Quality of Life Research (2018) 27:2393–2402
https://doi.org/10.1007/s11136-018-1898-0

Procedures to develop a computerized adaptive test to assess patient-reported physical functioning

Erin McCabe[1] · **Douglas P. Gross**[2] · **Okan Bulut**[3]

Accepted: 31 May 2018 / Published online: 7 June 2018

Abstract

Purpose The purpose of this paper is to demonstrate the procedures to develop and implement a computerized adaptive patient-reported outcome (PRO) measure using secondary analysis of a dataset and items from fixed-format legacy measures.

Methods We conducted secondary analysis of a dataset of responses from 1429 persons with work-related lower extremity impairment. We calibrated three measures of physical functioning on the same metric, based on item response theory (IRT). We evaluated efficiency and measurement precision of various computerized adaptive test (CAT) designs using computer simulations.

Results IRT and confirmatory factor analyses support combining the items from the three scales for a CAT item bank of 31 items. The item parameters for IRT were calculated using the generalized partial credit model. CAT simulations show that reducing the test length from the full 31 items to a maximum test length of 8 items, or 20 items is possible without a significant loss of information (95, 99% correlation with legacy measure scores).

Conclusions We demonstrated feasibility and efficiency of using CAT for PRO measurement of physical functioning. The procedures we outlined are straightforward, and can be applied to other PRO measures. Additionally, we have included all the information necessary to implement the CAT of physical functioning in the electronic supplementary material of this paper.

Keywords Computerized adaptive testing · Physical function · Assessment · Patient-reported outcome measure · Work disability · Item response theory

Introduction

Patient-reported outcome (PRO) measures are instruments that patients complete to provide information on aspects of their health status relevant to their quality of life, such as symptoms, and physical, mental and social health and functioning [1]. PRO measures are considered valuable sources of information for evaluating health services, establishing treatment effectiveness, and informing clinical decision-making. The routine clinical use of PRO measures may positively influence patient-provider communication, shared decision-making, and satisfaction with care [2].

Efficiency is often prioritized in clinical environments. A clinically useful PRO measure should be short and easy to administer, and gather enough information to show clinically meaningful differences within or between patients [3]. Item response theory (IRT) and computerized adaptive testing (CAT) are innovative methods to achieve this balance of efficiency and precision [4]. A CAT is essentially an individualized assessment tool. Patients receive a subset of items

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11136-018-1898-0) contains supplementary material, which is available to authorized users.

✉ Erin McCabe
emccabe@ualberta.ca

Douglas P. Gross
dgross@ualberta.ca

Okan Bulut
bulut@ualberta.ca

[1] Department of Rehabilitation Science, University of Alberta, 3-48 Corbett Hall, Edmonton, AB T6G 2G4, Canada

[2] Department of Physical Therapy, University of Alberta, 2-50 Corbett Hall, Edmonton, AB T6G 2G4, Canada

[3] Centre for Research in Applied Measurement and Evaluation, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada

from a larger bank of questionnaire items. The CAT system chooses the items that are most relevant to the patient's own health level, based on their answers to previous items. Therefore, the CAT system substantially reduces the length of the assessment, while the precision and comparability of the measurements are maintained [5].

Developing a new CAT can be a resource-intensive process. A CAT requires a bank of well-written items, which then must be calibrated using real patient response data. There are well-supported initiatives underway using rigorous methods to develop item banks that comprehensively cover a range of health constructs and patient populations. A well-known example is the Patient-Reported Outcomes Measurement Information System (PROMIS) supported by the National Institutes of Health [6]. While these initiatives will likely revolutionize PRO measurement, it will be years before item banks become available for most health constructs and patient populations. This is a problem for healthcare organizations that wish to implement CAT for measuring PROs.

An alternative approach is to transform existing fixed-format PRO measures (herein, *legacy measures*) into CAT by using an existing dataset of patient responses to calibrate the instrument. Researchers have demonstrated the feasibility of using retrospective secondary analysis of existing datasets to develop CAT for PROs [7–9]. This process is an attractive alternative, because it capitalizes on the extensive body of work to develop high-quality PRO measures, and repurposes PRO data collected in prior clinical research. Therefore, fewer resources are required to develop and implement the CAT.

The purpose of this paper is to demonstrate procedures to develop and implement a CAT using items from legacy measures. We apply these procedures to (1) create and test an operational CAT for assessing perceived physical functioning (CAT-PF), (2) evaluate various CAT designs, and (3) share the necessary information for practitioners to implement the CAT-PF.

Methods

Study design

For this study, we adopted the five-stage framework proposed by Thompson and Weiss for developing a CAT, which is outlined in Table 1 [10]. We discuss the implementation of this framework in the following sections.

Study participants

In this study, we conducted a secondary analysis of retrospective data collected from a Canadian provincial worker's compensation database—Alberta Workers' Compensation Board (WCB-Alberta). In the original study by Gross et al., data were extracted from WCB-Alberta databases on all claimants undergoing assessment at a work rehabilitation centre [11]. Claimants are typically referred for assessment in the sub-acute phase of recovery from injury, with the aim of determining readiness for work and further rehabilitation needs. At the time of the assessment, WCB claimants completed several PRO measures assessing quality of life, perceived pain using an 11-point visual analog scale (VAS) [12], disability, health status, and body-part-specific measures of physical function [11]. Demographic data were also collected.[1]

We used a subset of data from 1429 participants who completed fixed-format versions of three legacy measures: two scales from the *Short-Form 36 Health Survey* [13] and the *Lower Extremity Functional Scale* (LEFS). All participants who filled out the LEFS were somehow limited in their physical functioning (mobility, lower extremity impairment); however, their primary site of injury was not necessarily the lower extremity [14]. Participant characteristics for this subset are described in Table 2.

