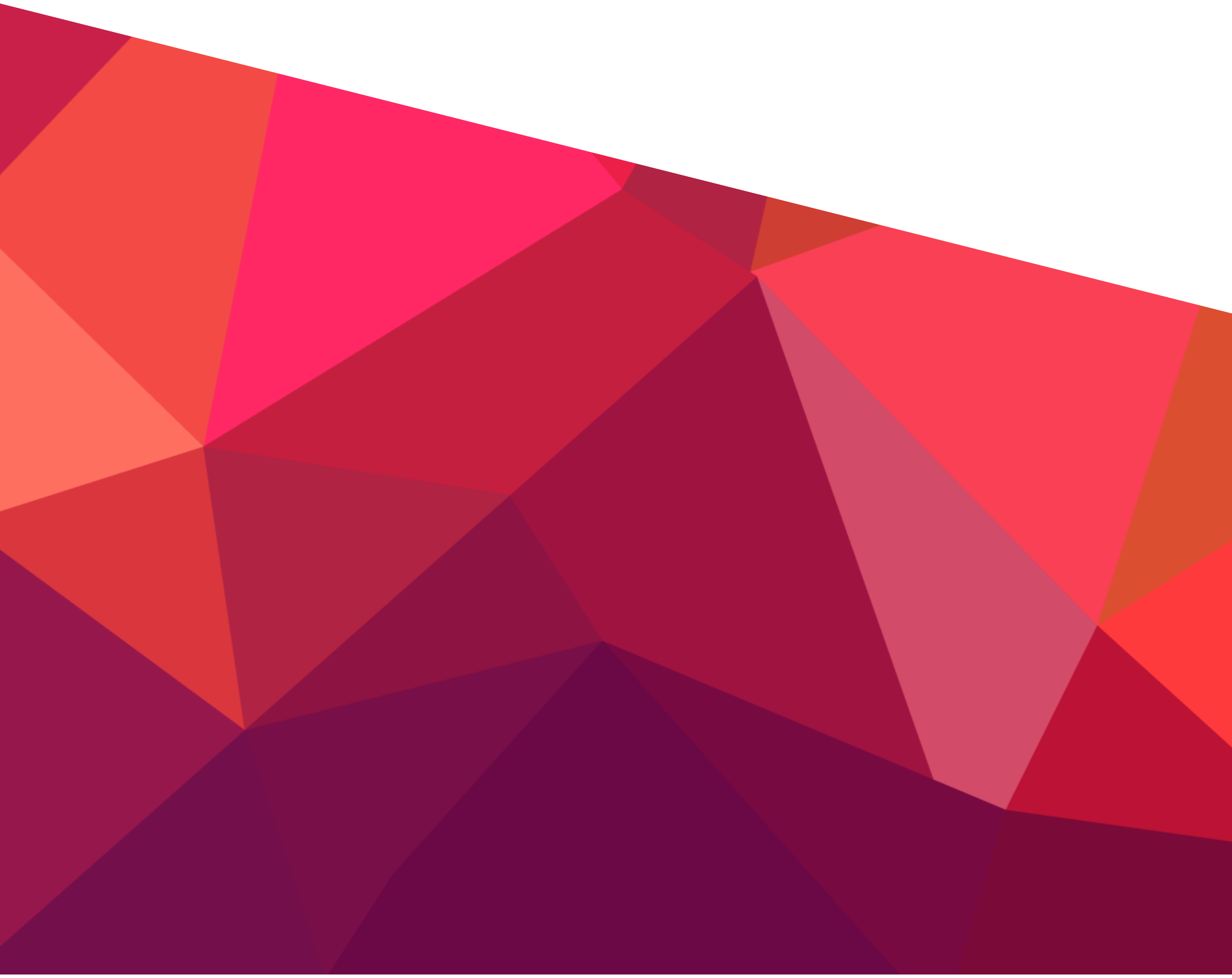
A COMPREHENSIVE GUIDE TO CREATING RELIABLE, RELEVANT AND ROBUST ASSESSMENTS

Conceptualization, Creation and Delivery of the Modern Day Assessment



mettl

CONTENT

A COMPREHENSIVE GUIDE TO CREATING RELIABLE, RELEVANT AND ROBUST ASSESSMENTS

01 SECTION

IN THIS eBook

In our interactions with clients, we've realized that there's a shroud of doubt over the world of assessments and their creation. An informed buyer and a credible vendor go a long way in creating the right ecosystem for a growing talent pool. Assessments are still largely considered simply a bunch of questions, with little visibility into its objectives and purpose. With this eBook we intend to shed some light on this little known, but infinitely impactful, world of assessments.

