Published in final edited form as:
Qual Life Res. 2010 February ; 19(1): 125–136. doi:10.1007/s11136-009-9560-5.

Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms

Seung W. Choi,
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 710 N. Lake Shore Dr, Chicago, IL 60611, USA

Steven P. Reise,
Department of Psychology, University of California, Los Angeles, Los Angeles, CA, USA

Paul A. Pilkonis,
Department of Psychiatry, University of Pittsburgh Medical Center, Pittsburgh, PA, USA

Ron D. Hays, and
Department of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; Health Program, RAND, Santa Monica, CA, USA

David Cella
Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, 710 N. Lake Shore Dr, Chicago, IL 60611, USA

Abstract

Purpose—Short-form patient-reported outcome measures are popular because they minimize patient burden. We assessed the efficiency of static short forms and computer adaptive testing (CAT) using data from the Patient-Reported Outcomes Measurement Information System (PROMIS) project.

Methods—We evaluated the 28-item PROMIS depressive symptoms bank. We used post hoc simulations based on the PROMIS calibration sample to compare several short-form selection strategies and the PROMIS CAT to the total item bank score.

Results—Compared with full-bank scores, all short forms and CAT produced highly correlated scores, but CAT outperformed each static short form in almost all criteria. However, short-form selection strategies performed only marginally worse than CAT. The performance gap observed in static forms was reduced by using a two-stage branching test format.

Conclusions—Using several polytomous items in a calibrated unidimensional bank to measure depressive symptoms yielded a CAT that provided marginally superior efficiency compared to static short forms. The efficiency of a two-stage semi-adaptive testing strategy was so close to CAT that it warrants further consideration and study.

Keywords

Computer adaptive testing; PROMIS; Item response theory; Short form; Two-stage testing

s-choi@northwestern.edu.

Introduction

Recent advances in item response theory (IRT) modeling and computing capabilities have led to an increased interest in item banking and associated computer adaptive testing (CAT) to measure patient-reported outcomes [1,2]. Applications of item banking and CAT are also increasing as a result of organized efforts such as the Patient-Reported Outcomes Measurement Information System (PROMIS) project funded by National Institute of Health (NIH) Roadmap Initiative [3]. One of the objectives of the initiative is to create practical tools that are capable of producing multiple short-form measures and CAT administrations on a common measurement metric so that studies using such tools can share and accumulate findings.

It is now possible to derive IRT-calibrated item banks that include large numbers of questions that individually and collectively represent a well-defined, unidimensional measure. Individual questions from these large banks can then be extracted, using various strategies, to create unique short forms of that measure. These short forms can be static, or they can be constructed adaptively in real time based on the respondent's answers to previous questions. Because only a subset of questions from a bank is needed to precisely estimate a person's score, these short forms (static and dynamic) offer more efficient measurement than a full-length measure. Because CAT selects items that maximize information about the person's likely score, it will generally require fewer items than a static short form to attain comparable precision. However, the increase in efficiency varies across item banks, and in some cases may not be sufficient to justify the added technical requirements for CAT administration. Furthermore, with multiple response option (i.e., polytomous) items, there are ways to customize static short forms to enhance their efficiency. For example, one can employ a two-stage semi-adaptive "branching" strategy where a common question is asked first, followed by a different set of questions based on the response to the first question [4]. Although such a strategy could be used with dichotomously scored items, it would be more complex to implement, because several items will be needed in the first stage (to make more than a binary decision) along with scoring and applying cut scores to make decisions for the second stage. In this paper, we compare various kinds of static short forms, including branching, and compare them to CAT in terms of their efficiency in recovering the full-bank score. For illustration purposes, we used the PROMIS Wave 1 depressive symptoms bank [5]. The efficiency of CAT in relation to full-length measures has been studied previously in the context of assessing depressive symptoms [6-8]. However, little is known to date with respect to the relative efficiency of static and dynamic short forms of comparable lengths in measuring depression.

This study was approved by the Institutional Review Board from Northwestern University (STU00013246).

Methods

PROMIS depression bank

The PROMIS Wave 1 depression bank consists of 28 polytomous items with five response categories (*Never*, *Rarely*, *Sometimes*, *Often*, and *Always*). Larger positive scores on the scale represent higher levels of depressive symptoms. Data from a total of 14,839 examinees were used for IRT calibration using the graded response model [9]. The calibration data included a general population sample of 782 full-bank testing cases taking the entire bank and 14,057 block-administration cases (6,213 general population respondents and 7,844 clinical respondents, of which 4,133 indicated that they had been told by a doctor or a health professional that they had depression) representing 14 domains with 7 items each. Blocks were created to include 3–4 common items shared between adjacent blocks for linking purposes. We conducted concurrent calibration on the entire sparse-matrix data encompassing both the full-bank and block-administration cases using MULTILOG [10].

The S-X^2 model fit statistics [11,12] were examined (based on the full-bank subset only) using the IRTFIT macro program [13]. All 28 items had adequate or better model fit statistics ($p > 0.05$). This preliminary scale was then re-centered (mean $\eta = 0.0$ and SD = 1.0) based on a scale setting subsample ($n = 3,060$) matching the marginal distributions of gender, age, race, and education in the 2000 US census [14]. The final bank item parameter estimates are presented in Appendix.

The scale represented by the 28 items was essentially unidimensional. A categorical confirmatory factor analysis (CFA) was conducted using Mplus [15] on a subset of 768 cases without missing responses. A single factor solution (based on polychoric correlations and the WLSMV estimator) produced a satisfactory model fit (CFI = 0.938, TLI = 0.995, RMSEA = 0.087, and SRMR = 0.032). The average absolute residual correlation was 0.026 (SD = 0.020). All absolute residual correlations but one ($r = 0.107$) were substantially below 0.10. The internal consistency reliability of the 28-item scale was .980.

The scale (bank) information function (see the outmost curve in Fig. 1) was shifted notably to the right (+1.5) in relation to the trait distribution (centered on 0.0). That is, there was an abundance of items covering moderate to severe levels of depression, yet very few extending to the lower (non-depressed) end of the continuum. The lowest category threshold value (i.e., the transition point from *Never to Rarely* and above) of the least symptomatic item in the bank ("I felt unhappy") was approximately −0.5. This implies that the measurement precision at the healthier end of the continuum (beyond 0.5 SD "less depressed" than the average person in the general population) is relatively poor.

First generation PROMIS CAT engine

We used the first generation PROMIS CAT engine [16] in this study and hence describe its components in some detail below. CAT selects items typically one at a time aiming to minimize the standard error by either maximizing Fisher's information (MFI) at the current trait estimate [17,18] or minimizing the posterior variance thereof [7-9]. Because trait estimates can be unstable at the early stages of CAT, the item selection based on interim theta estimates can be suboptimal (especially for short CATs) [19,20]. Several researchers have proposed methods to compensate for unstable interim theta estimates and to minimize the likelihood of optimizing at wrong places on the trait continuum when selecting items [19,21,22]. Veerkamp and Berger [22] demonstrated that selecting items based on the information function weighted by the likelihood of trait values, known as the maximum likelihood weighted information (MLWI), is superior to the traditional MFI criterion. van der Linden and Pashley [21] also examined a similar approach, the maximum posterior weighted information (MPWI), which used the posterior distribution in lieu of the likelihood distribution. Another approach utilizes a theta estimator with minimum mean square error [23], the expected a posteriori (EAP) estimator, instead of the maximum likelihood estimator (MLE).