Stage 1: Feasibility, applicability, and planning studies

Thompson and Weiss [10] recommend evaluating the applicability and feasibility of CAT within the context where it is intended to be used. The target construct and target population should also be defined. Practitioners who prioritize efficiency and convenience in PRO measurement should consider implementing CAT because fewer items need to be administered to obtain an accurate measurement. The proliferation of smartphones and personal computers has made it more feasible to implement CAT. Patients can complete a PRO before arriving in the clinic and scores can automatically be entered into an electronic medical record. Importantly, the scores from a CAT are comparable with scores from their legacy measures [15]. Barriers to implementing CATs for PRO assessment should be considered at this stage. Potential barriers include the availability of computer devices in the clinic, the capacity for training and support of practitioners and staff, and patient privacy policies with respect to electronic data collection and storage.

[1] More information about the study participant characteristics and the conditions under which data were collected is available in the original study publication [11].

Table 1 An overview of steps for the development of a computerized adaptive test for patient-reported outcome measurement

Stage	Work	Key questions
1. Feasibility, applicability, and planning studies		
Research practical and business considerations related to implementing CAT	Develop a CAT implementation plan	Is it feasible to implement CAT in this context? Are there substantial benefits to implementing CAT in this context?
	Define target population and target construct	What is the target construct and target population for the CAT?
2. Develop an item bank		
Utilize items from legacy measures	Create an item bank, format items for CAT system	Do we have items from legacy measures that could potentially be combined to create an item bank for CAT, and a dataset of responses to those items?
3. Pretest and calibrate item bank		
Preliminary item analysis	Descriptive statistics for items and scales	Do these items have good psychometric properties?
Test IRT assumptions	Unidimensionality, local independence, monotonicity	Are the assumptions of IRT met?
Fit an IRT model	Estimate item parameters, check IRT model fit, differential item functioning	How well does the IRT model fit the data? Do the items show bias in any subgroups?
Item calibration for CAT	Test information function, conditional standard error of measurement curve	How well does the item bank measure over the range of the target construct?
4. Determine specifications for the CAT		
Evaluate CAT performance	Real data simulations varying CAT design elements	What values of CAT design elements should be used for different measurement purposes?
5. Publish live CAT		
Implementation of live CAT	Development of CAT system	Is there an open-source (cost-effective) option for implementing CAT?

Adapted from Thompson and Weiss's [10] framework for the development of a CAT and recommendations proposed by Smits et al. [18]

Target population

The intended use of the CAT-PF is an outpatient physical rehabilitation setting, where PRO measures of physical functioning are often collected to monitor the effectiveness of treatment, inform clinical decision-making, and for program evaluation [16]. The target population for the CAT-PF is individuals with a work-related sub-acute or chronic musculoskeletal injury (4–6 weeks after initial injury or longer).

Target construct

The concepts and terminology of the World Health Organization (WHO)'s International Classification of Functioning, Disability, and Health (ICF) framework be used as a basis for discussion of rehabilitation outcome assessment [3, 17]. Accordingly, we operationally defined the target construct for the CAT-PF—'perceived physical functioning'—as self-reported capability to perform physical activities (e.g., mobility, lifting) and to participate in social roles (e.g., ability to perform usual work).

Stage 2: Develop an item bank

Candidate items

An item bank for CAT should have a sufficient number of items that measure precisely across all relevant levels of the target construct [18]. Therefore, we combined items from three legacy measures that potentially measure a single underlying construct to create a preliminary item bank. These legacy measures are described below.

SF-36 Physical Function and Role-Physical scales

We used 14 items from the Medical Outcomes Study Short Form 36 Health Survey—version 2 (SF-36), which is one of the most commonly used generic health status measures in the world [19]. There are ten items in the SF-36 Physical Function scale (PhysFunc) which ask about limitations in different mobility activities (e.g., walking, going up stairs, bending, lifting) [20]. The PhysFunc items have a 3-point response scale: 1-"Yes, limited a lot," 2-"Yes, limited a little," 3-"No, not limited at all" [13]. The SF-36 Role-Physical scale (RolePhys) has four items. They assess the impact of

Table 2 Characteristics of study participants (n = 1429)

Characteristic	Mean (SD) or number of WCB claimants (%)
Age (years)	41.9 (11.9)
Time since injury (days)	142.9 (283.4)
Scale scores[a]	
Perceived Pain-Visual Analog Scale	4.9 (2.6)
Lower Extremity Functional Scale	40.3 (16.9)
SF-36 Role Physical Subscale	18.2 (4.9)
SF-36 Physical Function Subscale	8.5 (4.0)
Primary diagnoses	
Fractures	182 (12.7)
Dislocations	24 (1.7)
Sprains/strains	571 (40.0)
Lacerations	94 (6.6)
Contusions	100 (7.0)
Nerve damage	28 (2.0)
Joint disorders	330 (23.1)
Others	100 (7.0)
Anatomical site of injury	
Neck	128 (9.0)
Trunk	241 (16.9)
Upper extremity	567 (39.7)
Lower extremity	320 (22.4)
Multiple site	6 (0.4)
Not specified	167 (11.7)
Education	
Did not complete high school	190 (13.3)
High school diploma	242 (16.9)
Post-secondary education	351 (24.6)
Unknown	646 (45.2)
Sex	
Male	983 (68.8)
Female	446 (31.2)

