Practical Assessment, Research, and Evaluation

Volume 12 *Volume 12, 2007* Article 1

2007

A Practitioner's Guide for Variable-length Computerized Classification Testing

Nathan A. Thompson

Recommended Citation

Thompson, Nathan A. (2007) "A Practitioner's Guide for Variable-length Computerized Classification Testing," *Practical Assessment, Research, and Evaluation*: Vol. 12 , Article 1.
DOI: https://doi.org/10.7275/fq3r-zz60
Available at: https://scholarworks.umass.edu/pare/vol12/iss1/1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 1, January 2007 ISSN 1531-7714

A Practitioner's Guide for Variable-length Computerized Classification Testing

Nathan A. Thompson
Thomson Prometric

Variable-length computerized classification tests, CCTs, (Lin & Spray, 2000; Thompson, 2006) are a powerful and efficient approach to testing for the purpose of classifying examinees into groups. CCTs are designed by the specification of at least five technical components: psychometric model, calibrated item bank, starting point, item selection algorithm, and termination criterion. Several options exist for each of these CCT components, creating a myriad of possible designs. Confusion among designs is exacerbated by the lack of a standardized nomenclature. This article outlines the components of a CCT, common options for each component, and the interaction of options for different components, so that practitioners may more efficiently design CCTs. It also offers a suggestion of nomenclature.

Variable-length computerized classification tests (CCT: Lin & Spray, 2000; Thompson, 2006; Parshall, Spray, Kalohn, & Davey, 2006) offer a sophisticated testing methodology designed to classify examinees into groups with the goal of maximizing the efficiency of the test, namely reducing classification error while using as few items as necessary. Variable-length CCT research primarily investigates new alternatives in CCT design that are proposed to further this goal. However, because several alternatives exist for the specification of each component in variable-length CCT, there are an extensive number of possible designs, creating possible confusion both in its development and in the relevant nomenclature. The purpose of this article is to organize variable-length CCT design and its nomenclature by first delineating the required components of a variable-length CCT, identifying commonly used alternatives for each component, and making recommendations regarding the interaction of components.

First, the term "computerized classification test" is a general expression that, upon initial inspection, can refer to any test for examinee classification administered by computer. However, it has been used in research to refer to a specific type of test, a variable-length computerized test (Lin & Spray, 2000; Parshall, Spray, Kalohn, & Davey, 2006). A variable-length test is a computerized assessment where not every examinee receives a test of the same length because the test will terminate when it has accomplished its purpose. In CCT, this occurs when the test is able to classify an examinee, while in computerized adaptive testing it often occurs when a point estimate of ability reaches a certain amount of precision.

It is proposed that the acronym VL-CCT be used to refer to variable-length CCT rather than using the term CCT, and the term CCT be reserved for the broader topic of classification exams administered by computer. Many classification tests administered by computer are not variable length,

such as linear-on-the-fly tests or computerized administrations of traditional fixed-form tests. Similarly, it is proposed that the term "adaptive," which is sometimes used to refer to variable-length computerized tests (Spray & Reckase, 1996), be reserved for a test that adapts to an examinee using *individual examinee information*, namely a test that selects items based on examinee responses. A test that maximizes information at a cutscore does not take the examinee into account, only aspects of the items, and is therefore not truly adaptive.

The broader term "classification" is used, following Parshall, Spray, Kalohn, and Davey (2006), rather than the often-used "mastery" (Kingsbury & Weiss, 1983; Yang, Poggio, & Glasnapp, 2006) because while most VL-CCT applications are for the commonly encountered mastery test, the VL-CCT methods are generalizable to classification for more than two groups (e.g., Spray, 1993; Eggen & Straetmans, 2000; Jiao, Wang, & Lau, 2004).

DESIGN COMPONENTS

VL-CCT requires the specification of five mandatory design components, similar to the components of computerized adaptive testing (Weiss & Kingsbury, 1984). The components are outlined in Table 1. Additional components addressing practical issues such as item exposure and content constraints are often imposed, but these are not mandatory for the creation of a VL-CCT, and will be addressed separately. Required VL-CCT components include:

1. Psychometric model
2. Calibrated item bank
3. Starting point
4. Item selection algorithm
5. Termination criterion (classification/scoring procedure).

Table 1: CCT Components

CCT Component	Available options	Example References
Psychometric Model	Classical Test Theory	Linn, Rock, & Cleary, 1972; Frick, 1992; Rudner, 2002
	Item Response Theory	Reckase, 1983; Kingsbury & Weiss, 1983; Lau & Wang, 1998; Eggen & Straetmans, 2000
Item Bank (peakedness not always reported)	Peaked	Xiao, 1999
	Not peaked	Kingsbury & Weiss, 1983; Finkelman, 2003; Yang, Poggio, & Glasnapp, 2006
Starting point	Default (LR = 1 or θ = 0.0)	Most extant research
	Previous information	Yang, Poggio, & Glasnapp, 2006
Item Selection	Estimate-based	Reckase, 1983; Kingsbury & Weiss, 1983; Eggen, 1999
	Cutscore-based	Spray & Reckase, 1994, 1996; Eggen, 1999
	Global (Mutual)	Weissman, 2004
Termination Criterion	SPRT	Reckase, 1983; Eggen, 1999; Eggen & Straetmans, 2000
	ACI	Kingsbury & Weiss, 1983; Eggen & Straetmans, 2000; Chang, 2006
	Bayesian decision theory	Vos, 2000; Glas & Vos, 2006

VL-CCT differs from computerized adaptive testing for point estimation of ability in that the termination criterion and the classification/scoring procedure are the same thing; the test is terminated when the examinee is classified. If the purpose of the test is a point estimate of examinee ability, then the two are considered separately (Weiss & Kingsbury, 1984).