Veerkamp and Berger [22] demonstrated that the MFI evaluated at the EAP estimate (instead of MLE) could realize a gain in efficiency similar to that found using MLWI. Choi and Swartz [24] demonstrated that MPWI in conjunction with EAP provides excellent measurement precision for polytomous item CAT, comparable to other computationally intensive Bayesian selection criteria. Therefore, the first generation PROMIS CAT engine employs the EAP theta estimator, the MPWI item selection criterion, and the standard deviation of the posterior distribution (PSD) as the standard error (SE) estimator.

Short-form creation strategies

Optimal assembly of tests from item banks is typically a multifaceted process (often with competing objectives) and can require advanced computational algorithms [25]. Focusing on

the efficiency and comparability with the full-bank score, we employed five measurement-centered (i.e., using measurement precision as the primary objective) selection strategies to create static short forms: (1) Expected information using normal weighting (“*E(Info) Normal*”); (2) Expected information using logistic weighting (“*E(Info) Logistic*”); (3) Expected information using uniform weighting (“*E(Info) Uniform*”); (4) Maximum correlation with full-length measures (“*Max R*”); and (5) Items with lowest average rank order in CAT administration (“*Min Avg CAT Rank*”).

The *E(Info) Normal* approach chooses items that maximize the “expected” information over a normal distribution. The information function of each item was weighted by a normal distribution (the assumed trait distribution) and integrated over theta. This approach places heavy weighting on the center of a target distribution. The rationale for this strategy is to maximize the overall measurement precision by focusing on the region where the majority of examinees are expected to be located.

The *E(Info) Logistic* approach uses the same idea but with less emphasis on the middle of the target distribution. The logistic density distribution is slightly flattened compared to the normal counterpart, thus providing more weight to both ends. Taking this to the extreme, we also employed an *E(Info) Uniform* weighting method which allows the item bank to determine the most informative items regardless of their location. This criterion may work well when the trait distribution and the scale information function are closely aligned.

The *Max R* selection criterion is devised using the full-bank score as the criterion, analogous to a stepwise selection procedure in the multiple regression analysis. We selected the first item that produced a scale score distribution with the maximum correlation with the full-bank scores. The next item to be selected is the one in conjunction with the first item that produces a scale score distribution with the maximum correlation. Because correlation is scale independent, this strategy may not guarantee that the two score distributions will have comparable dispersions; therefore the short-form scores can be highly correlated with the full-bank scores and yet have a different standard deviation [26]. To account for this possibility, we examined a “scale-dependent” variant of *Max R* that minimizes the root mean squared difference (RMSD) between the full-bank and short-form scores.

Selecting items for short forms can be guided by the item selection behavior of CAT [27]. This method attempts to mimic CAT by selecting the most frequently administered items (i.e., administered in early stages). This can be implemented by administering the entire item pool to a sample of simulated (or real) examinees drawn from a target population. Thus, every simulee takes the entire bank but in a potentially different order. For each item, the rank in which the item was administered is averaged over all simulees. We then select a set of items with the minimum average CAT presentation rank (*Min Avg CAT Rank*).

Finally, for benchmarking purposes, we also used two naive item-selection criteria as the potential lower and upper bounds, respectively: (a) a method selecting items at random, and (b) a hypothetical CAT with all items selected to maximize information at the respondent’s full-bank score (“*CAT-Full Theta*”). The random selection method was replicated over a large number of iterations (1,000) to derive empirical distributions. The hypothetical CAT was analogous to selecting optimal items in retrospect knowing the full-bank score.

Two-stage branching

We explored a simple two-stage branching technique as another benchmark criterion. We selected the first (“locator”) item using the same procedure as the PROMIS CAT (“*CAT-MPWI*”), maximizing the expected information over the prior distribution. We then assessed the posterior distribution corresponding to each of the five response options on the first item.

Finally, for each posterior distribution, we selected a fixed sequence of 27 items maximizing the expected information over the given posterior distribution. Functionally, this is equivalent to the *CAT-MPWI* procedure becoming static after the first item.

Post hoc simulations

We rank ordered the 28 items in the bank according to the six short-form creation strategies described previously: E(Info) Normal, E(Info) Logistic, E(Info) Uniform, Max R, Min RMSD, and *Min Avg CAT Rank*. The first three criteria require only the item parameter estimates, whereas the latter three also involve simulated (or real) item responses. For the latter three, we drew a random sample of 10,000 simulees from the unit normal distribution and simulated item responses to all 28 items using the bank item parameters. It is worth noting that simulated responses were used for form creation, so the real data (i.e., the full-bank cases) can be used later for evaluation purposes without capitalizing on chance. For the *Min Avg CAT Rank* method, we adaptively administered the entire bank to the simulees using the Firestar CAT simulator [16] with the EAP theta estimator using a unit normal prior and the MPWI item selection criterion. We then computed the average CAT administration rank for each item. Based on each of the six different orderings of items, we then constructed 28 fixed-length forms consisted of $i = 1, 2,\ldots, 28$ items, respectively. We examined the concordance of the rank ordering using Kendall's *W* coefficient [28].

The static forms and CAT were then *administered* in post hoc simulations to the full-bank respondents without missing responses ($n = 768$). After administering each item, the interim theta and SE estimates were recorded as final estimates for the static forms and CAT with the given number of items. We examined all possible test lengths, because the purpose here was not to create specific fixed-length short forms, but to determine likely efficiencies of short forms of varying lengths, created using different methods.

Evaluation criteria

Using the full-bank score as the reference, we have chosen the following evaluation criteria: (1) the correlation and RMSD with the full-bank scores, (2) the mean PSD, (3) the extent to which the mean and SD of the full-bank scores are recovered, (4) the percentage of extreme responses (e.g., *Never* on all items administered), and (5) the classification consistency (at or beyond a threshold indicative of a high level of depressive symptoms). For classification consistency, we used an arbitrary threshold (cutoff) point of $\eta = 1.0$ or one standard deviation unit above the mean of the PROMIS general population sample. The classification consistency was estimated initially using Cohen's kappa [29]. As an alternative measure, we also computed Yule's coefficient of colligation, *Y* [30], which is considered independent of the marginal distributions [31]. We then examined the sensitivity, specificity, positive predictive value, and negative predictive value [32] of the classification results.

