Published in final edited form as:

Arch Phys Med Rehabil. 2012 July ; 93(7): 1153–1160. doi:10.1016/j.apmr.2012.02.008.

Performance of an Item Response Theory-Based Computer Adaptive Test in Identifying Functional Decline

Andrea L. Cheville, MD, MSCE, Kathleen J. Yost, PhD, Dirk R. Larson, MS, Katiuska Dos Santos, MD, Megan M. O'Byrne, BA, Megan T. Chang, BA, Terry M. Therneau, PhD, Felix E. Diehn, MD, and Ping Yang, MD, PhD

Departments of Physical Medicine and Rehabilitation (Cheville, Dos Santos), Health Sciences Research (Yost, Larson, O'Byrne, Therneau, Yang), and Radiology (Diehn), Mayo Clinic, Rochester, MN; and Mayo Medical School, Rochester, MN (Chang)

Abstract

Objective—To achieve a low respondent burden and increase the responsiveness of functional measurement by using an item response theory-based computer adaptive test (CAT), the Activity Measure for Post-Acute Care (AM-PAC) CAT.

Design—Two-year prospective cohort study.

Setting—Telephonic assessments from a quaternary medical center.

Participants—Patients (N = 311) with late-stage lung cancer (LC).

Interventions—Monthly assessments for up to 2 years. Disease progression was determined via record abstraction. Anchor-based responsiveness techniques were used to compare AM-PAC-CAT score changes between global rating of change (GRC) question response levels, as well as between intervals when adverse clinical events or symptom worsening did and did not occur. Distribution-based responsiveness assessments included calculation of the standardized effect size (SES) and standardized response mean (SRM).

Main Outcome Measures—AM-PAC-CAT, symptom numerical rating scales, and a GRC.

Results—Administration time averaged 112 seconds over 2543 interviews. AM-PAC-CAT score changes became more positive as GRC responses reflected more improved states: a lot worse (−11.62), a little worse (−1.92), the same (−.10), a little better (1.01), and a lot better (2.82). Score changes were negative when associated with adverse clinical events. The SES and SRM for score differences between 1 to 2 and 9 to 10 months prior to death were −.87 and −1.13, respectively. The minimally important difference estimate was defined by the mean CAT session SE at 2.0.

Conclusions—The AM-PAC-CAT imposes a low, <2-minute, respondent burden, and distribution- and anchor-based methods suggest that is moderately responsive in patients with late-stage LC.

Keywords

Epidemiologic measurement; Mobility limitation; Neoplasms; Psychometrics; Rehabilitation

© 2012 by the American Congress of Rehabilitation Medicine

Reprint requests to Andrea L. Cheville, MD, MSCE, 200 First St SW, Rochester, MN 55905, Cheville.andrea@mayo.edu.

An audio podcast accompanies this article.

Listen at www.archives-pmr.org.

No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit on the authors or on any organization with which the authors are associated.

Item response theory (IRT)-based assessment may be a promising means to overcome our longstanding failure to effectively detect and introduce countermeasures to stabilize early disablement among patients with progressive disease.^{1,2} Clinically administered assessments are time consuming and expensive. Patient self-reported measures, on the other hand, are easier to administer but may suffer from limited or uncharacterized responsiveness.³ Responsiveness is key to either approach's success because sensitivity to change is a paramount requirement for effective longitudinal monitoring.^{4,5} Extended fixed-length instruments may offer enhanced responsiveness, but the associated respondent burden may render repeated assessments impractical.

IRT-based functional measures are theorized to offer the advantages of brevity and responsiveness over classical test theory-based, fixed-length instruments.⁶ IRT orders items along a hypothesized unidimensional latent continuum. This ordering creates the opportunity to administer different items to the same or different patients and generate comparable scores. The ordering also underlies a salient strength of IRT-modeled item banks, the ability to create computer adaptive tests (CATs) whose algorithms administer the most potentially informative items based on prior responses. CATs, due to their ability to restrict questioning to a limited number of discriminative items, offer the opportunity to be more responsive than fixed-length assessment tools and to reduce respondent burden.

Although the advantages of enhanced responsiveness have spurred IRT-based measurement initiatives including the National Institutes of Health Patient-Reported Outcomes Measurement Information System (PROMIS),⁶ these initiatives are relatively new and empirical support of the clinical benefits of IRT-based measurement have yet to be established. In contrast to the PROMIS physical function item bank, the Activity Measure for Post-Acute Care (AM-PAC) bank, and more specifically its related CAT, have been evaluated in clinical practice.⁷ However, neither the Activity measure for post acute care computer adaptive test (AM-PAC-CAT) nor any other physical function CAT has been evaluated where it may address a great clinical need—the detection of early functional decline among patients with progressive disease.

