Ensuring content validity of patient-reported outcomes: a shadow-test approach to their adaptive measurement

Seung W. Choi · Wim J. van der Linden

Accepted: 10 July 2017 / Published online: 14 July 2017 © Springer International Publishing AG 2017

Abstract

Purpose Most computerized adaptive testing (CAT) applications in patient-reported outcomes (PRO) measurement to date are reliability-centric, with a primary objective of maximizing measurement efficiency. A key concern and a potential threat to validity is that, when left unconstrained, individual CAT administrations could have items with systematically different attributes, e.g., sub-domain coverage. This paper aims to provide a solution to the problem from an optimal test design framework using the shadow-test approach to CAT.

Methods Following the approach, a case study was conducted using the PROMIS® (Patient-Reported Outcomes Measurement Information System) fatigue item bank both with empirical and simulated response data. Comparisons between CAT administrations without and with the enforcement of content and item pool usage constraints were examined.

Results The unconstrained CAT exhibited a high degree of variation in items selected from different substrata of the item bank. Contrastingly, the shadow-test approach delivered CAT administrations conforming to all specifications with a minimal loss in measurement efficiency.

Conclusions The optimal test design and shadow-test approach to CAT provide a flexible framework for solving complex test-assembly problems with better control of their domain coverage than for the conventional use of CAT in PRO measurement. Applications in a wide array of PRO domains are expected to lead to more controlled and balanced use of CAT in the field.

Keywords Optimal test design · Shadow-test approach to CAT · Mixed-integer programming · Patient-reported outcomes measurement information system (PROMIS) · Content validity

Introduction

In the assessment of PROs, fixed test forms and computerized adaptive tests (CATs) are often used interchangeably on the premise that both formats meet the same test specifications. In conventional CAT, however, items are selected mainly from a statistical perspective, whereas fixed forms are often created to meet more clinical and non-statistical requirements (e.g., sub-domain coverage, types of rating scales, negative vs. positive wording, etc.). PRO measures designed to assess one highly-specialized domain at a time can be measured very efficiently through CAT provided the assumption that any combination of the items produces comparable scores varying only in the level of measurement precision (i.e., are tau-equivalent). CAT programs in such disciplines as educational testing, on the other hand, often constrain the selection of their items severely to strictly conform to very detailed test specifications by means of content balancing. An analog of them would be a CAT for the measurement of depression covering the checklist of major symptoms and substrata with specific number of items apportioned. A potential threat to the validity of such PRO measures exists when they have to cover sub-domains that are conceptually distinct. In such cases, when left unconstrained, individual CATs rendered in real time will tend to be associated with systematically different sub-domains. An important word of caution to avoid creating any unwarranted expectation: content balancing is still predicated on sufficient unidimensionality.
and not designed to rectify any bias or distortion introduced to the measure due to substantial multidimensionality ignored during the calibration phase. In what follows, we use the shadow-test approach to CAT, founded on the optimal test design framework [1, 2], to guarantee real-time generated tests that do satisfy exactly the same, possibly complex set of test specifications for all patients while still optimizing measurement precision. Also, we show how the approach can be used to provide flexible solutions to the problem of creating PRO measures for various test formats nevertheless conforming to the same specifications and content requirements. More specifically, this paper aims to (i) introduce the optimal test design framework, (ii) show how it can be used to integrate all clinical, statistical, practical, and logical requirements for tests into their administration, (iii) use the shadow-test approach to construct PRO measures with a fully-adaptive, multistage, linear on-the-fly format or any hybrid version of them, (iv) review computational aspects of the shadow-test approach (i.e., constrained optimization via mixed-integer programming), and (v) discuss practical implementation issues in the context of PRO measurement.