Results

Comparison of short forms vs. CAT

Table 1 lists the 28 bank items from low to high levels of depressive symptoms—from "I felt unhappy" to "I felt I had no reason for living." The numbers under the first five columns are the orders in which items were selected by different short-form selection strategies (Kendall's $W = .712$). The numbers in bold indicate the top eight items. The *Min RMSD* criterion produced the same ordering of items as the *Max R* and hence was not presented. The item orders from the first three expected information criteria were similar to each other ($W = .972$). The *Max R* criterion selected items in a similar order to the *Min Avg CAT Rank* criterion for the first several items but deviated notably for the rest of the items ($W = .773$). The two criteria (that utilized simulated responses for form creation) selected most of their top items from the less

depressive end of the scale conforming to the simulee distribution (cf. $\eta = 0.0$ in Fig. 1). All five criteria selected the same item as the first item ("I felt depressed."), which was also the locator item selected under the two-stage branching strategy (see Table 1 for the items comprising the complementary short forms in the second stage).

Figure 2 shows the correlations between scale scores from the CAT and short forms of various lengths and the full-bank score based on all 28 items. The inner rectangular box in Fig. 2 highlights the comparison for the test lengths of 4–12 items, possible minimum and maximum test lengths for CAT and short forms. The PROMIS CAT (*CAT-MPWI*) yielded correlations as high as the hypothetical benchmark CAT (*CAT-Full Theta*), indicating that the *CAT-MPWI* performed as well as one can expect for the given item bank and the trait distribution. The correlations with the full-bank score were slightly less than 0.96 for the two CAT procedures with only five items. The static forms by the *Max R* and *Min Avg CAT Rank* criteria performed well and would require six items to attain the same correlation. The three expected information short-form selection procedures did not perform as well as the previous four procedures and would require at least eight items to attain the same level of correlation. A similar pattern of performances emerged when the RMSD was used as a criterion measure (data not presented).

Figure 3 shows the mean posterior standard deviation (PSD) for test lengths 4–12. The *CAT-Full* Theta and *CAT-MPWI* performed the best followed by the *Max R* and *Min Avg CAT Rank* criteria. With as few as five items, the CAT attained a mean PSD of a little over 0.30. The *Max R* and *Min Avg CAT Rank* criteria would require one additional item to reach the same level of mean PSD. The three expected information selection procedures would require at least eight items to attain the same level of measurement precision.

For more detailed comparisons, we focused on the test length of eight items—that is, the length of the PROMIS wave 1 depressive symptoms short form. Table 2 summarizes the comparison of various CAT and short forms of eight items. The PROMIS depressive symptoms short form happened to have the same set of eight items as the two expected information criteria (Normal and Logistic) and hence produced the same results in Table 2. The *CAT-MPWI* yielded the highest correlation (.977), trailing the hypothetical benchmark CAT (*CAT-Full Theta*, .978) barely, followed closely by the *Two-Stage* (.975), *Min Avg CAT Rank* (.971), *Max R* (.967), and the three expected information criteria (ranged from .954 to .958). The rankings remained unchanged when we considered the RMSD and mean PSD—the *CAT-Full Theta* and *CAT-MPWI* ranked at the top and the three expected information criteria at the bottom.

A somewhat different pattern of performance emerged when it came to the classification consistency estimated by Cohen's kappa and Yule's *Y* statistics. Again, we used an arbitrary theta value of 1.0 as the cut point (for illustration purposes and not based on standard setting) and measured the extent of agreement in classifications with the full-bank score. Although the pattern is somewhat unclear, the *E(Info) Normal* and *E(Info) Logistic* criteria performed somewhat better (Cohen's $\kappa = .925$ and Yule's $Y = .949$ for both) than the others (κ ranged from .855 to .901; Y ranged from .899 to .931). This result is reflective of the distribution of item information functions and the specific location of the cut score considered. The two expected information criteria would provide maximal information between the center of the bank information function ($\eta = 1.5$) and the peak of the weight functions ($\eta = 0.0$). Contrastingly, the *Max R* criterion ($\kappa = .857$; $Y = .899$) and the CAT-based methods, the *CAT-MPWI* ($\kappa = .855$; $Y = .900$), *CAT-Full Theta* ($\kappa = .862$; $Y = .903$), and *Min Avg CAT Rank* ($\kappa = .893$; $Y = .927$), would be optimized for the distribution of individual scores rather than for the bank information function.

A close inspection of the sensitivity, specificity, positive predicted value, and negative predicted value also showed that the relatively low kappa values were mostly due to larger false-negatives. This is evident in that the differences among the methods were the most salient in the sensitivity—that is, true-positives over the sum of true-positives and false-negatives. The classification consistency results observed here are not to be generalized beyond the specific item bank and the cut score location considered. An analysis using a different cut score (0.0) revealed that the differences in kappa and *Y* almost vanished (κ ranged from .868 to .891; *Y* ranged from .869 to .898), and the *E(Info) Normal* and *E(Info) Logistic* criteria were no longer the top performers.

As for the percentage of cases endorsing the lowest response category (*Never*) on all of the items administered, the two CAT procedures and the *Max R* criterion were the most effective and performed similarly in minimizing the percentage of floor responses. On the other hand, the three expected information criteria did not perform as well in this regard. With the test length of eight items, the two CAT procedures, *Two-Stage*, and the *Max R* criterion yielded about 14% floor responses, whereas the expected information procedures produced about 24%. Combining this with the results on classification consistency, it appears that the three expected information procedures are concentrating measurement on the upper region where the bank is rich in information and not providing adequate coverage at the lower end where most people are located.

With only a single round of adaptation, the two-stage branching technique came very close to the fully adaptive CAT. This procedure exhibited further improvements over the *Max R* and the *Min Avg CAT Rank* procedures in terms of the correlation, RMSD, mean PSD, and percentage of floor responses. Inspecting the composition of items in the complementary short forms in the second stage, however, suggests that fewer branches might have been adequate. Although the items were selected in different orders, the seven-item complementary short forms overlapped considerably with an exception of the branch stemming from the lowest response category (*Never*). It appears that the performance gain for the two-stage branching strategy over other static forms has been achieved mostly through adapting for those without depressive symptoms. This also demonstrated that the locator item ("I felt depressed") was a very efficient indicator of how examinees would respond to the entire bank. As displayed in the item response functions (Fig. 4), this item provides good measurement for a wide range of theta values (roughly from −0.5 to 2.5), which is almost as large as the effective measurement range of the entire bank (cf. Fig. 1). The correlation between the locator item and the full-bank score was 0.827 (or 0.807 with the item removed from the full-bank score). About 68% (or 65% with correction for overlap) of the total variance in the full-bank scores was accounted for by the responses on the locator item.

This also reveals that the PROMIS depressive symptom bank demonstrates construct homogeneity [33]. That is, the scores derived from the bank reflect variation on a single well-defined measure. To further elaborate on this point, we randomly selected a set of eight items 1,000 times from the bank and generated the evaluation criteria (see Table 2). Evidently, some (extreme) random tests did very well on the evaluation criteria. Over 1,000 replications, the correlations with the full-bank ranged from 0.934 to 0.972, with a median of 0.959. Although those correlations might have been somewhat spurious because they were part-whole correlations (i.e., items comprising the short forms also contributed to the full-bank score), the magnitude and range of correlations indicate that the items are essentially interchangeable. Thus, if a clinician selects questions with most relevant content to a target population without relying on statistical information, the scores from such measures are expected to be highly correlated with the full-bank score.

