in pneumonia patients. These predictive scores were based on five predictor variables: tracheostomy status, chronic obstructive pulmonary disease (COPD), enteral nutrition, chronic steroid use, and isolation of Pseudomonas aeruginosa from culture within the last 6 months. At scores greater than 11, the prediction score showed a sensitivity of 52.4% and a specificity of 84.9% in the	These predictive scores demonstrate the validity of good criteria in identifying patients at high risk for isolation of Pseudomonas aeruginosa, which can help guide therapeutic management and appropriate antibiotic selection. The use of these prediction scores can reduce unnecessary use of broad-spectrum antibiotics, which can reduce the risk of drug toxicity, resistance, and other negative consequences	**1.** Development of Prediction Scores: This study successfully developed a clinical prediction score that can identify isolation of Pseudomonas aeruginosa in pneumonia patients. These predictive scores were based on five predictor variables: tracheostomy status, chronic obstructive pulmonary disease (COPD), enteral nutrition, chronic steroid use, and isolation of Pseudomonas aeruginosa from culture within the last 6 months. **2.** Score Validity: This prediction score showed significant predictive validity with a sensitivity of 52.4% and a specificity of 84.9% in the validation cohort. The accuracy of the score was 70.5%, and the area below the recipient operating characteristic curve (AUROC) was 0.77 in the validation cohort. 3. Clinical Benefits: The use of these prediction scores can help direct therapeutic management and appropriate antibiotic selection, as well as reduce unnecessary use of broad-spectrum antibiotics. This can reduce the risk of drug toxicity, resistance, and other negative consequences.

		validation cohort. The score accuracy was 70.5%, and the area under the recipient operating characteristics curve (AUROC) was 0.77 in the validation cohort		
12.	The predictive validity of cognitive ability and personality tests toward police on-the-job training. By [59]- Journal of Indexing and matrics, Norway	Cognitive ability tests (Scales: lst,verbal admin, lt-no), Shapes basic personality test, These tests are used in the selection process of the Police Polytechnic (Norway) through an external provider (Aon Assessment / cut-e)	"On-the-job training performance": the instructor assesses the *work behaviors of the trainees for two years after the initial selection,* The regression approach also controls academic performance in police academy as a covariate	1. The cognitive and personality tests used do not show the expected predictability. 2. In regression that controlled for academy performance, the relationship between cognitive test scores and work performance was reversed (negative). 3. The authors suggest a cautious interpretation due to the relatively small sample and the limitations of the psychometric documentation of the tests used.

Table 3 Discussion

Cross-context General Patterns

Table 3 combines various domains, namely organization, leadership, clinical/medical, neuropsychology, sleep habits, language development, addiction, and police performance. Although the criteria (outcomes) are very different, three consistent patterns appear:

1. Multipredictor models excel, especially when the criteria are complex.

The clearest example is in the detection of brain amyloid: a combination of demographic, cognitive, ApoE, and blood biomarkers results in a high AUC (≈0.82–0.90), going beyond simple reference models. Similar patterns are seen in FMS+ age for injury risk, as well as MAA+ education for cognitive capacity.

2. The quality and "proximity" of the criteria determine the magnitude of validity.

When the criterion is close to the construct being measured, e.g. MAA with a series of attention tests; CBR (0.15) with early language development milestones (AUC ≈0.87); or DE ultrasound diaphragm with weaning success, the validity of the criteria tends to be moderate-strong. Conversely, when the construct-criterion distance is further or there is a range restriction, the validity is weakened (e.g., cognitive/personality tests for police OJT that even reverse direction after controlling IP in academia). So the closer the criterion is to the measured construct, the stronger the validity

3. External validation and calibration improve the reliability of claims.

Amyloid studies show good practices (external validation, calibration curve, Brier score). This is in contrast to the context of job selection where psychometric documentation is limited or small samples (Skoglund) make interpretation have to be careful.

