IBM® Kenexa® Skills Assessments on Cloud

Validation and Reliability

Smarter Workforce

IBM

TABLE OF CONTENTS

SKILLS TESTING

IBM® gives you the power to identify and select the most talented candidates. It includes over 1,100 validated assessments for clerical, software, call center, financial, healthcare, industrial and technical job classifications.

TEST FORMAT

IBM Kenexa skills assessments are delivered in one of the following formats:

- **Multiple-choice** - assessments present the test taker with a question and from two to five possible answers. Usually, only one of the possible answers is correct. An example would be the Basic Office Skills assessment. Some multiple-choice assessments contain images or audio.
- **Interactive Assessments** - these assessments simulate the environment of an actual desktop application. On each question, test takers are presented with a task and must correctly perform a step or series of steps to complete the task. If the task is completed correctly, the question is scored as correct. An example of an interactive (or sometimes referred as a "Software" test) is the Microsoft Outlook 2013 assessment. Interactive tests support the vast majority of short cut keys; applicants will not be adversely scored for using the most efficient means possible to answer questions.
- **Exception tests** - these assessments do not the traditional scoring algorithms found in a multiple-choice test. The best example of exception tests would be the typing or data entry assessments, as they measure a person's typing or data entry skills not in terms of correct or incorrect, but rather in words per minute.
- **Samples** - IBM provides a series of non-scoring environments for Writing Samples, Translation Samples, and Code Samples. Gauging writing skills (as well as coding and translation) can be a highly subjective endeavor. While you may be able to write a grammatically correct sentence, it does not guarantee that you are a "good writer", so we leave it to the customer to determine the level of writing effectiveness represented by the candidate's sample. We do believe though, that providing a controlled, immediate writing environment helps to assess the skills of the writer, in context. It is for this reason that we provide this environment and encourage online customers to use it, free of charge.

The number of questions depends on the number of tasks in that subject area. The test format also varies. Most assessments are multiple-choice (e.g. Basic Office Skills or C++ Programming) and most software tests are interactive. Some of the software tests are offered in three versions: Normal User, Power User and Whole Test. These different test levels are available so that you can give a test that is appropriate for the skill level you require.

- **Normal User Test:** These tests consist of about 30 questions. Normal User tests are appropriate for testing skills at a standard level. Although Normal User tests contain beginner, intermediate and advanced questions, all questions pertain to commonly used functions.
- **Power User Test:** These tests consist of about 25 advanced questions. Power User tests are appropriate for testing advanced skills of a test-taker. Use these tests to separate the skills of your highest scoring Normal User Test Takers. Power User tests should only be given when you want to concentrate on advanced skills.
- **Whole Test:** These tests consist of about 55 questions. They combine the questions from the Normal User and the Power User into one test. Use the Whole Test to get the most comprehensive skills assessment.

Test Components

Interactive and multiple-choice assessments have the following components:

- **Question Type:** Each test is generally broken down into 4-7 subject areas. These are referred to as the Question Type. This organizes the test in a coherent manner, according to skill sets. Each Type should have several questions assigned to it, although they do not have to be grouped together within a test. Example: Formatting would be a type in a Microsoft Word test.
- **Question Task:** Each question should have a unique Task. The task is a general description of what that question is about. Example: Bold would be a task in a Microsoft Word test.
- **Question Level:** Each question is to be assigned a Level. The levels we use are Basic, Intermediate and Advanced. Example: Bold would be considered a Basic task in a Microsoft Word test.

Test Levels

- **Basic:** Question focuses on a fundamental concept regarding this topic. Qualified test takers should answer the majority of Basic Questions correctly. A test taker with only Basic understanding of the topic may require supervision to complete more intermediate and advanced tasks within this topic.
- **Intermediate:** Question focuses on a solid understanding of the core concepts regarding this topic. A test taker with an Intermediate understanding of the topic is likely to be capable of working on most projects with only moderate assistance, usually with advanced concepts.

- **Advanced:** Question focuses on a superior understanding of the concepts regarding this topic. A test taker with an Advanced understanding of the topic is likely to be capable of mentoring others on even the most complex projects.

LENGTH OF TIME TO COMPLETE AN ASSESSMENT

How long it takes to complete an assessment can be determined by different factors, such as the test taker's familiarity with the subject matter, the number of questions on the assessment, and the testing computer being used. On average, it takes about 40 seconds per question. Average times for specific assessments can be provided upon request.

Typically, IBM does not impose time limits on assessments. However, it will be noted when each test taker started the test, finished the test, and how long it took the test taker to answer each question (loading time per question is not included).