Discussion

Our focus in this study was on examining the efficiency of measurement-centered CAT and short-form selection strategies. However, the process of constructing short forms is more holistic and multifaceted than selecting based on measurement precision alone. Content balancing, for example, has important practical implications, especially for measures with a high degree of conceptual breadth [34]. The relative efficiency of different methods was evaluated largely based on the correlation with the full-bank score. However, correlations between scores based on subsets of items from the item bank and full-bank scores tend to be inflated, especially when the item bank is small.

Determining the optimal test length for a short form is not entirely a psychometric consideration either; however, it is greatly affected by the item bank characteristics (e.g., the scale information function). The number of items needed for a desired level of measurement precision can vary considerably at different locations on the trait continuum and also by test administration format. Figure 1 displays the maximum attainable information over the trait continuum as a function of the test length. There are 27 curves stacked on top of each other under the scale information function (i.e., the outmost curve). The curve right underneath the scale information function is formed by different sets of 27 items that provide maximum information at the given trait level. Taking a more interesting example near the bottom, the curve labeled “3” represents the total amount of information that would result if we select the best (most informative) three items at each of the trait locations on the continuum (note the best three items will potentially differ at different trait levels). This curve shows that in theory, we could attain the standard error of 0.3 for the examinees with the true trait score between 0.0 and 2.5 with only three items (and six items for SE = 0.2). It also shows that it will take almost the entire bank to attain the same level of measurement precision at −1.0.

These curves in Fig. 1 reflect the best-case scenarios, thus in practice, more items will be needed to attain the same level of precision, because the true trait level of examinees is unknown. It may take 1–2 items under CAT to drill down to a neighborhood of the true trait level and then three additional items to attain the target standard error for most examinees in the theta region between 0.0 and 2.5. Overall, static forms will require a larger number of items because in general they attempt to optimize larger areas at once. However, the relative inefficiency of static forms can be in general largely due to poor measurement in both extremes. The eight-item PROMIS short form provided measurement precision almost as good as the CAT in the middle to upper region of the distribution. However, the CAT provided much better precision at both extremes.

In post hoc simulations using real data, the full-bank score is a practical choice for benchmarking. However, it is of limited utility when it comes to examining the recovery of the known thetas or the conditional precision of measurement. Figure 5 displays the conditional precision as measured by the root mean squared error (RMSE) for the four testing strategies at selected theta locations (1,000 simulees per location). For the theta range from 0 to +3, the four strategies provided excellent measurement precision (RMSE < 0.3). The CAT and two-stage branching strategy yielded considerably higher precision at both extremes (−2 and +4). However, considering the distribution of respondents, the highest impact on performance appeared to be at theta = 0.0 where the two adaptive strategies performed somewhat better than the static short form.

CAT can be an efficient way to measure if a large pool is available collectively covering a broad range of the continuum with individual items covering narrow, targeted trait levels. However, those seem to be uncommon prescriptions for item banks using polytomous items. Item pools with polytomous items tend to be small (20–30 items) [35], and individual items

cover relatively large areas compared to the effective measurement range for the entire bank [2]. For example, the first item chosen by all of the procedures ("I felt depressed") with five response options (*Never* to *Always*) covers a wide range of the continuum, not considerably narrower than the entire measurement range of the bank (see Fig. 4).

We demonstrated that two-stage testing with one locator item and five branches of seven items each can come very close to CAT in terms of measurement efficiency. The correlation between the *CAT-MPWI* and two-stage testing was extremely high (0.992). The two-stage branching also produced the second highest correlation with the full-bank score, only outperformed by the *CAT-MPWI*. One of the key advantages of two-stage testing is that it can be implemented under paper-and-pencil administrations, which can provide a very practical, efficient solution when computer access is limited. Involving only two stages (and the first stage involving only a single item), the potential challenge of self routing can be minimized.

We proposed a new short-form selection criterion, the *Max R*. Although the psychometric properties of this method compared to other procedures are subject to further investigation in future studies, we are assured that this procedure is (by definition) very efficient in maximizing the correlation with the full-bank score (although the utility of this objective is debatable). As a matter of fact, this procedure recovered the dispersion of the full-bank score more accurately (i.e., 0.965 comparable to 0.962 for the full-bank score) than other procedures that produced slightly reduced standard deviations (see Table 2). Another (and perhaps more important) potential for the *Max R* method is that an external measure (rather than the full-bank score) can serve as the criterion for selecting short-form items—for example, an existing *gold standard* measure can be used to select a short form for a new measure. In this use, the issue of inflated part-whole correlations is less relevant. One of the criticisms of abbreviated forms is that they have frequently been developed without careful, thorough examination of their validity compared to their full-length form [36]. The *Max R* has a lot of potential in this regard, e.g., it can be used to maximize the validity coefficient of new measures against their established, external *gold standard* criteria.

Acknowledgments

This study was supported in part by NIH grant PROMIS Network (U-01 AR 052177-04, PI: David Cella). The Patient-Reported Outcomes Measurement Information System (PROMIS) is a National Institutes of Health (NIH) Roadmap initiative to develop a computerized system measuring patient-reported outcomes in respondents with a wide range of chronic diseases and demographic characteristics. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Northwestern University, PI: David Cella, Ph.D., U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, Ph.D., U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, Ph.D., U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, Ph.D., U01AR52170; and University of Washington, PI: Dagmar Amtmann, Ph.D., U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, Ph.D., Lawrence Fine, MD, DrPH, Louis Quatrano, Ph.D., Bryce Reeve, Ph.D., William Riley, Ph.D., and Susana Serrate-Sztein, Ph.D. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the web site at www.nihpromis.org for additional information on the PROMIS cooperative group.

Appendix

See Table 3.

References

1. Bjorner JB, Chang CH, Thissen D, Reeve BB. Developing tailored instruments: item banking and computerized adaptive assessment. Quality of Life Research 2007;16(Suppl 1):95–108. [PubMed: 17530450]