Optimal test assembly

Optimal test assembly is an approach to constructing a test or a battery of tests with a potentially elaborate list of specifications, from a large variety of item pools, with a result that is optimal with respect to an objective chosen by the test assembler [1–3]. The approach involves mathematical modeling of test specifications, including content coverage and psychometric requirements as well as practical considerations, which often compete with each other and thus are difficult to satisfy simultaneously. Although the idea of modeling test specifications may seem far-fetched, it is actually the only viable solution for test-assembly problems with a complex set of specifications. For example, if an item pool consists of $N$ items and a test of $n$ is to be constructed, the number of possible combinations is equal to $\binom{N}{n}$, which is an astronomical number already for low values of $N$. Although imposing constraints with respect to content sub-domains, item types, rating scales, and so forth reduces the set containing feasible combinations, the set is too complex and still much too large to find an optimal solution by trial and error. Conventional test-assembly methods involve sequential optimization on the premise that selecting one local optimal item at a time leads to a global optimum for the final test. Optimal test assembly, on the other hand, uses the methodology of mixed-integer programming (MIP) to simultaneously select a full set of items conforming to all constraints to be optimal with respect to the objective of choice. The MIP methodology [4] has been developed and rigorously researched in the field of mathematical programming for a class of problems broadly known as constrained combinatorial optimization problems. Many instances of this class are found throughout business, trade, and industry and include such problems as machine scheduling in manufacturing, vehicle routing in transportation, portfolio assembly in finance, and crew scheduling in the airline industry [5, p. 89]. The “Appendix” section to this article provides a more technical overview of the process to formulate shadow-test assembly as a combinatorial optimization problem and calculate its solution within the optimal test-assembly framework.

Shadow-test approach to CAT

In conventional CAT, items are selected from the item pool sequentially, one item at a time. The methods currently available to balance the test content across examinees vary considerably with respect to the types of constraints handled and the level of conformity to the specifications guaranteed, ranging from simple content rotation strategies and thus are difficult to satisfy simultaneously. Although the idea of modeling test specifications may seem far-fetched, it is actually the only viable solution for test-assembly problems with a complex set of specifications. For example, if an item pool consists of $N$ items and a test of $n$ is to be constructed, the number of possible combinations is equal to $\binom{N}{n}$, which is an astronomical number already for low values of $N$. Although imposing constraints with respect to content sub-domains, item types, rating scales, and so forth reduces the set containing feasible combinations, the set is too complex and still much too large to find an optimal solution by trial and error. Conventional test-assembly methods involve sequential optimization on the premise that selecting one local optimal item at a time leads to a global optimum for the final test. Optimal test assembly, on the other hand, uses the methodology of mixed-integer programming (MIP) to simultaneously select a full set of items conforming to all constraints to be optimal with respect to the objective of choice. The MIP methodology [4] has been developed and rigorously researched in the field of mathematical programming for a class of problems broadly known as constrained combinatorial optimization problems. Many instances of this class are found throughout business, trade, and industry and include such problems as machine scheduling in manufacturing, vehicle routing in transportation, portfolio assembly in finance, and crew scheduling in the airline industry [5, p. 89]. The “Appendix” section to this article provides a more technical overview of the process to formulate shadow-test assembly as a combinatorial optimization problem and calculate its solution within the optimal test-assembly framework.

Shadow-test approach to CAT

In conventional CAT, items are selected from the item pool sequentially, one item at a time. The methods currently available to balance the test content across examinees vary considerably with respect to the types of constraints handled and the level of conformity to the specifications guaranteed, ranging from simple content rotation strategies and thus are difficult to satisfy simultaneously. Although the idea of modeling test specifications may seem far-fetched, it is actually the only viable solution for test-assembly problems with a complex set of specifications. For example, if an item pool consists of $N$ items and a test of $n$ is to be constructed, the number of possible combinations is equal to $\binom{N}{n}$, which is an astronomical number already for low values of $N$. Although imposing constraints with respect to content sub-domains, item types, rating scales, and so forth reduces the set containing feasible combinations, the set is too complex and still much too large to find an optimal solution by trial and error. Conventional test-assembly methods involve sequential optimization on the premise that selecting one local optimal item at a time leads to a global optimum for the final test. Optimal test assembly, on the other hand, uses the methodology of mixed-integer programming (MIP) to simultaneously select a full set of items conforming to all constraints to be optimal with respect to the objective of choice. The MIP methodology [4] has been developed and rigorously researched in the field of mathematical programming for a class of problems broadly known as constrained combinatorial optimization problems. Many instances of this class are found throughout business, trade, and industry and include such problems as machine scheduling in manufacturing, vehicle routing in transportation, portfolio assembly in finance, and crew scheduling in the airline industry [5, p. 89]. The “Appendix” section to this article provides a more technical overview of the process to formulate shadow-test assembly as a combinatorial optimization problem and calculate its solution within the optimal test-assembly framework.

