ACT Research Report Series 93-7

Multiple-Category Classification Using a Sequential Probability Ratio Test

Judith Spray

December 1993

ACT

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243

Multiple-Category Classification Using a Sequential Probability Ratio Test

Judith A. Spray

American College Testing

variable, X, represents a single Bernoulli trial and is distributed as Bin{$P(\theta_i)$,1}. Then,

$$\pi(\theta_i) = \text{Prob}(X = x | \theta = \theta_i) = P(\theta_i)^x \, Q(\theta_i)^{1-x}, \text{ where}$$

$$x = \begin{cases} 1, \text{ correct response} \\ 0, \text{ incorrect response} \end{cases}.$$

For this test item, the probability of observing $X = x$ under the alternative hypothesis is $\pi(\theta_1)$. Under the null hypothesis, the probability of observing $X = x$ is $\pi(\theta_0)$. The functions, $\pi(\theta_1)$ and $\pi(\theta_0)$, are called likelihood functions of x, and a ratio of these two functions, $L(x) = \pi(\theta_1)/\pi(\theta_0)$, is called a likelihood ratio.

Two error probabilities, α and β, can be defined, where

$$\text{Prob(choosing } H_1 \text{ if } H_0 \text{ is true)} = \alpha$$

and

$$\text{Prob(choosing } H_0 \text{ if } H_1 \text{ is true)} = \beta.$$

Wald (1947) stated that even though the nominal error rates, α and β, are established prior to testing, the actual error rates observed in practice, α^* and β^*, are bounded from above by functions of the nominal rates, or $\alpha^* \leq \alpha/(1-\beta)$ and $\beta^* \leq \beta/(1-\alpha)$. Wald (1947) also defined two likelihood ratio boundaries that are functions of α and β. These boundaries are A and B, where the lower boundary $= B \geq \beta/(1-\alpha)$ and the upper boundary $= A \leq (1-\beta)/\alpha$.

According to Wald's SPRT, item responses are observed in sequence, $x_1, x_2, \ldots, x_n$, and following each observation, the likelihood ratio, $L(x_1, x_2, \ldots, x_n | \theta_0, \theta_1)$, is computed, assuming conditional independence, where

$$L(x_1, x_2, \ldots, x_n | \theta_0, \theta_1) = \frac{\pi_1(\theta_1)\, \pi_2(\theta_1) \ \ldots \ \pi_n(\theta_1)}{\pi_1(\theta_0)\, \pi_2(\theta_0) \ \ldots \ \pi_n(\theta_0)}.$$

Abstract

Sequential probability ratio testing (SPRT), which usually is applied in situations requiring a decision between two simple hypotheses or a single decision point, is extended to include situations involving k decision points and [$(k + 1)$-choose-2] sets of simultaneous, simple hypotheses, where $k > 1$. The multiple-decision point or multiple-category SPRT procedure can be used to classify examinees into $k + 1$ categories using computer adaptive methods. Computer simulations utilizing a 200-item pool of previously calibrated test items show that the multiple-category SPRT method controls misclassification error rates adequately, provided that the number of decision points is not too large.

Multiple-Category Classification Using a Sequential Probability Ratio Test

Wald's (1947) sequential probability ratio testing (SPRT) procedure has been used with cognitive tests to classify examinees into one of two categories (e.g., pass/fail, master/nonmaster, certified/noncertified) (Reckase, 1983). In other words this procedure is useful for determining whether an examinee more likely belongs to one of two states or conditions: either an individual has ability or latent trait greater than or equal to some minimum value, δ or that same individual has ability less than the minimum value, δ. The value, δ, is frequently called a passing score or decision point.

