Utilizing Multidimensional Computer Adaptive Testing to Mitigate Burden With Patient Reported Outcomes
Michael Bass, MS1, Scott Morris, PhD2, Richard Neapolitan, Ph.D1
1Northwestern University, Chicago, Il, 2Illinois Institute of Technology, Chicago, Il

Abstract
Utilization of patient-reported outcome measures (PROs) had been limited by the lack of psychometrically sound measures scored in real-time. The Patient Reported Outcomes Measurement Information System (PROMIS) initiative developed a broad array of high-quality PRO measures. Towards reducing the number of items administered in measuring PROs, PROMIS employs Item Response Theory (IRT) and Computer Adaptive Testing (CAT). By only administering questions targeted to the subject’s trait level, CAT has cut testing times in half. The IRT/CAT implementation in PROMIS is unidimensional in that there is a separate set of questions administered for each measured trait. However, there are often correlations among traits. Multidimensional IRT (MIRT) and multidimensional CAT (MCAT) provide items concerning several correlated traits, and should ameliorate patient burden. We developed an MIRT model using existing PROMIS item banks for depression and anxiety, developed MCAT software, and compared the efficiency of the MCAT approach to the unidimensional approach.

Note: Research reported in this publication was supported in part by the National Library of Medicine of the National Institutes of Health under Award Number R01LM011962.

Introduction
There is increasing pressure for integration of the patient’s perspective in clinical research. An indication of this growth is the recent establishment of the Patient-Centered Outcomes Research Institute (PCORI) by Congress. PCORI’s goals include identifying the best available evidence to help patients and their healthcare providers make more informed decisions. Patient-reported outcomes (PROs) offer the patient’s perspective. The notion of the pre-eminence of the patient perspective has been prominent in literature for at least 20 years. The research literature contains ample evidence for not relying solely on clinician reporting of patients’ subjective experiences. Discordance between patient and clinician reports of Health-Related Quality of Life (HRQL) and health status has been widely reported. In general, clinicians have reported fewer symptoms and lower symptom severity compared to patient reports.

There is also increasing evidence that PROs can provide more information about toxicity and symptoms than physician or adverse event (AE) reports. Studies have shown that, for some subjective symptoms (e.g., in sexual function), clinician ratings are no better than chance in predicting PROs of the same symptoms. These findings support the need for tools to collect patient-reported symptoms and HRQL. The inclusion of PRO measures of AEs and other symptoms in clinical trials and other research recently has been recommended as a means to improve the quality and completeness of data, provide a more comprehensive reflection of patient experiences, and improve the operational efficiency of conducting studies.

The need to reduce patient burden in measuring PROs has long been a concern in health outcomes research. This concern has led researchers away from reliance on classical testing models and toward item response theory (IRT) and computer adaptive testing (CAT). These methods have been used successfully for decades in education, licensure, personality assessment, and selection of military personnel. By only administering questions targeted to the individual’s ability or trait level, testing times have on average been cut in half while at the same time improving overall test reliability. The efficiencies long noted in educational testing and credentialing are equally applicable in the measurement of PROs. Underlying every CAT is a bank of items “calibrated” to an IRT model. This model allows scores estimated from any subset of the entire bank to be expressed in the same metric. Therefore, although different individuals take different subsets of items, their scores are directly comparable. In a CAT-based measure, an initial item from the bank is administered. A person’s response to that item provides a rough estimate of the person’s level of the trait being measured. For example, a patient who initially chooses the response, “Very Much,” when asked, “How much did pain interfere with your day to day activities?” is likely to have a high level of pain interference. The items subsequently administered would be those known to be effective in discriminating among people experiencing high pain. On the other hand, an individual who responded “Never” to this question would be asked additional questions known to be good at discriminating among people with low levels of pain interference. Because the CAT only administers the items that are most informative for each individual, patients can typically
complete a PRO CAT in an average of five questions with reliability similar to a traditional 15–25-item survey measure. Reducing the number of items on each instrument lowers the burden on patients and allows the measurement of more PROs, thereby supporting more comprehensive assessments.