This study was designed to estimate the AM-PAC-CAT's responsiveness, respondent burden, and minimally important difference (MID) among patients with late-stage lung cancer (LC). Low detection rates of the almost universal disablement that afflicts patients with late-stage cancer make them a clinically relevant study population.^{1,2}

METHODS

Participants

A sample of 315 patients with stage IIIB or IV nonsmall cell LC or extensive stage small cell LC was targeted for enrollment between May 2008 and May 2009 from an eligible pool of participants identified through a previously described screening approach developed for the Mayo Clinic Epidemiology and Genetics of Lung Cancer Program.⁸ At least 5 attempts were made to contact all potentially eligible subjects by telephone on different days of the week and times of day. Those who provided verbal informed consent were consecutively enrolled regardless of their current treatment status. Subjects were required to be fluent in English, be able to converse on the telephone, and have an intact mental status (with a Folstein Mini-Mental State Examination score ≥ 25).

Proxy responses from caregivers were not permitted; however, participants were permitted to solicit assistance or corroboration from their caregivers. Participants were followed until

dropout, death, or the study's April 15, 2010 closure. The study was approved by the Mayo Clinic Institutional Review Board.

Data Collection Schedule

Data were collected via telephone by 1 of 4 research assistants or the principal investigator (A.L.C.). Attempts to contact a participant by telephone were reinitiated 25 days after the participant's previous contact and were continued daily at different times of the day for 2 weeks. Thereafter, interviews continued on alternate weeks until a participant's death was confirmed or they communicated a desire to drop out of the study. All telephone interviews followed a script that ordered patient-reported outcome (PRO) administration as follows: (1) AM-PAC-CAT, (2) symptom numerical rating scales, and (3) a global rating of change (GRC) question. Electronic medical record (EMR) abstraction continued throughout the study.

Patient-Reported Outcomes

AM-PAC-CAT physical and movement activities domain (hereafter referred to as the AM-PAC-CAT)—The AM-PAC-CAT is derived from a traditional fixed-length measure, the AM-PAC, which has demonstrated excellent reliability and validity.^{9–11} Comparable reliability and validity have been achieved administering the AM-PAC using a CAT platform.¹² The AM-PAC-CAT physical and movement activities domain utilized in this study was established through factor, modified parallel, and Rasch analysis.^{7,10}

The AM-PAC-CAT bank contains 101 items that query respondents regarding how much difficulty they experience performing physical activities, such as going up steps. Response options include none, a little, a lot, and unable. The AM-PAC-CAT questions and responses were read to participants over the phone. The response selected was entered into the computer by the interviewer, and the CAT algorithm selected the next question. Participants were given the ability to skip any item. Testing continued until an AM-PAC-CAT session SE fell below 2.0 or until a participant had answered 10 questions. The 2.0 SE threshold was based on Jette et al's⁷ report of 1.99 as the average SE among patients with AM-PAC-CAT basic mobility scores of 50 to 69, comparable with mean values among the study population. IRT-based scores (ie, theta) were transformed to T scores, which have a mean $\pm$ SD of 50 $\pm$ 10 in the reference population.

GRC question—A single question with 5 response categories (a lot worse, a little worse, the same, a little better, and a lot better) was used to assess participants' impression of how, relative to their previous evaluation, their ability to function had changed. GRCs that assess functional status have established validity in a range of chronic disease states, including cancer.^{5,13}

Symptom numerical rating scales—Considerable evidence demonstrates the value of single-item PRO assessments for describing symptoms,¹⁴ and these have been extensively validated in patients with cancer.¹⁵ Participants rated their pain and fatigue over the 7 days preceding each assessment point with 11-point numerical rating scales ranging from 0 (none) to 10 (as bad as it can be).

Electronic Medical Record

Two physicians abstracted data from the Mayo Clinic EMR. Data were initially abstracted by a single physician (K.D.S.) and then verified by a second physician (A.L.C.). Each participant's Charlson Index and medical comorbidities at study entry were established through review of medical and surgical history sections of all clinical notes generated in the 3 years preceding study entry, as well as review of all assigned *International Classification*

of Diseases–9th Revision codes. Comorbidities of potential relevance to participants' functional status were grouped into 5 categories: coronary artery disease, chronic obstructive pulmonary disease, stroke, neurologic/psychiatric disorder, and musculoskeletal disorder.

LC stage and date of initial diagnosis were obtained from medical, surgical, and radiation oncology notes and confirmed through previously described Epidemiology and Genetics of Lung Cancer Program procedures.^{16,17} New or progressive brain and bone metastases were identified on imaging study reports. Review of medical and radiation oncology, as well as orthopedic surgery notes, established whether bone metastases were symptomatic. Dates of disease progression were determined from medical oncology notes, and corroborated through imaging study reports. Uncertainties regarding imaging reports and clinician notes were resolved through image review by a radiologist (F.E.D.).

Ascertainment of Vital Status

Patients' vital status was verified through death certificates, the Mayo Clinic EMR, next-of-kin reports, the Mayo Clinic Tumor Registry, and the Social Security Death Index website. The study cohort's vital status was followed for 6 months after assessments stopped on May 31, 2010.

Statistical Analyses

Descriptive statistics—Means and SDs were calculated for continuous variables and proportions for binary variables to describe the study cohorts' characteristics at enrollment.

