Qual Life Res (2013) 22:1085–1092
DOI 10.1007/s11136-012-0215-6

Evaluation of a role functioning computer adaptive test (RF-CAT)

M. Anatchkova · M. Rose · J. Ware · J. B. Bjorner

Accepted: 22 May 2012 / Published online: 14 June 2012

Abstract

Objectives To evaluate the validity and participants' acceptance of an online assessment of role function using computer adaptive test (RF-CAT).

Methods The RF-CAT and a set of established quality of life instruments were administered in a cross-sectional study in a panel sample ($n = 444$) recruited from the general population with over-selection of participants with selected self-report chronic conditions ($n = 225$). The efficiency, score accuracy, validity, and acceptability of the RF-CAT were evaluated and compared to existing measures.

Results The RF-CAT with a stopping rule of six items with content balancing used 25 of the available bank items and was completed on average in 66 s. RF-CAT and the legacy tools scores were highly correlated (.64–.84) and successfully discriminated across known groups. The RF-CAT produced a more precise assessment over a wider range than the SF-36 Role Physical scale. Patients' evaluations of the RF-CAT system were positive overall, with no differences in ratings observed between the CAT and static assessments.

Conclusions The RF-CAT was feasible, more precise than the static SF-36 RP and equally acceptable to participants as legacy measures. In empirical tests of validity, the better performance of the CAT was not uniformly statistically significant. Further research exploring the relationship between gained precision and discriminant power of the CAT assessment is needed.

Keywords Role function · Computer adaptive test · Patient-reported outcome · Health-related quality of life

Introduction

In recent years, there has been an increased interest in the measurement and use of patient-reported outcomes (PRO) in clinical research and health care delivery studies and in the implementation of modern psychometric approaches for the improvement of precision and efficiency of PRO tools [1]. Computerized adaptive testing (CAT) based on item response theory models allows the selection of the most appropriate items for each respondent. [2–5]. This is a promising strategy in the development of improved health outcome measures [6, 7]. Calibration of all items on the metric of the general underlying dimension (e.g., the impact of health on role functioning) allows the computation of one overall impact score, while at the same time, items from all relevant content areas can be tailored using item selection algorithms that can take model parameters and content into account. The IRT approach has been promoted by the NIH-sponsored patient-reported outcomes measurement information system (PROMIS) initiative,

M. Anatchkova (✉) · M. Rose · J. Ware
Department of Quantitative Health Sciences, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, MA 01665-002, USA
e-mail: Milena.Anatchkova@umassmed.edu

M. Rose
Medical School, Charité, University Medicine Berlin, Berlin, Germany

J. Ware
John Ware Research Group Inc, Worcester, MA, USA

J. B. Bjorner
National Institute of Occupational Health, Copenhagen, Denmark

J. B. Bjorner
i3 QualityMetric Inc, Lincoln, RI, USA

aiming to develop precise and efficient measures of patient-reported symptoms, functioning, and health-related quality of life [1, 8]. A growing number of independent researchers in different health areas have also turned to this approach for improving PRO measurement. As a result, many item banks assessing various health-related quality of life (HRQOL) and functional areas have been developed [3, 9–16]. In addition, many reports exist on the results of empirical simulation studies exploring psychometric characteristics of CAT tools based on these banks, but simulation studies may favor CAT results, because the same items and participants that are used to build the CAT are then used to estimate person scores [3, 17]. Validity studies using independent samples in a variety of settings are needed to further evaluate CAT's validity and performance in the area of HRQOL; however, such results are still relatively rare.