02 SECTION

WHY THIS eBook

assessment,

noun [U]

the process of making a judgment or forming an opinion, after considering something or someone carefully

Let’s break this definition down and analyze further. To ensure that the judgment or opinion is sound and objective, and is an accurate reflection of ‘something’ or ‘someone’, the assessment process itself needs to be carefully decided. The problem of devising a good assessment, then, boils down to adopting the right process, which is further dependent on what needs to be considered about the ‘someone’ or ‘something’ – what skills, competencies, and abilities are to be considered in forming an opinion about the subject.

For the purpose of this eBook, we’ll only be considering assessments that are made for individuals, in a professional or academic setting.

03 SECTION THE PURPOSE OF ASSESSMENTS

We've all been taking assessments, of different shapes and sizes, since our childhood. Some Herman, Aschbacher, and Winters (1992) point out that "People perform better when they know the goal, see models, know how their

performance compares to the standard." Assessments are designed to provide an objective evaluation of skills or competencies of a test taker that can then serve as a reliable reflection of the test taker's ability. In a professional or academic setting, comparative assessments can be either of two types:

a) Norm-referenced assessments, commonly understood as the percentile system, exposes the relative standing of candidatesacross the measured competencies

b) Criterion-referenced assessments pit test takers against a pre-determined benchmark.

The choice of assessment type naturally depends on the objective of the exercise. In either form, an assessment is required to highlight the differences in performances of test takers.

When to use criterion referenced assessment and when to use norm-referenced?

According to Popham, J. W. (1975). Educational evaluation. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., Criterion referenced assessments are useful to determine whether each test taker has achieved specific skills or competencies, presumably after some training on those skills or competencies. Norm-referenced assessments, on the other hand, would be more useful in a professional setting where

the objective is to rank test takers relative to others in the test pool. It helps to discriminate

between high and low achievers.

04 SECTION
ASSESSMENTS VS GOOD ASSESSMENTS

Unfortunately, it's never really possible to make a completely accurate judgment or opinion about someone, since there are factors that are out of our control. Assessments, at best, probe students on a sample set of question or problems that will require the use of skills or competencies that are to be measured. The responses are then analyzed to arrive at an inference about their ability, either against a benchmarked score, or relative to the performance of other test takers in the pool. Even such measurements are bound to be somewhat inaccurate due to 'measurement errors', or variations in human performance that are not in our control.

A good assessment – one that does the above with as little error as possible, requires certain qualities.

One that does the above with as little error as possible requires certain qualities.

Variance

A fair assumption that rests at the foundation of all assessment exercises is that people have different levels of skills or competencies. For norm-referenced assessments in particular, discriminating between test takers on the measured criteria is vital. If a test is too tough, most test takers would score low, and if it is too easy, most test takers would score high, thus making discrimination difficult. For this reason, good assessments shouldn't either be too easy or too tough.

Reliability and Validity

Reliability is an indicator of the ability of an assessment to produce consistent results, over time as well as within the assessment, among the different factors tested. Naturally, a high reliability score means that the test produces results that are consistent over time.

However, it is important to understand that high reliability does not by itself mean that the test gives the right results. Accuracy of a test, or how exactly the test measures what needs to be measured, is reflected by the validity score of a test. To understand reliability and validity better, let's look at the following example.

A weighing scale is to be tested for validity and reliability. A pre-weighed block of 100 Kgs is used for the test.

Case 1

The weighing scale reads 70 Kgs each of the 5 times that the block is weighed. This shows that the weighing scale is perfectly reliable, since it is consistent in the measurement that is made over the 5 measurement instances. However, the weighing scale is not valid since it does not accurately measure the weight of the block.

RELIABLE, NOT VALID

A perfectly reliable assessment might not necessarily be a perfectly valid assessment.