Highlights per Domain

1. Organization & HR

   In Chen & Zhang's article (involution → turnover intention): the new instrument shows predictive validity and clear criteria (correlation +); the implications are practical—managing the perception of "endless competition" can lower exit intentions. Similarly, the Sjöberg & Grill article (work-sample leadership): valid for the dimension of "Consideration" (ρ≈0.33 from subordinates; positive from superiors). This emphasizes the importance of task/observation-based tests rather than just cognitive tests. Meanwhile, the article Skoglund (police selection): evidence of underperformance—set after being controlled by the academy's IP, cognitive relationship—field performance becomes negative. Selection redesign (more task simulations, situational judgement, and non-cognitive assessments) is recommended.

2. Clinical/medical
   1) Eskander (diaphragmatic ultrasound): DE (right cut-off >4.1 mm; left >5.5 mm) differentiates success–failure weaning (AUC ≈0.74–0.83); DTF/TPIA is invalid → focus the protocol on DE.
   2) Le Scouarnec (Alzheimer's): a high-AUC and well-calibrated multi-biomarker model; a strong example of prediction validity + criteria with external validation—ready for pre-PET or pre-CSF triage.
   3) Maskov (Pseudomonas score): five clinical variables give AUROC ≈0.77 with high specificity (≈85%); aids antibiotic stewardship (reduces unnecessary broad-spectrum).

3. Neuropsychology & developmental

   In the article Jeong & Ireland (MAA): significant correlation with WAIS/D-KEFS/CPT → validity of good criteria; regression shows predictive power for attention. Nyman's article

(CBR 0.14–0.15): high concurrent & predictive validity for languages aged 30–36 months; optimal criterion 0.14 balances sensitivity–specificity well. And Miranda's article (ASD – childhood language → adolescent outcomes): pragmatic language predicts all domains (social, adaptive, behavioral), explaining ≈36% of ADL variance—a strong argument for early pragmatic intervention.

4. Sleep behavior & occupational health
   In Miyagawa's article (bedtime procrastination): age is a strong predictor of BP; The Japanese version of BPS is reliable and valid and correlates with insomnia/sleep quality—a good population screening tool. While Huang & Liu (FMS – police) article: cut-off 13.5; AUC ≈0.70—sufficient for injury risk screening; scores decrease with age → age-based prevention programs are recommended.

What Do These Findings Mean?

1. For measuring tool design: select relevant, varied, and objectively measurable criteria. Use external validation whenever possible; report calibration (not just AUC/correlation).
2. For practical decisions:– Clinical: prioritize proven indicators (DE for weaning; Pseudomonas scores; amyloid biomarker panels) for triage and rational therapy.– HR/selection: add work-sample/observation and context indicators (involution, leadership consideration), reduce reliance on generic cognitive tests.– Education & rehabilitation: use CBR and MAA as screening tools and setting intervention targets (pragmatic language in ASD).
3. Under/overperformance signals: Skoglund shows underperformance (high prediction but not binding field performance), while some simple instruments with strong criteria (CBR, DE) are accurate.

Limitations to Note

Many studies use a limited sample or a single context (one institution/unit), so generalizations must be careful. Some results only reported correlations without incremental validity tests or decision-curve analysis, and there was no always a group bias test (gender/age/location).