One way to test the composite hypothesis that either the examinee has latent ability less than δ versus that the examinee has latent ability greater than or equal to δ, is to consider simple hypotheses, H_0 or H_1, regarding the unidimensional latent trait or ability (θ_i) of the examinee taking the test. These simple hypotheses can be written as

$$H_0: \ \theta_i = \theta_0$$

vs.

$$H_1: \ \theta_i = \theta_1 \ ,$$

where θ_i is an unknown parameter of the distribution of the dichotomous response to a particular test item, X (Silvey, 1975). Usually, θ_0 and θ_1 represent decision points that correspond to lower and upper limits, respectively, of the passing criterion or threshold, δ, where $\theta_0 < \delta < \theta_1$. The SPRT can then be used to test the composite hypotheses, H_0: $\theta_i < \delta$ versus H_1: $\theta_i \geq \delta$ by considering two weaker hypotheses, say $\omega_0 = \{\theta:\theta \leq \theta_0\}$ and $\omega_1 = \{\theta:\theta \geq \theta_1\}$ (Silvey, 1975; Wald, 1947).

In the case of cognitive testing, X can be assumed to follow a binomial distribution. If $P(\theta_i)$ is the probability that examinee i responds correctly to an item, and $Q(\theta_i) = 1 - P(\theta_i)$ is the probability of an incorrect response from examinee i, then, for this single item, the random

variable, X, represents a single Bernoulli trial and is distributed as Bin{$P(\theta_i)$,1}. Then,

$$\pi(\theta_i) = \text{Prob}(X = x|\theta = \theta_i) = P(\theta_i)^x \, Q(\theta_i)^{1-x} ,$$

where

$$x = \begin{cases} 1, & \text{correct response} \\ 0, & \text{incorrect response} . \end{cases}$$

For this test item, the probability of observing $X = x$ under the alternative hypothesis is $\pi(\theta_1)$. Under the null hypothesis, the probability of observing $X = x$ is $\pi(\theta_0)$. The functions, $\pi(\theta_1)$ and $\pi(\theta_0)$, are called likelihood functions of x, and a ratio of these two functions, $\mathbf{L}(x) = \pi(\theta_1)/\pi(\theta_0)$, is called a likelihood ratio.

Two error probabilities, α and β, can be defined, where

$$\text{Prob(choosing } H_1 \text{ if } H_0 \text{ is true)} = \alpha$$

and

$$\text{Prob(choosing } H_0 \text{ if } H_1 \text{ is true)} = \beta.$$

Wald (1947) stated that even though the nominal error rates, α and β, are established prior to testing, the actual error rates observed in practice, α^* and β^*, are bounded from above by functions of the nominal rates, or $\alpha^* \leq \alpha/(1-\beta)$ and $\beta^* \leq \beta/(1-\alpha)$. Wald (1947) also defined two likelihood ratio boundaries that are functions of α and β. These boundaries are A and B, where the lower boundary= $B \geq \beta/(1-\alpha)$ and the upper boundary = $A \leq (1-\beta)/\alpha$.

According to Wald's SPRT, item responses are observed in sequence, $x_1, x_2, \ldots, x_n$, and following each observation, the likelihood ratio, $\mathbf{L}(x_1, x_2, \ldots, x_n|\theta_0,\theta_1)$, is computed, assuming conditional independence, where

$$\mathbf{L}(x_1, x_2, \ldots, x_n|\theta_0,\theta_1) = \frac{\pi_1(\theta_1)\ \pi_2(\theta_1)\ \ldots\ \pi_n(\theta_1)}{\pi_1(\theta_0)\ \pi_2(\theta_0)\ \ldots\ \pi_n(\theta_0)} .$$

The likelihood ratio is then compared to the boundaries, A and B. If $\mathbf{L}(x_1, x_2, ..., x_n|\theta_0,\theta_1) \geq A$, then H_1 is accepted and the examinee is classified as $\theta_i \geq \delta$. If $\mathbf{L}(x_1, x_2, ..., x_n|\theta_0,\theta_1) \leq B$, then H_0 is accepted and the examinee is classified as $\theta_i < \delta$. If $B < \mathbf{L}(x_1, x_2, ..., x_n|\theta_0,\theta_1) < A$, no decision is made and another item response must be observed if a decision is to be made with the specified error rates.

