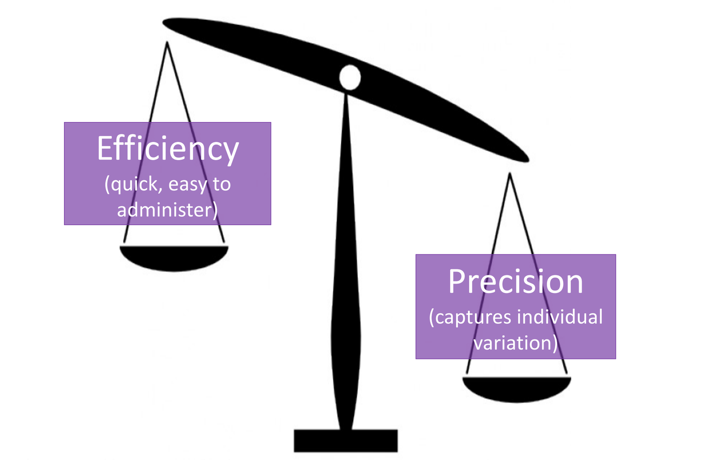
Developing computerized adaptive tests to improve the efficiency of patient-reported outcome assessment: Clinically feasible procedures

Erin McCabe[1], Douglas Gross[2], Okan Bulut[3]

[1]Department of Rehabilitation Sciences, University of Alberta, Edmonton, [2]Department of Physical Therapy, University of Alberta, Edmonton, [3]Department of Educational Psychology, University of Alberta, Edmonton.

INTRODUCTION

The time and resources related to administering, scoring and recording patient-reported measures can limit their use in clinical settings.



A clinically useful patient-reported outcome measure will be maximally efficient and precise.

Computerized adaptive testing addresses the challenges of patient-reported outcome measurement.

OBJECTIVE

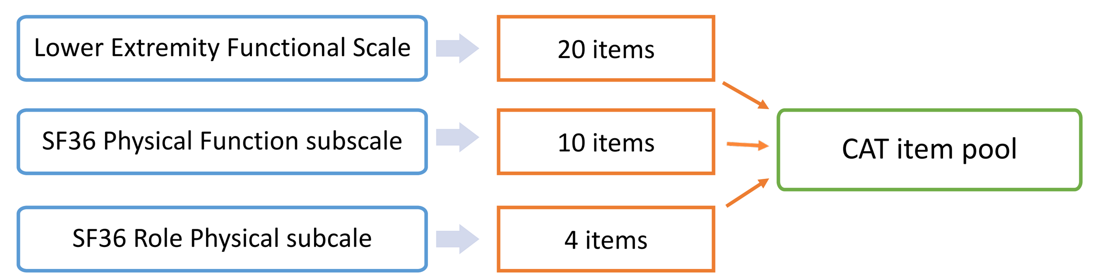
Demonstrate clinically feasible procedures to develop a computerized-adaptive patient-reported outcome measure.

COMPUTERIZED ADAPTIVE TESTING

- In **computerized adaptive testing** (CAT), patients receive a unique set of items from a large item bank targeted towards their own health status.
- CAT successively selects questions, based on what is known about the patient from their previous responses.
- CAT is designed to stop when it reaches a pre-determined threshold (e.g. maximum number of items, or a minimum standard error of measurement).
- The individualized test produces a reliable measurement with far fewer items than traditional questionnaires.

METHODS

We used the items in the 'Lower Extremity Functional Scale', the Medical Outcomes Study Short Form-36 'Role-Physical', and 'Physical-Function' subscales to create a CAT for physical functioning.



We analyzed an existing dataset of responses (n=1,429) to the scales, collected from workers with lower extremity impairment. Data analysis was conducted using the *mirtCAT* packages in R and RStudio.

Our key steps were (Figure 1):

1. We tested the items to ensure they are appropriate for use in CAT.
2. Calibration of the items using our dataset of responses.
3. We conducted computer simulations using the *mirtCAT* package in R. These allowed us to approximate how the CAT would perform with real respondents. We had two aims:
   (1) inform the design elements of our CAT
   (2) evaluate the performance of CATs of varying lengths

We developed the working CAT, using the *mirtCAT* package in R. This software is freely available.

Figure 2. CAT uses items that have been calibrated to a single metric (continuum of physical functioning), the person's scores will be on the same scale.

