NIH Public Access
Author Manuscript
Educ Psychol Meas. Author manuscript; available in PMC 2011 January 27.

Published in final edited form as:
Educ Psychol Meas. 2010 December 1; 70(6): 1–17. doi:10.1177/0013164410387338.

A New Stopping Rule for Computerized Adaptive Testing

Seung W. Choi,
Northwestern University Feinberg School of Medicine

Matthew W. Grady, and
University of Texas at Austin

Barbara G. Dodd
University of Texas at Austin

Abstract

The goal of the current study was to introduce a new stopping rule for computerized adaptive testing. The predicted standard error reduction stopping rule (PSER) uses the predictive posterior variance to determine the reduction in standard error that would result from the administration of additional items. The performance of the PSER was compared to that of the minimum standard error stopping rule and a modified version of the minimum information stopping rule in a series of simulated adaptive tests, drawn from a number of item pools. Results indicate that the PSER makes efficient use of CAT item pools, administering fewer items when predictive gains in information are small and increasing measurement precision when information is abundant.

Keywords

Computerized Adaptive Testing (CAT); Stopping Rule; Testing Burden

According to Wainer (2000) an adaptive test can be considered complete after a predetermined number of items have been administered, when a predetermined level of measurement precision has been reached, or when a predetermined length of time has elapsed. The two most commonly used methods for determining when a computerized adaptive test is complete are the fixed length and variable length stopping rules.

Under a fixed length stopping rule, an adaptive test is terminated when a predetermined number of items have been administered. Accordingly, all examinees are administered the same number of items, regardless of the degree of measurement precision achieved upon termination of the test. The primary advantage of the fixed length stopping rule is its simplicity. However, one consequence of the implementation of the fixed length stopping rule is that examinees will be measured with different degrees of precision, with larger measurement error typically occurring at extreme trait levels. Additionally, a fixed length stopping rule may limit the efficiency of an adaptive test through the unnecessary administration of items that contribute little information about examinee trait level.

In contrast, variable length stopping rules typically seek to achieve a certain degree of measurement precision for all examinees, even when doing so means that some examinees are given more items than others. Two types of variable length stopping rules have been

Send correspondence to: Seung W. Choi, s-choi@northwestern.edu, Northwestern University Feinberg School of Medicine, Department of Medical Social Sciences, 710 N. Lake Shore Dr., Chicago, IL 60611, USA, Phone: (312) 503-9803, Fax: (312) 503-9800.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

used (Dodd, Koch, & De Ayala, 1993). These are the standard error (SE) stopping rule and the minimum information stopping rule. Of these, the most commonly used has been the SE stopping rule, which terminates an adaptive test when a predetermined standard error has been reached for the most recent examinee trait estimate (Boyd, Dodd, & Choi, 2010). One advantage of the SE stopping rule is that, when the item pool information function if relatively flat, it typically yields near equivalent measurement precision across the examinee trait continuum. A disadvantage of the SE stopping rule is that it may limit the efficiency of an adaptive test by unnecessarily administering additional items to examinees for which a predetermined standard error criterion cannot be met. This negates the trademark efficiency of the adaptive test and results in an unnecessary testing burden for examinees, which may be especially undesirable when the adaptive test serves as a brief screening device, as is the case for some medical and psychiatric CAT applications (Gardner et al., 2004). A second potential disadvantage of the SE stopping rule is that it may limit measurement precision by terminating the adaptive test, even though informative items are still available for administration. In health outcomes research a SE stopping rule that is conditional on the trait estimate has been used because there is a desire to measure individuals with severe medical problems more accurately than those with less severe symptoms (Ware et al., 2000, 2003; Ware, Gandek, Sinclair, & Bjorner, 2005).

Under the minimum information stopping rule, the CAT is terminated when there are no more available items capable of providing a predetermined minimum level of information for an examinee at the most recent trait estimate. This approach attempts to prevent the administration of items that contribute little information about the examinee trait level. The primary strength of the minimum information stopping rule is that it tends against the needless administration of items to examinees for which a high degree of measurement precision is not possible. A weakness of the minimum information stopping rule is that it typically delivers less accurate measurement precision than the minimum standard error stopping rule, due to its reluctance to administer additional items when information is relatively scant. Dodd, Koch, and De Ayala (1989) compared the minimum SE and minimum information stopping rule approaches in simulated adaptive tests based on the graded response model. Results indicated that the SE stopping rule yielded higher correlations between known and estimated abilities, while also producing considerably fewer floor/ceiling cases. Additionally, Dodd, Koch, and De Ayala (1993) found that, for adaptive test scored with the partial credit model, the SE stopping rule produced fewer floor/ceiling cases, while also resulting in shorter tests with greater measurement precision than the minimum information stopping rule.

Whether they are fixed or variable length, all stopping rules attempt to yield adaptive tests that are both accurate and efficient. However, the difficult balance between measurement precision and testing efficiency can subject existing stopping rules to two potential problems that are of particular interest in the current study. The first of these occurs when an adaptive test is stopped, even though desired gains in information would result from the administration of additional items. In this scenario, measurement precision is unwittingly sacrificed in favor of a shorter test. Both the fixed length and minimum SE stopping rules are prone to this problem because their respective stopping criteria do not consider potential gains in measurement precision that would result from the administration of more items. This is problematic because, in some cases, meaningful gains in measurement precision may be both desirable and possible with the administration of only one or two additional items.

