Published in final edited form as:

Int J Test. 2020 ; 20(2): 146–168.

Stopping Rules for Computer Adaptive Testing When Item Banks Have Nonuniform Information

Scott B. Morris¹, Michael Bass², Elizabeth Howard¹, Richard E. Neapolitan³

¹Department of Psychology, Illinois Institute of Technology, Chicago, IL USA

²Department of Medical Social Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL USA

³Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL USA

Abstract

The *standard error (SE)* stopping rule, which terminates a *computer adaptive test (CAT)* when the SE is less than a threshold, is effective when there are informative questions for all trait levels. However, in domains such as patient reported outcomes, the items in a bank might all target one end of the trait continuum (e.g., negative symptoms), and the bank may lack depth for many individuals. In such cases, the *predicted standard error reduction (PSER)* stopping rule will stop the CAT even if the *SE* threshold has not been reached, and can avoid administering excessive questions that provide little additional information. By tuning the parameters of the PSER algorithm, a practitioner can specify a desired tradeoff between accuracy and efficiency. Using simulated data for the PROMIS *Anxiety* and *Physical Function* banks, we demonstrate that these parameters can substantially impact CAT performance. When the parameters were optimally tuned, the PSER stopping rule was found to outperform the *SE* stopping rule overall and particularly for individuals not targeted by the bank, and presented roughly the same number of items across the trait continuum. Therefore, the PSER stopping rule provides an effective method for balancing the precision and efficiency of a CAT.

Keywords

patient reported outcome; item response theory; computer adaptive testing; stopping rule

INTRODUCTION

Computer adaptive tests (CATs) are an attractive option for assessment contexts where measurement must be both precise and efficient. By tailoring questions to each examinee, a CAT allows for much shorter tests without sacrificing measurement precision. CAT methods have been used successfully for decades in education and professional licensure (Melican, Breithaupt, & Zhang, 2010; Rudner, 2010; Sands & Waters, 1997). More recently, there has

been an increase in the use of CATs for assessing non-cognitive domains (Cella et al., 2010; Gibbons, Bower, Lovell, Valderas, & Skevington, 2016; Stark, Chernyshenko, Drasgow, & White, 2012). As the use of CAT expands, it is important for testing professionals to understand the unique challenges of assessments in other contexts.

The current research explores the use of CAT in assessing patient reported outcomes (PROs). PROs provide information about a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a physician, and include subjective reports of symptoms, physical and social functioning, and health-related quality of life (Snyder, Jensen, Segal, & Wu, 2013). PROs have become an important component of healthcare practice and health outcomes research, and their use is on the rise worldwide (Alonso et al., 2013; Castrejón, Gossec, & Carmona, 2015; Falavigna et al., 2017; Vilagut et al., 2015).

When collecting PROs, there is frequently concern about the burden on patients of completing lengthy questionnaires. This is an ideal context for the use of CATs, which can often provide highly reliable scores with a small number of items (Gershon & Cook, 2011). However, the nature of PRO constructs creates several challenges for implementing a CAT. Item pools for PROs are often relatively small and demonstrate a restricted range of trait locations (Revicki, Chen, & Tucker, 2015). The potential benefits of CAT may be unrealized if there are many individuals for whom the item bank is not well aligned.

Concerns about bank alignment may be further exacerbated when measures are adapted across cultures, because the distribution of the trait may differ from the country in which the measure originated (Apers et al., 2016; Rubin, Peyrot, & Siminerio, 2006). An item bank that is well aligned to the trait levels common in one country may function less well in another setting because trait levels in the second country are not well represented in the item bank (e.g., Shimazu, Schaufeli, Miyanaka, & Iwata, 2010).

Testing professionals have long recognized the need to re-evaluate psychometric properties when assessments are applied across new cultures or nations (International Test Commission, 2017). Of particular relevance here is the need to consider differences in the range of the trait being assessed, and its implications for the mode of assessment (International Test Commission, 2018, recommendation 2.7). Before implementing a CAT, it is important to assess the alignment of the item bank with the trait distribution in the target population, and to consider the implications of misalignment on CAT performance.

One important decision in the design of a CAT is the stopping rule (Wainer, Dorans, Flaugher, Green, & Mislevy, 2000). A common rule is to stop the CAT when the *standard error* (*SE*) of the trait estimate falls below a threshold (e.g., $SE < 0.3$). If the bank has a uniform distribution of item locations, this stopping rule will guarantee a desired level of measurement precision for all examinees (Babcock & Weiss, 2009; Wainer et al., 2000). However, PRO banks often have limited depth at some trait levels. For example, measures of anxiety and depression tend to focus on negative symptoms, and physical function measures focus on physical limitations. Consequently, item banks in these domains have few items targeted toward positive states.

When item banks have limited depth at some trait levels, the *SE* stopping rule can be inefficient (Choi, Grady, & Dodd, 2010). If there are few (or no) items that provide information at a particular trait level, an excessive number of items may be required to reach the *SE* cutoff. In fact, the CAT might administer all items in the bank without ever achieving the desired level of precision. CATs often have a preset maximum item limit to prevent this from happening. However, such rules are *ad hoc*, and are not likely to generalize to all persons or banks.

A more theoretically justified approach would stop the CAT when additional items are unlikely to further improve the trait estimate. Choi, Grady & Dodd (2010) proposed the *predicted standard error reduction* (PSER) stopping rule, which terminates a CAT when there are no remaining items that are likely to substantially improve measurement precision. That is, if an examinee's trait level is far removed from the location of the remaining items, it is unlikely that the person's response would change the current trait estimate. Even though the desired level of precision may not have been reached for that individual, additional questions are unlikely to improve the situation, and therefore it is preferable to stop the CAT rather than ask additional questions that have little utility.

While prior work has demonstrated the effectiveness of the PSER stopping rule (Choi et al., 2010), further gains in efficiency may be possible by tuning the PSER algorithm to the demands of a particular situation. In the current paper, we build on the work of Choi et al. (2010) by exploring how the choice of parameter values impacts the accuracy and efficiency of the PSER stopping rule. Further, we provide guidance and a software tool to assist practitioners in choosing optimal values of these parameters, thereby supporting the practical implementation of the PSER method.

Item Banks for Patient Reported Outcomes

PRO constructs differ in important ways from the kinds of characteristics traditionally assessed in achievement and educational contexts. Revicki, Chen, and Tucker (2015) note that many health outcome measures are unipolar in nature, being primarily meaningful at one end of the trait continuum. For example, depression is defined by feelings of sadness, hopelessness, and worthlessness that are indicative of functional impairment. Low depression is defined not by positive affect but by the absence of these negative symptoms (Pilkonis et al., 2011).