RELIABLE, VALID

Case 2

Let's assume the scale reports the following readings when the block is measured 5 times.

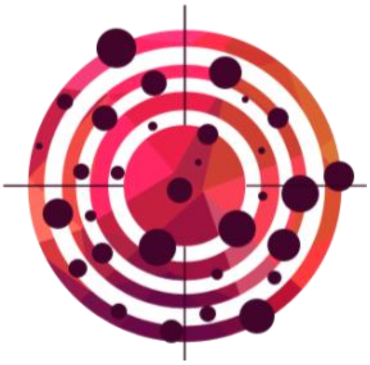
98 Kgs, 99 Kgs, 100 Kgs, 101 Kgs, 102 Kgs.

Such a weighing scale is not perfectly reliable since it is not consistent across measurement instances, but it is reasonably valid since the measurements lie in an acceptable range of the right measurement – 100 Kgs.

UNRELIABLE, VALID

Case 3

If the scale reads 100 Kg for each of the 5 instances, then the scale is both perfectly reliable & valid. In simple terms, reliability refers to the precision of an assessment while validity points towards the accuracy of the test.



UNRELIABLE, NOT VALID

Integrity and Transparency

Since test items are created by humans and there is always a possibility for personal biases to creep in, during the development of a test, multiple subject matter experts should be involved in the creation.

Once the test is developed, the content and scoring system needs to be reviewed by independent technical reviewers to ensure test efficacy.

Standardization

Post development, the test needs to be administered to a representative sample size to understand the typical scores that are obtained. It also helps in setting benchmarks that can then be used to judge test takers against.

Security

Content security is one of the biggest concerns in test development. Since tests might be administered over the course of a few days, there is a distinct possibility for questions to get leaked and be used to gain an unfair advantage over initial test takers. Tests need to contain a healthy mix of questions from different question banks to ensure that test takers get different versions of the same test, while maintaining the validity and reliability of the test. Another method to ensure test security is by using adaptive testing, where test takers are administered tests on the basis of individual aptitude/skill, ensuring that no two assessments are alike.

05 SECTION DESIGNING GOOD ASSESSMENTS

Now that we have identified the qualities a good assessment should possess, we can Look at the steps involved in creating such an assessment. Understandably, the efficacy of the assessment depends on the attention given to it during its creation. To standardize the process and lend assessments the credibility they require, the design process is broken down into steps, each of which focuses on a specific aspect of the assessment. Arguably, higher the number of steps involved in its creation, more is the apparent focus given to these specific aspects. Here below, we outline the 9 step process that is followed by Mettl while designing an assessment.

Step 1 - Requirement Gathering

This is the first phase in the assessment development process. Assessment team receives/gathers prerequisites for an assessment through one or more of the following means

- ***Market Analysis*** – The market refers to external stakeholders. Let's consider the recruitment market. Umpteen number of different profiles are being recruited for, and the internal team gauges the demand for each of these profiles. Based on the criticality of profiles, the internal assessment team goes about deciding which profile to create assessments for.

- ***Client Interactions*** – Often, clients already have an idea about what they want to test candidates on. In such cases, the assessment team interacts with the client to understand the requirement in-depth and sits down to devise a good assessment that will cater to these particular requirements.

- ***Request from Internal stakeholders*** - The assessment team also receives requirements from internal stakeholders, like the sales team or marketing team, who are in direct touch with clients or prospective clients, and might have a better inkling about their pulse and potential requirements.

Once the requirements are received, they are analyzed by studying objectives, the target audience, and the assessment criteria. Typically, clients already have an idea of the criteria that the assessment should be based on. But in cases when such criteria are not defined upfront, the assessment team needs to work with the client to understand the requirements and then decide on the criteria.

Step 2 - SME & ITR Acquisitions

This is the 2nd phase in the assessment development process where internal expertise (author) for authoring/creating the assessments needs to be evaluated. If such expertise is not available internally, Job Descriptions (JDs) for external hiring of Subject Matter Experts (SMEs) and Independent Technical Reviewers (ITRs) are created. This has to be followed by acquiring a list of potentials for SMEs and ITRs. Candidates are shortlisted and a pool of SMEs and ITRs for specific job roles is created.