Assessment of responsiveness with anchor-based methods—Two anchor-based approaches were used to evaluate the AM-PAC-CAT's responsiveness. The first compared changes in AM-PAC-CAT scores with participants' GRC responses. Research suggests that intervals for assessing responsiveness based on a GRC should allow for meaningful change to occur but should minimize recall bias.¹⁸ We therefore only used responses collected between 3 and 6 weeks apart. The association between GRC responses and change in AM-PAC-CAT scores over assessment intervals was evaluated using a linear model in which the GRC question was incorporated as a class (ie, discrete) variable with 5 categories. A linear model was fit in a generalized linear models framework using generalized estimating equations (GEEs) to account for the within-subject correlation due to multiple responses per subject. The parameter estimates from this model correspond to the mean changes in the AM-PAC-CAT scores for each category of the GRC response. The association between the GRC and change in the AM-PAC-CAT score was estimated with a Spearman correlation coefficient. GRC responses were also used to estimate the areas under receiver operating characteristic curves (AUCs), as proposed by Deyo et al.¹⁹ Binary variables were created by dichotomizing the sample at each GRC response level. Logistic regression with GEEs was used to estimate the AUCs.

Methodologies to assess responsiveness have focused on positive change.^{20–22} Lacking a criterion standard to assess responsiveness to negative change, we used the occurrence of clinical events and symptom worsening as anchors, because both have been empirically linked to functional decline.^{23,24} There is precedent for this approach.^{20,25} We compared AM-PAC-CAT score changes when clinical events or symptom worsening did or did not occur during an assessment interval. We selected clinical events that could be reliably identified in the EMR: the development of brain and symptomatic bone metastases, and cancer progression. For the purpose of assessing the responsiveness of the AM-PAC-CAT scores, a window around their collection was used to determine the development of metastases. If a new diagnosis of brain or bone metastases was present in the 60 days before or after a measurement, then it was associated with that measurement. This approach was

based on the fact that metastases may have been present but not yet detected. We examined the impact of worsening pain and fatigue on AM-PAC-CAT scores because both symptoms have been associated with functional decline in cancer populations with symptoms expressed as change per month.²⁶ We examined the magnitude of AM-PAC-CAT score changes in the presence of these events, and whether an event was associated with an AM-PAC-CAT score decline of 2, 5, and 10 points, which were less than the 13-point change in AM-PAC-CAT scores established by Tao et al,²⁷ to reflect a large decline in functionality. Linear models were used for continuous outcomes and logistic regression for binary outcomes. GEEs were used to account for the within-subject correlation.

Calculation of distribution-based responsiveness statistics—We calculated 3 effective size statistics.²¹ These included the standardized effect size (SES), or Cohen d ,^{28,29} the standardized response mean (SRM),³⁰ and the responsiveness retrospective (RR) coefficient.³¹ The SES and SRM express mean change as functions of the baseline SD and the SD of change, respectively. The SES therefore helps to estimate responsiveness in a varied population, while the SRM helps to estimate responsiveness in a variably changing one. For each statistic, values of 0.2, 0.5, and 0.8 indicate small, moderate, and large changes, respectively,^{19,32} which reflect comparable levels of responsiveness. SES and SRM numerators were calculated by subtracting AM-PAC-CAT scores collected 1 to 2 months from those collected 9 to 10 months prior to death, the interval when AM-PAC-CAT scores declined most precipitously and consistently. SRMs were calculated for score changes associated with each GRC response level. The RR numerator requires the variance from a stable population.³¹ We used the scores collected >12 months prior to death, because graphic assessment revealed stability over this interval.

Minimally Important Differences

Anchors used to estimate MID for a PRO should have at least a moderate correlation (>0.3) with the PRO.^{33,34} We computed the correlation between AM-PAC-CAT change scores and the GRC question. Provided the correlation was 0.3, patients were categorized as a little worse or a little better based on their response to the GRC question. Mean AM-PAC-CAT change scores in the little worse/little better GRC categories are estimates of the MID. The MID estimates were then compared with the SE of the AM-PAC-CAT to confirm that the magnitude of the MID was larger than the measurement error detectable by the measure.

RESULTS

Participants

Three hundred and eleven (87%) of the 357 patients who were invited to participate enrolled. Their demographic, clinical, and cancer-specific characteristics are listed in table 1. A total of 2543 telephone interviews were completed, an average $\pm$ SD of 8 ± 5.8 per patient, range 1 to 22. AM-PAC-CAT sessions lasted 112 ± 81.9 seconds on average. Seventy-eight percent of interviews, excluding the initial interviews, occurred within 6 weeks of the previous interview. Within the 2 years of follow-up, 202 (65%) participants died, and 51 (16%) dropped out, 39 (13%) after completing only 1 AM-PAC-CAT session. Six months after the cessation of AM-PAC-CAT administration, at the time of data analysis, 20 additional participants had died. Forty-nine percent of participants were interviewed at least once within 2 months of death. Figure 1 illustrates the progressive decline in the mean AM-PAC-CAT scores over the last year of life among the 222 deceased participants, and it is similar to published function-time plots for patients with cancer.^{35,36}

Anchor-Based Assessment of Responsiveness

Responsiveness assessed through the GRC—Mean AM-PAC-CAT score changes increased in association with the GRC responses that reflected changes to better status. On average, the scores of participants who characterized their functional status as a lot worse decreased by 11.62, while those who characterized their functional status as a little worse declined by 1.92. The mean score change of participants who reported being the same was $-.10$. Average score changes of participants characterizing their function as a little better and a lot better were 1.01 and 2.82, respectively. Figure 2 illustrates a dot plot with all AM-PAC-CAT score changes for each response option and lists the total number of sessions and participants providing data for each response level. AM-PAC-CAT score changes were moderately correlated with participants' responses to the GRC ($\rho=.30$, $P<.001$).

