Journal of Computerized Adaptive Testing

Volume 13 Number 2
June 2026

Item Selection Rules for Content Adaptive Progress Testing

Gergo Pinter, Nuno Santos, José Miguel Pêgo, Daniel Zahra, Steven Ashley Burr

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 13, Number 2, June 2026

DOI 10.7333/2606-1302007

ISSN 2165-6592

Item Selection Rules for Content Adaptive Progress Testing

Gergo Pinter, University of Plymouth, UK
Nuno Santos, IT Solutions, Braga, Portugal.
José Miguel Pêgo, University of Minho, Braga, Portugal.
Daniel Zahra and Steven Ashley Burr, University of Plymouth, UK

Content adaptive progress testing (CAPT) introduces a new approach to longitudinal testing of candidates who take progress tests as part of their educational program. CAPT assessments adapt between subsequent assessments, tailoring items on an individual level to assist each candidate in demonstrating knowledge across many topics which are all mapped to the overall course learning outcomes. In this way, success is measured by topic completion on an individual level. Some of the key benefits include a truly personalized learning experience, with individual feedback throughout, while also tracking each candidate's progress against a single long-term goal of demonstrating attainment across all required topics. For the learning experience to be personalized, there needs to be automatic software selection of different items for each candidate based on their performance across a sequence of assessments. The advantages, disadvantages and practical implementation of different item selection rules are discussed. There is a need to ensure that no item is repeated to the same candidate in any test: No two items in the same test belong to the same topic for the same candidate, a minimum number of topics per broader area are administered in every assessment, and selection of incomplete topics is prioritized over complete topics. These rules need to be clear and transparent to motivate appropriate learning.

Keywords: adaptive testing, assessment, content adaptive progress testing, fairness, item selection, personalized testing

Content adaptive progress testing (CAPT) refers to the process of selecting items for a candidate's next assessment based on their performance in previous assessments. As such, CAPT assessments are adaptive *between* assessments, rather than other forms of computer adaptive testing (CAT) that are adaptive *within* an assessment—selecting the next item presented to the candidate based on the level of difficulty correctly answered in previous items (Burr et al, 2023). CAPT has been implemented for summative assessment of the 2023-24 cohort at Peninsula Medical School (PMS) with an aim of delivering personalized assessment aligned to each individual student's progress and stage of learning (Burr et al., 2022). While highly applicable in medical education, the concepts of CAPT are generalizable to other longitudinal assessments and have been presented without specific reference to medical education where appropriate. The specifics of the PMS CAPT system are detailed in a separate section for reference.

There are several advantages of introducing CAPT: (1) Personalized assessment and revisiting content where a student underperforms provides scaffolding of learning with respect to the zone of proximal development principle (Allal & Ducrey, 2000); (2) revisiting content to assure attainment across all learning outcomes (topics) without allowing compensation; (3) personalizing diagnostic feedback to optimize learning; (4) learning can be clearly and transparently aligned to the curriculum and students' development; (5) students can complete assessments at different times; and (6) items can be reused more efficiently across the years and cohorts of students. Potential disadvantages include that: (1) it might be more difficult to identify narrow content areas that cohorts of candidates are struggling with at a specific time because they do not all see the same items at the same time; and (2) it might be more difficult to identify specific items which might need to be reviewed.

With traditional (non-adaptive) progress testing, all candidates take several assessments each year, where each assessment is integrated and computer-delivered, with all candidates in all stages taking the same items that are all set at the qualifying standard (Schuwirth & Van der Vleuten, 2012). Thus, with progress testing, the required standard is transparent from the start of the program and growth in knowledge toward achieving the final standard is monitored throughout the program. One criticism of progress testing has been its misalignment with candidates' progression in education and lack of individualized feedback that is provided to candidates. The main reasons for implementing CAPT are a desire to both (1) support learning and (2) assure attainment in each topic that can be assessed in a final achievement test, such as a national licensing examination. The aim is to ensure that every candidate attains an acceptable breadth of knowledge that is comprehensive by eliminating compensation; and thereby to help all candidates become safe generalists prepared for their first job in clinical practice (Burr et al., 2022).

In contrast, when candidates are administered adaptive progress tests, they are generally presented different subsequent items based on their past performance in that assessment. CAPT differs from these assessments in two primary ways. First, CAPT assessments are only adaptive between assessments. Items are tailored on a candidate-specific basis, but this adapting happens before (and not during) an assessment, such that all the items to be administered are set before the assessment is administered.