The IBM team continually works with clients' business needs to drive the performance and development of products and services. Customization is available, including look and feel, custom reports, integration into Management Systems, and proprietary test content.

CUSTOMER GUIDELINES

IBM recommends that an organization establish and utilize a consistent standard hiring process when making hiring decisions. Test results are most helpful in making a hiring decision when they correspond with a candidate's history, interview impressions, and reference checks to provide a reliable and detailed picture of the candidate.

CUSTOMER RESPONSIBILITY

IBM tests have been developed through a rigorous content validation strategy. Most focus on real-life scenarios and knowledge-based actions to assess the current level of a *particular skill set*. For example, IBM currently has many tests applicable for an Administrative Assistant. Organizations select those tests that cover the important facets of the job. The important thing to note is that each of these tests must reflect actual skills used on the job. Although the test may be content valid (yield an accurate and representative indication of the skill tested), test validity may be compromised if incorrectly administered.

Customer Compliance

While the customer may rely on IBM's assurance of content validity, the test administrator has some responsibilities to ensure the predictive validity of the test. Listed below are some issues the test administrator should adhere to in order to maintain test validity:

- *The First Step of Customer Compliance* - Administer skills tests to only those applicants that will be employing those skills on the job. For example, a Java 2 test may be administered to a candidate that will be expected to conduct programming in Java 2. Conversely, the Java 2 test should not be administered to a candidate applying for a position, such as a Payroll Manager, that will not be programming in Java 2. A job analysis should be conducted to determine the skills required for each position and only tests that apply to those skills should be administered to applicants.
- *The Second Step of Customer Compliance* - Review the content of each test before administering. The content of the test must fit the skills required for the position. If, upon review, a segment of the content of a test is found to fall outside of the core responsibilities of the job, contact IBM. IBM will be happy to discuss customization of test content for specific organizational needs. Taking advantage of IBM's customization opportunities will also ensure compliance with the EEOC Guidelines for test administration.
- *The Third Step of Customer Compliance* - Review the skills that are job applicable and which will be tested. Can the skill feasibly be learned in a brief on the job training period? The EEOC Guidelines specifically state that a pre-employment test should not cover skills that conceivably could be learned in a brief, on the job orientation.

Picking the Right Assessment

Vendors will undoubtedly claim that one kind of test can solve all of your hiring problems by predicting everything you need to know about an applicant's performance. Fortunately, his or her company just so happens to have that test; the truth is that test doesn't often exist. The right approach is to fit the right tool to the right need. Use different tests in conjunction to strive for a rounded understanding of each applicant.

In order to pick the appropriate assessment for the position, you need to have a firm idea of what you want the test(s) to measure and how you are going to use the test results in your hiring process. A common way to achieve that is by conducting a job analysis or reviewing the job description (which is based on a job analysis).

There are many different methods for gathering job analysis data including focus group interviews with subject matter experts, structured questionnaires, and direct observation.

- **Identify Job Requirements:** Identify each Job Requirements of the position.

  Example: You may have a Job Description for a position that "Knowledge of Human Resources".

- **Identify Position Subject Matter Experts:** Identify the personnel qualified to match job requirements to test content.

  Example: Who understands what "Knowledge of Human Resources" means for the specific position?

- **Match Job Requirements to Test Titles:** Determine which tests best fit each job requirement.
  - What competencies are most important to the success of the position?
  - What level of mastery of the subject is required at entry to this job?
  - Does the applicant have the potential to learn this skill quickly?
  - Does the applicant need to have up-to-date knowledge of this the day they start?
  - Will the hired person be expected to walk in tomorrow and take over these tasks with only orientation?
- **Test Review:** Once you have narrowed down your list of potential titles, have the Position Subject Matter Experts review the tests to determine the extent to which the tests are relevant to successful job performance.

Example: After reviewing the description and outline of test, you are considering the Human Resources Basics test and the Interviewing and Hiring Concepts. Each test is reviewed and bother tests are considered potentially valuable.

- **Pilot Testing:** Administer the test(s) to a small number of incumbents and, if available, a small number of novices. Following the test, the examinees are interviewed and their opinions on each of the items solicited. Additionally, where sufficient numbers exist, some preliminary statistical analyses can be conducted.

  Example: Both tests are taken by incumbents and it is determined that only the Human Resources Basics test appropriately fulfills the requirement.

  - Did the test content seem relevant to the position?
  - Did the test scores correlate to job performance?
  - Did the incumbents do well on most sections, but poorly on a specific content area?
  - Is that content area essential at hire? Is that part more or less relevant?
  - Did the candidate score well on the basic and intermediate level items, but poorly on advanced?

