Item Selection Methods Based on Multiple Objective Approaches for Classifying Respondents Into Multiple Levels

Applied Psychological Measurement

2014, Vol. 38(3) 187–200

© The Author(s) 2013

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621613509723

apm.sagepub.com

Maaïke M. van Groen¹, Theo J. H. M. Eggen^{1,2},
and Bernard P. Veldkamp²

Abstract

Computerized classification tests classify examinees into two or more levels while maximizing accuracy and minimizing test length. The majority of currently available item selection methods maximize information at one point on the ability scale, but in a test with multiple cutting points selection methods could take all these points simultaneously into account. If for each cutting point one objective is specified, the objectives can be combined into one optimization function using multiple objective approaches. Simulation studies were used to compare the efficiency and accuracy of eight selection methods in a test based on the sequential probability ratio test. Small differences were found in accuracy and efficiency between different methods depending on the item pool and settings of the classification method. The size of the indifference region had little influence on accuracy but considerable influence on efficiency. Content and exposure control had little influence on accuracy and efficiency.

Keywords

multiple objective approaches, computerized adaptive testing, multiple level classification, item selection, sequential probability ratio test

Originally, computerized adaptive tests (CATs) were developed for obtaining an efficient estimate of an examinee's ability, but Weiss and Kingsbury (1984), Lewis and Sheehan (1990), and Spray and Reckase (1994) showed that CATs can also be used for classification problems (Eggen & Straetmans, 2000). In these computerized classification tests (CCTs), the main interest is not in obtaining an estimate but in classifying the examinee into one of multiple categories (e.g., pass/fail or master/nonmaster). CCT can be used to find a balance between the number of items administered and the level of confidence in the correctness of the classification decision (Bartroff, Finkelman, & Lai, 2008). In CCT, the administration of additional items

¹Cito, Arnhem, the Netherlands

²Twente University, Enschede, the Netherlands

Corresponding Author:

Maaïke M. van Groen, Cito, Psychometric Research Centre, Amsterdamseweg 13, 6814 CM Arnhem, the Netherlands.
Email: Maaïke.vanGroen@cito.nl

stops when enough evidence is available for making a decision. As in Eggen and Straetmans (2000), the focus in the current article is on classifying examinees into one of three (or even more) categories.

In adaptive classification testing, item selection is based on the examinee's previous responses, which tailors the item selection to the test taker's ability. Several item selection methods are described in the literature (see, for example, Eggen, 1999; Thompson, 2009). The design of the item selection method determines partly the efficiency and accuracy of the test (Thompson, 2009). Current methods select items based on one point on the scale and are often not adaptive in selecting items. However, if several cutting points are specified, gathering as much information as possible at all cutting points while considering the examinee's proficiency may be desirable. By doing so, information is gathered throughout a larger part of the ability scale. Especially at the beginning of the test, uncertainty exists about the ability of the examinee, which implies that gathering information at a range of points on the scale would be beneficial.

The article is organized as follows: First, details are given regarding computerized classification testing. Then some of the current and newly developed item selection methods are described. The performance of the methods was compared using simulation studies. The final section of this article gives concluding remarks.

Classification Testing

Computerized classification testing can be used if a classification decision has to be made about the level of an examinee in a certain domain. CCT was used to place students in one of three mathematics courses of varying difficulty in the Netherlands (Eggen & Straetmans, 2000), but can also be used if a decision such as master/nonmaster is required. An advantage of classification testing is that shorter tests can be constructed, while maintaining the desired accuracy (Thompson, 2009). Reducing the number of items is important because the testing time is reduced, fewer items have to be developed, security problems are reduced, and item pools have to be replenished less often (Finkelman, 2008). Adaptive classification testing shares with CAT the advantage of adapting the test to the ability of the examinee. This possibly reduces the examinee's frustration because fewer too easy or too hard items are administered and a larger set of items is selected from the item pool. However, examinees can experience that the items in a CAT are difficult (Eggen & Verschoor, 2006) when compared with a regular test in which an able student answers only relatively easy items. This drawback of CAT as well as CCT was overcome by Eggen and Verschoor (2006) by selecting easy items. CCT also shares the drawback with CAT that examinees cannot change answers to previously administered items.