The Patient Reported Outcomes Measurement Information System (PROMIS) is an initiative, funded by the National Institutes of Health (NIH) Common Fund, to improve assessment of self-reported symptoms and other health-related quality of life domains. CAT-based PROs have become much more accessible through PROMIS. A primary goal for PROMIS was the creation of CATs to enable efficient, psychometrically robust assessment of PROs for a wide range of health outcomes research. To achieve this goal, customized banks of items for measuring PROs were developed and tested at 12 primary research and data collection sites. Initially PROMIS developed and validated 27 adult and 9 pediatric item banks across the domains of physical, mental and social health. Many PROMIS instruments were developed by scientists who had previously authored well-known PROs including the Short Form-36, the Functional Assessment of Chronic Illness Therapy, and the Health Assessment Questionnaire. In practice, PROMIS banks have shown greater reliability, fewer ceiling and floor effects and more responsiveness to change than the “gold standard” instruments.

The PROMIS initiative lays the foundation for widespread health outcomes screening to take place in the waiting room of healthcare providers. These screenings can coincide with the current activities that usually occur while a patient is waiting for their healthcare provider, such as filling out or updating their past medical history and insurance forms. The data collected from these screenings can be incorporated into the patient’s electronic health record (EHR) alongside other clinical patient data. This will give healthcare providers a more accurate picture of their patients overall health and healthcare needs. Healthcare providers can make more informed decisions with regards to mental health referrals or treatment options.

In order to realize these benefits, PRO screenings must capture maximum information with minimum patient burden. PROMIS is unidimensional in that there is a separate item bank and therefore a separate set of questions administered for each measured trait. This has been a good initial step in addressing the issue of patient burden, but in the real world things are not always as simple as the one dimensional approach; that is, there are correlations between many of the traits assessed by the PROMIS system. For example, depression and anxiety are closely related, which means many questions that ascertain level of depression can also provide information about the level of anxiety. A multidimensional IRT (MIRT) model represents items concerning several correlated traits. A multidimensional CAT (MCAT) selects items from an item pool such that the items selected maximize information provided on several correlated traits. Such a model would significantly reduce the number of items administered in measuring PROs and thereby ameliorate the patient’s burden of answering many questions.

Researchers have recently realized the acute value of the MCAT approach in health services. In this research we investigated whether MCAT can lessen patient burden relative to unidimensional CAT (UCAT) in the domain of PROs. We performed this investigation using the related traits “anxiety” and “depression.” First, we developed an MIRT model for these two traits using existing PROMIS item banks. Then we developed MCAT software to administer questions based on this MIRT model. We compared the performance of this MCAT approach to the unidimensional approach currently used by PROMIS using 20,000 simulated individuals. We found that the MCAT consistently provided comparable estimates of anxiety and depression with fewer items, relative to administering separate UCATs.

**IRT and CAT Review**

Since the methodology concerns IRT and CAT, we first briefly review these two.

**Item Response Theory.** In item response theory (IRT), we estimate an individual’s traits from the individual’s responses. Each trait is a skill, knowledge, or quality in some domain. For example, a trait may be “arithmetic ability,” while the responses are answers to arithmetic problems. Traits are hidden variables; responses are observed variables. Figure 1 shows a Bayesian network (BN) representing an IRT model with one trait $\theta$ (unidimensional) and five items.

The probability distribution of correctly answering a dichotomous item given the level of a trait is commonly modeled using the following two parameter logistic evidence model:

$$P(I_j = \text{Right}) = \frac{1}{1 + e^{-a_j (\theta - b_j)}}, \quad (1)$$
where $b_j$ measures the location of the item on the latent trait continuum (e.g., whether it is easy or difficult), and $a_j$ is a discrimination parameter reflecting how sensitive the item is to differences in the latent trait.

![Figure 1. A unidimensional IRT model.](image)

In the case of PROs there is no “correct” response. Instead, a polytomous IRT model is used to represent the probability of each option on the response scale. The probability distribution of a polytomous item with $m+1$ categories given trait $\theta$ in the graded response model\(^{36}\) is computed as follows:

$$P_{jk}(\theta) = \frac{1}{1 + e^{-a_j (\theta - b_k)}} = P(I_j \geq k), \quad k = 1, 2, ..., m$$

$$P_{j,m+1}(\theta) = P(I_j \geq m + 1) = 0$$

where $b_{j1} < b_{j2} < \cdots < b_{jm}$. The probability of selecting option $k$, $k = 0, 2, ..., m$, is given by

$$P(I_j = k) = P(I_j \geq k) - P(I_j \geq k + 1).$$

Traits in a given domain can be related, which means many items that ascertain the level of one trait can also provide information about the level of other traits. For example, many items that ascertain depression level can also provide information about anxiety level\(^{37}\). These relationships can be modeled using an MIRT model. In the case of MIRT, Equation 1 for the two parameter dichotomous logistic model is as follows:

$$P(I_j = \text{Right}) = \frac{1}{1 + e^{-a_j \theta - b_j}}$$

(2)

where $\theta = (\theta_1, \theta_2, ..., \theta_n)$ is a vector of traits and $a_j$ is a vector of discrimination parameters.