Implementation

Tests can reliably and accurately measure some job-related skills that might be difficult to measure during interviews. Even the most well designed structured interview can involve some interviewer bias. Determining when assessments are used in the hiring process is often dependent on how important the skills results will affect the hiring decision.

Below are some questions you may want to consider to assist in determining the best approach to implementing assessments in your hiring process:

- Are the assessments primarily used to save time because of the amount of applicants?
- Are the assessments primarily used to review candidates objectively? (E.g. Jane may consider herself an expert in Microsoft Excel whereas Frank may consider himself to be proficient. In reality, Jane and Frank may have about the same level of experience, but their self perception is limited to their experience and exposure.)
- Are you testing onsite or remotely?
- Do you currently utilize testing in your hiring process?
- Will administering too many assessments irritate excellent prospective candidates, causing drop rate to increase more than is acceptable?

Test Implementation

Multiple Hurdles - test takers must pass two or more tests or process to continue in the hiring process.

This approach may be necessary if a certain licensure or certification component is required for the position.

If the cost of test administration is high, this method may reduce cost by utilizing less costly methods to the majority and administer more expensive tests to a smaller population.

If there are too many hurdles, you may end up with a candidate pool without any of the genuinely useful qualifications.

Sequential Hurdles – test takers must pass one test or process to continue the hiring process.

Early rounds are often utilized to reduce candidate pool. The reduced candidate pool may still contain unqualified candidates, but the amount will be greatly reduced.

Total Assessment - test takers are administered every test and process in the hiring process. The information gathered is used in a flexible manner.

This allows a high score on one test to be counterbalanced with a low score on another.

You want to make sure you don't offset a score of a required element with a preferred element.

Many companies look to initially reduce the candidate pool by reviewing any "must have" requirements as knockout items where these "must have" requirements cannot be offset by some other competency. This could include certain certification requirements, ability to work particular hours, required knowledge or disposition…etc. Then, they administer a group of tests and/or procedures that when looking at the candidate as a whole will have a well balanced view of the candidate prior to hiring. For example, you may have a candidate that has:

- High scores in all knowledge tests
- Appropriate experience based on resume, reference checks, and referrals
- Great enthusiasm exhibited in the interview

- Minimal management experience and below average scores on leadership assessments
- What is most important to the position?
- Can any of the first three areas compensate for the last?
- Is one item more critical to the success of the position?

Scoring Guidelines

How candidates' results are interpreted will vary depending on the purpose of testing. Therefore, it is possible that every position within a company would have different levels of acceptable scores. Should you wish to determine cut-rates for our tests, it is your responsibility to set up a separate study in a position-specific context in order to comply with EEOC guidelines.

Test score reports include question-by-question results. Consider the levels and types of items missed in conjunction with the needs for the position. To ensure that you are using all tests in a valid manner, please check individual question results and compare them to your needs.

Typical Result Information

Result Item	Definition
Date:	Date the test was taken.
Elapsed Time:	Total time spent on questions (this will not include any time required to load a question so that someone with a slower internet connection will not be negatively impacted. With the exception of Call Center Data Entry, audio tests do NOT include audio playing time in the elapsed time.
Questions Correct:	Number of Correct Questions out of the total number of questions.
Score:	This is the test taker's score (in most instances, this is generated from the number of correct questions divided by the total number of questions.)

Cut Scores

Where tests can be scored as right/wrong, it is possible to establish cut scores using Angoff's procedure. In developing cut scores (passing scores), it is common to use a two-pass Angoff procedure. In the first pass, a group of SMEs rate each item according to the following directions: "Suppose you have a roomful of 100 certification candidates who just meet the standard we discussed and who haven't taken the test before. How many of these 100 minimally-acceptable candidates will be able to answer the question correctly?" An average rating for each of the items is computed. In the second pass, the same SMEs are given the mean rating for each item and a facilitator guides discussion until consensus is reached. Summing these ratings and multiplying that value by the total number of items indicates the minimum number of items these experts feel a minimally qualified candidate will get correct.

Where they cannot be scored as such (e.g., where they are normed against some comparison group) it is still possible to set cut scores judgmentally or empirically. Either way, it involves determining the scores obtained by minimally acceptable candidates. In practice, most companies set their cut scores as high as they can, given their existing applicant flow.

IBM does not provide cut scores (pass/fail requirements); how candidates' results are interpreted will vary greatly depending on the purpose of testing. Companies should set their own proficiency levels and cut scores to reflect the needs of the position. Therefore, it is possible that every position within a company would have different levels of acceptable scores.

