WHITE PAPER

Rapid Item Development (RapID™): Using Intelligent Templates to Fast-Track Item Bank Expansion

Lisa Sallstrom, CAE, PMP
Vice President, Certification & Membership

Gabriela Welch, CAE
Director, Certification

Frank Perna, CAE
Senior Manager, Test Development

Carolina Cruz
Manager, Test Development

Rachael Jin Bee Tan, PhD
Psychometrician

ASCM
ASSOCIATION FOR SUPPLY CHAIN MANAGEMENT

Abstract

Rapid Item Development (RapID™) is a new, patent pending methodology that uses intelligent templates to accelerate the development of exam questions for high-quality standardized tests and certification exams. Developed by ASCM, the Association for Supply Chain Management (formerly APICS), RapID™ is used with calculation-based items as an approach to quickly develop new items with known statistical properties[1]. It addresses many of the common challenges and inefficiencies in test publishing and exam development. RapID™ utilizes some of the theory behind Automated Item Generation (AIG), while eliminating the need to continuously pretest newly-created items. As discussed in more detail below, RapID™ uses a cloning process to develop psychometrically-sound items from an intelligent template. ASCM conducted a pilot program to demonstrate the RapID™ approach, and the outcomes are provided here to reveal learnings and best practices.

[1] Although it is impossible to predict the exact statistical performance of new items, RapID™ methodology has been shown to produce new items that consistently perform within the predicted range on multiple metrics.

Introduction

Why is there a need for a new rapid item bank expansion approach? When managing a professional certification program, it is considered good business practice to ensure the relevancy of content with respect to changing industry needs. According to a 2015 Financial Times article[2], "We have no choice but to match our own pace of work to the demands of a superfast globalized business world," argues Sir Martin, Chief Executive of Marketing Services Group WWP. "You have to be responsive; you shouldn't attempt to fight it or slow the pace down." In the certification business, market acceptance of a program can be tied to its relevancy and how well the certification reflects current industry job requirements. The rate of change can vary depending on the industry. For example, keeping pace with innovation and advancement in the computer technology or medical field may require content updates more frequently than other industries that change at a slower pace.

Furthermore, a robust pool of items increases test security and reduces the impact of a security breach because new items can be quickly substituted with minimal disruption to exam administration. Organizations can also reduce item exposure by having more items from which to choose during exam form development.

RapID™ item development can support certification business advancements such as adaptive testing, which requires at least three times[3] as many items to test for miniscule variances in difficulty while covering the entire breadth and depth of the content outline. Another benefit of RapID™ is that it has the potential to reduce language translation costs, because translation of one root item (template) can enable the staff to create many clones. For almost all organizations, an increase in item production directly corresponds to an increase in test development costs. Depending on the item writing process, the overhead cost of developing a single statistically-verified, scored item can range from several hundred to several thousand dollars. And to achieve scored status, new items require pretesting, which involves statistical validation and management of test publishing cycles and pretest tails/unscored item sets to obtain maximum throughput. RapID™ can help eliminate the need to repeat pretesting for variations on a specific topic or methodology, which increases opportunities to pretest different item types and levels of thinking.

RapID™ is a strategic approach to quickly develop new items that do not have to undergo the pretest process before being used in scored positions on an exam. This reduces test development costs while keeping exam content current and increasing test security. In addition, the ramp-up of test items and a growing item bank will allow organizations to expand their testing windows, even as far as expanding to on-demand testing. This can be a significant benefit to the organization as well as to candidates. With on-demand testing, candidates have the flexibility of taking the exam on their schedule, without missing the opportunity to test, or needing to re-study if they missed set windows. In addition, on-demand testing allows the opportunity for more frequent statistical analysis, approval of statistically valid exam items, and decreased risk of item exposure, among the many benefits.