One part of the CCT procedure determines whether testing can be stopped and a decision can be made before the maximum test length is reached. Popular and well-tried methods are based on the sequential probability ratio test (SPRT). The SPRT (Wald, 1947/1973) was first applied to classification testing by Ferguson (1969) using classical test theory and by Reckase (1983) using item response theory (IRT). The SPRT has been applied to CAT and multistage testing (Luecht, 1996; Mead, 2006; Zenisky, Hambleton, & Luecht, 2010). Other available methods (Thompson, 2009) are not considered in this study.

In CCT, a cutting point is specified between each pair of adjacent levels. The indifference regions are set around these points, which account for the uncertainty of the decisions, owing to measurement error, regarding examinees with ability values close to the cutting point (Eggen, 1999). If multiple cutting points are specified with accompanying indifference regions, it would be strange if the indifference regions of different cutting points overlapped. Overlapping indifference regions implies that classification into one of three levels is admissible for examinees

with an ability within the overlapping regions and that uncertainty exists about decisions regarding the three levels. In this situation, test developers should reconsider the number of cutting points and the size of the indifference regions. However, in practice, this is not always possible.

For applying the SPRT to a classification problem, two hypotheses are formulated for each cutting point θ_c based on the boundaries of the accompanying indifference region (Eggen, 2010):

$$H_0 : \theta < \theta_c - \delta_{c1}, \quad (1)$$

$$H_a : \theta > \theta_c + \delta_{c2}, \quad (2)$$

in which θ denotes ability and δ_c the widths of the indifference regions. These are set equal to δ . To avoid overlapping indifference regions, δ should be smaller than half the difference between adjacent cutting points.

Item responses are modeled using IRT, in which a relation is specified for the score on an item depending on item parameters and the examinee's ability (Van der Linden & Hambleton, 1997). The relationship between giving a specific score to an item ($x_i = 1$ correct, $x_i = 0$ incorrect) and an examinee's is modeled with a probability function. The model used here is the two-parameter logistic model (Birnbaum, 1968/2008), in which the probability of a correct response is given by

$$P(x_i = 1 | \theta) = \frac{\exp(a_i[\theta - b_i])}{1 + \exp(a_i[\theta - b_i])} = p_i(\theta), \quad (3)$$

where a_i represents the discriminating power of item i , b_i difficulty, and θ ability.

A prerequisite for CCT is a calibrated item bank that is suitable for the specific testing situation. In a calibrated item bank, the fit of the model is established, and estimates of the item parameters are available, with items with inappropriate difficulty or low discrimination parameters removed.

In IRT, the probability of an examinee's responses to test items is conditionally independent given the latent ability parameter. Inference about the ability of an examinee can be drawn from the likelihood of the responses after k items are administered (Eggen, 1999) using

$$L(\theta; \mathbf{x}) = \prod_{i=1}^k p_i(\theta)^{x_i} [1 - p_i(\theta)]^{1-x_i}, \quad (4)$$

in which $\mathbf{x} = (x_1, \dots, x_k)$ denotes the vector of responses to the administered items.

When the SPRT is applied to classification testing, the likelihood ratio of both hypotheses after k items are administered (Eggen, 2010) is used as the test statistic:

$$LR(\theta_c + \delta; \theta_c - \delta) = \frac{L(\theta_c + \delta; \mathbf{x})}{L(\theta_c - \delta; \mathbf{x})}. \quad (5)$$

Decision rules are applied for making the decision to continue testing or to make the decision that the performance is at a level below or above the specific cutting point:

$$\text{administer another item if} \quad \beta(1 - \alpha) < LR(\theta_c + \delta; \theta_c - \delta) < (1 - \beta)/\alpha; \quad (6)$$

$$\text{ability below the cutting point if} \quad LR(\theta_c + \delta; \theta_c - \delta) \leq \beta/(1 - \alpha); \quad (7)$$

$$\text{ability above the cutting point if} \quad LR(\theta_c + \delta; \theta_c - \delta) \geq (1 - \beta)/\alpha, \quad (8)$$

where α and β are small constants that specify acceptable decision errors (Eggen, 1999). In practice, a maximum test length is set to ensure that testing stops at some point. If the maximum test length is reached, the examinee is classified as performing above the cutting point if the likelihood ratio is larger than the midpoint of the interval of Equation 6. If multiple cutting points are specified and the decision is made that the ability is above cutting point θ_c , the same procedure is applied for cutting point θ_{c+1} .