A second potential problem for existing stopping rules occurs when items are unnecessarily administered to examinees for which more precise trait estimates are unlikely. Both the fixed length and minimum SE stopping rules are prone to this problem. This can be especially problematic when the match between the item pool information function and the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

examinee trait distribution is poor, and thus, high degrees of measurement precision are unachievable for large numbers of examinees. The minimum information stopping rule is also subject to the efficiency problem, albeit in a different way. Because it is unguarded against administering excessive amounts of items to examinees for which informative items are abundant, the implementation of the minimum information rule may limit the efficiency of the adaptive test when the match between the item pool information function and the examinee trait distribution is good. To prevent this, the minimum information stopping rule may be combined with the minimum SE stopping rule. However, combining these approaches will limit the ability of the minimum information stopping rule to decrease standard error beyond the minimum limit imposed by the SE component.

Figure 1 can be used to illustrate potential inefficiencies of the minimum SE and fixed length stopping rules. This figure displays the maximum attainable information over the trait continuum, for the 28-item PROMIS depressive symptoms pool on which a portion of the current study was based (Choi et al., 2010). The vertical bars above the horizontal axis show the locations of the category threshold parameters across all items. The outermost curve represents the item pool information function, while the inner curves represent the maximum information that would result from the administration of a given number of items. Note that, at a theta level of −1, the maximum attainable information peaks near 10, regardless of the number of items administered. In this situation, both the minimum SE and fixed length stopping rules would administer high numbers of items, even though doing so would result in diminishingly small gains in measurement precision. Figure 1 also demonstrates the potential sacrifices in measurement precision that result from the use of the minimum SE stopping rule. For instance, at a theta level of 1, the minimum SE stopping rule would terminate the test after administering only a few items, even though substantial gains in measurement precision could be achieved through the administration of only one or two additional items.

Figure 2-a further illustrates the aforementioned efficiency problem associated with minimum SE and fixed length stopping rules, by displaying the minimum attainable standard error as a function of test length, at five trait levels for the same 28 item pool. Note that, for a theta level of −2, the minimum attainable standard error is above that which would typically be required to terminate the test under the minimum SE stopping rule. For this trait level, the minimum SE stopping rule would administer all 28 items in the pool, even though the desired standard error is unobtainable. A fixed length stopping rule would also administer the maximum number of items to examinees at this trait level, resulting in an unnecessarily long test. Figure 2-a also illustrates the potential sacrifices in measurement precision that result from the use of the minimum SE stopping rule. For instance, at a theta level of 1, the minimum SE stopping rule would terminate the test after administering less than five items, even though the favorable match between the examinee trait level and the item pool would allow for a considerable further reduction in standard error.

Figures 1 and 2-a suggest that the potential problems inherent in both fixed length and minimum SE stopping rules stem from the fact that both rules are relatively insensitive to the relationship between examinee trait level and the item pool information function. For instance, by considering only the final standard error achieved, the SE rule fails to acknowledge that, for some examinees, the amount of information provided by the item pool prohibits a high degree of measurement precision. Though minimum information stopping rules are sensitive to potential changes in measurement precision, using them in combination with a standard error rule can limit their ability to decrease standard error beyond the minimum limit imposed by the SE component. Accordingly, it seems that a more sensitive approach to stopping an adaptive test would involve a more systematic evaluation of the potential change in measurement precision that would result from the administration of

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

additional items. Figure 2-b illustrates this point by displaying the change in standard error as a function of test length, at five points along the examinee trait level, for the 28 item pool used in the previous examples. Note that the decreases in standard error that result from the administration of additional items are similar for each level of theta. This suggests that the size of the predicted change in measurement precision may be a more stable criterion for terminating a test than would be the overall standard error, which may vary considerably across the ability continuum.

For a number of CAT applications, including medical and psychiatric assessments, there is strong motivation to reduce testing burden. For these applications, where adaptive tests are typically short and testing efficiency is paramount, the needless administration of items is especially problematic (Gibbons, et al., 2008; Haley, et al., 2008). Additionally, medical and psychiatric CATs are typically designed to target a more narrow range of trait levels than most achievement or aptitude tests. For examinees that fall outside the targeted trait level range, a high degree of measurement precision is either not necessary or untenable. For examinees whose trait levels fall within the targeted range, a higher degree of measurement precision is desirable. For these CAT applications, the ideal stopping rule would facilitate a high degree of measurement precision for examinees whose trait levels fall within the targeted range (for which the item pool is designed), and limit testing burden for those examinees whose trait levels fall outside the targeted range. The stopping rule introduced in the current study is designed for use in such situations and will be subsequently described in greater detail.

The current study proposes a new stopping rule that attempts to address the efficiency and measurement precision problems associated with commonly used stopping rules. The new stopping rule, called the predicted standard error reduction stopping rule, (PSER) seeks to balance the dual concerns of measurement precision and testing efficiency by considering the predicted change in measurement precision that would result from the administration of additional items. This is accomplished by evaluating the predictive posterior distribution and the expected incremental improvement in precision that would result from subsequent items administrations. Under this approach, an adaptive test is terminated when the predicted gain in measurement precision brought on by the administration of an additional item is below a predetermined criterion. Alternatively, the adaptive test is allowed to continue when the predicted gain in measurement precision brought on by the administration of an additional item is greater than the preset criterion. The purpose of the research is to explore the properties of the PSER and assess its performance with regard to the number of items administered and the accuracy of theta estimation in comparison with existing stopping rules.