Shadow-test approach to CAT

In conventional CAT, items are selected from the item pool sequentially, one item at a time. The methods currently available to balance the test content across examinees vary considerably with respect to the types of constraints handled and the level of conformity to the specifications guaranteed, ranging from simple content rotation strategies and thus are difficult to satisfy simultaneously. Although the idea of modeling test specifications may seem far-fetched, it is actually the only viable solution for test-assembly problems with a complex set of specifications. For example, if an item pool consists of $N$ items and a test of $n$ is to be constructed, the number of possible combinations is equal to $\binom{N}{n}$, which is an astronomical number already for low values of $N$. Although imposing constraints with respect to content sub-domains, item types, rating scales, and so forth reduces the set containing feasible combinations, the set is too complex and still much too large to find an optimal solution by trial and error. Conventional test-assembly methods involve sequential optimization on the premise that selecting one local optimal item at a time leads to a global optimum for the final test. Optimal test assembly, on the other hand, uses the methodology of mixed-integer programming (MIP) to simultaneously select a full set of items conforming to all constraints to be optimal with respect to the objective of choice. The MIP methodology [4] has been developed and rigorously researched in the field of mathematical programming for a class of problems broadly known as constrained combinatorial optimization problems. Many instances of this class are found throughout business, trade, and industry and include such problems as machine scheduling in manufacturing, vehicle routing in transportation, portfolio assembly in finance, and crew scheduling in the airline industry [5, p. 89]. The “Appendix” section to this article provides a more technical overview of the process to formulate shadow-test assembly as a combinatorial optimization problem and calculate its solution within the optimal test-assembly framework.
the trait estimate for the examinee is updated ($\hat{\theta}_1$); and (4) the optimal test-assembly engine then assembles a new shadow test satisfying all previous constraints to be optimal at $\hat{\theta}_1$, with one additional constraint, i.e., inclusion of the item(s) administered in the previous stage(s). The cycle is repeated until the intended test length has been reached.

The examinee only sees the items actually administered to her/him. The full forms from which they were selected remain completely hidden; hence, the name shadow test. As each shadow test satisfies all test specifications, so does each adaptive test. Likewise, as each shadow test was selected to be optimal at the updated trait estimate and only its best free item at the estimate was administered, the final trait estimate is optimal.

The shadow-test approach also provides a flexible mechanism to control the level of adaptivity to render different test formats with the same test specifications. Although maximum adaptivity is realized when the shadow test is reassembled immediately upon each administered item, the same mechanism allows for assembling any conceivable testing format as a special case [5]. A few possible formats with reduced levels of adaptivity are produced as: (1) a single shadow test assembled to target a specific trait level(s) and administered in whole to all examinees, which is equivalent to a fixed form test (the lowest level of adaptivity); (2) an individualized shadow test constructed for each examinee targeting his/her score from a previous administration and presented in its entirety (linear on-the-fly test: LOFT); (3) an individual shadow test reassembled to be optimal at the examinee’s trait update at some predetermined points only (on-the-fly multistage test [MST]); and (4) any combinations of the above. Figure 2 shows graphically how a 12-item test with varying levels of adaptability can be constructed to render different test formats using the mechanism of freezing and reassembling the shadow test. As the approach guarantees complete satisfaction of the same set of test specifications for different test formats, it offers a unique opportunity to compare their relative efficiencies [8]. An empirical example of the relative efficiency of several test formats with an increasing degree of adaptivity will be presented later in this article.

**Item-exposure/usage control**

The maximum-information item-selection criterion, which increases test efficiency in CAT, can also cause overexposure of a small proportion of items while underexposing the rest. Results like this are a critical concern in high-stakes testing for item security reasons. Although item security may not be of the utmost concern in PRO measurement, highly-unbalanced exposure rates amount to inefficient item pool use and is likely to give rise to content coverage concerns. In extreme cases, when the typically greedy item-selection algorithm is left unconstrained, the majority of items in the pool may remain inactive. Furthermore, maximum-information algorithms favor items with large positive error in their slope parameters estimates, which leads to underestimation of the standard errors of the trait estimates in fixed-length CAT and premature termination of variable-length CAT [9]. These undesirable consequences can be mitigated by introducing an item-exposure control mechanism [10].

A simple randomization strategy can reduce unbalanced item utilization to a certain extent [11]; however, more precise control is obtained by monitoring and adjusting the item-exposure rates directly in real time. The shadow-test approach addresses the problem by adding random item-eligibility constraints to the test-assembly model so that items with higher exposure rates have higher probabilities of being temporarily ineligible for the examinees [12]. Prior to the test, for each examinee, random Bernoulli experiments are conducted by the CAT engine to determine which items to constrain as temporarily ineligible from the item pool for the examinee. As the method permanently
updates the probabilities of ineligible for all items and their updates are adaptive (i.e., the probabilities decrease if their exposure rates increase, and conversely), it guarantees exposure rates for all items in the pool tightly bound below a maximum exposure rate set a priori (e.g., 0.25).