The second main difference is in the mechanism of adapting items. There already exist some adaptive testing formats which aim to ensure topic coverage of items, such as cognitive diagnostic adaptive testing (CD-CAT) (Collares, 2022), Multistage testing (MS) (Yan, et al., 2014), and the National Council Licensure Examination for Registered Nurses (NCLEX-RN®) (NCSBN, 2023). However, these all still involve difficulty adapting to select the next item or group of items. In contrast, CAPT ensures topic coverage without difficulty adapting and requires each candidate to achieve across a blueprint (achievement in all areas, with no compensation across topics) rather than determining a level of ability overall (CAT, NCLEX-RN®), or within topic areas (CD-CAT), or across modules (MS). In this way, it is also distinct in that it emphasizes topic coverage over assessment of mastery of given topics (e.g. Bloom, 1968).

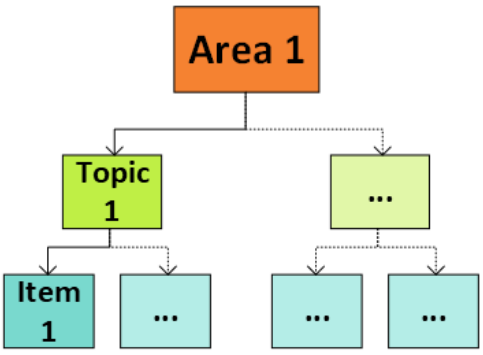
CAPT Structure

Adaptation in assessment must be done in a structured way to ensure understanding across stakeholder groups, transparency, and quality assurance of the process. This can be facilitated by adopting a hierarchical knowledge structure within the item bank, alignment of topics across the curriculum and assessment, and clear categorization of items within the bank from which assessments are built.

Item Bank Structure

As with traditional progress testing and CAT, CAPT relies on a large item bank of multiple-choice items. The content covered in the Peninsula Medical School (PMS) CAPT assessments are organized in a hierarchical structure, as depicted in Figure 1. At the top level, there are areas. Each area covers a large section or module of teaching and corresponding knowledge. Each area is further divided into topics. Each topic falls under only a single area and covers specific learning objectives. Finally, each unique item falls under a single topic. For example, in a medical program, assessment *items* on cardiac arrest would fall under the "Cardiac Arrest" *topic*, which could belong to the "Cardiovascular System" *area*. To allow for multiple items on the same topic, each topic is populated with multiple different items covering the same narrow area of knowledge incorporating a learning outcome. The use of areas can be omitted should there be little in common between the topics in an item bank since selection criteria are set at the topic level.

Figure 1. Example Item Bank Structure of Items and Topics Belonging to an Area
Each unique item falls under a single topic and each topic belongs to a single area



CAPT assessments are populated with items on an individual candidate level prior to each exam date and the content of these assessments depends on a candidate's performance on previous CAPT assessments.

Constructive Alignment

Constructive alignment is introduced in CAPTs through topic blueprinting, which restricts the appearance of items on some topics in specific assessments across the learning continuum. Each blueprint contains only the topics that are appropriate for the assessments being populated at each developmental stage. For example, it might be desirable to restrict possible items to only topics that are expected to have been covered by the time an assessment is scheduled. It is also possible to restrict entire areas if no topics that fall under an area are selected in the blueprint.

As candidates progress through their course, they move to a different blueprint with more topics included than the previous one. In this way, new topics can be introduced at the point by which they are expected to have attained knowledge on these topics. The number of topics selected in subsequent blueprints only increases, and once a topic is selected it remains selected on all future blueprints. In the final blueprint, all topics are selected so that candidates can be administered items on any of the topics in the CAPT item bank.

This focus on adapting by content is based on a curriculum topic map comprising topics that are taught and assessed. Constructing subsequent assessments individualized to candidates'

prior achievement in topic areas ensures achievement across all topics without compensation, and explicit alignment to the topics in the curriculum. The focus is also explicitly on adapting by content, not difficulty, as in other CAT models; and importantly, without trying to adapt by both simultaneously. In trying to achieve adaptation by both at the same time, one must be prioritized, such that practically, an assessment of finite length cannot contain the same range of item difficulties within the same range of content and still require achievement of every topic to a given level of ability (Burr et al. 2023; Luecht, De Champlain, & Nungester, 1998).

Topic Status

As candidates answer items on the various blueprinted topics, a topic status is assigned on a candidate-specific basis for all topics. This personalizes the value of feedback for the individual, allowing them to diagnose areas of underperformance.

Each time a candidate correctly answers an item, the associated topic is considered complete for them, even if they subsequently answer a different item on the same topic incorrectly in a later test. However, these proposed rules for what constitutes a complete topic could easily be altered to suit other applications. In its current implementation, all topics that are selected in the blueprint fall under one of three categories for each candidate: complete, incomplete, and unencountered topics where incomplete and unencountered topics both fall under the category of outstanding topics (yet to be completed).

Complete Topics.

1. The candidate has been administered an item on this topic at least once before.
2. On at least one of those occasions, the candidate answered the item correctly.

Outstanding Topics.

Incomplete topics.

1. The candidate has been administered an item on this topic at least once before.
2. The candidate has only given incorrect answers each time they have had an item on this topic.