Method

Stopping Criteria

The PSER stopping rule is based on the predictive posterior variance of theta, which is also the objective function of the minimum expected posterior variance (MEPV) item selection criterion (Owen, 1975; Thissen & Mislevy, 2000; van der Linden & Pashley, 2000). The MEPV (or equivalently maximum expected posterior precision) criterion can be expressed as follows:

$$i_k \equiv \arg\min_j \left\{ \sum_{r=1}^{m_j} p_j(r|u_{i_1}, u_{i_2}, \cdots, u_{i_{k-1}}) \mathrm{Var}(\theta|u_{i_1}, u_{i_2}, \cdots, u_{i_{k-1}}, U_j=r): j \in R_k \right\},$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

where $p_j(r|u_{i1},u_{i2},\ldots,u_{ik-1})$ denote the posterior predictive distribution (van der Linden & Pashley, 2000), which is the probability of giving response r to item j with m_j categories given the previous response history, and $\mathrm{Var}(\theta|u_{i1},u_{i2},\ldots,u_{ik-1},U_j = r)$ is the posterior variance for item j with predicted response category r. This objective function is the Bayesian counterpart of the maximum information selection criterion (Thissen & Mislevy, 2000).

In conjunction with the minimum SE stopping rule, the PSER defines two additional threshold parameters. For the current study, these are referred to as the hyper and hypo parameters. The hyper parameter determines a minimum expected reduction in standard error. If satisfied the "stop" cue called by the minimum SE stopping rule will be overridden and the adaptive test will continue. The hypo parameter also specifies a minimum expected reduction in standard error. If unsatisfied the "continue" cue called by the SE stopping rule will be overridden and the adaptive test will terminate. Suppose that the hyper and hypo parameters are set equal to 0.03 and 0.01, respectively. When the standard error associated with the current theta estimate falls below the target stopping threshold, the minimum SE stopping rule calls a stop cue. Then the PSER assesses if any remaining item in the pool is expected to reduce the standard error by 0.03 or more. An affirmative answer overrides the stop cue and continues the CAT. Likewise, when a continue cue is called because the minimum standard error threshold is unsatisfied, the PSER assesses if any remaining item in the pool is expected to reduce the standard error by at least 0.01. A negative answer will override the continue cue and the CAT will be stopped.

The performance of the PSER stopping rule was compared to two existing procedures: the minimum SE stopping rule and the minimum information stopping rule. The minimum information stopping rule was modified and combined with the minimum SE stopping rule such that it terminates an adaptive test either when no item remaining in the pool can provide a predetermined minimum level of information or when a predetermined standard error criterion has been satisfied. The purpose of this modification was to avoid administering additional items beyond achieving a predetermined level of measurement precision. Thus the minimum information threshold is used to assess an exit condition without meeting a standard error threshold but is not used to sustain beyond achieving the standard error threshold. While the minimum SE stopping rule was included in the current study primarily to provide a basis of comparison for the PSER stopping rule in terms of measurement precision, the minimum information stopping rule was included to provide a basis of comparison for the PSER stopping rule in terms of the number of items administered. It is important to note that the primary purpose of the current study was not to compare the minimum SE and minimum information rules on measurement precision or number of items administered

Construction of Item Pools

Item pools with 30 to 60 items have been found adequate for successful CAT procedures with polytomous items (Dodd, Koch, & De Ayala, 1989; Dodd, Koch, & De Ayala, 1993). For the current study, two real pools of 28 and 95 items, respectively, served as models to generate small (30) and large (90) item pools. Real item pools were based on actual items from a real medical CAT item pool. All items had five response categories and calibrated according to the graded response model (Samejima, 1969). Two types of item pool distributions were also generated: matched vs. mismatched. The item pool information function for the 28-item pool was shifted markedly (+1.5) to the right in relation to the trait distribution (mismatched), whereas the 95-item pool was roughly centered on the trait mean (matched). A total of four banks were generated by crossing the item pool sizes (30 and 90) and the information distributions (matched and mismatched).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

CAT Simulations

A sample of 1,000 simulees was drawn from the standard normal distribution to serve as known theta values. Conventional procedures were used to simulate the item responses for the four item pools according to the graded response model. The item response data were then submitted to CAT simulations under the three stopping rules. The standard normal distribution was chosen as the prior for θ for the expected a posteriori (EAP) estimation and the minimum expected posterior variance (MEPV) item selection criterion. The minimum standard error threshold was set equal to 0.3. The hyper and hypo parameters for the PSER stopping rule were set equal to 0.03 and 0.01, respectively. The minimum information threshold for the minimum information stopping rule was set equal to 0.5 (Dodd, Koch, & De Ayala, 1989). The logic for this choice of information threshold was, if the threshold was set too low, the minimum information stopping rule (as it is combined with the minimum SE stopping rule for the current study) would essentially simplify to the minimum SE stopping rule. This would be undesirable because the minimum information stopping rule is included in the current study primarily so that it can be compared to the PSER stopping rule in terms of the number of items administered. Under all three stopping rules the CAT was restricted in length between a minimum of three and a maximum of 15 items. Because the current study simulates a medical or psychiatric CAT application, item overexposure was not seen as problematic (Bjorner, Chang, Thissen, & Reeve, 2007), and therefore, no item exposure control procedure was implemented. The above steps were replicated 100 times to measure random variations in the results. The two main outcome variables of interest were the average number of items administered and the measurement precision (i.e., RMSE, bias). To further examine the results conditional on theta, the study was replicated using a separate sample of simulees generated from fixed theta intervals (−2.5 to 2.5 by 0.5), 1000 per theta point totally 11,000 simulees.