Another item-usage control method, known as z-stratification [13], is available for use in conjunction with the item-ineligibility method. The method partitions the item pool into several strata based on the slope parameters (z) of the items. It then constrains the item-selection process to begin in the lowest stratum and finish in the highest, basically deferring the selection of the better z values until the trait estimate has converged nearly completely. The rationale is to use high-information items at the most beneficial point in testing. The shadow-test approach provides a straightforward mechanism to incorporate the stratified procedure: (1) the test-assembly model is constrained to include a predefined number of items from each stratum; and (2) a position constraint is imposed on the order of the strata so they are presented in order from the lowest to the highest discrimination.

Method

In the first part of the empirical section, we will demonstrate (1) the extent to which unconstrained CAT administrations without content balancing can differ from each other, and (2) how the shadow-test approach enforces the required distribution of content across individual CAT administrations. The second part of the section will focus on the possible consequences of uncontrolled item pool usage. More specifically, we will then investigate (1) the extent to which unconstrained CAT administrations without item-exposure control lead to highly-skewed item usage, and (2) how the addition of exposure/item-usage control can ensure more balanced use of the items at minimal costs (e.g., no or little reduction in measurement precision). Finally, we will demonstrate how the shadow-test approach can produce the same test with different adaptive formats using a sample item bank from the Patient-Reported Outcomes Measurement Information System (PROMIS®), illustrating both its effectiveness and efficiency.

PROMIS fatigue bank

An ideal case for the purpose of demonstrating the shadow-test approach is a relatively large item bank with multiple content sub-domains. The PROMIS fatigue item bank consists of 95 items measuring two conceptual areas: an individual’s fatigue experience and its impact on daily living [14]. The item bank includes the 13-item FACIT-fatigue (FACIT-F) [15] as a legacy or validity measure and employed two different rating scales to score items either on intensity (1 = Not at all to 5 = Very much) or frequency (1 = Never to 5 = Always). Previous studies showed the PROMIS fatigue item bank to be dominated by a strong general factor. Although its developers obtained better fit using a bifactor model with two additional group factors, Fatigue Experience (40 items) and Fatigue Impact (55 items), they used a unidimensional IRT model to eventually calibrate the item bank. A recent study confirmed the dominance of the general factor but also suggested that one of the group factors could be further divided into Impact on Social/Recreational Activities (8 items) and Impact on Mental/Cognitive Activities (15 items) [16]. One implication of this observed dimensional structure is that some of the items in the item bank may exhibit local dependence (LD) when scored using the unidimensional IRT parameter estimates.

Sample data

The PROMIS Waive I data that responded to the full fatigue bank (n = 803) were obtained from the HealthMeasures consortium [17]. The dataset was used to examine the item-usage rates and distribution of items selected across the different sub-domains under unconstrained CAT administrations. Simulated response data were also used for more extensive comparisons of different CAT methods which are explained in the following section.

Comparison of CAT methods

To compare the first-generation PROMIS CAT engine [18–20] (Unconstrained CAT) with the shadow-test approach to CAT, a fixed-length CAT of 12 items was simulated under the following conditions.

1. Unconstrained CAT (Method 1): a 12-item fixed-length CAT with no constraints; the maximum posterior weighted information (MPWI) item-selection criterion; the expected a posteriori (EAP) trait estimator; and an initial prior mean of θ = 0.

2. Shadow-test approach to CAT (Method 2): a 12-item fixed-length CAT as in Method 1 with the following constraints

(a) 6 items from Fatigue Experience,
(b) 6 items from Fatigue Impact,
(c) 1 or 2 items from Impact on Mental/Cognitive Activities,
(d) 1 or 2 items from Impact on Social/Recreational Activities,
(e) 6 items on Intensity rating scale,
(f) 6 items on Frequency rating scale,
Shadow-test approach to CAT as a linear on-the-fly multi-stage test (MST) with three stages (Method 4): a 12-item fixed-length CAT as in Method 3 with the shadow test (re)assembled prior to administering 1st, 5th, and 9th items only

Shadow-test approach to CAT as a linear on-the-fly test (LOFT) (Method 5): a 12-item fixed-length test as in Method 4 with a single shadow test assembled for each examinee prior to the 1st item only and frozen for the rest of the test

The five CAT methods described above are arranged from the unconstrained (Method 1) to the most constrained (Method 5). With additional constraints imposed on the content, usage, and format, some reduction in the measurement precision was expected for the more constrained methods. Because each of the CAT formats above can be implemented using the shadow-test approach, it offers a unique opportunity to compare their relative measurement precision. The test length of 12 was chosen to emulate a unique opportunity to compare their relative measurement precision. The test length of 12 was chosen to emulate a half-length rendering.