This study builds on our previous work where we developed and tested a role functioning item bank covering three content domains (family, social, and work), following a previously described [6] multistage approach. Our conceptualization was inspired by the biopsychosocial model of health and disability and the International Classification of Functioning, Disability and Health (ICF) [18]. A series of focus groups with participants with diverse educational and ethnic background were used to explore relevance and importance of roles, perception of the impact of health on role functioning, and elicit feedback on the format and content of sample items. Based on this work, we established a theoretical model and generated 87 new items with health attribution and a four-week recall period [19]. In a previously reported confirmatory factor analyses (sample size $n = 2{,}500$), we tested the assumptions of unidimensionality and local independence that are critical for IRT analyses. A bi-factor model with item loadings for the overall health impact factor was retained, allowing us to consider the item bank sufficiently unidimensional for applications that require unidimensionality, such as IRT [20]. To estimate the item parameters for each domain on a common metric, we used the generalized partial credit model (GPCM) [21]. The final item bank had a total of 64 acceptably fitting, and the IRT model items covering three general content areas (family, social, and occupational roles). We used computer simulations with real data to compare the psychometric merits of alternative strategies for programming CAT assessments of role functioning [22]. A fixed six-item stopping rule balancing items from the three content areas (family, occupation, and social) was retained for comprehensive content coverage [20]. Some preliminary validation work was also completed with the data from the calibration study. Results suggested that the full bank and the CAT assessment discriminated equally well between clinical groups and general health status. In addition, the static SF-36 RP scale was also comparable to the CAT in its ability to differentiate between examined groups. The CATs extend the range of the continuum covered with high precision compared to the SF-36 RP scale, but the gains in precision and coverage of the CATs did not lead to significantly better discrimination of groups [23].

Here, we evaluate the validity of a CAT (RF-CAT) based on the RF item bank in an independent sample of general population and participants with selected self-reported chronic conditions. Our objectives were to (1) assess administrative efficiency, content range coverage, and measurement precision; (2) assess known group construct validity (e.g., the ability of the instruments to differentiate between groups of patients known to differ in their level of role functioning); and (3) evaluate participants' acceptance of the tool.

Methods

Participants

Participants were recruited for the study via the Internet by a panel company (www.YouGovPolimetrix.com). We aimed to recruit a sample that was stratified across age groups with equal gender representations and race and ethnicity quotas representing the US population. Half of the sample was designed to include participants with selected self-reported chronic conditions (asthma, heart disease, diabetes, and auto-immune) and the other half was recruited as general population, but excluding participants with these conditions. All participants completed a consent form and received an incentive for their participation in the study in the form of "polling points."

Instruments

The main focus of this research was to evaluate the RF-CAT, based on a newly developed RF bank of 64 items. The bank was designed to assess the role functioning of participants within various relevant roles in the domains of family, occupational, and social life. To avoid presenting irrelevant items to participants, skip patterns were used in the CAT, so that each participant only answered questions that were relevant to his/her social roles. For comparison, we included several existing scales measuring different domains of role functioning. The complete Role Physical scale (RP Scale) of the SF-36 Health Survey [24] assesses limitations with work or other daily activities due to physical problems. The presenteeism questions of the World Health Organization's Heath and Work Performance Questionnaire (HPQ) [25] allow the assessment of absolute and relative presenteeism. Absolute presenteeism is conceptualized as the actual performance of an individual in relation to possible performance, while relative presenteeism is conceptualized as the ratio of actual performance to the performance of most workers at the same job. The short form of the Work Limitations Questionnaire (WLQ) [26] measures the degree to which employed individuals are experiencing on-the-job limitations due to their health problems and health-related productivity loss. Respondents were also asked to complete the SF-12v2 Health Survey [27], the CDC-HRQOL-14 [28]—two legacy tools including questions on role functioning, which were used in comparisons with user acceptance, assessed through self-report.

Statistical analyses

This paper reports the evaluation and validity test results of an RF-CAT based on an item bank from previously described item bank development process [19, 20]. In this study, we evaluated the real-life RF-CAT for item usage, efficiency (average time to complete the RF-CAT), measurement accuracy, range of measured levels (ceiling and floor effects), concordant and discriminant validity, and participants' acceptance. In addition, we compared the discriminant validity of the RF-CAT to some of the legacy measures of role functioning and work performance.