[a] Calculated from sum total of raw scores for the items in the scale

a physical health problem on a person's capability of participating in their regular daily activities (work, accomplishing tasks) [13]. These items have a 5-point response scale ranging from 1-"All of the time," to 5-"None of the time" [13]. We felt there was sufficient evidence to combine the PhysFunc and the RolePhys scales because in factor analytic studies of the complete SF-36 instrument, their items consistently load on the same factor, usually defined as *physical health* [19].

Lower Extremity Functional Scale

The Lower Extremity Functional Scale (LEFS) was the other scale used in this study [14]. Binkley et al. developed the LEFS as a PRO measure of physical functioning in patients with a lower extremity impairment [14]. The LEFS was designed for use in clinical practice and as a research tool, and is reported to be a reliable, valid, and responsive tool for assessing physical function in populations of patients with musculoskeletal conditions [14, 21]. The LEFS asks patients to rate the level of difficulty they would have with each of 20 different physical activities. It has a 5-point response scale from 0-"extreme difficulty or unable to perform," to 4-"no difficulty." The items in the LEFS legacy version have a common opening phrase, "Today, do you or would you have any difficulty at all with..." We had to adapt the items so they could be read as stand-alone items on a computer screen.

Stage 3: Pretest and calibrate item bank

To use IRT, our item bank must meet certain requirements. We should check that our items function well (preliminary item analysis), that they are free of bias (differential item functioning), that they meet the assumptions of IRT, that they adequately fit the chosen IRT model. These steps ensure the quality and integrity of the CAT system. A detailed description of the pre-testing methods is provided in the electronic supplementary material (Online resource 1: Pretesting the item bank). All analyses were performed using R statistical software. R is a freely available language and environment for statistical computing [22].

We checked for potentially problematic items by examining the content and statistical properties of the items and legacy measures. We analyzed the items for differential item functioning (i.e., bias across population subgroups) by sex, age, education level, and diagnosis. The three key assumptions of IRT were tested: unidimensionality, local independence (LI), and monotonicity. Items are considered *unidimensional* when responses are a function of a single characteristic or variable (i.e., perceived physical functioning) [23]. We assessed dimensionality of the items from our legacy measures by fitting a one-factor confirmatory factor analysis (CFA) model to the data. The LI assumption is met when individuals' responses to items are not related to their response to other items after the target construct has been taken into account [23]. We evaluated the assumption of LI by checking for correlations between the residuals from the CFA model. The *monotonicity* assumption was checked by visual inspection of empirical plots, the data should follow the shape of the function specified by the IRT model [23].

Fitting an IRT model

The item bank will need to be calibrated using a particular IRT model. IRT is a measurement framework for the design, analysis, and scoring of instruments measuring various latent traits and health-related constructs (e.g., physical ability, pain, quality of life). When a set of items

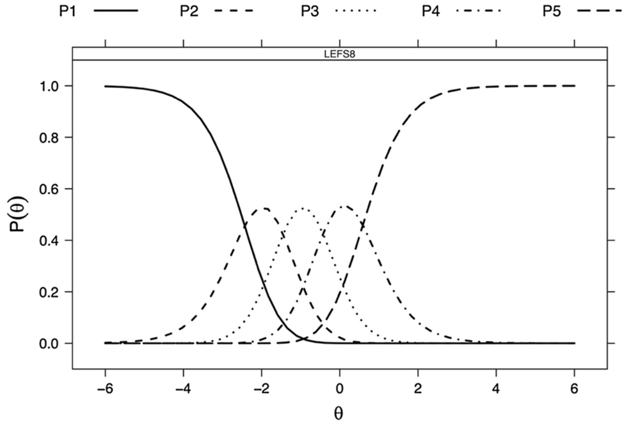

Fig. 1 An example of response option characteristic curves from GPCM. This plot was generated from item 8 of the Lower Extremity Functional Scale (*Note* P1–P5 refer to the response options in this particular item). The *x*-axis shows the latent trait (i.e., physical functioning) on a logistic scale ranging from −6 to 6, while the *y*-axis shows the probability of choosing a particular response option (labeled as P1–P5). This figure shows that respondents whose latent trait level is low (e.g., −4) are more likely to select the first response option (P1), whereas respondents whose latent trait level is high (e.g., 4) are more likely to select the last response option (P5). The steepness (i.e., slope) of the curves in this figure indicates item discrimination (i.e., the steeper the curve, the more discriminate the item) while the location of the curves on the *x*-axis shows the response category thresholds

has been calibrated to an IRT model, all scores based on any subset of those items become comparable. Readers who are interested in learning more about the basics of IRT should consult one of the many good resources on the topic [23, 24].