Results

Real-data analysis

The real-data analyses were based on a subset of the PROMIS Wave 1 data [17] with complete responses on all 95 items (n = 669). Local dependence was assessed using the Q3 statistic [21], which identified 99 pairs of items as spuriously correlated (|Q3| > 0.3). The highest Q3 statistic was associated with the following pair of items: How fatigued were you when your fatigue was at its worst? and the following item: How often did you have to push yourself to get things done because of your fatigue? The same item was used only for 21.7% of the cases under Method 3 with no locally-dependent (LD) item pairs and no globally-dependent (GD) item pairs. The rest of the test

Comparing the content-constrained CAT (Method 2) and the CAT constraining both content and usage (Method 3) shows that over 40% of the item bank (39 of 95 items) was never used under Method 2, whereas only about 6% (6 items) remained inactive under Method 3. The distribution of item-exposure rates for Method 2 was highly-skewed and all 669 instances of CAT in the sample included the following item: How often did you have to push yourself to get things done because of your fatigue? The same item was used only for 21.7% of the cases under Method 3 with LD item pairs and no LD item pairs.

Simulated data analysis

For more detailed comparisons of the CAT methods and their effectiveness, simulated data were generated as follows: (1) θ values were generated over a grid, [−2.0, 3.0], with increments of 0.5 and 1000 replications at each θ point, to serve as the true level of fatigue for a total 11,000 simulees; and (2) each CAT method was used to simulate adaptive test administrations from the PROMIS fatigue item bank resulting in a response matrix of dimension 11000 × 95. Figure 3 displays the conditional root mean square error (RMSE) functions for the first three CAT methods: the unconstrained (Method 1), content constrained (Method 2), and content and usage constrained (Method 3); and the RMSE was calculated conditioning on each θ point as follows: \( \text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\theta}_i - \theta)^2}{1000}} \). The RMSE functions for the three CAT methods behaved as expected throughout the θ range. (Incidentally, PROMIS has adopted the T-score metric as a linear transformation of θ, i.e., T-score = θ × 10 + 50.) First, the RMSE function was lowest for the unconstrained CAT (Method 1) while some reduction of measurement accuracy was discernible for content-constrained CAT (Method 2), especially at the lower (healthier) end of the fatigue continuum. The latter was primarily...
due to the distribution of items in the PROMIS fatigue item bank, which favors the higher end of the fatigue continuum. To some extent, however, reduction of the measurement accuracy observed at the lower end was also due to the fact that, as further presented below, Method 2 and 3 constrained the presence of higher-information. Second, adding the item-ineligibility constraints (Method 3) had practically no impact on the accuracy of measurement in the upper half of the fatigue continuum where information was plentiful.

Table 1: Average number of items selected by category—real-data results ($n = 669$)

<table>
<thead>
<tr>
<th>Category</th>
<th>Item count</th>
<th>Specification</th>
<th>Unconstrained CAT (Method 1)</th>
<th>Constrained CAT (Method 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Domain: impact</td>
<td>55</td>
<td>6</td>
<td>6.12</td>
<td>2.91</td>
</tr>
<tr>
<td>Domain: experience</td>
<td>40</td>
<td>6</td>
<td>5.88</td>
<td>2.91</td>
</tr>
<tr>
<td>Impact: mental</td>
<td>15</td>
<td>1 or 2</td>
<td>0.01</td>
<td>0.20</td>
</tr>
<tr>
<td>Impact: social</td>
<td>8</td>
<td>1 or 2</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td>Rating scale: intensity</td>
<td>48</td>
<td>6</td>
<td>9.08</td>
<td>1.58</td>
</tr>
<tr>
<td>Rating scale: frequency</td>
<td>47</td>
<td>6</td>
<td>2.92</td>
<td>1.58</td>
</tr>
<tr>
<td>Slope: low [1.17–2.92]</td>
<td>31</td>
<td>4</td>
<td>0.54</td>
<td>1.70</td>
</tr>
<tr>
<td>Slope: medium [2.92–3.52]</td>
<td>34</td>
<td>4</td>
<td>0.65</td>
<td>1.32</td>
</tr>
<tr>
<td>Slope: high [3.52–4.77]</td>
<td>30</td>
<td>4</td>
<td>10.81</td>
<td>2.65</td>
</tr>
<tr>
<td>Legacy: FACIT-F</td>
<td>13</td>
<td>1 or 2</td>
<td>2.55</td>
<td>0.81</td>
</tr>
<tr>
<td>LD item Pairs ($</td>
<td>Q3</td>
<td>&gt; 0.3$)</td>
<td>99</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 3: Conditional root mean square error (RMSE) functions for different CAT configurations