Responsiveness estimates based on the AUCs were derived using participants' GRC responses to dichotomize the AM-PAC-CAT scores. AUCs estimated for GRC response cutoffs indicating a status change were as follows: a lot worse (.86), a little worse (.64), a little better (.63), and a lot better (.69).

AM-PAC-CAT score changes with clinical events and symptom changes—

Odds ratios associated with the presence of clinical events in the occurrence of a 2-, 5-, or 10-point decline in AM-PAC-CAT scores are listed in table 2. Worst pain, average pain, and fatigue, expressed as change per month, were significantly associated with AM-PAC-CAT score declines of 2, 5, and 10 points. The identification of new brain metastases was associated with 5- and 10-point score reductions, while the development of symptomatic bone metastases was associated with 2- and 5-point reductions. Associations between progression of disease and AM-PAC-CAT score declines did not achieve statistical significance.

A similar pattern was noted when mean AM-PAC-CAT reductions were compared between participants who did and did not experience clinical events or symptom worsening. As listed in table 3, with the exception of progression of disease, AM-PAC-CAT score changes were significantly larger and negative when an event occurred during the assessment interval.

Distribution-Based Assessment of Responsiveness

The SES for the difference in participants' AM-PAC-CAT scores collected 1 to 2 months and 9 to 10 months prior to death was $-.87$. The SRM for change in this time period was -1.13 .

SRMs for AM-PAC-CAT score changes grouped by participants' GRC responses were notably lower. SRMs associated with the GRC responses of a lot worse, a little worse, the same, a little better, and a lot better were $-.95$, $-.49$, $-.03$, $.30$, and $.58$, respectively. The RR coefficient for MIDs of 1.0, 1.5, and 2.0 was $.41$, $.62$, and $.83$, respectively.

MID Estimation

AM-PAC-CAT change scores were moderately correlated with participant responses to the GRC ($\rho = .30$, $P<.001$); therefore, it was appropriate to use the GRC to estimate the MID. The mean AM-PAC-CAT change score in the a little worse GRC category was 1.92 and the mean change score in the a little better category was 1.01. Thus, estimates of the MID for the AM-PAC-CAT were approximately 1 to 2 points on a T score scale.

Table 4 lists the percentile and end-range values of the individual SEs generated by the CAT algorithm. These cluster near 2, reflecting the fact that the CAT was programmed to administer questions until the session-specific SE fell below 2, or a total of 10 items had

been administered. Yost et al³⁴ advocate use of the SE to either bound or define the MID when, as in this case, it exceeds the estimated MID range derived by other means.

DISCUSSION

To the best of our knowledge, this is the first effort to characterize the responsiveness, MID, and respondent burden of an IRT-based CAT in evaluating function over time in a cancer population. We used anchor- and distribution-based methods, as has been recommended by other investigators,^{28,33,37} and found the AM-PAC-CAT to be moderately responsive with SRMs in the range of 0.4 to 0.6 among patients characterizing themselves as a little changed. Consistent declines in mean AM-PAC-CAT scores over intervals that spanned participants' development of brain or bone metastases, or worsening symptoms, lend further support for responsiveness. Also, participants whose AM-PAC-CAT scores dropped by 2, 5, or 10 were significantly more likely to have experienced these events. Use of the AM-PAC-CAT imposed a low respondent burden because session durations averaged less than 2 minutes.

Previous AM-PAC-CAT Responsiveness Estimates

This investigation differs from prior studies in that it strove to examine the potential utility of the AM-PAC-CAT as a screening tool to aid the clinical challenge of identifying functional loss among vulnerable patients. Jette et al⁷ previously evaluated the instrument's responsiveness; however, they focused on patients receiving outpatient physical therapy who generally improved, and they based their estimates on 2 time-points. We deliberately targeted patients who were likely to decline and repeatedly assessed them for up to 2 years. Despite these differences, Jette's SRM estimate, .68, also suggests moderate responsiveness.

Interpretation of Responsiveness Statistics

Our choice to estimate 3 separate effect size-based estimates was driven by differences of opinion as to which is least influenced by extraneous factors unrelated to responsiveness. For example, the SES and SRM standardize mean score changes by an expression of population variance and are therefore influenced by population heterogeneity.³⁸ The RR estimate, which has been cited as a potential superior responsiveness statistic,¹⁹ may be most informative regarding the AM-PAC-CAT's sensitivity to early functional losses, because it standardizes the MID, which is often smaller than mean differences in clinical trials, by variance in a clinically stable population. Our RR estimates, assuming MIDs of 1.0, 1.5, and 2.0, were .41, .62, and .83, respectively, and agreed with most of our other SRM estimates in suggesting moderate responsiveness for the AM-PAC-CAT.