Results

Figure 3 shows the item pool information functions for the 30- and 90-item matched and mismatched pools. The matched information functions were symmetrical around 0 and provided roughly consistent, maximal information in the range from −1 to +1 on the theta scale. The mismatched information functions had roughly the same shape but were shifted to the right about 1.5 logits. Consequently, the measurement precision in the mismatched pools, regardless of size, would be inevitably poor at lower theta levels.

Of particular interest in the current study were the performances of the PSER and SE stopping rules in two specific scenarios. The first scenario occurred when a standard error of 0.3 had been achieved, but sufficiently informative items were still available for administration. In these cases, the SE stopping rule would terminate the adaptive test, while the PSER stopping rule would continue administering items. In the current study, these are referred to as “hyper” cases. The second scenario of interest occurred when a standard error of 0.3 had not been achieved, but no sufficiently informative items were available for administration. In this scenario, the SE stopping rule would continue administering items, while the PSER stopping rule would terminate the adaptive test. These are referred to as a “hypo” cases.

Number of Items Administered

Table 1 summarizes the number of items administered for the three stopping rules by information distribution and pool size averaged across the 100 replications. The minimum information stopping rule administered the fewest items on average. Because the minimum information stopping rule terminates the CAT when either the minimum information threshold is unsatisfied or the minimum standard error has been reached, by design it always

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

administers fewer items than the minimum standard error stopping rule. As a result, the differences in the number of items administered between the two stopping rules are always hypo cases where the CAT is terminated before reaching the predetermined standard error threshold. As expected, the differences were smaller when the information distribution matched the trait distribution and the item pool was larger. On average, the minimum information stopping rule administered 1.68 fewer items than the minimum standard error stopping rule under the 30-item mismatched pool, whereas the difference was merely 0.037 under the 90-item matched pool condition. Overall, the PSER stopping rule administered fewer items (−0.635) than the minimum standard error stopping rule under the 30-item mismatched item pool condition, and slightly more items (+0.436) under the 90-item matched item pool condition. The differences between the PSER and minimum standard error stopping rules, however, are determined by the prevalence of both hyper and hypo cases and hence warrant a close examination. Table 2 summarizes the prevalence of hyper and hypo cases under different conditions averaged across the 100 replications. The average percentage of hyper cases ranged from 28.4 to 45.8. The highest percentage was observed under the 90-item matched pool condition. The average percentage of hypo cases ranged from 1.2 to 18.4. The highest percentage was observed under the 30-item mismatched condition.

The number of items administered for the PSER and minimum standard error stopping rules differed for hyper and hypo cases. For hyper cases, the PSER stopping rule administered an average of about one item more than the SE stopping rule, regardless of the size or distribution of the item pool. This was because the PSER rule could override the minimum standard error criterion in favor of increasing measurement precision where possible. For hypo cases, the PSER averaged two to five items fewer than the SE stopping rule, depending on the size and distribution of the item pool. For the 30 item mismatched item pool condition, the PSER averaged five items less than the SE. For both matched items pool conditions, the PSER administered about three fewer items than the SE. For the 90 item mismatched pool condition, the PSER administered about two fewer items than the SE. This is because the PSER was capable of assessing the potential standard error reduction that would result from the administration of the additional items.

In terms of the minimum number of items administered, the PSER averaged about one item more than the SE stopping rule for hyper cases. This was true for all item pool conditions. For hypo cases, the PSER administered anywhere from five to seven items fewer than the SE, depending on the size and distribution of the item pool. For the 30 item matched pool, the minimum number of items administered was about seven fewer than for the PSER than for the SE. For both 90 item pools, the difference was five items in favor of the PSER. For the 30-item mismatched pool, the minimum number of items administered was six fewer for the PSER than for the SE.

In terms of the maximum number of items administered, the PSER averaged about two to three items more than the SE, depending on the size and distribution of the item pool. For hypo cases, the maximum number of items administered was about one item fewer for the PSER, regardless of the size and distribution of the item pool.