**Computer Adaptive Testing.** Computer adaptive testing (CAT)\(^{38}\) improves test measurement quality and efficiency by striving to administer optimal items to each examinee. CAT selects items sequentially to minimize the standard error of the estimate of $\theta$. Commonly employed item selection criteria used in CAT include maximizing Fisher’s information (FI)\(^{39}\), and using the Kullback-Leibler information\(^{40}\). Chen et al.\(^{41}\)
In unidimensional IRT, Fisher's information for a polytomous item given trait level \( \theta \) is as follows:

\[
I_j(\theta) = \sum_{k=0}^{m} \left[ \frac{P_{jk}(\theta)(1 - P_{jk}(\theta)) - P_{j,k+1}(\theta)(1 - P_{j,k+1}(\theta))}{P_{jk}(\theta) - P_{j,k+1}(\theta)} \right]^2.
\]

As indicated in this expression, information is a function of the individual’s standing on the latent variable \( \theta \), which is unknown. To address the uncertainty inherent in the estimate of \( \theta \), we computed the maximum posterior weighted information (MPWI)\(^{42}\), where the expected value of \( I_j(\hat{\theta}) \) is taken across the posterior distribution of based on the individual’s responses so far. The next item chosen is the one that maximizes the MPWI.

In the case of MIRT, FI takes the form of a matrix, and the determinant of the information matrix serves as a summary measure of the information regarding the set of traits\(^{29}\). Thus, the next item is chosen that maximizes the posterior weighted determinant of the information.

**Methods**

The current study presents the result of a Monte Carlo simulation designed to evaluate the reduction in test length when using multidimensional versus unidimensional CAT. In order to represent a typical PRO assessment situation, we simulated responses to the actual PROMIS Anxiety and Depression banks. The PROMIS banks were developed using unidimensional IRT analyses. Therefore, an initial step toward building the simulation was to estimate the item parameters of a 2-dimensional MIRT model.

**Measures.** The PROMIS item banks were developed though a mixture of qualitative and quantitative methods. Rigorous scale construction practices were followed, with substantial input from domain experts, as well as revision based on respondent feedback. Statistical item analyses were conducted using both classical and IRT-based methods to identify the best items for inclusion in the final bank. Cella et al.\(^{25}\) The banks consist of 29 anxiety items and 28 depression items from the current PROMIS item banks.

For these items, respondents are asked to report the frequency of experiences during the past seven days. Examples of item content are as follows: “I was easily startled” (anxiety), and “I felt worthless” (depression). Responses were indicated on a 5-point scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always).

**Estimating the MIRT Model.** Calibrating an IRT model involves obtaining estimates of item parameters described in Equation 2. Unidimensional item parameters were available from prior work on the development of the PROMIS banks\(^{25}\). For the MCAT, a preliminary analysis was conducted to obtain the necessary item parameters. Item parameters for a 2-dimensional Graded Response Model were estimated using the program IRTPRO.

This step utilized existing data collected for the development of PROMIS instruments\(^{25}\). The sample was 52% female with a median age of 50. The racial composition was 82% white, 9% black, 8% multi-racial, and 1% other. Nine percent were Latino/Hispanic\(^{22}\). The current study used a subsample of 7945 adults sampled from the general population who completed the anxiety and depression items.

The parameter estimates from the MIRT model were found to correspond closely to those from the unidimensional analyses (all correlations > .9). Items tended to have large discrimination parameters, indicating that each item was a strong indicator of the underlying trait. The average a parameter value was 2.78 (ranging from 1.41 to 3.76) for Anxiety and 3.23 (ranging from 2.21 to 4.72 for Depression).

Item location (b) parameters indicated that the items tended to reflect higher levels of anxiety and depression. For example, among individuals with a slightly below average level of the trait (0.5 SD below the mean), the most common response selected was the lowest option on the 5-point scale. This indicates that the items in this bank are primarily sensitive to moderate to high levels of anxiety and depression, and are less able to distinguish among individuals with low levels of anxiety and depression.

