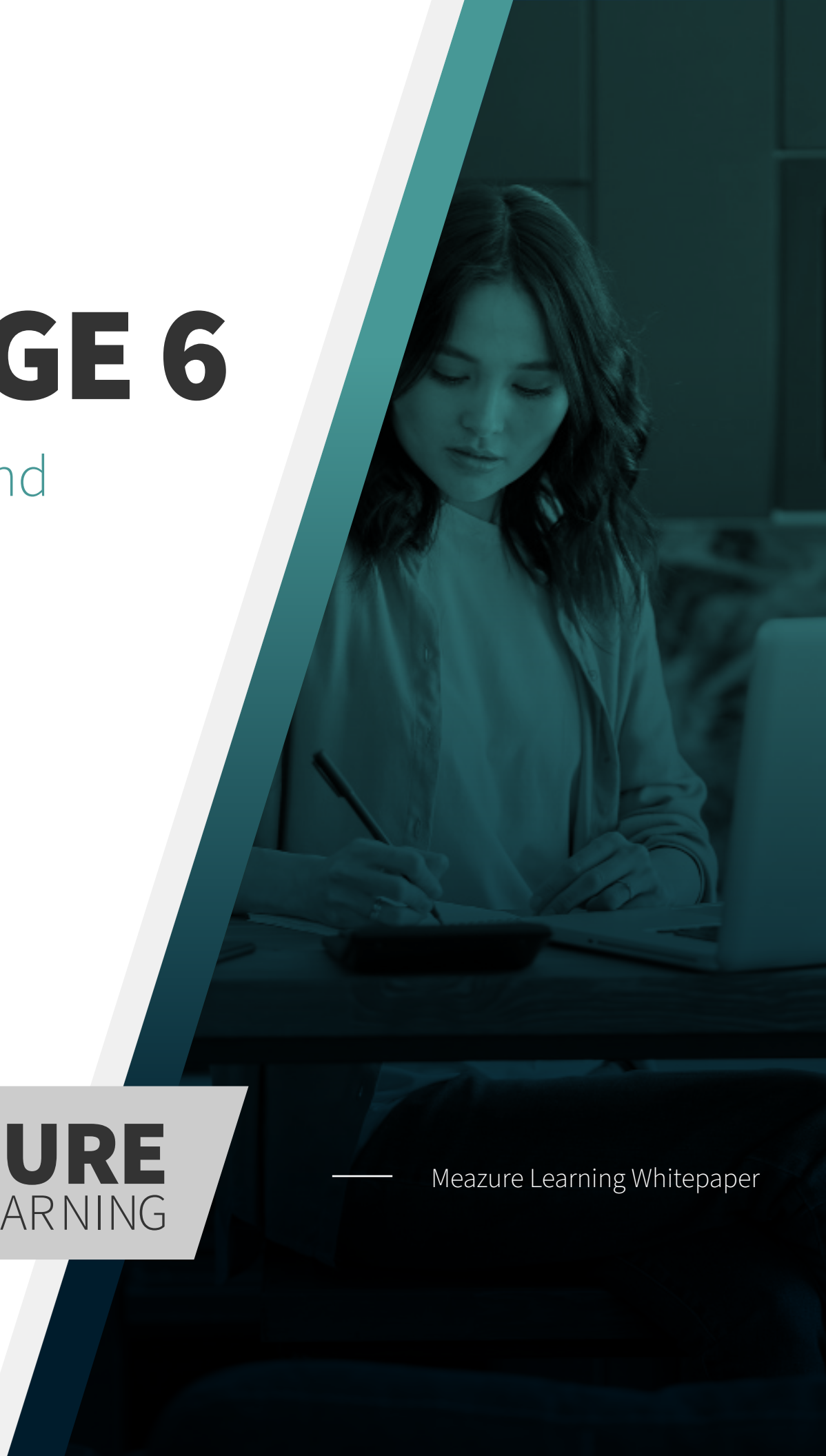
STAGE 6

Scoring and Reporting



MEAZURE LEARNING

Meazure Learning Whitepaper

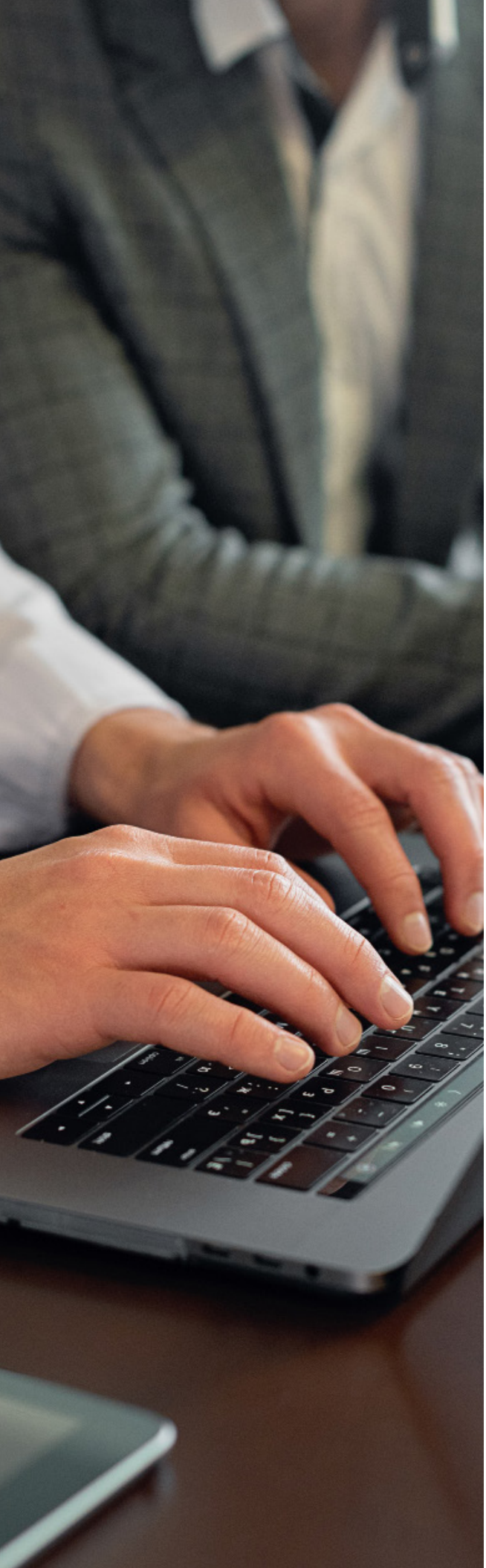
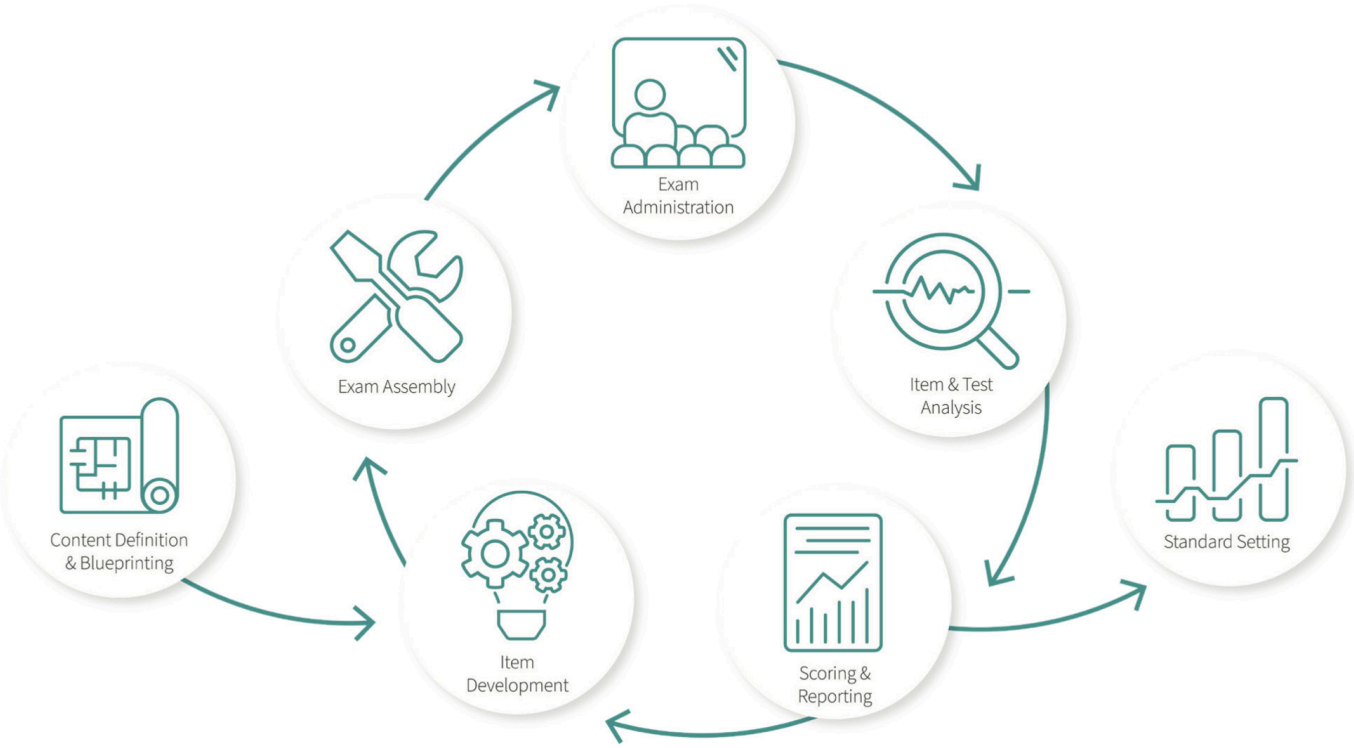


Table of Contents

Introduction

The Assessment Life Cycle is a way of organizing the processes involved in creating valid assessments into a series of easy-to-understand, logical stages.



At this stage of the cycle, you have finished designing and validating your exam; and have presented the approved exam form (or forms) to your candidates. The exam has been written; and candidates' response data have been vetted to ensure that the items that compromise the exam follow psychometric best practices in terms of their difficulty and ability to discriminate between high- and low-scoring candidates.