If δ increases, the difference between the likelihoods is larger and thus, more uncertainty is allowed for making the decision, which implies less accurate decisions and shorter tests. Eggen (1999) found for the situation with one cutting point, increasing the acceptable error rate by increasing α and β had little effect on the proportion of correct decisions (PCD), but increasing δ influenced classification accuracy.

Current Item Selection Methods

Several item selection methods can be used in CCT (see Eggen, 1999; Luecht, 1996; Stocking & Swanson, 1993; Thompson, 2009). Most methods were developed for tests with classification into two levels, but a few methods were proposed for tests with more levels. The majority of item selection methods are based on Fisher information (Van der Linden, 2005). Other types of information such as Kullback–Leibler information (Eggen, 1999) and mutual information (Weissman, 2007) can also be used but are not included here. If maximizing Fisher information is the objective, the optimization function becomes

$$\max_{i \in V_i} I_i(\theta), \quad (9)$$

where V_i denotes the set of items still available for administration. In Equation 9, the information I provided by the $k + 1$ th item is maximized:

$$I_i(\theta) = a_i^2 p_i(\theta) [1 - p_i(\theta)]. \quad (10)$$

A method currently used is maximizing information at the current ability estimate $\hat{\theta}$. The accuracy of this estimate is related to the number of items available for estimation (Hambleton, Swaminathan, & Rogers, 1991), which causes the method to select items that are potentially not optimal at early stages of the test. The advantage of this method is that items are selected adaptively to the examinee.

A number of methods are available for classification tests with more than two levels (Eggen & Straetmans, 2000; Wouda & Eggen, 2009). Maximization of test information at the middle of the cutting points and at the nearest cutting point (Spray, 1993) are just two approaches (Eggen & Straetmans, 2000). The first method determines the middle of the cutting points nearest to the current estimate and maximizes information at that point. The second method optimizes at the cutting point located nearest to the ability estimate. Both methods base their item selection on the ability estimate, which is considered an advantage in educational settings.

As Weissman (2007) concluded, choosing an item selection method in conjunction with the SPRT is not straightforward. Spray and Reckase (1994) concluded that maximizing information at the cutting score, for classifying into one of two levels, results, on average, in shorter tests than does selecting items at the current ability estimate. Thompson (2009), however, concluded that this method is not always the most efficient option. Wouda and Eggen (2009) compared methods that maximize information at the middle of the cutting points, at the nearest cutting point, and at the ability estimate using simulations and found for the situation with two cutting

points that maximization at the middle of the cutting points resulted not only in the most accurate, but also in the longest, tests.

The methods described thus far all select items based on some optimal statistical criterion. In practical testing situations, however, item exposure and content control also have to be considered. Item overexposure can be a safety concern and several methods have been developed for dealing with it (e.g., see Sympson & Hetter, 1985). Content control mechanisms can ensure that the assembled tests meet test specifications; for example, 10 items should measure Domain A and at least 12 Domain B. For an extended overview of methods that deal with content restrictions, see Van der Linden (2005).

The methods described in this section maximize information at one point on the latent scale using the ability estimate. One objective is formulated using this estimate and it is this objective that is optimized. An alternative approach is to maximize information on all cutting points simultaneously. If an objective is formulated for each cutting point, they can be combined using a multiple objective approach. These approaches combine several objectives into one objective function using various methods. The developed methods all take the ability estimate into account. The advantage of these approaches is the more precise measurement at all cutting points. Multiple objective approaches were used for optimal test design in multidimensional testing (Veldkamp, 1999) and exposure control in CAT (Veldkamp, Verschoor, & Eggen, 2010).

Item Selection Based on Multiple Objective Approaches

Veldkamp (1999) described six approaches for combining multiple objectives: weighting methods (WM) ranking or prioritizing methods, goal programming (GP) methods, global-criterion (GC) methods, maximin methods, and constraint-based methods. These approaches are adapted for classification testing with multiple cutting points. The methods are first described and then adapted for CCT.

