Azim Premji University

The Interplay of Validity and Reliability in Educational Assessments

The Interplay of Validity and Reliability in Educational Assessments[i]

Any assessment represents a systematic sampling of a student's attainment of knowledge, capacities, values, and dispositions. It captures only a limited portion of what a student knows or can demonstrate. Consequently, the interpretation of assessment results involves making inferences from a student's observed performance on specific tasks to their overall competency in a subject, as illustrated in Figure 1. The specific tasks—denoted as Assessment 1, Assessment 2, and Assessment 3 in Figure 1—may include various forms of assessment, such as written or performance-based tasks.

For assessments to be effective, an assessment must be both valid and reliable.

- Validity refers to whether the test truly assesses what it is intended to assess.
- Reliability refers to the degree of consistency in results obtained across different sets of question papers, evaluators, assessment methods, and testing occasions.

Understanding validity and reliability is essential for educators, administrators, and policymakers to ensure that assessments are fair, meaningful, and reflective of student learning. This article aims to explore the concepts of validity and reliability, examine their significance in educational assessments, and discuss a checklist for ensuring both validity and reliability in assessments.

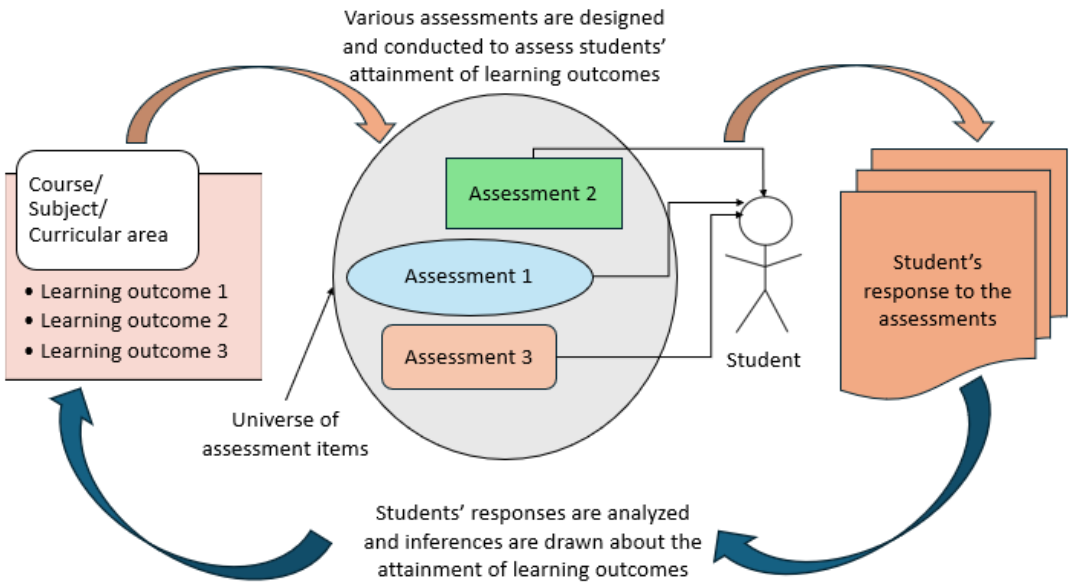


Figure 1: Ensuring the quality of assessments is crucial for obtaining useful and accurate evidence about students' learning levels

Concept of Validity

In common parlance, validity is often defined as the ability of an item or a test to "*assess what it intends to assess*." For instance, if you intend to assess a person's ability to drive, the test should genuinely assess whether the person can drive—and to what degree of proficiency. However, there are more sophisticated conceptualizations of the term validity.

In measurement theory, validity is defined as the extent to which the scores from a measure accurately represent the variable they are intended to measure [1]. In the context of education, the variable we often aim to assess is student learning, or more specifically, the knowledge, capacities, values, and dispositions a student demonstrates in relation to a predetermined syllabus and timeframe. This could include students' levels of learning in subjects such as Languages, Mathematics, Science, Social Science, or even in areas like Arts, Music, and Physical Fitness.