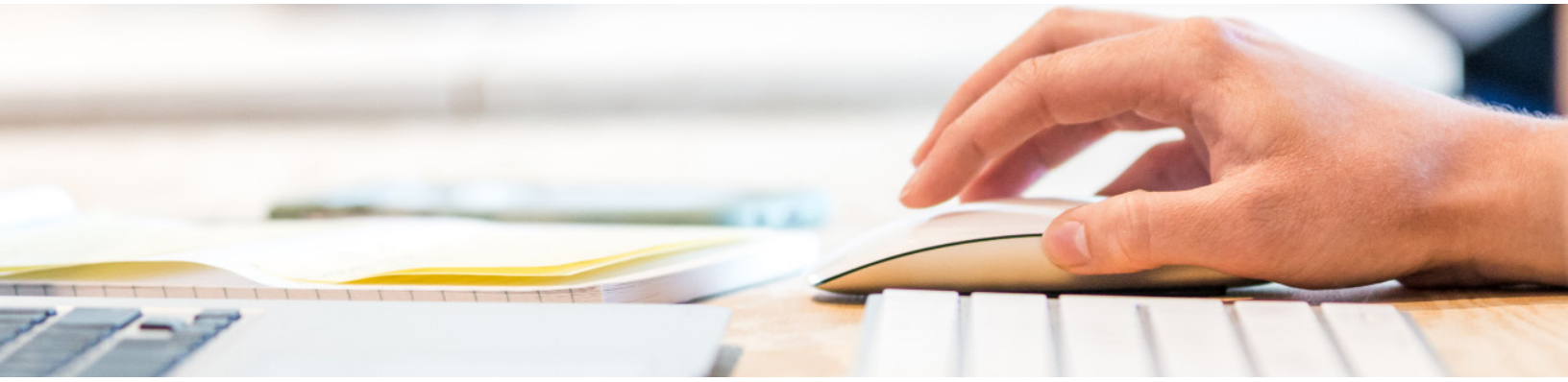
Now that all of this is done, it is important to revisit the now-vetted set of exam items; and calculate candidates' final exam scores efficiently and accurately. Stage Six of the Assessment Life Cycle involves calculating candidates' scores on the exam they completed; and generating reports that will communicate these results in the clearest and most valid manner possible.

The purpose of this white paper is to explain the process of generating final candidate scores and reports; and to explore the considerations that should go into designing and disseminating various score reports that meet each client's unique needs.

1 | Scoring Candidate Responses

When dealing with criterion-referenced assessments (i.e., those that are generally used in a credentialing or licensure context), the key goal of scoring is to compare the performance of each candidate to an established numeric "standard" that differentiates "competent" from "not-yet-competent" performance .

In order to generate a score for each candidate that can be compared to this numeric standard, graders must find a way to express each candidate's level of performance across all items (or tasks) in the assessment as a number. Depending on how the assessment is structured, the exact process of deriving a final numeric score for each candidate may differ. That said, most assessment methods can be numerically scored using one of two general approaches: "Objective scoring" or "judgemental scoring". Each of these forms of scoring will be described in detail below:



OBJECTIVELY-SCORED ASSESSMENTS

What is an "Objectively-Scored" Assessment? The Example of Multiple-Choice Items

An objectively-scored assessment is any assessment where candidate responses on a given item can be evaluated by comparing that response to a keyed "correct" (or "best") answer. For example, consider a multiple-choice examination. When grading such an exam, the grader generally compares the answers that candidate selected for each individual question to an 'answer key' that contains the answer that the candidate should have selected for that question (i.e., the answer that is considered correct, or clearly best among the various answer options). Using this document, the grader can assign each candidate a score for every question by simply checking to see whether or not the candidate answered the question with the same option that is listed as being "correct" in the answer key. The sum of these item-level scores represents the candidate's total exam score.

In most multiple-choice exams, we assume that every question is equally 'valuable' (that is, each question is considered to be an equal or parallel form of the underlying construct that is being tested). As a result, in most cases, every question on the exam is weighted to be worth the same number of points. Following on this idea, we can derive the following basic scoring system: All candidates start with a total exam score of "0.0". Each candidate receives a score adjustment of "+1.0" for every question they answer correctly (i.e., select the same answer that is given by the answer key); and receive a score adjustment of "+0.0" for every question they either fail to answer or answer incorrectly (i.e., select one of the 'distractor' answer options). The candidate's total score on the exam is then calculated as the number of questions in the exam that they answered correctly (for example, 42 points of a possible 60).

This may be the most common approach to scoring multiple-choice examinations. However, depending on the purpose of the assessment, it is also possible to weight different questions differently. For example, a question may only be worth half as much as other questions in the exam (so may result in a score adjustment of "+0.5" if answered correctly). Moreover, in some assessments, candidates may actually be penalized for giving an incorrect answer (for example, giving an incorrect response may result in a score adjustment of "-1.0").

Other Types of Objectively-Scored Assessments:

Multiple-choice exams are perhaps the most well-known and widely-used form of objectively-structured assessment methods; however, other types do exist. For example, a candidate may be presented with an exam containing 'multiple select 'questions (in which several alternatives may be correct for a single question), 'fill-in-the-blank' questions (in which a small number of words or numbers are compared to those keyed as "correct"); or 'image-based/hotspot' tasks (in which the candidate must click on the relevant area of a diagram based on the question prompt).

Although the specific format of the question may vary, all of these approaches rely on objectively-scored assessment methods; as each may be effectively and accurately scored by a grader who has no subject matter expertise. This is only possible because the answers in these assessments can be expressed in a simple format; therefore, all the grader needs is access to an answer key that lists the correct answer option for each question.

Why Use Objectively-Scored Assessments?

As mentioned, the distinguishing feature of objectively-scored assessments is that they do not require subject matter expertise in order to score. In fact, they can often be scored automatically by a computer! As a result, these assessments are highly practical and very popular; and are especially common when working with larger samples of candidates.

The Process of Grading Objectively-Scored Assessments

Before actually proceeding with any scoring, a number of quality assurance steps should be undertaken to ensure that the candidate response data file that will be scored is complete and accurate. This 'quality assurance' step can include several steps, such as:

- Checking to make sure that every response is matched to a specific item (usually indicated by the column of the data file), and to a specific candidate (usually indicated by the row of the data file).
- Checking to make sure that the data is in the correct format; and confirming the correct number of columns (i.e., items or tasks) and rows (i.e., candidate) are present.
- Ensuring the correct answer key is included and referenced correctly by the data file.
- Checking the data file against the list of candidates to ensure that every candidate who should be scored is present and accounted for.

