Behavior Research Methods (2022) 54:2565–2578
https://doi.org/10.3758/s13428-021-01757-5

Improving test efficiency for a grid multidimensional computerized classification test by the application of a conditional latent-trait distribution to a sequential probability ratio test

Tien-Hsiang Liu[1] · Cheng-Te Chen[2] · Chung-Ping Cheng[3] · Ching-Lin Shih[1]

Accepted: 23 November 2021 / Published online: 16 December 2021

Abstract

The measurement efficiency of a multidimensional computerized adaptive testing (MCAT) can be improved by taking the correlations between the dimensions into account during the item selection and latent-trait estimation procedures (Segall, 1996; Wang & Chen, 2004). Although a multidimensional computerized classification test (MCCT), which was based on a multidimensional itemresponse model, was previously found more efficient than its unidimensional counterpart, the difference was negligible (Seitz & Frey, 2013); the researchers had adopted a sequential probability ratio test (SPRT) as the termination criterion in this MCCT study. To make a classification decision on each dimension, which is called a grid classification (Wang et al., 2019), only items that loaded on that dimension were used to calculate the likelihood ratio, which squandered the available information of the correlations between the dimensions. The current study utilizes such useful information to improve the measurement efficiency of the MCCT by applying a conditional distribution of the latent-trait estimates and then including all the administered items to calculate the likelihood ratio in the SPRT. The performance of this newly proposed method was evaluated through a series of simulation studies. The results showed that the proposed method can sizably improve the measurement efficiency of an MCCT by saving 1% to 32% of the test length in comparison with the SPRT when the two test dimensions are at least moderately correlated. The findings and further applications of this study are discussed.

Keywords Multidimensional computerized classification test · Sequential probability ratio test · Conditional latent trait distribution

Every day, various tests and questionnaires are used in many fields; these include achievement tests, self-reported psychological/clinical assessments, and organizational personnel selection instruments. These tests can be divided into two categories: (1) based on the purposes of the measurement and (2) the way that the test scores are interpreted. Included in the first category is the ranking of the trait level of the test-takers in an ascending or descending manner along a continuum that is being measured, which is called norm-referenced testing. Examples of this include many language proficiency tests (e.g., the Test of English as a Foreign Language, TOEFL) and many achievement tests in schools (e.g., when a "curve" is applied in the grading). The other category of tests includes those in which an examinee's trait level is compared to a set of pre-specified criteria to make a pass/fail, master/non-master, or basic/proficient/advanced decision, which is called criterion-referenced testing. Examples of the latter type of test, such as a test for a driver's license, teacher certification, or a depression screening instrument, intrinsically classify examinees into one of two or more mutually exclusive categories and are also commonly described as classification tests (Spray & Reckase, 1996).

Generally, these two kinds of tests were originally implemented in a paper-and-pencil format in which all the examinees were asked to respond on an identical test form. Benefitting from the rapid progress in computer technology and

✉ Ching-Lin Shih
educls@g-mail.nsysu.edu.tw

[1] Institute of Education and Center for Teacher Education, Assessment Research Center, National Sun Yat-sen University, Kaohsiung 804, Taiwan

[2] Department of Educational Psychology and Counseling, National Tsing Hua University, Hsinchu City, Taiwan

[3] Department of Psychology, National Cheng Kung University, Tainan City, Taiwan

the increased availability of computers (particularly personal computers) in the latter half of the twentieth century, the way that tests were delivered and administered changed. For example, computerized adaptive testing (CAT) (van der Linden & Glas, 2000) was developed to estimate respondents' latent traits on a continuum with shortened length tests that can reach precisions comparable to their paper-and-pencil counterparts. Based on similar logic, computerized classification tests (CCTs; Thompson, 2009; Huebner & Fina, 2015) were developed to facilitate making accurate classification decisions for examinees more efficiently (Spray & Reckase, 1996) and can be used to find a balance between the level of confidence in the accuracy of a classification decision and the number of items that need to be administered (Bartroff et al., 2008). For both kinds of computerized testing, a concern in measurement efficiency is crucial for test administrators and practitioners. This concern is even more pressing when a test measures multiple dimensions simultaneously.

Many of the commonly used classification instruments were designed to measure multiple dimensions simultaneously and make classification decisions on each of the dimensions separately. For example, the second version of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 1989) contains ten clinical scales and three main validity scales that measure an adult's personality and psychopathology. A total of 567 true/false items were used to measure ten clinical symptoms, such as hypochondria, depression, hysteria, and mania. After completing the inventory, a set of transformed T-scores are used to indicate the level of clinical symptoms on each of the MMPI-2 scales, with scores below 50 points representing low levels or normality. A multidimensional instrument, such as the MMPI-2, is mostly (if not always) designed to comprehensively screen for all the domains/symptoms simultaneously, rather than for only one or some of them. Under such a scenario, the required number of items for the multidimensional measure is expected to be sufficiently large to reach a qualified precision necessary for classification decisions. The more dimensions being measured, the more items will be required. As a consequence, more dimensions are likely to yield a longer questionnaire and hence increase the response burden for the participants, which might in turn encourage participants to not complete instruments, yielding lower response rates and poorer data quality for studies (Jones, 2014; Rolstad et al., 2011). Therefore, how to improve measurement efficiency is an important issue for multidimensional CCTs (MCCTs; van Groen et al., 2016).

In order to improve the measurement efficiency of MCCTs, we can follow the example provided by multidimensional computerized adaptive testing (MCAT; Segall, 1996; Luecht, 1996). Both MCCTs and MCATs are similarly based on multidimensional item-response theory. The measurement efficiency of MCATs was found to be largely improved over their multiple unidimensional counterparts by taking the correlations between dimensions into account in latent trait estimation and item selection procedures (Segall, 1996; Wang & Chen, 2004). Though it was expected that the required items for an MCCT would be fewer than for comparable multiple unidimensional CCTs (i.e., multiple UCCTs), this is not the case according to the current MCCT literature (e.g., Seitz & Frey, 2013).

Because the termination criteria highly determine the measurement efficiency of CCTs, they have been repeatedly investigated in recent years when investigating measurement efficiency. One of the most commonly used termination criteria in the CCT field is the sequential probability ratio test (SPRT; Wald, 1947). Through testing two simple hypotheses, the SPRT is used to decide whether an examinee's θ is greater than a specified point above the cutoff score or less than another specified point below the cutoff score (Thompson, 2009). This method was first applied to implement decision-making with CATs by Reckase and his colleagues (Reckase, 1983; Spray & Reckase, 1996). The SPRT has been further investigated in many studies during the last two decades (Eggen, 1999; Eggen & Straetmans, 2000; Finkelman, 2008; Thompson, 2009; van Groen et al., 2014, 2016). The popularity of the SPRT is mostly due to its ease of implementation and its usefulness in classifying examinees into two or more categories (Eggen & Straetmans, 2000; Spray & Reckase, 1996). However, these studies were mostly focused on unidimensional rather than multidimensional CCTs.

To investigate whether the SPRT can be applied to an MCCT, Spray et al. (1997) used it on the two-dimensional ACT Mathematics Test to make a single overall decision. They specified a passing rate on a reference test and obtained an equivalent latent passing score by solving for the latent trait vector. Two distinct curves that were approximately parallel were then defined to create an indifference region for a two-dimensional CCT to make an overall decision for the participants. However, the vectors that satisfied these two curves might not have yielded identical likelihood values for each administered item, so they concluded that an extension of the unidimensional SPRT to its multidimensional case was not feasible. Moreover, van Groen et al. (2016) conducted three simulation studies to compare several item-selection procedures and classification methods for a within-item multidimensional item pool (i.e., all the items simultaneously measured multiple dimensions) to make an overall decision. By applying the reference composite (RC) method (Reckase, 2009) to place all the projected examinees' latent trait vectors on a unidimensional line, all the examinees could then be ranked on the RC axis. The simulation results supported that their proposed method had the same characteristics as the unidimensional SPRT. Furthermore, the multidimensional SPRT resulted in more accurate decisions and longer tests than its unidimensional counterpart (van Groen et al., 2016). Specifically, the average test length (ATL) for the MCCTs and UCCTs ranged from 44 to 50 and 38 to 50, respectively, in which 50 was the maximum test length by design. The percentages of correct classifications (PCC) for the MCCTs (ranging from .865 to .895) were higher than those of the UCCTs (ranging from .837 to .851). Recently, van Groen et al. (2019) extended their previous studies to classify examinees by making a decision per dimension or making an overall decision based on all dimensions, under both the between-item and within-item multidimensional models. They compared two termination criteria: the SPRT and the confidence interval method (Kingsbury & Weiss, 1979). Through two simulation examples, they concluded that the SPRT tended to result in longer tests but more accurate decisions.