- Tests must be proven highly relevant to the job duties and/or highly related to job performance to be defensible if you use it to knock out candidates.
- High cut scores reduce candidate volume (need to evaluate your candidate volume and a sample of test-takers to get a good cut score level for setting it where you get adequate flow).
- High cut scores are more likely to have adverse impact (imbalance in the pass rate of protected groups).
- Low cut scores provide little utility.

Ideally, you should generate a consistent benchmarking process using employee or applicant results specific for the position for which you are testing. Alternatively, you can set a temporary low passing score, so that only very poor performers (below average) are removed from your eligible population while utilizing other parts of your hiring process that you deem crucial to the success of the position as the major factor(s) of any hiring decision. You can also just use test results as information to learn more about the test taker's strengths and weaknesses and not as a qualifying factor.

You want to make sure that setting any passing scores is well documented. Why the skill or behaviors being tested was determined as requirements of the position, the personnel involved in setting the passing score and why they were qualified to determine the passing score, and any data or methods involved in the determination of passing scores.

CUSTOMIZATION

IBM can customize system according to your needs; customized reports, integration into management systems, creating proprietary tests, validation studies, job analysis, and training sessions. To request more information about our services, please contact us at https://www.ibm.com/us-en/marketplace/employee-assessments

VALIDATION AND RELIABILITY FOR SKILLS ASSESSMENTS

IBM assessments enable organizations to accurately assess the skills and capabilities of applicants during the pre-employment screening process. In order for the results to be beneficial, two conditions must be met. First, the test must measure what it claims to measure consistently and reliably. The development of a valid and reliable test for personnel selection purposes requires adherence to best practices in test development. Second, the attribute measured by the assessment must be relevant to the position and should typically be a requirement of the job upon employment (i.e., not something that can be learned on the job). IBM is committed to ensuring internal content validity. The test administrator is responsible for ensuring that a test is appropriate for a position. Before reviewing the development and validation of this particular assessment, some background on measurement is provided.

RELIABILITY

As mentioned, to be useful a measure must be both reliable and valid. What is reliability? *Reliability* refers to the degree of **dependability, consistency, or stability** of a measure. Reliability can be measured in a number of ways. For example, *test-retest* reliability is the consistency of a score over time. *Internal consistency reliability* refers to the inter-relatedness of items within a scale. Items measuring the same thing should be highly correlated with one another. *Inter-rater* reliability refers to the consistency between two judges (e.g., interviewers). Inter-rater reliability is not the same as agreement as two raters could be perfectly reliable and never make the exact same rating (i.e., if one judge consistently rated the target one point higher than the other judge across all facets).

Reliability is important because it sets the upper limit on validity. A test cannot be more valid than it is reliable. Think of it this way: a test cannot predict performance better than it predicts itself. In most cases, an estimate of .70 is considered the minimally acceptable value for reliability (although higher values are better, factors like range restriction may reduce the observed reliability below the .70 threshold).

Reliability is in part a function of test length, such that longer measures tend to be more reliable. However, there is a decreasing benefit to adding items, i.e., the biggest improvement in reliability comes when going from 1 item to 2, the second largest improvement occurs when going from 2 items to 3, etc. The number of items required to reach acceptable levels of reliability varies as a function of what is being measured and how well the test was developed.

In the majority of cases where a well-developed assessment is administered more than once, the scores will be very similar (test-retest reliability). Although scores are likely to be very close, several reasons explain why the scores are rarely identical. First, the time lab between test administrations can influence the test-retest reliability of an assessment. Generally, the longer the lag between test administrations the lower the reliability. Second, interventions occurring between test administrations can influence the test-retest reliability of the assessment. Someone who has attended training on a piece of office software during the interim can be expected to improve his/her scores on that assessment taken a second time. Third, transient factors, such as distractions in the testing environment, the test takers' physical well-being, and their mood on the date of administration can affect test scores. Some of these factors can be controlled; others cannot. The *standard error of measurement* is used to place a confidence interval around an observed score (indicating the range within which a new test score is most likely to fall).

Reliability is a necessary but insufficient condition for a test to be useful, a measure must also be valid. What is validity? *Validity* refers to the **accuracy of the inference** drawn from a measure. Tests are administered so that we might learn something about the knowledge, skills, abilities, or personality of the person taking the test. The extent to which the test allows us to draw accurate inferences or conclusions about such elements is validity. For example: a typing test is valid if it allows us to draw accurate inferences about a person's typing ability.