Many health-related assessments that were initially developed for specific treatment populations are being adapted for broader use in standard patient care practice. While efforts have been made to expand the range of PRO measures in order to provide reliable scores across the trait continuum (e.g., Bruce, Fries, Lingala, Hussain, & Krishnan, 2013), many PRO measures still primarily differentiate among individuals at the negative end of the continuum and provide less information for high-functioning individuals.

An additional characteristic of PRO is that the range of potential items is relatively narrow (Revicki et al., 2015). While there might be a nearly infinite number of ways to reformulate questions assessing mathematical reasoning or reading comprehension, the pool

of potential items for constructs such as anxiety, depression or fatigue is much more limited. Consequently, item pools for PROS are often relatively small.

A consequence of these two characteristics is that a PRO item bank might contain few or no items that are informative at some trait levels. When evaluating the feasibility of a CAT in a new setting, it is useful to examine bank sensitivity in relation to the distribution of the trait in the target population.

A useful tool for evaluating bank sensitivity at various traits levels is the bank information function (i.e., the test information if all items were administered). We demonstrate this issue using the *Anxiety* bank from the Patient Reported Outcomes Measurement Information System (PROMIS) (Pilkonis et al., 2011). In Figure 1, we see that for individuals with trait values less than -0.25 , there is substantially less information than there is for individuals with values of about 0, and there is practically no information for individuals with values less than -1.0 .

This kind of non-uniform distribution of item information is not uncommon in PRO measures. Many measures of health outcomes were developed to identify individuals in need of treatment, and therefore most items differentiate primarily at the negative end of the continuum (Barthel et al., 2017; Janssen et al., 2016; Pilkonis et al., 2011; Rose, Bjorner, Becker, Fries, & Ware, 2008). Consequently, item banks that are developed to be in line with traditional measures will often have information functions similar to that depicted in Figure 1.

Figure 1 also displays a histogram of anxiety trait scores for a representative sample of the general U.S. population. We can see that much of the population exhibits low levels of anxiety, for which the bank provides little information.

Due to the mismatch between the population distribution and the bank information function, the *SE* stopping rule will not work well for individuals with negative trait values. Morris, Bass, Lee, and Neapolitan (2016) showed that for individuals with low trait levels ($\theta < 0$), the *root mean square error* (*RMSE*) can actually increase with the number of items administered. Because their trait levels fall far below the location of any item, individuals with negative θ values would consistently answer *never* to each item. As they answer more questions, they are increasingly drawn to the lowest possible score, regardless of where in the negative range their trait falls. These individuals might be asked all items in the bank before stopping, while never obtaining an accurate measure of their trait.

To further illustrate the issue, we simulated responses on the PROMIS *Anxiety* CAT for 100 examinees at each of several trait levels ranging from θ equals -4 to 4 . Figure 2 shows the average number of questions using the *SE* < 0.3 stopping rule at varying θ levels. Individuals with low trait values are indeed asked all 29 questions.

For a non-uniform item bank such as this, we need some mechanism in addition to the *SE* threshold for stopping the CAT. PROMIS currently uses the heuristic to ask a minimum of 4 and a maximum of 12 questions. However, a theory-based mechanism may work better and

be more adaptable to different banks and contexts. One alternative approach is the *predicted standard error reduction* (PSER) stopping rule (Choi et al., 2010).

Computer Adaptive Testing

Rather than administering the same items to everyone, CAT improves measurement efficiency by administering an optimal set of items tailored to each individual. A variety of item selection procedures exist (Chen, Ankenmann, & Chang, 2000). In the Bayesian approach (Choi & Swartz, 2009; van der Linden, 1998), a prior probability density function is placed on the value of the trait and is conditionally updated each time a response is provided. The standard deviation of the posterior trait distribution provides the *SE* of the trait estimate. Therefore, a useful item selection strategy is to administer the item that is expected to produce the smallest posterior standard deviation.

CAT also requires a stopping rule (Babcock & Weiss, 2009; Wainer et al., 2000). A variable-length stopping rule attempts to achieve a given level of precision for all individuals, and administers as many items as needed for each individual. The most common variable-length rule terminates the CAT when *SE* of the trait estimate falls below a threshold representing the desired level of precision. That threshold is often taken to be 0.3, which corresponds to a reliability of 0.9 (Nunnally & Bernstein, 1995).

The *SE* stopping rule can be problematic when applied to item banks with non-uniform information across the applicable trait distribution. If there are trait levels for which the bank has few informative items, an examinee might be administered many items that do not improve the trait estimate. An optimal stopping rule would end the CAT early in such cases.

The PSER Stopping Rule

Using the Bayesian approach, we place a prior distribution on the trait value θ , which is either the unit normal distribution or an empirical distribution based on population data. We discretize this distribution so that there are r distinct values of θ . If we let *Data* be the answers to the questions so far administered, the posterior variance of θ is given by

$$Var(\theta | Data) = \sum_{i=1}^r \theta_i^2 P(\theta_i | Data) - \left(\sum_{i=1}^r \theta_i P(\theta_i | Data) \right)^2. \quad (1)$$

The *SE* stopping rule terminates the CAT when the square root of this value is less than a threshold, which is ordinarily 0.3. If we let X be a random variable denoting an available item with m alternatives, the expected value of the variance, given we administer item X , is given by

$$\begin{aligned}
E_X(Var(\theta|Data)) &= \sum_{j=1}^m \left(\sum_{i=1}^r \theta_i^2 P(\theta_i|x_j, Data) - \left(\sum_{i=1}^r \theta_i P(\theta_i|x_j, Data) \right)^2 \right) P(x_j \\
&\quad | Data) \\
&= \sum_{j=1}^m \left(\sum_{i=1}^r \theta_i^2 P(x_j|\theta_i) P(\theta_i|Data) - \frac{\left(\sum_{i=1}^r \theta_i P(x_j|\theta_i) P(\theta_i|Data) \right)^2}{\sum_{i=1}^r P(x_j|\theta_i) P(\theta_i|Data)} \right).
\end{aligned} \tag{2}$$

The CAT presents the item X for which this expected value is smallest. Therefore, the expected reduction in the SE for the next item is given by

$$ExpRed = \sqrt{Var(\theta|Data)} - \min_X \sqrt{E_X(Var(\theta|Data))}, \tag{3}$$

where the minimum is taken over all items X still available. $ExpRed$ is the amount by which we can expect the SE to decrease if we present another item. If $ExpRed$ is small, we may not want to present another item even if a threshold for the SE has not been reached, while if it is large we may want to present another item even if a threshold has been reached. These considerations motivate the *predicted standard error reduction (PSER)* stopping rule (Choi et al., 2010).