Psychometric Model

The first step in the technical development of a VL-CCT is the selection of a psychometric model that will be used as a basis for the remaining components. Both prevailing psychometric theories, classical test theory (CTT) and item response theory (IRT) can be used in the development of VL-CCT. Both require the sampling of examinees to obtain calibrations of item parameters, but the use of CTT requires the additional constraint of being able to distinguish between members of the intended groups through other means.

CTT requires distinguishable groups because VL-CCT with a classical model is designed to have separate item parameters for each group (Frick, 1992; Rudner, 2002), specifically the difficulty or proportion-correct statistics. For example, a test for licensure may obtain a sample of licensed, practicing professionals and another sample of students being educated to enter the profession. Theoretically, more practicing professionals will answer the item correctly, and the test is designed to make use of this differentiation. CTT offers the advantage of conceptual simplicity and applicability to small samples, but can still be used to design highly efficient VL-CCTs (Rudner, 2002).

IRT is actually a family of models that offers a powerful alternative to CCT, but with the drawback that it requires a much larger calibration sample, up to 1,000 examinees (Wainer & Mislevy, 2000). IRT evaluates the probability of a correct response across all levels of ability (θ), with fine distinction rather than only across broad groups. A standard-setting study must be performed to determine a cutscore or cutscores on θ. While the majority of VL-CCT research has employed dichotomous IRT (e.g., Reckase, 1983; Spray & Reckase, 1996; Eggen, 1999) it is also possible to use polytomous IRT (Lau & Wang, 1998) such as the generalized partial credit model (Muraki, 1992). An advantage of IRT is that it places items and examinees on a common scale (θ), as well as the cutscore. This generally applicable scale facilitates sophisticated methods for the remaining components.

Calibrated item bank

The optimal characteristics of the item bank to be used are determined by other components of the VL-CCT. Obviously, the calibration procedure depends upon the psychometric model selected, and the structure of the item bank is derived from the content specifications of the test, often based on a job analysis. The ranges of desired item statistics depend on the intended item selection algorithm. For example, if the item selection algorithm will select items with difficulty values near the cutscore on the θ metric, then many items will be needed with difficulty parameters in this region. If the item selection algorithm will match item difficulty to estimated examinee θ, then a wide range of difficulty parameters are necessary because there may be a wide range of examinee ability. Unfortunately, detailed information regarding the item bank, especially the very relevant characteristic of peakedness, is not always reported in VL-CCT research.

An important question to ask when developing the item bank is as to how many items will be needed. The answer lies in several issues. If the test is very high stakes and only a small amount of classification error can be tolerated, such as a medical licensure examination, then more items are needed than a test whose stakes are lower, such as a test of corporate training retention. If IRT is used, then fewer items are needed if the items provide a high amount of information, which can occur when they have relatively high discrimination values or a polytomous IRT model. Ideally, simulation studies will be carried out to determine the efficiency of the proposed procedure, which includes the characteristics of the item bank. If the results of the simulation study are not satisfactory, other choices of components may be explored.

Starting point

If previous information is available concerning individual examinees, this can be used to modify the starting point of the test (Weiss & Kingsbury, 1984; Yang, Poggio, & Glasnapp, 2006). For example, if the test must be periodically retaken for recertification in a profession, then previously certified examinees may warrant a higher initial estimate of ability than examinees that are taking the test for the first time. Their previous score itself may be used, if available. This is not often the case, though, and the default values of a termination criterion are usually used, such as a likelihood ratio of 1.0 (an even ratio) or examinee ability at the mean of the population.

Item selection algorithm

The most basic item selection algorithm for VL-CCT is random item selection (Kingsbury & Weiss, 1983). At each point in the test, an item is randomly selected from the bank. Unfortunately, this makes no use of known information regarding neither the items nor the examinees, and therefore sacrifices efficiency.

A more appropriate approach is intelligent item selection, where the computer evaluates the unadministered items in the bank and decides which would be the "best" to administer next. While this is conceptually straightforward, it is not operationally so, as there are different methods to quantify exactly what constitutes "best." Moreover, certain methods are more appropriate for or can only be used with certain termination criteria, item bank characteristics, and psychometric models. Given the number of methods available, and the differences in appropriateness regarding other components, a large amount of VL-CCT research as focused on item selection (e.g., Spray & Reckase, 1994; Eggen, 1999; Lin & Spray, 2000).