To evaluate measurement accuracy, we examined the descriptive characteristics and the plots of 95 % confidence intervals (±1.96* standard error of measurement) against norm-based scores for three tools: the RF-CAT assessment (six items), the SF-36 RP scale (four items) scored with IRT parameters derived from our earlier work, and a simulated RF-CAT with four items. For the simulated CAT, we used real data simulation to select four out of the six items administered by the real-life CAT. All IRT scores were computed in a normed metric with a mean of 50 and a standard deviation of 10. To evaluate "ceiling" and "floor" effects, we examined the data from the CAT for cases in which all administered items received the highest or the lowest score.

We used correlation analyses to determine the concurrent validity of the RF-CAT by examining its association with the SF-36 RP scale, the WLQ, and HPQ presenteeism scale. In order to explore the discriminant validity of the RF-CAT, we selected participants with self-reported chronic conditions associated with varying levels of role functioning impairment (asthma, heart disease, diabetes, and auto-immune disease) and compared their scores to participants who did not suffer from any of the selected conditions using an analysis of variance procedure in SAS and relative validity coefficients for comparisons with results from established tools [29]. The same set of analyses was performed for participants with different scores on self-reported general health status evaluated through the general health items of the SF-36 survey. Participants were classified in 5 different groups of general health (poor, fair, good, very good, and excellent). In addition, for employed participants, we examined the ability of the measures to differentiate between groups of people who had taken no sick days, one sick day, or more than one sick day over the last month. For each comparison, relative validity (RV) estimates were obtained by dividing the F-statistic of the comparison CAT measure by the F-statistic for the real-life RF-CAT with six items. The F-statistic for a measure will be larger when the measure produces a larger average separation in scores for groups being compared or has a smaller within-group variance, or both. The RV coefficient for each measure in a given test describes, in proportional terms, the empirical validity of that scale, relative to the most valid scale in that test.

Participants' acceptance of the tools was evaluated through descriptive analyses of the responses to the user acceptance questions. Participants were randomly assigned to complete the user evaluation questions after the RF-CAT, the SF12v2, or the CDC-HRQOL, so we examined the differences in evaluations to see whether evaluations for different tool will be different. In addition, we performed a qualitative evaluation of comments provided by participants presented to an open-ended question format from the user evaluation tool asking them to "provide additional comments including suggestions on how to improve the survey."

Results

Demographic characteristics

A total of 503 registrations were recorded. After examining the records, we excluded 59 participants who had completed less than 90 % of administered items (21 of these dropped out with completion rates below 50 %). The results presented here are based on the remaining 444 records. The mean age of respondents was 50 (SD = 16) years (range, 18–88), 51 % were female, 79 % Caucasian, 16 % of the sample had a high school or lower education, 25 % were college graduates, and 23 % had postgraduate degrees (Table 1). Two subsamples were drawn: a general population sample ($n = 219$) and a chronic disease sample ($n = 225$) comprised of respondents indicating at least one of four conditions (asthma $n = 55$, heart disease $n = 21$, diabetes $n = 46$, auto-immune $n = 33$, and multiple condition $n = 64$).

Table 1 Characteristics of the sample ($n = 444$)

Age	50 (18–88)
	% (n)
Gender	
Female	50.68 (225)
Ethnicity	
Hispanic or Latino	7.21 (32)
Race	
Black or African–American	7.67 (34)
White	79.91 (354)
Asian or from the Indian subcontinent	2.93 (13)
American Indian/Alaskan native	1.81 (8)
Native Hawaiian or other Pacific Islander	.68 (3)
Other	4.97 (22)
Education	
Eighth grade or less	.68 (3)
Some high school	1.81 (8)
High school graduate	13.77 (61)
Some college	35.44 (157)
College graduate	24.60 (109)
Postgraduate education or degree	23.48 (104)
Family status	
Living alone	19.5 (86)
Living with a partner	39.91 (176)
Living with a partner and children	26.98 (119)
With family other than partner/children	9.75 (43)
Single parent	3.85 (17)
Family income	
Less than $5,000	1.37 (6)
$5,001–$20,000	11.42 (50)
$20,001–$45,000	26.03 (114)
$45,001–$75,000	22.37 (98)
$75,000–100,000	16.21 (71)
More than $100,000	15.30 (67)
Chronic condition	
Asthma	12.56 (55)
Diabetes	10.50 (46)
Coronary artery disease (heart disease)	4.79 (21)
Auto-immune disease	7.53 (33)
Multiple conditions selected	14.61 (64)
No, I do not have any of these conditions	50 (219)
Employment status	
Working at a paying job full time	34.23 (152)
Working at a paying job part time	9.68 (43)
Self-employed	8.56 (38)
Student	6.53 (29)
Unemployed for health reasons	9.91 (44)
Retired	24.10 (107)
Laid off or unemployed	6.08 (27)
A full-time homemaker	8.56 (38)
Other	2.93 (13)