The PSER rule has three parameters: *SE threshold*, *hyper*, and *hypo*. The *SE threshold* specifies the desired level of precision (e.g., $SE < 0.3$). If the expected reduction is less than *hypo*, the CAT stops even if the *SE threshold* has not been reached. In this way, the CAT will stop if there is little available information left for the subject, and pointless questions will not be asked. If the expected reduction is greater than *hyper*, the CAT will continue even if the *SE threshold* has been reached. In this way, if we have the potential to reduce the SE substantially, then the CAT will proceed.

The PSER rule can be algorithmically described as follows:

```

if  $ExpRed > hyper$  then
    keep going
elseif  $SE < threshold$  or  $ExpRed < hypo$ 
    stop
else
    keep going
endelse

```

There are three interesting special cases of this rule. First, if $hyper = \infty$, the only change from the *SE* stopping rule is that we also stop if the expected reduction is not sufficiently large. Second, if $hyper = hypo$, then the *threshold* is eliminated from the algorithm. That is,

we keep going if *ExpRed* *hyper*; otherwise we stop. Third, if *hyper* = ∞ and *hypo* = 0, the algorithm is equivalent to the *SE* stopping rule.

The *hypo* and *hyper* parameters provide a flexible mechanism to adapt the PSER algorithm to a variety of assessment contexts and item banks. It is unlikely that fixed values of these parameters will apply to all situations, because the amount of information provided by an individual item varies widely across different types of measures. Self-reported items often provide considerable information. For example, many adaptive PRO measures can achieve good precision with only about 5 items on average (Gershon & Cook, 2011), and so each item will produce a relatively large decrease in *SE*. Contrast this with items on a multiple-choice knowledge test, where even an adaptive test might require dozens of items, each of which will produce a much smaller *SE* improvement than observed in the PRO context.

Choi et al. (2010) demonstrated the usefulness of this stopping rule using one set of *hyper* and *hypo* parameters. In order to more fully explore the impact of optimal values of these parameters, we conducted a simulation of CAT precision and efficiency under a wide range of *hyper* and *hypo* levels using the PROMIS *Anxiety* and *Physical Function* item banks.

METHOD

We conducted an empirically-based Monte Carlo simulation of CAT performance using the item parameters from two operational PROMIS item banks (*Anxiety* and *Physical Function*) and the empirical distribution of these traits. Although the simulations were informed by actual CATs, we chose to simulate responses rather than using actual examinee responses. This approach has the advantage that we know the true trait levels and therefore can evaluate the accuracy of the estimates produced by the CAT. Further, the available data from the CAT development (Cella et. al 2010) and from operational use of the CATs does not provide complete responses to all possible items, making it impossible to simulate alternate stopping rules.

Measures

The PROMIS item banks were developed through a rigorous process including input from subject matter experts as well as statistical item analyses (Cella et al., 2010). The *Anxiety* and *Physical Function* banks have 29 and 121 items, respectively. Respondents were asked to report the frequency of experiences, during the past seven days, on a 5-point scale (1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always).

Calibration of graded response model (GRM) item parameters was based on a diverse sample of 21,133 respondents from the United States (Cella et al., 2010). Women comprised 55.3% of the sample. Age ranged from 18 to 89 years ($M=50.7$, $SD=18.2$). The sample was 78.1% Caucasian, 10.7% African American, 0.6% Native American, 0.6% Asian, and 10% other or mixed-race. 11.8% reported Hispanic/Latino ethnicity across all racial groups.

CAT Development

The CATs examined in the current study were built using the item parameters from the current operational versions of the *Anxiety* and *Physical Function* banks available through PROMIS. Software was developed for this study to administer the CAT. *Expected a posteriori* (EAP) estimates of θ were computed using a unit normal prior. The CAT selected items to maximize the PSER criterion using Equation 3.

Simulation Procedure

Participant trait levels were generated to match those of a representative sample of the US population, using data from the PROMIS development study (Cella et al., 2010). *Anxiety* trait values were simulated for 6212 examinees with values ranging from -1.6 to 3.5 ($M=0.01$, $SD=0.91$). *Physical Function* trait values were generated for 7863 examinees with values ranging from -3.9 to 2.2 ($M=0.04$, $SD=0.91$). These values were treated as the true latent trait values for the purpose of the simulation. For each simulated θ , potential responses to all items were generated according to the conditional probability distribution based on the IRT model.

In normal operation, a CAT will have a stopping rule, and will administer items until this rule has been reached. However, because this simulation examined a variety of stopping rules, we let the CAT administer all items in each bank. We then applied a range of stopping rules to the simulated data.

Outcome Measures

Two criteria were used to evaluate the performance of the CATs. *RMSE* was computed as a measure of accuracy. Specifically, we calculated the square root of the average squared difference between the EAP estimate and the actual θ . The second criterion was the efficiency of each CAT, operationalized as the average of the number of items presented to each examinee, which is either the number of items required to reach the stopping rule cutoff or the maximum number of items in the bank if the stopping rule was not reached.

Methods Compared

We compared the performance of two versions of the *SE* stopping rule and several versions of the PSER stopping rule. An unconstrained *SE* stopping rule (referred to as $SE < 0.3^1$) terminated the exam when the posterior standard deviation fell below 0.3 or all items in the bank were administered. We also applied the constrained *SE* stopping rule currently used in the PROMIS CATs, which terminates when $SE < 0.3$, with the proviso that at least 4 questions and no more than 12 questions are asked. For the PSER methods, we investigated the combinations of *hyper* and *hypo* values shown in Table 1 along with a *SE* threshold of 0.3.

¹Additional simulations conducted using alternate SE cutoffs ranging from 0.2 to 0.4 produced a similar pattern of results, and are not reported here to conserve space. Complete results are available on request from the first author.

RESULTS

Table 2 shows the results for *Anxiety* for the various stopping rules. The second and third columns show the average number of items presented and the *RMSE* over all simulated individuals, while the fourth and fifth columns show those values only for the individuals not targeted by the item bank (i.e., individuals with low anxiety). Figure 3 shows this same information graphically. For readability we only show the results for $hypo = 0, 0.01$, and 0.015 ; the other levels largely overlap with these values.

Figure 3a shows distinct curves associated with the three *hypo* levels, with differing points along each curve reflecting different *hyper* values. Focusing on the *RMSE* above $.25$, it is clear that $hypo = .015$ can produce scores of comparable accuracy with fewer items than the other *hypo* levels.