Intelligent item selection methods can generally be classified into two types, cutscore-based and estimate based (Thompson, 2006). Cutscore-based methods seek to maximize the amount of information that items provide at the cutscore (with IRT), or equivalently to differentiate between the two groups divided by the cutscore (classical). Estimate-based methods seek select the next item based on an estimate of examinee ability, regardless of the cutscore location.

When the psychometric model is classical test theory, three cutscore-based methods have been proposed (Rudner, 2002): maximum discrimination, information gain, and minimum expected cost. Maximum information is the conceptually simplest; the next item chosen is that which offers the best discrimination between the groups, or has the greatest difference in the proportion P of correct responses in each group from the calibration sample. Rudner (2002) quantifies this as

$$M_i = \left| \log \frac{P(z_i = 1 \mid m_k)}{P(z_i = 1 \mid m_{k+1})} \right| \qquad (1)$$

where $z_i = 1$ is a correct response to item i, and the upper mastery group being considered is m_{k+1}. For example, an item that is correctly answered by 0.80 of masters m_{k+1} and 0.30 of nonmasters m_k is more discriminating than an item that was correctly answered by 0.60 of masters and 0.50 of nonmasters.

Information gain evaluates the same concept through a more sophisticated quantification that employs Shannon's information (1948) index of entropy,

$$H(S) = \sum_{k=1}^{K} - P_k \log_2 P_k \, . \qquad (2)$$

Information gain is maximized with the greatest reduction of entropy,

$$H(S_0) - H(S_i) \qquad (3)$$

where $H(S_0)$ is the current level of entropy and $H(S_i)$ is the expected entropy after item i, expressed as

$$H(S_i) = P(z_i = 1)H(S_i \mid z_i = 1) + P(z_i = 0)H(S_i \mid z_i = 0) \qquad (4)$$

Information gain increases with a greater difference between the group difficulty statistics (Rudner, 2002).

Minimum expected cost is a Bayesian criteria that fits with the Bayesian decision theory framework that is applied to VL-CCT with classical

test theory. It assumes that the test user can arbitrarily specify the relative cost or loss of each type of classification error, and seeks to minimize the expected cost, as calculated with the posterior probabilities of the examinee's classification. If c_{21} is the cost of making a classification decision in group 2 (d_2) when the examinee is actually in mastery group 1 (m_1), and c_{12} is vice versa, then the expected cost is

$$B = c_{21}P(d_2 \mid m_1)P(m_1) + c_{12}P(d_1 \mid m_2)P(m_2) \quad (5)$$

Items are selected to minimize expected cost after administration,

$$MEC = B(X=1)P(X=1) + B(X=0)P(X=0) \quad (6)$$

where the probability of each response is multiplied by the expected cost B if the examinee were to respond with the given response (Rudner, 2002). The probability of a response is assumed to be (Rudner, 2002)

$$P(x=X) = P(x=X \mid m_1)P(m_1) + P(x=X \mid m_2)P(m_2) \quad (7)$$

Three cutscore-based methods have also been suggested with IRT (Lin & Spray, 2000): maximum Fisher information, maximum Kullback-Liebler information, and a log-odds ratio. Maximum Fisher information seeks to maximize information at a single point, given the probability of a correct response P and incorrect response Q (Embretson & Reise, 2000)

$$I_i(\theta) = \left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2 \Big/ P_i(\theta)Q_i(\theta). \quad (8)$$

whereas Kullback-Liebler information evaluates information across a region θ_0 to θ_1 around the cutscore (Eggen, 1999)

$$K_i(\theta_1 \| \theta_0) = P_i(\theta_1)\log\frac{P_i(\theta_1)}{P_i(\theta_0)} + Q_i(\theta_1)\log\frac{Q_i(\theta_1)}{Q_i(\theta_0)} \quad (9)$$

Lin and Spray (2000) additionally suggested a transformation of the ratio between the probability of a correct response for a point above the cutscore and a point below the cutscore, as defined by the SPRT termination criterion. Note that the three are conceptually equivalent; an item with the greatest information (slope) at the cutscore is the item that will have the greatest information in a small region around the cutscore, or produce the greatest differences in probability values on either side of the cutscore. Therefore, they perform comparably (Lin & Spray, 2000).

Likewise, Fisher and Kullback-Liebler information can be used as criteria for estimate-based item selection (Reckase, 1983; Spray & Reckase, 1994; Eggen, 1999). The equation of the calculation remains the same, but is now calculated with respect to the current estimate of examinee θ at each point in the test. The conceptual equivalence also carries over to this application; the item with the highest information at the current θ estimate is also the item with the highest information in a small region around the current θ estimate. Estimate-based item selection can also be referred to as adaptive item selection, because it makes use of individual examinee information, namely the response vector, in an effort to adapt the test to the individual examinee. Cutscore-based item selection is sometimes called sequential selection.

An additional method that is too broad to fit either item selection type is *mutual information*, which evaluates item information across the range of θ. Because it is so broad, it is not as applicable to a situation where the exact neighborhood of desired information is known, such as a single cutscore. However, it is quite useful when information is needed across a wider range of θ, such as the beginning of a CAT when little is known regarding examinee θ, or when there are several cutscores (Weissman, 2004).