Any test administered with the SPRT procedure is, by its very nature, adaptive in that examinees with different abilities (i.e., different values of θ_i) could have different expected test lengths, n_i, the number of items that must be administered before a classification is made. Typically, those examinees with $\theta_i \leq \theta_0$ or $\theta_i \geq \theta_1$ will have shorter expected test lengths than those with $\theta_0 < \theta_i < \theta_1$.

To facilitate the SPRT procedure for criterion-referenced testing, the value of δ usually corresponds to a minimum proportion, $p(\delta)$, of m items in the item pool that an examinee is expected to answer correctly in order to be classified as $\theta_i \geq \theta_1$. If $p(\delta)$ is known *a priori*, then δ can be found by solving for δ in the expression, $p(\delta) = 1/m \; \Sigma \; P_j(\delta)$, $j = 1, 2, ..., m$. The item functions, $P_j(\delta)$, are typically expressed as 3-parameter logistic item response functions with known (i.e., calibrated) item parameters.

Values for θ_1 and θ_0 are selected according to the precision that is desired. Values of θ_0 and θ_1 that are close to each other imply high precision, while greater differences in θ_0 and θ_1 imply less precision. Normally, θ_1 and θ_0 are selected to be equidistant from δ, although this is not a necessary condition for the SPRT procedure. The region from θ_0 to θ_1 is known as the *indifference region* because there is usually an amount of indifference associated with the classification made for individuals within that region. The distance, $|\theta_1 - \theta_0|$, is the *width* of the indifference region. Test length is a function of this region; for fixed values of α and β, a

larger indifference region results in shorter expected test lengths for all examinees (Reckase, 1983; Spray & Reckase, 1987).

Within the context of an adaptive test, the m items in the item pool are usually ranked 1 through m on the basis of *item information* at the decision point, $p(\delta)$ or equivalently, δ, and then administered in sequence to each examinee. Therefore, many examinees could receive some of the same items as all other examinees taking the sequential test. Because this is usually undesirable from a test security standpoint, some randomization scheme can be employed to assure that item-exposure rates (i.e., the number of times that any item is presented to examinees) are controlled.

In addition there is usually some maximum number of test items or maximum test length (MTL), that, from a practical standpoint in terms of testing time, can be presented to a single examinee. Frequently, a forced classification is made once this maximum number of items has been reached and no classification under the likelihood ratio test has occurred. Typically, after reaching this maximum test length, $\log\{\mathbf{L}(x_1, x_2, ..., x_{max})\}$ is compared to the log of the SPRT boundaries, A and B. Classification is then made according to some distance rule, for example by $\text{MIN}\{|\log\mathbf{L}(x_1, x_2, ..., x_{max})-\log A|, |\log\mathbf{L}(x_1, x_2, ..., x_{max})-\log B|\}$. For tests where MTL is fairly small, forced classifications can occur for many examinees. The effect of forced classification on the SPRT procedure is to alter the actual classification error rates, α^* and β^*, reducing the classification accuracy.

Multiple Categories

When a testing situation requires classification into more than two categories, such as into one of a number of entry level courses, a modified SPRT procedure can be used (Wetherill,

1975). The purpose of this paper is to describe one such multiple-category modification and to report on the results of computer-simulated SPRTs requiring multiple classifications.

A Sequential Probability Ratio Test Involving Two Decision Points

Suppose that the purpose of a SPRT is to classify an examinee into one of k+1 categories (e.g., hierarchically ordered mathematics courses), where k is the number of decision points required. For the following discussion, it is assumed that $k = 2$. The three categories of possible mutually exclusive classification are $\theta_i < \delta_1$, $\delta_1 \leq \theta_i < \delta_2$, or $\theta_i \geq \delta_2$. The values of δ_1 and δ_2 are established or known *a priori*. However, because the usual SPRT tests hypotheses about single values of θ_i defined by the endpoints of the indifference region, such a region must be constructed around each decision point. One such endpoint can be chosen midway between δ_1 and δ_2 or (δ_1 + MIDIST) where MIDIST = (δ_2 - δ_1)/2. This θ value is labeled θ_2, while another θ value (θ_1) can be chosen, such that $\theta_1 = \delta_1$ - MIDIST. This result gives an indifference region around δ_1 of size $|\theta_2 - \theta_1|$, = 2 X MIDIST. A similar indifference region can be constructed around δ_2 using θ_2 and θ_3 as indifference region endpoints, where $\theta_3 = \delta_2$ + MIDIST. These three values of θ form the set $\{\theta_1, \theta_2, \theta_3\}$, where $\theta_1 < \theta_2 < \theta_3$. Once these values of θ are established, three sets of SPRT hypotheses can be formulated:

$$H_1: \theta_i = \theta_1 \qquad H_2: \theta_i = \theta_2 \qquad H_3: \theta_i = \theta_1$$
$$H_{\bar{1}}: \theta_i = \theta_2' \qquad H_{\bar{2}}: \theta_i = \theta_3' \qquad H_{\bar{3}}: \theta_i = \theta_3'$$

All three sets of hypotheses are tested after each item response is obtained, and the following decisions are made, based on the results of these tests:

Decision 1 is made ($\theta_i < \delta_1$) when H_1 and H_3 are both accepted.

Decision 2 is made ($\delta_1 \leq \theta_i < \delta_2$) when $H_{\bar{1}}$ and H_2 are both accepted.

Decision 3 is made ($\theta_i \geq \delta_2$) when $\mathbf{H_2}$ **and** $\mathbf{H_3}$ are both accepted.

Otherwise, testing continues.

For each SPRT, test items can no longer be ranked for sequential administration by item information *at a single decision point* because there is more than one such point. A reasonable compromise is to rank items by item information at the decision point that is closer to an estimate of the examinee's ability based upon the responses to previous items. For this study, a Bayes estimate of θ_i is obtained (Owen, 1975) for each examinee after each item response, and the viable test items remaining in the pool are then ranked, by item information, at this decision point and administered. The process continues until a decision is reached for each examinee.

Establishing error rates. In order to test the set of three hypotheses given above, desired error rates must be provided. These are used to derive the critical values for the likelihood ratio tests. Let $p_{h|j}$ designate the probability that $\theta = \theta_h$ is accepted, given that $\theta = \theta_j$ is correct, $h = 1, 2, 3$; $j = 1, 2, 3$. The power of any single SPRT is $p_{h|h}$ or $p_{j|j}$, and for simplicity, let $p_{h|h} = p_{j|j}$ for all h and j. It makes intuitive sense to allow the error rates, $p_{h|j}$, $h \neq j$, to vary as a function of the distance between θ_h and θ_j. Specifically, the desired error rate should be less when the distance between θ_h and θ_j is greater. If d_{hj} represents the distance (i.e., the absolute difference) between θ_h and θ_j, $h \neq j$, then $|D_h| = \Sigma\ 1/d_{hj}$, summed over j, represents the norm of these distances. Then a possible set of error rates with these properties are

$$p_{h|j} = (1 - p_{h|h})\ d^{-1}_{hj}/|D_h|,\ h \neq j\ .$$

Establishing likelihood ratio boundaries. The likelihood ratio boundaries used to make one of the three decisions mentioned above follow straightforward from the simple SPRT procedure involving two categories. In order to test H_0: $\theta = \theta_h$ versus H_1: $\theta = \theta_j$, the upper boundary is $p_{j|j}/p_{j|h}$ and the lower boundary is $p_{h|j}/p_{h|h}$, $h = 1, 2, 3$; $j = 1, 2, 3$; $h \neq j$.