Weighting Methods

A straightforward method for optimizing several objective functions involves combining them into one objective function to which weights can be added to give various objectives varying importance (Veldkamp, 1999). The weighted deviation model (Swanson & Stocking, 1993), in which the deviations from the goal values are combined using weights, is probably the most well-known application of WM to test construction problems. Instead of weighting the deviations, weighting is applied to the objectives. The decision between Levels 1 and 2 could be considered to be more important than the decision between Levels 2 and 3. Specification of different weights for the two decisions ensures that more information is gathered at the first cutting point. Varying weights while administering the test gives more weight to specific objectives at various testing stages. In this study, the weight for a specific cutting point increases if the ability estimate is closer to the cutting point. This implies that item selection is adapted to the proficiency level of the examinees. The resulting objective function is

$$\max \sum_{c=1}^C \frac{1}{|\hat{\theta} - \theta_c|} I_i(\theta_c), \quad \text{for } i \in V_i. \quad (11)$$

Ranking or Prioritizing Methods

If certain objectives are more important than others, ranking or prioritizing methods can be used. Ranking methods require all objectives be ranked according to their perceived importance (Ignizio, 1982). In the first step, the most important objective is optimized. In the second step, a constraint is added that ensures that the value of the first objective is close to the target value obtained and the second objective is optimized. This process continues until all objectives are optimized (Veldkamp, 1999). In most tests using the SPRT, ranking or prioritizing methods cannot be used because no differences in the importance of certain cutting points are specified.

Goal Programming

In both methods discussed so far, the goal was to find the optimal solution. GP methods focus on achieving specific target values (Veldkamp, 1999). However, achieving all target values specified a priori is not always possible. In those situations, the preferred solution is calculated (Veldkamp, 1999). GP methods minimize the deviations between what was aspired to and what is actually accomplished (Ignizio, 1982). The combined objective function specifies the deviations from the targets and the priorities for achieving each objective (Mollaghasemi & Pet-Edwards, 1997). Several goal function approaches are described in the literature such as Van der Linden's (2005) framework for optimal test design and the normalized weighted absolute deviation heuristic (Luecht, 1996).

Veldkamp (1999) proposed that, in the absence of prespecified targets, the test assembler starts with an intuitive guess and use the procedure iteratively. In CCT, no targets are available for information, but gathering as much information as possible is preferred. One possibility is to compute the sum of the information each available item can provide at each cutting point and at the current ability estimate. The item with the largest sum is selected. Weights can be added before calculating the sum. The resulting objective function then becomes what Luecht (1996) calls a composite objective function,

$$\max \sum_{s \in V_s} w_s I_i(\theta_s), \text{ for } i \in V_i, \quad (12)$$

where w_s denotes the weight for scale point s and $V_c = \{\theta_1, \dots, \theta_C, \hat{\theta}\}$. In this study, all cutting points were considered equally important so all weights were set equal. However, a utility function can be used to weight the objectives as well as weights based on policy decisions.

Global-Criterion Methods

GC methods optimize all objectives separately and combine the results into one global criterion. First, all objectives are optimized resulting in optimal values for every objective (Veldkamp, 1999). Second, the results are combined into a global criterion. The value of this global criterion is then optimized. The method for combining the results is specified a priori. When this method is applied to CCT, the first step is to optimize the objectives for each cutting point separately. One possibility is to consider the items that provide the most information at the different cutting points and the current ability estimate. Several methods for combining the results are possible. The separate objectives were combined by calculating the sum of the information at the cutting points for all items that provide the most information at one of the selected points on the ability scale. The combined objective then becomes

$$\max \sum_{c=1}^C I_i(\theta_c), \text{ for } i \in V_{\max}, \quad (13)$$

where $V_{\max}$ denotes the set of available items, which provide the most information at one of the cutting points or the current ability estimate. This differs from the previous method in that the optimal values for the objectives are combined and then the global optimum is used instead of using nonoptimal values and then combining them into the goal function. Weights have not been included in this study but can be added.

Maximin Methods

MA methods can be used if a maximum value has to be found on multiple points (Boekkooi-Timminga, 1989). A lower boundary is set on the target of the objectives. This boundary is then maximized (Van der Linden, 2005). If the objectives are on the same scale, the method ensures that unexpected extreme values for one or more objectives do not occur.