As mentioned, the process of actually scoring candidates' responses on objectively-scored exams is fairly straightforward. In most cases, these data are scored by comparing each candidate's response to the keyed "correct" answer; and assigning one point for every correct answer - and zero points otherwise. This will usually result in the creation of a data matrix that matches the candidate response matrix; but which replaces candidates' actual responses (e.g., 'A', 'B', 'C') with the points they were allocated for each response (e.g., 1, 0, 1). See below for an example of this:

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Candidate 1	A	B	C	C	B	A	B	C	B	D
Candidate 2	A	C	A	D	B	C	B	C	A	C
Candidate 3	B	C	A	C	D	A	B	C	A	D
Candidate 4	A	B	B	C	B	A	B	B	A	D
Candidate 5	A	B	D	D	C	A	D	C	B	A
Candidate 6	C	B	A	D	D	C	B	A	A	D
Candidate 7	D	B	D	A	D	A	D	D	A	C
Candidate 8	A	A	D	C	A	A	D	C	C	A
Candidate 9	A	D	A	C	D	A	B	B	D	D
Key	A	B	A	C	D	A	B	B	A	D

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Candidate 1	1	1	0	1	0	1	1	0	0	1
Candidate 2	1	0	1	0	0	0	1	0	1	0
Candidate 3	0	0	1	1	1	1	1	0	1	1
Candidate 4	1	1	0	1	0	1	1	1	1	1
Candidate 5	1	1	0	0	0	1	0	0	0	0
Candidate 6	0	1	1	0	1	0	1	0	1	1
Candidate 7	0	1	0	0	1	1	0	0	1	0
Candidate 8	1	0	0	1	0	1	0	0	0	0
Candidate 9	1	0	1	1	1	1	1	1	0	1

Once this process has been completed, it is relatively simple to calculate candidates' exam scores using the data matrix that was just created. Usually, this is just the sum of 1's that the candidate received. That said, in order to ensure the accuracy of this scoring procedure, it is a common best practice to simultaneously score data using two different data analysis platforms (for example, Excel and R). Only when the results from these two platforms match perfectly can the score data be considered accurate.

JUDGEMENTALLY-SCORED ASSESSMENTS:

What is a "Judgementally-Scored" Assessment?

Unlike objectively-scored assessments, judgementally-scored assessments require some degree of subject matter expertise on the part of the grader. This is because the questions that candidates are asked to answer (or the tasks that they are asked to complete) are more open-ended than the questions or tasks that are used in objectively-scored assessments. For example, a candidate may be asked to answer an 'essay-style' question that requires them to explain the process they would follow in a given situation (or otherwise demonstrate their line of reasoning). Alternatively, a candidate may be asked to produce a diagram or audio recording; or complete a roleplay or other performance-based exercise.

Given the open-ended nature of these assessments, candidates will often produce a wide range of answers. Consequently, it's often not possible to express the 'correct' answer to these types of assessments using a simple answer key. In order to score these assessments, a grader must instead draw on their own expertise in the subject that is being assessed. Because some level of familiarity or expertise in the subject area is required, the group of graders for judgementally-scored assessments is generally drawn from a known pool of subject matter experts in the field. These experts then score the assessment by comparing the response that each candidate gave against a set of pre-established general criteria for quality e.g.,, a grading rubric).

Why Use Judgementally-Scored Assessments?

At a glance, it may seem impractical to use judgementally-scored assessments (as objectively-scored assessments offer a much more straightforward and cost-effective approach to assessing candidates). However, the open-ended nature of these assessments can allow graders to assess the depth of a candidate's knowledge and thinking in a way that is often not effectively captured using multiple-choice questions or other objectively-scored assessment metrics. Therefore, depending on the underlying construct or subject area that is being assessed, these types of assessments may be more or less appropriate. For example, judgementally-scored assessments may be worth considering when assessing candidates' critical reasoning abilities, soft skills, or ability to complete complex, multi-step tasks.

The Process of Grading Judgementally-Scored Assessments

Because of the complexity involved in scoring judgementally-scored questions, some data 'translation' needs to occur before candidates' raw data can be scored. In completing the assessment, each candidate will provide a "raw" response. This raw response could be in the form of a written essay, video or audio recording, or simulated work deliverable.

In order to score candidates on the broader assessment, these raw responses will need to be rated by graders first (often subject matter experts in the field). In doing so, the graders will produce a numeric set of candidate scores across the various items or tasks in the exam, which can then be used to derive a final examination score.

As if this isn't complex enough, a psychometrician may be faced with multiple ratings for each item/task in the exam. Often, this is due to two (non-exclusive) reasons:

- First, the grading rubric may require graders to score each item on multiple different criteria. For example, a written response may be graded for its content, its grammar, and its style. Each of these may be weighted differently in order to calculate the final item grade (e.g., 60% content; 20% grammar; 20% style).
- Second, because this kind of grading is somewhat subjective, it is often advisable have multiple graders score the same responses. This helps to ensure that candidates' scores are less biased and relatively consistent, no matter who graded them.

As a result, the "translated" data file may actually contain multiple ratings for each item/task in the exam (which must be averaged or otherwise handled appropriately in order to derive a numeric score for each candidate across each item or task).

Once all of this has been taken care of, the final scoring and quality assurance for judgmentally-scored items are much the same as they were for objectively-scored items. That said, because the data from these questions are much more complex to begin with, the potential for error is also generally higher. This, of course, is all the more reason to include a 'double scoring' step for these items (for example, redundantly scoring the data using both Excel and R).