Step 3 - Develop Assessment Design Framework/Blueprint

Assessments manager and the in-house design team are involved in the development of Assessment Framework/Blueprint based on the requirements obtained as described before. The design would involve decision to be made on the following parameters:

- Duration of the test
- Total questions/items per assessment
- Item/question distribution:

Section-wise analysis of the assessments	Topic-wise allocation of questions in an assessments	Difficulty levels to be used in the assessment	Distribution of marks based on the weightage for each difficulty level

Defining Difficulty Levels – Bloom's Taxonomy

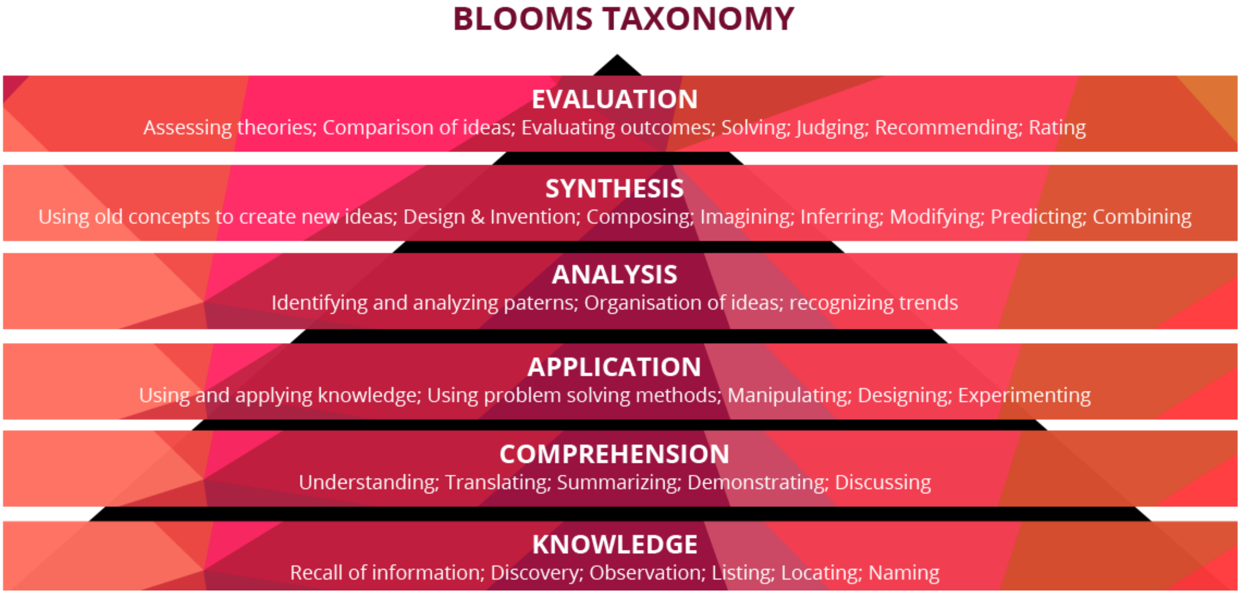
Questions/items in an assessment require different levels of thinking skills. Bloom's taxonomy, named after Benjamin Bloom, who chaired the committee of educators that devised it, classifies the different objectives that educators set for students. It divides educational objectives into three domains – cognitive ("learning/knowledge"), affective ("feeling/heart") and psychomotor ("doing/hands"). Within each of these domains, learning capability is further classified according to the skill level required. For instance, you'll see that within the cognitive domain, 'application' is a learning level that presupposes prior "knowledge".

Bloom's taxonomy is divided into two levels

Lower Order Thinking Skills (LOTS)	Higher Order Thinking Skills (HOTS)
Contains knowledge/recall, comprehension/understanding and application. For entry level profiles, LOTS are considered.	Contains application, analysis, synthesis, and evaluation. For intermediate and advanced levels of profiles, LOTS and HOTS both are considered.

At Mettl, for the purpose of assessment creation, our assessments team sticks to the Cognitive domain hierarchy of Bloom's taxonomy. Questions/items are classified according to the learning skill expected to answer them. Depending on who the test is meant for, and what learning skills are to be tested, question items are pulled from the data bank to put together the right assessment.