Different IRT models can be used, depending on the item response options and other properties. We chose the generalized partial credit model (GPCM) to calibrate the item bank [20]. The GPCM is useful because it can accommodate items with different response option labels and it estimates two types of parameters for each item: a unique discrimination for each item and the response option category thresholds (analogous to item difficulty) [23]. Examining the item discrimination parameter can give us an indication of the power of the item to distinguish between individuals with different levels of the health-related construct (i.e., how well the item can distinguish between patients with different levels of physical functioning) [23]. Response option category thresholds indicate the level of perceived physical functioning necessary to choose one response option over another (e.g., whether a patient chooses “A little bit of difficulty” over “No difficulty,” or “Quite a bit of difficulty” over “moderate difficulty”). The response threshold parameters and the level of health construct are estimated on the same logistic scale (typically ranging from −5 to 5), which enables us to compare the items in relation to the patient.

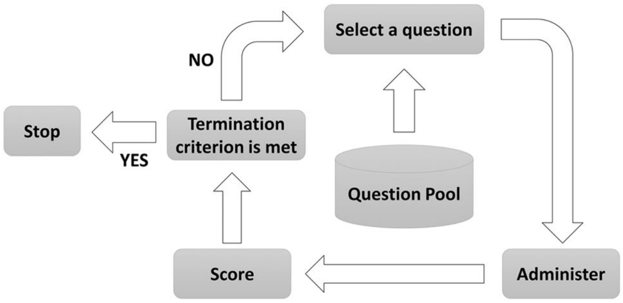

Fig. 2 A typical CAT administration

We evaluated the fit of GPCM to the data by examining the parameters and response option characteristic curves (OCCs) for each item [22, 25]. Figure 1 shows an example of an OCC from a single item. We also compared the actual responses from participants to their expected responses (predicted by the GPCM) through visual inspection of empirical plots and statistical analysis. The impact of poorly fitting items on the IRT parameters was assessed by comparing the fit of the GPCM model to a restricted GPCM (model with the potentially problematic items removed).

Item calibration

Based on the pre-testing results, we selected the items for the CAT-PF item bank, with the goal of balancing statistical fit with retaining an adequate number of items to capture all relevant levels of perceived physical functioning. We calibrated the final CAT item bank using the GPCM. We generated a plot of the test information function (TIF) and of the conditional standard error of measurement (cSEM) to examine the functioning of the item bank for individuals over a range of perceived physical functioning levels.

Stage 4: Determine specifications for the CAT

CAT administration and scoring

In order to understand the following procedures, in this section we explain how the CAT administration and scoring system works. A typical CAT administration is illustrated in Fig. 2. First, the CAT system presents a starting item for all patients. This item either can be the same for all patients or randomly selected from the item bank for each patient. Based on the response option selected by the patient, the CAT system calculates an interim estimate of physical functioning. Then, the interim estimate is used for deciding the next item for the patient. The CAT system selects the most optimal item from the item bank and presents it to the patient. This cycle continues until a termination criterion (a pre-determined stopping rule) is met. The termination

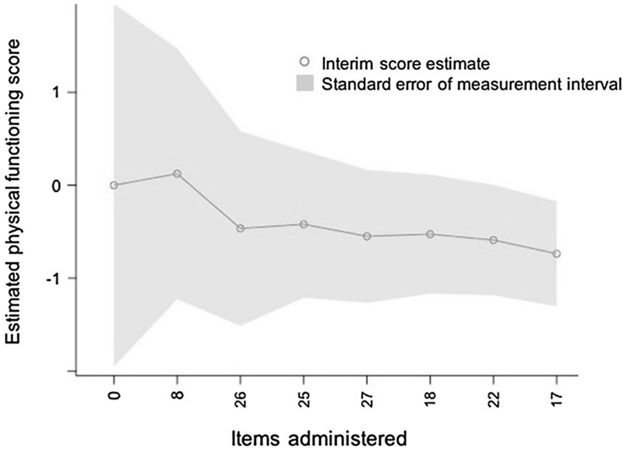

Fig. 3 Estimation of interim scores and final physical functioning score estimates in CAT

criterion may be based on a maximum test length (i.e., a predetermined number of items), a threshold of precision (i.e., degree of certainty or confidence that the score accurately reflects the patient's perceived physical functioning), or a combination of test length and precision.

Figure 3 demonstrates how the scoring in a CAT system works for a particular patient. The patient's level of physical functioning is set to zero (see the *y*-axis) at the beginning because it represents the mean IRT-score of physical functioning for the full-item bank administration during the item calibration step, and is the best guess of the CAT-PF for all individuals prior to administering any items. After each item is administered (see the *x*-axis), the patient's level of physical functioning is estimated. Because it is unlikely to measure physical functioning precisely with only a few items, the cSEM for the interim estimate is relatively wide at the beginning. As the patient takes more items, the estimate of the level of physical functioning becomes more precise (the cSEM interval becomes smaller). Once the CAT reaches the stopping point, a final estimate of physical functioning is obtained and reported.