In particular, for the case involving two decision points and three categories,

if $\mathbf{L}(x_1, x_2, ..., x_n|\theta_1,\theta_2) \leq p_{1|2}/p_{1|1}$ and $\mathbf{L}(x_1, x_2, ..., x_n|\theta_1,\theta_3) \leq p_{1|3}/p_{1|1}$, Decision 1 ($\theta_i<\delta_1$) is made;

if $\mathbf{L}(x_1, x_2, ..., x_n|\theta_2,\theta_3) \leq p_{2|3}/p_{2|2}$ and $\mathbf{L}(x_1, x_2, ..., x_n|\theta_2,\theta_1) \geq p_{2|2}/p_{2|1}$, Decision 2 ($\delta_1\leq\theta_i<\delta_2$) is made; and

if $\mathbf{L}(x_1, x_2, ..., x_n|\theta_2,\theta_3) \geq p_{3|3}/p_{3|2}$ and $\mathbf{L}(x_1, x_2, ..., x_n|\theta_3,\theta_1) \geq p_{3|3}/p_{3|1}$, Decision 3 ($\theta_i\geq\delta_2$) is made.

A Sequential Probability Ratio Test Involving *k* Decision Points

In general, suppose that the purpose of an SPRT is to classify an examinee into one of $k + 1$ categories, where k is the number of decision points required. For the following discussion, it is assumed that $k \geq 2$. The $k + 1$ categories of mutually exclusive classification are $\theta_i < \delta_1$, $\delta_1 \leq \theta_i < \delta_2$, $\delta_2 \leq \theta_i < \delta_3$, ... , $\theta_i \geq \delta_k$. Once again, the values of δ_1, δ_2, δ_3, etc. are established *a priori*. These might represent the criteria for receiving class grades *A*, *B*, *C*, and so on. In order to perform the necessary SPRTs, $k + 1$ values of θ must be established in the manner described previously (i.e., the values of θ represent midpoints between adjoining decision points). These θ values are used to test [($k + 1$)-choose-2] simple SPRTs of the form, θ_1 versus θ_2, θ_1 versus θ_3, ..., θ_1 versus θ_{k+1}, θ_2 versus θ_3, ..., θ_k versus θ_{k+1}. Error rates, $p_{h|j}$, and likelihood ratio boundaries remain the same as described previously.

The number of tests necessary for acceptance before a decision is made (i.e., before an examinee is classified into one of the $k + 1$ categories) is k. As before, if item administration is terminated before a classification is made (e.g., MTL is reached), then the region of classification containing an estimate of the examinee's ability, θ_i, can be obtained and used to place the examinee into one of the $k + 1$ categories. The same situation applies in the $k \geq 2$ case

as it did in the 2-category case, in terms of the effect of forced classification on the SPRT procedure and the actual error rates. Classification into categories is not as accurate when item administration ends when the MTL value is reached.

All [(k + 1)-choose-2] sets of hypotheses are tested, and the following decisions are made, based on the results of these tests:

Decision 1 is made ($\theta_i < \delta_1$) when the k tests of H: $\theta_i = \theta_1$ are accepted.

Decision j is made ($\delta_{j-1} \leq \theta_i < \delta_j$) when the k tests of H: $\theta_i = \theta_j$, j = 2, 3, ... , k, are accepted;

Decision k + 1 is made ($\theta_i \geq \delta_k$) when the k tests of H: $\theta_i = \theta_{k+1}$ are accepted.

Results of Computer Simulations

Computer simulations were conducted to determine if the multiple-category SPRT procedure produced classifications that were characteristic of a simple, one-decision-point SPRT. In other words, did the multiple-category SPRT produce classification error rates and average test lengths that were greatest at the decision points? Would the error rate appear to be controlled appropriately by the specified power, and if so, by what amount?

A calibrated 200-item pool was used to simulate multiple-category SPRT classifications via computer. Items were calibrated with the BILOG computer program (Mislevy & Bock, 1984). Mean estimates of the a-, b-, and c-parameters for the item pool were 1.18, .48, and .16, respectively. Four computer simulations were performed. Simulation I (the simple SPRT) required a single decision point k = 1,{δ = .05; or $p(\delta)$ = .43} with 3 sizes of the indifference region: (-.20, .30), (-.45, .55), and (-.95, 1.05) with power (i.e., $p_{h|h} = p_{j|j}$) = .90. Simulation II consisted of 2 decision points of δ_1 = -1.05 and δ_2 = 1.05, or $p(\delta_1)$ = .23 and $p(\delta_2)$ = .75, respectively, again with .90 power.