In addition, Seitz and Frey (2013) proposed an MCCT for between-item multidimensionality; in other words, multiple dimensions were being measured simultaneously with unidimensional items. To make classification decisions for each dimension, which is called a grid classification (Wang et al., 2019), they adopted a multiple unidimensional version of the SPRT as the termination criterion. Through a series of simulation studies, they found that an intrinsically multidimensional CCT resulted in a similar PCC with a shorter ATL than its multiple unidimensional counterparts. Specifically, for a two-dimensional test with a between-dimension correlation equal to 0.85 and only one cut score on each dimension, the mean PCC and mean ATL for the UCCTs and MCCTs were 78.95%, 47.20, 79.24%, and 46.47, respectively. Given that the PCCs were comparable, the differences in test length between the MCCTs and UCCTs were less than about one item (over an average of 47.20 items), which indicates that the efficiency of the MCCTs did not show much improvement over the UCCTs. In sum, both of the studies used multidimensional versions of the SPRT as the termination criterion in the MCCTs and found that the measurement efficiency remained similar to their unidimensional counterparts.

These unexpected results were mainly due to these studies actually using a multiple unidimensional version of the SPRT, as it functioned the same way as the unidimensional CCTs, one dimension at a time. By doing so, the correlations between the dimensions were not being taken into account in the termination criterion of the MCCTs. More specifically, the items loaded on a dimension only provided information for making classification decisions on that dimension, not for any others. Without auxiliary information delivered from the other dimensions, the measurement efficiency of the MCCTs is hardly improved, which causes difficulty for practitioners using these algorithms to make informative decisions efficiently. To this end, this study aimed to improve the measurement efficiency of grid MCCTs by including auxiliary information from other dimensions. Specifically, a conditional latent-trait distribution is applied to the multidimensional version of the SPRT. Through also incorporating the information of the correlations between the dimensions into the estimation of the examinees' latent trait into the calculation of the likelihood ratio in the SPRT, this method was expected to improve the measurement efficiency of grid MCCTs, and its performance was investigated through a series of simulation studies.

This article is organized as follows. First, multidimensional item-response theory is introduced, followed by the unidimensional and multidimensional versions of the SPRT. Second, the way we apply a conditional latent-trait distribution to the SPRT is demonstrated. Next, the cut-score-based item selection procedure is also presented. After that, we describe the simulation study that was conducted to evaluate the performance of the original SPRT and the newly proposed SPRT in terms of the percentage of correct classifications and average test lengths in grid MCCTs. Finally, several suggestions based on the findings of this study are made.

Multidimensional item-response theory

The multidimensional item response model describes the relationship between an examinees' latent trait vector, item parameter vector, and the probability of a correct response on an item. For example, taking the three-parameter multidimensional item-response model (M3PL), the probability of examinee j endorsing item i can be calculated as follows (Segall, 1996):

$$\mathrm{P_i}\left(\boldsymbol{\theta}_{\mathbf{j}}\right) = \mathrm{c_i} + \frac{1 - \mathrm{c_i}}{1 + \exp\left[-\mathbf{a}_{\mathrm{i}}'\left(\boldsymbol{\theta}_{\mathbf{j}} - \mathrm{b_i} \bullet \mathbf{1}\right)\right]}, \tag{1}$$

where $\boldsymbol{\theta}_{\mathbf{j}}$ stands for the latent trait vector of examinee j, and $\mathbf{a}_{\mathrm{i}}' = \left(\mathrm{a_{i1}}, \mathrm{a_{i2}}, \dots, \mathrm{a_{ip}}\right)$ represents the item discrimination vector of item i. For dimension r that item i was not intended to measure, $\mathrm{a_{ir}} = 0$. If all the items have multiple nonzero elements in $\mathbf{a}_{\mathrm{i}}$, then the test has within-item multidimensionality; when only one nonzero element exists in $\mathbf{a}_{\mathrm{i}}$, the test has between-item multidimensionality (Wang & Chen, 2004). The symbols $\mathrm{b_i}$ and $\mathrm{c_i}$ represent the item difficulty and pseudo-chance parameter for item i. The symbol **1** is a p × 1 unit vector, which indicates the same item difficulty is used on each dimension that is being measured.

Sequential probability ratio test (SPRT)

The SPRT method was used to make a binary classification decision through testing the two simple hypotheses listed below (Nydick, 2014):

$$\mathrm{H_0}: \theta_{\mathrm{j}} = \theta_0 - \delta = \theta_{\mathrm{L}},$$
$$\mathrm{H_1}: \theta_{\mathrm{j}} = \theta_0 + \delta = \theta_{\mathrm{U}},$$

where θ_j is an unknown latent trait parameter underlying the responses to the items for examinee j, and θ_0 is the cutoff point that is used to distinguish a pass from a fail. With these hypotheses, the SPRT creates an indifference region around θ_0 with a width equal to 2δ, where θ_L and θ_U stand for the lower and upper bounds of this region (Spray & Reckase, 1996). To make a decision with high accuracy, a narrow indifference region (i.e., a smaller δ) should be specified, yet a longer test is required to make such a decision (Reckase, 1983).

After creating the two simple hypotheses and deciding on the size of δ, the SPRT method calls for a likelihood ratio test to make a pass/fail decision for each examinee. For an examinee that has already been administered k items, the likelihood that this examinee will obtain a specific response vector is defined as follows:

$$\mathrm{L}\left(\theta_j;\mathbf{u}_k\right) = \prod_{i=1}^{k} \mathrm{P}_i\left(\theta_j\right)^{u_i}\left(1-\mathrm{P}_i\left(\theta_j\right)\right)^{1-u_i}, \tag{2}$$

where $\mathbf{u}_k$ is the examinee's response vector on k items; $\mathrm{P}_i(\theta_j)$ denotes the probability that an examinee with latent trait θ_j endorses item i; and $u_i = 1$ or 0 indicates a correct or incorrect response on item i, respectively. The likelihood ratio (LR) is then calculated as follows:

$$\mathrm{LR} = \frac{\mathrm{L}\left(\mathbf{u}_k|\theta_j = \theta_U\right)}{\mathrm{L}\left(\mathbf{u}_k|\theta_j = \theta_L\right)} = \frac{\prod_{i=1}^{k} \mathrm{P}_i\left(\theta_U\right)^{u_i}\left[1-\mathrm{P}_i\left(\theta_U\right)\right]^{1-u_i}}{\prod_{i=1}^{k} \mathrm{P}_i\left(\theta_L\right)^{u_i}\left[1-\mathrm{P}_i\left(\theta_L\right)\right]^{1-u_i}}, \tag{3}$$

where the numerator and the denominator calculate the likelihood of obtaining a response vector $\mathbf{u}_k$ at the upper and lower bounds of the indifference region, respectively. After that, the LR is compared to two decision points A and B, which are defined as $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$, where α and β represent the nominal Type I and Type II error rates, respectively. If LR ≥ A, the examinee's θ is more likely to be above, rather than below, the cutoff score, and H_1 is accepted. The examinee is then classified as a "pass" on the classification test ($\theta_j > \theta_0$), and the test is terminated. In contrast, if LR ≤ B, the examinee's θ is more likely to be below, rather than above, the cutoff score, and H_0 is accepted. The examinee is then classified as a "fail" on the classification test ($\theta_j < \theta_0$), and the test is terminated. If B < LR < A, then no decision is made, and another item is selected and administered, unless the maximum test length has been reached (Spray & Reckase, 1996).

Grid multidimensional computerized classification test (grid MCCT)