For an assessment to be valid, it must accurately reflect the construct it aims to assess. For instance, if a science test is intended to assess a student's ability to apply scientific reasoning but only includes factual recall questions like “What is the boiling point of water?”. In this case, students’ performance on such assessments will not accurately represent their attainment of the skill of scientific reasoning. Therefore, such assessment lacks validity. According to the Standards of Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], National Council for Measurement in Education [NCME], 2014) [1], there are five sources of validity- test content, response process, internal structure, relations to other variables and testing consequences. ***In this article, we will discuss validity evidence based on test content and response process.***

Test Content Validity

It refers to the *degree of alignment between the content of the assessment and the underlying construct it is intended to assess*. This includes examining whether the assessment tasks accurately and comprehensively reflect the Knowledge, Capacities, Values and Dispositions (KCVD) the test is designed to assess. A high-quality assessment should ensure that all relevant components of the construct are represented, while excluding elements that are not directly related to the construct. A test construct refers to the concept or characteristic that a test is specifically designed to assess. Examples of test constructs include scientific temper, collaboration, communication, and critical thinking. These constructs represent the underlying KCVD that an assessment aims to assess through carefully designed tasks or questions.

1. ***Construct underrepresentation*** occurs when an assessment fails to include key elements of the intended construct. This can lead to an incomplete or misleading assessment of a student's competencies. For example, a reading comprehension test that only includes vocabulary questions and omits inference or main idea questions underrepresents the

broader construct of reading comprehension. Scores from such a test may not fully reflect a student's true ability in the intended domain, thus limiting the validity of the assessment.

2. ***Construct irrelevance*** refers to the extent to which performance on an assessment is influenced by factors that are unrelated to the construct being assessed. These extraneous elements can distort the interpretation of test results. Examples include the use of unnecessarily complex language in math word problems or the inclusion of culturally biased content that disadvantages certain groups. When irrelevant variables influence test performance, it becomes difficult to determine whether scores accurately reflect a student's ability in the intended construct, undermining fairness and validity.

Response Process Validity

It indicates the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers. In other words, it assesses whether the test elicits the kind of thinking or performance that aligns with the construct. For example, if a science assessment aims to assess scientific reasoning but students are instead using test-taking strategies (Solving old question papers, practicing textbook questions, familiarization with test format, etc.) or guessing due to unclear questions, then the response process does not align with the intended construct. Similarly, if a test is designed to assess mathematical reasoning, it is essential to determine whether test takers are genuinely engaging with the material through reasoning, rather than merely applying standard algorithms. Discrepancies between intended and actual response processes can compromise the validity of the assessment and may signal the need to revise item format, assessment item or marking scheme.

In subsequent sub-sections, the issues of validity that arise due to lack of test content and response process validity evidence are described in detail.

Issues of Validity

Is it possible to measure learning?

One of the fundamental questions in education is whether learning itself can truly be measured. Learning is often seen as a dynamic, multifaceted process involving not only knowledge acquisition but also the development of capacities, values and dispositions. While assessments can measure specific content knowledge or capacities at a given time, they may not fully capture a student's broader or long-term learning, such as their critical thinking, creativity, or aesthetics appreciation. Additionally, learning is influenced by a range of factors, including personal experiences, environmental context, student specific variables and social influences, which can complicate how effectively we measure it. As a result, assessments might fail to capture the full extent of a student's learning or fail to measure the outcomes that are not directly observable and assessable.

In such a context—where learning is recognized as complex, multi-dimensional, and influenced by numerous internal and external factors—the validity of an assessment becomes a relative concept rather than an absolute one. It cannot be assumed that a single assessment can capture all facets of student learning. For instance, a written test may validly assess a student's ability to solve algebraic equations, but it may not validly assess their collaborative skills, or creative expression or how to apply the knowledge of algebra in a real-life situation. These latter dimensions may require alternative modes of assessment—such as observations, portfolios, or performance-based tasks. Thus, in practical terms, the validity of any assessment is bounded by what is being assessed, how it is being assessed, and the purpose of the assessment. Rather than claiming that an assessment is universally valid, it is more accurate to say that it is valid for assessing a specific construct, in a specific context, and for a specific purpose. This nuanced understanding encourages educators and assessment designers to be critically reflective about the tools they use, to clarify what their assessments aim to capture, and to recognize what they might unintentionally leave out.