A good starting value has to be found for the boundary. This value should be low enough to ensure feasibility and high enough to ensure that the calculations do not consume unreasonable amounts of time. In CCT, a lower boundary can be set on the information at the cutting points and the ability estimate provided by the items that were administered thus far. The item was selected that maximized the boundary.

Constraint-Based Methods

Constraint-based methods require prioritizing the objectives (Veldkamp, 1999). One objective is optimized, and the other objectives are reformulated into constraints. Additional constraints can be added to specify other test characteristics such as amount of testing time, content control using, for example, Van der Linden's (2005) framework. In CCT, using a constraint-based method implies that one cutting point is considered the most important. This cutting point is formulated as an objective, and constraints are formulated for the remaining cutting points. In the present study, no cutting point was considered to be most important, so this method was not applied here.

Simulation Studies

Using simulation studies, average test length (ATL) and classification accuracy were investigated for various item selection methods. Classification accuracy was defined as the PCD. The influence of the size of the indifference region on ATL and PCD is investigated in the second part. Simulations with different values for α and β are not reported here because increasing the acceptable error rate had little effect on the PCD. This is in line with Eggen's (1999) findings; he reported the same finding for simulations with two cutting points. The effects of content constraints and exposure control were investigated in the last section.

Methods based on WM, GP, GC, and MM were included in the simulation studies. Three existing item selection methods were also included: selecting the item that maximizes information at the current ability estimate (AE), selecting the item that maximizes information at the middle of the nearest set of cutting points (MC; Eggen & Straetmans, 2000), and selecting the item that maximizes information at the nearest cutting point (NC; Spray, 1993). Random item selection (RA) was included to serve as a baseline for the ATL and the PCD.

The characteristics of the item pool were expected to influence the ATL and PCD of the eight item selection methods. Two item pools were investigated. The first pool was simulated with item parameters generated with $a \sim N(1.50, 0.50)$ with $a > 0$ and $b \sim U(-3.00, 3.00)$. One thousand items were generated for the item pool, and maximum test length was set at 40 items. The specifications of this pool result into a rather "ideal" situation. One thousand examinees were randomly drawn from $\theta \sim N(0.00, 1.00)$. This was replicated 100 times for each item selection method. The ATL and PCD strongly depend on the number of defined cutting points; thus, simulations using two, three, and four cutting points are presented here. The cutting points were set at the 33th and 66th percentiles for two cutting points; the 25th, 50th, and 75th percentiles for three cutting points; and the 20th, 40th, 60th, and 80th percentiles for four cutting points of the population distribution. In the study, $\delta = 0.10$, and $\alpha = \beta = .05$.

A second item pool consisted of 250 items from a real test. The parameters of the items in the pool are from a mathematics test for adult education in the Netherlands (Eggen & Straetmans, 2000). The test was used to place students in one of three courses. Using a standard setting procedure, the cutting points were set at -0.13 and 0.33 . Testing was stopped after 40 items or less to ensure comparability with the simulations with the simulated item pool. The acceptable error rates (α , β) were set at 0.05 and δ at 0.10 . The distribution of θ for generating examinee ability was set equal to the estimated population distribution. One thousand examinees were simulated with $N \sim (0.294, 0.522)$. The items had a mean item difficulty of 0.00 , and the mean discrimination was 3.09 . The simulations were executed for the eight item selection methods and were replicated 100 times.

Simulations With a Simulated Item Pool

The results for the simulated item pool simulations are summarized in Table 1. The ATL and the PCD are provided in the table. First, Table 1 clearly indicates that on average at least 32 items are required before tests are terminated by the SPRT method if two cutting points were specified. Second, the PCD was just above the specified accuracy level. The differences in the PCD were rather small between the different item selection methods. However, the random method was 8% less accurate than the other methods. Depending on the item selection method, almost 35 items were required with three cutting points. AE was the most efficient method. The differences in the PCD were also small, but random item selection resulted, in additional incorrect decisions for 8% of the examinees. A minimum of 85% of the examinees were classified correct using one of the other item selection methods.