Without regard to hyper and hypo cases, the PSER administered slightly more items than the SE stopping rule for the two 90-item pools and for the 30-item matched pool. However, this was likely due to the fact that these three pools resulted in relatively high numbers of hyper cases and low numbers of hypo cases. For the 30-item mismatched pool, the PSER administered fewer items than the SE stopping rule. This is likely due to the fact that the 30 item mismatched pool yielded fewer hyper cases, while also producing the highest number of hypo cases.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

RMSE

The RMSEs produced by the three stopping rules are closely tied to the number of items administered in the CAT. Table 3 summarizes the RMSEs for the three stopping rules under different conditions averaged across the 100 replications. The PSER produced the lowest RMSEs under all conditions with an exception of the 30-item mismatched pool condition where the PSER and minimum standard error stopping rules produced the same RMSEs. This is a noteworthy result because the PSER administered an average of 0.635 fewer items (or equivalently about 10 percent fewer items).

As with the number of items administered, the RMSEs produced by the three stopping rules differed for hyper and hypo cases. Again, hyper cases can occur only under the PSER stopping rule, whereas hypo cases can occur under both the PSER and minimum information stopping rules. For hyper cases, the PSER stopping rule produced lower RMSEs than the SE stopping rule, regardless of the size or distribution of the item pool. This was due to the fact that the PSER could override the SE stopping criterion when additional informative items were available in the pool. For hypo cases, however, the PSER stopping rule produced slightly higher RSMEs than the SE stopping rule. This was because, for hypo cases, the PSER was willing to sacrifice a small degree of measurement precision in order to limit the test length.

Results Conditional on Theta

As mentioned previously, the RMSE differences are due to the numbers of hyper and hypo cases produced by the PSER and the number of hypo cases by the minimum information stopping rule under the four item pools. Table 4 summarizes the prevalence of hyper and hypo cases for the simulees generated from fixed intervals (N=11,000). To get better sense of the implications of using the PSER criterion the RMSEs and the changes in the number of items administered (compared to the SE stopping rule) are presented for the hyper and hypo cases in the table. The prevalence of hyper and hypo cases is directly related to the item pool distribution and size, and can vary across trait levels. Figure 4-a illustrates the prevalence of hyper and hypo cases as a function of theta. The curves representing the prevalence of hyper and hypo cases showed inverse relationships and revealed some distinctive patterns: the prevalence of hyper cases is the highest at the center of the item pool information function; and the prevalence of hypo cases is peaked where the item pool information function floors.

Figure 4-b shows the average number of items administered by information distribution and pool size for the three stopping rules conditional on theta. For both the 30-item and 90-item mismatched item pool conditions, the minimum standard error stopping rule administered the maximum number of items (i.e., 15) to almost all simulees at or below theta = −1.5. For the 30-item mismatched item pool condition, the PSER administered about six items less than the minimum standard error stopping rule. For the 90-item pool condition, the PSER administered about 2 items less. The fact that the PSER administered more items under the 90-item mismatched pool condition than the 30-item condition indicates that the hypo (H−) threshold may need to be adjusted based on the pool size when the information distribution does not matched the trait distribution. The minimum information stopping rule administered substantially less items at or below theta = −1.5 than the other two stopping rules, implying that the measurement precision at lower theta levels under this stopping rule will be unacceptably poor. This also indicates that the minimum information threshold value of 0.5 has been set too high for the two mismatched item pools. The PSER administered about 0.6 more items than the other two stopping rules at or above theta = 0.0 where the item pools provided good information. For the matched pool conditions regardless of size, the PSER administered about 0.6 more items than the other stopping rules in the middle. The

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

minimum standard error stopping rule administered about two more items at the two extreme theta levels under the 30-item matched item pool condition.

Figure 4-c displays the RMSEs produced by the three stopping rules for each level of bank size and information conditional on theta. As expected from the number of items administered examined previously, the PSER produced lower RMSEs where the item pool information was at its maximum: at or above theta=0.0 for the mismatched item pool conditions and in the middle for the matched item pool conditions. For the mismatched item pool conditions, seemingly inconsistent patterns were observed for the minimum information stopping rule at the bottom end. For instance, at theta=−1.5 the minimum information stopping rule appeared to be associated with the smallest RMSEs. However, these patterns were artifacts of having degenerative theta estimates produced under the minimum information stopping rule. Almost all true theta values of −1.5 or lower ended up with an estimate of −1.43 and hence the small RMSE for theta=−1.5. Bias was also examined for the three stopping rules under each experimental condition by theta level (data not shown). With an exception of the minimum information stopping rule showing larger bias values at the lower end, no substantial differences were observed among the three stopping rules.

In order to examine the potential risk of terminating the CAT prematurely under the PSER we closely examined the hypo cases individually under all item pool conditions by studying the impact of terminating the CAT early on the recovery of the true theta. Figure 4-d displays box-and-whisker plots showing the change in the absolute error $|\theta - \hat{\theta}|$ incurred for the hypo cases, as a result of administering fewer items compared to the SE stopping rule. For each hypo case, the change in absolute error was calculated as $|\theta - \hat{\theta}_{PSER}| - |\theta - \hat{\theta}_{SE}|$. Under this formation, large positive difference values represent premature CAT termination on the part of the PSER relative to the SE stopping rule. Interestingly, the expected change is minimal and the mean absolute error is actually smaller for the PSER at some theta locations. There are, however, individual cases displaying a substantial amount of increase in error by reducing the CAT up to 6 items.