2. Thissen D, Reeve BB, Bjorner JB, Chang CH. Methodological issues for building item banks and computerized adaptive scales. Quality of Life Research 2007;16(Suppl 1):109–119. [PubMed: 17294284]
3. Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, et al. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Medical Care 2007;45(5 Suppl 1):S3–S11. [PubMed: 17443116]
4. Belov DI, Armstrong RD. A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. Applied Psychological Measurement 2008;32(2):119–137.
5. Pilkonis, PA.; Choi, SW.; Reise, SP.; Stover, AM.; Riley, WT.; Cella, D. The development of scales for emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, Anxiety, and Anger. in preparation
6. Fliege H, Becker J, Walter O, Bjorner J, Klapp B, Rose M. Development of a computer-adaptive test for depression (D-CAT). Quality of Life Research 2005;14(10):2277–2291. [PubMed: 16328907]
7. Gardner W, Shear K, Kelleher K, Pajer K, Mammen O, Buysse D, et al. Computerized adaptive measurement of depression: A simulation study. BMC Psychiatry 2004;4(1):13. [PubMed: 15132755]
8. Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagiolini A, Grochocinski VJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services 2008;59(4): 361–368. [PubMed: 18378832]
9. Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph 1969;17
10. Thissen, D.; Chen, W-H.; Bock, RD. Multilog (version 7) [Computer software]. Scientific Software International; Lincolnwood, IL: 2003.
11. Kang T, Chen T. Performance of the generalized S-X^2 item fit index for polytomous IRT models. Journal of Educational Measurement 2008;45(4):391–406.
12. Orlando M, Thissen D. Further investigation of the performance of S-X^2: An item fit index for use with dichotomous item response theory models. Applied Psychological Measurement 2003;27(4): 289–298.
13. Bjorner, JB.; Smith, KJ.; Orlando, M.; Stone, C.; Thissen, D.; Sun, X. IRTFIT: A macro for item fit and local dependence tests under IRT models. Quality Metric, Inc.; Lincoln, RI: 2006.
14. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley WT, Hays RD. Representativeness of the PROMIS Internet Panel. Journal of Clinical Epidemiology. in press.
15. Muthen, LK.; Muthen, BO. Mplus user's guide. 1998.
16. Choi SW. Firestar: Computerized adaptive testing simulation program for polytomous IRT models. Applied Psychological Measurement 2009;33(8):644–645. [PubMed: 20011609]
17. Lord, FM. Applications of item response theory to practical testing problems. Erlbaum; Hillsdale, NJ: 1980.
18. Weiss DJ. Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement 1982;6(4):473–492.
19. Chang H-H, Ying Z. A global information approach to computerized adaptive testing. Applied Psychological Measurement 1996;20(3):213–229.
20. Passos, V. Lima; Berger, MPF.; Tan, FE. Test design optimization in CAT early stage with the nominal response model. Applied Psychological Measurement 2007;31(3):213–232.
21. van der Linden, WJ.; Pashley, PJ. Item selection and ability estimator in adaptive testing. In: van der Linden, WJ.; Glas, CAW., editors. Computerized adaptive testing: Theory and practice. Kluwer Academic; Boston, MA: 2000. p. 1-25.
22. Veerkamp WJJ, Berger MPF. Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics 1997;22(2):203–226.
23. Bock RD, Mislevy RJ. Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement 1982;6(4):431–444.
24. Choi SW, Swartz RJ. Comparison of CAT item selection criteria for polytomous items. Applied Psychological Measurement 2009;33(6):419–440. [PubMed: 20011456]
25. van der Linden W. Optimal assembly of psychological and education tests. Applied Psychological Measurement 1998;22(3):195–211.

26. Reise SP, Henson JM. Computerization and adaptive administration of the NEO PI-R. Assessment 2000;7(4):347–364. [PubMed: 11151961]
27. Hol AM, Vorst HCM, Mellenbergh GJ. Computerized adaptive testing for polytomous motivation items: administration mode effects and a comparison with short forms. Applied Psychological Measurement 2007;31(5):412–429.
28. Kendall MG, Babington SB. The problem of m rankings. The Annals of Mathematical Statistics 1939;10(3):275–287.
29. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20(1):37–46.
30. Yule GU. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society 1912;75:579–652.
31. Warrens M. On association coefficients for 2 × 2 tables and properties that do not depend on the marginal distributions. Psychometrika 2008;73:777–789. [PubMed: 20046834]
32. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. British Journal of Medicine 1994;309:102.
33. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. Annual Review of Clinical Psychology 2009;5:1–25.
34. Reise SP, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. Quality of Life Research 2007;16(Suppl 1):19–31. [PubMed: 17479357]
35. Dodd BG, Koch WR, De Ayala RJ. Operational characteristics of adaptive testing procedures using the graded response model. Applied Psychological Measurement 1989;13(2):129–143.
36. Smith GT, McCarthy DM, Anderson KG. On the sins of short-form development. Psychological Assessment 2000;12(1):102–111. [PubMed: 10752369]

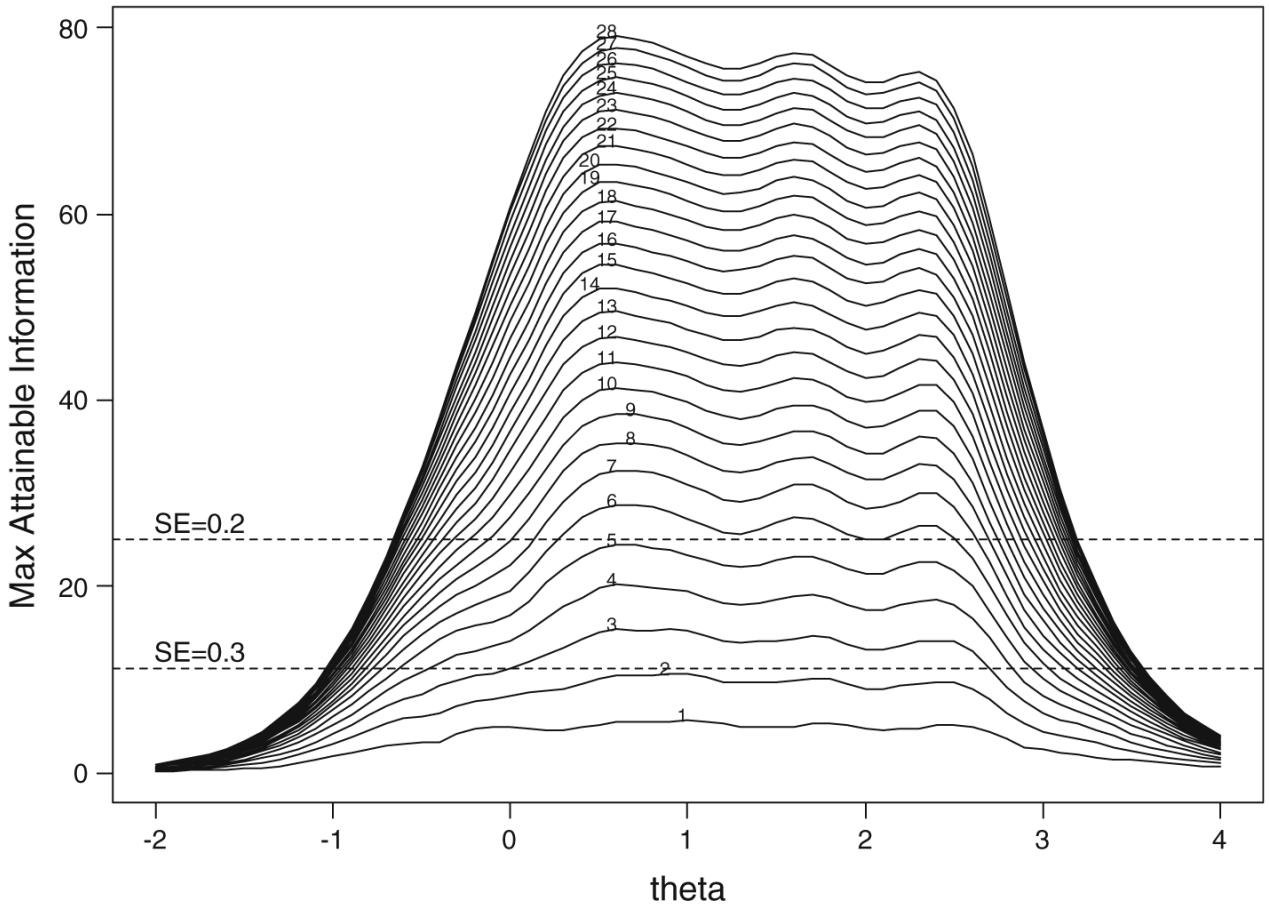

Fig. 1.
Scale (Bank) information function (*the outmost curve*) and maximum attainable information curves (*the inner curves*)