Large *hyper* values produced shorter, less accurate tests (i.e., the upper left end region of each curve), while smaller *hyper* values produced longer, more accurate tests (i.e., the lower right region of the curve). Differences among the levels of the *hyper* parameter reflect the inherent tradeoff between test length and precision; options that reduced *RMSE* inevitably required more items. Setting $hyper = hypo$ did not perform well, resulting in a large increase in test length relative to the *RMSE* reduction. Test length can be minimized by setting $hyper = \infty$, with only a slight reduction in accuracy; conversely, setting *hyper* only slightly above *hypo* ($+ .005$) improved accuracy with only a slightly longer test.

We see from Table 2 and Figure 3 that with $hypo = 0.015$ and $hyper = 0.025$, PSER performed as well as or better than both *SE* and PROMIS in terms of both the average number of items asked and *RMSE*. The advantage of the PSER rule is even more apparent when we consider only the θ values in the range not targeted by the item bank. For individuals with low anxiety (θ between -2 and -1), *SE* asks 23.23 questions on the average, PROMIS asks 11.17, while PSER only asks 6.16. So, the PSER has dramatically reduced the burden for these individuals.

Figure 4 shows the number of questions and *RMSE* by level of anxiety for the best-performing PSER method ($hypo = .015$, $hyper = .025$) and the PROMIS stopping rule. The accuracy of the two approaches is fairly similar across trait levels, with PSER slightly better at high anxiety levels and PROMIS slightly better at low anxiety levels. Larger differences were observed in test length. While the PSER rule asked about the same number of items across the entire trait continuum, the number of questions with the PROMIS rule was substantially higher at low anxiety. For very low anxiety levels, the PROMIS method administered nearly twice as many items with little or no improvement in accuracy.

Table 3 and Figure 5 show the results for the *Physical Function* bank for the various stopping rules. With $hypo = 0.015$ and $hyper = 0.025$, PSER obtains *RMSE* as good as (within $.002$) or better than both *SE* and PROMIS, with fewer items on average. Furthermore, when we consider only trait levels not targeted by the bank (i.e., individuals high on physical function), the performance of PSER is substantially better.

Figure 6 displays the number of items administered and *RMSE* at different levels of physical function for the PSER and PROMIS stopping rules. The two rules generally produced similar *RMSE*, but differed in the number of items administered at the extremes of the trait continuum. For very high-functioning individuals, the PROMIS stopping rule tended to administer more items (almost twice the PSER method in the worst condition). In contrast, at very low levels of physical function, PROMIS tended to ask one or two fewer questions than PSER, likely due to the PSER rule administering additional questions after the *SE* threshold had been met.

Tuning PSER Parameters

Rather than assuming that the parameter values obtained in the current study are universally applicable, we recommend that the values be tuned to a particular item bank while considering the objectives of the assessment. Although a simulation such as the one reported in this paper would be the optimal way to select *hyper* and *hypo* values, the programming involved is complex and time-consuming, making this approach impractical for many users. In this section, we describe general guidelines and provide a graphical tool for selecting parameter values that does not require extensive simulation.

Because the *hypo* parameter determines when the CAT can stop early (i.e., before the specified *SE* cutoff has been reached), it will have the greatest impact on CAT efficiency. Larger *hypo* values produce a more efficient CAT (fewer items) at the expense of slightly less precise estimates for some examinees, whereas *hypo* values close to zero favor precision over efficiency. Reasonable values for *hypo* depend on the expected *SE* change when the bank is functioning well, and will therefore likely depend on the specific parameters of the bank.

Once the optimal *hypo* value is determined, the *hyper* parameter can be selected to specify when additional questions should be asked beyond the *SE threshold*. The *hyper* value must be equal to or greater than *hypo*, with lower values producing a slightly longer, more precise test, and larger values producing a shorter, less precise test. The choice will depend on the goals of the assessment. In many contexts (e.g., assessing patient reported outcomes), the primary concern is reducing test length and minimizing the burden on examinees. In such cases, setting *hyper*= ∞ (or more practically, *hyper* = 1.0 when θ has standard normal distribution) will not allow additional questions after the *SE threshold* has been reached, and therefore will yield the most parsimonious test. Where both precision and efficiency are of equal concern, our simulations suggest that setting *hyper* about $1.67 \times hypo$ provides a good balance of precision and parsimony. In contrast, setting *hyper* at a lower value (say, *hyper* = $1.1 \times hypo$) should allow more precision with only a slightly longer test. The most precise scores would be obtained by setting *hyper* = *hypo*; however, we do not recommend this extreme value, because our simulations found that it can in some cases substantially increase the length of the CAT. Ultimately, the choice of value for *hyper* depends on the goal of the assessment and the relative importance of parsimony versus precision in a particular setting.

In order to assist in the application of the PSER stopping rule, we developed a graphical tool to assist in the selection of *hypo* and *hyper* values without conducting an extensive simulation (R code is available in the file *PSER simulator.R* at <https://>

github.com/ScottMorrisIIT/PSER). Using several types of input that will be available to many practitioners, the tool provides graphs similar to those in Figures 3 and 5, which show the precision/efficiency tradeoff at a variety of *hypo* and *hyper* values. The user can then select the configuration that best suits the specific situation.

In order to apply the PSER algorithm, we need a sequence of items administered by the CAT, a *SE* estimate after each item is administered, and the expected *SE* change for the next item. The code allows for several different types of input, which can be used to approximate these values.

At minimum, the CAT developer will have access to the item parameters for all items in a bank. Because the item characteristics determine the bank information, this will be a major determinant of the CAT performance. From the item parameters, the code computes item information and an optimal sequence of items at several θ levels. Tracing through this sequence and summing the item information as each additional item is administered, the expected *SE* from a CAT of length k can be computed from the inverse of the summed item information. The PSER algorithm can then be applied to this sequence to determine the expected *SE* and CAT length at each θ . The results are then averaged across θ to obtain the expected *SE* and CAT length for a particular hypo/hyper combination. One remaining question is what distribution to assume for θ when computing the average. If θ is expected to be approximately normal, a weighted average can be computed assuming a normal density. Alternately, if data with trait estimates are available, this can be used to obtain an empirical θ distribution to use in computing the weighted average. The code currently implements this approach using item parameters from the Graded Response Model, although the code can be readily adapted to compute item information for other IRT models.

An approximation based only on item parameters has limitations. In particular, it does not reflect the diversity of options available when implementing a CAT, such as alternate trait estimation approaches, alternate item selection criteria, item exposure controls, and content constraints. While accounting for all of these factors is beyond the scope of the current research, our code can be combined with a tailored simulation that incorporates the configuration of a particular CAT. That is, the user could simulate examinees responding to a specific CAT, saving the sequence of *SE* and expected *SE* as each item is administered. This sequence can then be used with the code we developed to evaluate the PSER algorithm with alternate *hypo/hyper* values.