2 | Equating and Scaling Scores

THE PROBLEM WITH RAW SCORES

A raw score usually reflects the sum of points received on all items comprising an assessment. Although there is a tempting simplicity to using raw scores in your final candidate score reports, there are some disadvantages to doing so.

One disadvantage of raw scores is that they are not always comparable across different test administrations. Let's imagine a situation in which two friends wrote the same 140-item credentialing exam in two consecutive years (i.e., one wrote the exam each year). Even though both exams are ostensibly graded out of a possible 140 points, the items that comprise these exams may be different. In order to account for the likelihood that one set of items will be at least a little harder or easier than the other set, two different passing marks may be set (this process is covered in more detail in the next white paper in this series).

Following this idea, imagine that one of the candidates scored 92 points one year and passed the exam; whereas the other (writing a different form) scored 93 points the following year, and failed. This seems illogical at a glance; and would likely confuse or concern the two friends. Since the ordered nature of the raw score metric seems to stand in contradiction to this interpretation, how is the candidate to understand their performance?

The problem is that the raw scores do not provide an interpretive framework against which candidates can compare their performance. Scaled scores mitigate these difficulties by mapping raw scores onto a score scale, with an established set of rules for interpretation. Once the test taker and stakeholders are introduced to the metric and the meaningfulness of particular values within it, the scores are readily interpretable in terms of norms, criteria, or a combination of the two. Moreover, because the interpretation of the scale is consistent over time, scores from different administrations can be directly compared.

EQUATING SCORES ACROSS DIFFERENT TEST FORMS

When you visit a different country, you often need to convert the money you're carrying into a different form of currency. For example, one Canadian dollar is not exactly equal to one American dollar. At its core, the process of creating scale scores is much the same. As this form of currency conversion: A score of 48 on 'Form 1' does not necessarily mean the candidate would also be expected to score 48 on 'Form 2' (as one

form may be more or less difficult). To figure out what this score of 48 would convert to, we need a sort of "translation key" that tells us what a score on one exam form would convert to.

This process of conversion is known in the psychometric world as score 'equating'. And there are actually several different approaches to it. Some of the most common include:

- Mean equating
- Linear equating
- Percentile equating
- Circle arc equating
- Item response theory (IRT) equating

Although the various approaches to equating differ in terms of the specific rules they use to convert scores (and the indices they rely on to create their 'translation keys'), the overall logic and goal is the same in all approaches: To ascertain how well a candidate on one form of the exam would do on a different form, based on the relative difficulty of the two forms.

For the purpose of a simple illustrations, let's consider the mean equating approach (in many ways, the simplest version of equating). In mean equating, "The difference observed at the mean…is defined to be constant throughout the score scale" (Kolen, 1988, p. 33). For example, two forms of an exam may be equated using the following formula:

$$XForm1 - MForm1 = XForm2 - MForm2$$

As illustrated, each candidate's score is re-expressed as their degree of divergence from the mean score on the examination form that the candidate wrote. So, if 'Candidate A' received a score 64.0 on 'Form 1' (which has a mean of 64.0); and 'Candidate B' received a score of 61.4 on 'Form 2' (which has a mean of 61.4), these two candidates would be considered to have received the same mean equated score. Following this idea, a candidate's score on 'Form 1' could be equated to 'Form 2' algebraically, using the following:

$$XForm2 = XForm1 - MForm1 + MForm2$$

USING STATISTICAL EQUATING TO CREATE SCALE SCORES

Once a 'translation key' for the two different test forms has been created using statistical equating, the psychometrician can go ahead and convert the raw scores to scale scores! Often, this is done using an approach like the one shown below:

Baseline Form Scores	Step 1: Equating New Scores	Step 2: Creating Scaled Scores
...	...	...
8	6	400
9	7	425
10	8	450
11	9	475
12	10	500
13	11	525
14	12	550
15	13	575
16	14	600
...	...	...

The first two columns in the table above represent the baseline score metric and the new form metric as determined through equating. Completing this equating step, the equivalent score on the new form has been calculated relative to baseline scores, so that performances can be directly compared across the two forms. For example, the passing score on the Baseline Form was previously defined as 12; and the equivalent score on the New Form is 10. The third column, labeled 'Step 2', then transforms the score scale into the desired range (i.e., one that has properties that facilitate valid score interpretation). In this case, the range encompasses at least 400 to 600 (and probably larger).

3 | Creating Score Reports

"Because validity, by definition, focuses on the interpretation of test scores, any feature of a score report that invites or encourages an interpretation that evidence or theory do not support has corrupted the validity of the assessment (p. 678)."

As the Handbook of Test Development so aptly states, effective score reporting is a matter of communicating the results of an examination is a way that encourages the most valid interpretation possible. Consequently, the work of a psychometrician not only involves communicating the most desired results; or even communicating in the clearest manner possible. This work also must involve careful due diligence to avoid encouraging the reader to misinterpret or overextend the information that they are receiving form their final report. This is true when creating individual score reports for candidates; as well as when creating summary reports for schools or the testing organization (both of which will be described in detail below):