To make multiple classification decisions, one for each dimension, in a between-item grid MCCT in which each item is designed to measure only one dimension, Seitz and Frey (2013) applied the SPRT to facilitate this goal. Though the conditions can be extended to multiple cutoff scores on each dimension, we focus on the two-category condition in this study to keep the conditions simple. The hypotheses being tested for the classification decision on the cutoff score along dimension d can be expressed as follows:

$$\mathrm{H}_0^{(d)} : \boldsymbol{\theta}_j^{(d)} = \boldsymbol{\theta}^{(d)} - \delta = \boldsymbol{\theta}_L^{(d)} \tag{4}$$

$$\mathrm{H}_1^{(d)} : \boldsymbol{\theta}_j^{(d)} = \boldsymbol{\theta}^{(d)} + \delta = \boldsymbol{\theta}_U^{(d)}, \tag{5}$$

where $\boldsymbol{\theta}_j^{(d)}$ indicates the examinee's latent trait vector, and a classification decision is currently being made on dimension d. As demonstrated in Seitz and Frey's (2013) MCCT study, the SPRT method's test statistic on dimension d can be expressed as follows:

$$\mathrm{LR}^{(d)} = \frac{\mathrm{L}\left(\mathbf{u}|\boldsymbol{\theta}_U^{(d)}\right)}{\mathrm{L}\left(\mathbf{u}|\boldsymbol{\theta}_L^{(d)}\right)} = \frac{\prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_U^{(d)}\right)^{u_i}\left(1-\mathrm{P}_i\left(\boldsymbol{\theta}_U^{(d)}\right)\right)^{1-u_i}}{\prod_{i=1}^{k} \mathrm{P}_i\left(\boldsymbol{\theta}_L^{(d)}\right)^{u_i}\left(1-\mathrm{P}_i\left(\boldsymbol{\theta}_L^{(d)}\right)\right)^{1-u_i}}, \tag{6}$$

where k indicates the number of items being administered, and $\boldsymbol{\theta}_U^{(d)}$ and $\boldsymbol{\theta}_L^{(d)}$ stand for the latent trait vector that only differs in dimension d and can be expressed as follows:

$$\boldsymbol{\theta}_U^{(d)} = \left[\hat{\theta}^{(1)}, \hat{\theta}^{(2)} \ldots \hat{\theta}^{(d-1)}, \theta_U^{(d)}, \hat{\theta}^{(d+1)} \ldots\right], \tag{7}$$

$$\boldsymbol{\theta}_L^{(d)} = \left[\hat{\theta}^{(1)}, \hat{\theta}^{(2)} \ldots \hat{\theta}^{(d-1)}, \theta_L^{(d)}, \hat{\theta}^{(d+1)} \ldots\right], \tag{8}$$

where $\theta_U^{(d)}$ and $\theta_L^{(d)}$ are the upper and lower bounds of the indifference region on dimension d. By comparing the elements in Eqs. 7 and 8, readers can find that these two vectors only differ in the elements on dimension d, whereas every other element is the same. To calculate the likelihood ratio in Eq. 6, all k items that already had been administered should theoretically be involved in the calculation. However, because Seitz and Frey (2013) were trying to make decisions in a between-item MCCT in which all the items were unidimensional, the likelihood ratio in Eq. 6 was simplified, and the revised equation is as follows:

$$\mathrm{LR}^{(\mathrm{d})} = \frac{\mathrm{L}\left(\mathbf{u}_{\mathrm{d}}|\theta_{\mathrm{U}}^{(\mathrm{d})}\right)}{\mathrm{L}\left(\mathbf{u}_{\mathrm{d}}|\theta_{\mathrm{L}}^{(\mathrm{d})}\right)} = \frac{\prod_{\mathrm{i}=1}^{\mathrm{k}_{\mathrm{d}}} \mathrm{P}_{\mathrm{i}}\left(\theta_{\mathrm{U}}^{(\mathrm{d})}\right)^{\mathrm{u}_{\mathrm{i,d}}}\left(1-\mathrm{P}_{\mathrm{i}}\left(\theta_{\mathrm{U}}^{(\mathrm{d})}\right)\right)^{1-\mathrm{u}_{\mathrm{i,d}}}}{\prod_{\mathrm{i}=1}^{\mathrm{k}_{\mathrm{d}}} \mathrm{P}_{\mathrm{i}}\left(\theta_{\mathrm{L}}^{(\mathrm{d})}\right)^{\mathrm{u}_{\mathrm{i,d}}}\left(1-\mathrm{P}_{\mathrm{i}}\left(\theta_{\mathrm{L}}^{(\mathrm{d})}\right)\right)^{1-\mathrm{u}_{\mathrm{i,d}}}}, \tag{9}$$

where $\mathbf{u}_{\mathrm{d}}$ are the responses to the k_{d} items that are designed to measure dimension d, which means only the items that are loaded on dimension d were really helpful for making a decision on dimension d (i.e., items cannot provide any information when making classification decisions on dimensions that are outside the dimension that they are proposed to measure). The between-item MCCT in their study was actually a set of multiple unidimensional CCTs, with one CCT for each dimension. This in turn hampers the expected improvement in measurement efficiency from the CCTs to the MCCT in which the latter can usually facilitate the improvement by extracting information from the other dimensions.

Sequential probability ratio test with conditional latent trait distribution (SPRT_C)

To extract auxiliary information from other dimensions when making a classification decision on dimension d, we can follow the example provided by the use of the MCATs. Because the multiple dimensions that are simultaneously measured in an MCAT are usually (if not always) correlated with each other, the correlations between the dimensions are taken into account during the estimation of the latent traits. An examinee's latent trait estimates on all the dimensions are therefore updated simultaneously in real time after an item is administered, which can help increase the measurement efficiency (Wang & Chen, 2004). In the current study, to improve the measurement efficiency of an MCCT, two components were introduced to use the information from the between-dimension correlations.

First, the latent trait distributions of the other dimensions are estimated conditioned on the upper and lower limits of the dimension that is currently being classified. The conditional latent trait distribution is pre-calculated and then taken into consideration in SPRT_C by applying this information to Eq. 6. Specifically, when making a decision on dimension d, the hypotheses being tested in the MCCT are as follows:

$$\mathrm{H}_0^{(\mathrm{d})} : \theta^{(\mathrm{d})} = {\theta_0}^{(\mathrm{d})} - \delta^{(\mathrm{d})} = {\theta_{\mathrm{L}}}^{(\mathrm{d})} \tag{10}$$

$$\mathrm{H}_1^{(\mathrm{d})} : \theta^{(\mathrm{d})} = {\theta_0}^{(\mathrm{d})} + \delta^{(\mathrm{d})} = {\theta_{\mathrm{U}}}^{(\mathrm{d})}, \tag{11}$$

where ${\theta_0}^{(\mathrm{d})}$, ${\theta_{\mathrm{L}}}^{(\mathrm{d})}$, and ${\theta_{\mathrm{U}}}^{(\mathrm{d})}$ indicate the cutoff point, lower bound, and upper bound of the indifference region on dimension d, respectively. The test statistic of SPRT_C is then expressed as follows:

$$\mathrm{LR}^{(\mathrm{d})} = \frac{\mathrm{L}\left\{\mathbf{u}\middle|\left[\left(\theta^{(\mathrm{d})} = {\theta_{\mathrm{U}}}^{(\mathrm{d})}\right), \left(\overline{\theta}^{(1)}, \overline{\theta}^{(2)}, \ldots, \overline{\theta}^{(\mathrm{d}-1)}, \overline{\theta}^{(\mathrm{d}+1)}, \ldots |{\theta_{\mathrm{U}}}^{(\mathrm{d})}\right)\right]\right\}}{\mathrm{L}\left\{\mathbf{u}\middle|\left[\left(\theta^{(\mathrm{d})} = {\theta_{\mathrm{L}}}^{(\mathrm{d})}\right), \left(\overline{\theta\prime}^{(1)}, \overline{\theta\prime}^{(2)}, \ldots, \overline{\theta\prime}^{(\mathrm{d}-1)}, \overline{\theta\prime}^{(\mathrm{d}+1)}, \ldots |{\theta_{\mathrm{L}}}^{(\mathrm{d})}\right)\right]\right\}}, \tag{12}$$

where $\overline{\theta}^{(\mathrm{m})}$and $\overline{\theta\prime}^{(\mathrm{m})}$ indicate the expected values of the latent trait distribution of dimension m conditional on ${\theta_{\mathrm{U}}}^{(\mathrm{d})}$ and ${\theta_{\mathrm{L}}}^{(\mathrm{d})}$, respectively. These values can be easily calculated before the test proceeds, given the indifference regions on all the dimensions and the correlations between the dimensions.

Second, instead of only using the administered items that loaded on the dimension being classified, SPRT_C subsumes all the items that have been administered into the calculation of the likelihood ratio. For example, in Eq. 12, the vectors of the expected values of the latent trait distribution that differ for every element are listed in the denominator and numerator, respectively, in $\mathrm{LR}^{(\mathrm{d})}$. By doing this, all the administered items can provide various levels of information to help make the classification decision on dimension d, no matter which dimension these items were designed to measure. By accumulating the information that was contributed by every administered item in this way, the resulting test information on each dimension is higher than its unidimensional counterpart.