Alternatively, if data from the administration of the CAT is available, this can be used to obtain the sequence of *SE* estimates after each item is administered. In most cases, the expected *SE* change on the next item will not be available (unless the CAT uses this as the item selection criterion), but we can approximate this with the actual change in *SE* from one item to the next. This approach has the advantage of evaluating the full configuration of the operational CAT with actual examinee responses. However, because the data will be limited to responses before the CAT's existing stopping rule is triggered, it is not likely to allow examination of the effect of administering additional items beyond the stopping rule (i.e.,

the *hyper* parameter). Nevertheless, it could still provide a reasonable basis for determining when the CAT should stop early (i.e., the *hypo* parameter).

DISCUSSION

The growing importance of PROs worldwide, combined with the increasing use of sophisticated psychometric methods such as CAT (Alonso et al., 2013), makes it essential that testing professionals be prepared to deal with the unique challenges of assessing health outcomes. In contrast to educational and achievement tests, PROs are often characterized by relatively small item pools and many measures were originally developed to differentiate primarily among levels of dysfunction in clinical populations (Pilkonis et al., 2011). Before using these measures in new settings, it is important to assess the alignment of the bank with the distribution of the trait in the target population. This type of analysis is relevant anytime a clinical measure is adapted for use in the general population, and becomes particularly important when implementing measures in a new country or culture (International Test Commission, 2018).

A useful tool for analyzing bank alignment is the bank information function (i.e., the test information if all items in the bank are administered). As illustrated in Figure 1, this type of graph depicts the range of trait levels at which a CAT can function effectively, as well as areas where the item bank provides insufficient information. This graph is particularly useful when paired with the distribution of the trait being assessed, as in Figure 1, in order to determine if the bank has sufficient information at common trait levels. An alternative, and possibly more intuitive approach would be to graph the expected standard error (i.e., $1/\sqrt{I(\theta)}$) in place of the information. R code for producing these graphs is available in the file *attainable information plots.R* available at <https://github.com/ScottMorrisIIT/PSER>

The alignment of item information with the trait distribution is an important element of test development and revision (Veldkamp & van der Linden, 2010). When adapting clinically-oriented tests for use in a broader population, expanding the range of item locations is often necessary, and examination of the bank information functions can provide useful information regarding areas where additional item development is needed (e.g., Bruce et al., 2013).

Expanding the item pool is not a trivial matter, requiring substantial time and resources to develop and evaluate new items. Further, many PRO constructs reflect unipolar traits that are meaningful primarily at one end of the trait continuum (Revicki et al., 2015). In such cases, the addition of items reflecting more positive states could potentially change the nature of the construct being measured (Pilkonis et al., 2011). For these reasons, practitioners need to be able to deal with existing limited item banks.

When an item bank has few informative items at a particular trait level, the CAT should terminate in order to avoid unnecessary burden on individuals for whom additional items provide little improvement in the trait estimate. Operational CATs often employ ad hoc maximum item constraints in order to ameliorate this problem. The PSER stopping rule provides a more direct and theoretically driven strategy to deal with this issue.

For the PROMIS *Anxiety* and *Physical Function* banks, our simulations showed that the PSER stopping rule with $hypo = 0.015$ and $hyper = 0.025$ outperformed the other stopping rules overall, and also when we consider only those individuals not targeted by the item bank. Importantly, the PSER stopping rule solves the problem of excessive test length for individuals not targeted by the item bank (as shown in Figures 4 and 6).

Because the *hypo* rule comes into play only when test information is low, differences between the PSER and SE stopping rules occur primarily at the extremes (i.e., low levels of anxiety and high levels of physical function). This is an attractive feature of the PSER rule – for individuals with moderate trait levels, where the bank has ample information, the procedure retains the benefits of the SE threshold stopping rule, guaranteeing that all examinees are measured with equal precision. It is only at the extremes, where this level of precision cannot be attained, that there is a need to stop the bank early.

Although the PSER stopping rule is an effective way of minimizing assessment burden for individuals who are poorly aligned with the item bank, it is not the only solution to this problem. One alternative approach would be to determine from preliminary analysis the attainable *SE* at each level of the trait, and to set a commensurate SE threshold that depends on the examinee's θ estimate (Riley, Conrad, Bezruczko, & Dennis, 2007). Further, in situations where the CAT is used to make a decision (e.g., passing or failing a licensure exam, or flagging a patient as needing further assessment due to a high Anxiety score), another option would be to stop the CAT when there is sufficient confidence in the classification decision (Bergstrom & Lunz, 1992; Gnamb & Batinic, 2011). Thus, the stopping rule might apply a relatively strict threshold (e.g., $SE < .2$) for theta values near the decision point, but less demanding threshold (e.g., $SE < .4$) for theta values far from the critical region.

These alternative approaches will necessarily be somewhat ad hoc, relying heavily on the expert judgment of the CAT developer to identify a fixed configuration of stopping rules. In contrast, by embedding the guiding principle into the stopping rule, the PSER approach minimizes judgment calls in the selection of trait ranges and varying threshold values, and ensures optimal flexibility in matching the stopping rule to each individual. Nevertheless, approaches based on varying *SE* thresholds would avoid some of the computational complexity of the PSER method (i.e., the need to compute expected *SE* change after each item), and might accomplish much of the improvement in efficiency with less computational overhead.

Our results demonstrate that the performance of the PSER algorithm depends on finding the optimal values of the *hypo* and *hyper* parameters. We note that the same values of *hypo* and *hyper* yielded similar superior performance for both the *Anxiety* and *Physical Function* banks. These banks are quite distinct in terms of the number of items and the shape of the bank information function: the 121-item *Physical Function* bank targets low trait levels, while the 29-item *Anxiety* bank targets high trait levels. Furthermore, as can be seen from Tables 2 and 3, when $hyper = \infty$, all investigated values of *hypo* yield superior performance over the standard SE stopping rule. This result moderately supports that the performance of the PSER is robust relative to the choice of *hypo* for at least PROs.

Nevertheless, we do not expect that these same values of *hypo* and *hyper* will generalize to all item banks and assessment contexts. CATs differ in a variety of ways that are likely to impact the optimal PSER configuration, including item characteristics (e.g., number of response items, item information), the degree of mismatch between the bank information and the distribution of the trait in the target population, as well as operational constraints on the CAT (e.g., item exposure control, enemy items, content constraints).