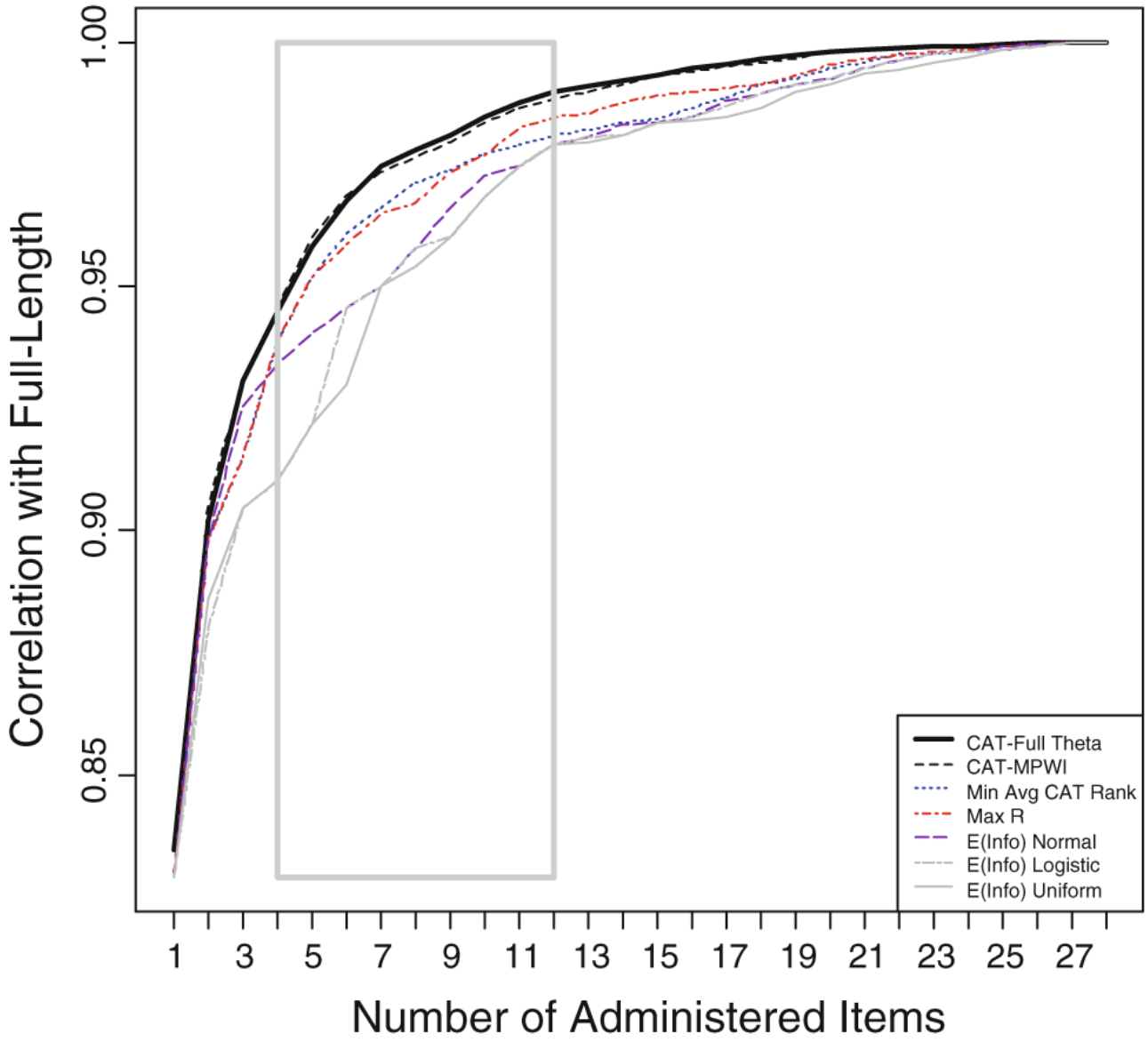

Fig. 2.
Correlation with the full-bank theta estimates as a function of test length (1 through 28 items)