The preliminary analysis also yielded estimates of Anxiety and Depression levels for each individual in the data. The distribution of both anxiety and depression were positive skewed, with approximately 70% of individuals...
obtaining trait values near the mean ($\theta = 0$) or slightly below ($\theta = -1$), and the remaining 30% demonstrating higher levels of the trait. The two traits were highly correlated, $r = 0.88$, $p < 0.001$.

**Monte Carlo Simulation.** The simulation consisted of generating data for 20,000 respondents who completed both the UCAT and MCAT versions of the PROMIS Anxiety and Depression measures.

We first generated the true trait levels of each respondent by sampling from the bivariate frequency distribution obtained during the IRT calibration step. This approach produced simulated data representative of anxiety and depression levels observed in the general population.

Each simulated respondent was administered three CATs: MCAT for anxiety and depression, UCAT for anxiety, and UCAT for depression. First, an initial item was selected by the CAT algorithm. Next, a probability distribution for the individual was generated according to that individual’s true trait levels and the IRT model for that item. The IRT model corresponding to each CAT (i.e., unidimensional or multidimensional) was used to generate a probability distribution. Next, a response was generated according to this probability distribution.

After each item, the CAT then updated the posterior distribution of $\theta$ based on the individual’s response, along with the expected a posteriori (EAP) estimate $\hat{\theta}$, and the standard error (SE) of this estimate. This updated theta distribution was used by the CAT algorithm to select the next item. This was repeated for the first 20 items administered by each of the CATs.

The CAT was designed to stop when the SE of each $\theta$ estimate falls below a threshold. The SE is a measure of uncertainty in the trait estimate (i.e., how far the estimate is expected to fall from the true trait level). Consequently, when the SE falls below the threshold, this indicates that the trait has been estimated to an acceptable level of precision, and is unlikely to change substantially on further testing. For the current study, we considered a variety of thresholds, ranging from 0.2 to 0.4 in increments of .025. This range was centered on the threshold of 0.3 currently implemented in the unidimensional PROMIS CATs.

Two primary outcomes of interest were the number of items required to achieve the stopping rule, and the accuracy of the resulting EAP estimates, operationalized as the root mean square error,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N}((\hat{\theta}_{Anx,t,i} - \theta_{Anx,t,i})^2 + (\hat{\theta}_{Dep,t,i} - \theta_{Dep,t,i})^2)}{N}},$$

where $\hat{\theta}_{t,i}$ and $\theta_{t,i}$ represent the EAP estimate and true value of trait $t$ for respondent $i$, and $N$ is the number of simulated respondents.

**Table 1.** Percent of simulated respondents meeting the stopping rule on both anxiety and depression within 20 items.

<table>
<thead>
<tr>
<th>SE Cutoff</th>
<th>MCAT</th>
<th>UCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>56%</td>
<td>63%</td>
</tr>
<tr>
<td>0.225</td>
<td>63%</td>
<td>64%</td>
</tr>
<tr>
<td>0.25</td>
<td>67%</td>
<td>67%</td>
</tr>
<tr>
<td>0.275</td>
<td>71%</td>
<td>72%</td>
</tr>
<tr>
<td>0.3</td>
<td>79%</td>
<td>76%</td>
</tr>
<tr>
<td>0.325</td>
<td>82%</td>
<td>80%</td>
</tr>
<tr>
<td>0.35</td>
<td>85%</td>
<td>86%</td>
</tr>
<tr>
<td>0.375</td>
<td>88%</td>
<td>89%</td>
</tr>
<tr>
<td>0.4</td>
<td>91%</td>
<td>90%</td>
</tr>
</tbody>
</table>
Results

An initial analysis indicated that not all simulated respondents reached the stopping rule within the 20 simulated items (see Table 1). This was particularly problematic for the more stringent stopping rules, which required a SE below 0.3. Further analyses revealed that the cases that failed to meet the stopping rule generally occurred when one or both of the traits were less than 0. As noted above, the items available in the Anxiety and Depression banks tend to be more sensitive to higher levels of these traits. The absence of items targeted toward lower levels of these traits limits the precision obtainable for individuals with below-average anxiety or depression. Thus, the administration of additional items is unlikely to change this result.