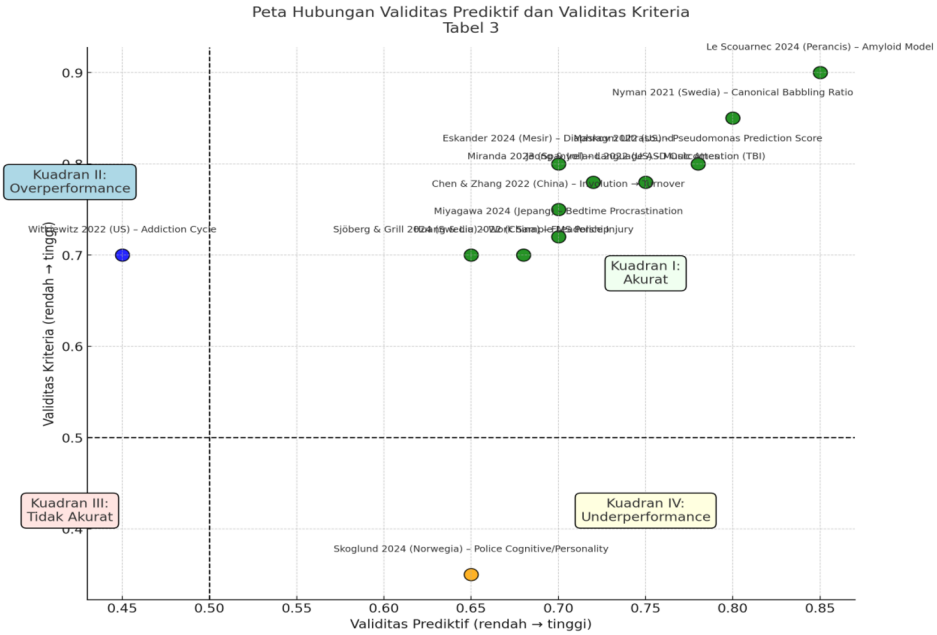

FIGURE 4. Predictive Validity and Criterion Validity Relationship Map (Table 3). The following figure shows the distribution of Table 3 articles into four quadrants (Accurate, Overperform, Inaccurate, Underperform). The majority of studies were in Quadrant I (Accurate), with one study in Quadrant II and one study in Quadrant IV, while Quadrant III was empty.

Summary Table of Articles Table 3

Article	**Quadrant**	**Brief Interpretation**
Chen & Zhang 2022 (China) – Involution → Turnover	I (Accurate)	Valid work involution instruments predict turnover intention; positive correlation with exit intention.
Sjöberg & Grill 2024 (Sweden) – Work Sample Leadership	I (Accurate)	Work-sample tests are valid for predicting 'consideration' behavior; correlation is stronger than subordinate (ρ≈0.33).
Eskander 2024 (Egypt) – Ultrasound Diaphragm	I (Accurate)	Diaphragmatic excursion (DE) significantly predicts weaning; DTF & TPIA are invalid.
Jeong & Ireland 2022 (US) – Music Attention (TBI)	I (Accurate)	MAA correlates with neuropsychological tests; validity of the criteria is good for the attention of TBI patients.
Miyagawa 2024 (Japan) – Bedtime Procrastination	I (Accurate)	BPS is valid and reliable; age is a strong predictor of sleep procrastination; positive correlation with insomnia.

Le Scouarnec 2024 (France) – Amyloid Model	I (Accurate)	Biomarker+demographic+cognitive model AUC 0.82–0.90; good external validation; precise calibration.
Huang & Liu 2022 (China) – FMS Police Injury	I (Accurate)	Low FMS score predicts sports injuries; cut-off 13.5; AUC≈0.70; useful for screening.
Nyman 2021 (Sweden) – Canonical Babbling Ratio	I (Accurate)	CBR ≥0.14 is valid concurrent & predictive of early language development; High AUC (0.87).
Miranda 2023 (Spanish) – Language ASD Outcomes	I (Accurate)	Pragmatic language is a strong predictor of social, adaptive, behavioral outcomes in ASD adolescents.
Maskov 2022 (US) – Pseudomonas Prediction Score	I (Accurate)	Valid clinical prediction score (AUROC≈0.77); high specificity; supporting antibiotic stewardship.
Witkiewitz 2022 (US) – Addiction Cycle	II (Overperformance)	Addiction cycle domains predict strong AUD outcomes 1–3 years; more relevant than AUD symptoms.
Skoglund 2024 (Norway) – Police Cognitive/Personality	IV (Underperformance)	Weak predictive cognitive/personality tests for police OJT; negative relationships after academy control.

Policy Implications (Concise)

1. Set a predictor package according to the context (clinic, HR, education), prioritizing the indicators closest to the criteria.
2. Require comprehensive validity reporting: discrimination (AUC/r), calibration, and external validation.
3. Develop predictive outcome-based intervention pathways (e.g., DE-based weaning protocols; prehab/conditioning for personnel with low FMS; early pragmatic intervention in ASD).