Fig. 4: Distribution of item-usage (exposure) rates for CAT without usage control (solid line) and CAT with usage control (dashed line)—simulated data ($n = 11,000$)

Table 2 displays the average number of items selected from each category conditional on the T-scores for the unconstrained case (Method 1). It highlights how unconstrained CAT drew items from different categories depending on the level of fatigue. On the other hand, the
items selected with constrained CAT (Methods 2 and 3) were fully compliant with the test specifications (shown in Column 3) across all levels of fatigue (with no exceptions, so data illustrating this finding are not presented). There are interesting patterns worth noting in Table 2. First, the distribution of items for Fatigue Impact and Experience was closely tied to the severity of fatigue. For the lower end (T-scores between 30 and 40), over 90% of items selected under the unconstrained CAT (Method 1) was from Fatigue Experience. This pattern is visually depicted in Fig. 5. The distribution of the two sub-domains reached an equilibrium at T-score of 45 and reversed for T-scores of 50 and higher. That is, Fatigue Impact dominated the measure at the upper half of the continuum for about 70%. Although this change of focus of measurement from one sub-domain to another across the fatigue continuum might be unsurprising or even desired, it is preferred to incorporate such features more explicitly into the design of the CAT, instead of just letting the item-selection algorithm determine the composition of the test rather implicitly. Second, it is also interesting to observe that the two rating scales (Intensity vs. Frequency) were used quite disproportionately at various levels of fatigue. For instance, when left unconstrained, about 9 to 11 items (out of 12) selected at T-score levels between 50 and 70 had the Intensity rating scale. Although this over-representation might be inconsequential, more precise control is possible, as shown in Fig. 3, with little impact on the accuracy of measurement at the upper half of the continuum. Third, administering LD item pairs was mostly concentrated at the both ends of the continuum, e.g., 7.07 pairs on average at T = 30, and 5.34 at T = 80, but rarely near its middle. This result implies that when information is scarce, unconstrained CAT (Method 1) is more prone to draw on LD items, which tend to have questions perceived by the respondents as redundant and lead to underestimated standard errors of measurement.

### Accuracy and adaptivity

The shadow-test approach provides a simple device to control the adaptivity of CAT through its refreshing–freezing mechanism. This facility allows for the test administrator to determine when and how often the shadow test should be refreshed given the item bank and extent of constraints to be imposed [22]. Accordingly, it offers a
means to compare the relative efficiency of different CAT formats with varying levels of adaptivity. The three test formats (Methods 3, 4, and 5) introduced earlier enforce the same exact test specifications pertaining to both the content and item-usage constraints but with decreasing adaptivity (Method 3 > Method 4 > Method 5). As a reminder, Method 3 is a fully-adaptive CAT in which the shadow test is reassembled upon administering each item and updating the trait estimate ($\hat{\theta}$); Method 4 is a multi-stage test (MST) constructed on the fly in which the shadow test is (re)assembled three times for 1st, 5th, and 9th items and frozen during the administration of in-between items; and Method 5 is a linear on-the-fly test (LOFT) where the shadow test is assembled once for each examinee optimized for the prior distribution, e.g., $N(0, 1)$, and frozen for the rest of the test. Figure 6 shows the conditional RMSE functions as a measure of relative efficiency for the three test formats. Surprisingly, the level of adaptivity realized in the three test formats had little impact on the accuracy of measurement, with a visible difference present only toward the both ends of the continuum between the fixed (Method 5) and two adaptive formats (Methods 3 and 4). The difference between the fully-adaptive (Method 3) and multistage format (Method 4) was practically nonexistent.