Unencountered topics.

1. The candidate has never been administered an item on this topic before.
2. Topics that are not selected in the blueprint do not fall under this category. Instead, those topics are not assigned a status since items under those topics are not yet available for assignment.

Item Selection Rules

The validity of software automatically selecting items compared with manual selection has been previously established (Lunz & Deville, 1996). In addition to the hierarchical item structuring, blueprint and topic status assignment, a framework of six item selection rules (ISRs) was established to clearly define how items would be chosen to populate the next assessment for each candidate. Since the item selection process is candidate-specific, each of these rules applies to the selection of items for an individual candidate in a specific test.

Outline

Each ISR is summarized below with further details for each ISR following the summary. For ease of reference, there is an accompanying summary table (Table 1).

ISR1. The same unique item is not repeated in any subsequent test.

ISR2. Previously voided items (including all items from voided tests) are not repeated in any subsequent test.

ISR3. Items are only selected from topics included in the stage-specific blueprint.

ISR4. No more than one item can be selected from the same topic in each test.

ISR5. A minimum of three items from each area are included in every test, selecting from complete topics only if necessary.

ISR6. Further items are selected only from outstanding topics, selecting from complete topics only if necessary. The balance between the items selected from outstanding topics is kept so that the ratio of items *selected* from incomplete topics and unencountered topics matches the ratio of the ***remaining*** (outstanding) incomplete and unencountered topics.

These six ISRs are applied in a hierarchical fashion, with each rule introducing additional sequential constraints on the way that assessments are populated. When applying each rule, all previous rules are also followed. Each ISR is detailed below, with specific details about how each rule was implemented.

Table 1. A Summary of Item Selection Rules (ISRs) with Advantages and Disadvantages

Item Selection Rule	Advantage/Disadvantage
ISR1: No repeated items.	Advantage: Promote comprehensive study. Disadvantage: Requires a vast item bank.
ISR2: No reuse of voided items.	Advantage: Promote fairness. Disadvantage: Additional data processing.
ISR3: Use item blueprinting.	Advantage: Promote test-to-course content alignment. Disadvantage: Requires blueprint review and maintenance.
ISR4: One item per topic per test.	Advantage: Allow scope for improvement between tests. Disadvantage: Limits opportunities to complete some topics earlier.
ISR5: Include three items per area per test, favoring incomplete topics.	Advantages: Balance within a test and promoting comprehensive study. Disadvantage: Possible small reduction in opportunities to complete topics.
ISR6: Balance remaining items from incomplete and unencountered topics.	Advantage: Increased introduction of new topics. Disadvantage: Technical complexity and tuning requirements.

Rationale for Item Selection Rules, Their Advantages, Disadvantages, and Implementation

ISR1

As with conventional approaches to assessment, ISR1 ensures unique items are never shown to the same candidate twice. This rule encourages candidates to engage with a broader range of topics rather than try to recall specific items that had been administered to them previously.

Advantage. This promotion of comprehensive study is a more desirable trait when training generalist practitioners.

Disadvantage. This practice requires a vast item bank to provide new items for candidates throughout their time on the program.

In practice. To ensure unique items are never repeated, the items available for selection are limited to those that each specific candidate has never been administered previously.

ISR2

ISR2 builds on ISR1 and addresses the details of item or assessment removal for individuals or cohorts. When an item or assessment is voided, all affected topic completion is reset to the state before the assessment with voided item(s) took place.

For example, when voiding a specific item (e.g., due to a change in best practices or item exposure), the relevant topic status for any candidates who were administered that item reverts to whatever it was prior to the current test. Since the affected candidates have in effect been given fewer opportunities to complete a topic than candidates who were not administered voided items, this can be considered if progression is assessed based on the total number of complete topics.

The same logic for reverting progression applies if a whole assessment must be voided, though in this case it is also conceivable to offer a repeat assessment opportunity since item assignments are unique to each candidate.

To account for these scenarios, the details of all the items (including voided ones) answered by candidates are retained. This data is stored, ensuring that candidates will not get a repeated unique item in future tests, even if the item was previously voided.

Advantages. This rule promotes fairness for all candidates. As with all assessments, the post-assessment process might identify specific items (or whole tests) that should be removed. With CAPT, these items might only have appeared for a single or small subset of the candidates being administered a particular test. Assessment integrity can be upheld by implementing this rule and devising a way to continue the post-assessment process, as is established for traditional knowledge assessments. ISR2 is also important since voided items can be updated and deemed suitable for addition to future tests. However, it will still be necessary to not repeat similar items to the same candidate (as per ISR1).

Disadvantage. This rule introduces some additional data processing to separately store voided items.

In practice. ISR2 is implemented at the item assignment stage. During the filtering of previously administered items (ISR1), additional elimination removes previously void items from the bank of items that can be assigned on a candidate-level basis.