CREATING CANDIDATE SCORE REPORTS

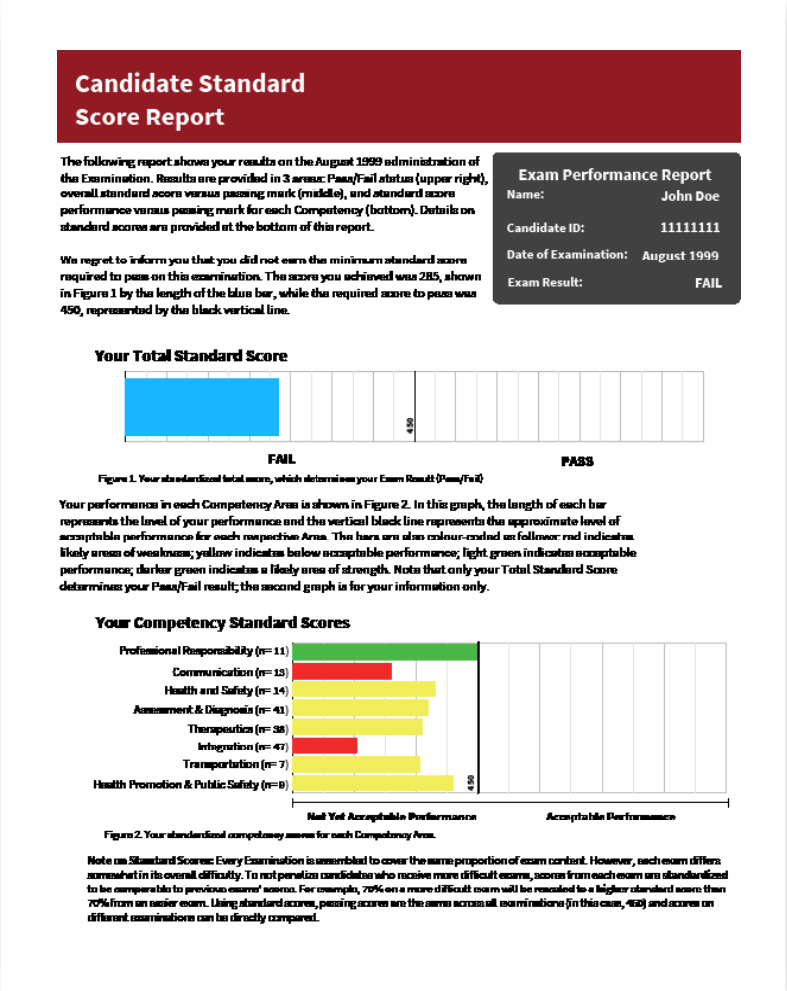
The reports that are most commonly produced for clients include individual candidate reports and group reports, in which data are aggregated often by province, and/or by educational institute.

To give you an example of the kind of information that may be included in a candidate score report, consider the example on the following page:

Here we can see a candidate's performance, relative to the minimal standard (or passing score) for the exam. We can also see whether or not the candidate passed by looking at the grey area in the upper-right hand corner of the report (sadly, we can see that this candidate was not successful). In the middle section of the report, we can also see the candidate's total standard score relative to the passing score (and note that the candidate was fairly close to passing – a standard score of 446, relative to the passing score of 475).

CREATING AGGREGATED SCORE REPORTS

Often, it's not just candidates who receive reports to document the results of an examination. It's also common to create reports to be sent to the testing organization; as well as the schools that candidates came from. These reports can help schools diagnose how their students did relative to the larger pool of candidates across the entire nation; and can also help the testing organization keep track of how the general cohort of candidates did compared to previous cohorts who took the exam.



Candidate Standard Score Report

The following report shows your results on the August 1999 administration of the Examination. Results are provided in 3 areas: Pass/Fail status (upper right), overall standard score versus passing mark (middle), and standard score performance versus passing mark for each Competency (bottom). Details on standard scores are provided at the bottom of this report.

We regret to inform you that you did not earn the minimum standard score required to pass on this examination. The score you achieved was 285, shown in Figure 1 by the length of the blue bar, while the required score to pass was 450, represented by the black vertical line.

Exam Performance Report

Name:	John Doe
Candidate ID:	11111111
Date of Examination:	August 1999
Exam Result:	FAIL

Your Total Standard Score

FAIL PASS

Figure 1. Your standardized total score, which determines your Exam Result (Pass/Fail)

Your performance in each Competency Area is shown in Figure 2. In this graph, the length of each bar represents the level of your performance and the vertical black line represents the approximate level of acceptable performance for each respective Area. The bars are also colour-coded as follows: red indicates likely areas of weakness; yellow indicates below acceptable performance; light green indicates acceptable performance; darker green indicates a likely area of strength. Note that only your Total Standard Score determines your Pass/Fail result; the second graph is for your information only.

Your Competency Standard Scores

Professional Responsibility (n= 11)
Communication (n= 13)
Health and Safety (n= 14)
Assessment & Diagnosis (n= 41)
Therapeutics (n= 38)
Integration (n= 47)
Transportation (n= 7)
Health Promotion & Public Safety (n=9)

Not Yet Acceptable Performance Acceptable Performance

Figure 2. Your standardized competency scores for each Competency Area.