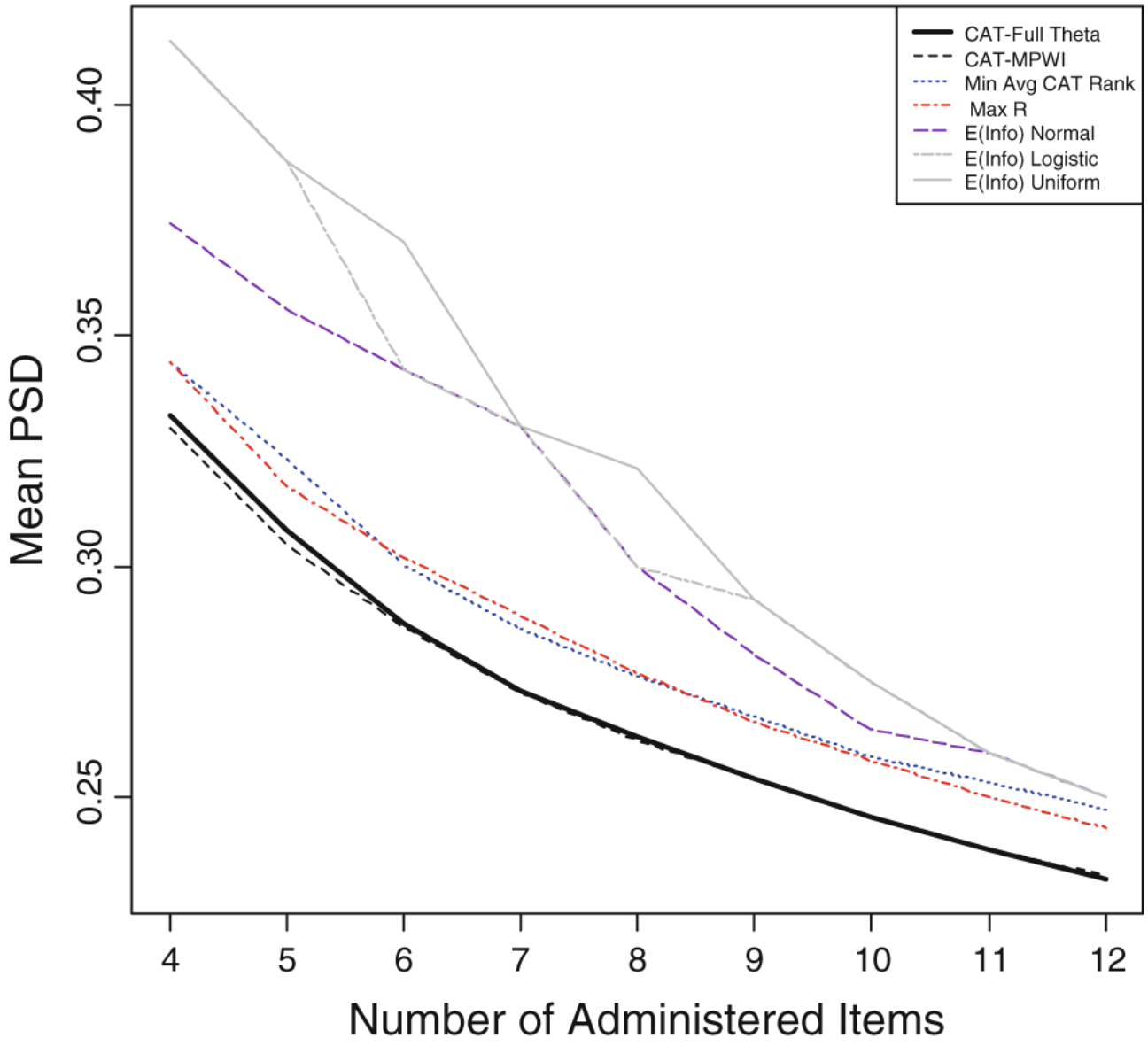

Fig. 3.
Mean posterior standard deviation (PSD) as a function of test length (4 through 12 items)

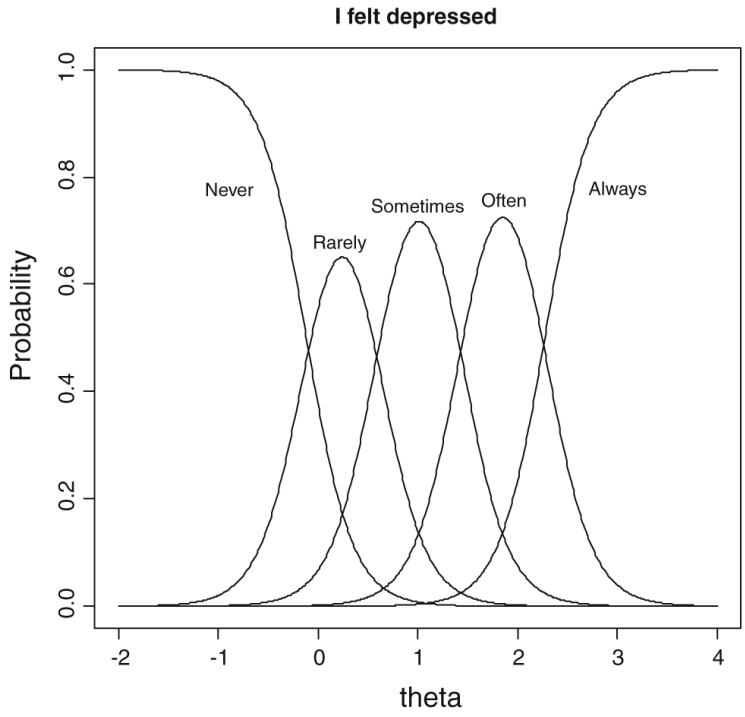

Fig. 4.
Item response functions for the locator question, “I felt depressed”

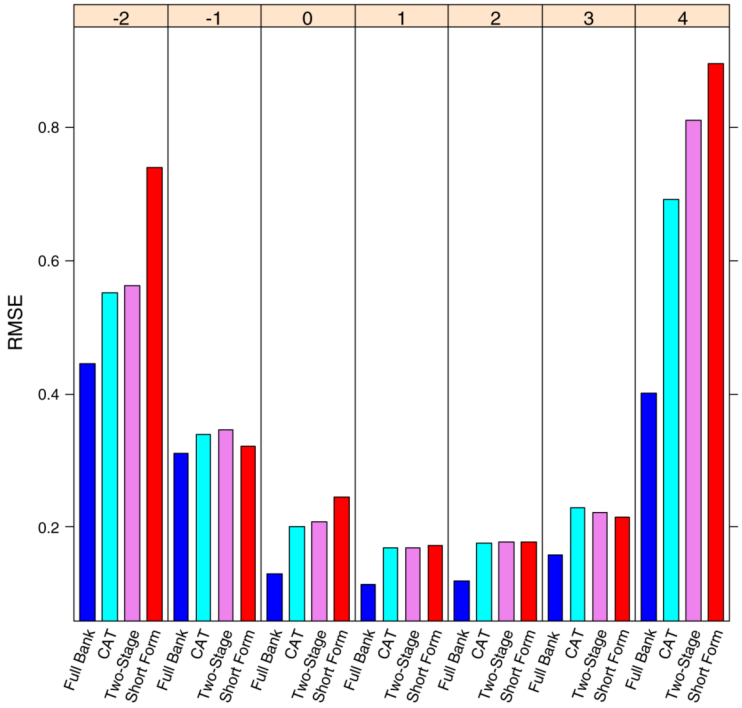

Fig. 5.

Root mean squared error (RMSE) of theta estimates at selected theta points. *Note*: the RMSEs were generated based on 1,000 simulees at each theta location

Table 1

Item selection order by short-form creation strategy

Item	Selection method					Two-stage (Locator—"I felt depressed")				
	E(Info) Normal	E(Info) Logistic	E(Info) Uniform	Max *R*	Min avg CAT rank	Never	Rarely	Sometimes	Often	Always
I felt unhappy*	**2**	**6**	**7**	**2**	**2**	**1**	**5**	**7**		
I felt discouraged about the future	9	10	10	**4**	**4**	**4**				
I felt disappointed in myself	10	11	11	**5**	**6**	**3**				
I felt sad*	**8**	**8**	**9**	**3**	**3**	**2**				
I felt emotionally exhausted	14	15	20	**6**	10	**5**				
I felt depressed*	**1**	**1**	**1**	**1**	**1**					
I felt pessimistic	23	23	25	7	22	**6**				
I felt lonely	20	20	24	12	14					
I felt that my life was empty	13	13	14	25	11					
I found that things in my life were overwhelming	12	12	12	**8**	**7**	**7**				
I had trouble feeling close to people	22	22	23	11	21					
I felt like a failure*	**3**	**5**	**5**	9	**5**		**1**	**5**	**5**	**4**
I withdrew from other people	17	18	19	15	18					
I felt that I had nothing to look forward to*	**7**	**7**	**6**	20	**8**		**4**	**4**	**4**	**5**
I felt that I was not needed	16	16	17	22	19					
I felt that I was to blame for things	19	19	18	17	16					
I felt helpless*	*5*	**4**	**4**	23	9		**2**	**3**	**3**	**3**
I felt that I was not as good as other people	26	27	27	18	26					
I felt worthless*	**4**	**2**	**3**	24	12		**3**	**2**	**2**	**2**
I had trouble making decisions	21	21	21	10	20					
I felt that nothing could cheer me up	11	9	**8**	26	13		**7**	**6**	**6**	**6**
I felt that nothing was interesting	18	17	15	19	17					
I felt guilty	28	28	28	13	25					