Termination criteria

Three termination criteria are utilized in VL-CCT: IRT-based confidence intervals, the sequential probability ratio test, and decision theory. Each offers substantially shorter tests than a conventional full-length fixed-form test while maintaining a

similar level of classification accuracy (Kingsbury & Weiss, 1983; Rudner, 2002). However, the appropriateness of each criterion and the maximization of its benefits depend on several other factors, such as the psychometric model and item selection algorithm. For example, the use of IRT-based confidence intervals requires a large bank of IRT-calibrated items, which in turn requires a large calibration sample.

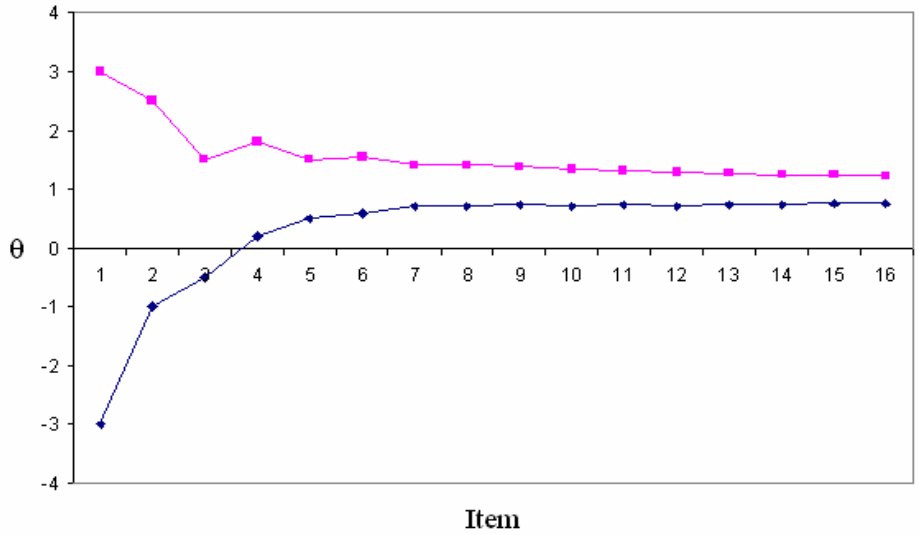
The confidence interval approach formulates the classification purpose as a statistical estimation problem (Eggen & Straetmans, 2000). The test is designed to obtain an estimate of θ for examinee $j, \hat{\theta}_j$, and determine if it is within a range of θ corresponding to membership in a certain group, where group membership is defined by ranges of θ that are delineated by cutscores. To quantify this definition, a confidence interval is constructed around $\hat{\theta}_j$ using the conditional standard error of measurement (CSEM), expressed as (Thompson, 2006)

$$\hat{\theta}_j - z_\alpha(CSEM) \le \theta_j \le \hat{\theta}_j + z_\alpha(CSEM) \quad (10)$$

where z_α is the normal deviate corresponding to a 1-α confidence interval. While this method was originally suggested with a Bayesian (Owen, 1975) estimation procedure (Kingsbury & Weiss, 1983; Spray and Reckase, 1996), it can also be used with maximum likelihood estimation (Eggen & Straetmans, 2000, Yang, Poggio, & Glasnapp, 2006). Similarly, while it originally suggested as a two-sided interval, the algorithm can also be designed as a one-sided interval with β-protection (Chang, 2005).

An example of this approach in the simplest, two-group mastery testing situation is to evaluate after each item whether the confidence interval is above or below the cutscore. If the interval is completely above the cutscore, the examinee can be classified as "Pass," and if completely below, as "Fail." If the interval contains the cutscore, another item is administered. An example of this is shown in Figure 1, where an examinee is administered 16 items before the confidence interval falls completely above a cutscore of 0.75.

Figure 1: Graphic depiction of confidence interval criterion with cutscore of 0.75



This approach was originally termed "adaptive mastery testing" (Kingsbury & Weiss, 1983), but "adaptive" is a term that is more appropriately reserved for estimate-based item selection algorithms, and the confidence interval approach does not require adaptive item selection as originally

suggested. Additionally, it is easily extended beyond the mastery testing situation to three or more categories (Eggen & Straetmans, 2000), so the inclusion of "mastery" in the term in unnecessarily limiting.

The SPRT formulates the classification purpose as a hypothesis testing problem (Eggen & Straetmans, 2000), comparing the ratio of the likelihoods of two competing hypotheses. In CCT, the likelihoods are calculated using the probability P of an examinee's response if each of the hypotheses were true, that is, if the examinee were truly a "pass" (P_2) or "fail" (P_1) classification. This is expressed in general form after n items as, where X is the observed response to item i:

$$LR = \frac{\prod_{i=1}^{n} P_{2i}^{X_i}(1-P_{2i})^{1-X_i}}{\prod_{i=1}^{n} P_{1i}^{X_i}(1-P_{1i})^{1-X_i}}. \qquad (11)$$

This ratio is then compared to two decision points A and B (Wald, 1947)

Lower decision point = $B = \beta / (1 - \alpha)$ (12)
Upper decision point = $A = (1 - \beta) / \alpha$ (13)

If the ratio is above the upper decision point after n items, the examinee is classified as above the cutscore. If the ratio is below the lower decision point, the examinee is classified as below the cutscore. If the ratio is between the decision points, another item is administered.