The following figure depicts the classification for better understanding.


BLOOMS TAXONOMY
EVALUATION
Assessing theories; Comparison of ideas; Evaluating outcomes; Solving; Judging; Recommending; Rating
SYNTHESIS
Using old concepts to create new ideas; Design & Invention; Composing; Imagining; Inferring; Modifying; Predicting; Combining
ANALYSIS
Identifying and analyzing paterns; Organisation of ideas; recognizing trends
APPLICATION
Using and applying knowledge; Using problem solving methods; Manipulating; Designing; Experimenting
COMPREHENSION
Understanding; Translating; Summarizing; Demonstrating; Discussing
KNOWLEDGE
Recall of information; Discovery; Observation; Listing; Locating; Naming

Step 4 - Define Assessment Methodology

This is a brainstorming phase where the author/SME, Design Contact, and Assessment Manager together decide on the item/question patterns to be set. The session tries to answer the “WHAT” of question assessment methodology. Inputs from clients are incorporated and based on the assessment blueprint; the cognitive domain of Bloom’s taxonomy is applied in deciding on the question types and the media to be used. Question types could be any or a combination of the following:

Simulation/Hands-on Based (Coding, Query Writing, Case Analysis etc)	Scenario based
Media based (Images/Video)	Fill in the blanks (FITB)
Comprehension	Logical reasoning
Chronological sequencing	Matching the columns

Step 5 - Assessment Development

The process of development involves test item creation and THREE rounds of reviews. The author/SME is provided with a set of guidelines for authoring the test items and for answering the "HOW" of the assessment methodology. The author/SME has to strictly avoid any sort of plagiarism in the content.

Once the creation is complete, test items are put through the prescribed rounds of reviews.

- ***ID Review**:* First is the Instructional Designing review by an in-house Instructional Designer (ID). Before reviewing the test items, the ID reviewer needs to check the submission for any plagiarized content. The ID checks the transition, flow, and clarity of the test items.

- ***Edit Review**:* The edit reviews are done by an in-house editor. The editor checks the test items for language correctness.

- ***ITR**:* After this, the test items are attempted by an independent technical reviewer (ITR) for functional/ technical correctness. ITR could be done internally or externally.

- ***Audit**:* Finally, the Content manager does an audit of these test items.

Step 6 - Assessment Validation

The test items are sent to client for validation. The feedback received from the client is incorporated and the assessment is finalized. In certain specific cases, the assessment is run through a small portion of population to check its validity.

Step 7 - Translation Process

One of the major hurdles organizations face in administering online assessments in India is the lack of linguistic flexibility available on assessment content. It is imperative for assessments firms to employ language experts who can translate the assessments for diverse linguistic communities that the assessment is to be administered to.

Step 8 - Upload and Link Generation

Once the assessment is created, the next step involves presenting the assessment to the client or the test taker. This involves uploading the content on a secure and stable platform and generating the link that would trigger the assessment.

The assessment link is then shared with Independent Technical Reviewers (ITRs) who are expected to the attempt the test for a last-mile check.

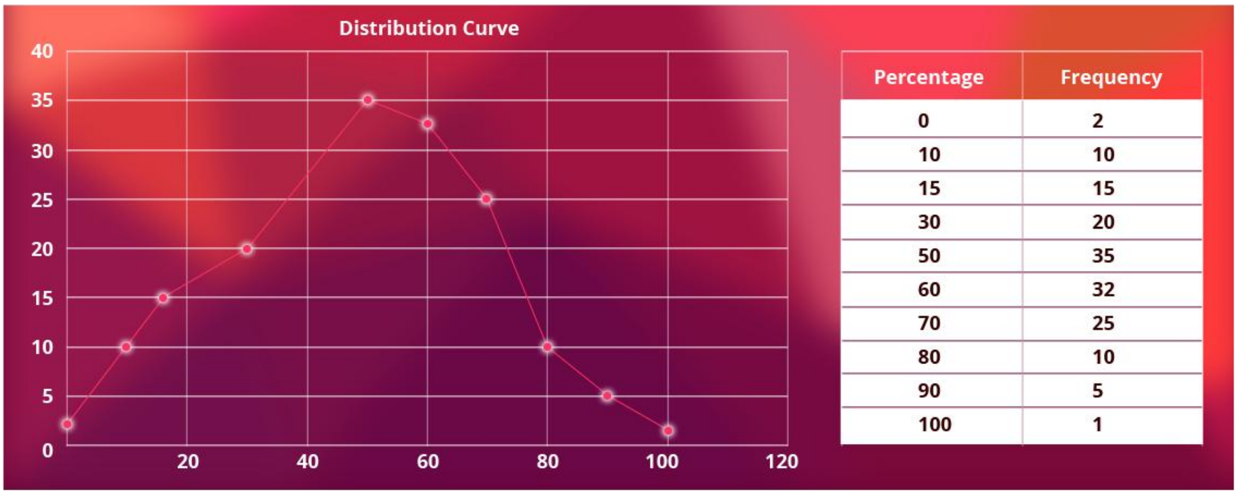
Step 9 - Feedback and on-going validation