CAT can be adapted to suit different PRO measurement contexts depending on the intended use of the scores. By changing the termination criterion, a practitioner can make trade-offs between desired measurement precision and efficiency of administration. For instance, in a busy clinical environment where PRO data are collected for the purpose of program evaluation (i.e., assessing groups), there may be a willingness to sacrifice a certain amount of precision in order to collect PRO data more efficiently. Alternatively, when a precise measure is desirable, for instance, when individual treatment decisions are being made, the practitioner may use a precision-based termination criterion of cSEM = 0.31, which is comparable to a traditional reliability of 0.90 [18]. Precision and test length can be combined to limit the length of a test for individuals who do not fall in the range of perceived physical functioning where the CAT-PF is most precise (usually those who score at the very high or very low levels of the target construct). The estimates for these individuals may never reach a pre-specified level of precision, so the maximum test length criterion may be added with the aim of decreasing the burden of administration.

Real data simulations

Real data simulations allow us to approximate the performance of CAT for hypothetical CAT design scenarios. In a real data simulation, the computer simulates the CAT-PF administration for each patient based on his or her actual responses to all items in the item bank and produces a physical functioning score estimate (i.e., 'CAT-PF scores') [10].

We conducted the real data simulations using the *mirtCAT* package (version 1.6.1) in R [22, 26]. In the simulations, we used the *maximum a posteriori* scoring algorithm to estimate patients' physical functioning level because it can estimate scores for all response patterns (e.g., for individuals whose responses are all at the extreme ends of the response scales). We chose the Maximum Fisher Information for item selection.

We tested several conditions where we manipulated the termination criteria. In the first series of CAT simulations, we set the CAT-PF to stop when it had administered a maximum number of items (i.e., maximum test length of 4, 8, 12, or 16 items). This allowed us to evaluate the number of items needed to attain a reasonably precise estimate. In the second series, we manipulated the precision-based termination criterion, at a cSEM = 0.15, as well as 0.22 and 0.31 (corresponding to traditional reliability coefficients of 0.95 and 0.90, respectively) [18]. In the final series, we tested combination rules, where a precision-based criterion of either 0.22 or 0.31 was combined with various maximum test length stopping criteria.

CAT evaluation

We evaluated the performance of the various CAT-PF designs by comparing the CAT-PF scores with IRT estimates of physical functioning calculated from responses to the full-item set (i.e., 'full-scores'). For each condition, we used the following evaluation criteria: (1) mean cSEM of CAT-PF scores (precision); (2) correlation of full-scores and CAT-PF scores (accuracy); (3) root mean square error of the difference and the average difference between CAT-PF scores and full-scores (precision); (4) average and range of the number of items administered to reach a pre-set cSEM (efficiency); (5) proportion of individuals who were administered the entire item bank (efficiency). The accuracy evaluation criteria for the 8 item CAT design are illustrated by Fig. 4.

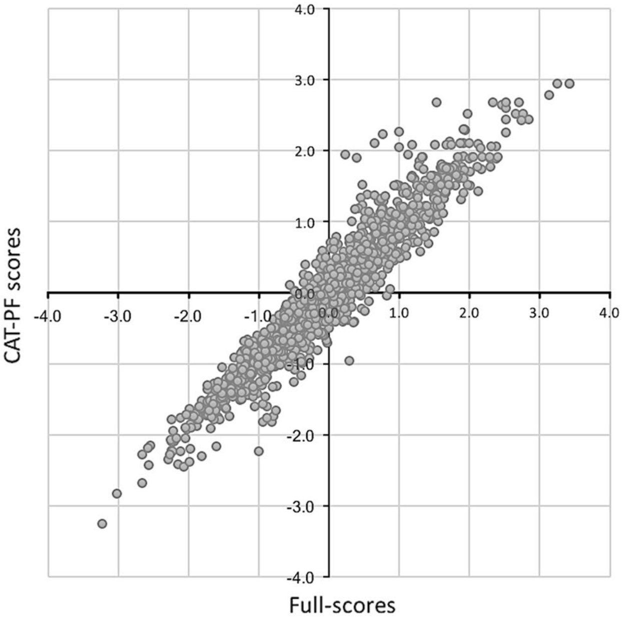

Fig. 4 Relationship between the full-scores and the scores from the CAT-PF using an 8-item maximum test length termination criterion. This figure illustrates the accuracy of the CAT-PF scores, that is, how closely they come to the scores from the full-item bank

Results

Pre-testing and calibration

The electronic supplementary material contains a detailed description of the item pre-testing results (Online resource 1: Pretesting the item bank). Briefly, preliminary analysis of the items did not reveal problematic items. All the items fit the GPCM model, and showed adequate unidimensionality (Table A3, Online resource 1), indicating that the items measure a single underlying construct. However, we detected local dependencies between some items, and we eliminated three of the four RolePhys items because of this. The remaining items performed similarly across the subgroups of sex, age, education, and diagnosis.

The final item bank for the CAT-PF consisted of 31 items, which we calibrated with the GPCM. The item parameters are available in Table A1 (Online Resource 1). The TIF and cSEM curves in Fig. 5 were calculated from the full-item bank administration. The curves indicated that our measure are most informative for individuals with physical functioning scores between − 3.0 and + 3.0.