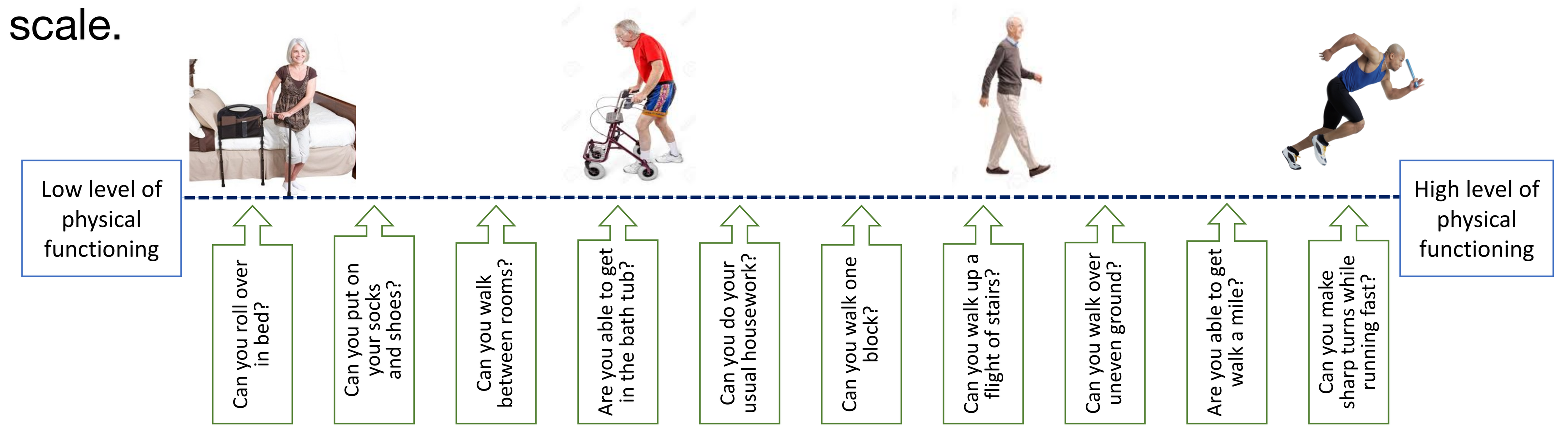


Figure 1. Key Steps in developing a CAT for clinical practice

1. **Select existing PROs to create a CAT item bank**
   - LEFS
   - SF36 Role Physical
   - SF36 Physical Function
2. **Obtain a dataset of responses**
   - Secondary analysis of a dataset, or
   - Collect data (responses to items)
3. **Calibrate the items for CAT using the dataset of responses**
4. **Evaluate the performance of various CAT designs**
   - Manipulate test length, precision, algorithms
5. **Develop the working CAT**
   - Use freely available software

Table 1. Performance of CAT of varying lengths.

Length of CAT	Correlation with legacy measures
4 Items	0.90
8 Items	0.95
12 Items	0.97
20 Items	0.99

Table 2. Performance of CAT when the precision stopping rule is manipulated.

Precision stopping point (SE)	Correlation with legacy measures	Number of items given mean [range]
0.20	0.99	21.7 [17-34]
0.25	0.96	11.3 [8-34]
0.30	0.94	7.0 [5-34]
0.35	0.92	4.6 [4-34]