Two SRM estimates fell well above the clustering between 0.4 and 0.6 that, in general, characterized our results. The first estimate, .95, was derived from the subgroup of patients who rated themselves as a lot worse. Prior investigators have advocated caution in the use of GRC-based responsiveness assessments, because they depend on patients' perception of change, which may be biased, particularly among participants selecting responses at scale extremes, and are more likely to reflect the current state.^{37,39} Mean AM-PAC-CAT score changes associated with the GRC responses of "a lot worse" and "a lot better" differed markedly, 11.62 versus 2.82, respectively. Cella et al⁴⁰ described unequal mean quality of life score changes between patients with cancer who selected negative (worsening) and positive (improving) GRC responses, and suggested that the greater magnitude of change required before patients described themselves as worsened may reflect denial. In our study, on average, patients did not characterize themselves as a lot worse until their AM-PAC-CAT score had decreased by roughly 12 points. Thus, the absolute magnitude of the SRM numerator, $D_X/SD(D_X)$, was markedly larger among this subgroup, and bias appears to have

inflated the SRM estimate. Further research is warranted to clarify whether the directionality of change is an important consideration in responsiveness assessments.

The second SRM estimate, -1.13 , derived by comparing AM-PAC-CAT scores collected 1 to 2 and 9 to 10 months prior to death, reflects the degree and consistency of participants' deterioration during this interval. Population characteristics drive SRMs such that homogenous, uniformly changing populations typically yield high SRMs by increasing the numerator (mean change) while reducing the denominator (SD of change), $D_x/SD(D_x)$.^{38,41} In contrast, heterogeneous, stable populations yield low SRMs. Many functional instruments have been evaluated in populations of the former type comprised of patients who will improve either because they (1) have just suffered a functionally catastrophic event (eg, stroke or hip fracture)⁴² or (2) have undergone treatments known to be effective (eg, knee arthroplasty).³⁰ Our comparison of AM-PAC-CAT scores at 1 to 2 and 9 to 10 months prior to death is analogous in that disablement was uniform and severe during this period. High SRMs (0.9 – 1.2) were estimated over a similar interval of marked decline among patients with amyotrophic lateral sclerosis, even though several of the functional measures had displayed only moderate responsiveness when used to assess mixed populations of changing and nonchanging patients.^{41,43,44} While the high SRM (-1.13) adds further support for the AM-PAC-CAT's responsiveness, it does not speak to our principal aim of characterizing the AM-PAC-CAT's responsiveness in a varied and variably changing population.

Our finding of AUCs ranging from $.63$ to $.86$ also suggests moderately good responsiveness, although it should be noted that Deyo and Inui,⁴⁵ who introduced the use of AUCs in responsiveness assessment, proposed that conventional parameters for AUC interpretation ($.50$ – $.75$ = fair, $.75$ – $.92$ = good) do not apply to responsiveness assessments, because external criteria like the GRC are not definitive “gold standards.”²⁰ As a result, the interpretation of AUCs derived with a single instrument from a single population remains subjective. Some context is afforded by AUCs reported for the Roland-Morris Disability Questionnaire, an instrument generally considered to have good responsiveness ($.67$ – $.81$).^{20,46} Based on these values, the AM-PAC-CAT would seem to have comparable responsiveness to the Roland scale, which has 24 items.

MID Estimation

The individual SEs estimated during each CAT session clustered at the upper bound of our anchor-based MID estimate of 2.0 . This reflects our a priori specification to cease CAT administration when the SE fell below 2.0 (75% of sessions) or after 10 items had been administered (25% of sessions). Reducing the SE stopping threshold and/or eliminating the 10-item cap would reduce the SEs and enhance resolution, but at the cost of an increased respondent burden. In addition, because the SE decrement achieved by administering each additional item steadily decreases, the number of items required to consistently reduce the SE below 1.5 may be substantial. Given these considerations, we suggest 2.0 as a pragmatic MID, for patients with late-stage cancer when the AM-PAC-CAT is administered with comparable CAT parameters. Yost et al,³⁴ in determining MIDs for the PROMIS short forms, similarly used the SE as the lower MID boundary whenever an SE exceeded the anchor-based MID estimates.

Methodologic Relevance

Our results demonstrate the feasibility of evaluating responsiveness under conditions that parallel a measure's intended application. Our intention was to examine the AM-PAC-CAT's sensitivity to small but meaningful functional loss among a population destined to become disabled in a variable manner. The longitudinal follow-up that is integral to the management of patients with late-stage cancer differs radically from the clinical situations in

which responsiveness is often assessed. Observational data gleaned during longitudinal follow-up may, therefore, provide opportunities to derive realistic responsiveness estimates that are more relevant to chronic disease management. Clinical events and symptoms, used as anchors in this study, are often routinely collected in clinical practice and complement GRC-based estimates. Additionally, our analytic approach using GEEs offers a means of adjusting for the inevitably correlated nature of data collected through longitudinal assessments. Prior investigators have generally handled this correlation by reporting separate responsiveness statistics for each assessment interval (eg, 0–3 and 0–6mo),^{43,47} or by using bootstrap or jackknife techniques.^{30,48} While GEEs are methodologically similar to the jackknife approach, they permit inclusion of unlimited assessment intervals, afford greater flexibility in specifying the correlation between repeated measures, and are more computationally efficient.