ISR3

ISR3 filters the assignable items through blueprinting. Each cohort's blueprint is tailored so that candidates are only administered items on topics appropriate to their progression.

Advantages. Narrowing item topics helps CAPT assessments align with the intended learning outcomes for each stage of the program. It also avoids testing on topics that candidates are not yet expected to have covered in either taught sessions or self-study.

Disadvantages. Blueprinting requires careful consideration and review to ensure that topic learning outcomes are covered. The rigid structure of blueprinting could limit the rapid inclusion of new areas if they are not covered by the existing topic list. This risk is diminished by the large number of existing blueprint topics, which might already be suitable for new item themes.

In practice. To ensure candidates are only administered items from the topics selected in their specific blueprints, item assignments are handled separately for each cohort. In doing so, all candidates who come under the same blueprint can have their individual sets of items assigned at the same time. For each batch of item assignments, a blueprint is selected along with the candidates for whom that blueprint applies. Once a blueprint is assigned, the bank of assignable items is filtered to include only the items from topics that are selected in that blueprint.

ISR4

ISR4 limits item selection so that each candidate is never administered multiple items on the same topic in a single test.

Advantages. Implementing ISR4 gives candidates scope for improvement between assessments and increases the number of opportunities for candidates to complete different topics in each test. CAPT aims to provide opportunities for candidates to demonstrate their learning in many topics. Therefore, there needs to be at least as many topics selected in each blueprint as there are individual items in each assessment. Asking only a single item on any topic in a single assessment gives candidates more opportunities to demonstrate this wide-ranging learning. ISR4 also reduces the potential for bias or the overemphasis of topics in a single test.

Disadvantage. Limiting an assessment to include at maximum a single item on a topic limits the number of opportunities to complete specific topics in a single test.

In practice. During item assignment for each candidate, all items that fall under the same topic as an item that has already been assigned are removed from the bank of items that can be assigned to that candidate.

ISR5

ISR5 was introduced to ensure a wide coverage of items across all areas. As per ISR5, each assessment includes at least three items on three different topics from each area so long as at least three topics are selected in the blueprint for that area. If fewer than three topics are selected in the blueprint for an area, then only as many items will be populated as there are topics selected for that area.

Advantage. ISR5 ensures a balanced representation of all areas in each assessment and discourages candidates from focusing study on only a limited number of areas. This further promotes learning toward a comprehensive understanding of the program throughout.

Disadvantage. CAPTs have been designed to maximize the chances for candidates to complete outstanding topics in each test. In practice, this is achieved by predominantly selecting items from outstanding topics. However, to satisfy ISR5, complete topics can be introduced if a candidate has fewer than three remaining outstanding topics in a specific area. In this case, outstanding topics would still be prioritized, with a new item from a complete topic added only when no outstanding topics remain to populate up to three items from an area. This might result in some inclusion of items on topics in areas that candidates are already proficient in and reduce their opportunity to focus on weaker areas. However, this will only be the case when they are close to completing some areas in their entirety.

In practice. ISR5 is the first rule where items are selected from the item bank and added to a candidate's test. Each area is populated with three items, all while following ISR1, ISR2, ISR3 and ISR4. For any areas with fewer than three topics selected, each candidate only receives an equivalent number of items to the number of topics selected in the blueprint for that area in this part of the item assignment.

ISR6

Following the population of areas with up to three items in accordance with ISR5, ISR6 concerns the assignment of the remainder of the items in each assessment. ISR6 details the additional balancing applied to address the limitations of random item assignment. Without further constraints after ISR5, it is possible that candidates would be assigned only items on incomplete topics or only items on unencountered topics. Both cases are undesirable owing to them potentially discouraging widespread study. The introduced balancing coerces the ratio of

the selected items from incomplete and unencountered topics to match the ratio of the remaining (outstanding) incomplete and unencountered topics for each candidate. For example, if a candidate has twice as many incomplete topics as unencountered topics remaining, then in the next assessment they can expect to be administered twice as many items on incomplete topics as on unencountered topics.

Advantages. One benefit of ISR6 is the increased inclusion of items on newly introduced (and therefore unencountered) topics as candidates move between stages and new topics are added to the blueprint. As brand-new topics are introduced in the blueprint, there will be a greater chance of including items from these new topics. This is because the number of outstanding unencountered topics will have increased, compared to outstanding incomplete topics, which have not increased in number. This provides candidates with more opportunities to complete newly introduced topics early on after their introduction. The introduction of ISR6 also aims to encourage exploration of these newly introduced topics, broadening candidates' knowledge and maximizing the number of opportunities they will get to complete topics.