Defining scope of content and its representation

Any assessment should fairly and accurately represent the full scope of learning as outlined in the curriculum to ensure test content validity. This means it should include a balanced sampling of the key competencies that students are expected to attain, rather than focusing disproportionately on certain competencies or neglecting others. When an assessment fails to achieve this balance, it compromises its content validity—that is, the extent to which it accurately reflects the intended competencies.

For example, as shown in Table 1, across all subjects, a large percentage of questions target lower order thinking skills, and there is a disproportionately high number of questions drawn directly from the textbook. Such a skewed blueprint limits the assessment of subject-specific competencies, thereby undermining the overall validity of the assessment.

To ensure content validity, a balanced test blueprint or comprehensive assessment framework is an essential tool. A blueprint acts as a planning document that maps out the structure of the assessment in advance. It guides the distribution of test items across different content areas (such as algebra, geometry, or statistics in mathematics), cognitive levels (such as remember, understand and application), and competencies (as specified in the curriculum). For example, in a science assessment, the blueprint might allocate 30% of the questions to conceptual understanding, 40% to application, and 30% to analysis and reasoning. It may also ensure proportional representation across various units like biology, physics, and chemistry. This structured approach helps assessment designers avoid overemphasizing topics that are easier to assess or more familiar to the test developers and instead maintain alignment with curricular goals. By using a blueprint, educators can create assessments that are not only valid but also transparent, equitable, and purposeful—providing students with a fair opportunity to demonstrate their learning across the full spectrum of what they are expected to know and do.

Cognitive Levels	English	Hindi	Math	Science	Social Science
Remember	43%	15%	32%	38%	69%
Understand	16%	19%	26%	52%	31%
Apply	25%	28%	40%	4%	-
Analyse	10%	19%	2%	2%	-
Evaluate	0%	2%	-	4%	-
Create	6%	17%	-	-	-
% questions from the textbook/ Previous year question papers	10%	17%	43%	9.4%	11%

Table 1: Distribution of questions across different cognitive levels [2]

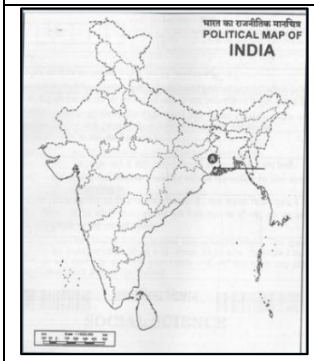
Quality Issues in Assessment items

When assessment items do not adequately cover the desired knowledge, capacities, values, and dispositions outlined in the curriculum, the assessment underrepresents the intended competencies, thereby failing to assess what it is meant to assess. Similarly, poor item design—such as ambiguous wording, culturally biased content, or irrelevant questions—can confuse students or cause them to misunderstand what is being assessed. As a result, the assessment may end up measuring lower order thinking skills or test-taking ability rather than the intended construct (e.g., critical thinking or scientific reasoning).

In the case of multiple-choice questions, if the distractors are obviously incorrect, students can easily eliminate them and may guess the correct answer without genuinely attaining the competency. This makes the item too easy and fails to differentiate between students who have mastered the competencies and those who have not.

In each of these scenarios—whether it is the underrepresentation of key competencies (Figure 4), poorly designed assessment items (Figure 2), or the use of implausible distractors (Figure 3)—the validity of the assessment is severely compromised.

This item aims to assess students' ability to locate geographical features—such as countries, states, physical landforms, water bodies, and important cities—on a map, using spatial understanding, directional skills, and appropriate map conventions. However, the item only assesses students' ability to locate precise places based on a few leading questions, thereby limiting the depth of the intended assessment of the competency.