Item	Selection method					Two-stage (Locator—"I felt depressed")				
	E(Info) Normal	E(Info) Logistic	E(Info) Uniform	Max *R*	Min avg CAT rank	Never	Rarely	Sometimes	Often	Always
I felt that I wanted to give up on everything	15	14	13	27	23				**7**	
I felt hopeless*	**6**	**3**	**2**	14	15		**6**	**1**	**1**	**1**
I felt ignored by people	25	26	26	21	27					
I felt upset for no reason	24	25	22	16	24					
I felt I had no reason for living	27	24	16	28	28					**7**
Cronbach's Alpha for eight-item short forms	.953	.953	.954	.954	.948	.947	.954	.954	.954	.950

Note: Items are ordered by mean category threshold values (from low to high levels of depressive symptoms). Numbers in bold indicate top eight items

*Indicates PROMIS short-form items. *E(Info) Normal*: Expected information using normal weighting, *E(Info) Logistic*: Expected information using logistic weighting, *E(Info) Uniform*: Expected information using uniform weighting, *Max R*: Maximum correlation with full-bank score, *Min Avg CAT Rank*: Minimum average CAT presentation rank. Two-Stage: A locator item ("I felt depressed.") followed by five different branches of seven items each depending on the responses to the locator item

Table 2

Comparison of performance—eight-item CAT and short forms (n = 768)

	CORR[a]	RMSD[b]	Mean PSD[c]	Kappa[d]	Sensitivity[e]	Specificity[f]	Positive predicted value[g]	Negative predicted value[h]	Yule's Y[i]	% Floor[j]	% Ceiling[k]	Mean Theta[l]	SD Theta[m]
PROMIS													
Short form	0.958	0.279	0.300	0.925	0.953	0.986	0.918	0.992	0.949	23.18	0.26	−0.085	0.951
CAT-MPWI	0.977	0.209	0.262	0.855	0.858	0.983	0.892	0.977	0.900	13.54	0.13	−0.095	0.958
Two-stage	0.975	0.217	0.267	0.901	0.915	0.986	0.915	0.986	0.931	13.67	0.39	−0.093	0.951
Selection criteria													
CAT-full theta	0.978	0.205	0.263	0.862	0.868	0.983	0.893	0.979	0.903	13.28	0.13	−0.097	0.962
Min avg CAT rank	0.971	0.229	0.276	0.893	0.934	0.980	0.884	0.989	0.927	15.89	0.13	−0.066	0.943
Max R	0.967	0.247	0.277	0.857	0.868	0.982	0.885	0.979	0.899	13.67	0.26	−0.087	0.965
E(Info) Normal	0.958	0.279	0.300	0.925	0.953	0.986	0.918	0.992	0.949	23.18	0.26	−0.085	0.951
E(Info) Logistic	0.958	0.279	0.300	0.925	0.953	0.986	0.918	0.992	0.949	23.18	0.26	−0.085	0.951
E(Info) Uniform	0.954	0.290	0.321	0.886	0.906	0.983	0.897	0.985	0.919	29.30	0.39	−0.091	0.943
Random Selection 1,000 Reps													
Minimum	0.934	0.224	0.286	0.784	0.793	0.964	0.795	0.967	0.847	13.80	0.13	−0.142	0.892
Q1	0.954	0.260	0.313	0.850	0.859	0.979	0.868	0.978	0.895	19.01	0.13	−0.088	0.922
Median	0.959	0.274	0.324	0.869	0.887	0.982	0.888	0.982	0.908	21.16	0.13	−0.073	0.933
Mean	0.958	0.276	0.325	0.868	0.885	0.982	0.888	0.982	0.908	21.37	0.18	−0.073	0.933
Q3	0.963	0.290	0.336	0.886	0.906	0.985	0.907	0.985	0.923	23.44	0.26	−0.057	0.945
Maximum	0.972	0.345	0.373	0.940	0.972	0.997	0.980	0.995	0.965	32.94	0.52	−0.008	0.973

[a] CORR: correlation with full-bank scores

[b] RMSD: root mean squared difference

[c] Mean PSD: mean posterior standard deviation

[d] Kappa: Kappa statistic as a measure of classification consistency (threshold: η = 1.0)

[e] Sensitivity: true-positives/(true-positives + false-negatives)

[f] Specificity: true-negatives/(true-negatives + false-positives)

[g] Positive predictive value: true-positives/(true-positives + false-positives)

[h] Negative predictive value: true-negatives/(true-negatives + false-negatives)

[i] Yule's Y coefficient of colligation

[j] % Floor: % minimum score for all items

[k] % Ceiling: % maximum score for all items

[l] Mean theta: mean of theta estimates

[m] SD theta: standard deviation of theta estimates

Table 3

PROMIS depression bank parameters (graded response model)

Item	Slope	Threshold1	Threshold2	Threshold3	Threshold4
I felt unhappy	3.483	−0.536	0.348	1.347	2.355
I felt discouraged about the future	3.183	−0.261	0.397	1.306	2.134
I felt disappointed in myself	3.093	−0.358	0.413	1.404	2.224
I felt sad	3.274	−0.499	0.406	1.413	2.376
I felt emotionally exhausted	2.685	−0.299	0.424	1.358	2.308
I felt depressed	4.343	−0.117	0.598	1.428	2.273
I felt pessimistic	2.381	−0.458	0.478	1.546	2.632
I felt lonely	2.588	−0.079	0.633	1.477	2.328
I felt that my life was empty	3.185	0.198	0.782	1.526	2.324
I found that things in my life were overwhelming	3.106	0.044	0.722	1.639	2.472
I had trouble feeling close to people	2.564	−0.038	0.693	1.653	2.584
I felt like a failure	3.970	0.204	0.796	1.649	2.296
I withdrew from other people	2.802	0.148	0.772	1.603	2.538
I felt that I had nothing to look forward to	3.932	0.305	0.913	1.594	2.412
I felt that I was not needed	2.920	0.204	0.891	1.655	2.528
I felt that I was to blame for things	2.736	0.073	0.810	1.803	2.673
I felt helpless	4.145	0.350	0.915	1.678	2.471
I felt that I was not as good as other people	2.333	0.186	0.947	1.729	2.633
I felt worthless	4.261	0.401	0.976	1.696	2.444
I had trouble making decisions	2.613	−0.023	0.868	1.864	2.826
I felt that nothing could cheer me up	3.657	0.312	0.982	1.782	2.571
I felt that nothing was interesting	2.834	0.141	0.907	1.846	2.875
I felt guilty	2.018	−0.050	0.926	2.000	2.966
I felt that I wanted to give up on everything	3.241	0.461	1.034	1.834	2.515
I felt hopeless	4.454	0.558	1.074	1.779	2.530
I felt ignored by people	2.364	0.210	0.987	1.906	2.934
I felt upset for no reason	2.549	0.194	1.012	2.013	3.127
I felt I had no reason for living	3.131	0.918	1.481	2.164	2.856