Disadvantages. With the introduction of ISR6, it is important to consider its effects on balancing from the perspective of candidates being administered exams. Candidates could become frustrated if they perceive that they are repeatedly presented with items on topics they have answered incorrectly in the past, so it will be important to review and perhaps tune the balancing in the future. In addition, ISR6 introduces arguably the greatest level of technical complexity out of all the ISRs. As such, it is important to clearly explain the benefits of applying ISR6 in a way that candidates can easily follow and understand.

In practise. When applying ISR6, it is not enough to calculate the ratio of outstanding incomplete and unencountered topics and allocate the number of items to assign from each group. This is because it might not be possible to populate an entire assessment from only these outstanding topics. Complete topics can be added in ISR5 (ensuring a minimum of three items per area) if these items cannot be made up from outstanding topics or if there are insufficient outstanding topics remaining to fill the entire assessment with 125 items. These items on complete topics are hereafter referred to as *forced complete topic* items.

Mathematical Implementation

It is important to calculate the required number of forced complete topic items and only assign items on outstanding topics to the remaining items. The forced complete topic items assigned in ISR5 can be determined by calculating the number of outstanding topics in each area. For each area with zero, one, or two outstanding topics, three, two, and one item(s) on complete topics will be added, respectively. Summing these contributions across all areas will give the number of items on complete topics that must be added due to ISR5.

The second contribution to items from complete topics comes from a lack of sufficient outstanding topics to populate the entire test. The number of items on complete topics introduced this way is equal to the total number of items in each assessment, less the number of outstanding topics. For example, a candidate being administered a 125-item assessment who has only 20 outstanding topics will see one item on each of those outstanding topics in their next test. The remaining 105 items on their next assessment will necessarily be made up of items on topics that they have already completed (forced complete topic items). It is important to note that the outstanding topics only include topics that are in the blueprint and not just the total list of topics. For example, a candidate with only 170 topics in their blueprint and with 100 complete topics, will be considered to have 70 outstanding topics regardless of the total number of topics that will be gradually introduced.

The total number of required forced complete topic items will always be the greater of the two contributions (either from ISR5 or from a lack of sufficient outstanding topics to populate

the entire test). Given the known forced number of items on complete topics, the balancing specified in ISR6 can be applied to the remaining items on outstanding topics.

Let the ratio r be the ratio of the number of incomplete outstanding ($I_{oustanding}$) to unencountered outstanding ($U_{oustanding}$) topics:

$$r = \frac{I_{oustanding}}{U_{oustanding}}. \quad (1)$$

The ratio of the items to add on incomplete topics (I_{add}) and the items to add on unencountered topics (U_{add}) should closely match the ratio, r, in the following way:

$$r = \frac{I_{oustanding}}{U_{oustanding}} \approx \frac{I_{add}}{U_{add}}. \quad (2)$$

In the case that there are fewer outstanding topics than the total number of items in an assessment (N), then an item on all outstanding topics will all be added to the next test. For this scenario, the ratio will match exactly, and balancing need not be considered. In other cases, with a known number of forced complete topic items (FC) to be added to the test, I_{add} and U_{add} can be calculated using two equations (Equations 3 and 4).

Equation 3 states that the sum of items on incomplete and unencountered topics is the total number of assessment items less the number of forced items on complete topics that must be added:

$$I_{add} + U_{add} = N - FC. \quad (3)$$

Equation 4 states that the ratio of incomplete and unencountered items to add should closely match the ratio of outstanding incomplete and unencountered items (from Equation 2):

$$\frac{I_{add}}{U_{add}} \approx r. \quad (4)$$

Then, by combining Equations 4 and 3, the values for U_{add} and I_{add} can be calculated:

$$I_{add} \approx r \times U_{add} \quad (5.1)$$

$$(r \times U_{add}) + U_{add} \approx N - FC \quad (5.2)$$

$$(1 + r) \times U_{add} \approx N - FC \quad (5.3)$$

$$U_{add} \approx \frac{N - FC}{(1 + r)} \quad (5.4)$$

$$I_{add} \approx r \times \frac{(N - FC)}{(1 + r)} \quad (5.5)$$

Equations 5.4 and 5.5 enable the desired number of items on incomplete and unencountered topics to be calculated precisely for each candidate. In many cases, the ratio will be a decimal, leading to a decimal number of items to add. Since an integer number of items must be added, U_{add} and I_{add} should be rounded to the nearest respective integers. However, as shown in Equation 3, U_{add} and I_{add} will sum to an integer ($N - FC$), so rounding one value up will always mean that the other is rounded down, except when both numbers end in .5. In the case that both U_{add} and I_{add} end in .5, it was decided to round the group with fewer outstanding topics up to the nearest integer and the group with a greater number of outstanding topics down to the nearest integer.

With U_{add} and I_{add} calculated, the remaining items in the assessment are populated, filling the assessment to N total items, prioritizing items on outstanding topics, all while adhering to ISR1, ISR2, ISR3 and ISR4.