[2] Financial Times (2015). Managing Fast & Slow in a World that Keeps Accelerating. Emma De Vita Retrieved from https://www.ft.com/content/30aa29f0-9a5d-11e4-8426-00144feabdc0

[3] Thompson, Nathan A., PhD. (2014). Requirements for Adaptive Testing. Retrieved from http://www.assess.com/wp-content/uploads/2014/03/Requirements-for-CAT-Testing.pdf

Traditional Item Writing Processes

In most high-stakes certification programs, items are typically generated in one of two ways: either through conventional item writing workshops, or by contracting with expert item writers serving as consultants. Both processes have become the industry standard, and both demand a significant amount of organizational resources, making them ripe for innovative advances in their approach.

Traditional item writing processes have benefits and drawbacks when compared and contrasted with innovative item writing processes such as RapID™.

Conventional Item Writing Workshops

In conventional on-site item writing workshops, an organization typically recruits a diverse group of eight or more subject matter experts (SMEs) and brings them together in one physical location. Over the course of a few days, this group will work face to face to develop new, raw items for pretesting. To recruit SMEs, an organization typically reaches out to its pool of certified or licensed individuals to seek volunteers. These writers are often incentivized through a combination of professional development credit, honoraria, travel reimbursement, and/or networking opportunities. They typically receive training prior to the item writing task, which might include organizational style guide standards. More seasoned item-writing volunteers often serve as coaches or mentors for newer, less-experienced writers.

An advantage of this process is that all volunteers are in one place, lending itself to less distraction and more motivation to complete tasks, thereby developing many new items at a given time. This process also serves as an opportunity for organizations to build and assess their base of SMEs for future engagement opportunities. While face-to-face item writing workshops are the most common approach to new item development, there are areas of constraint associated with the process. Availability of volunteers is a big concern, and finding individuals that have the necessary skills, as well as several days to commit to the project, can be a challenge. According to an article that appeared in The NonProfit Times[4] , volunteering is at a 10-year low. Volunteers' time has become increasingly scarce and it has become more and more difficult to engage volunteers. Cost is also a significant consideration. In addition to the inherent cost of meeting logistics (travel, food, meeting space, etc.), there is the cost of staff time and resources required to oversee the process. It can be a labor-intensive endeavor for many organizations.

Contracted Item Writing Experts

A common alternative to in-person workshops is for organizations to hire one or more external SME contractors to write new items. Expectations, payment and terms are detailed in a contract, along with the requested number of items and associated content domains. This option is usually chosen for item development throughput expediency (conversion to scored items), which allows the exam development committee to focus on form content quality and review. Contracting is also a good option for organizations that lack the resources to manage face-to-face item writing workshops.

However, when using a contractor, it may be difficult to identify SMEs who are able and willing to write the number of items needed within the specified time frame. As a workaround, organizations may choose to contract with a test development and/or administration vendor, which can prove to be costly and unsustainable in the long term. With both the face-to-face writing workshop and contracted remote writers, all newly-written items must go through a pretesting process to ensure they are psychometrically valid before becoming operational as a scored item. This is a lengthy process which usually takes anywhere from 12-24 months, largely dependent on exam volume and delivery processes to generate the minimum number of required item exposures for psychometric validity.

[4] Clolery, Paul. (2014). Troubling Numbers in Volunteering Rates. Retrieved from https://www.thenonprofittimes.com/npt_articles/troubling-numbers-in-volunteering-rates/

Innovative Item Writing Processes

Automation is transforming the way item banks are developed, supplementing traditional item development activities with a significantly faster and more efficient process. Leveraging technology, testing organizations are becoming able to diversify the means by which new items are developed.

Automated Item Generation

Automated Item Generation (AIG) is a relatively new process that organizations are exploring to mass generate exam items with the assistance of computer technology. It typically requires SMEs to create complex cognitive models that are used to develop item templates from which dozens, sometimes even hundreds, of items can be produced. Since this process is very different from traditional item writing methods, staff and SMEs must be trained in developing a cognitive modeling procedure and how to use specialized software to generate item clones. Often AIG software must be purchased or licensed, but in some cases, organizations choose to create and maintain their own proprietary software, which results in additional overhead costs. Furthermore, items generated through AIG still require a pretest period to validate statistics prior to converting them to scored. Therefore, while this procedure can exponentially generate a large number of items, it is still constrained by a lengthy pretest timeline.