Discussion

Stopping is one of the three key components in CAT (Thissen & Mislevy, 2000). Determining when to stop can be more complex than whether or not a predetermined number of items have been administered or a desired level of measurement precision has been attained. This is particularly true when the match between the item pool and the examinee trait distribution is poor. In the present study we introduced a new stopping rule based on the predictive posterior variance. This statistic is computed as part of most fully Bayesian item selection procedures, including the minimum expected posterior variance criterion (MEPV, or equivalently maximum posterior precision: Owen, 1975; Thissen & Mislevy, 2000; van der Linden & Pashley, 2000). The predictive posterior variance can be used to determine whether or not administering one or more additional items from the pool is likely to reduce the standard error below a certain level.

Based on the conventional minimum standard error stopping rule, the PSER asks one basic question: is it worth continuing? The “worth” is conceptualized as the amount of reduction in uncertainty, and measured against two threshold values, the hyper (H+) and hypo (H−) thresholds. Prior to satisfying a standard error cutoff, the minimum worth is defined by the H− threshold, whereas it is defined by the H+ threshold once the standard error cutoff has been satisfied. Initial results were promising: the PSER shortened the CAT when information was scarce without unduly compromising measurement precision; and it attained higher precision when information was abundant without unduly lengthening the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

CAT. Because the PSER stopping rule is capable of evaluating the predicted reduction in standard error associated with the administration of additional items, it seemed to take better advantage of the item pool.

Our comparison of the new stopping rule against the two existing procedures was performed under a specific set of threshold values (e.g., a minimum SE of 0.3; an H+ threshold of 0.03, an H− threshold of 0.01, a minimum information threshold of 0.5 etc.). Because the results are expected to vary depending on the threshold parameters chosen, the generalizability of the present study is limited and subject to the reasonability of the particular threshold parameters chosen. One approach to overcome the limitation is to examine the results by systematically varying the threshold values. However, a more fundamental solution would be to develop a method to derive optimal threshold values for specific test settings. We set the H+ threshold at three times as large as the H− threshold, because the risk associated with terminating the CAT after satisfying a minimum standard error threshold was considered smaller than the risk associated with prematurely terminating a CAT prior to satisfying the minimum standard error threshold. The optimal H+ and H− threshold values can be determined numerically by defining an objective function based on the expected cost of administering additional items (which can differ by the current test length and measurement precision), testing burden, measurement precision, and other measurable and/or quantifiable factors. Computer simulations can be used to search for optimal threshold values that optimize a predetermined objective function for a given item pool and other imposed testing parameters or settings.

In evaluating the PSER criterion, understanding the behavior of hypo cases is of particular interest because of the potential for aberrant cases in which the CAT is terminated at a very early stage with a much larger error in the trait estimate than the SE stopping criterion. Hypo cases occur where information is scarce. Therefore, given that the provisional theta estimate is in the neighborhood of the true value, a steep descent in error is highly unlikely after administering a few items. In the event that the provisional and true theta values are not converging rapidly, the risk of terminating the CAT prematurely is present under the PSER criterion. As a safeguard against premature terminations, the CAT is sustained at least until a preset minimum number of items have been administered. Then the question is whether the minimum is sufficiently long for the provisional theta estimate to fall in the neighborhood of the true value. Conversely, setting the minimum too high negates the relative efficiency of the PSER and CAT in general. Optimizing various components of CAT is by no means a substitute for having an item bank with ideal characteristics, e.g., a large pool of items collectively covering a broad range of the continuum with individual items covering narrow, targeted trait levels. However, for many psychological and health outcomes constructs those seem to be uncommon prescriptions for item banks. For example, it is intrinsically difficult to have items that measure the lower (healthier) end of the construct of depression. Items that capture low levels of depression overlap with transient negative affect, which is present across a range of psychopathology other than depression (for review, see Watson, 2009).

Finally, the stopping rule proposed in the current study is designed primarily for CAT applications where test length is short, and where reduction of testing burden is of great importance. Short adaptive tests simulated in this study may not provide the equality or degree of measurement precision desirable for a high stakes achievement or aptitude test. Future research on the PSER stopping rule should also consider implementing item exposure control procedures as this may be desirable for some short CAT applications.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Acknowledgments

This research was supported in part by an NIH grant PROMIS Network (U-01 AR 052177-04, PI: David Cella) for the first author

References

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

Boyd, AM.; Dodd, BG.; Choi, SW. Polytomous models in computerized adaptive testing. In: Nering, ML.; Ostini, R., editors. Handbook of polytomous item response theory models. New York NY: Routledge; 2010. p. 229-255.

Choi SW, Reise SP, Pilkonis PA, Hays RD, Cella D. Efficiency of static and computer adaptive short forms compared to full length measures of depressive symptoms. Quality of Life Research 2010;19:125–136. [PubMed: 19941077]

Dodd BG, Koch WR, De Ayala RJ. Operational characteristics of adaptive testing procedures using the graded response model. Applied Psychological Measurement 1989;13:129–143.

Dodd BG, Koch WR, De Ayala RJ. Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. Educational and Psychological Measurement 1993;53:61–77.