Table 3 shows the contribution of various cross-domain instruments in strengthening the validity of predictive and criterion values, which are relevant to the sustainable development agenda. Within the framework of SDG 3: Good Health and Well-being, findings from research in the medical and neuropsychological fields confirm the important role of accurate metrics for improving health services. For example, the use of diaphragmatic ultrasound in predicting the weaning success of pediatric patients, biomarker and cognitive models for early detection of Alzheimer's, and Music-based Attention Assessment (MAA) in traumatic brain injury (TBI) patients provide evidence that high-validity instruments can support the diagnosis, intervention, and rehabilitation process. Strong validity in this context has direct implications for patient safety, treatment effectiveness, and improved quality of life, which is in line with SDG 3's mission to ensure healthy lives and support the well-being of all people of all ages.

Meanwhile, research in the field of organization and employment is closely related to SDG 8: Decent Work and Economic Growth. Instruments such as workplace involution questionnaires to predict turnover intention, work sample tests that assess leadership behavior, and Functional Movement Screen (FMS) to detect the risk of injury in police personnel, show how predictive validity and criteria can be used to strengthen the quality of the work environment. Through early identification of burnout risks, increased leadership effectiveness, and injury prevention, organizations are able to reduce losses due to employee turnover, increase productivity, and create safer and more sustainable working conditions. Thus, Table 3 confirms that the development and validation of instruments not only provides academic benefits, but is also an integral part of the strategy to achieve SDG 3 and SDG 8.

5. CONCLUSION

The results of this systematic literature review (SLR) show that predictive validity and criterion validity are fundamental aspects in assessing the effectiveness of selection instruments, academic assessments, and measurement tools in the health and employment sectors.

First, in the context of higher education (Table 1), there is no single indicator that is universally superior in predicting study success. The multipredictor model—which combines high school grades, grade rankings, and aptitude and achievement tests—is consistently more stable than a single test. The HSGPA and class rankings tend to provide strong predictive power of GPA and graduation, while general ability and achievement tests add value primarily to mapping a student's initial readiness.

Second, in the alternative instrument for university entrance selection (Table 2), evidence shows that non-cognitive factors such as independent learning strategies, motivation, attribution, and vocational experience also contribute significantly to academic achievement. This underscores the importance of a more holistic approach to selection, relying not only on cognitive tests, but also on psychosocial factors and broader contextual experiences.

Third, in instruments outside the realm of formal education (Table 3), predictive validity and criteria have proven to be relevant in various fields: medical, neuropsychological, organizational, and employment. Health instruments such as diaphragmatic ultrasound or Alzheimer's biomarker models support the achievement of SDG 3 (Good Health and Well-being), while work and organizational instruments such as involution questionnaires, leadership tests, and FMS support SDG 8 (Decent Work and Economic Growth). These findings expand the scope of validity as a concept that is not only academic, but also applicable to human well-being and work productivity.

Overall, the SLR emphasizes that predictive validity and criteria should not be seen as mere statistical coefficients, but rather as strategic instruments to support fair selection, targeted education and health policies, and the creation of a safe and productive work environment. This cross-contextual approach confirms that the success of predictions and the relevance of criteria are an important foundation for the achievement of the Sustainable Development Goals (SDGs)

6. ACKNOWLEDGMENTS

The author would like to thank Prof.Dr. Komarudin, M.Si. as the Rector of Jakarta State University, Prof. Dr. Wardani Rahayu, M.Si. as the Head of the PEP Study Program, and especially to Dr. Achmad Ridwan, M.Si. and Prof. Dr. Dinny Devi Triana, M.Pd. as the promoter and co-promoter for their valuable guidance and input until this article was realized. Also, I would like to thank Ir. Edwar Syarif, M.T. as my life partner who always accompanies

me during the difficulties in completing this article, and to my beloved children. Thank you all for your valuable help and input.