Rapid Item Development

The RapID™ methodology introduced here also uses a systematic approach to item cloning, an exciting and innovative way to quickly augment the item pool. Unlike AIG, however, once a process is established to ensure psychometric viability, new items can be automatically generated via a tested item template with little strain on resources[5].

RapID™: Item Cloning Without Pretesting

Rapid Item Development capitalizes on the advantages of a cloning process that overcomes the need for any pretesting required to develop scored items. RapID™ entails a four-step process. Once these steps are completed, new items can be created from the initial template without the need for further pretesting.

1. Identify or create a root item.
2. Develop a template from the root item.
3. Clone additional items from the template.
4. Conduct statistical validation of initial clones.

We will outline each of the four steps below, walking through the RapID™ methodology from the formation of a root item through the statistical vetting of cloned items.

[5] A proven item template is one that produced items displaying similar statistical properties, verified by the item difficulty and discrimination parameters gathered from pretesting.

Step 1: Identify or create a root item.

The RapID™ methodology begins with the creation or identification of a root item that provides the starting point for development of the item template. Some guidelines are listed below for identifying the root item:

1. When creating or identifying a root item, ensure that the item is relevant to the exam body of knowledge, aligns with the exam blueprint, and provides value to the content that is being tested.
2. As with any newly-created item, follow pre-established organizational item writing style guidelines.
3. To create a dynamic template from which many item clones can be generated, it is best for the root item to have several different variables that can be manipulated. For example, ASCM has many items that present numerical inputs to use for calculation purposes. These numerical inputs are used to create formulas for the key and each plausible distractor.

The example below illustrates the importance of the variety in numerical inputs.

The selection of a pre-existing scored item from your item pool has several advantages, including style guide adherence, linkages to the exam content outline, and previously- validated psychometric statistics. When selecting an existing item, choose an item that met psychometric standards during its most recent administration (based on your organization's acceptable ranges for performance statistics).

Refer to Figure 1 for an example root item on the Cost of Goods Sold (COGS) concept. This item was retired from the ASCM Certified in Production and Inventory Management (CPIM) Part 1 exam for supply chain professionals. It performed well while on the exam, meeting the psychometric standards for item difficulty and discrimination established by ASCM.

Figure 1: RapID™ Root Item:

The question below is based on the following information:

Revenue	$1,000
Material	$360
Overhead	$60
Labor	$180
General and Administrative (G&A) Expenses	$200

Which of the following amounts represents the costs of the goods sold (COGS)?

A. $420
B. $600*
C. $620
D. $800
*represents the answer key

Step 2: Develop a template from the root item.

To develop a template from the root item, the key calculation must be known, as well as calculations to generate all the remaining plausible distractors. Strong distractors help ensure that an item discriminates well, and that test-savvy candidates are not able to guess the correct answer solely by process of elimination. For example, if there are only two ways to manipulate the variables in the item stem, the root item is too easy because more than one distractor will be implausible and can be quickly eliminated as incorrect.

The template should identify the variables, provide the calculation and rationale for each answer option, (i.e., key and distractors), and define any variable constraints. Variable constraints help ensure distractors are plausible and the item stem provides realistic information. If computer software is used, constraints are required to define the range of potential values for each variable. If SMEs are asked to clone items from a template, variable constraints promote standardization and provide additional quality control. Please refer to Figure 2 for the template developed from the root item in Figure 1. In this example, the constraints that must be followed are listed with each variable to ensure that plausibility is maintained. The placeholder shows the variable combination used to calculate each answer option.

Figure 2: RapID™ Item Template:

The question below is based on the following information:

Revenue	$V (must be greater than w+x+y)
Material	$W (must be greater than x and y)
Overhead	$X (must be less than w)
Labor	$Y (must be less than w)
General and Administrative (G&A) Expenses	$Z (must not less than v)

Which of the following amounts represents the costs of the goods sold (COGS)?

Placeholder	Rationale
A. W+X	(Material + Labor)
B. W+X+Y	(Material + Labor + Overhead)*
C. W+X+Z	(Material + Labor + G&A)
D. W+X+Y+Z	(Material + Labor + Overhead + G&A)

Step 3: Clone additional items from the template.