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;4:13. [PubMed: 15132755]

Gibbons RD, Weiss DJ, Kupfer DJ, Frank F, Fagiolini A, Grochocinski VJ, Bhaumik DK, Stover A, Bock RD, Immekus JC. Using computerized adaptive testing to reduce the burden of mental health assessment. Psychiatric Services 2008;59:361–368. [PubMed: 18378832]

Haley SM, Gandek B, Siebens H, Black-Schaffer RM, Sinclair SJ, Tao W, Coster WJ, Ni P, Jette AM. Computerized adaptive testing for follow-up after discharge from inpatient rehabilitation: II. Participation outcomes. Archives of Physical Medicine and Rehabilitation 2008;89:275–283. [PubMed: 18226651]

Owen RJ. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association 1975;70:351–356.

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph 1969:17.

Thissen, D.; Mislevy, RJ. Testing algorithms. In: Wainer, H., editor. Computerized adaptive testing: A primer. 2. Hillsdale NJ: Lawrence Erlbaum; 2000. p. 101-133.

van der Linden, WJ.; Pashley, PJ. Item selection and ability estimation in adaptive testing. In: van der Linden, WJ.; Glass, CAW., editors. Computerized adaptive testing: Theory and practice. Netherlands: Kluwer Academic Publishers; 2000. p. 271-288.

Wainer, H., editor. Computerized adaptive testing: A primer. 2. Hillsdale, NJ: Lawrence Erlbaum Associates; 2000.

Ware JE, Bjorner JB, Kosinski M. Practical implications of item response theory and computerized adaptive testing. Medical Care 2000;38:II73–II82. [PubMed: 10982092]

Ware JE, Kosinski M, Bjorner JB, Bayliss MS, Batenhorst A, Dahlof CGH, et al. Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Quality of Life Research 2003;12:935–952. [PubMed: 14651413]

Ware JE, Gandek B, Sinclair SJ, Bjorner JB. Item response theory and computerized adaptive testing: Implications for outcomes measurement in rehabilitation. Rehabilitation Psychology 2005;50:71–78.

Watson D. Differentiating the mood and anxiety disorders: A quadripartite model. Annual Review of Clinical Psychology 2009;5:221–247.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

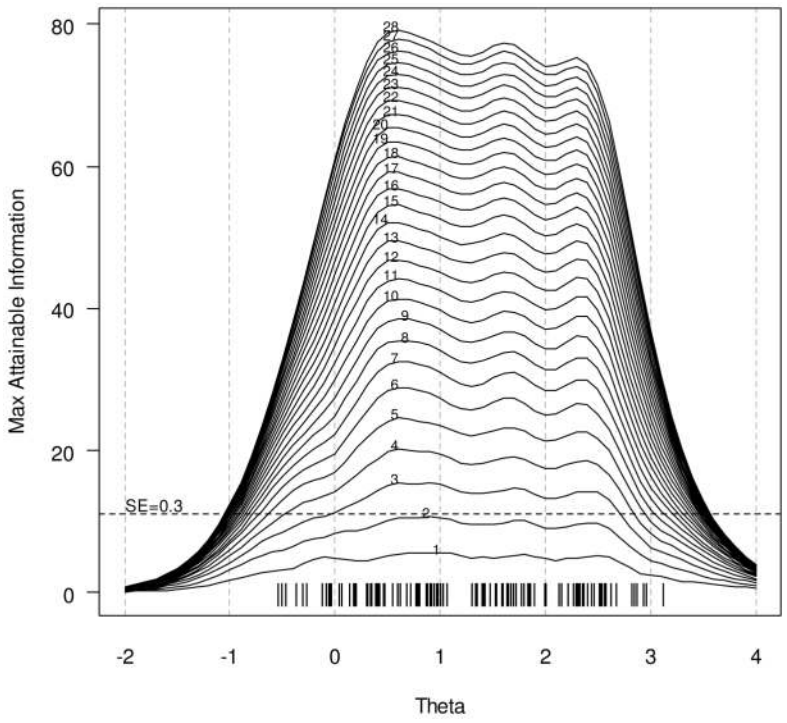

Figure 1.
Maximum attainable information curves.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