VALIDITY

Additionally, validity is a property of the inference and not of the test. People often ask: "Is this test valid?" The answer should always be "It depends." Validity depends on the use of the test. A test designed to predict success in sales is valid to the extent it allows one to draw accurate conclusions about an individual's ability to sell. Whereas the test would yield valid conclusions when given to applicants for a sales job, it would be very unlikely to produce valid predictions when administered to applicants for a firefighter's job. The same test can be valid or not, depending on how and where it is used.

Another important point is that there are degrees of validity. Validity is not an either/or proposition - the more accurate the inference, the greater the validity. Even tests with modest validity can have great utility (i.e., value) if the employer can be highly selective or if it can be leveraged across a great number of applicants.

In days past, the field of psychology used to talk about types of validity. Today, most have adopted the "unitary" view of validity, meaning there is only validity – the degree of accuracy in the inferences drawn. What were types of validity are now recognized as strategies for accumulating evidence that a measure is valid. The more evidence of validity, and the more varied the strategies for accumulating it, the better.

Content Validity

One strategy for establishing validity is called *content validity.* Content validity is established when it can be demonstrated that the items comprising the measure are representative of the domain it is intended to measure. Items should be present in proportion to their existence in the domain. Take the example of a test designed to assess a candidate's ability to read (a) safety materials, (b) operating procedures, and (c) routine correspondence. If safety materials represent 50% of what an incumbent reads, then items measuring safety-related reading comprehension should comprise 50% of the total measure. Subject matter experts are typically used to ensure that the items are clear, accurate and representative.

IBM validates its skills assessments using Content Validation, which is an appropriate validation method according to the Uniform Guidelines on Employee Selection Procedures(1978) adopted by the Equal Employment Opportunity Commission, the Department of Labor, the Department of Justice, and the Civil Service Commission.

IBM formalizes its skills content validation process with a Content Validity Ratio (CVR) as well as a Content Validation Form (CVF). In order for tests to be released, IBM requires a CVR of .99 or higher as per C. H. Lawshe (1975). As per Barrett (1992, 1996), we require our SMEs to review the tests via a series of questions for the test as a whole, on an item-by item analysis, and potential issues to review.

Criterion-Related Validity

Another strategy supporting the accuracy of an inference is known as *criterion-related validity*. Criterion-related validity is established when the measure is shown to be significantly related to an important outcome variable or criterion. Although there are many different criterion-related validity designs, the two most commonly used are predictive and concurrent. In a concurrent validation study, incumbents are tested and measures of their performance are gathered at more or less the same time. In a predictive validation study, applicants are tested and the performance of those hired is measured at some later date (after performance has stabilized and can be reliably measured). In both cases, validity is established when the test scores are found to relate to performance (i.e., higher tests scores are generally associated with higher performance). Both predictive and concurrent designs have their strengths and weaknesses and both are appropriate when evaluating the validity of a measure in a selection context.

Construct Validation

A third strategy supporting validity is that of *construct validation*, which is established by demonstrating relationships between the measure and other theoretically meaningful variables (and by demonstrating no relationship where there should be none). These other variables might be different measures of the same construct, or antecedents expected to be related to the variable, etc. Convergent validity is often shown by demonstrating a high correlation between two different measures of the same construct. Similarly, discriminant validity can be shown by demonstrating low correlations between different constructs.

Face Validity

A fourth strategy for supporting validity is *face validity*, which indicates whether a test taker believes that a test is an appropriate measure of future work performance. Although face validity does not support the overall validity of a test, it can improve applicant perceptions of the testing process. IBM assessments are designed to test applicant skill using highly realistic test stimuli (e.g., measuring skills in the *Microsoft* office suite using an interface that mimics the actual computer programs). The realism of this testing process promotes high face validity.

ADVERSE IMPACT

Although there are several standards for determining *adverse impact*, the four-fifths rule is the most common. This 4/5th's standard holds that adverse impact is indicated when the hiring rates for a minority group falls below 4/5's (or 80%) of the majority group. For example, if the selection ratio for whites is 30% (meaning 30% of those who apply are hired), then the selection ratio for other groups should be at least 24% (30 x .80). Although adverse impact should be avoided where possible, its presence does not preclude the use of an assessment. It does require the employer to demonstrate that the assessment is job related and that no less discriminatory alternatives exist. As an example, strength tests are often used to hire fire fighters because of the high physical strength requirements of that job. On average, females do not score as well as males on these tests. As a result, using physical tests to hire firefighters may create adverse impact against women. Given the importance of strength for the job and the lack of alternatives, many fire departments use strength tests even though they may create adverse impact.

It is important to highlight that even if adverse impact potential is not indicated in assessment measurements other aspects of the hiring process may cause adverse impact. In all cases, the potential for adverse impact should be monitored carefully.