REFERENCES

[1] S. Azwar, *Reliabilitas dan Validitas*, 4th ed. 2022.
[2] D. S. Naga, *Teori Sekor Pada Pengukuran Mental*, 3rd ed. Jakarta: PT Nagarani Citrayasa, 2022.
[3] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser, "Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective," *J. Acad. Mark. Sci.*, vol. 40, no. 3, pp. 434–449, 2012, doi: 10.1007/s11747-011-0300-3.
[4] Siswanto, "Validitas Sebagai Alat Penentu Kehandalan tes Hasil Belajar," *J. Pendidik. Akunt. Indones.*, vol. VI, no. 1, pp. 107–117, 2008.
[5] C. Cunningham, F. Patterson, and J. Cleland, "A literature review of the predictive validity of European dental school selection methods," *Eur. J. Dent. Educ.*, vol. 23, no. 2, pp. 73–87, 2019, doi: 10.1111/eje.12405.
[6] W. B. Armstrong, "The association among student success in courses, placement test scores, student background data, and instructor grading practices," *Community Coll. J. Res. Pract.*, vol. 24, no. 8, pp. 681–695, 2000, doi: 10.1080/10668920050140837.
[7] D. Goldhaber and U. Özek, "How Much Should We Rely on Student Test Achievement as a Measure of Success?," *Educ. Res.*, vol. 48, no. 7, pp. 479–483, 2019, doi: 10.3102/0013189X19874061.
[8] O. L. Liu, L. Frankel, and K. C. Roohr, "Assessing Critical Thinking in Higher Education: Current State and Directions for Next-Generation Assessment," *ETS Res. Rep. Ser.*, vol. 2014, no. 1, pp. 1–23, 2014, doi: 10.1002/ets2.12009.
[9] P. Olszewski-Kubilius and S. Corwith, "Poverty, Academic Achievement, and Giftedness: A Literature Review," *Gift. Child Q.*, vol. 62, no. 1, pp. 37–55, 2018, doi: 10.1177/0016986217738015.
[10] Asrijanty, "Validitas Prediktif Bakat Skolastik dan Prestasi Belajar sebagai Kriteria Seleksi Masuk Perguruan Tinggi," *J. Pendidik. dan Kebud.*, vol. 20, no. 4, 2014, doi: 10.24832/jpnk.v20i4.163.
[11] S. G. A. van Herpen, M. Meeuwisse, W. H. A. Hofman, S. E. Severiens, and L. R. Arends, "Early predictors of first-year academic success at university: pre-university effort, pre-university self-efficacy, and pre-university reasons for attending university," *Educ. Res. Eval.*, vol. 23, no. 1–2, pp. 52–72, 2017, doi: 10.1080/13803611.2017.1301261.
[12] A. Vista and G. S. Alkhadim, "Pre-university Measures and University Performance: A Multilevel Structural Equation Modeling Approach," *Front. Educ.*, vol. 7, no. February, pp. 1–8, Feb. 2022, doi: 10.3389/feduc.2022.723054.
[13] Kementerian Pendidikan Dasar dan Menengah RI, "Peraturan Menteri Pendidikan Dasar dan Menengah Nomor 9 Tahun 2025 tentang Tes Kemampuan Akademik (TKA).," p. https://bpmpsultra.kemdikbud.go.id/wp-content/uplo, 2025.
[14] United Nations Educational, "Education for sustainable development: A roadmap.," *Unesco*, p. https://unesdoc.unesco.org/ark:/48223/pf0000374802, 2020.
[15] World Health Organization., "World health statistics 2021: Monitoring health for the SDGs, sustainable development goals." 2021.
[16] International Labour Organization, *Work for a brighter future: Global commission on the future of work. International Labour Organization.* 2019.
[17] J. M. Rothstein, "College performance predictions and the SAT," *J. Econom.*, vol. 121,