The choice of parameters may also be influenced by the purpose of the assessment. Although we concentrated on the values of *hypo* = 0.015 and *hyper* = 0.025 for the purpose of comparison, they are not necessarily the values a practitioner may want to use. As we can see in Figures 3 and 5, by varying the values of *hyper* and *hypo*, we obtain a tradeoff between *RMSE* and number of questions asked. One strength of the PSER stopping rule is that the user can vary these parameters to optimize the precision/efficiency tradeoff in light of the goals of the assessment, opting for a slightly longer measure if greater precision is desired, or accepting slightly less precision if reducing respondent burden is of greater importance. However, this flexibility also necessitates a method to choose the appropriate *hyper* and *hypo* parameters for a particular setting. The R code described in the preceding section provides a graphical tool to assist in tuning the PSER stopping rule for any particular bank, and thereby supports practical implementation of the method.

Acknowledgments

Authors Note

This work was supported by National Library of Medicine grants R01LM011962 and R01LM011663.

REFERENCES

- Alonso J, Bartlett SJ, Rose M, Aaronson NK, Chaplin JE, Efficace F, ... for the PROMIS International group. (2013). The case for an international patient-reported outcomes measurement information system (PROMIS®) initiative. *Health and Quality of Life Outcomes*, 11(1), 210. 10.1186/1477-7525-11-210 [PubMed: 24359143]
- Apers S, Kovacs AH, Luyckx K, Thomet C, Budts W, Enomoto J, ... Isachd, on behalf of the A.-I. consortium and. (2016). Quality of Life of Adults With Congenital Heart Disease in 15 Countries: Evaluating Country-Specific Characteristics. *Journal of the American College of Cardiology*, 67(19), 2237–2245. 10.1016/j.jacc.2016.03.477 [PubMed: 27173035]
- Babcock B, & Weiss DJ (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Presented at the GMAC Conference on Computerized Adaptive Testing.
- Barthel D, Otto C, Nolte S, Meyrose A-K, Fischer F, Devine J, ... Ravens-Sieberger U (2017). The validation of a computer-adaptive test (CAT) for assessing health-related quality of life in children and adolescents in a clinical sample: study design, methods and first results of the Kids-CAT study. *Quality of Life Research*, 26(5), 1105–1117. 10.1007/s11136-016-1437-9 [PubMed: 27830512]
- Bergstrom BA, & Lunz ME (1992). Confidence in Pass/Fail Decisions for Computer Adaptive and Paper and Pencil Examinations. *Evaluation & the Health Professions*, 15(4), 453–464. 10.1177/016327879201500407
- Bruce B, Fries J, Lingala B, Hussain YN, & Krishnan E (2013). Development and assessment of floor and ceiling items for the PROMIS physical function item bank. *Arthritis Research & Therapy*, 15(5), 1.

- Castrejón I, Gossec L, & Carmona L (2015). The EULAR Outcome Measures Library: an evolutionary database of validated patient-reported instruments. *Annals of the Rheumatic Diseases*, 74(2), 475–476. 10.1136/annrheumdis-2014-206485 [PubMed: 25452307]
- Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, ... Hays R (2010). Initial adult health item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS™) network: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179–1194. 10.1016/j.jclinepi.2010.04.011 [PubMed: 20685078]
- Chen S-Y, Ankenmann RD, & Chang H-H (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, 24(3), 241–255. 10.1177/01466210022031705
- Choi SW, & Swartz RJ (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, 33(6), 419–440. 10.1177/0146621608327801 [PubMed: 20011456]
- Choi Seung W., Grady MW, & Dodd BG (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, 70(6), 1–17. 10.1177/0013164410387338 [PubMed: 21278821]
- Falavigna A, Dozza DC, Teles AR, Wong CC, Barbagallo G, Brodke D, ... Riew KD (2017). Current Status of Worldwide Use of Patient-Reported Outcome Measures (PROMs) in Spine Care. *World Neurosurgery*, 108, 328–335. 10.1016/j.wneu.2017.09.002 [PubMed: 28893693]
- Gershon RC, & Cook K (2011). Use of computer adaptive testing in the development of machine learning algorithms. *Pain Medicine (Malden, Mass.)*, 12(10), 1450–1452. 10.1111/j.1526-4637.2011.01235.x
- Gibbons C, Bower P, Lovell K, Valderas J, & Skevington S (2016). Electronic Quality of Life Assessment Using Computer-Adaptive Testing. *Journal of Medical Internet Research*, 18(9), e240. 10.2196/jmir.6053 [PubMed: 27694100]
- Gnambs T, & Batinic B (2011). Polytomous Adaptive Classification Testing: Effects of Item Pool Size, Test Termination Criterion, and Number of Cutscores. *Educational and Psychological Measurement*, 71(6), 1006–1022. 10.1177/0013164410393956
- International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). Retrieved from https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf
- International Test Commission. (2018). ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations. Retrieved from https://www.intestcom.org/files/guideline_diverse_populations.pdf
- Janssen SJ, Paulino Pereira NR, Raskin KA, Ferrone ML, Hornicek FJ, van Dijk CN, ... Schwab JH (2016). A comparison of questionnaires for assessing physical function in patients with lower extremity bone metastases: Physical Function in Bone Metastases. *Journal of Surgical Oncology*. 10.1002/jso.24400
- Melican GJ, Breithaupt K, & Zhang Y (2010). Designing and Implementing a Multistage Adaptive Test: The Uniform CPA Exam. In van der Linden WJ & Glas CAW (Eds.), *Elements of Adaptive Testing* (pp. 167–189). 10.1007/978-0-387-85461-8_9
- Morris S, Bass M, Lee M, & Neapolitan RE (2016). Advancing the efficiency and efficacy of patient reported outcomes with multivariate computer adaptive testing. *JAMA*.
- Nunnally JC, & Bernstein IH (1995). *Psychometric theory*. New York, NY: McGraw-Hill.
- Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, & PROMIS Cooperative Group. (2011). Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS(R)): Depression, anxiety, and anger. *Assessment*, 18(3), 263–283. 10.1177/1073191111411667 [PubMed: 21697139]
- Revicki DA, Chen W-H, & Tucker C (2015). Developing item banks for patient-reported health outcomes. In Reise SP & Revicki DA (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. (pp. 334–363). Routledge.
- Riley BB, Conrad KJ, Bezruczko N, & Dennis ML (2007). Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: the GAIN substance problem scale. *Journal of Applied Measurement*, 8(1), 48–64. [PubMed: 17215565]