SECTION – E

(Map based questions) 1 + 2 = 3

13.1 On the given outline Political Map of India, identify the place marked as 'A' with the help of following information and write its correct name on the line marked near it :

(A) The place where Indian National Congress Session was held in September, 1920. 1

13.2 On the same given Map of India, locate and label the following :

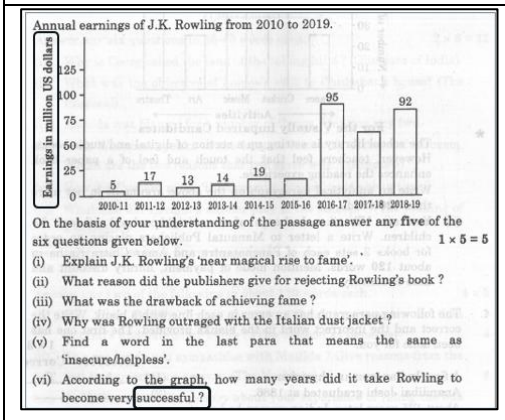
I. (a) Ramagundam Thermal Plant 1

OR

(b) Pune Software Technology Park. 1

II. Chennai (Meenambakam) International Airport. 1

Figure 2: Sample assessment item 1 [3]

The graph (stimulus) is open to interpretation, as the question relates to the author becoming a successful person. The question assumes that maximum earnings are an indicator of success. A few questions are not based on the graph. Also, the graph used in the text could follow the Indian currency system.	The correct answer to this question is recycling. All three distractors are implausible. Chemical treatment, which is more suited for hazardous waste, and composting, which is suitable for biodegradable waste, would serve as better distractors. Additionally, "burying" and "dumping" can be replaced with "landfilling." One of the best solutions to manage non-biodegradable wastes is (A) burning (B) dumping (C) burying (D) recycling.
	Option B is the expected answer; however, options A, C, and D also imply the same as option B. Therefore, all options can be considered correct in this context. The pair of equations $kx + 2y = 5$ and $3x + y = 1$ will have unique solution if a) k = 0 b) k ≠ 6 c) k = 2 d) k = 3 The distractors can be- k = 6 (Takes the condition for unique solution as = instead of ≠); k ≠ 0 (Confuses the condition for consistency with the condition for leading coefficients) and k is any natural number (Does not know the conditions for consistency)
Figure 3: Sample assessment item 2 [3]	***Figure 4: Sample assessment item 3 & 4 [3]***

Concept of Reliability

Reliability *refers to the dependability of test results*. The test results will be more dependable only when they are consistent and uniform across various instances [1]. These various instances are depicted in Figure 5. In practical terms, if a student consistently answers a range of addition problems correctly (e.g., “4 + 9,” “6 + 7,” “8 + 5”), a teacher can reliably infer that the student has developed the skill of adding two single-digit numbers. Similarly, in physical education, if a student consistently performs well in endurance tasks (like running, jumping, and aerobic exercises), the test is likely reliable in measuring their physical fitness. ***There can be different types of reliability as follows:***