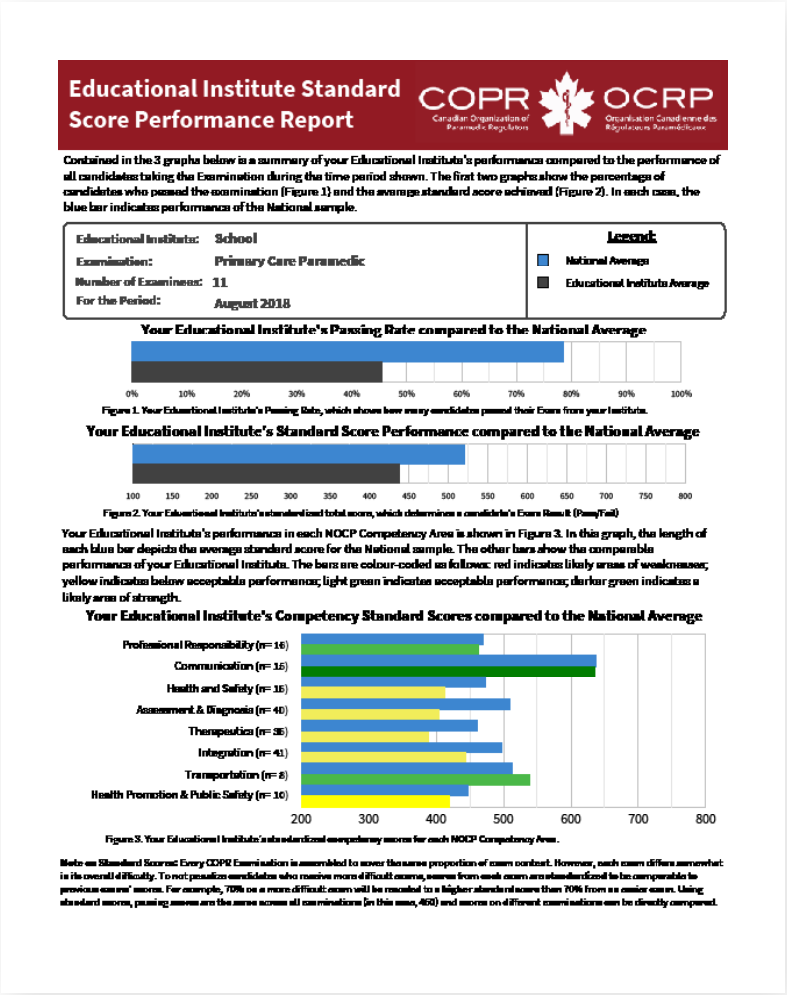
Note on Standard Scores: Every Examination is assembled to cover the same proportion of exam content. However, each exam differs somewhat in its overall difficulty. To not penalize candidates who receive more difficult exams, scores from each exam are standardized to be comparable to previous exams' scores. For example, 70% on a more difficult exam will be rescaled to a higher standard score than 70% from an easier exam. Using standard scores, passing scores are the same across all examinations (in this case, 450) and scores on different examinations can be directly compared.

To give you an idea of how such an aggregated report would look, consider the following example report:

In this example, we can see the average total examination score that candidates at the school in question received (relative to the average score candidates across the entire nation received). For ease of interpretation, these values are presented both in the original raw score units, as well as the scale score units. Below these values, we can also see the average (scale) score that candidates from this school received, relative to the average score candidates across the nation received. Note that this score is also broken down by each competency area on the exam. This information allows the school to get a sense of how their students did relative to students at other schools across the country; as well as diagnose which specific areas their program may be excelling (or failing) to prepare candidates for.

QUALITY ASSURANCE OF FINAL SCORE REPORTS

Earlier in this report, we already covered the importance of doing a solid quality assurance check for candidate scores. However, even if this process was carefully followed, the process of creating the reports may introduce further errors. Therefore, score reports should also be subject to a quality assurance check prior to their dissemination. This can involve checking the values in the report against the final (already assured) candidate data file, to ensure that the correct scores are associated with the correct names; and that the correct comparison values (e.g., national averages) are used.



Educational Institute Standard Score Performance Report

COPR Canadian Organization of Paramedic Regulators | OCRP Organisation Canadienne des Régulateurs Paramédicaux

Contained in the 3 graphs below is a summary of your Educational Institute's performance compared to the performance of all candidates taking the Examination during the time period shown. The first two graphs show the percentage of candidates who passed the examination (Figure 1) and the average standard score achieved (Figure 2). In each case, the blue bar indicates performance of the National sample.

Educational Institute:	School
Examination:	Primary Care Paramedic
Number of Examinees:	11
For the Period:	August 2018

Legend: National Average; Educational Institute Average

Your Educational Institute's Passing Rate compared to the National Average

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Figure 1. Your Educational Institute's Passing Rate, which shows how many candidates passed their Exam from your Institute.

Your Educational Institute's Standard Score Performance compared to the National Average

100 150 200 250 300 350 400 450 500 550 600 650 700 750 800

Figure 2. Your Educational Institute's standardized total score, which determines a candidate's Exam Result (Pass/Fail)

Your Educational Institute's performance in each NOCP Competency Area is shown in Figure 3. In this graph, the length of each blue bar depicts the average standard score for the National sample. The other bars show the comparable performance of your Educational Institute. The bars are colour-coded as follows: red indicates likely areas of weaknesses; yellow indicates below acceptable performance; light green indicates acceptable performance; darker green indicates a likely area of strength.

Your Educational Institute's Competency Standard Scores compared to the National Average

Professional Responsibility (n= 16)
Communication (n= 16)
Health and Safety (n= 16)
Assessment & Diagnosis (n= 40)
Therapeutics (n= 35)
Integration (n= 41)
Transportation (n= 8)
Health Promotion & Public Safety (n= 10)

200 300 400 500 600 700 800

Figure 3. Your Educational Institute's standardized competency scores for each NOCP Competency Area.

Note on Standard Scores: Every COPR Examination is assembled to cover the same proportion of exam content. However, each exam differs somewhat in its overall difficulty. To not penalize candidates who receive more difficult exams, scores from each exam are standardized to be comparable to previous exams' scores. For example, 70% on a more difficult exam will be rescaled to a higher standard score than 70% from an easier exam. Using standard scores, passing scores are the same across all examinations (in this case, 450) and scores on different examinations can be directly compared.