- Rose M, Bjorner JB, Becker J, Fries JF, & Ware JE (2008). Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS). *Journal of Clinical Epidemiology*, 61(1), 17–33. 10.1016/j.jclinepi.2006.06.025 [PubMed: 18083459]
- Rubin RR, Peyrot M, & Siminerio LM (2006). Health Care and Patient-Reported Outcomes: Results of the cross-national Diabetes Attitudes, Wishes and Needs (DAWN) study. *Diabetes Care*, 29(6), 1249–1255. 10.2337/dc05-2494 [PubMed: 16732004]
- Rudner LM (2010). Implementing the Graduate Management Admission Test Computerized Adaptive Test. In van der Linden WJ & Glas CAW (Eds.), *Elements of Adaptive Testing* (pp. 151–165). 10.1007/978-0-387-85461-8_8
- Sands WA, & Waters BK (1997). Introduction to ASVAB and CAT. In Sands WA, Waters BK, & McBride JR (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 3–9). Washington, DC, US: American Psychological Association.
- Shimazu A, Schaufeli WB, Miyanaka D, & Iwata N (2010). Why Japanese workers show low work engagement: An item response theory analysis of the Utrecht Work Engagement scale. *BioPsychoSocial Medicine*, 4, 17. 10.1186/1751-0759-4-17 [PubMed: 21054839]
- Snyder CF, Jensen RE, Segal JB, & Wu AW (2013). Patient-reported outcomes (PROs): Putting the patient perspective in patient-centered outcomes research. *Medical Care*, 51(8 0 3), S73–S79. 10.1097/MLR.0b013e31829b1d84 [PubMed: 23774513]
- Stark S, Chernyshenko OS, Drasgow F, & White LA (2012). Adaptive Testing With Multidimensional Pairwise Preference Items: Improving the Efficiency of Personality and Other Noncognitive Assessments. *Organizational Research Methods*, 15(3), 463–487. 10.1177/1094428112444611
- van der Linden WJ (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201–216.
- Veldkamp BP, & van der Linden WJ (2010). Designing Item Pools for Adaptive Testing. In van der Linden WJ & Glas CAW (Eds.), *Elements of Adaptive Testing* (pp. 231–245). 10.1007/978-0-387-85461-8_12
- Vilagut G, Forero CG, Adroher ND, Olariu E, Cella D, & Alonso J (2015). Testing the PROMIS® Depression measures for monitoring depression in a clinical sample outside the US. *Journal of Psychiatric Research*, 68, 140–150. 10.1016/j.jpsychires.2015.06.009 [PubMed: 26228413]
- Wainer H, Dorans NJ, Flaugher R, Green BF, & Mislevy RJ (2000). *Computerized Adaptive Testing: A Primer - 2nd Edition* (2nd ed.). Retrieved from <https://www.questia.com/library/1822386/computerized-adaptive-testing-a-primer>

Figure 1.
Maximum Attainable Test Information (Solid Line) and Trait Frequency (Histogram) for the
PROMIS *Anxiety* Trait.

Figure 2.
Expected Number of Items Required to Reach the $SE < 0.3$ Stopping Rule (Solid Line) and Full Bank Fisher Information (Dashed Line) at Each Level of θ for the *Anxiety* Item Bank.

Figure 3.

RMSE vs. Average Number of Items Presented for the *Anxiety* Trait for the Full Range of the Trait (a) and for Trait Levels Underrepresented in the Item Bank (b).

Figure 4. Average Number of Items Administered (Solid Line) and RMSE (Dotted Line) for Each Value of θ for the *Anxiety* Trait Using the PROMIS and PSER Stopping Rules.

Figure 5.

RMSE vs. Average Number of Items Presented for *Physical Function* for the Full Range of the Trait (a) and for Trait Levels Underrepresented by the Bank (b).

Figure 6. Average Number of Items Administered (Solid Line) and RMSE (Dotted Line) for Each Value of θ for the *Physical Function* Bank Using the PROMIS and PSER Stopping Rules.

Table 1

The Combinations of Values of Hypo and Hyper Used in the PSER in the Simulation Study

hypo		Hyper					
0	-	.005	.01	.015	.02	∞	
.005	.005	.01	.015	.02	.025	∞	
.01	.01	.015	.02	.025	.03	∞	
.015	.015	.02	.025	.03	.035	∞	
.02	.02	.025	.03	.035	.04	∞	
.025	.025	.03	.035	.04	.045	∞	
.03	.03	.035	.04	.045	.05	∞	

Table 2

Average Number of Items Presented and RMSE for the Anxiety Bank

Method	Avg. # Items All θ	RMSE ALL θ	Avg. # Items $\theta < -1$	RMSE $\theta < -1$
PROMIS	5.88	0.279	11.17	0.325
SE=0.3	7.9	0.286	23.23	0.341
SE=0.3: Hypo=0: Hyper=0.005	13.93	0.218	23.94	0.316
SE=0.3: Hypo=0: Hyper=0.01	13.93	0.218	23.94	0.316
SE=0.3: Hypo=0: Hyper=0.015	9.76	0.248	23.35	0.331
SE=0.3: Hypo=0: Hyper=0.02	9.76	0.248	23.35	0.331
SE=0.3: Hypo=0: Hyper=Infinity	7.9	0.286	23.23	0.341
SE=0.3: Hypo=0.005: Hyper=0.005	12.04	0.219	13.7	0.308
SE=0.3: Hypo=0.005: Hyper=0.01	12.04	0.219	13.7	0.308
SE=0.3: Hypo=0.005: Hyper=0.015	7.87	0.249	13.12	0.324
SE=0.3: Hypo=0.005: Hyper=0.02	7.87	0.249	13.12	0.324
SE=0.3: Hypo=0.005: Hyper=0.025	6.51	0.271	13	0.334
SE=0.3: Hypo=0.005: Hyper=Infinity	6.01	0.287	13	0.335
SE=0.3: Hypo=0.01: Hyper=0.01	12.04	0.219	13.7	0.308
SE=0.3: Hypo=0.01: Hyper=0.015	7.87	0.249	13.12	0.324
SE=0.3: Hypo=0.01: Hyper=0.02	7.87	0.249	13.12	0.324
SE=0.3: Hypo=0.01: Hyper=0.025	6.51	0.271	13	0.334
SE=0.3: Hypo=0.01: Hyper=0.03	6.51	0.271	13	0.334
SE=0.3: Hypo=0.01: Hyper=Infinity	6.01	0.287	13	0.335
SE=0.3: Hypo=0.015: Hyper=0.015	6.42	0.257	6.28	0.301
SE=0.3: Hypo=0.015: Hyper=0.02	5.06	0.279	6.16	0.312
SE=0.3: Hypo=0.015: Hyper=0.025	5.06	0.279	6.16	0.312
SE=0.3: Hypo=0.015: Hyper=0.03	4.66	0.291	6.16	0.313
SE=0.3: Hypo=0.015: Hyper=0.035	4.66	0.291	6.16	0.313
SE=0.3: Hypo=0.015: Hyper=Infinity	4.57	0.294	6.16	0.313
SE=0.3: Hypo=0.02: Hyper=0.02	5.06	0.279	6.16	0.312
SE=0.3: Hypo=0.02: Hyper=0.025	5.06	0.279	6.16	0.312
SE=0.3: Hypo=0.02: Hyper=0.03	4.66	0.291	6.16	0.313
SE=0.3: Hypo=0.02: Hyper=0.035	4.66	0.291	6.16	0.313
SE=0.3: Hypo=0.02: Hyper=0.04	4.66	0.291	6.16	0.313
SE=0.3: Hypo=0.02: Hyper=Infinity	4.57	0.294	6.16	0.313
SE=0.3: Hypo=0.025: Hyper=0.025	4.69	0.288	4.94	0.323
SE=0.3: Hypo=0.025: Hyper=0.03	4.29	0.299	4.94	0.324
SE=0.3: Hypo=0.025: Hyper=0.035	4.29	0.299	4.94	0.324
SE=0.3: Hypo=0.025: Hyper=0.04	4.29	0.299	4.94	0.324
SE=0.3: Hypo=0.025: Hyper=0.045	4.2	0.302	4.94	0.324
SE=0.3: Hypo=0.025: Hyper=Infinity	4.2	0.302	4.94	0.324
SE=0.3: Hypo=0.03: Hyper=0.03	4.29	0.299	4.94	0.324
SE=0.3: Hypo=0.03: Hyper=0.035	4.29	0.299	4.94	0.324