Even fewer tests were terminated by the SPRT with four cutting points. AE, MC, and NC resulted in the lowest ATL. Depending on the item selection method, 80% to 83% of the classifications were correct. RA classified accurately in 68% of the tests. A comparison of the simulations using two, three, and four cutting points showed that more items were required before testing were terminated by the SPRT if more cutting points were specified. In addition, the currently used item selection methods tended to require fewer items if the number of cutting points was increased. In addition, the classification accuracy decreased by 2% to 4% when an additional cutting point was specified. Specifying more cutting points also implied that the multiple objective methods have to take more cutting points into account, which could have resulted in longer tests if the distances between the current ability estimate and the cutting points increased.

Simulations With the Mathematics Item Pool

ATL was much smaller in the simulations with the mathematics pool (Table 2). RA was clearly outperformed by the other methods. Eleven additional items were required before a

Table 1. Results From Simulations With a Simulated Item Pool.

Item selection method	Two CP		Three CP		Four CP	
	ATL	PCD	ATL	PCD	ATL	PCD
RA	39.533	0.820	39.776	0.745	39.868	0.676
AE	32.646	0.906	34.861	0.866	35.938	0.826
MC	32.694	0.902	34.989	0.862	36.038	0.827
NC	32.721	0.908	35.009	0.867	36.170	0.828
WM	34.153	0.907	37.201	0.867	38.359	0.830
GP	33.065	0.907	37.296	0.863	39.364	0.818
GC	35.602	0.902	39.275	0.855	39.961	0.809
MA	33.259	0.902	36.856	0.853	38.444	0.798

Note. CP = cutting points; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion; MA = maximum.

Table 2. Results From Simulations With a Mathematics Item Pool.

Item selection method	ATL	PCD
RA	31.599	0.875
AE	20.213	0.915
MC	20.666	0.912
NC	20.593	0.916
WM	20.923	0.917
GP	20.831	0.914
GC	22.571	0.909
MA	22.514	0.908

Note. ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion.

classification was made if RA was used instead of AE. The shortest tests were produced by AE, NC, GP, and WM. The PCD was the highest for WM. NC, AE, GP, and MC had a slightly lower PCD. RA resulted in the lowest PCD. Most methods classified more accurately than was specified by α and β .

Simulations With Various Delta Values

To investigate the effect of the size of the indifference region on the ATL and PCD, the simulations were repeated with different values for δ . The effect of the size of the indifference region was investigated for the simulated and the mathematics item pool. The investigation of δ was limited to the range 0.050 to 0.400 for the simulations with the simulated pool and for the simulations with the mathematics pool to the range 0.025 to 0.225. The simulations with the simulated item pool were performed with two cutting points. As described previously, setting δ to larger values does not make any sense if that implies that the indifference regions of different cutting points would overlap. The results are displayed for RA, AE, WM, and GC. The results of the other methods that were not included were similar to the results of the presented methods except for RA.

Figure 1. Results from simulations with a simulated item pool for different sizes of the indifference region.

Note. The solid black line denotes RA, the solid gray line WM, the dotted black line AE, and the dotted gray line GC. RA = Random item selection; WM = Methods based on weighting; AE = ability estimate; GC = global criterion; PCD = proportion of correct decisions; ATL = average test length.

Simulations with a simulated item pool. The PCD for the simulations with a simulated item pool is displayed in the left part of Figure 1. The difference in the PCD appeared to be related to the size of the indifference region. If δ was set at rather large values, the PCD decreased. The results also indicated that the difference in accuracy between random item selection and other item selection methods was only slightly influenced by the value of δ . The ATL of the simulations with a simulated item pool is plotted in the right part of the figure for different values of δ . The number of items administered before a classification was made was clearly influenced by δ . ATL decreased if the size of the indifference region was increased. The ATL decreased with 11 items if RA is used, but ATL decreased with up to 27 items if a different method was used.

Simulations with the mathematics item pool. The PCD for the simulations with the mathematics item pool is displayed in the left part of Figure 2. In contrast to the simulations with a simulated item pool, the PCD decreased if the size of the indifference region was increased to 0.10. The difference in the PCD was rather small between different values of δ , but depending on the item selection method, if δ was set higher than 0.175, the PCD dropped below the desired accuracy level as specified by α and β . If RA was used, the PCD was for all investigated values of δ below the desired accuracy level.