Study Limitations

These findings should be interpreted and generalized cautiously in light of persistent uncertainty regarding the interpretation of responsiveness statistics and the degree of responsiveness needed to justify a measure's clinical adoption.^{37,49} Whether our estimates will apply to other cancer and non-cancer populations remains uncertain. Additionally, participants characterized themselves as a lot worse/a lot better during only 6% of assessment intervals (n = 101) (see fig 2). Therefore, the mean AM-PAC-CAT score changes associated with these GRC responses and all statistics incorporating them may be unstable.

CONCLUSIONS

The weight of evidence collected using distribution- and anchor-based methods suggests that the AM-PAC-CAT is responsive to functional decline in patients with late-stage LC and imposes minimal respondent burden. The AM-PAC-CAT and similar IRT-based CATs offer a promising means of screening for early and progressive disability among vulnerable patients. The AM-PAC-CAT's MID in patients with late-stage LC is 2.0, as established by the median individual SE.

Acknowledgments

Supported by the National Institutes of Health (grant no. KL2 RR024151-01).

List of Abbreviations

AM-PAC	Activity Measure for Post-Acute Care
AM-PAC-CAT	Activity measure for post acute care computer adaptive test
AUC	area under receiver operating characteristic curve
CAT	computer adaptive test
EMR	electronic medical record
GEE	generalized estimating equation
GRC	global rating of change
IRT	item response theory
LC	lung cancer
MID	minimally important difference

PRO	patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System
RR	responsiveness retrospective
SES	standardized effect size
SRM	standardized response mean

References

1. Cheville AL, Beck LA, Petersen TL, Marks RS, Gamble GL. The detection and treatment of cancer-related functional problems in an outpatient setting. *Support Care Cancer*. 2009; 17:61–7. [PubMed: 18478275]
2. Cheville AL, Troxel AB, Basford JR, Kornblith AB. Prevalence and treatment patterns of physical impairments in patients with metastatic breast cancer. *J Clin Oncol*. 2008; 26:2621–9. [PubMed: 18509174]
3. Cheville AL, Basford JR, Troxel AB, Kornblith AB. Performance of common clinician- and self-report measures in assessing the function of community-dwelling people with metastatic breast cancer. *Arch Phys Med Rehabil*. 2009; 90:2116–24. [PubMed: 19969178]
4. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol*. 1989; 42:403–8. [PubMed: 2659745]
5. Eurich DT, Johnson JA, Reid KJ, Spertus JA. Assessing responsiveness of generic and specific health related quality of life measures in heart failure. *Health Qual Life Outcomes*. 2006; 4:89. [PubMed: 17125512]
6. Cella D, Riley W, Stone A, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol*. 2010; 63:1179–94. [PubMed: 20685078]
7. Jette AM, Haley SM, Tao W, et al. Prospective evaluation of the AM-PAC-CAT in outpatient rehabilitation settings. *Phys Ther*. 2007; 87:385–98. [PubMed: 17311888]
8. Li Y, Sheu CC, Ye Y, et al. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *Lancet Oncol*. 2010; 11:321–30. [PubMed: 20304703]
9. Coster WJ, Haley SM, Andres PL, Ludlow LH, Bond TL, Ni PS. Refining the conceptual basis for rehabilitation outcome measurement: personal care and instrumental activities domain. *Med Care*. 2004; 42(1 Suppl):I62–72. [PubMed: 14707756]
10. Haley SM, Coster WJ, Andres PL, et al. Activity outcome measurement for postacute care. *Med Care*. 2004; 42(1 Suppl):I49–61. [PubMed: 14707755]
11. Siebens H, Andres PL, Pengsheng N, Coster WJ, Haley SM. Measuring physical function in patients with complex medical and postsurgical conditions: a computer adaptive approach. *Am J Phys Med Rehabil*. 2005; 84:741–8. [PubMed: 16205429]
12. Haley SM, Fragala-Pinkham M, Ni P. Sensitivity of a computer adaptive assessment for measuring functional mobility changes in children enrolled in a community fitness programme. *Clin Rehabil*. 2006; 20:616–22. [PubMed: 16894805]
13. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. Transition questions to assess outcomes in rheumatoid arthritis. *Br J Rheumatol*. 1993; 32:807–11. [PubMed: 8369891]
14. Cleeland CS, Mendoza TR, Wang XS, et al. Assessing symptom distress in cancer patients: the M.D. Anderson Symptom Inventory. *Cancer*. 2000; 89:1634–46. [PubMed: 11013380]
15. Buchanan DR, O'Mara AM, Kelaghan JW, Minasian LM. Quality-of-life assessment in the symptom management trials of the National Cancer Institute-supported Community Clinical Oncology Program. *J Clin Oncol*. 2005; 23:591–8. [PubMed: 15659506]
16. Sugimura H, Nichols FC, Yang P, et al. Survival after recurrent nonsmall-cell lung cancer after complete pulmonary resection. *Ann Thorac Surg*. 2007; 83:409–17. discussion 417–8. [PubMed: 17257962]