no. 1–2, pp. 297–317, 2004, doi: 10.1016/j.jeconom.2003.10.003.
[18] J. V. D. Agostino and S. J. Powers, "Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis," *Am. Educ. Res. J.*, 2009, doi: 10.3102/0002831208323280.
[19] R. C. Atkinson and S. Geiser, "Reflections on a Century of College Admissions Tests," vol. 38, no. 9, pp. 665–676, 2009, doi: 10.3102/0013189X09351981.
[20] M. H. Razmi, M. Khabir, and S. A. Tilwani, "A Meta-analysis on the predictive validity of graduate record examination (GRE) general test," *Int. J. Lang. Test.*, vol. 11, no. 2, pp. 51–63, 2021, [Online]. Available: https://www.ijlt.ir/article_138057.html
[21] S. Messick, "Validity of Test Interpretation and Use," *ETS Res. Rep. Ser.*, vol. 1990, no. 1, pp. 1487–1495, 1990, doi: 10.1002/j.2333-8504.1990.tb01343.x.
[22] E. Mishler, "Validation in Inquiry-Guided Research: The Role of Exemplars in Narrative Studies," *Harv. Educ. Rev.*, vol. 60, no. 4, pp. 415–443, 1990, doi: 10.17763/haer.60.4.n4405243p6635752.
[23] B. Flynn, R. G. Schroederb, and S. Sakakibara, "A framework for quality management research and an associated measurement instrument," *J. Oper. Manag. 11*, vol. 11, pp. 339–366, 1994.
[24] C. l. Kimberlin and almut G. Winterstein, "Validity and reliability of measurement instruments used in research Carole," *Res. Fundam.*, vol. 65, 2008, doi: 10.2146/ajhp070364.
[25] C. R. Reynolds, R. A. Altmann, and D. N. Allen, "Introduction to Psychological Assessment," *Mastering Mod. Psychol. Test.*, pp. 1–47, 2021, doi: 10.1007/978-3-030-59455-8_1.
[26] D. E. Huber, A. L. Cohen, and A. Staub, "A'compensatory selection' effect with standardized tests: Lack of correlation between test scores and success is evidence that test scores are predictive of success," *PLoS One*, vol. 17, no. 5 May, pp. 1–14, 2022, doi: 10.1371/journal.pone.0265459.
[27] E. et al. Alarcón-Bustamante, "From missing data to informative GPA predictions: Navigating selection process beliefs with the partial identifiability approach," *Br. J. Math. Stat. Psychol.*, 2025.
[28] S. A. Almalki, A. B. H. AlJameel, Z. Alghomlas, T. Alothman, and F. Alhajri, "Assessing the predictive validity of pre-admission criteria on dental students' academic performance: a cross-sectional study," *BMC Oral Health*, vol. 24, no. 1, pp. 1–16, 2024, doi: 10.1186/s12903-023-03839-6.
[29] M. et al. Jaehn, "Predictive validity of Admission tests and Educational Attainment on Preclinical Academic Performance-a Multisite Study," *BMC Med. Educ. J.*, 2025.
[30] S. Azwar, "Kualitas Tes Potensi Akademik Versi 07a," *J. Penelit. dan Eval. Pendidik.*, vol. 12, no. 2, 2008, doi: 10.21831/pep.v12i2.1429.
[31] I. I. Krisna, "Prediksi Tes Bakat Skolastik Terhadap Prestasi Belajar Siswa Sekolah Menengah Atas," *Indones. J. Educ. Assesment*, vol. 1, no. 1, 2018, doi: 10.26499/ijea.v1i1.5.
[32] G. H. Alnahdi, "Aptitude tests and successful college students: The predictive validity of the general aptitude test (GAT) in Saudi Arabia," *Int. Educ. Stud.*, vol. 8, no. 4, pp. 1–6, 2015, doi: 10.5539/ies.v8n4p1.
[33] R. Ferdhiana, I. Julita, A. Rusyana, and N. Salwa, "Hubungan Indeks Prestasi Kumulatif (IPK) dengan Nilai Ujian Akhir Nasional (UAN) : Studi Kasus di FMIPA Unsyiah," *Statistika*, vol. 15, no. 1, pp. 17–23, 2015.
[34] F. A. Abdelfattah, O. S. Obeidat, Y. A. Salahat, M. B. BinBakr, and A. A. Al Sultan, "The predictive validity of entrance scores and short-term performance for long-term success in engineering education," *J. Appl. Res. High. Educ.*, vol. 14, no. 4, pp. 1272–