As seen in Table 2, the choice for an item selection method influenced the PCD, but Figure 2 indicates that this also holds for different values of δ . In the right part of Figure 2, the ATL is shown. The ATL decreased to 10 or 11 items depending on the item selection method if δ was increased, except for RA. Although test length decreased a lot, δ was only slightly increased because the PCD had to remain above the specified accuracy level.

Simulations With Content and Exposure Control

Thus far, the simulations were limited to the situation in which no content specifications were specified and no action taken to avoid overexposure. In actual testing programs, constraints

Figure 2. Results from simulations with the mathematics item pool for different sizes of the indifference region.

Note. The solid black line denotes RA, the solid gray line WM, the dotted black line AE, and the dotted gray line GC. RA = Random item selection; WM = methods based on weighting; AE = ability estimate; GC = global criterion; PCD = proportion of correct decisions; ATL = average test length.

have to be met for the content of the test, and attention is also paid to item exposure. In adaptive testing, implementing content or exposure control often results in longer tests. Eggen and Straetmans (2000) considered content and exposure control for the mathematics item pool. Their simulations were replicated in this study for the eight item selection methods.

The Kingsbury and Zara (1989) approach was used to select 16% of the items from subdomain mental arithmetics/estimation, 20% from measuring/geometry, and the other items from the other domains in the curriculum. The item was selected from the domain for which the difference between the desired and achieved percentage of items selected thus far was the largest.

Exposure control was implemented using a simplified form of the Sympton and Hetter (1985) method. When an item was selected, a random number g was drawn from the interval $(0, 1)$. When $g > 0.5$, the item was administered; if not, another item was selected by the item selection method. The rejected item was not admissible for the respondent for the remainder of the test.

A different procedure was implemented to select the first 3 items. An examinee was presented a relatively easy item from the item pool. Fifty-four items were denoted as easy items. Depending on the implemented content control method, an easy item was selected for each domain, or 3 relatively easy items were selected at random.

Simulations were run implementing content (C) and exposure (E) control with the maximum test length set at 25, $\delta = 0.10$, and $\alpha = \beta = .05$. For random item selection, simulations were limited to the situation in which no content and exposure control was implemented, but Items 1 to 3 were selected from the set of relatively easy items.

The results for these simulations are given in Table 3. Content control had limited influence on the ATL and the PCD. Exposure control had not only a slightly larger influence on the ATL but also had a low impact on the PCD. Implementing content and exposure control resulted in the longest tests and the least accurate decisions, and had little influence on the ranking of the item selection methods.

Table 3. Results From Simulations With and Without Content Constraints and Exposure Control.

Selection	No C, no E		C		E		C + E	
	ATL	PCD	ATL	PCD	ATL	PCD	ATL	PCD
RA	23.103	0.838						
AE	17.122	0.896	17.343	0.895	18.218	0.885	18.439	0.886
MC	17.330	0.890	17.685	0.891	18.314	0.883	18.494	0.883
NC	17.667	0.896	17.701	0.897	18.561	0.887	18.648	0.888
WM	17.987	0.897	17.969	0.896	18.768	0.889	18.897	0.889
GP	17.451	0.893	17.675	0.895	18.392	0.884	18.580	0.884
GC	18.985	0.885	18.970	0.889	19.779	0.881	20.196	0.881
MA	18.529	0.885	18.845	0.885	19.256	0.878	19.609	0.877

Note. C = content constraints; E = exposure control; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion.

Discussion

Four item selection methods were developed for this study and were compared with current methods for classifying examinees into multiple categories. The new methods consider the multiple cutting points when selecting items. Simulations were used to investigate the effect of the item selection methods on the PCD and the ATL using the mathematics item pool.

In the first series of simulations, a simulated item pool was used. Two, three, and four cutting points were specified. Random item selection resulted in a lower PCD, but differences in test length and accuracy were small for the other methods. The differences in the PCD for different item selection methods were also small. A second series of simulations was investigated using the mathematics item pool. The differences in the ATL and PCD were also small except for random item selection.