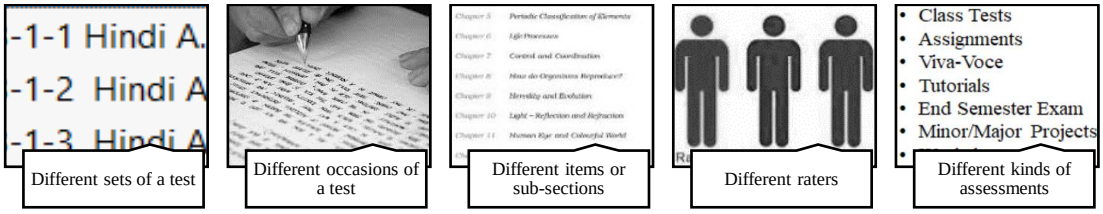


Figure 5: Types of reliability

1. ***Different sets of a test***- This involves administering two different but equivalent (parallel) forms of the same test to the same group of students. The results are then compared to determine whether the two forms yield consistent results.
2. ***Different occasions of a test***- This type of reliability assesses the consistency of measurements when the same test is administered to the same group of students at two different points in time. A high test-retest reliability indicates that the test produces similar results when given on two separate occasions.
3. ***Different items or sub-sections***- This involves assessing the consistency of responses within a single test or assessment. It is typically measured using techniques like Cronbach's alpha [4], which quantifies the extent to which items within a test measure the same underlying construct. This coefficient estimates the average correlation among all items in the test, providing a numerical value—typically ranging from 0 to 1—that reflects the extent to which the items are interrelated. A higher Cronbach's alpha value (generally above 0.70) suggests strong internal consistency, implying that the items reliably assess the intended construct. Conversely, a low alpha value may indicate that some items are not aligned with the overall purpose of the assessment, potentially compromising the reliability of the test.
4. ***Different raters***- This refers to the consistency of measurements when different raters evaluate the same set of students' responses. It is often used in situations where subjective judgment is involved, such as scoring constructed response questions (open-ended questions).
5. ***Different kinds of assessments***- Classroom assessment comprises of a series of informal and formal assessments including observation, projects, portfolio, debate, and exhibition. Reliability in classroom context is understood in terms of consistency and uniformity properties. Consistency is ensured when there is uniformity in students' performance across multiple methods of assessments.

An important question here to ask is whether the measures of reliability of scores in terms on different sets of tests, raters, occasions, sub-sections, and types apply similarly for all purposes of assessment that includes classroom-based assessment, entrance examination, psychometric test, board exam and large-scale assessments. This will be clear through following examples-

- Test-retest reliability fundamentally contradicts the idea of classroom-based assessment where it is expected that students' performance will remain consistent over two occasions of the same test. In a classroom set up, students learning is seen in a continuum and a teacher will want their students to progress towards the desired competency over a period of time. On the other hand, test-retest reliability will be more appropriate for psychometric tests.

- Cronbach's alpha measures how consistently items in a test reflect the construct. It looks at the variation in responses—greater variation among items generally means better reliability. If all students answer a question either correctly or incorrectly, it does not add to this variation and does not help improve the test's reliability. However, in a classroom setting, such questions can still be valuable for identifying learning gaps or misunderstandings. Cronbach's alpha is especially useful for measuring reliability in large-scale assessments and selection tests.

Therefore, it is important to understand the purpose of a test to infer the extent to which test scores are dependable. Also, it is important to note that the test scores depend on various factors such as quality of assessment instrument, the way the test is administered, and the approach used for evaluating the students' responses. Any inference on the reliability of assessment based on students' scores will be inappropriate without closely looking into these factors.

Figure 6 shows all the processes involved in assessment design, administration, and reporting. The process of designing assessment is influenced by various systematic and random errors. Systematic errors occur due to flaws in the measuring instrument. In this case, such systematic errors can be in the form of errors in either the assessment items (unambiguous item stem, implausible distractors for MCQs, unclear stimulus, etc.) or the marking schemes (inadequate definition of parameters of evaluation and their definition). Random errors are unpredictable fluctuations in measurements that occur due to various uncontrollable factors. In this context, these arise due to learner specific factors (students' motivation, concentration, and fatigue) and assessment environment specific factors (poor lightning, noise distractions and uncomfortable examination room temperature).