Because the outcome measures could be computed only for cases that met the stopping rule, respondents who did not satisfy this requirement were excluded from subsequent analyses. To ensure comparability, a respondent was included in the results only when the stopping rule was reached for both the UCAT and MCAT for a particular cutoff. However, the number of examinees was allowed to vary across cutoffs, so that more data was available for less stringent cutoffs (e.g., N=16,817 for SE<0.4) than for more stringent cutoffs (e.g., N=11,262 for SE < 0.2).

The average number of items required to reach the stopping rule is presented in Table 2. The MCAT required fewer items at each of the cutoff levels (differences were all significant, **p<0.001**). On average the MCAT required 1.2 fewer item than the UCAT. Furthermore, despite the reduced number of items, the MCAT and UCAT resulted in equally accurate trait estimates, as indicated by similar RMSE values. The greater efficiency of the MCAT is illustrated in Figure 2, which plots the RMSE against the average number of items required for each stopping rule. The lower curve for the MCAT represents greater accuracy given the same number of items, relative to the UCAT.

**Table 2.** Number of Items Required to Reach the Stopping Rule on Both Traits and RMSE of UCAT and MCAT.

<table>
<thead>
<tr>
<th>SE</th>
<th>Cutoff N</th>
<th># Items To Cutoff</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCAT</td>
<td>MCAT</td>
<td>UCAT</td>
<td>MCAT</td>
</tr>
<tr>
<td>0.2</td>
<td>11262</td>
<td>16.3</td>
<td>14.9</td>
</tr>
<tr>
<td>0.225</td>
<td>12454</td>
<td>12.6</td>
<td>11.6</td>
</tr>
<tr>
<td>0.25</td>
<td>12727</td>
<td>10.1</td>
<td>9.3</td>
</tr>
<tr>
<td>0.275</td>
<td>13137</td>
<td>8.8</td>
<td>7.6</td>
</tr>
<tr>
<td>0.3</td>
<td>13910</td>
<td>7.7</td>
<td>6.6</td>
</tr>
<tr>
<td>0.325</td>
<td>14565</td>
<td>7.4</td>
<td>5.7</td>
</tr>
<tr>
<td>0.35</td>
<td>15489</td>
<td>6.9</td>
<td>5.3</td>
</tr>
<tr>
<td>0.375</td>
<td>16343</td>
<td>6.2</td>
<td>5.2</td>
</tr>
<tr>
<td>0.4</td>
<td>16817</td>
<td>5.9</td>
<td>4.8</td>
</tr>
<tr>
<td>Average</td>
<td>9.1</td>
<td>7.9</td>
<td>0.37</td>
</tr>
</tbody>
</table>
Discussion

The results of the Monte Carlo simulation clearly demonstrate the benefits of multidimensional adaptive testing for patient reported outcomes. Consistent with prior research\textsuperscript{28}, MCAT was found to produce accuracy similar to UCAT while reducing the number of items administered.

The absolute reduction is the test length was considerable. The UCAT required 9 items on the average, while the MCAT required only 8 items. This corresponds to an 11% reduction of the test length, which represents an important reduction in patient burden. However, the reduction in test length observed here is less than that has been found in other MCAT applications\textsuperscript{28}. This may be due to the nature of the items examined in study. In general, each item provides considerable information about the trait. The small number of items needed to obtain precise trait estimates may limit the potential for refinements such as MCAT to reduce test length to the same degree as that achieved when there are a large number of items.

A strength of the current simulation is the use of actual items from the widely used PROMIS anxiety and depression banks to build the simulation, thereby enhancing the realism of the simulation. Nevertheless, the results represent only these two traits, and may not generalize to other PRO measures.

A distinctive characteristic of anxiety and depression is their very high correlation ($r=0.88$). It is possible that MCAT will demonstrate less benefit over UCAT when less correlated traits are modeled, because each trait will provide less information about the other. On the other hand, EAP trait estimates tend to be slightly biased toward the mean of the prior distribution, and this effect is heightened in MCAT when the traits are highly correlated. Therefore, MCAT with less correlated traits may exhibit less bias and greater accuracy. Further research is needed to replicate these findings with other combinations of traits.

The sample of participants for the current study was mostly white (82%), and the generalizability of these findings across ethnic groups merits further study. Mean differences across groups should not affect the performance of MCAT assessments. However, if ethnic subgroups interpret and respond to items differently, this will affect the parameters of the IRT model, and consequently the accuracy of the IRT-based adaptive tests. Although there is evidence that measures of depression may function differentially across ethnic groups\textsuperscript{43,44}, little is known about the impact of these differences on UCAT and MCAT measures.

References