While the P parameters have often been subscripted with 0 and 1 to match the traditional hypothesis testing notation of H_0 and H_1, not every VL-CCT with the SPRT is limited to the two hypotheses of pass and fail. If there are two or more cutscores being used to classify examinees into three or more groups, there will be at least three P values that need to be specified (Eggen & Straetmans, 2000; Rudner, 2002; Jiao, Wang, & Lau, 2004). If the notation begins at 1 rather than 0, these values will be numbered sequentially and more appropriately reflect the multiple-cutscore situation.

Several methods have been offered for specifying the P_1 and P_2 parameters. Originally, the SPRT was developed with equivalent P_1 and P_2 parameters for each step in the test (Wald, 1947), and was first applied thus to educational measurement (Ferguson, 1969). They can be allowed to vary by estimating the proportion of the population in each group correctly answering the question, using classical difficulty statistics (Frick, 1992; Rudner, 2002). Weitzman (1982a, b) suggested averaging classical difficulty statistics across multiple subgroups within each group.

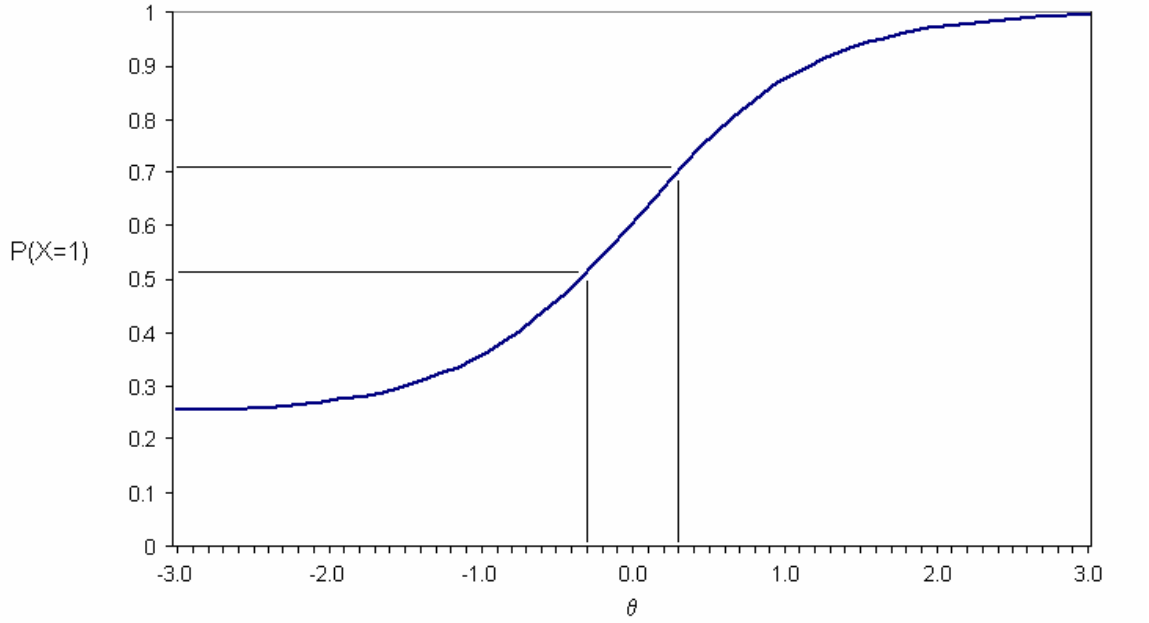
The most commonly used method (Reckase, 1983; Parshall, Spray, Kalohn, & Davey, 2006) is to define as the probability of a correct response to each item from an examinee with corresponding ability levels θ_1 and θ_2, as calculated with an IRT item response function. These two ability levels are defined in the mastery testing situation as the lowest acceptable θ for an examinee that passes the test and the highest acceptable θ for an examinee that fails the test. The range between the two values is termed the "indifference region," and is often determined in practice by adding and subtracting a small constant δ from the cutscore (Eggen, 1999; Eggen & Straetmans, 2000). This approach is depicted graphically in Figure 2, with a cutscore of 0.0 and $\delta = 0.3$, making for $P_1 = 0.52$ and $P_2 = 0.72$. A wider indifference region will lead to increased error but decreased test length (Reckase, 1983; Eggen, 1999).

Both the SPRT and IRT confidence intervals are easily extended to multiple cutscores, and although two methods have been suggested for a multiple-cutscore SPRT (Sobel & Wald, 1949; Armitage, 1950), they are equivalent when given an even comparison (Govindarajulu, 1987). Jiao, Wang, and Lau (2004) compared the two in a simulation study, but with different indifference regions, which caused the differing results found. VL-CCT is a highly efficient method of classifying examinees into three or more groups (Spray, 1993; Xiao, 1999; Eggen & Straetmans, 2000; Yang, Poggio, & Glasnapp 2006).