Given a test designed to measure multiple latent traits that follows a multivariate normal distribution, if the latent trait vector $\boldsymbol{\theta}$ is divided as $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta_1} \\ \boldsymbol{\theta_2} \end{bmatrix}$, then $\boldsymbol{\theta}$ follows a multivariate normal distribution with mean vector $\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}$ and a variance-covariance matrix $\Sigma = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}$. The conditional distribution of $\boldsymbol{\theta_1}$ given $\boldsymbol{\theta_2} = \mathbf{a}$ follows a multivariate normal distribution with mean and variance-covariance equal to $\overline{\mathbf{u}}$ and $\overline{\Sigma}$, respectively (Eaton, 1983). That is,

$$\mathrm{P}\left(\boldsymbol{\theta_1}|\boldsymbol{\theta_2} = \mathbf{a}\right) \sim \mathrm{N}\left(\overline{\mathbf{u}}, \overline{\Sigma}\right), \tag{13}$$

where

$$\overline{\mathbf{u}} = u_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\mathbf{a} - \mathbf{u}_2\right), \tag{14}$$

$$\overline{\Sigma} = \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}}. \tag{15}$$

For example, taking a two-dimension $\boldsymbol{\theta}$, the conditional distribution of θ_1 given θ_2 can be expressed as follows (Jensen, 2000):

$$\mathrm{P}\left(\theta_1\middle|\theta_2 = \mathrm{a}\right) \sim \mathrm{N}\left(\mathrm{u}_1 + \frac{\sigma_1}{\sigma_2}\rho\left(\mathrm{a} - \mathrm{u}_2\right), \left(1 - \rho^2\right)\sigma_1^2\right), \tag{16}$$

where ρ stands for the correlation between the two dimensions, and (u_1, σ_1) and (u_2, σ_2) stand for the means and standard deviations of the distributions of θ_1 and θ_2, respectively. Assuming the two-dimensional latent traits follow a multivariate normal distribution with $u = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where ρ stands for the correlation between the dimensions, and the cutoff point and δ on dimension θ_2 are set at 2 and 0.2, respectively, which results in an indifference region of [1.8, 2.2]. The updated upper and lower bounds of θ_1 are $[\theta_1|\theta_2 = 2.2] \sim N(1.76, 0.36)$ and $[\theta_1|\theta_2 = 1.8] \sim N(1.44, 0.36)$, while the corresponding likelihood ratio can be expressed as the following:

$$LR^{(2)} = \frac{L\Big\{\mathbf{u}\Big|\big[(\theta_1|\theta_2 = 2.2\ ),(\theta_2 = 2.2)\big]\Big\}}{L\Big\{\mathbf{u}\Big|\big[(\theta_1|\theta_2 = 1.8\ ),(\theta_2 = 1.8)\big]\Big\}} = \frac{L\Big\{\mathbf{u}\Big|\boldsymbol{\theta} = \Big[\Big(N(1.76,0.36),\ \ 2.2\Big]\Big\}}{L\Big\{\mathbf{u}\Big|\boldsymbol{\theta} = \Big[\Big(N(1.44,0.36),\ \ 1.8\ \Big]\Big\}}, \tag{17}$$

where $(\theta_1|\theta_2)$ is a distribution rather than a scalar. To calculate the likelihood, the expectation of the distribution is used in this study. In SPRT_C, classifying examinees into one of two mutually exclusive categories on dimension d can proceed with the following steps:

1. Estimate the latent trait vectors that are conditioned on $\theta_U^{(d)}$ and $\theta_L^{(d)}$ and their corresponding likelihood ratio $LR^{(d)}$ according to Eq. 12.
2. Compare $LR^{(d)}$ and make classification decisions according to the following rules.
3. Accept $\theta^{(d)} > \theta_0^{(d)}$ and classify the examinee as passed if $LR^{(d)} \geq A$;
4. Accept $\theta^{(d)} < \theta_0^{(d)}$ and classify the examinee as failed if $LR^{(d)} \leq B$;
5. Select and administer one more item if $B < LR^{(d)} < A$;
6. When the test has reached the pre-specified maximum test length, the MCCT is forced to make a decision according to the following rules:

   Classify the examinee as passed if $|\log LR^{(d)} - \log A| < |\log LR^{(d)} - \log B|$, OR Classify the examinee as failed if $|\log LR^{(d)} - \log A| \geq |\log LR^{(d)} - \log B|$.

Item selection

In determining the measurement efficiency of an MCCT, the item selection procedure is as important as the termination criterion. The item selection procedure attempts to choose an item that can provide the maximum information at a specific point on the θ scale. According to the different points along the θ continuum that are used to calculate the information, item selection procedures can be divided into two types. The first type uses the maximization of the item information at the current θ estimate. Studies using this approach include those of Kingsbury and Weiss (1983) and Reckase (1983). The second type uses the maximization of the item information at the cut-score point. The two studies by Spray and Reckase (1994, 1996) use this approach. In the current study, we adopted the latter approach for the reasons described below.

To make classification decisions in an adaptive test, Reckase (1983) used the SPRT as the termination criterion, and the items were selected to maximize the information at the previous latent trait estimate. However, Reckase stated that the SPRT assumes the probability of an endorsement is the same for all items, which makes it unreasonable if items are selected to maximize information at a latent-trait level. In addition, when the item-response function is a three-parameter logistic model (3PL; Birnbaum, 1968), the SPRT statistic was found to be non-monotonic with respect to the classification bound (Spray & Reckase, 1994). Also, by comparing these two types of item selection procedures in a CCT when using the SPRT as the termination criterion, selected items that yield the highest information on the cutoff point tend to result in higher measurement efficiency than the selected item having the most information based on the current latent-trait estimates (Spray & Reckase, 1994; Thompson, 2009). Therefore, it is recommended to select items at the classification bound rather than the current latent trait estimate when using the SPRT as the termination criterion of the classification tests, which is what was done in the current study.

Simulation study

Design

A two-dimension grid MCCT that makes binary classification decisions on each of the dimensions was employed in this simulation study. To compare the measurement efficiency of SPRT_C with the SPRT that was used in Seitz and Frey (2013), denoted as SPRT in the study, three key independent variables were manipulated: (a) the correlation between the dimensions was set at 0.0, 0.5, or 0.8, which indicated a zero, medium, or high correlation, respectively; (b) the cutoff points were set at (−3.0, −3.0), (−2, −2), (−1.5, −1.5), (0.0, 0.0), (1.5, 1.5), (2, 2), (3.0, 3.0), (3.0, 1.5), (3.0, 0.0), (0.0, 1.5), or (−3.0, −1.5), where the first and the second number in the parentheses stand for the cutoff point of the first and second dimension, respectively; and (c) the termination criterion was the SPRT method or the newly proposed SPRT_C method. The dependent variables were the percentage of correct classifications (PCC) and the average test length (ATL). The reasons for manipulating these independent variables and their respective levels are explained below.

Correlations between dimensions

It has been found that the efficiency of an MCAT increases as the correlations between the dimensions increase (Wang & Chen, 2004). The higher the correlations, the more efficient the MCAT. It was of interest as to whether similar findings would hold for an MCCT, and therefore, this variable was included. Three levels of correlations were manipulated in this study: 0.0, 0.5, and 0.8, which indicated uncorrelated, moderately correlated, and highly correlated dimensions, respectively. It was expected that SPRT_C would have better performance than SPRT as the between-dimension correlation increased; the higher the correlation, the better the performance of SPRT_C. For the conditions with different levels of correlations, the performance of SPRT_C could be inferred from the results of this study.

Cutoff points

Nydick (2014) found that the performance of a CCT with an optimal Fisher information (FI)-based item selection algorithm and SPRT depended on the location of each examinee relative to the cut point. Therefore, the locations of the cut points were used in the current study. In addition, both symmetric and asymmetric cutoff points were known to have been used in practice, so these two kinds of cutoff points were part of the manipulation.

Termination criterion

The termination criterion is used to decide whether a test should stop or not, which greatly determines the measurement efficiency of a CCT. In addition, this research proposes a new termination criterion SPRT_C. Therefore, this variable was included in the current study.

The PCC was calculated as the average percentage of correct classifications over all examinees. For each examinee, the classification decision correctness was coded as one when both dimensions were classified correctly, and it was zero when any of the dimensions were classified incorrectly. The ATL was calculated as the average number of items that were required for a classification decision across the examinees. To better depict the performance of SPRT_C relative to SPRT on the two dependent variables, we calculated the relative efficiency (RE) on the PCC and ATL for these two methods as follows:

$$\mathrm{RE} = \frac{\mathrm{N}_{\mathrm{SPRT\_c}}}{\mathrm{N}_{\mathrm{SPRT}}} \tag{18}$$

where $\mathrm{N}_{\mathrm{SPRT\_c}}$ and $\mathrm{N}_{\mathrm{SPRT}}$ stand for the PCC of SPRT_C and SPRT, respectively, when calculating the relative efficiency on the PCC; $\mathrm{N}_{\mathrm{SPRT\_c}}$ and $\mathrm{N}_{\mathrm{SPRT}}$ stand for the ATL of SPRT_C and SPRT, respectively, when calculating the relative efficiency of the ATL. For the condition in which the two methods have similar PCCs (i.e., an RE of the PCC close to 1), an RE value in the ATL smaller than one indicates a shorter test for SPRT_C than SPRT. The smaller the RE value in the ATL, the more efficient SPRT_C is. For the condition in which the PCCs are not similar for both methods, the PCC and ATL should be taken into consideration simultaneously. Consequently, the PCC per item that divides the PCC by the ATL to depict the average PCC that each item can contribute is calculated for both methods. Another index that combines the information of both the PCC and ATL is the loss function (Vos, 2000), which is defined as follows:

$$\mathrm{Loss} = 100 * 1_{\mathrm{w}} + \mathrm{L}, \tag{19}$$

where 1_{w} is a binary indicator variable that takes a value of 1 or 0 when the examinee was classified incorrectly or correctly, respectively; and L is the number of items that were administered to the examinee. The constant 100 is a penalty for an incorrect classification. A higher incorrect classification rate, as well as a long test length, results in a higher value of the loss function. Hence, a lower value indicates the better performance of a method.