This was somewhat unexpected even considering the fact that information functions for polytomous items are often multi-modal, covering a wide range of trait levels [19]. Therefore, we examined a few possibilities that might have obscured the potential difference in the accuracy of measurement between the test formats, including: (1) the 12-item test was long enough for the least-adaptive (or fixed) format (LOFT: Method 5) to provide good measurement; (2) the constraints resulted in the use of less informative items even for the formats with higher levels of adaptivity; and (3) the first shadow test created using maximum prior/posterior weighted information (MPWI) [19] instead of maximum Fisher information (MFI) as the objective function, covered the middle portion of the fatigue scale already reasonably well. Additional simulations were conducted to compare the three levels of adaptivity realized in CAT, MST, and LOFT under the following three factors: (1) 12 items versus 6 items; (2) constrained versus unconstrained; and (3) MPWI versus MFI. The constraints for the 6-item conditions were generated by halving the number of items required in the specification (see Column 3 of Table 2). The unconstrained conditions were carried out removing all content and item-usage constraints (i.e., Method 1). Finally, the MFI criterion replaced with the MPWI criterion as the objective function when constructing the shadow test. That is, the initial shadow test was optimized for a single value, $\theta = 0$ (or $T$-score = 50) rather than an entire prior distribution.

Figure 7 summarizes the additional simulation results graphically. The eight panels in the figure, labeled (a) through (h), correspond to the eight conditions created by crossing the three factors described above (i.e., the test length, constraints, objective function). For example, Panels (a) to (d) are for the constrained conditions (i.e., with imposing both content and usage constraints), while (e) to (h) are for the unconstrained conditions. All three factors generally revealed small effects on the relative efficiency of the three adaptivity levels. More substantial effects were found for the unconstrained conditions, (e) to (h). The least-adaptive format (LOFT) was generally less accurate at both extremities of the continuum, but especially so at the lower end. However, under all four conditions, the difference between CAT and MST was negligible for all practical purposes, with CAT slightly outperforming MST under the unconstrained conditions with the longer 12-item test. We believe these generally small differences between the CAT and MST formats to be the result of the short sub-tests for the MST (4 items per stage), possibly in combination with the highly informative nature of the polytomous items.

Conclusions

Most CAT applications in PRO measurement to date are reliability-centric, with a primary objective of maximizing measurement efficiency. In other testing applications, specification of item-selection algorithms in CAT entails consideration of content coverage in addition to increasing the precision of measurement. Consequently, item-selection is guided by well-defined and detailed test
specifications to ensure that the set of items administered in each CAT administration meets all of the requirements of the test specifications [23, pp. 80–81]. A key concern and potential threat to the validity in CAT is the lack of compliance of the individual assessments rendered in real time with the test specifications. So far, establishing content-oriented validity in CAT has not been seen as critical in PRO measurement where items still tend to be selected primarily on the basis of their statistical properties and a construct assumed to be homogeneous. However, a similar concern should arise when the item bank measuring a PRO contains multiple sub-domains and/or other attributes that differentiate them empirically. When such structures are neglected, they may actually dominate the measure for all examinees or change its nature along the continuum that is measured. In either case, the interpretability of scores and their changes can be compromised. To that end, the U.S. Food and Drug Administration guidance for PROs recommended that content validity be established before any attempts to interpret their construct validity, reliability, or sensitivity to change [24, p. 12].

So, what are the costs of constraining item selection in CAT? Generally, they depend on the severity of the constraints and the depth of the item bank. For the PROMIS fatigue item bank, the loss of information due to imposing a total of 109 constraints (10 constraints pertaining to content coverage and 99 related to LD) was the equivalent of a little over one item. One additional item would have been required to make up for the item-usage constraints. Both findings justify the conclusion that the RMSE functions observed in our study for a 12-item shadow CAT constrained for content, LD, as well item usage were roughly equivalent to what they would have been for a 10-item unconstrained CAT (see Fig. 3). This comparison is for measurement accuracy only. On the other hand, for most contexts, we do not expect the precision gained by leaving out some of the content constraints to compensate for the compromise in validity.

**Fig. 7** Conditional RMSE for three test formats by type of CAT (constrained vs. unconstrained), test length (12 vs. 6), and objective function (MPWI vs. MFI)
Unlike content balancing, item-usage control is generally not a concern in PRO measurement where test security is not a risk. Nevertheless, highly-skewed use of the item bank (albeit not necessarily a validity issue) should raise a practical concern if the majority of items would never be used. Suppose a bank contains two items, A and B, that are equivalent in all substantive aspects. However, Item A is always preferred by the selection algorithm because of its slightly elevated information function relative to Item B. If so, it may seem reasonable to either drop Item B from the bank (knowing it will never be selected) or allow the CAT engine to select Item B with some positive probability (acknowledging the fact that the choice between them is not of any substantive importance). After all, items must have made it to an item bank for some good reasons. Monitoring and controlling item usage should therefore be useful as part of item bank management. The empirical results presented in the paper suggest that this can be done effectively without significant loss in measurement precision.