Using an established template, one will be able to create multiple clone items. Clones should be identical in format, with the only changes made being to the different item variables. It is also extremely important to adhere to the pre-determined variable constraints to ensure that all aspects of the new item remain plausible. Even small changes to the template's language, format or presentation may result in variability in the statistical performance of clones. Because creating new scored items without pretesting is one of the goals of RapID™, it is critical to follow the item template. Figure 3 shows an item clone that was created from the template in Figure 2.

Figure 3: Item created by cloning RapID™ Template:

The question below is based on the following information:

Revenue	$1,500
Material	$400
Overhead	$50
Labor	$155
General and Administrative (G&A) Expenses	$220

Which of the following amounts represents the costs of the goods sold (COGS)?

A. $450
B. $605*
C. $670
D. $825
*represents the answer key

Step 4: Verify statistical performance of clones.

As with any newly-developed item, the items cloned using the RapID™ process should undergo an initial statistical analysis to validate the performance of a template. Once the performance of a template item is validated, multiple clones can be created and used in scored positions on future exam administrations without needing to pretest. Approved item templates are those whose clones perform within a psychometrically- acceptable range to the root item on multiple statistical measures.

It is recommended that the performance of at least three clones be verified before approving the use of a template for mass item generation. These three items can be referred to as beta clones, with subsequent clones becoming immediately operational after successful performance of the beta clones has been verified. To establish a consistent testing environment, it is recommended that organizations administer the beta clones concurrently using multiple pretest tails/sets on the same or parallel base forms. This data collection design helps protect against sample changes and other sources of variance that may be introduced over time. If concurrent testing is not possible, there should be an individualized plan for administering at least three beta clones from the same item template during a reasonable timeframe.

To ensure the success of the RapID™ process, it is critical to verify the statistical performance of beta clones, which includes having a large enough sample size to draw defensible conclusions when interpreting pretest results. Each of the beta clones must be administered to an adequate sample of candidates before making statistical comparisons of performance across the clones and root item. All ASCM exams have relatively high candidate volume, which permits the use of item response theory (IRT) scoring. IRT is a powerful statistical model that allows for sample-independent comparisons of candidate and item performance. To maintain a stable IRT scoring scale, ASCM typically collects a minimum of 250 candidate responses to each pretest item before running statistical analyses on examination data.

Approving an item template occurs by comparing the actual and predicted performance of the beta clones on three statistical indices. The first index is the IRT item difficulty, or b parameter. In general, item difficulty values for an ASCM exam range from -4 to +4 logits, with higher values indicating more difficult items.

In addition to judging clone performance by the IRT b parameter, ASCM also compares clone performance using classical test theory (CTT) statistics. Unlike IRT parameters, CTT statistics are sample-dependent and vary depending on the proficiency level of the candidates taking the exam. An item's CTT difficulty value (*p*-value) reflects the proportion of candidates who answered the item correctly on a single exam form during a specific administration window. An item's *p*-value is 0 if no candidates answered the item correctly, and 1 if all candidates answered it correctly.

The second CTT index is an item's discrimination value, which represents the correlation between item and exam performance. ASCM measures item discrimination using the point-biserial correlation coefficient. Discrimination values range from -1 to +1. A +1 occurs when all high performers (large total exam score) answer an item correctly and all low performers (small total exam score) respond incorrectly. The inverse results in a discrimination of -1. Larger discrimination values are desirable because they indicate a strong, positive relationship between answering an item correctly and performing well on the examination.

What constitutes similar item performance across the three indices must be determined by the organization and should be based on existing psychometric/statistical guidelines and thresholds used in decision making, such as those used when assessing item quality.

ASCM uses the following statistical guidelines to help determine whether the beta clones and root item are performing similarly:

- The *p*-values should be within +/-0.10 of each other.
- The discrimination statistics should be within +/-0.15 of each other.
- The IRT b values should be within +/-0.60 logits of each other.