As to other design aspects of this study, the two-dimensional item pool contained 600 unidimensional items in which one half measured the first dimension and the other half measured the second dimension. The item parameters were generated from the following distributions for both dimensions: a ~ N(1, 0.25^2), b ~ U(−3.6, 3.6), c ~ U(0, 0.3), which was adopted from Chen et al. (2000). For each condition, the two-dimensional latent trait vectors were generated from a multivariate normal distribution for 5000 examinees with a mean vector equal to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and a variance-covariance matrix as $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, where ρ is the correlation between the dimensions that were manipulated in the study. The response data were generated according to the multidimensional three-parameter item-response model (Eq. 1). Because an analysis of variance (ANOVA) was applied to the PCC and ATL, the 5000 examinees were divided into five replications in which each replication contained 1000 examinees. To make sure the final classification decisions were made according to items that were selected from both dimensions and to address the issue of test validity, the minimum and maximum test length on each dimension were set at three and 30 for each examinee, which yielded the shortest and longest test containing six and 60 items, respectively. As for the parameters of SPRT and SPRT_C, α and β were usually each set at .05 or .10; whereas δ was usually set from .1 to .4 to represent a small to large indifference region (Eggen, 1999; Finkelman, 2008; Seitz & Frey, 2013; van Groen et al., 2016). In this study, the parameters α, β, and δ were set at .05, .05, and .2, respectively, as suggested in the preceding research.

Results

The results of the PCC, ATL, PCC per item, and loss for both methods under the different levels of correlation and various cutoff points are presented from left to right in Table 1. The relative efficiencies of the PCC and ATL are listed as well. Below, the results are divided into sections describing the PCC and ATL, which are introduced separately.

Percentage of correct classifications (PCC)

When the two dimensions were highly correlated (ρ = 0.8), both methods yielded almost perfect classifications at the extreme symmetric cutoff points (−3.0, −3.0 and 3.0, 3.0). However, as the cutoff went toward the origin on the two-dimensional space (0.0, 0.0), the PCC decreased to 82.08% and 87.14% for SPRT_C and SPRT, respectively. For other

Table 1 Percentage of correct classification (PCC), average test length (ATL), PCC per item, and LOSS function under various conditions

		PCC			ATL			PCC per item		Average losses	
ρ	Cutoff point	SPRT_C	SPRT	RE	SPRT_C	SPRT	RE	SPRT_C	SPRT	SPRT_C	SPRT
0.8	(3.0, 3.0)	99.78	99.78	1.00	12.11	17.16	0.71	8.24	5.82	12.33	17.38
	(2.0, 2.0)	97.58	97.82	> 0.99	14.79	20.21	0.73	6.60	4.84	17.21	22.39
	(1.5, 1.5)	94.64	95.36	0.99	17.04	24.36	0.70	5.55	3.91	22.40	29.00
	(0.0, 0.0)	82.08	87.14	0.94	24.13	33.76	0.71	3.40	2.58	42.05	46.62
	(−1.5, −1.5)	93.80	95.12	0.99	18.22	26.85	0.68	5.15	3.54	24.42	31.73
	(−2.0, −2.0)	96.96	97.94	0.99	16.19	21.72	0.75	5.99	4.51	19.23	23.78
	(−3.0, −3.0)	99.74	99.88	> 0.99	11.62	16.71	0.70	8.58	5.98	11.88	16.83
	(3.0, 1.5)	97.54	97.58	> 0.99	14.72	19.29	0.76	6.63	5.06	17.18	21.71
	(3.0, 0.0)	93.34	93.32	> 1.00	21.25	24.68	0.86	4.39	3.78	27.91	31.36
	(0.0, 1.5)	90.82	90.56	> 1.00	24.21	28.36	0.85	3.75	3.19	33.39	37.80
	(−3.0, −1.5)	97.60	97.58	> 1.00	16.53	21.54	0.77	5.90	4.53	18.93	23.96
0.5	(3.0, 3.0)	99.68	99.76	> 0.99	16.78	17.16	0.98	5.94	5.81	17.10	17.40
	(2.0, 2.0)	97.94	98.14	> 0.99	19.54	20.06	0.97	5.01	4.89	21.60	21.92
	(1.5, 1.5)	94.88	95.20	> 0.99	22.92	24.35	0.94	4.14	3.91	28.04	29.15
	(0.0, 0.0)	83.92	86.96	0.97	30.01	33.34	0.90	2.80	2.61	46.09	46.38
	(−1.5, −1.5)	94.16	95.08	0.99	25.66	26.73	0.96	3.67	3.56	31.50	31.65
	(−2.0, −2.0)	97.28	97.72	> 0.99	21.46	21.73	0.99	4.53	4.50	24.18	24.01
	(−3.0, −3.0)	99.76	99.86	> 0.99	16.01	16.72	0.96	6.23	5.97	16.25	16.86
	(3.0, 1.5)	97.38	97.42	> 0.99	16.72	19.23	0.87	5.83	5.07	19.34	21.81
	(3.0, 0.0)	93.08	93.06	> 1.00	21.80	24.34	0.90	4.27	3.82	28.72	31.28
	(0.0, 1.5)	90.64	90.98	> 0.99	25.30	28.23	0.90	3.58	3.22	34.66	37.25
	(−3.0, −1.5)	97.26	97.54	> 0.99	18.12	21.43	0.85	5.37	4.55	20.86	23.89
0.0	(3.0, 3.0)	99.86	99.86	1.00	17.20	17.20	1.00	5.80	5.80	17.34	17.34
	(2.0, 2.0)	97.88	97.88	1.00	20.32	20.32	1.00	4.82	4.82	22.44	22.44
	(1.5, 1.5)	94.68	94.68	1.00	24.42	24.42	1.00	3.88	3.88	29.74	29.74
	(0.0, 0.0)	86.88	86.88	1.00	33.41	33.41	1.00	2.60	2.60	46.53	46.53
	(−1.5, −1.5)	95.94	95.94	1.00	26.52	26.52	1.00	3.62	3.62	30.58	30.58
	(−2.0, −2.0)	97.84	97.84	1.00	21.72	21.72	1.00	4.51	4.51	23.88	23.88
	(−3.0, −3.0)	99.80	99.80	1.00	16.73	16.73	1.00	5.96	5.96	16.93	16.93
	(3.0, 1.5)	97.34	97.34	1.00	19.28	19.28	1.00	5.05	5.05	21.94	21.94
	(3.0, 0.0)	93.00	93.00	1.00	24.49	24.49	1.00	3.80	3.80	31.49	31.49
	(0.0, 1.5)	90.90	90.90	1.00	28.21	28.21	1.00	3.22	3.22	37.31	37.31
	(−3.0, −1.5)	97.82	97.82	1.00	21.43	21.43	1.00	4.56	4.56	23.61	23.61

Note. *ρ* = correlation between dimensions, *SPRT* sequential probability ratio test, *SPRT_C* sequential probability ratio test with conditional latent trait distribution, *RE* relative efficiency. *RE > 1.00 indicates 1.01 > RE > 1.00; RE > 0.99 indicates 1.00 > RE > 0.99* cutoff points, such as the middle symmetric, (−1.5, −1.5) and (1.5, 1.5), and asymmetric, (3.0, 1.5), (−3.0, −1.5), (0.0, 1.5), and (3.0, 0.0), both methods yielded PCCs greater than 90% (ranging from 90.56% to 97.60%). For the ρ = 0.8 conditions, the relative efficiency ranged from 0.94 to 1.00, which indicated that SPRT_C performed comparably to SPRT. A similar pattern was found in the moderately correlated condition (ρ = 0.5). The highest PCCs were found for the extreme symmetric cutoff points, whereas the lowest PCC was found at the cutoff (0.0, 0.0). When the two dimensions were uncorrelated, SPRT_C performed exactly the same as SPRT. In general, the resulting relative efficiency ranged from 0.94 to 1.00, which means SPRT_C yielded PCCs fairly comparable to SPRT in all the simulated conditions.