1285, 2022, doi: 10.1108/JARHE-04-2021-0126.
[35] J. Baron and M. F. Norman, “SATs,ACHIEVEMENT TESTS,AND HIGH-SCHOOL CLASS RANK AS PREDICTORS OF COLLEGE PERFORMANCE,” *Educ. Psychol. Meas.*, vol. 9, no. 2, pp. 183–205, 1992.
[36] H. Stumpf and J. C. Stanley, “Group data on high school grade point averages and scores on academic aptitude tests as predictors of institutional graduation rates,” *Educ. Psychol. Meas.*, vol. 62, no. 6, pp. 1042–1052, 2002, doi: 10.1177/0013164402238091.
[37] W. S. Zimmerman, H. Parks, K. Gray, and W. B. Michael, “THE VALIDITY OF TRADITIONAL COGNITIVE MEASURES AND OF SCALES OF THE STUDY ATTITUDES AND METHODS SURVEY IN THE PREDICTION OF THE ACADEMIC SUCCESS OF EDUCATIONAL OPPORTUNITY PROGRAM STUDENTS,” pp. 465–470, 1977.
[38] A. Ali and U. Ali, “Predictability of engineering students’ performance at the University of Engineering and Technology, Peshawar from admission test conducted by educational testing and evaluation agency (ETEA), NWFP, Pakistan,” *Procedia - Soc. Behav. Sci.*, vol. 2, no. 2, pp. 976–982, 2010, doi: 10.1016/j.sbspro.2010.03.137.
[39] D. Dani and A. Sroyer, “Pengaruh Nilai Ujian Nasional Terhadap Indeks Prestasi Kumulatif Mahasiswa FMIPA Universitas Cenderawasih,” *J. Nas. Strukt.*, pp. 122–131, 2016, [Online]. Available: https://ejournal.uncen.ac.id/index.php/SM/article/view/310
[40] A. Almarabheh, M. H. Shehata, A. Ismaeel, H. Atwa, and A. Jaradat, “Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study,” 2022, *frontiersin.org*. doi: 10.3389/fmed.2022.971926.
[41] P. Jeffrey, R. Harris, and J. Sherman, “Quality improvement: A practical nursing program’s admission test,” *Nurse Educ. Today*, vol. 73, no. October 2018, pp. 65–70, 2019, doi: 10.1016/j.nedt.2018.11.016.
[42] H. Meyer, S. Zimmermann, J. Hissbach, D. Klusmann, and W. Hampe, “Selection and academic success of medical students in Hamburg, Germany,” *BMC Med. Educ.*, vol. 19, no. 1, pp. 1–15, 2019, doi: 10.1186/s12909-018-1443-4.
[43] J. H. Paris and R. Heiser, “An Examination of the Predictive Validity of a Measure of College Admissions Applicants’ Attributions of Success and Failure,” *J. Postsecond. Student Success*, vol. 1, no. 4, pp. 112–140, 2022, doi: 10.33009/fsop_jpss130205.
[44] J. C. Pérez-González, G. Filella, A. Soldevila, Y. Faiad, and M. J. Sanchez-Ruiz, “Integrating self-regulated learning and individual differences in the prediction of university academic achievement across a three-year-long degree,” *Metacognition Learn.*, vol. 17, no. 3, pp. 1141–1165, 2022, doi: 10.1007/s11409-022-09315-w.
[45] T. Donnon, E. O. Paolucci, and C. Violato, “The predictive validity of the MCAT for medical school performance and medical board licensing examinations: A meta-analysis of the published research,” *Acad. Med.*, vol. 82, no. 1, pp. 100–106, 2007, doi: 10.1097/01.ACM.0000249878.25186.b7.
[46] T. de Boer, F. Van Rijnsoever, and H. de Bresser, “Dear admission committee. . .: Which moves in application essays predict student master grades?,” *PLoS One*, vol. 19, no. 6 June, pp. 3–5, 2024, doi: 10.1371/journal.pone.0304394.
[47] D. Amelung, S. Zegota, L. Espe, T. Wittenberg, T. Raupach, and M. Kadmon, “Considering vocational training as selection criterion for medical students: evidence for predictive validity,” *Adv. Heal. Sci. Educ.*, vol. 27, no. 4, pp. 933–948, 2022, doi: 10.1007/s10459-022-10120-y.
[48] Q. Chen and Y. Zhang, “Development of Questionnaire on the Sense of Workplace Involution for Newly Recruited Employees and Its Relationship with Turnover