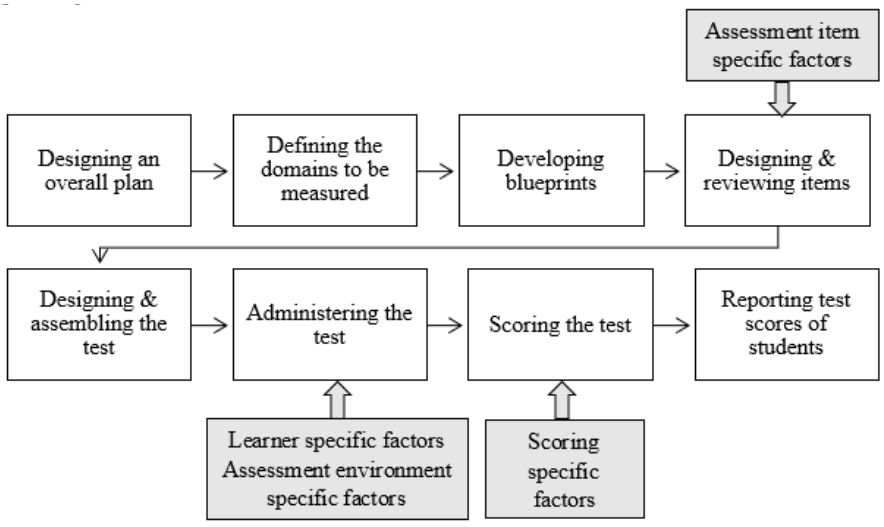


Figure 6: Processes in assessment design, administration, and reporting

<u>In subsequent sub-sections, the issues of reliability that arise due to these random and systematic errors are described in detail.</u>

Issues of Reliability

Teacher's explicit or implicit bias

For a teacher, it is important to base her judgements and decisions about students on assessment results that are dependable. A wide range of assessments will provide more sources of information, with demonstrable evidence for validity and more dependable decisions can be made. This also ensures providing opportunity to students to best demonstrate their achievement of competencies through varied methods of assessment without considering teacher's explicit or implicit bias. An example of multiple methods of assessment used for competency is shown in Figure 7.
Reliability across multiple methods of assessment refers to the consistency of assessment outcomes when different methods are used to assess the same construct. In other words, it assesses how well different methods agree with each other in providing consistent results. High reliability suggests that the methods are capturing the true underlying construct being assessed, while low reliability indicates inconsistency and uncertainty in the assessments.

Class 5 EVS Competency- Describes the interdependence among animals, plants and humans.	**Group Discussion:** Students share real-life examples of how living beings are interconnected.
	Worksheet: Includes drawings or short answers on ecosystem interdependence.
	Role Play: Students act out roles of animals, plants, and humans to explore their relationships.

Figure 7: Multiple methods of assessment for assessing a competency

Gathering evidences of students learning through just one method of assessment may provide misleading results about students learning, thereby causing a threat to reliability. Both the nature of competency and the cohort of students do not lend themselves for one-time assessment. Addressing the uniqueness of students requires a diverse range of assessment methods that consider individual strengths, learning styles, interests, and developmental stages. Similarly, holistic assessment of competencies is crucial for gaining a comprehensive understanding of student's knowledge, capacities, values and dispositions.

Inaccurate assessment items

Students' ability to respond to a question not only depends on the extent of attainment of the desired competencies but also how well questions and tasks are phrased. Ambiguous or complex phrasing and inaccessible language can confuse students leading to inaccurate responses (example in Figure 8). Similarly, implausible distractors in the case of multiple-choice questions help students to deduce the answer rather than providing the response through the intended cognitive processes (example in Figure 9). In such a scenario, the probability of getting accurate evidences of students learning is always uncertain. Since the trigger chosen for receiving students' response itself is faulty,

one can never be sure about the consistency in assessment results. This is analogous to measuring something using a measuring tape that is made up of elastic material, the measurement here might be inconsistent because of the non-standard nature of the measuring tool.

This question is ambiguous and be rephrased in following manner- There are 12 fruits in a basket, out of which 4 are mangoes and the rest are oranges. A boy picked up 3 fruits at random and the outcome of his choice is not known. Find the mean and variance of the number of oranges which is possibly left in the basket.	Question stem and distractors have matching words that give clues. Distractors are sub-sets of each other.
31) There are 12 fruits in a basket, out of which 4 are mangoes and the rest are oranges. A boy picks 3 fruits at a random from the basket. Find the mean and variance of the number of oranges.	7. Who were called 'Chapmen'? A. Book seller B. Paper seller C. Workers of printing press D. Seller of 'penny chap books'
Figure 8: Ambiguous item stem [5]	***Figure 9: MCQ with implausible distractors [3]***

Subjectivity in evaluation

Constructed response questions allow students to provide responses that are open ended and qualitative in nature. However, the interpretation of these responses can vary widely among evaluators, as it often depends on their individual perspectives, experiences, and expectations. Such discrepancy in interpretations inhibits the dependability of test results as more than one evaluator will grade the same response differently. Similarly, a single evaluator may provide inconsistent grading for the same response on two different occasions due to lack of clearly stated criteria of assessment as discussed in Figure 10. Such discrepancies are stated as interrater and intrarater reliability issues respectively.