An ANOVA of the mean PCC revealed that the three-way interaction effect of the correlation between dimensions, cutoff points, and termination criteria was significant [F(5.76, 34.65) = 14.608, partial η^2 = .712, η^2 = .005]. Simple main effects between the two termination criteria in each condition of the correlation between the dimensions by the cutoff points were also examined. Only one of them [ρ = 0.8, cutoff points = (0.0, 0.0)] had a mean difference of 5.06% favoring the SPRT criterion after a Bonferroni correction. The results also indicated that SPRT and SPRT_C performed quite comparably, except for a high correlation between the dimensions and a cutoff point close to zero.

Average test length (ATL)

When the correlation between dimensions was 0.8, SPRT_C generally yielded a shorter test length than SPRT, with a decrement ranging from 3.43 to 9.63 items (i.e., 14% to 32% of the test length), which caused the RE of the ATL to range from 0.68 to 0.86. When the two dimensions were moderately correlated, the RE of the ATL increased and ranged from 0.85 to 0.99 (i.e., saving 1% to 15% of the test length). For uncorrelated dimensions, SPRT_C yielded an identical test length as SPRT. In general, SPRT_C resulted in a shorter test length than SPRT. The higher the between-dimension correlation, the shorter the SPRT_C test.

An ANOVA of the mean ATL revealed that the three-way interaction effect of the correlation between dimensions, cutoff points, and termination criteria was significant [F(6.679, 12.84) = 21.694, partial η^2 = 0.988, η^2 = 0.010]. With a Bonferroni correction, simple main effects between the two termination criteria in each condition of the correlation between dimensions by the cutoff points were all significant and favored the SPRT_C criterion (mean differences ranging from 3.43 to 9.63 when ρ = 0.8, and from 0.38 to 3.32 when ρ = 0.5) except for the conditions when ρ = 0, as well as when ρ = 0.5 and the cutoff point = (2, 2), (−2, −2), (1.5, 1.5), and (−1.5, −1.5) (mean differences ranging from 0 to 2.926). The results again indicated that the test lengths were shorter (i.e., the measurement efficiency was better) for SPRT_C than for SPRT when the two dimensions were correlated moderately or highly. The higher the correlation, the better the measurement efficiency.

To simultaneously take the PCC and ATL into consideration in comparing these two methods, the PCCs that each item could contribute were also calculated. As previously, both methods were found to yield the highest PCCs per item at the extreme symmetric cutoff points and the lowest PCCs per item at the origin. For the conditions in which the dimensions were moderately correlated, these two methods yielded similar PCCs per item, where SPRT_C resulted in a slightly higher PCC for each item than its SPRT counterpart. Under the highly correlated conditions, for the cutoff points for which both methods yielded comparable PCCs, (3.0, 3.0), (−3.0, −3.0), (0.0, 1.5), and (3.0, 0.0), each item could contribute 3.74 to 8.57 and 3.17 to 5.97 of the PCC in SPRT_C and SPRT, respectively. Even for other cutoff points in which SPRT_C exhibited slightly lower PCCs, (1.5, 1.5), (0.0, 0.0), and (−1.5, −1.5), each item could contribute PCCs ranging from 3.33 to 5.63 in SPRT_C relative to their SPRT counterparts that ranged from 2.53 to 5.97. In general, SPRT_C yielded comparable to higher measurement efficiency than SPRT under all the simulated conditions. The higher the correlation, the more efficient SPRT_C, which indicated that SPRT_C could generally take advantage of the correlation between the dimensions to improve the measurement efficiency of an MCCT.

It is noteworthy that SPRT_C yielded lower PCCs than SPRT on the cutoffs (0.0, 0.0) and (−1.5, −1.5), respectively. To determine the reason underlying this phenomenon, all the examinees were divided into three categories according to the results of the classification decisions: correct on both dimensions, correct on one dimension, and incorrect on both dimensions. The results are listed in Table 2. For each category, the number of examinees was counted, and their corresponding ATL was calculated. In general, SPRT_C accumulated fewer examinees than SPRT in the first category, with most of them falling into the second category. For example, for the ρ = 0.8 conditions, SPRT_C accumulated 348 examinees less than SPRT in correct decisions on both dimensions; 276 out of them were correctly classified on only one dimension.

Furthermore, we plotted the first 1000 examinees with their corresponding classification decisions on a two-dimensional plane for SPRT_C and SPRT on two cutoffs under highly correlated conditions (ρ = 0.8). The results at cutoffs (0.0, 0.0) and (−1.5, −1.5) are plotted in Figs. 1 and 2, respectively. An examinee who was classified correctly on both dimensions was marked as one;

Table 2 Number count of examinees and their averaged test length of different type of classification decision results under various conditions

		Correct on both dimensions				Correct on one dimension				Incorrect on both dimensions			
		# of examinees		ATL		# of examinees		ATL		# of examinees		ATL	
ρ	Cutoff point	SPRT_C	SPRT	SPRT_C	SPRT	SPRT_C	SPRT	SPRT_C	SPRT	SPRT_C	SPRT	SPRT_C	SPRT
0.8	(3.0, 3.0)	4989	4989	12.04	17.10	10	10	44.50	43.00	1	1	60.00	60.00
	(2.0, 2.0)	4879	4891	14.23	18.84	100	99	34.10	37.50	-	-	-	-
	(1.5, 1.5)	4732	4768	15.37	23.19	234	221	45.38	48.00	34	11	54.79	57.82
	(0.0, 0.0)	4104	4357	19.76	31.53	833	594	43.71	48.11	63	49	49.68	57.96
	(−1.5, −1.5)	4690	4756	16.37	25.68	264	234	45.01	49.28	46	10	52.52	59.80
	(−2.0, −2.0)	4848	4897	16.09	21.10	130	97	34.78	39.30	-	-	-	-
	(−3.0, −3.0)	4987	4994	11.52	16.68	9	6	47.33	47.67	4	-	53.25	-
	(3.0, 1.5)	4877	4879	14.23	18.84	123	121	34.10	37.50	-	-	-	-
	(3.0, 0.0)	4667	4666	20.45	23.84	333	334	32.39	36.41	-	-	-	-
	(0.0, 1.5)	4541	4528	23.19	27.20	457	465	34.22	39.34	2	7	40.00	55.57
	(−3.0, −1.5)	4880	4879	16.09	21.10	120	121	34.78	39.30	-	-	-	-
0.5	(3.0, 3.0)	4984	4988	16.69	17.11	16	12	45.50	39.00	-	-	-	-
	(2.0, 2.0)	4897	4907	16.16	18.76	94	91	37.45	36.92	-	-	-	-
	(1.5, 1.5)	4744	4762	21.58	22.00	224	214	46.92	46.50	32	24	54.94	54.54
	(0.0, 0.0)	4196	4348	27.55	31.55	780	629	42.51	44.88	24	23	55.00	56.00
	(−1.5, −1.5)	4708	4754	24.30	25.83	271	240	47.24	43.97	21	6	51.57	56.17
	(−2.0, −2.0)	4864	4886	17.55	20.99	120	113	38.38	38.82	-	-	-	-
	(−3.0, −3.0)	4988	4993	15.92	16.68	9	7	51.56	45.43	3	-	50.33	-
	(3.0, 1.5)	4869	4871	16.16	18.76	131	129	37.45	36.92	-	-	-	-
	(3.0, 0.0)	4654	4653	20.95	23.44	346	347	33.34	36.40	-	-	-	-
	(0.0, 1.5)	4532	4549	24.10	27.00	465	445	36.79	40.34	3	6	52.00	60.00
	(−3.0, −1.5)	4863	4877	17.55	20.99	137	123	38.38	38.82	-	-	-	-
0.0	(3.0, 3.0)	4993	4993	17.18	17.18	7	7	37.29	37.29	-	-	-	-
	(2.0, 2.0)	4894	4894	18.80	18.80	106	106	36.98	36.98	-	-	-	-
	(1.5, 1.5)	4734	4734	23.44	23.44	262	262	41.61	41.61	4	4	60.00	60.00
	(0.0, 0.0)	4344	4344	31.75	31.75	632	632	44.01	44.01	24	24	55.63	55.63
	(−1.5, −1.5)	4797	4797	25.84	25.84	197	197	42.21	42.21	6	6	57.00	57.00
	(−2.0, −2.0)	4892	4892	21.05	21.05	105	105	38.82	38.82	-	-	-	-
	(−3.0, −3.0)	4990	4990	16.69	16.69	10	10	37.60	37.60	-	-	-	-
	(3.0, 1.5)	4867	4867	18.80	18.80	133	133	36.98	36.98	-	-	-	-
	(3.0, 0.0)	4650	4650	23.58	23.58	349	349	36.46	36.46	1	1	60.00	60.00
	(0.0, 1.5)	4545	4545	26.98	26.98	448	448	40.17	40.17	7	7	55.86	55.86
	(−3.0, −1.5)	4891	4891	21.05	21.05	109	109	38.82	38.82	-	-	-	-

Note. ρ = correlation between dimensions, *ATL* average test length, *SPRT* sequential probability ratio test, *SPRT_C* sequential probability ratio test with conditional latent trait distribution, “-” = 0

otherwise, the examinee was marked as zero. In Fig. 1, all the zeros were distributed near the origin. More specifically, the zeros were distributed evenly in the four quadrants in SPRT, whereas most of the zeros were scattered in the second and fourth quadrants in SPRT_C. That is, for examinees with one ability above the average and the other below the average, the application of between-dimensional correlation in SPRT_C might have a negative influence on PCC. For cutoff (−1.5, −1.5), where the examinees are plotted in Fig. 2, the same findings still hold.