Efficiency and Item usage

The RF-CAT was programmed with a content balanced stopping rule mandating administration of six items—two from each content area (family, occupational, and social life). Using this stopping rule across all participants, the RF-CAT selected for administration 25 of the 64 items in the item bank. The average time for the completion of the RF-CAT was 66 s (SD = 47 s; range, 12–596 s) (participants ($n = 6$) with registered time over 15 min were excluded in this calculation, as it was determined in these cases most likely there were technical difficulties and the system timed out).

Measurement range and score accuracy

The RF-CAT administered six items with five response options, and there were no respondents selecting only the lowest or only the highest response option for all items, meaning there were no ceiling and floor effects. For the simulated CAT with four items, there was no floor effect, while 27 % of participants were at the ceiling of the scale. For the SF-36 RP scale, 3 % of participants were at the floor and 28 % of participants at the ceiling of the scale. The score accuracy achieved by the actual RF-CAT with six items was better than the score accuracy of the SF-36 RP scale. The RF-CAT covered the score range 30–50 with reliability corresponding to Chronbach's alpha of .95, while at the same reliability level, the SF-36 RP scale covered the score range 38–45, resulting in an improved precision of the RF-CAT compared to the SF-36 RP scale over a range of more than one standard deviation. Using the available data from RF-CAT with six items, we also simulated a four-item CAT for a head-to-head comparison with the SF-36 RP scale. The four-item CAT still had better precision than the SF-36 RP scale, covering the score range of 33–48 with reliability of .95, but did not provide score assessment in the higher end of the continuum.

Validity

RF-CAT and all legacy measures produced scores that were significantly different for participants with and without chronic conditions and across levels of self-reported general health (Table 2). As expected, participants without the selected chronic conditions had higher role functioning scores. The scores also increased with better levels of self-reported general health. The relative validity coefficients indicated that only the WLQ had lower ability to differentiate across the selected external criteria compared to the RF-CAT. Scores based on the SF-36 RP scale differentiated equally well between groups as the RF-CAT, despite the lower levels of precision demonstrated in Fig. 1.

Table 2 Discriminant validity results

	No	Poor		Fair		Good		Very good		Excellent		*F*	*P*	RV*	RV 95 % CI	R^2
		$n = 31$		$n = 82$		$n = 156$		$n = 125$		$n = 47$						
	Items	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD					
a General health status																
RF-CAT (6 items)	6	33.35	4.48	37.86	5.30	43.77	6.11	47.38	5.95	50.32	5.43	74.87	<.0001	1.00		0.4
RF-CAT (4 items)	4	33.25	5.35	37.17	5.16	42.88	5.51	45.90	5.10	48.13	4.77	72.44	<.0001	0.97	.88–1.05	0.4
SF-36 RP scale	4	25.64	8.40	35.50	10.49	47.20	9.00	51.09	8.05	53.36	7.34	88.27	<.0001	1.18	.92–1.45	0.44
SF-36 RP Theta	4	32.10	5.13	37.47	5.33	43.86	5.62	46.79	5.72	48.91	5.03	80.42	<.0001	1.07	.86–1.31	0.43
Work performance measures																
HPQ Absolute presenteism	1	29.26	25.40	58.33	23.85	73.74	19.35	82.01	16.17	84.78	16.15	54.86	<.0001	0.73	.49–1.06	0.35
HPQ Relative presenteism	2	0.81	0.46	0.88	0.28	1.07	0.32	1.11	0.27	1.14	0.23	12.48	<.0001	0.17	.09–.30	0.11
WLQ Productivity loss	8	10.34	4.71	7.05	4.38	3.85	3.12	2.73	2.86	1.99	2.49	30.21	<.0001	0.40	.25–.64	0.3