Further Adverse Impact Analyses can be conducted upon your request provided that the collection of demographic data has been built into the test environment deployed for your account(s).

Non-White Races include American Indian or Alaska Native (Not Hispanic or Latino), Asian (Not Hispanic or Latino), Black or African American (Not Hispanic or Latino), Hispanic or Latino, Hispanic/Latino (White Race Only), Native Hawaiian or Other Pacific Islander (Not Hispanic or Latino), and Two or More Races (Not Hispanic or Latino)

*Greater than 100% Adverse Impact Ratio indicates that the minority group scored higher than the majority.

RESEARCH METHODS FOR SKILLS ASSESSMENTS

We encourage our customers, partners, and employees to provide suggestions and insight into our assessment research to improve our customers' selection process. In addition to ensuring the continued satisfaction of our clients, feedback is instrumental in establishing internal benchmarks that drive the performance of our associates, products, and services. Over 80% of our development is from customer requests; we research every request made.

Every quarter, IBM analysts research specific industries for potential testing growth as well as requests yielding a library of individual titles for potential development. Particular industry focus may be determined through Sales Feedback, Industry Publications, Industry Experts, Competitors, and/or Customers.

A Four-step Research Process includes:

Step 1: Title Research: Research from Customers, Industry Contacts, and Industry Research is conducted to make sure the test we are proposing to develop is focused correctly.

Step 2: Competitor Research: Once the title research is complete, an assessment of competitor's holdings is made. A series of data is gathered.

Step 3: Customer Feedback: Following the composition of a test title, description, a listing of the positions it is aimed at, as well as an outline of the test content, we forward this information to customers, and request answers to a series of questions.

Step 4: Accept/Decline: IBM reviews the research data for evaluation to either accept or decline for development on a quarterly basis.

DEVELOPMENT PROCESS FOR SKILLS ASSESSMENTS

IBM assessments target such areas as computer-oriented knowledge (from software to databases), call center, accounting and finance, office skills and industrial, among others. All content is developed, reviewed, documented, and maintained through a proprietary online Content Management System and encounters a ten-step development process. In order for a title to be released, all appropriate steps must be completed.

Phase 1: Contract Subject-Matter Experts (SME) to Create and Review Assessments

The qualifications for subject matter experts are determined by the subject matter; for certain technologies, 10-15 years of high-level experience is not possible. We review both our existing Industry Resources that we have utilized in the past that retain a high SME score. For new SMEs, we review resumes with current experience, professional associations, education, certifications, and training experience...all relative to the particular subject matter.

Phase 2: Test Author Creates Initial Question Set

IBM works with respected experts in each field and combine our industry research and their expertise in determining our test content.

Phase 3: Subject-Matter Experts Review Test Content

In order to ensure that the content developed is representative, we enlist the help of a team of Subject Matter Experts to review the test as well as to make suggestions for inclusions that may have been overlooked. Subject-matter experts will review the content to ensure that all questions and answers are correct and determine whether the question set and skill level is a representative sample of the subject matter being measured. This process of writing and reviewing results in a test that is meaningful, appropriate, and useful in assessing the knowledge of the test taker in a particular skill.

Phase 4: Internal Review of Subject-Matter Experts' Feedback

The IBM validation team reviews the reports generated by the reviewers noting what questions the author must correct or defend.

Phase 5: Author Revisions Based off of Subject-Matter Experts' and Internal Feedback

The author is provided with the compiled feedback to make any corrections or to state why questions that have been challenged should stay as is. Once the test has gone through this process, our internal validation team examines the test for bias and balance within the test questions, answers, types, and tasks.

Phase 6: Initial Validation Review

IBM tests are content validated and focus on real-life scenarios and knowledge-based actions to assess the skill level of a ***particular skill set***. Our skills assessments do not determine whether an applicant has the right demeanor or personality. You get to pick and choose those tests that cover the facets of the job description that are important to the needs of the position. Multiple members of the validation team confirm that the assessment meets the criteria established which include, but are not limited to:

- Review Reports to verify that the overall review was positive. If there are too many disagreements, further SMEs may be required.
- Review Reports for all suggested changes and view the results of those suggested changes; make sure author provided documentation for why they did or did not modify the test based on reviewers' comments. If the author did not provide adequate documentation, the test is sent back to author.
- The SMEs followed the Writing and Reviewing Guidelines.