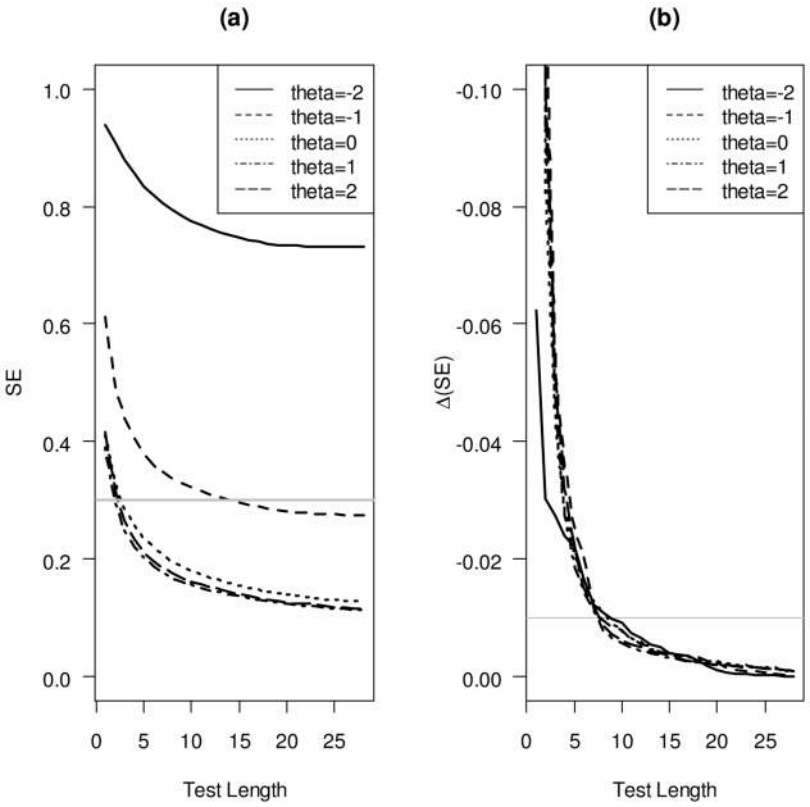

Figure 2.
(a) Minimum standard error and (b) changes in standard error by test length at different theta locations.
Note: Horizontal reference lines drawn at 0.3 (left) and −0.01 (right)

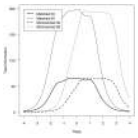

Figure 3.
Item bank information functions.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

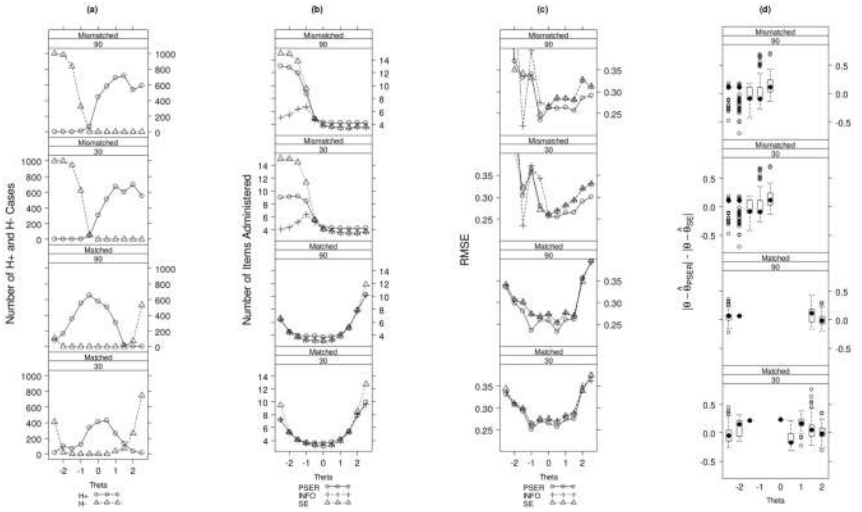

Figure 4.

(a) Number of hyper (H+) and hypo (H−) cases; (b) average number of items administered; (c) RMSE; and (d) difference in absolute error for hypo (H−) cases compared to the minimum standard error stopping rule.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 1

Number of Items Administered Averaged Across 100 Replications

		PSER		INFO		SE	
Distribution	**Size**	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Mismatched	30	5.519	0.062	4.474	0.051	6.154	0.129
	90	5.756	0.092	4.493	0.063	5.681	0.118
Matched	30	4.182	0.041	3.896	0.045	3.993	0.059
	90	4.190	0.044	3.717	0.047	3.754	0.053

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 2

Percentage of Hyper and Hypo Cases by Information Distribution and Item Pool Size Averaged Across 100 Replications

		Hyper Cases				Hypo Cases			
Distribution	**Size**	**Mean**	***SD***	**Min**	**Max**	**Mean**	***SD***	**Min**	**Max**
Mismatched	30	30.1	1.5	26.5	33.1	18.4	1.2	15.9	21.7
	90	35.4	1.6	31.7	38.8	13.0	1.0	10.7	16.0
Matched	30	28.4	1.5	24.1	32.1	3.2	0.6	2.0	5.4
	90	45.8	1.5	41.2	48.6	1.2	0.4	0.4	2.3

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 3

Root Mean Squared Error (RMSE) of Theta Estimates Averaged Across 100 Replications

		PSER		INFO		SE	
Distribution	**Size**	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Mismatched	30	0.311	0.009	0.339	0.010	0.311	0.008
	90	0.290	0.008	0.326	0.009	0.298	0.008
Matched	30	0.273	0.007	0.282	0.007	0.281	0.007
	90	0.260	0.006	0.276	0.006	0.275	0.006

Educ Psychol Meas. Author manuscript; available in PMC 2011 January 27.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

Table 4

Prevalence (%), Changes in Number of Items Administered (NIA), and Root Mean Squared Error (RMSE) for Hyper and Hypo Cases

		Hyper Cases			Hypo Cases		
Distribution	**Size**	**%**	**Mean* Change in *NIA* (*SD*)**	***RMSE***	**%**	**Mean* Change in *NIA* (*SD*)**	***RMSE***
Mismatched	30	30.8	1.03 (0.17)	0.071	32.9	−5.56 (1.00)	0.313
	90	33.1	1.04 (0.21)	0.063	28.7	−2.11 (0.77)	0.242
Matched	30	17.5	1.04 (0.29)	0.003	14.1	−3.86 (1.91)	0.044
	90	29.3	1.03 (0.16)	0.025	6.5	−2.81 (1.25)	0.030

*Compared to SE stopping rule

Educ Psychol Meas. Author manuscript; available in PMC 2011 January 27.