The third criterion is a Bayesian decision theory framework (Lewis & Sheehan, 1990; Sheehan & Lewis, 1992). In this approach, loss or utility structures must be defined that take into account the probability of the possible classification scenarios for each examinee and the costs associated with each, as represented in Eq. 5, with the additional term of the cost of administering

another item. Vos (2000) claimed that a threshold loss structure is conceptually appropriate, but a linear or squared-error loss function may be used, as it more appropriately represents the high-stakes classification testing situation. It is much less of an concern to license a doctor whose true score is slightly below the cutscore than one whose true score is substantially below.

Figure 2: P_1 and P_2 specification with Reckase's (1983) method



Because there are an infinite number of possible criterion functions, no specific function best represents this approach, which has led to the development of optimization methods (Vos, 2000). This infinite number of functions is both an advantage and a drawback of Bayesian decision theory. Proponents point out that it presents greater flexibility and allows various costs to be explicitly taken into account. However, the selection of a function introduces a certain amount of arbitrariness. Moreover, both the SPRT and IRT confidence intervals allow costs of misclassification to be specified via nominal error rates.

Bayesian decision theory VL-CCTs have been traditionally applied with a classical binomial model for item responses, because when used in combination with a beta examinee distribution, the number-correct score alone can be used to calculate expected loss at future stages of the test (Vos, 2000). Items are selected and tests are terminated to minimize the specified loss function. However, item response theory and adaptive item selection techniques have recently been incorporated with this termination criterion (Glas & Vos, 2006).

PRACTICAL CONSTRAINTS

In addition to the mandatory technical components of VL-CCT, a non-mandatory but essential element is practical constraints on the computer algorithms. In most cases, it is desirable to place constraints on item exposure, so that the most informative items are not over-exposed and therefore compromised (Kalohn & Spray, 1998). Additionally, it is often the case that the test content is distributed across definite content areas, and the VL-CCT must be developed to take this into account (Eggen & Straetmans, 2000). For public relations purposes, constraints on test length may be instilled to prevent tests that are too short, as a VL-CCT can make decisions with only a few items or continue indefinitely for examinees of certain ability levels (Parshall, Spray, Kalohn, & Davey, 2006). However, while these serve important

purposes, their psychometric effect is not substantial (Eggen & Straetmans, 2000). Obviously, a minimum test length will increase the average test length over all examinees, and a maximum test length will decrease average test length. Similarly, item exposure constraints and content constraints generally only serve to increase average test length.

An additional constraint that has been suggested is to design the VL-CCT with truncation rules (Finkelman, 2003). The purpose of this is to address the situation where an examinee has a θ level very near the cutscore, and the VL-CCT may very well continue until the item bank is exhausted without making a classification decision. A component may be designed into the VL-CCT to recognize this situation and determine when the remaining items in the bank do not have enough information to make a decision either way, even if the examinee answered all of the remaining items correctly (or incorrectly). This differs from a maximum test length constraint in that is not arbitrarily set, e.g., 50 items, but empirically justified and therefore more legally defensible.

ROBUSTNESS

VL-CCTs have been found to be quite robust with respect to certain violations of assumptions. Research has focused on the role of IRT assumptions in VL-CCTs with the SPRT termination criterion, as it is used more often than the confidence interval and Bayesian decision theory criteria. The assumptions of IRT are more restrictive than classical test theory, and therefore have a greater possibility for detrimental effect. Two primary assumptions can have an effect on average test length and classification accuracy: the number of dimensions and number of parameters.

Reckase (1983) evaluated VL-CCT efficiency and accuracy when data was simulated with a three-parameter IRT model, but the VL-CCT assumed a one-parameter model. This effectively lowered the cutscore by 1.5 θ units, greatly decreasing classification accuracy. Similar results were found by Kalohn and Spray (1999) and Jiao and Lau (2003), regardless of any practical constraints imposed. Using a two-parameter model when the three-parameter model is more appropriate does not have such dire consequences (Jiao & Lau, 2003). Violations of this type are easy to account for by simply using a more appropriate IRT model.

Employing unidimensional IRT when data is actually modeled with two-dimensional IRT has a less dramatic effect. Spray, Abdel-fattah, Huang, and Lau (1997) and Lau (1996) found that this has only a minimal effect on classification error. However, it does require more items to make a decision, and the increase in items is related to a decrease in the correlation between dimensions (Lau, 1996).

Of course, even if the correct model is chosen, there is a possibility that the item calibration procedure did not accurately estimate item parameters. Similar to what was found regarding the number of dimensions, Spray and Reckase (1987) found that item parameter estimation error had no effect on classification error, but simply required more items for the test to make a decision, which is important to remember when weighing number of IRT parameters against sample size.

VL-CCTs with an SPRT termination criterion can perform efficiently even when the parameters being used are not actually estimated. Huang, Kalohn, Lin, and Spray (2000) investigated the situation where a large item pool is available, but only a small subset has been calibrated with IRT while the majority of the items have classical statistics. VL-CCTs with classical parameters transformed to IRT performed just as accurately as VL-CCTs with the known IRT parameters that were generated for the simulation, while maintaining the same ATL, demonstrating that the efficiency of SPRT-based VL-CCTs is quite robust.