Changing the format of adaptive testing (e.g., CAT to MST) is a major undertaking. As demonstrated in this paper, such changes are easily prepared and implemented using the refreshing-freezing mechanism available for the shadow-test approach. For example, on-the-fly MST is the same as a CAT with the shadow test temporarily frozen during the administration of groups of consecutive items [5]. It should be noted though that longer periods of frozen shadow tests lead to loss of statistical accuracy of the measurements. Nevertheless, perhaps, one practical benefit of freezing the shadow test could be easy forward and backward navigation for the examinees to preview and review test content within block of items, for instance, small blocks of items from the same sub-domain presented on the same screen.

Finally, constrained CATs are especially suited for PROs with a relatively complex domain hierarchy and larger numbers of items but still sufficiently unidimensional. Conversely, they may be less needed for measuring highly-specialized domains with small numbers of interchangeable items, for which conventional CATs without constraints can be very effective. Content balancing aims to improve content validity by ensuring that the sub-domains are represented consistently across individual instances of CAT. However, ex post facto content balancing through the shadow-test approach or other alternatives may not undo any bias or distortion introduced to the measure when multidimensionality is ignored during the calibration phase. If there are multiple dimensions underlying a measure, a fundamental solution is to split them out and calibrate them separately or use a multidimensional model. Constraining the selection of highly-redundant, LD items to an examinee may improve face validity at least and content validity to a certain extent, because otherwise the examinee may perceive LD items as being overly redundant and the content domain represented by them being oversampled. However, limiting selection of LD items may not be a fundamental solution, which may also involve multidimensional models allowing nuisance factors or correlated residuals.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix

The process of creating and solving a series of combinatorial optimization problems using mixed-integer programming (MIP) is presented in Fig. 1 (labeled as Optimal Test Assembly / MIP). In what follows, we provide a brief overview of the process to make it easy for the reader to access. However, more comprehensive accounts are available [1, 2, 25]. MIP addresses a class of mathematical programming problems with both integer and real-valued variables for which we seek a solution that maximizes or minimizes an objective function, while satisfying a set of constraints. The combinatorial optimization problem pos- tulates a finite set of alternatives from which an optimum solution can be determined based on the objective values associated with the alternatives. Nearly every conceivable automated test-assembly problem can be formulated in this class in the following four steps [1, chapter 3]. (1) Identifying the decision variables: Choosing appropriate decision variables is critical as it governs how expediently we can denote and identify the subset of feasible solutions from all potential combinations. The most common choice is binary variables, $x_i \in \{0, 1\}, i = 1, 2, \ldots, I$ where $x_i = 1$ if Item $i$ is selected in the test, or $x_i = 0$ otherwise. In addition, we often add one or more real-valued variables for technical reasons; (2) Modeling the constraints: The choice of binary decision variables also facilitates formulating constraints at the various levels distinguished in the test. For example, the total test length can be constrained to be $L$ by imposing the linear constraint $\sum_{i=1}^I x_i = L$. Likewise, the length for any sub-category, $c$, can be constrained to $l_c$, by imposing $\sum_{i \in V_c} x_i = l_c$ where $V_c$ defines the subset of items in the pool for categorical item attribute $c$ (e.g., a content sub-domain). Quantitative constraints can be imposed just as easily. For example, the total test time needed to complete the test can be constrained to $T$ or less by imposing $\sum_{i=1}^I t_i x_i \leq T$, where $t_i$ is the expected time needed to

 Springer
answer Item $i$; (3) **Modeling the objective:** An objective function provides a basis for determining an optimum solution among potential alternatives. The most common choice in CAT is maximizing information at the current solution among potential alternatives. The most common function provides a basis for determining an optimum solution: The preceding three steps are taken prior to the onset of the CAT administration, whereas this final step is performed in real time and involves submitting the model to a MIP solver. Modern MIP algorithms search the solution space iteratively for an optimum. However, explicitly enumerating all solutions is impractical given the astronomical size of the space for a typical test-assembly problem. Fortunately, MIP solvers preprocess the problem and then find the best settings for their implicit enumeration algorithms. Nowadays, highly efficient algorithms are typically “branch and cut,” that is, know how to cut off entire regions of the search space where an optimum appears to be impossible [1, p. 81].

**References**