Among the above three statistics, ASCM relies most heavily on the IRT difficulty value to determine whether to approve an item template. For the IRT difficulty value, similar performance is achieved if the absolute value of the displacement statistic (the difference between the actual and predicted difficulty of the item) is less than 0.60 logits. This is the same threshold ASCM uses when evaluating item performance to decide when to "un-anchor" a scored item's difficulty parameter when calibrating pretest items[6]. It is not recommended for organizations to adopt these thresholds without considering their existing psychometric guidelines for assessing item quality. An organization's psychometric staff or qualified consultants should be involved when deciding what is the most defensible option for assessing clone performance and approving templates for mass item generation without pretesting.

Pilot Study

ASCM recently conducted a pilot test of the RapID™ process using 13 item templates. These item templates were developed from 12 statistically-proven items and one new item that was identified as a good candidate for a root item. Three beta clones were created from each template and concurrently tested on different pretest tails appended to the same scored set of items. The forms were administered over six months and an average of 250 candidate responses were gathered on each beta clone.

Tables A, B, and C, shown below, contain the *p*-values, point-biserial correlation coefficients, and IRT b values of the root items and beta clones. Each table also provides the maximum observed difference in each statistic's performance across the following categories: (1) all items, (2) the root item and the individual clones, and (3) among the clones.

[6] O'Neill, T., Peabody, M., Tan, R.J.B., Du, Y. (2013). How much item drift is too much? Rasch Measurement Transactions, 27:3, pages 1423-1424.

CTT Item Difficulty Results

Concerning CTT item difficulty, nine of the 13 templates met the ASCM criteria: "The *p*-values should be within +/- 0.10 of each other," according to the statistics across all items, which calculates the maximum *p*-value difference between all pairs of four items for a template. The slightly larger difference in observed *p*-values for templates 4, 6, and 9 (0.12) was observed between a clone and the root item, with all clones for templates 6 and 9 performing very similarly to one another. This suggests that the differences may be influenced by the age of the root item calibration, with clones from older items being easier due to exposure over time to the template. The clones for template 13, which was new and did not have a root item, all performed differently from one another, indicating this template is not appropriate for mass clone generation.

Table A: Proof of Concept: *p*-value Results

Item	Root Item	Clone 1	Clone 2	Clone 3	Max Difference
1	0.86	0.83	0.88	0.88	0.05
2	0.76	0.81	0.83	0.83	0.07
3	0.65	0.69	0.67	0.63	0.06
4	0.59	0.71	0.63	0.66	0.12
5	0.72	0.68	0.76	0.74	0.08
6	0.83	0.93	0.95	0.95	0.12
7	0.54	0.52	0.46	0.51	0.08
8	0.72	0.72	0.74	0.73	0.02
9	0.56	0.68	0.65	0.65	0.12
10	0.87	0.91	0.89	0.88	0.04
11	0.88	0.9	0.89	0.89	0.02
12	0.85	0.94	0.93	0.92	0.09
13	New	0.73	0.67	0.55	0.18

Average difference: 0.04 **Median difference: 0.03**

CTT Item Discrimination Results

Concerning CTT item discrimination, all but two pilot templates met the ASCM criteria: "The discrimination statistics should be within +/-0.15 of each other." Templates 5 and 7 had varying discrimination values across the clones, but only for template 7 were these values unacceptably low (i.e., 0.10 and 0.05 for Clones 1 and 2). All other discrimination values were within acceptable performance ranges. Since discrimination is not used during the exam scoring process, the monitoring of this statistic is important to ensure acceptable value ranges are consistently observed.

Table B: Proof of Concept: r_{pb}-value Results

Item	Root Item	Clone 1	Clone 2	Clone 3	Max Difference
1	0.47	0.53	0.41	0.42	0.12
2	0.47	0.53	0.47	0.45	0.08
3	0.29	0.33	0.27	0.31	0.06
4	0.40	0.48	0.39	0.41	0.09
5	0.36	0.43	0.20	0.40	0.23
6	0.41	0.46	0.31	0.38	0.15
7	0.30	0.10	0.05	0.37	0.32
8	0.45	0.38	0.43	0.35	0.10
9	0.32	0.31	0.31	0.34	0.03
10	0.33	0.34	0.28	0.39	0.11
11	0.36	0.38	0.43	0.42	0.07
12	0.37	0.28	0.35	0.38	0.10
13	New	0.27	0.33	0.27	0.06