17. Visbal AL, Williams BA, Nichols FC 3rd, et al. Gender differences in non-small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002. *Ann Thorac Surg.* 2004; 78:209–15. discussion 215. [PubMed: 15223430]
18. Turner D, Schunemann HJ, Griffith LE, et al. Using the entire cohort in the receiver operating characteristic analysis maximizes precision of the minimal important difference. *J Clin Epidemiol.* 2009; 62:374–9. [PubMed: 19013766]
19. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Control Clin Trials.* 1991; (4 Suppl):142S–58S. [PubMed: 1663851]
20. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chronic Dis.* 1986; 39:897–906. [PubMed: 2947907]
21. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol.* 2000; 53:459–68. [PubMed: 10812317]
22. Tuley MR, Mulrow CD, McMahan CA. Estimating and testing an index of responsiveness and the relationship of the index to power. *J Clin Epidemiol.* 1991; 44:417–21. [PubMed: 2010785]
23. Williamson GM, Schulz R. Activity restriction mediates the association between pain and depressed affect: a study of younger and older adult cancer patients. *Psychol Aging.* 1995; 10:369–78. [PubMed: 8527058]
24. Given B, Given C, Azzouz F, Stommel M. Physical functioning of elderly cancer patients prior to diagnosis and following initial treatment. *Nurs Res.* 2001; 50:222–32. [PubMed: 11480531]
25. Meenan RF, Anderson JJ, Kazis LE, et al. Outcome assessment in clinical trials. Evidence for the sensitivity of a health status measure. *Arthritis Rheum.* 1984; 27:1344–52. [PubMed: 6391495]
26. Given CW, Given BA, Stommel M. The impact of age, treatment, and symptoms on the physical and mental health of cancer patients. A longitudinal perspective. *Cancer.* 1994; 74:2128–38. [PubMed: 8087780]
27. Tao W, Haley SM, Coster WJ, Ni P, Jette AM. An exploratory analysis of functional staging using an item response theory approach. *Arch Phys Med Rehabil.* 2008; 89:1046–53. [PubMed: 18503798]
28. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. *J Clin Epidemiol.* 1997; 50:79–93. [PubMed: 9048693]
29. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A comparison of the sensitivity to change of several health status instruments in rheumatoid arthritis. *J Rheumatol.* 1993; 20:429–36. [PubMed: 8478847]
30. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. *Med Care.* 1990; 28:632–42. [PubMed: 2366602]
31. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis.* 1987; 40:171–8. [PubMed: 3818871]
32. Cohen, J. Statistical power analysis for the behavioral sciences. Hillsdale: Lawrence Erlbaum Associates; 1988.
33. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol.* 2008; 61:102–9. [PubMed: 18177782]
34. Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. *J Clin Epidemiol.* 2011; 64:507–16. [PubMed: 21447427]
35. Lunney JR, Lynn J, Foley DJ, Lipson S, Guralnik JM. Patterns of functional decline at the end of life. *JAMA.* 2003; 289:2387–92. [PubMed: 12746362]
36. Costantini M, Beccaro M, Higginson IJ. Cancer trajectories at the end of life: is there an effect of age and gender? *BMC Cancer.* 2008; 8:127. [PubMed: 18454854]
37. Liang MH. Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Med Care.* 2000; 38(9 Suppl):II84–90. [PubMed: 10982093]

38. Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol.* 1997; 50:869–79. [PubMed: 9291871]
39. Guyatt GH, Norman GR, Juniper EF, Griffith LE. A critical look at transition ratings. *J Clin Epidemiol.* 2002; 55:900–8. [PubMed: 12393078]
40. Cella D, Hahn EA, Dineen K. Meaningful change in cancer-specific quality of life scores: differences between improvement and worsening. *Qual Life Res.* 2002; 11:207–21. [PubMed: 12074259]
41. Wallace D, Duncan PW, Lai SM. Comparison of the responsiveness of the Barthel Index and the motor component of the Functional Independence Measure in stroke: the impact of using different methods for measuring responsiveness. *J Clin Epidemiol.* 2002; 55:922–8. [PubMed: 12393081]
42. Latham NK, Mehta V, Nguyen AM, et al. Performance-based or self-report measures of physical function: which should be used in clinical trials of hip fracture patients? *Arch Phys Med Rehabil.* 2008; 89:2146–55. [PubMed: 18996244]
43. De Groot IJ, Post MW, Van Heuveln T, Van Den Berg LH, Lindeman E. Measurement of decline of functioning in persons with amyotrophic lateral sclerosis: responsiveness and possible applications of the Functional Independence Measure, Barthel Index, Rehabilitation Activities Profile and Frenchay Activities Index. *Amyotroph Lateral Scler.* 2006; 7:167–72. [PubMed: 16963406]
44. Frihagen F, Grotle M, Madsen JE, Wyller TB, Mowinckel P, Nordsletten L. Outcome after femoral neck fractures: a comparison of Harris Hip Score, Eq-5d and Barthel Index. *Injury.* 2008; 39:1147–56. [PubMed: 18656868]
45. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. *Health Serv Res.* 1984; 19:275–89. [PubMed: 6746293]
46. Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. *Med Care.* 2010; 48:1007–14. [PubMed: 20856144]
47. Amjadi SS, Maranian PM, Paulus HE, et al. Validating and assessing the sensitivity of the Health Assessment Questionnaire-Disability Index-derived Short Form-6D in patients with early aggressive rheumatoid arthritis. *J Rheumatol.* 2009; 36:1150–7. [PubMed: 19369459]
48. Chang E, Abrahamowicz M, Ferland D, Fortin PR. Comparison of the responsiveness of lupus disease activity measures to changes in systemic lupus erythematosus activity relevant to patients and physicians. *J Clin Epidemiol.* 2002; 55:488–97. [PubMed: 12007552]
49. King MT. A point of minimal important difference (MID): a critique of terminology and methods. *Expert Rev Pharmacoecon Outcomes Res.* 2011; 11:171–84. [PubMed: 21476819]