Validation, Performance and Progression Monitoring

CAPT facilitates the goal of allowing candidates to demonstrate knowledge in each of the topics deemed necessary by the program. As with all adaptive assessments, validation ensures correct delivery and fairness. Since CAPT adapts between assessments, many validation steps can take place before an assessment is even delivered. Before delivery, all student-specific item assignments are checked against all six ISRs. This adapting validation check is an unusual prospect and contrasts with CAT assessments where adaptation typically takes place during an assessment and makes independent validation more challenging.

Performance is ultimately decided by whether a candidate has successfully completed all topics after the end of the final CAPT assessment. However, it is also possible to track and monitor interim progress and set targets at progression decision points.

Peninsula Medical School Specifics

While the CAPT methods described are general and applicable to other areas where progress testing is desired, the core aspects of CAPT assessments have been tailored to the medical program at Peninsula Medical School (PMS). Students at PMS take four progress tests in each of the first four years of the five-year (five stage) undergraduate Bachelor of Medicine Bachelor of Surgery (BMBS) program. The first two tests are formative. All students take the same items in the first formative test, and the second formative assessment adapts by content. Topic completion progress during the first two formative tests is reset before the first summative assessment (third overall test) where all students again take the same items as each other (with no repeat items from the formative tests) and the second summative assessment (fourth overall test) adapts by content. The subsequent 12 tests, over the course of the next three years of the program, are all summative, with items selected adaptively. This approach was chosen to gradually introduce new students to the CAPT system. Thresholds for passing depend on completing the number of topics at any given stage that predict reaching the cumulative target of completing all topics before entering the final year of the program, during which students are eligible to take the Medical School Applied Knowledge Test (MS AKT).

The current PMS item bank is largely aligned to the UK MS AKT content map (General Medical Council, 2021) and comprises 18 areas (clinical areas) at the highest level in the CAPT hierarchical structure. Each clinical area is sectioned into three sub-areas: conditions, presentations, and PMS topics. The first two align with the areas covered by the GMC content map, while the final sub-area covers additional teaching that is required by the program. Currently, there are 644 topics that each fall under a single sub-area and clinical area. In their first stage (CAPTs 01-04), only 167 topics are selected in the blueprint. This rises to 412 topics in the second stage (CAPTs 05-08) and all 644 are selected from the start of stage 3 (CAPTs 09-16).

Currently, each CAPTs assessment lasts three hours by default and contains 125 items. This means that each student is expected to answer 1,750 summative items throughout the 14 summative assessments. Therefore, using the most basic, very primitive estimate, an average performance of 36.8% (644/1750) is expected to allow a student to complete all topics.

From the perspective of a student taking a test, each CAPT appears indistinguishable from a traditional progress test, although all students will face a different set of 125 items on adaptive tests, with the specific items being determined by their past performance. CAPT necessarily requires a large item bank and item selection rules to provide each student multiple opportunities to answer the required topics during subsequent assessments.

Discussion

This article described the item selection rules used to govern the implementation of content adaptive progress tests. While the specific implementation is for students studying medicine, the concept is highly generalizable and could easily be adapted to other applications.

The Impact of Item Selection Rules in Context

Instead of constraining adaptation within an assessment and adapting by difficulty (van der Linden & Reese, 1998; Bengs et al., 2018), each CAPT assessment adapts by content between assessments, selecting, for each candidate, 125 items based on their previous performance. This challenges the conventional view that adaptation can only occur based on difficulty, within a single assessment, and that adaptation is limited by content balancing constraints (Reckase et al., 2019). In the future, other types of items, in addition to multiple-choice questions, could be introduced since item selection is based on content rather than difficulty.

To facilitate the introduction of CAPT has required developing assessments within the framework of knowledge hierarchy, constructive alignment, and a detailed categorization of our item bank, infrastructure that underpinned the development of the set of item selection rules that forms the focus of this article.

The same unique item is never repeated (even if it was voided previously), to prevent any unfair advantage from aberrant item pre-knowledge (Belov, 2014; Barnard, 2015). Only one item can appear on a topic in each assessment undertaken by a candidate, to provide the opportunity to learn from feedback before retesting, and this also limits cueing between items (i.e., enemy constraints; Bengs et al., 2018). Only topics that are selected in a candidate's stage-specific blueprint will appear, to improve alignment between curricular teaching and assessment, contrasting with no direct stage-of-curriculum alignment for traditional progress tests (Plessas, 2015). There will be at least three items on each area (where at least three topics selected in the stage-specific blueprint), and other items will come from outstanding topics where possible to promote maintaining a breadth of learning. The ratio of added items on incomplete to unencountered topics matches (as closely as possible) the ratio of outstanding incomplete to outstanding unencountered topics. This maximizes the opportunity to successfully complete all required content without compensation (Burr et al., 2023). As such, it could be argued that CAPT functions to adapt both "observations" to reduce compensation, and "claims" of feedback to students and staff across tests (Levy et al., 2023). Again, arguably this occurs within a framework that can be considered to be a "hybrid locus of control," influenced by both direct examiner end-destination requirements and indirect examinee choice (Levy et al., 2023), according to their individualized learning path to achieve that destination.