Method	Avg. # Items All θ	RMSE ALL θ	Avg. # Items $\theta < -1$	RMSE $\theta < -1$
SE=0.3; Hypo=0.03; Hyper=0.04	4.29	0.299	4.94	0.324
SE=0.3; Hypo=0.03; Hyper=0.045	4.2	0.302	4.94	0.324
SE=0.3; Hypo=0.03; Hyper=0.05	4.2	0.302	4.94	0.324
SE=0.3; Hypo=0.03; Hyper=Infinity	4.2	0.302	4.94	0.324

Table 3

Average Number of Items Presented and RMSE for the Physical Function Bank

Method	Avg. # Items All θ	RMSE ALL θ	Avg. # Items $\theta > 1$	RMSE $\theta > 1$
PROMIS	5.08	0.25	8.72	0.332
SE=0.3	7.47	0.273	29.88	0.33
SE=0.3: Hypo=0: Hyper=0.005	14.52	0.185	32.89	0.272
SE=0.3: Hypo=0: Hyper=0.01	14.52	0.185	32.89	0.272
SE=0.3: Hypo=0: Hyper=0.015	9.57	0.222	30.29	0.306
SE=0.3: Hypo=0: Hyper=0.02	9.57	0.222	30.29	0.306
SE=0.3: Hypo=0: Hyper=Infinity	7.47	0.273	29.88	0.33
SE=0.3: Hypo=0.005: Hyper=0.005	11.17	0.187	10.7	0.277
SE=0.3: Hypo=0.005: Hyper=0.01	11.17	0.187	10.7	0.277
SE=0.3: Hypo=0.005: Hyper=0.015	6.21	0.223	8.11	0.311
SE=0.3: Hypo=0.005: Hyper=0.02	6.21	0.223	8.11	0.311
SE=0.3: Hypo=0.005: Hyper=0.025	4.97	0.247	7.72	0.33
SE=0.3: Hypo=0.005: Hyper=Infinity	4.12	0.274	7.7	0.335
SE=0.3: Hypo=0.01: Hyper=0.01	11.17	0.187	10.7	0.277
SE=0.3: Hypo=0.01: Hyper=0.015	6.21	0.223	8.11	0.311
SE=0.3: Hypo=0.01: Hyper=0.02	6.21	0.223	8.11	0.311
SE=0.3: Hypo=0.01: Hyper=0.025	4.97	0.247	7.72	0.33
SE=0.3: Hypo=0.01: Hyper=0.03	4.97	0.247	7.72	0.33
SE=0.3: Hypo=0.01: Hyper=Infinity	4.12	0.274	7.7	0.335
SE=0.3: Hypo=0.015: Hyper=0.015	5.89	0.229	6.29	0.314
SE=0.3: Hypo=0.015: Hyper=0.02	4.66	0.252	5.91	0.332
SE=0.3: Hypo=0.015: Hyper=0.025	4.66	0.252	5.91	0.332
SE=0.3: Hypo=0.015: Hyper=0.03	4.21	0.263	5.89	0.336
SE=0.3: Hypo=0.015: Hyper=0.035	4.21	0.263	5.89	0.336
SE=0.3: Hypo=0.015: Hyper=Infinity	3.8	0.279	5.89	0.337
SE=0.3: Hypo=0.02: Hyper=0.02	4.66	0.252	5.91	0.332
SE=0.3: Hypo=0.02: Hyper=0.025	4.66	0.252	5.91	0.332
SE=0.3: Hypo=0.02: Hyper=0.03	4.21	0.263	5.89	0.336
SE=0.3: Hypo=0.02: Hyper=0.035	4.21	0.263	5.89	0.336
SE=0.3: Hypo=0.02: Hyper=0.04	4.21	0.263	5.89	0.336
SE=0.3: Hypo=0.02: Hyper=Infinity	3.8	0.279	5.89	0.337
SE=0.3: Hypo=0.025: Hyper=0.025	4.4	0.256	5.05	0.34
SE=0.3: Hypo=0.025: Hyper=0.03	3.95	0.267	5.04	0.344
SE=0.3: Hypo=0.025: Hyper=0.035	3.95	0.267	5.04	0.344
SE=0.3: Hypo=0.025: Hyper=0.04	3.95	0.267	5.04	0.344
SE=0.3: Hypo=0.025: Hyper=0.045	3.68	0.276	5.03	0.345
SE=0.3: Hypo=0.025: Hyper=Infinity	3.54	0.282	5.03	0.345
SE=0.3: Hypo=0.03: Hyper=0.03	3.95	0.267	5.04	0.344
SE=0.3: Hypo=0.03: Hyper=0.035	3.95	0.267	5.04	0.344

Method	Avg. # Items All θ	RMSE ALL θ	Avg. # Items $\theta > 1$	RMSE $\theta > 1$
SE=0.3; Hypo=0.03; Hyper=0.04	3.95	0.267	5.04	0.344
SE=0.3; Hypo=0.03; Hyper=0.045	3.68	0.276	5.03	0.345
SE=0.3; Hypo=0.03; Hyper=0.05	3.68	0.276	5.03	0.345
SE=0.3; Hypo=0.03; Hyper=Infinity	3.54	0.282	5.03	0.345