Average difference: 0.07 **Median difference: 0.05**

IRT Item Difficulty Results

Ten of the 13 pilot templates performed within ASCM's acceptable bounds for IRT item difficulty, which requires IRT b values to be within +/-0.60 logits of each other. Template 13, which performed poorly according to the CTT item difficulty criteria, also had varied IRT difficulty values for all clones, further demonstrating it is not a good candidate for an item template. Templates 6 and 12 produced clones that performed similarly to each other and were within acceptable IRT difficulty ranges, with the maximum difference resulting from comparisons between the clones and the root item.

Table C: Proof of Concept: IRT Results

Item	Root Item	Clone 1	Clone 2	Clone 3	Max Difference
1	-0.89	-0.55	-0.94	-0.99	0.44
2	-0.33	-0.39	-0.56	-0.57	0.25
3	0.37	0.35	0.44	0.58	0.23
4	0.72	0.22	0.59	0.42	0.51
5	0.12	0.39	-0.06	0.01	0.44
6	-0.51	-1.67	-2.00	-1.99	1.49
7	0.92	1.13	1.38	1.12	0.46
8	0.12	0.20	0.02	0.08	0.17
9	0.71	0.37	0.51	0.50	0.34
10	-1.01	-1.30	-1.08	-1.03	0.29
11	-1.21	-1.16	-1.12	-1.11	0.09
12	-0.82	-1.74	-1.66	-1.57	0.92
13	New	0.09	0.44	0.94	0.84

Average difference: 0.26 **Median difference: 0.17**

It is important to note that some initial pretesting is still required, and item performance must still be monitored on an ongoing basis. Best practice for optimal item management recommends that the performance of the items be checked periodically to ensure clones are performing as expected and meet organizational psychometric standards[7]. ASCM conducts IRT fit analyses whenever pretest items are calibrated, which is the optimal time to confirm scored item performance and flag any clones that deviate from expected values. While it is critical to ensure that cloned items continue to perform similarly to each other and the root item, content should also be monitored to ensure templates stay current. As with any item, clones can eventually become obsolete based on the advancement of the exam program and changes in the field.

For ASCM, the potential of items generated from a template is tied directly to the number of plausible numerical inputs. More complex templates with several variable options can directly affect and increase the number of potential cloned items that can be generated from that template. Once the process is operationalized within an organization's test development approach, SMEs can focus on the creation of templates, allowing item bank growth to occur exponentially.

[7] Knapp, J., Anderson, L., Wild, C. (2015). Certification: The ICE Handbook, 2nd Edition, Chapter 7, pages 166

Discussion

The goals of credentialing organizations can differ quite drastically, but the need for item growth is a challenge that is shared industry-wide. The RapID™ methodology for fast-tracking item bank expansion is a viable, statistically valid means for achieving this growth. The successful outcome of the ASCM pilot study proves that RapID™ is an effective process for organizations to meet ever-changing business needs and achieve a number of strategic business goals, including strengthening exam security, decreasing long-term item development costs, and maximizing volunteer efficiency. Through the use of RapID™ templates, key concepts and content areas can be introduced more quickly and tested more efficiently, thereby allowing organizations to accurately assess candidate ability in an accelerated business environment.

The RapID™ Item Development process is a proprietary, patent-pending methodology owned by ASCM. For more information or details on licensing the process, please contact Certification@ascm.org

ABOUT ASCM

The Association for Supply Chain Management (ASCM) is the global leader in supply chain organizational transformation, innovation and leadership. As the largest nonprofit association for supply chain, ASCM is an unbiased partner, connecting companies around the world to the newest thought leadership on all aspects of supply chain. ASCM is built on a foundation of APICS certification and training spanning 60 years. Now, ASCM is driving innovation in the industry with new products, services and partnerships that enable companies to further optimize their supply chains, secure their competitive advantage and positively influence their bottom lines. For more information, visit ascm.org.

ASSOCIATION FOR SUPPLY CHAIN MANAGEMENT