Our validation team examines each question within each test for bias according to EEOC Guidelines. For each question, the following is considered as well as the test at large:

- Is there any language within this question that excludes any member or segment of the population?
- Is the language of this question slanted toward any member or segment of the population?
- Will this question result in an adverse impact for any member or segment of the population?
- Does the methodology employed by this question lean toward a biased benefiting of any member or segment of the population?
- Does the methodology employed by this question lean toward adversely impacting any member or segment of the population?
- May the test result in an adverse impact to any member or segment of the population?
- May the test result in a biased benefiting of any member or segment of the population?
- If any of these questions are answer in the affirmative, the test is reworked or rewritten.

Issues of Balance

Each of the test questions is assigned a skill level; basic, intermediate or advanced. Our validation team goes to great lengths to determine that the skill level assigned to each question is accurate, utilizing the aforementioned Individual Question Documentation Report as well as the reviewers' insights.

The accuracy of the skill level is necessary to produce, not only useful evaluations for our customers, but to accurately reveal the test takers level of knowledge. This ensures that each question is representative of both the skill being test as well as the level of knowledge required by the task.

The validation team also evaluates the percentages of levels, assuring that the test is composed of questions that are appropriately balanced, yielding the opportunity for test takers to reveal their level of knowledge and preventing a too difficult test from resulting in skewed low scores that are not useful to the customer, or test taker.

Phase 7: Quality Assurance

The Quality Assurance Plan presents a framework to ensure delivery of quality products and services. Each assessment is reviewed in a minimum of five different formats (Online Development Server, 4.0 CD-ROM administrator, 3.5 CD-ROM administrator, Online Content Management System-QA Report, and the Online Content Management System -Review Report) by several Quality Assurance members. If questions arise, the Quality Assurance team consults the Subject Matter Experts assigned to this assessment. Each Question, Answer, Task, and Topic undergoes a nine-step review process covering Question Duplication, Types Allocation, Spelling, Grammar, Possession, Punctuation, Structure Format & Structure Consistency, Scoring, and Code/Pictures.

If the test has introduced new functionality specifications, such as a new "type" of test format, the test is also reviewed and compared to the functionality specifications.

Phase 8: Final Validation Review

The test goes through a final round of validation review. All issues discovered in this review must be satisfied prior to release.