Fig 1. Mean AM-PAC-CAT scores as a function of time in months among deceased participants during the last year of life.

Fig 2.
Dot plot of AM-PAC-CAT scores associated with different GRC responses.

Table 1**Study Cohort Demographic, Clinical, and Cancer-Specific Characteristics**

Characteristics	Total N = 311
Demographics and comorbidities	
Age	65.4 ± 10.9
Sex, female	153 (49.2)
Charlson Index at study entry	8.7 ± 2.5
Neurologic or axis I psychiatric disorder	22 (7.1)
Musculoskeletal disorder	109 (35.1)
CAD	69 (22.2)
COPD	101 (32.5)
Primary caregiver	
Spouse	238 (76.5)
Child	39 (12.5)
Parent	7 (2.3)
Sibling	8 (2.6)
Friend	1 (0.3)
Other	18 (5.8)
Cancer characteristics	
Stage at enrollment	
IIIB	40 (12.9)
IV	238 (76.5)
Extensive stage SCLC	33 (10.6)
Brain metastases	116 (58.5)
Bone metastases—symptomatic	103 (33.1)
Documented POD	193 (62.1)
Number of radiation treatments during study interval	1.7 ± 1.6
Symptom characteristics	
Pain average	2.0 ± 2.2
Pain worst	3.2 ± 3.3
Fatigue	4.4 ± 2.7
Dyspnea	3.3 ± 2.9

NOTE. Values are mean ± SD or n (%).

Abbreviations: CAD, coronary artery disease; COPD, chronic obstructive pulmonary disease; POD, progression of disease; SCLC, small cell lung cancer.

Table 2

Odds Ratios for Presence of Symptom Worsening and Adverse Clinical Events When AM-PAC-CAT Scores Fell by 2, 5, or 10 Points During an Assessment Interval

Covariate	2-Point Decline (n = 459)			5-Point Decline (n = 199)			10-Point Decline (n = 64)		
	OR	95% CI	P	OR	95% CI	P	OR	95% CI	P
Previous AM-PAC-CAT score	1.05	1.04–1.07	<.001	1.06	1.04–1.08	<.001	1.03	0.99–1.08	.147
Pain change/month	1.12	1.06–1.19	<.001	1.18	1.08–1.28	<.001	1.20	1.03–1.41	.020
Fatigue change/month	1.16	1.11–1.21	<.001	1.20	1.13–1.28	<.001	1.37	1.25–1.51	<.001
New or progressive symptomatic bone metastasis	1.38	1.03–1.84	.029	1.65	1.11–2.46	.013	1.72	0.82–3.60	.151
New or progressive brain metastasis	1.28	0.91–1.80	.163	1.77	1.16–2.70	.008	2.15	1.11–4.15	.023
Progression of disease	1.20	0.99–1.46	.058	1.30	0.97–1.73	.075	0.99	0.56–1.77	.986

Abbreviations: CI, confidence interval; OR, odds ratio.

Table 3

Mean AM-PAC-CAT Score Changes Associated With the Occurrence of Symptom Worsening or Adverse Clinical Events During an Assessment Interval

Covariate	Estimate	SE	P
Previous AM-PAC-CAT score	−0.09	0.02	<.001
Pain change/month	−0.32	0.09	<.001
Fatigue change/month	−0.45	0.07	<.001
New or progressive symptomatic bone metastasis	−1.05	0.42	.013
New or progressive brain metastasis	−1.18	0.54	.031
Progression of disease	−0.40	0.22	.067

Table 4

Distribution of CAT Session Specific SE Estimates

Quantile (%)		SE Estimate
100	Maximum	13.27
99		10.52
95		3.57
90		2.48
75	3rd quartile	2.01
50	Median	1.91
25	1st quartile	1.83
10		1.80
5		1.78
1		1.71
0	Minimum	1.65