CONCLUSIONS

While there is no question as to the fact that VL-CCTs offer a substantial advantage in terms of shorter tests than a conventional fixed-length approach, the specific components utilized directly affects the extent of this advantage. Knowledge of the available options and the limitation of choices due to the situation, such as a small sample size, is essential in the development of an efficient and defensible VL-CCT. An additional area of choices is presented by the practical constraints that are most likely warranted.

The most important aspect to take into account when selecting VL-CCT components is the resources of the testing program. As previously mentioned, IRT-based confidence intervals is not a viable termination criterion for a testing program that does not have sufficient sample sizes for IRT. The next most important aspect to consider is that some components are more appropriate for use with certain other components. For example, the SPRT works most effectively with cutscore-based item selection because it increases the P_2-P_1 difference (Lin & Spray, 2000), while IRT confidence intervals work most effectively with estimate-based item selection because it decreases the CSEM. Therefore, the SPRT is more appropriate for an item bank with a relatively peaked information function, while IRT confidence intervals are more appropriate for a bank with a relatively uniform information function. If only a certain item bank is available, it may then effectively determine the termination criterion and item selection method to be used.

Several points of contention remain in CCT research, the most important of which is the most effective termination criterion. Spray and Reckase (1996) found the SPRT to outperform IRT confidence intervals, but cutscore-based item selection was used for both methods even though it is less than optimal for the confidence interval criterion. Chang (2005) viewed the confidence interval approach more favorably, so a consensus has not been reached regarding the two. Additionally, Bayesian decision theory methods (Vos, 2000) have not been compared to optimal VL-CCTs with the SPRT or IRT confidence intervals. However, the SPRT is the uniformly most powerful test of two competing hypotheses (Spray & Reckase, 1996), as well as being applicable with any psychometric model (Reckase, 1983; Lau & Wang, 1998; Rudner, 2002). This allows it the advantage of producing very efficient tests even with small sample sizes (Frick, 1992; Rudner, 2002).

Again, item selection methods have only small differences in terms of efficiency when assessing similar information (Spray & Reckase, 1994; Eggen, 1999; Lin & Spray, 2000; Rudner, 2002). The differences are much greater between cutscore-based and estimated-based algorithms than within (Eggen, 1999). These results imply that, in some cases, a less sophisticated methodology may be more appropriate. Mutual information, the most sophisticated method, is appropriate early in the test or when there are several cutscores (Weissman, 2004).

The same is true regarding the psychometric model. VL-CCTs designed with classical test theory can classify examinees accurately and with few items (Frick, 1992; Rudner, 2002). However, if the sample size allows, the advantages of IRT are well known (Embretson & Reise, 2000). Additionally, if polytomously scored items are applicable, decisions can be made with even fewer items than dichotomous IRT (Lau & Wang, 1998), as polytomous IRT offers more information and across a wider range (Dodd, De Ayala, & Koch, 1995). Moreover, a VL-CCT with IRT item parameters estimated from classical statistics also works efficiently (Huang, Kalohn, Lin, & Spray, 2000). If dichotomous IRT is used, it is important to employ the appropriate number of parameters, as using fewer parameters than is needed effectively lowers the cutscore (Reckase, 1983; Jiao & Lau, 2003).

While most research concerning VL-CCT assumes the common situation of mastery testing, where the purpose of the test is to classify the examinee as "pass" or "fail," the methods outlined herein are easily extendable testing situations to more than one cutscore (Spray, 1993; Eggen & Straetmans, 2000; Weissman, 2004). Similarly, they can be extended to testing situations where the psychometric model is polytomous (Lau & Wang, 1998; 1999), such as the partial credit model (Masters, 1982) and generalized partial credit model (Muraki, 1992).

It is also important to consider VL-CCT in the broader view of classification testing. These tests are often high-stakes, and great care should be taken to develop legally defensible tests and to safeguard the integrity of the items, the test, and scores. This in turn serves to safeguard the value of the credential, which is of paramount importance as the credential is the primary product being sold by organizations. Though a large amount of income may be generated through ancillary revenue sources such as continuing education, it is the credential itself that holds the greatest value to all stakeholders. This protection is the ultimate

purpose of the practical testing constraints, and should be considered as constraints are imposed.

VL-CCT remains a subject of current research. Specific topics include psychometric model (Lau & Wang, 1998; 1999; Rudner, 2002), multiple classifications (Jiao, Wang, & Lau, 2004; Yang, Poggio, & Glassnap, 2006), item selection algorithms (Weissman, 2004), specification of termination criteria (Chang, 2006) and combinations of methods (Glas & Vos, 2006). Additionally, research on computerized adaptive testing for point estimation of ability is often relevant (e.g., Weissman, 2006). Future research will continue to investigate new methodologies and refine current ones to offer incremental increases in test efficiency and accuracy. The pervasiveness of the classification testing situation in today's society ensures that VL-CCT will remain an important type of computerized testing.

REFERENCES

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, 12*, 137-144.