Ans. ANALYTICAL PARAGRAPH **Note: Analysis to be based on the given input only.** **No extra credit to be awarded for any additional information to the given content.** **Content - 2 Marks** **Analysis – 2 Marks** **Expression - 1 Mark**	The assessment criteria, including presentation, content, and language, should be clearly defined, and the descriptions for each level must be explicitly stated.

Figure 10: Superficial marking scheme [3]

Such scoring specific discrepancies can be controlled through setting the marking scheme in advance which clearly defines the criteria, and their descriptions related to different levels of attainment for each criterion. In the context of large-scale assessments such as board exams, it becomes crucial to train the evaluators in using the marking schemes and rescoring a sample of scripts by multiple

evaluators. Marking schemes should be clear, comprehensive, detailed and also provide scope for alternative responses.

Interplay of Validity and Reliability

Validity and reliability are the two cornerstones of high-quality assessment. While each has its own distinct definition and purpose, they are deeply interconnected and often overlap in both theory and practice. Understanding this overlap is essential for designing meaningful assessments.
Another important dimension of assessment quality is fairness, which refers to the principle that all students should have an equal and unobstructed opportunity to demonstrate their performance on the construct(s) being assessed [1]. Universal design emphasizes the need to develop tests that are as accessible and usable as possible for all test takers within the intended population, regardless of characteristics such as gender, age, language background, culture, socioeconomic status, or disability. This approach requires careful consideration at every stage of test development and implementation [1].

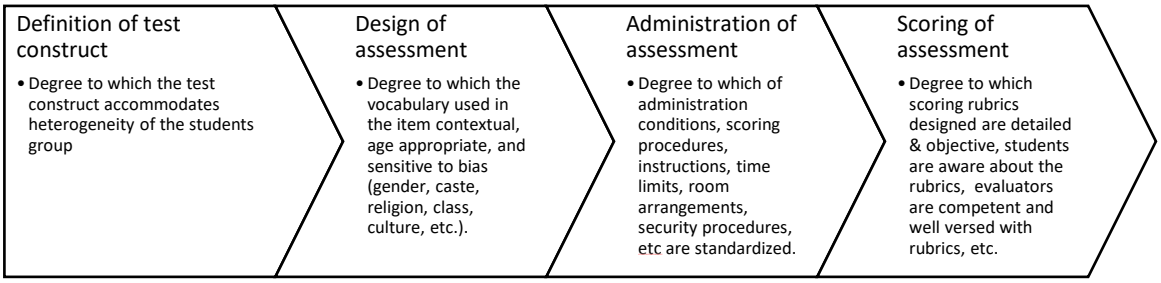


Figure 11. Ensuring fairness across all processes of assessment development

A few examples explaining the interlinkage of Validity, Reliability, and Fairness:

- Using ***multiple methods of assessment***, such as written tests, projects, presentations, and observations, helps ensure content validity by capturing a wider range of knowledge, comprehension, values, and dispositions. This ensures the assessment reflects what students actually know and can do, thereby reducing construct underrepresentation. It supports reliability by balancing out inconsistencies that may arise from relying on a single method and enhances fairness by giving all students an equal opportunity to demonstrate their learning in ways that align with their strengths and needs.
- A ***blueprint*** enhances validity by ensuring the assessment covers the intended learning outcomes—without underrepresenting or overemphasizing any part of the curriculum. It contributes to reliability by bringing consistency to the design process and helping to balance item difficulty and types across different versions or sections of the test. Additionally, a blueprint supports fairness by ensuring all students are assessed on content they were taught and given equal opportunities to demonstrate their understanding. Blueprint can also support fairness by providing an informed balance of questions belonging

to lower order thinking skills and higher order thinking skills to ensure fair chance to maximum number of test takers.