	No. items	Condition		No condition reported		*F*	*P*	RV*	RV 95 % CI	R^2
		$n = 219$		$n = 219$						
		Mean	SD	Mean	SD					
b. Chronic condition										
RF-CAT (6 items)	6	41.19	6.96	46.19	7.06	55.5	<.0001	1.00		0.11
RF-CAT (4 items)	4	40.42	6.50	44.73	6.27	50.02	<.0001	0.90	.78–1.01	0.10
SF-36 RP scale	4	41.68	12.26	48.89	10.29	44.27	<.0001	0.80	.53–1.12	0.09
SF-36 RP Theta	4	40.99	7.08	45.49	6.67	47.08	<.0001	0.85	.57–1.22	0.10
Work performance measures										
HPQ Absolute presenteism	1	66.35	26.88	77.05	19.66	21.4	<.0001	0.39	.15–.73	0.05
HPQ Relative presenteism	2	1.02	0.36	1.06	0.27	1.99	ns	0.04	.002–.20	0.00
WLQ Productivity loss	8	4.99	4.22	3.36	3.49	12.89	<.0004	0.23	.05–.47	0.04

	No	No sick days		1 sick day		More than 1 sick day		*F*	*P*	RV*	RV 95 % CI
		$n = 54$		$n = 84$		$n = 42$					
	Items	Mean	SD	Mean	SD	Mean	SD				
c. Sick days in last 4 weeks (employed participants only)											
RF-CAT (6 items)	6	45.88	5.21	47.45	6.20	42.72	6.22	8.91	<.0002	1.00	
RF-CAT (4 items)	4	44.81	4.83	46.04	5.26	41.81	5.51	9.30	<.0001	1.04	.85–1.30
SF-36 RP scale	4	50.56	6.78	51.60	7.07	46.57	10.96	5.47	<.005	0.61	.13–1.42
SF-36 RP Theta	4	45.88	5.21	47.45	6.20	42.72	6.22	5.02	<.005	0.56	.17–1.31
Work performance measures											
HPQ absolute presenteism	1	83.51	14.16	79.39	13.19	75.47	16.85	3.72	<.02	0.42	.06–1.94
HPQ relative presenteism	2	1.12	0.27	1.12	0.28	1.06	0.32	0.65	ns	0.07	.006–.75
WLQ productivity loss	8	2.98	3.03	2.64	2.50	4.52	4.26	4.86	<.008	0.55	.14–1.48

* Relative Validity

For the subsample of employed participants (part-time and full-time employment), the RF-CAT successfully differentiated between people who had taken sick days in the past months and those who have not, as did the other measures. While once again there was a trend for the RF-CAT to differentiate better between the selected groups as indicated by the higher F values, the bootstrapped CI coefficients were very wide and suggested the measures did not vary in the sensitivity to differences.

Participants' acceptance

Participants' evaluations of the tools system were positive overall, with no differences in ratings observed between the