Intention," *Int. J. Environ. Res. Public Health*, vol. 19, no. 18, 2022, doi: 10.3390/ijerph191811218.

[49] A. Sjöberg and M. Grill, "A validity study of a work sample test of leadership behavior using supervisor and subordinate ratings as criteria," *Scand. J. Psychol.*, vol. 65, no. 1, pp. 129–135, 2024, doi: 10.1111/sjop.12954.

[50] A. M. Eskander, A. M. Abd-Elhameed, N. M. Osman, S. M. Magdy, and G. E. ElKess, "Role of diaphragmatic ultrasound in predicting weaning success from mechanical ventilation in pediatric intensive care unit," *Egypt. J. Radiol. Nucl. Med.*, vol. 55, no. 1, 2024, doi: 10.1186/s43055-024-01285-0.

[51] E. Jeong and S. J. Ireland, "Criterion-Related Validation of a Music-Based Attention Assessment for Individuals with Traumatic Brain Injury," *Int. J. Environ. Res. Public Health*, vol. 19, no. 23, 2022, doi: 10.3390/ijerph192316285.

[52] S. Miyagawa, T. Sato, and S. Maeda, "Demographic predictors of bedtime procrastination in the Japanese population," *Sleep Biol. Rhythms*, vol. 22, no. 2, pp. 199–206, 2024, doi: 10.1007/s41105-023-00508-7.

[53] L. Le Scouarnec *et al.*, "Development and assessment of algorithms for predicting brain amyloid positivity in a population without dementia," *Alzheimers. Res. Ther.*, vol. 16, no. 1, p. 219, 2024, doi: 10.1186/s13195-024-01595-5.

[54] X. Huang and H. Liu, "Criterion Validity of Functional Movement Screen as a Predictor of Sports Injury Risk in Chinese Police Staff," *Int. J. Environ. Res. Public Health*, vol. 19, no. 12, 2022, doi: 10.3390/ijerph19126992.

[55] A. Nyman, S. Strömbergsson, and A. Lohmander, "Canonical babbling ratio – Concurrent and predictive evaluation of the 0.15 criterion," *J. Commun. Disord.*, vol. 94, no. November, 2021, doi: 10.1016/j.jcomdis.2021.106164.

[56] K. Witkiewitz and E. Al, "Constructs derived from the addiction cycle predict alcohol use disorder treatment outcomes and recovery 3 years following treatment," *Psychol. Addict. Behav.*, vol. 14, no. 1, pp. 3–5, 2022, doi: 10.1037/0893-164x.14.1.3.

[57] A. Miranda, C. Berenguer, I. Baixauli, and B. Roselló, "Childhood language skills as predictors of social, adaptive and behavior outcomes of adolescents with autism spectrum disorder," *Res. Autism Spectr. Disord.*, vol. 103, no. September 2022, 2023, doi: 10.1016/j.rasd.2023.102143.

[58] Y. Maskov, J. M. Sanders, B. Tilahun, S. A. Hennessy, J. Reisch, and M. Johns, "Derivation and Validation of a Clinical Prediction Score to Identify the Isolation of Pseudomonas in Pneumonia," *Microbiol. Spectr.*, vol. 10, no. 3, 2022, doi: 10.1128/spectrum.00424-22.

[59] T. H. Skoglund, "The predictive validity of cognitive ability and personality tests toward police on-the-job training," *Int. J. Police Sci. Manag.*, vol. 26, no. 2, 2024.