Real data simulations

The full-scores estimates of the participants ranged from − 3.24 to 3.42. The CAT-PF evaluation results under various designs are found in Table 3. In design 4, 16 items were administered (52% of the item bank), and CAT-PF score accuracy was 0.987. The accuracy of CAT-PF scores increased from $r = 0.914$ to 0.955 when we changed the maximum test length from 4 to 8 items (design 1 and 2). When a precision-based criterion of 0.22 was used (design 7), CAT-PF administered between 12 and 31 items per participant, 10% of the sample required the maximum of 31 items, and accuracy was high ($r = 0.984$). In design 8, an average of 8 items were administered (range 6–31, $r = 0.955$) with

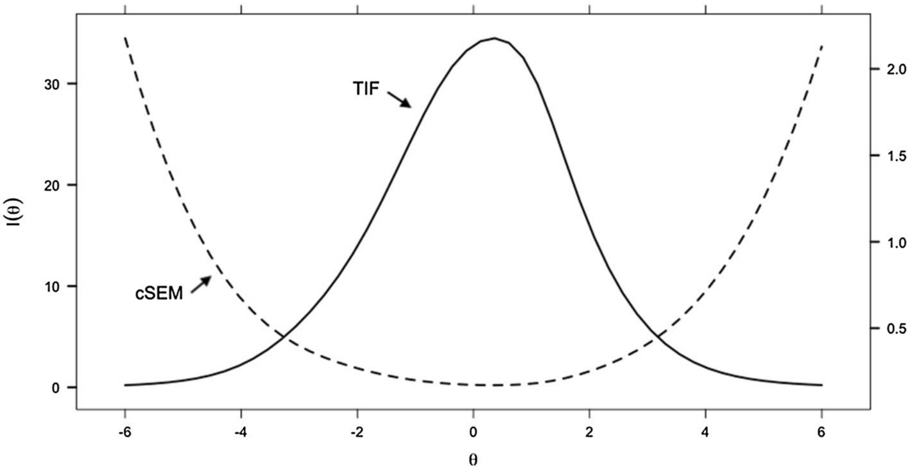

Fig. 5 The conditional standard error of measurement (cSEM) and the test information function (TIF) curves for the CAT-PF item bank. The *x*-axes in the curves represent the latent trait continuum (in this case, physical functioning) on the same logistic scale as the item response characteristic curve in Fig. 1. In IRT, the reliability (test information, on the left *y*-axis) and the precision (cSEM, on the right *y*-axis) of a test vary across the latent trait continuum. Therefore, the TIF and cSEM curves are useful in understanding how a measure will perform, that is, how informative and precise it will be at a particular level of health-related construct

Table 3 Performance of the physical function computerized adapted test (CAT-PF) under various simulation conditions

Design number	Test length[a]	cSEM[b]	CAT-PF score mean[c]	Mean cSEM[d]	Accuracy[d]	RMSD[e]	Bias[f]	Items used[g]	Range[h]	Proportion receiving max items[i]
Manipulating test length termination criteria (maximum items)										
1	4		0.0215	0.342	0.914	0.398	− 0.014			
2	8		0.0085	0.266	0.955	0.291	− 0.001			
3	12		0.0073	0.233	0.976	0.213	0.000			
4	16		0.0061	0.215	0.987	0.154	0.001			
5	20		0.0048	0.202	0.994	0.110	0.002			
Manipulating precision termination criteria (cSEM)										
6		0.15	0.0071	0.187	1.000	0.000	0.000	31.0	31–31	1.00
7		0.22	0.0042	0.221	0.984	0.175	0.003	15.7	12–31	0.10
8		0.28	0.0114	0.275	0.955	0.292	− 0.004	7.9	6–31	0.02
9		0.31	0.0155	0.300	0.939	0.336	− 0.008	6.1	5–31	0.01
10		0.35	0.0183	0.332	0.924	0.373	− 0.011	4.6	4–31	0.01
Combining termination criteria										
11	8	0.22	0.0085	0.266	0.955	0.291	− 0.001	8.0	8–8	1.00
12	12	0.22	0.0073	0.233	0.976	0.213	0.000	12.0	12–12	1.00
13	16	0.22	0.0069	0.225	0.982	0.186	0.000	13.7	12–16	0.26
14	4	0.31	0.0215	0.342	0.914	0.398	− 0.014	4.0	4–4	1.00
15	8	0.31	0.0162	0.304	0.935	0.348	− 0.009	5.6	5–8	0.11
16	12	0.31	0.0162	0.301	0.939	0.337	− 0.009	5.8	5–12	0.03
17	16	0.31	0.0162	0.301	0.939	0.337	− 0.009	5.9	5–16	0.02

[a]maximum number items to be administered before CAT are terminated, [b]precision cSEM threshold, [c]mean of the estimated CAT-PF scores, [d]average cSEM of the estimated CAT-PF scores, [d]correlation of full-scores with CAT-PF score, [e]root mean square-difference between full-scores and CAT-PF scores, [f]average difference between full-scores and CAT-PF scores, [g]average number of items administered, [h]minimum and maximum number of items administered, [i]proportion of individuals who were administered the maximum number of items

precision level of cSEM = 0.28. We found that combining the precision-based (cSEM = 0.31) and maximum test length (8 items) termination criteria (design 15) was effective in improving efficiency, while maintaining a desirable level of precision (accuracy $r = 0.935$, cSEM = 0.304). Increasing the maximum test length beyond 8 items at this precision level did not substantially improve accuracy or precision of the CAT-PF scores (design 16 and 17).