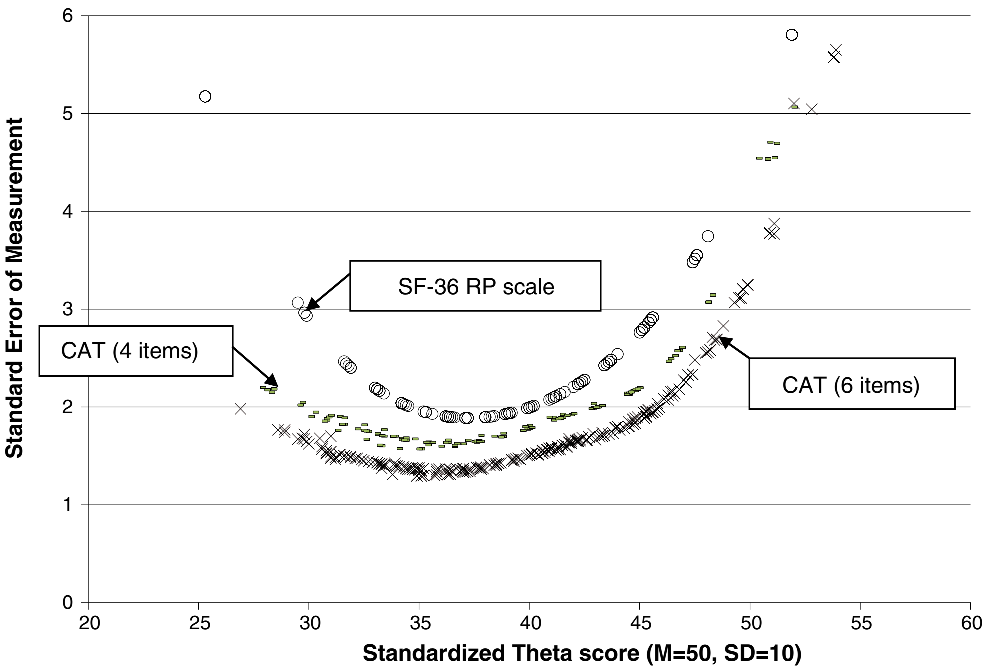

Fig. 1 Measurement precision of CAT's in relation to SF36RP scale

Table 3 Qualitative comment by measure evaluated

	SF12v2	CDC-HRQOL	RF-CAT	Total
Positive evaluations	20	14	12	46
Negative evaluations	2	2	2	6
Recommendations	13	19	19	51
Status clarification	4	9	11	25

RF-CAT and static assessments, indicating the RF-CAT was as well accepted as the SF12v2 and the CDC-HRQOL. Overall, 46 % of participants found the assessment to be useful or somewhat useful, 60 % found it to be at least somewhat relevant, 95 % found the length to be appropriate, 98 % found it easy or very easy to complete, 75 % were very willing, and another 18 % somewhat willing to answer the questions again.

A total of 127 comments were provided in response to the open-ended question and of these, 39 were completed after SF12v2, 44 after the CDC-HRQOL survey, and 44 after the RF-CAT. Comments were classified by their content into one of four categories (negative evaluation ($k = 6$), positive evaluation ($k = 46$), recommendations for change ($k = 51$), and comments on medical status of participants ($k = 26$). Once again positive evaluations were more prevalent than negative ones. Participants noted the easiness of completion, the computer graphics, and brevity of completion as positive characteristics of the assessments. The majority of the recommendations were focused on providing more detail in the questions, as participants felt that sometimes questions are very vague and do not provide an accurate evaluation of their health status. Some recommendations were also provided for better software and appearance solutions, but these tended to be inconsistent and even sometimes contradictory (e.g., both smaller and larger fonts were recommended). More positive evaluations and fewer recommendations were provided for the SF12v2 (see Table 3.).

Discussion

The results of this study provided evidence that the RF-CAT is a feasible, efficient, valid, and well-accepted measure. Findings are consistent with the preliminary validity testing that we performed as part of the item bank development process [19, 20, 23]. In the current independent sample, the RF-CAT assessment and the SF-36 RP scale once again had comparable ability to discriminate between different clinical groups and general health status, despite the fact that the RF-CAT improved precision of assessment over a wider range of scores. In this study, we explored the performance of a six-item CAT based on a stopping rule derived from earlier simulation studies and a four-item CAT for direct face-to-face comparison with the SF-36 RP scale. Both measures performed well with the six-item CAT being slightly better as could be expected by a longer measure.

As the item bank contained items assessing occupational role functioning, we also compared the RF-CAT to established work performance measures assessing work productivity (WLQ) and presenteeism (HPQ). The RF-CAT differentiated better than the WLQ between the groups of patients with and without any chronic conditions and groups with different levels of self-reported general health. While the RV coefficients were higher for the RF-CAT in the sick days analyses, the confidence intervals for the differences between the measures were so wide that this difference could not be considered important.