It was expected beforehand that taking multiple cutting points into account would decrease the ATL and increase the PCD. The simulation results, however, show that currently available item selection methods classify as accurate and efficient as the multiple objective methods. This is probably caused by taking all cutting points into account in the later stages of the test. In the later stages, we already know in which part of the scale a classification is likely to be made, but the methods also take the other parts of the scale into account. This can be solved by restricting the number of cutting points that is considered in item selection after a number of items are administered. It would also be interesting to use a multiple objective approach in the starting phase of the test and then switch to one of the currently available methods. By using such an approach, the advantages of both types of methods are exploited; first, a broad part of the scale is considered, and then the more accurate estimate can be used for item selection.

Comparing the simulations with a simulated pool and the mathematics item pool shows that characteristics of the item pool, distribution of ability, settings of the classification method, and the number of cutting points all influence test length and accuracy. This suggests that simulation studies should be executed during the development process for a classification test.

Test length and accuracy were influenced by the size of the indifference regions. The simulations were repeated with different specifications for the indifference region. The ATL and the PCD decreased in the simulations with a simulated and a mathematics item pool. Test length and accuracy were only slightly influenced by content and exposure control in the simulations

with the mathematics item pool. This finding is in line with Eggen and Straetmans's (2000) findings.

In the present study, item selection methods were included that considered the examinee's ability. In addition to the obvious psychological and educational advantages of adaptive item selection, some initial simulations with item selection methods that were not adaptive showed that the number of items required before making a decision and the accuracy of the classifications was comparable or even better with adaptive methods than with methods that are not adaptive.

Investigating whether the current results can be replicated if different and larger item pools are used or if the characteristics of the examinees are changed would be interesting. In the present study, only one method was developed per multiple objectives approach, but other methods are possible. These should be investigated using simulation studies with different item pools, different SPRT settings, and different examinee characteristics.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Bartroff, J., Finkelman, M. D., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*, 73, 473-486. doi:10.1007/s11336-007-9053-9
- Birnbaum, A. (2008). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Charlotte, NC: Information Age. (Original work published 1968)
- Boekkooi-Timminga, E. (1989). *Models for computerized test construction* (Unpublished doctoral dissertation). Twente University, Enschede, the Netherlands.
- Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, 23, 249-261. doi:10.1177/01466219922031365
- Eggen, T. J. H. M. (2010). Three-category adaptive classification testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 373-387). New York, NY: Springer. doi:10.1007/978-0-387-85461-8
- Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60, 713-734. doi:10.1177/00131640021970862
- Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, 30, 379-393. doi:10.1177/0146621606288890
- Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computerassisted branched test for a program of individually prescribed instruction* (Unpublished doctoral dissertation). University of Pittsburgh, PA.
- Finkelman, M. D. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, 33, 442-463. doi:10.3102/1076998607308573
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Ignizio, J. P. (1982). *Linear programming in single and multiple objective systems*. Englewood Cliffs, NJ: Prentice Hall.
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing. *Applied Measurement in Education*, 2, 359-375. doi:10.1207/s15324818ame0204_6

- Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386. doi:10.1177/014662169001400404
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404. doi:0.1177/014662169602000406
- Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*, 185-187. doi:10.1207/s15324818ame1903_1
- Mollaghasemi, M., & Pet-Edwards, J. (1997). *Making multiple objective decisions*. Los Alamitos, CA: IEEE Computer Society Press.
- Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York, NY: Academic Press.
- Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Report No. ACT-RR-93-7). Iowa City, IA: American College Testing.
- Spray, J. A., & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test*. Paper presented at the National Meeting of the National Council on Measurement in Education, New Orleans, LA.
- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292. doi:10.1177/014662169301700308
- Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166. doi:10.1177/014662169301700205
- Sympton, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
- Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793. doi:10.1177/0013164408324460
- Van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer. doi:10.1007/0.387.29054.0
- Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York, NY: Springer.
- Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement, 36*, 253-266. doi:10.1111/j.1745-3984.1999.tb00557.x
- Veldkamp, B. P., Verschoor, A. J., & Eggen, T. J. H. M. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicológica, 31*, 335-355.
- Wald, A. (1973). *Sequential analysis*. New York, NY: Dover. (Original work published 1947)
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375. doi:10.1111/j.1745-3984.1984.tb01040.x
- Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67*, 41-58. doi:10.1177/0013164406288164
- Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.
- Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355-372). New York, NY: Springer. doi:10.1007/978-0-387-85461-8_18