OTHER CONSIDERATIONS IN SCORE REPORTING

Including Sub-Scores for Specific Competency Area

One of the main uses of a score report is to help candidates diagnose where they went wrong. It's often the case that candidates want to learn more about their strengths and weaknesses in the area that is being tested.

To meet this need, many certification and licensure tests include summaries of candidates' scores for each competency that is tested in the exam. This information is meant to help unsuccessful candidates identify the areas where they need to study further in order to be successful in the future. Following this idea, some testing organizations only provide these reports to candidates who are unsuccessful. That said, these reports can do more than just help failing candidates diagnose their areas for improvement; they can also help successful candidates identify possible areas for their own future professional development.

Although these sub-area scores can be very informative, it bears noting that a minimum number of items within each category is usually required before any such diagnostic interpretations can be statistically justified. The reason for this is twofold. First, there is a probabilistic (bot not certain) relationship that exists between candidate ability and performance on test questions. That is, increases in candidate ability increase the likelihood that test takers will answer questions correctly; but it does not guarantee a correct response. As a result, the more questions that are asked, the more confidence can be vested in the conclusion that higher scores really do indicate higher ability.

The second reason is that performance on individual test questions are intended to generalize to a larger domain of knowledge, skill, and abilities. Because this domain spans a range of content, several questions are required in order to effectively assess candidates on all aspects of the domain being tested (as opposed to just a narrow facet thereof).

Unfortunately, there are no definitive guidelines as to how many items are required before sub-scale reporting can be considered reliable. Some factors to consider include the discrimination indices of the individual items (for a reminder of what item discrimination is, please refer to the previous white paper in this series). though in most cases, it's the discrimination relative to the rest of the subscale items that is important, not relative to the test domain as a whole. In any case, the higher the discrimination, the fewer the numbers of items required. The complexity of the subscale is also a key factor. When the domain defined by the subscale is small, fewer items are required.

Providing Information on Candidate Sub-Groups

From the perspective of the schools or testing organizations, it is often helpful to consider performance differences among different groups of candidates. For example, a testing organization may be interested in how candidates from different schools did (in order to shed light on differences in the quality of education and learning across educational institutes); or may be interested in demographic differences, such as domestic versus international students.

As with subscale reporting, a minimum number of observations is required for effective sub-group reporting. The larger the number of observations (individuals) within the groups being compared, the more likely that observations will relate to group membership (rather than being observed due to chance alone). Moreover, having more data points helps to keep candidate results confidential (as it may be possible to deduce the performance of specific individuals if there are not many in individuals in the subgroup being reported on).

Next Stage | Standard Setting

The next and final stage in the Assessment Life Cycle surrounds the process of establishing a passing point for the newly-created exam (called “standard setting”). Using a process such as the Angoff or Ebel method, subject matter experts rate the difficulty of the examination content; and compare this to an agreed-upon definition of the minimally-competent candidate This stage will be discussed in more detail in the next (and final) white paper in this series.

Conclusion

In summary, the Assessment Life Cycle is a way of organizing the processes involved in creating valid assessments into a series of easy-to-understand, logical stages. The focus of this white paper was to detail the fundamental steps and key processes that are involved in the sixth of these stages (i.e., scoring and reporting). As covered, this involves a number of specific steps, including:

- Reviewing candidate data to ensure its completeness and accuracy
- Scoring candidate responses
- Creating scaled scores via statistical equating
- Creating candidate-level score reports
- Creating aggregated score reports
- Reviewing the newly-created reports to ensure their accuracy

Following these best practice steps – and the Assessment Life Cycle in general – will help ensure that your assessment program is valid and defensible; affording the greatest possible benefit to both your test-takers and your organization.

Let Meazure Learning help you apply the Assessment Life Cycle to your assessment program. Meazure Learning offers a full range of products and services that cover every step and process. Our clients agree: we know testing; and we will work hard to make sure that your testing program is the best that it can be.

To explore this opportunity—or for more information—please feel free to contact us at:
meazurelearning.com/services

List of psychometric services offered in Assessment Life Cycle Stage 6

At Meazure Learning, we provide a host of services to our clients that encompass each of the Assessment Life Cycle stages. Below is a list of psychometric services that Meazure Learning offers specifically for **Stage 6: Scoring and reporting**

Service	Description
Candidate Scoring Services	Once candidate exam data has been collected, it must be scored, so that it can be determined how candidates did relative to the established passing point on the exam. At Meazure Learning, we offer a variety of scoring services that follow best practices to ensure the accuracy and validity of our clients' exam results. This includes redundant scoring across multiple platforms; as well as independent quality assurance checks to verify the accuracy of the results. We offer these services for a variety of different types of assessments, including both objective- and subjectively-scored assessment metrics.
Score report creation	The key to effective score reporting is to communicate candidates' results in the most timely, accurate, clear, and valid manner possible. At Meazure Learning, we offer a wide variety of tailored score report services to our clients. This includes candidate-, school-, and national-level reports that include a variety of metrics to help ascertain and diagnose candidate performance.
Statistical equating and scale score creation	It is often advisable to convert candidates' raw exam scores into a standardized scale score (so that scores are comparable across different testing administrations). Doing so also allows a testing organization to set passing points on two forms of an exam that are tailored to the relative difficulty of the exam items. At Meazure Learning, we have expertise in a variety of equating methods, including mean, linear, percentile, circle arc, and IRT-based methods. We are able to leverage these approaches to help ensure that candidate scores for our client's exams transcend any single exam administration; and can be effectively compared and communicated.