The RF-CAT was as well accepted as established self-report measures like the SF12v2 and the CDC-HRQOL scale, suggesting that the lower response burden did not have a significant impact on subjective evaluations of the tool. All measures were found to be useful by about half of participants in the study, possibly reflecting the inclusion of participants with no role impairment for whom the measure is of limited relevance. Qualitative evaluations of the measure also generated some suggestions for improvement, which could be useful in further refinement of existing items and/or development of new ones.

Some of the results of this study were in line with expectations for improvements in measurement brought by the use of computer adaptive testing: the measure was feasible, and it did provide improved precision of assessment and has the potential for better tailoring of questions to each individual participant. However, these gains did not lead to some of the expected practical advantages in the current study. Namely, the gains in precision did not lead to universally improved ability to discriminate between selected known groups of participants, nor did a lower number of questions lead to better acceptance by participants.

To some degree, these findings can be explained by some limitations in our study. We used an Internet-based sample and relied entirely on self-report for the assessment of all criterion variables. In addition, some of our relative validity analyses used smaller sample sizes, leading to very wide confidence intervals for relative validity coefficients, even when the differences in the observed values were substantial.

On the other hand, these findings also raise some interesting questions regarding the relationship between improved measurement precision and practical implications of assessment. It would be interesting, for example, to determine the degree of improvement in precision required to achieve gains in discriminant ability of a tool at the group level. More studies are also needed to evaluate the advantages of improved measurement precision in settings where individual level of assessment is needed.

As one of the few studies to conduct a validity test of a CAT in a field study and a head-to-head comparison between an IRT-based CAT measure and an established tool, our report provides some important findings and raises interesting questions. Further studies are needed to address these questions in different settings and populations with varying degrees of role functioning impairment and across other HRQOL domains. Methodological studies exploring the relationship between gains in measurement precision and practical differences in tool performance can inform decisions on when the use of computer adaptive tests is desirable.

Acknowledgments The project described was supported by Award Number K01AG028760 from the National Institute on Aging. Partial salary support for Dr. Anatchkova is provided by the National Institutes of Health grant 1U01HL105268-01.The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health.

References

1. Cella, D., Riley, W., Stone, A., 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.
2. Gandek, B., Sinclair, S. J., Jette, A. M., & Ware Jr., J. E. (2007). Development and initial psychometric evaluation of the participation measure for post-acute care (PM-PAC). *American Journal of Physical Medicine & Rehabilitation, 86*(0894-9115; 1), 57–71.
3. Haley, S. M., Gandek, B., Siebens, H., et al. (2008). Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. *Archives of Physical Medicine and Rehabilitation, 89*(2), 275–283.
4. Mulcahey, M. J., Haley, S. M., Duffy, T., Pengsheng, N., & Betz, R. R. (2008). Measuring physical functioning in children with spinal impairments with computerized adaptive testing. *Journal of Pediatric Orthopedics, 28*(0271-6798; 3), 330–335.
5. Wilkie, D. J., Judge, M. K., Berry, D. L., Dell, J., Zong, S., & Gilespie, R. (2003). Usability of a computerized PAINReportIt in the general public with pain and people with cancer pain. *Journal of Pain and Symptom Management*, *25*(0885-3924; 3), 213–224.
6. Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the headache impact test (HIT). *Quality of Life Research, 12*(8), 913–933.
7. Bayliss, M. S., Dewey, J. E., Dunlap, I., et al. (2003). A study of the feasibility of internet administration of a computerized health survey: The headache impact test (HIT). *Quality of Life Research, 12*(8), 953–961.
8. Cella, D., Yount, S., Rothrock, N., et al. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11.
9. Haley, S. M., Fragala-Pinkham, M., & Ni, P. (2006). Sensitivity of a computer adaptive assessment for measuring functional mobility changes in children enrolled in a community fitness programme. *Clinical Rehabilitation, 20*(7), 616–622.
10. Hart, D. L., Deutscher, D., Werneke, M. W., Holder, J., & Wang, Y. C. (2010). Implementing computerized adaptive tests in routine clinical practice: Experience implementing CATs. *Journal of Applied Measurement, 11*(3), 288–303.
11. Hart, D. L., Wang, Y. C., Cook, K. F., & Mioduski, J. E. (2010). A computerized adaptive test for patients with shoulder

impairments produced responsive measures of function. *Physical Therapy, 90*(6), 928–938.