- Clearly articulated ***assessment tasks or questions***, when written in a clear, age-appropriate, and unambiguous manner, are more likely to assess the intended KCVD, minimizing the influence of construct-irrelevant factors. This improves validity by ensuring the assessment truly reflects the target construct. Such clarity also enhances reliability by reducing confusion and promoting consistent responses across students. Moreover, fairness is strengthened when students from diverse backgrounds can easily understand what is being asked, preventing language or cultural differences from becoming barriers to performance.
- ***Rubrics*** with clearly defined criteria and performance levels or comprehensive and detailed marking schemes increase response process validity by aligning scoring closely with the intended competencies. They enhance reliability by providing a consistent framework for assessment, thereby reducing subjectivity and variability among assessors. At the same time, rubrics support fairness by making expectations transparent and eliminating bias or subjectivity in the evaluation process.

References

1. AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.
2. Azim Premji University. (2020). A Study of Class X Board Exams: CBSE, 2020. CBSE & Assessment group.
3. CBSE Previous Year Question Papers. Retrieved from https://www.cbse.gov.in/cbsenew/question-paper.html
4. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International journal of medical education, 2, 53.
5. Department of School Education (Pre-university) Previous Year Question Papers. Retrieved from https://pue.karnataka.gov.in/english

Annexure: Review Parameters for Assessment Items and Marking Scheme for Ensuring Validity & Reliability

Item Stem

1. Is the item assessing the desired curricular goals, competencies and learning outcomes particular subject?
2. Is the item aligned to the test construct?
3. Is the information in the stem factually and conceptually accurate?
4. Does the stem avoid complex, ambiguous, and/or tricky language?
5. Is the stem free of grammatical errors?
6. Does the stem use negative phrase only when necessary?

MCQ Distractors

1. Are distractors based on reasonable misconceptions and errors? Are distractors plausible?
2. Is there one, and only one, clearly correct answer?
3. Are options independent of each other? Are there no options with the same meaning?
4. Are options similar in structure in terms of degree of specificity, and/or length?
5. Is the correct answer not clued by the item stem, such as absolutes or words repeated in both the stem and options?

Stimulus

1. Is the stimulus necessary, relevant, and useful to answer the question?
2. Is the stimulus likely to be interesting/engaging to students? Is it age appropriate?
3. Is the stimulus clear, accurate and sufficient to answer the item?
4. Does the stimulus contain appropriate and accurate labels?
5. Is the stimulus significantly free from copyright issues?

Item Bias and Sensitivity Issues

1. Is the item inclusive and representative in nature?
2. Does the item avoid reiterating or representing anything regressive in nature?
3. Does the item contain controversial information or any misinformation which can be contested in a socio-political, cultural, or historical context?
4. Does the item reflect current social ethos and progressiveness that we have collectively achieved as a society?

Marking scheme

1. Does the marking scheme also take care of alternate answer/s?
2. Is the weightage of marks proportional to the volume of the content and the time taken by the student to respond to a question?
3. Are clear-cut directions given to reduce the subjectivity in the marking scheme?
4. Is there a minimum value point given in the marking scheme for a particular question?
5. Is there alignment between the cognitive process expected to attempt an item and cognitive process expected in the solution as provided in the marking scheme?

[i] This article is co-authored by Shilpi Banerjee and Aanchal Chomal. They work in the School of Continuing Education at Azim Premji University. They can be reached at shilpi.banerjee@azimpremjifoundation.org and aanchal@azimpremjifoundation.org.
This article can be cited as-
The Interplay of Validity and Reliability in Educational Assessments, Assessment resources, 2025, Azim Premji University

The Interplay of Validity and Reliability in Educational Assessments