Challenges Compared to Other Progress Tests

In some cases, personalized assessment is no longer compatible with the types of post-assessment analysis applied to non-adaptive assessments where the same set of items are attempted by an entire cohort. For example, as part of the post-assessment process, it is more challenging to identify items for review based on performance metrics such as the point-biserial correlation or facility (Varma, 2006). This limitation can be offset by incorporating the ability for candidates to flag individual items with a review request during the assessment, for later consideration by staff, or by weighting traditionally used metrics by the number of times a specific item has been included in an assessment. It is also offset by the ability to provide tailored feedback to each individual candidate based not only on an individual assessment, but topic coverage across multiple assessments. This reflects both the adaptation between assessments in CAPT and the developing achievement of topics by the candidate. There is potential

here to explore how feedback is utilized by candidates to structure future learning, and how it differs from feedback provided by other forms of assessment.

As each candidate is presented items that are different to those encountered by other candidates, any voided items might not be the same across all candidates. As such, it is important to consider the total number of opportunities each candidate is given to correctly answer items on incomplete topics when it comes to setting pass thresholds based on the total number of completed unique topics. Where items are voided due to circumstances outside a candidate's control, mitigation might take the form of replacement opportunities to answer items or a reduction in the total number of topics candidates are expected to complete.

The implementation of CAPT highlights the principal findings of Stocking et al. (1993), that the success of automatic item selection depends on having an adequate number of items, a classification system for items, and the quality of the items. CAPT assessments rely on the abundance of new, unique items for all topics in the item bank. When candidates answer an item on a topic incorrectly, there must be a new item on that topic that can be included in a subsequent test. A careful check must be made so that there are no topics without new, unique items for each candidate. This might require new items to be written before ISRs can be applied to assign items for the next CAPT assessment. Monitoring of item usage to ensure adequacy, and uniform exposure, along with the possibility of developing a parallel alternative bank, all require further consideration (Barnard, 2015). Adequacy checks can be easily introduced as part of validation checks on the ISRs.

As a new type of assessment, CAPT does not benefit from an item bank that has been used for many years. Item information takes time to ascertain (as would difficulty information in a difficulty-adapting progress test). With continued examinations, this would become less of an issue in time as item history grows. Student preparation to answer certain topics might lead to a greater lack of engagement with items about other topics within an assessment. In the future, this could be monitored using differential rapid-guessing rates, with a threshold for intervention to improve validity (Wise, 2023). It should also be possible to measure changes in traits across assessments for individuals (Tai et al., 2023), and also for cohorts. In the future, such measures could inform both student support and curriculum design. For example, item response theory might be used not to try to adapt assessment construction by both content and difficulty, but to provide more information on the difficulty of each item, within and between topics and candidates, which could in turn inform teaching and learning activities to address knowledge gaps and failure to demonstrate a level of ability within a given topic.

Implications for Learning

All CAPTs appear indistinguishable in format from a traditional progress test at point of delivery, only the selection of items is different (Burr et al., 2022). Herein, the item selection rules are detailed, explaining their rationale and potential impact on motivating learning. Topics are based on the patient presentations and conditions listed by the GMC for the national UK Medical Licensing Assessment undertaken in the final year of study. The blueprint of topics that could be administered in any of the assessments in a year is made available to candidates at the start of the year, so there is alignment to the curriculum.

While all candidates start with the same items in their first summative assessment, the topics they are asked in subsequent assessments become personalized to their needs, to maximize their chances to complete all the required topics. Candidates will know that their incomplete topics will come up again with different items in future tests. All specialities will be represented in all tests. Overall, these factors combine to discourage compensation and mean that candidates cannot avoid areas of weakness. This supports all candidates to become safe generalists and thus be prepared for foundation practice by the end of the program.

Feedback between tests enables candidates to learn more about topics that still need to be completed. The granular and detailed cataloging of questions in the item bank in turn allows candidates also to see and track their progress across various elements of the curriculum on an individual level, tracking their progress and also highlighting areas of improvement. As with traditional progress testing, not all content needs to be covered in teaching before it is tested if it is expected that candidates will have engaged with areas in self-study. But curricular alignment by progressive accumulation of topics by stage is possible and serves to increase the acceptability of the assessments to candidates and teaching staff. It is important that candidates know in advance what topics could be selected for them in any assessment within that year and know the threshold number of topics required in order to pass their stage.