12. Hart, D. L., Werneke, M. W., Wang, Y. C., Stratford, P. W., & Mioduski, J. E. (2010). Computerized adaptive test for patients with lumbar spine impairments produced valid and responsive measures of function. *Spine (Phila Pa 1976)*, *35*(24), 2157–2164.
13. Turner-Bowker, D. M., Saris-Baglama, R. N., Anatchkova, M., & Mosen, D. M. (2010). A computerized asthma outcomes measure is feasible for disease management. *The American Journal of Pharmacy Benefits, 2*(2), 119–124.
14. Anatchkova, M. D., Saris-Baglama, R. N., Kosinski, M., & Bjorner, J. B. (2009). Development and preliminary testing of a computerized adaptive assessment of chronic pain. *The Journal of Pain, 10*(9), 932–943.
15. Becker, J., Fliege, H., Kocalevent, R. D., et al. (2008). Functioning and validity of a computerized adaptive test to measure anxiety (A-CAT). *Depress Anxiety, 25*(12), E182–E194.
16. Kopec, J. A., Badii, M., McKenna, M., Lima, V. D., Sayre, E. C., & Dvorak, M. (2008). Computerized adaptive testing in back pain: Validation of the CAT-5D-QOL. *Spine (Phila Pa 1976)*, *33*(12), 1384–1390.
17. Kosinski, M., Bjorner, J. B., Ware, J. E., Jr., Sullivan, E., & Straus, W. L. (2006). An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. *Journal of Clinical Epidemiology, 59*(7), 715–723.
18. World Health Organization. (2002). *Towards a common language for functioning, disability and health: ICF: The international classification of functioning, disability and health*. Geneva: World Health Organization (WHO).
19. Anatchkova, M. D., & Bjorner, J. B. (2010). Health and role functioning: The use of focus groups in the development of an item bank. *Quality of Life Research, 19*(1), 111–123.
20. Anatchkova, M. D., Ware, J. E., & Bjorner, J. B. (2011). Assessing the factor structure of a role functioning item bank. *Quality of Life Research, 20*, 745–758.
21. Muraki, E. (1997). Generalized partial credit model. In V. D. Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 153–164). New York, NY: Springer.
22. Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research, 16*(Suppl 1), 95–108.
23. Anatchkova, M. D., & Bjorner, J. B. *Item calibration of a generic role functioning item bank*. ISPOR 13th Annual European Congress, Prague. http://www.ispor.org/research_study_digest/details.asp.
24. Ware, J. E., Jr., & Dewey, J. (2000). *How to score version two of the SF-36 health survey*. Lincoln, RI: QualityMetric Incorporated.
25. Kessler, R. C., Barber, C., Beck, A., et al. (2003). The world health organization health and work performance questionnaire (HPQ). *Journal of Occupational and Environmental Medicine, 45*(2), 156–174.
26. Lerner, D., Amick III, B. C., Rogers, W. H., Malspeis, S., Bungay, K., & Cynn, D. (2001). The work limitations questionnaire. *Medical Care*, *39*(0025-7079; 1), 72–85.
27. Ware, J. E., Jr., Kosinski, M., Turner-Bowker, D. M., & Gandek, B. (2002). *How to score version two of the SF-12 health survey*. Lincoln, RI: QualityMetric Incorporated.
28. Moriarty, D. G., Zack, M. M., & Kobau, R. (2003). The centers for disease control and prevention's healthy days measures: Population tracking of perceived physical and mental health over time. *Health and Quality of Life Outcomes, 1*(1), 37.
29. McHorney, C. A., Ware, J. E., Jr., & 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*, 247–263.