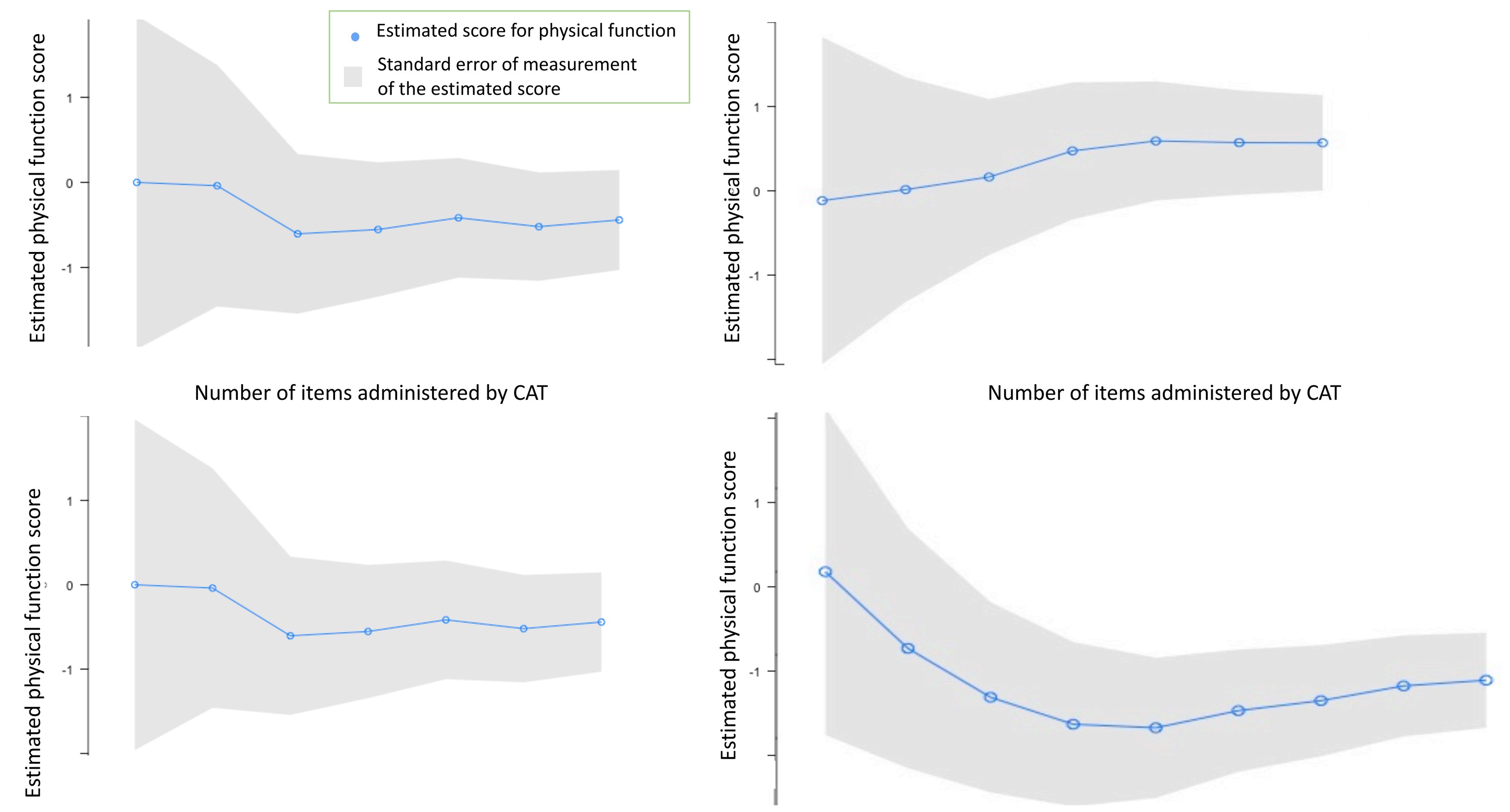


Figure 3. Examples of CAT scoring responses from four individuals. The estimated score is in blue, while the grey area represents the standard error of measurement (SE) for the estimated score. This illustration shows how the score estimate changes, and the estimate becomes more precise (SE decreases) as items are administered.

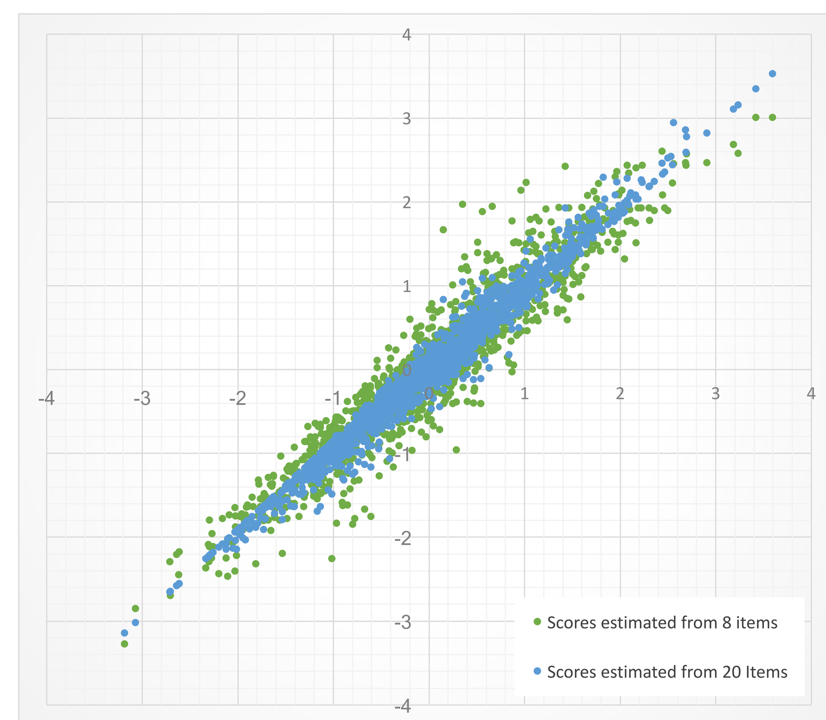


Figure 4. Relationship Between Scores from the legacy measures to CAT scores.

This figure illustrates the impact of manipulating test length on precision. For a CAT with a maximum test length of 8 items, the correlation between the full questionnaires and the CAT scores is lower. If a more precise measurement is needed, then a longer test of 20 items will produces a more reliable measure.

CONCLUSION

CATs are an efficient option for patient-reported outcome measurement. We have demonstrated clinically feasible procedures for developing and implementing CATs into clinical practice.

13th Annual Spotlight on Research Breakfast, Edmonton, AB. Nov 22, 2017