Conclusion

Classification tests are commonly used in education, psychology, and personnel selection fields. Taking the educational field for instance, the Preliminary Scholastic Aptitude Test (PSAT), which is a nationwide, multiple-choice test in the United States, is used to assist in identifying students’ academic strengths and weaknesses mainly in reading and mathematics and provide them with practice and assistance for the Scholastic Aptitude Test (SAT), which is typically taken the following

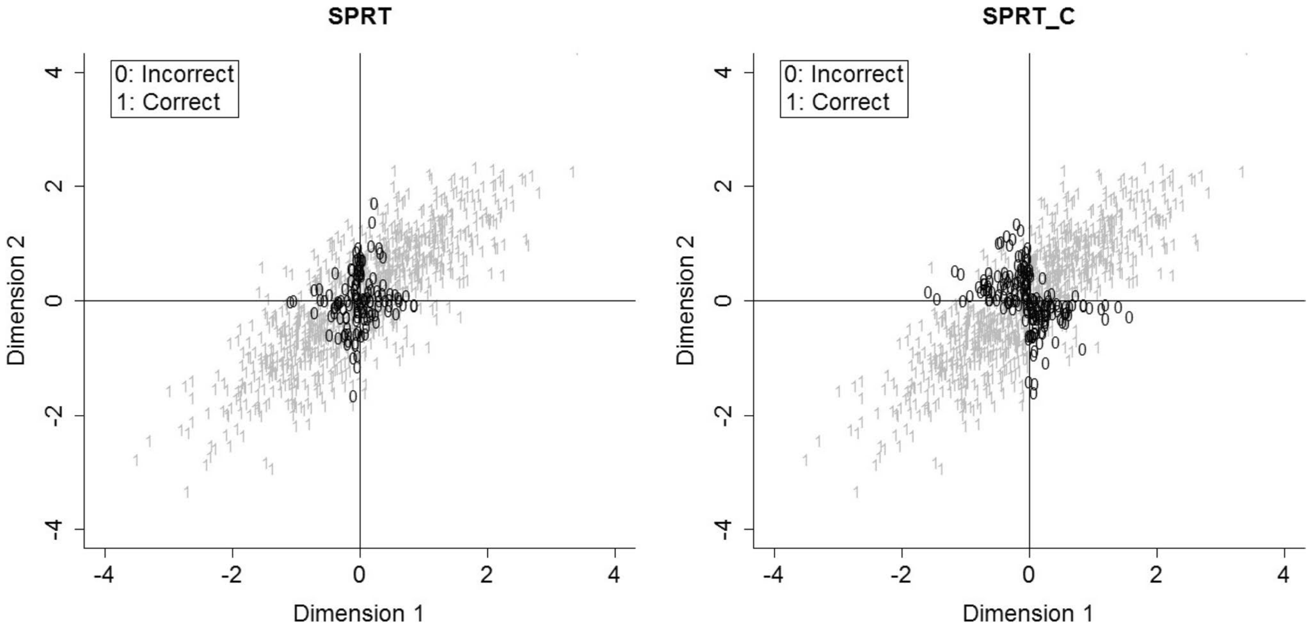

Fig. 1 Plots of correctness of classification decisions for 1000 examinees under between-dimension $\rho = 0.8$ and cutoff (0.0, 0.0)

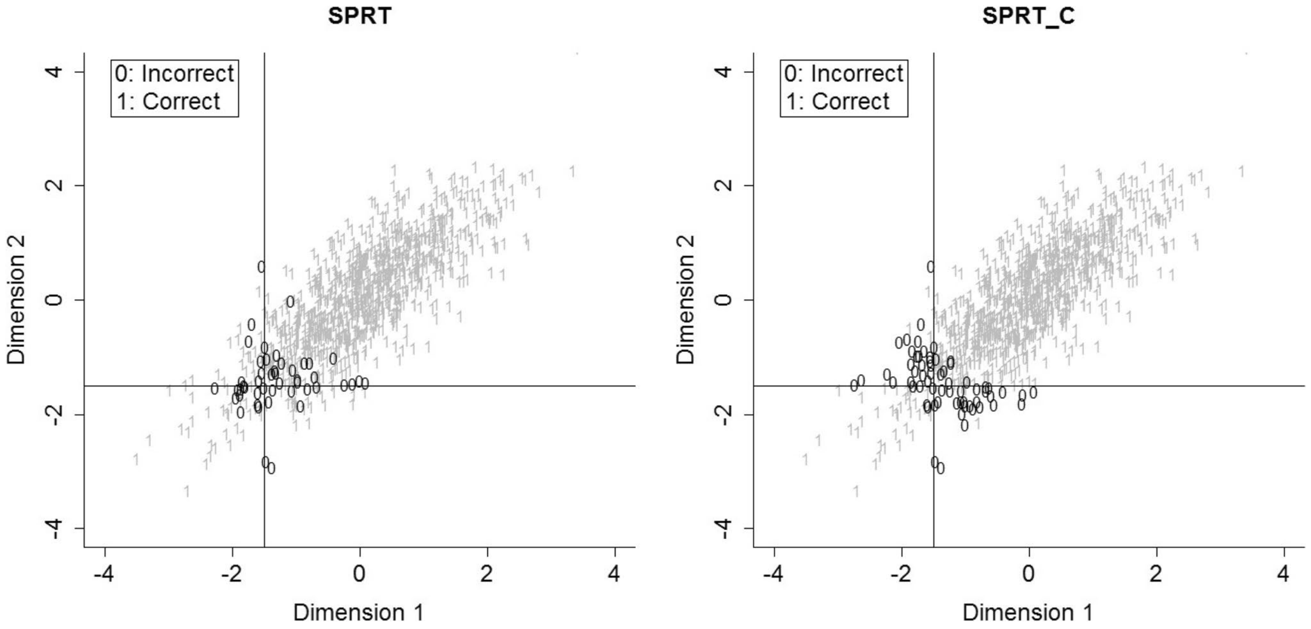

Fig. 2 Plots of correctness of classification decisions for 1000 examinees under between-dimension $\rho = 0.8$ and cutoff (−1.5, −1.5)

academic year. Operating under the assumption that a test can measure more than one latent trait, this kind of test can be used to make decisions on multiple dimensions. In reality, there are numerous testing programs that now require reporting subscale scores for different objectives (Yao & Boughton, 2007), including making decisions. These tests help decision-makers on a daily basis, and the results directly affect the lives of countless individuals. It is certainly of great value to improve the measurement efficiency of these classification tests, especially when a test is multidimensional.

For the conditions in which multiple latent traits are to be measured simultaneously, an MCAT applies the collateral information of the between-dimensional correlations to the estimation of the latent traits, which has been found to largely improve its measurement efficiency (Wang & Chen, 2004). In the current study, to apply such collatera information to an MCCT, a conditional latent trait distribution was used in the SPRT. This approach is easy to implement in the SPRT and can be effective. Through a compendious simulation study, the proposed SPRT_C was found to largely shorten the test length (up to 32%) when the two dimensions were highly correlated; it was able to save up to 15% of the test length when the two dimensions were moderately correlated. As expected, more items can be saved when the correlation between the dimensions moves from moderate to high. Even when the dimensions were less correlated or uncorrelated, SPRT_C could perform in quite a similar manner as its multi-unidimensional counterpart (i.e., SPRT in this study). Furthermore, the performance of SPRT_C was found to depend on the location of examinees relative to the cutoff point. The mean ability on both dimensions was both zero in this study, where SPRT_C resulted in the longest and shortest ATL on the cutoff point (0.0, 0.0) and extreme points (−3.0, −3.0) and (3.0, 3.0), respectively. The results confirmed Nydick's (2014) findings in which he focused on CCT that was based on a unidimensional binary response model. Within the SPRT_C method, although the expected values of the conditional latent trait distribution are required, all these values can be calculated easily before the test proceeds. Therefore, this method can be used widely to make classification decisions in actual practice.