Stage 5: Implementing the CAT-PF

The final step was to develop the working CAT-PF using the *mirtCAT* package in R (version 1.6.1) [26]. We have shared information necessary to implement the CAT-PF as electronic supplementary material. We have included the R code (Online resource 2), item calibrations (Online resource 3), and the item text (Online resource 4). The costs of implementing the CAT-PF can be reduced by using freely available online CAT platforms (e.g., Concerto: Open-Source Online Adaptive Testing Platform http://concertoplatform.com [27]).

Discussion

In this paper, we have outlined the feasible and effective processes we used to develop a CAT for assessment of patient-reported physical functioning. We recycled previous work in this area by conducting a secondary analysis of existing data, and using items from legacy measures—which conserved resources.

In the real data simulations, we found that we could administer fewer items and maintain the accuracy of the CAT-PF scores. This demonstrates the potential of CAT-PF to be efficient for measuring perceived physical functioning. Many authors have reported similar findings in simulation studies of CAT for rehabilitation outcome assessment [3, 28–31]. Further, the authors who administered a live CAT in validity studies have found that CAT performs well and that it is well accepted by patients [32, 33].

Practitioners who wish to implement the CAT-PF can use results of the simulations to inform the set-up of the CAT termination criteria. For group-level measurement, a highly efficient CAT design would be to use a combination termination criteria including: (1) maximum test length of 8 items;

and (2) precision termination criteria of cSEM = 0.31. The practitioner can expect the accuracy and precision of the scores to be maintained, with far fewer items than if the fixed-format legacy measures were used.

For making individual treatment decisions, a more precise measurement may be desirable. In this case, practitioners should use a precision-based termination criterion (cSEM = 0.22). A gain in overall efficiency could be expected, with the average number of items administered being 15.7 (SD = 5.8), with some individuals requiring as few as 12 items to reach a CAT-PF score estimate with this degree of precision. The accuracy ($r = 0.98$) and precision (mean cSEM = 0.22) of the scores would not be greatly impacted.

Strengths and limitations

We had a sufficiently large sample size to calibrate the items for CAT using IRT [23]. We used legacy measures with well-established measurement properties to create our CAT item bank, the items fit the GPCM well, and the item bank as a whole measured precisely over the range of perceived physical functioning in the study participants. However, we found local dependencies between some items, and only some of those items were eliminated because of our concern with maintaining an adequate number of items in the item bank.

There are a few limitations to this study which practitioners should be aware of when implementing CAT-PF. The first is that this study was conducted using real data simulations, and thus, the CAT-PF may perform differently when implemented in clinical practice. Second, the ability to calculate separate scores for each legacy measure is lost when the items are combined into a single item bank for CAT. This may be a concern for practitioners who are interested in measuring participation-related and activity-related physical functioning as separate constructs. A third consideration is that we have not provided evidence for the use of CAT-PF to assess change within an individual over time (i.e., responsiveness). Finally, the CAT-PF item calibrations are specific to the population in this study. As with any measurement instrument, the practitioner must use clinical judgement to decide whether their intended patient population and context is similar enough to our study sample.

Conclusions

In summary, we have presented procedures to develop a CAT for measuring perceived physical functioning. Our procedures can be applied to other patient-reported outcomes. We shared the information necessary to implement the CAT-PF in clinical practice.

Funding The Workers' Compensation Board of Alberta provided the data for this research and funded the original study from which data were obtained.

Compliance with ethical standards

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

Ethical approval All study procedures were approved by the Health Research Ethics Board at the University of Alberta (Pro00073119).

Informed consent Not applicable, this study involved a secondary analysis of an existing dataset.

References

1. Canadian Institute for Health Information. (2009). PROMs background document.
2. Valderas, J. M., Kotzeva, A., Espallargues, M., Guyatt, G., Ferrans, C. E., Halyard, M. Y., et al. (2008). The impact of measuring patient-reported outcomes in clinical practice: A systematic review of the literature. *Quality of Life Research, 17*(2), 179–193. https://doi.org/10.1007/s11136-007-9295-0.
3. Jette, A. M., & Haley, S. M. (2005). Contemporary measurement techniques for rehabilitation outcomes assessment. *Journal of Rehabilitation Medicine, 37*(6), 339–345. https://doi.org/10.1080/16501970500302793.
4. Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research, 6*(6), 595–600. https://doi.org/10.1023/A:1018420418455.
5. Weiss, D. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social, 2*(1), 1–23.
6. Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., et al. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011.
7. Walker, J., Bohnke, J. R., Cerny, T., & Strasser, F. (2010). Development of symptom assessments utilising item response theory and computer-adaptive testing-A practical method based on a systematic review. *Critical Reviews in Oncology/Hematology, 73*(1), 47–67. https://doi.org/10.1016/j.critrevonc.2009.03.007.
8. Flens, G., Smits, N., Carlier, I., van Hemert, A. M., & de Beurs, E. (2016). Simulating computer adaptive testing with the mood and anxiety symptom questionnaire. *Psychological Assessment, 28*(8), 953–962. https://doi.org/10.1037/pas0000240.
9. Smits, N., Zitman, F. G., Cuijpers, P., Den Hollander-Gijsman, M. E., & Carlier, I. V. (2012). A proof of principle for using adaptive testing in routine outcome monitoring: The efficiency of the mood and anxiety symptoms questionnaire—anhedonic depression CAT. *BMC Medical Research Methodology*. https://doi.org/10.1186/1471-2288-12-4.
10. Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research and Evaluation, 16*(1), 1–9.