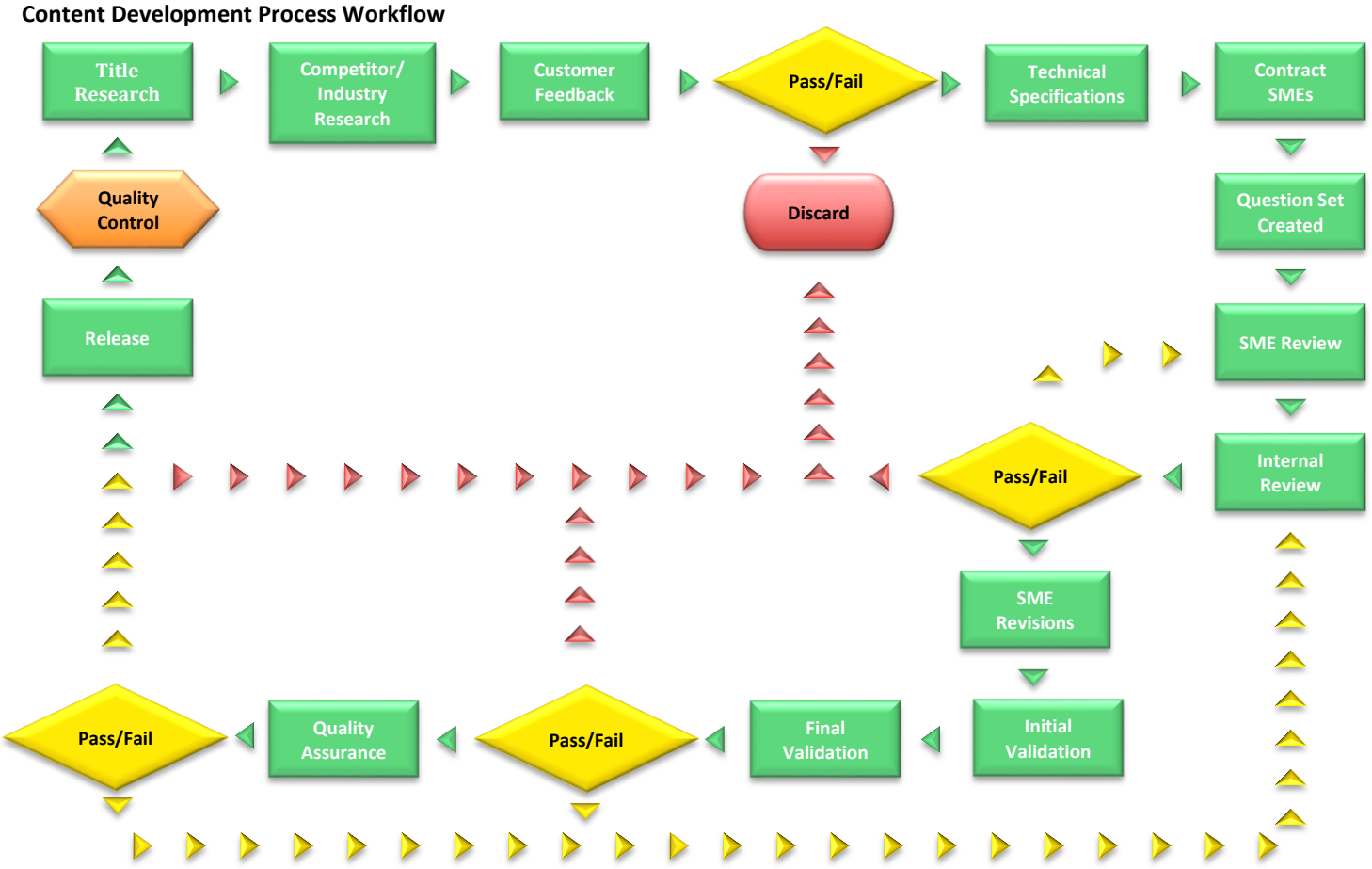
Phase 9: Release

As long as the previous 8 Phases have been conducted and completed fully, with no outstanding issues, the test is posted to the development server, re-tested for functionality and navigation and released to customers with any applicable marketing materials.

Phase 10: Quality Control

IBM is committed to providing our customers with quality assessments. To ensure that our tests are accurate and current, we review tests regularly. All tests that are modified re-encounter Phases 6-9 at minimum. Every quarter in our online newsletter, we notify customers of the tests that we have modified due to customer/test taker feedback, time-sensitive titles, and excessive usage re-writes.

Content Development Process Workflow



APPENDIX A: RESOURCES REGARDING PERSONNEL ASSESSMENT

The following are general reference documents and recommended readings regarding specific topics and issues relating to personnel testing and assessment.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. 1985. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Arvey, R.D., and R.H. Faley. 1988. *Fairness in Selecting Employees.* Reading, MA: Addison-Wesley.

Boudreau, J. 1996. Cumulative Supplement to Employment Testing Manual. Boston: Warren, Gorham & Lamont.

Bureau of National Affairs Policy and Practice Series. 1992-1995. *Fair Employment Practices Manual #8.* Washington, DC: Author.

Bureau of National Affairs. 1990. The Americans with Disabilities Act: A Practical and Legal Guide to Impact, Enforcement, and Compliance. Washington, DC: Author.

Douglas, J.A., D.E. Feld, and N. Asquith. 1989. *Employment Testing Manual.* Boston, MA: Warren, Gorham & Lamont.

Equal Employment Opportunity Commission and U.S. Department of Justice. 1991. *Americans with Disabilities Act Handbook*. Washington, DC: Author.

Equal Employment Opportunity Commission. 1978. The Office of Personnel Management, U.S. Department of Justice and U.S. Department of Labor (1979). *Questions and Answers Clarifying and Interpreting the Uniform Guidelines on Employee Selection Procedures.* (1988).

Equal Employment Opportunity Commission. 1992. A Technical Assistance Manual on the Employment Provisions (Title I) of the Americans with Disabilities Act. Washington, DC: U.S. Government Printing Office.

Equal Employment Opportunity Commission. 1992. EEOC Technical Assistance Manual on Employment Provisions of the Americans with Disabilities Act; *ADA Enforcement Guidance: Preemployment Disability Related Questions and Medical Examinations*.

French, W.L. 1990. *Human resources management* (2nd edition). Houghton Mifflin Co.: Boston, MA.

Guion, R.M. 1997. Assessment, Measurement, and Prediction for Personnel Decisions. Mahwah, NJ: Lawrence Erlbaum Associates.

Society for Industrial and Organizational Psychology, Inc. 1987. *Principles for the Validation and Use of Personnel Selection Procedures* (3rd edition). College Park, MD: Author.

U.S. Department of Justice. 1991. *The Americans with Disabilities Act: Questions and Answers.* Washington, DC: Civil Rights Division, U.S. Department of Justice.

U.S. Department of Labor, Employment and Training Administration. 1993. *JTPA: Improving Assessment: A Technical Assistance Guide.* Washington, DC: Author.

U.S. Department of Labor, Employment and Training Administration. 1999. *Testing and Assessment: an Employer's Guide to Good Practices.* Washington, DC: Author.