Chang, Y.-c. I. (2005). Application of Sequential Interval Estimation to Adaptive Mastery Testing. *Psychometrika, 70*, 685-713.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 19*, 5-22.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 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*, 713-734.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction.* Unpublished doctoral dissertation, University of Pittsburgh.

Finkelman, M. (2003). *An Adaptation of Stochastic Curtailment to Truncate Wald's SPRT in Computerized Adaptive Testing* (CSE Report 606). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research, 8*, 187-213.

Glas, C.A.W., & Vos, H.J. (2006). *Testlet-Based Adaptive Mastery Testing* (Computerized Testing Report 99-11). Newtown, PA: Law School Admission Council.

Govindarajulu, Z. (1987). *The sequential statistical analysis of hypothesis testing, point and interval estimation, and decision theory.* Columbus, OH: American Sciences Press.

Huang, C.-Y., Kalohn, J.C., Lin, C.-J., and Spray, J. (2000). *Estimating Item Parameters from Classical Indices for Item Pool Development with a Computerized Classification Test.* (Research Report 2000-4). Iowa City, IA: ACT, Inc.

Jiao, H., & Lau, A. C. (2003). *The Effects of Model Misfit in Computerized Classification Test.* Paper presented at the annual meeting of the National Council of Educational Measurement, Chicago, IL, April 2003.

Jiao, H., Wang, S., & Lau, C. A. (2004). *An Investigation of Two Combination Procedures of SPRT for Three-category Classification Decisions in Computerized Classification Test.* Paper presented at the annual meeting of the American Educational Research Association, San Antonio, April 2004.

Kalohn, J. C., & Spray, J. A. (1998). *Effect of item selection on item exposure rates within a computerized classification test.* Paper presented at the annual meeting of the National Council of Educational Measurement, San Diego, CA, April 1998.

Kalohn, J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement, 36*, 47-59.

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 theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data.* Unpublished doctoral dissertation, University of Iowa, Iowa City IA.

Lau, C. A., & Wang, T. (1998). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego.

Lau, C. A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367-386.

Lin, C.-J. & Spray, J.A. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test.* (Research Report 2000-8). Iowa City, IA: ACT, Inc.

Linn, R. L., Rock, D. A., & Cleary, T. A. (1972). Sequential testing for dichotomous decisions. *Educational & Psychological Measurement, 32*, 85-95.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Parshall, C.G., Spray, J.A., Kalohn, J.C., & Davey, T. (2006). *Practical considerations in computer-based testing.* New York: Springer.

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: Academic Press.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures.* Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

Shannon, C.E. (1948). A mathematical theory of communication, *Bell System Technical Journal,* 27, 379-423 and 623-656, July and October. Available online: http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html

Sheehan, K., & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. *Applied Psychological Measurement, 16,* 65-76.

Sobel, M. & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics, 20,* 502-522.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (Research Report 93-7). Iowa City, Iowa: ACT, Inc.

Spray, J. A., Abdel-fattah, A. A., Huang, C., and Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional* (Research Report 97-5). Iowa City, Iowa: ACT, Inc.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (Research Report 87-17). Iowa City, IA: ACT, Inc.

Spray, J. A., & Reckase, M. D. (1994). *The selection of test items for decision making with a computerized adaptive test.* Paper presented at the Annual Meeting of the National Council for Measurement in Education (New Orleans, LA, April 5-7, 1994).

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 & Behavioral Statistics,21*, 405-414.

Thompson, N. A. (2006). Variable-length computerized classification testing with item response theory. *Clear Exam Review, 17*(2), 21-26.

Vos, H.J. (2000). A Bayesian Procedure in the Context of Sequential Mastery Testing. *Psicológica, 21,* 191-211.

Wainer, H., & Mislevy, R.J. (2000). Item response theory, calibration, and estimation. In Wainer, H. (Ed.) *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

Weissman, A. (2006). A Feedback Control Strategy for Enhancing Item Selection Efficiency in Computerized Adaptive Testing. *Applied Psychological Measurement, 30*, 84-99.

Weitzman, R. A. (1982a). Sequential testing for selection. *Applied Psychological Measurement, 6*, 337-351.

Weitzman, R. A. (1982b). Use of sequential testing to prescreen prospective entrants into military service. In D. J. Weiss (Ed.), *Proceedings of the 1982 Computerized Adaptive Testing Conference.* Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1982.

Yang, X., Poggio, J.C., & Glasnapp, D.R. (2006). Effects of Estimation Bias on Multiple-Category Classification with an IRT-Based Adaptive Classification Procedure. *Educational and Psychological Measurement, 66*, 545-564.

Xiao, B. (1999). Strategies for computerized adaptive grading testing. *Applied Psychological Measurement, 23*, 136-146.

Citation

Thompson, Nathan A. (2007). A Practitioner's Guide for Variable-length Computerized Classification Testing. *Practical Assessment Research & Evaluation*, 12(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1

Authors

Correspondence concerning this paper should be addressed to

Nathan A. Thompson
Thomson Prometric
1260 Energy Lane
Saint Paul, MN 55108

nathan.thompson [at] thomson.com