There are no aggregate grades associated with this approach to progress testing (Schuwirth & Van der Vleuten, 2012) as decisions are based on whether candidates have completed a sufficient number of topics, at their stage, to be on track to complete them all before the final year (Burr et al., 2022). Candidates do not need to complete all the topics listed in Stage 1 in order to pass. Candidates also know in advance the rules that determine how items are selected for them, and know that once a topic is selected in the blueprint it will remain for them in all subsequent stages, and they will need to complete it eventually. The transparent selection of items according to content across a sequence of progress tests thus facilitates assessment *as* learning (Bennett, 2010).

References

Allal, L., & Ducrey, G. P. (2000). Assessment of—or in—the zone of proximal development. *Learning and instruction*, *10*(2), 137-152. DOI

Barnard, J.J. (2015). Implementing a CAT: The AMC experience. *Journal of Computerized Adaptive Testing, 3*(1), 1-12. DOI

Belov, D. I. (2014). Detecting item preknowledge in computerized adaptive testing using information theory and combinatorial optimization. *Journal of Computerized Adaptive Testing, 2*(3), 37-58. DOI

Bengs, D., Brefeld, U., & Kröhne, U. (2018). Adaptive item selection under matroid constraints. *Journal of Computerized Adaptive Testing, 6*(2), 15-36. DOI

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A Preliminary theory of action for summative and formative assessment. *Measurement, 8*, 70-91. DOI

Bloom, B. S. (1968). Learning for mastery. instruction and curriculum. *Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints, Number 1*. Evaluation comment, *1*(2), n2.

Burr, S. A., Gale, T., Kisielewska, J., Millin, P., Pêgo, J. M., Pinter, G., Robinson, I. M., & Zahra, D. (2023). A narrative review of adaptive testing and its application to medical education. *MedEdPublish, 13*(221), 1-8. DOI

Burr, S.A., Kisielewska, J., Zahra, D., Hodgins, I., Robinson, I., Millin, P., Gale, T., Santos, N., & Pêgo, J. M. G. M. (2022). Personalising knowledge assessments to remove compensation and thereby improve preparation for safe practice--developing content adaptive progress testing. Preprint. DOI

Collares, C. F. (2022) Cognitive diagnostic modelling in healthcare professions education: an eye-opener. *Advances in Health Science Education: Theory and Practice. 27*(2):427-440. DOI

General Medical Council. (2021). Medical Licensing Assessment content map Medical Licensing Assessment content map. WebLink

Levy, R., Behrens, J. T., & Mislevy, R. J. (2023). An extended taxonomy of variants of computerized adaptive testing. *Journal of Computerized Adaptive Testing, 10*(1), 1-21. DOI

Luecht, R. M., De Champlain, A., Nungester, R. J. (1998) Maintaining content validity in computerized adaptive testing. *Advances in Health Science Education: Theory and Practice. 3*(1): 29–41 DOI

Lunz, M. E., & Deville, C. W. (1996). Validity of item selection: A comparison of automated computerized adaptive and manual paper and pencil examinations. *Teaching and Learning in Medicine: An International Journal*, *8*(3), 152-157. DOI

National Council of State Boards of Nursing, NCSBN. (2023). *NCLEX-RN® Test Plan*; April 2023. Chicago: NCSBN.

Plessas, A. (2015). Validity of progress testing in healthcare education. *International Journal of Humanities Social Sciences and Education, 2*(8), 23-33.

Reckase, M., Ju, U., & Kim, S. (2019). How adaptive is an adaptive test: Are all adaptive tests adaptive? *Journal of Computerized Adaptive Testing, 7*(1), 1-14. DOI

Schuwirth, L. W., and van der Vleuten, C. P. (2012). The use of progress testing. *Perspectives on medical education, 1*, 24-30. DOI

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). Application of an automated item selection method to real data. *Applied Psychological Measurement*, *17*(2), 167-176. DOI

Tai, M. H., Cooperman, A.W., DeWeese, J. N., & Weiss, D. J. (2023). How do trait change patterns affect the performance of adaptive measurement of change? *Journal of Computerized Adaptive Testing, 10*(3), 32-58. DOI

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*(3), 259–270. DOI

Varma, S. (2006) Preliminary item statistics using point-biserial correlation and *p*-values. *Educational Data Systems Inc.: Morgan Hill CA. 16*(07), 1-7.

Wise, S. L. (2023). Expanding the meaning of adaptive testing to enhance validity. *Journal of Computerized Adaptive Testing, 10*(2), 22-31. DOI

Yan, D., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. Yan, C. Lewis, & A. A. von Davier (Eds.), *Computerized multistage testing: Theory and applications* (pp. 3–20). New York: Chapman and Hall/CRC.

Author Address

gergo.pinter@plymouth.ac.uk

Citation

Pinter, G., Santos, N., Pêgo, J. M., Zahra, D., & Burr. S. A. (2026).
Item selection rules for content adaptive progress testing.
Journal of Computerized Adaptive Testing, 13(2). 7–20. DOI 10.7333/2606-1302007