By shortening the test length, the measurement efficiency of an MCCT can be improved and convey many advantages. First, because the average test length of making classification decisions for examinees was shorter in SPRT_C, this means fewer items are required to make classification decisions with SPRT_C. Because the items that had the most information at the cutoff point were selected, SPRT and SPRT_C tended to yield the same sequence of selected items. Hence, items that were selected and examined by SPRT but not by SPRT_C (usually in the late stages of the test) will be exposed less and replaced less frequently. Therefore, the effort and workload of maintaining an item bank, as well as writing new items, can be lowered to some extent. This can help to achieve cost-effectiveness for many test programs. Second, respondents might be more willing to participate in a survey if they are told the length of a test is shorter. In addition, the tedium and carelessness of respondents can be diminished to some extent, which can in turn reduce the likelihood of invalid responses (Forbey & Ben-Porath, 2007; Schmidt et al., 2003). Moreover, a shorter test length can introduce a reduction in administration time, as well as patients' and clinicians' burdens, which is quite valuable for clinical assessments (e.g., mental health) (Gibbons et al., 2008).

A multidimensional classification test can be used to make an overall decision over multiple dimensions or multiple decisions with one for each of the dimensions. Though the former approach is more popular in the MCCT literature (Spray et al., 1997; van Groen et al., 2016), this study aimed to go with the latter approach for a number of reasons. First, for educational and psychological usage, this approach can provide more detailed diagnostic information for practitioners. For example, a natural science test might contain items that are designed to measure at least one of a number of multiple disciplines, such as physics, chemistry, biology, geology, and astronomy. The results of the test can be an overall pass/fail decision on natural science. Nevertheless, making decisions for each dimension can provide score profiles to help a teacher diagnose students' strengths or weaknesses in these more specific disciplines (Luecht, 1996). Second, the procedures that follow the decisions can be more efficient and cost-effective. With such subscale information, subsequent lessons or plans can be more appropriately designed. For example, for jobs that require multiple skills, knowing applicants' profile information on each skill can help human resource staff to appropriately arrange further training courses for newly recruited employees. All of these examples show how valuable this subscale information can be and why it should be collected whenever possible.

As to the choice of values for parameters α, β, and δ, it is generally recommended to use .05, .05, and .2, respectively, since many studies used the set of values and found it performed well in the CCT scenario (Eggen, 1999; Finkelman, 2008; Seitz & Frey, 2013; van Groen et al., 2016). The cutoff points used here were set to investigate the performance of the SPRT_C, therefore various combinations were manipulated. In reality, the cutoff point should be set according to the purpose of classification. For example, a common consensus to identify gifted and talented children is to set the cutoff points at 2.0 (or 1.96) standard deviations (SD) above the population mean on each dimension of the intelligence test. To identify disabled students for further remedial teaching, the cutoff points can be set at 1.5 or 2.0 SDs below the population mean on latent continuums.

There might be limitations when applying the SPRT_C to a practical scenario. For example, the effect of α, β, and δ to the SPRT_C was investigated through another small simulation study that only contains cutoff points (−3.0, −3.0), (0.0, 0.0), and (3.0, 3.0). All three parameters were found to have a limited effect on PCC for both SPRT and SPRT_C methods. Parameter α and β showed their effects on ATL for cutoff points (0.0, 0.0), (−3.0, −3.0), and (3.0, 3.0), (0.0, 0.0), respectively, whereas the size of δ, as well as its interaction with the other two parameters, can shorten ATL sizably. The results were not included in this paper due to the findings were quite similar to the current study, but a note should be made here. For shorter tests (e.g., less than 10 items), the differences in the performance between the SPRT and the SPRT_C become smaller, which might imply a limitation that the benefit of using the SPRT_C might be diminished for short tests. Additionally, the minimum and maximum test lengths on each dimension were set at 3 and

30 for each examinee, yet that might be not good enough to address the issue of test validity. Instead, the content balancing procedure might be more useful, and its effects on the SPRT_C should be investigated in further studies.

As for further research, this study focused on an MCCT that proceeds with dichotomous items. How SPRT_C performs for an MCCT with polytomous items can be further investigated. Additionally, several item selection procedures and termination criteria have been proposed to improve measurement efficiencies, such as the expected log-likelihood ratio method (Nydick, 2014) and Kullback-Leibler information (Eggen, 1999) for item selection procedures; and the generalized likelihood ratio (GLR; Bartroff et al., 2008) and SPRT with stochastic curtailment (SCSPRT; Finkelman, 2008) for test termination criteria. How the findings of the SPRT_C can be further applied as well as the improvement of measurement efficiency when combined with these methods can be further explored. In addition, the item selection procedure used in this study selects items that yielded maximum information at cutoff points, which means it is a nonadaptive procedure. Some studies (e.g., van Groen et al., 2016) have tried to make classification decisions by selecting the item that results in the largest decrement in the volume of the confidence ellipsoid at the current latent-trait level (i.e., adaptive) in the MCAT scenario (Segall, 1996). The performance of applying this kind of item selection strategy to MCCT combined with SPRT_C is of interest. Moreover, the SPRT_C method showed slightly poorer performances than SPRT for examinees whose two latent traits fell on different sides of the mean of latent trait distribution (usually at the origin) when the between-dimension correlation was high and the cutoff was set at (0.0, 0.0). Though the cutoff points were usually set at the extreme rather than the mean of the latent continuum, our findings imply that the SPRT_C method should be used with caution when the cutoff points are set at or near the mean of the latent trait distribution. How this drawback can be corrected in SPRT_C needs further investigation. Furthermore, whether the measurement efficiency of an MCCT system that provides an overall classification decision with subscale classification or diagnosis information can be improved by SPRT_C is also of interest.

Acknowledgments C.L.S. was supported by a Ministry of Science and Technology (Taiwan) grant MOST 105-2410-H-110 -048. The authors declare no competing financial interests.

References

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its application to computerized adaptive testing. *Psychometrika*, *73(3)*, 473-486.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Addison-Wesley.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) manual for administration and scoring*. University of Minneapolis Press.

Chen, S.-Y., Ankenmann, R. D., & Chang, H. -H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24(3)*, 241-255.

Eaton, Morris L. (1983). *Multivariate Statistics*: a Vector Space Approach. John Wiley and Sons.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23(3)*, 249-261.

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*(5), 713-734.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics*, *33*(4), 442-463.

Forbey, J. D., & Ben-Porath, Y. S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. *Psychological Assessment, 19(1)*, 14-24.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric services*, *59*(4), 361-368.

Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior research methods*, *47(2)*, 549-561.

Jensen, J (2000). *Statistics for Petroleum Engineers and Geoscientists*. Elsevier.

Jones, W. P. (2014). Enhancing a short measure of Big Five personality traits with Bayesian scaling. *Educational and Psychological Measurement*, *74*(6), 1049–1066.

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive testing strategy for mastery decisions* (Research Report 79–5). Minneapolis, M.N.: University of Minnesota Press.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257-283). Academic Press.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20(4)*, 389-404.

Nydick, S. W. (2014). The sequential probability ratio test and binary item response models. *Journal of Educational and Behavioral Statistics*, *39(3)*, 203-230.

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). Academic Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health*, *14(8)*, 1101-1108.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8(2)*, 206-224.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331-354.

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling*, *55(1)*, 105-123.

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.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics*, *21(4)*, 405-414.

Spray, J. A., Abdel-fattah, A. A., Huang, C.-Y., & Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional*. (Research Report No.97-5). Iowa City, IA: ACT.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69(5)*, 778-793.

van der Linden, W. J., & Glas, C. (Eds.). (2000). *Computer adaptive testing: Theory and practice*. Kluwer Academic Publishers.

van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2014). Item Selection Methods Based on Multiple Objective Approaches for Classifying Respondents Into Multiple Levels. *Applied Psychological Measurement*, *38(3)*, 187–200. https://doi.org/10.1177/0146621613509723

van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement*, *40*(6), 387-404.

van Groen, M. M., Eggen, T. J., & Veldkamp, B. P. (2019). Multidimensional Computerized Adaptive Testing for Classifying Examinees. In *Theoretical and Practical Advances in Computer-based Educational Measurement* (pp. 271-289). Springer.

Vos, H. J. (2000). Bayesian procedure in the context of sequential mastery testing. *Psicologica, 21*, 191-211.

Wald, A. (1947). *Sequential analysis*. John Wiley.

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 450-480.

Wang, Z., Wang, C., & Weiss, D. (2019). *Grid Multi-classification Adaptive Classification Testing with Multidimensional Polytomous Items*. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/209022. Accessed 1 Aug 2021.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31(2)*, 83-105.

Open Practices Statement The data for all simulation experiments are available at https://osf.io/v4a3j, and none of the simulation experiments were preregistered.