11. Gross, D. P., Zhang, J., Steenstra, I., Barnsley, S., Haws, C., Amell, T., et al. (2013). Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers. *Journal of Occupational Rehabilitation, 23*(4), 597–609. https://doi.org/10.1007/s10926-013-9430-4.
12. Finch, E., Brooks, D., Stratford, P. W., & Mayo, N. (2002). *Physical rehabilitation outcome measures: A guide to enhanced clinical decision making* (2nd edn.). Toronto: Canadian Physiotherapy Association.
13. McHorney, C., Ware, J. E., & Raczek, A. E. (1993). The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care, 31*(3), 247–263. https://doi.org/10.2307/3765819.
14. Binkley, J. M., Stratford, P. W., Lott, S. A., & Riddle, D. L. (1999). The Lower Extremity Functional Scale (LEFS): Scale development, measurement properties, and clinical application. *Physical Therapy, 79*(4), 371–383. https://doi.org/10.1097/00005650-198903001-00007.
15. Gibbons, L. E., Feldman, B. J., Crane, H. M., Mugavero, M., Willig, J. H., Patrick, D., Schumacher, J., Saag, M., Kitahata, M. M., & Crane, P. K. (2011). Migrating from a legacy fixed-format measure to CAT administration: calibrating the PHQ-9 to the PROMIS depression measures. *Quality of Life Research, 20*(9), 1349–1357. https://doi.org/10.1007/s11136-011-9882-y.
16. Riddle, D. L., & Stratford, P. W. (2013). *Is this change real? Interpreting patient outcomes in physical therapy*. Philadelphia, PA: F.A. Davis Co. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=550421.
17. World Health Organization. (2002). Towards a common language for functioning, disability and health ICF. International classification. WHO/EIP/GPE/CAS/01.3.
18. Smits, N., Paap, M. C. S., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-1821-8.
19. McDowell, I. (2006). *Measuring health: A guide to rating scales and questionnaires* (3rd edn.). New York: Oxford University Press.
20. Syddall, H. E., Martin, H. J., Harwood, R. H., Cooper, C., & Sayer, A., A (2009). The SF-36: A simple, effective measure of mobility-disability for epidemiological studies. *The Journal of Nutrition, Health & Aging, 13*(1), 57–62. https://doi.org/10.1007/s12603-009-0010-4.
21. Mehta, S. P., Fulton, A., Quach, C., Thistle, M., Toledo, C., & Evans, N. A. (2016). Measurement properties of the lower extremity functional scale: A systematic review. *The Journal of Orthopaedic and Sports Physical Therapy, 46*(3), 200–216. https://doi.org/10.2519/jospt.2016.6165.
22. R Core Team. (2016). R: A language and environment for statistical. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
23. de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press. Retrieved from https://www.library.ualberta.ca/catalog/4469201.
24. Baker, F. (2001). The basics of item response theory. ERIC clearinghouse on assessment and evaluation. University of MD. Retrieved April 21, 2018, from http://echo.edres.org:8080/irt/baker/.
25. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06. doi.
26. Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1–39. https://doi.org/10.18637/jss.v071.i05. doi.
27. The University of Cambridge Psychometrics Centre. (n.d.). Concerto. Open-source online adaptive testing platform. Retrieved from http://concertoplatform.com/.
28. Hart, D. L., Mioduskib, J. E., Wernekec, M. W., & Stratford, P. W. (2006). Simulated computerized adaptive test for patients with lumbar spine impairments was efficient and produced valid measures of function. *Journal of Clinical Epidemiology, 59*, 947–956.
29. Haley, S. M., Coster, W. J., Andres, P. L., Kosinski, M., & Ni, P. (2004). Score comparability of short forms and computerized adaptive testing: Simulation study with the activity measure for post-acute care. *Archives of Physical Medicine and Rehabilitation, 85*(4), 661–666. https://doi.org/10.1016/j.apmr.2003.08.097.
30. Jette, A. M., McDonough, C. M., Haley, S. M., Ni, P., Olarsch, S., Latham, N., et al. (2009). A computer-adaptive disability instrument for lower extremity osteoarthritis research demonstrated promising breadth, precision, and reliability. *Journal of Clinical Epidemiology, 62*(8), 807–815. https://doi.org/10.1016/j.jclinepi.2008.10.004.
31. McDonough, C. M., Ni, P., Peterik, K., Marfeo, E. E., Marino, M. E., Meterko, M., et al. (2017). Improving measures of work-related physical functioning. *Quality of Life Research, 26*(3), 789–798. https://doi.org/10.1007/s11136-016-1477-1.
32. Anatchkova, M., Rose, M., Ware, J., & Bjorner, J. B. (2013). Evaluation of a role functioning computer adaptive test (RF-CAT). *Quality of Life Research, 22*(5), 1085–1092. https://doi.org/10.1007/s11136-012-0215-6.
33. Hart, D. L., Wang, Y.-C., Stratford, P. W., & Mioduski, J. E. (2008). Computerized adaptive test for patients with lumbar spine impairments produced valid and responsive measures of function. *Journal of Clinical Epidemiology, 61*(4), 1113–1124.