Feedback is a critical part of the assessment creation process. All the steps prior to the administration of a test are focused on minimizing errors in measurement. But, the real litmus test of an assessment happens when the actual target audience takes an assessment built this way. Feedback from test takers is taken at the end of each assessment to gauge the effectiveness and subjective experience of the test takers. Further validation is done using the following stages of analysis.

- Overall analysis
- Section-wise
- Topic-wise analysis
- Item-wise analysis (discriminate)

To lend more perspective to the analysis and the objective behind it, let's take the simple case of an overall test analysis. The objective is to understand if the items/questions serve to amply discriminate between the abilities of the test takers. Once the test has been administered to a critical mass of test takers (higher the better), a distribution curve is plotted of the percentages (in case of a norm-referenced test) against their frequency, i.e. the number of test takers that achieve a certain percentage score.

Consider the following data that maps the percentages obtained by test takers against their frequency.



Percentage	Frequency
0	2
10	10
15	15
30	20
50	35
60	32
70	25
80	10
90	5
100	1

In this case, we see that the data follows a normal distribution. Higher the number of test takers that fall within a standard deviation of -1 and +1, better the test is accepted to be. If the curve shows an anomaly, where a disproportionately high number of test takers score a very low or very high percentage of marks, thus shifting the apex of the curve either towards the left or right of the graph, the test would be considered erroneous and corrective action would be required.

An even deeper analysis is conducted by doing a similar study on individual test items/questions to see how well the question discriminates among the test takers. By its nature, statistical analysis of assessments for on-going validation requires considerably elaboration, and to keep this eBook palatable, we'll not delve into its details.

06 SECTION QUALITY CHECK

Quality checks should be done at each and every stage, to continually weed out inconsistencies or errors in the assessment content. Gaps, minor and major, are identified, communicated and rectified. The table below represents the categorization of minor and major gaps.

MINOR GAPS
Instructional clarity needs to be improved
Article usage, comma, and period usages

MAJOR GAPS
Incorrect assessment design framework/ design framework not mapping to the requirements
Content not mapping the design framework and target audience
Inappropriate scenarios/situations/code
Technical Issues
ID issues - transition, flow, instructional clarity is missing
Language issues - grammatical, spelling, subject-verb agreement

MERCER | mettl

ABOUT US

Mercer I Mettl is a Saas based assessment platform that enables organizations to create customized assessments for use across the entire employee lifecycle, beginning with pre-hiring screening and candidate skills assessment, training and develoment programs for employees/students, certification exams, contests and more.

TRY FOR FREE

INDIA OFFICE

+91-9555114444
Plot85, Sector 44, Gurgaon,
Haryana, India - 122003

US OFFICE

+1-650-614-1816
Mettl Technologies Inc. 113
Barksdale Professional Center,
Newark, Delaware 19711, USA

Australia:	+61390699664
Indonesia:	+6285574678938
Singapore:	+6531386714
South Africa:	+27875517192
UAE:	+9718000320460
UK:	+441422400843

contact@mettl.com

Be sure to carefully read and understand all of the disclaimers, limitations and restrictions before using the assessment services, reports, products, psychometric tools or the company systems or website. Read the complete disclaimer here.

MARSH & McLENNAN COMPANIES