Simulation III consisted of 3 decision points at -.95, 0.05, and 1.05, or $p(\delta_1) = .24$, $p(\delta_2) = .43$, and $p(\delta_3) = .74$, also with power = .90. Finally, simulation IV required 4 decision points at -1.05, -.55, .55, and 1.05, or $p(\delta_i) = .23, .31, .58$, and $.75$, $i = 1,2,3,4$, with .90 power. For each simulation, 3 different values of MTL, the maximum test length or maximum number of items to be administered before a forced classification was made, were used. These were 10, 20, and 50. For any single set of simulation conditions, a sequential test was administered 100 times to an examinee with known ability, θ_i, where θ_i varied systematically from -3.0 to +3.0 in increments of .25.

Two outcome measures were tabulated over each set of 100 replications. *Classification Error Rate (CE Rate)* was the number of times that a simulated examinee with a known ability, θ_i, was misclassified, either before MTL items were presented or after MTL items were administered and a forced classification was made. *Average Test Length (ATL)* was the average number of test items administered before an examinee was classified.

Simulation I. Figures 1, 2, and 3 show *CE Rates* for Simulation I ($k = 1$) for three sizes of the indifference region, respectively; (-.20, .30); (-.45, .55); and (-.95, 1.05), respectively, for the three values of MTL. Figures 4,5, and 6 show *ATL* for the same conditions, also respectively.

CE Rate peaked at or near the single decision point, $\delta = 0.05$, regardless of the value of MTL (See Figures 1, 2, and 3). Classification Error was slightly greater for the largest indifference region and was also greater for lower values of θ (See Figure 3). For all three indifference regions, CE Rate decreased as MTL increased. The ATL function reached a peak at or near $\delta = .05$. As expected, values of ATL increased when MTL increased and when the

width of the indifference region decreased. Slightly elevated ATL levels in the upper ability region of θ were noted under all conditions. See Figures 4, 5, and 6.

Simulation II. Figures 7 and 8 show Classification Error Rate and ATL for Simulation II ($k = 2$) for the three values of MTL. The decision points were $\delta_1 = -1.05$ and $\delta_2 = 1.05$, or $p(\delta_1) = .23$ and $p(\delta_2) = .74$. The errors once again tended to peak at or near the two decision points and were minimized in the tails and in between the two decision points. There was a tendency for the error to be higher at the lower decision point, δ_1. Values of the ATL also reached maximums at or near the decision points, although there were some exceptions for very low values of θ.

Simulation III. For Simulation III ($k = 3$), Figures 9 and 10 show CE Rate and ATL, again for the three values of MTL. These figures are consistent with the $k = 2$ situation, in that CE Rate and ATL reached maximums at the three decision points. Once again, the error at the lowest decision point, δ_1, tended to be slightly higher than at the remaining two decision points. Misclassification was greatest for the shortest test (i.e., when MTL = 10). The ATL peaked dramatically at the lowest decision point for ATL = 50, and, to a lesser extent, when MTL = 20. The average length of the test increased considerably with the added decision point (see Figure 10 versus 8).

Simulation IV. Figures 11 and 12 show Classification Error Rate and ATL for this simulation condition. The error plot shows the familiar patterns in which the greatest misclassification occurred at the three decision points. The ATL was greatest at the two lowest decision points but peaked again at δ_3 and δ_4, as expected.

Comparisons across simulations

In order to compare the two outcome measures better across simulations, expected values of CE and ATL were computed by assuming that θ was distributed as N(0,1). Note that this assumption was **not** necessary in order to conduct the multiple-category SPRT simulations. The results appear in Table 1. This table shows that classification error rate and average test length usually increased with the number of decision points. The exception was the $k = 2$ case where MTL = 20 and 50.

Summary and Conclusions

The extension of the SPRT procedure to multiple decision points for classification appears to work as expected. Error rates appeared to be controlled, for the most part, for values of θ away from the decision points in a manner similar to the $k = 1$ case or simple SPRT. Recall that in the simple case, the SPRT procedure guarantees that classification errors, $\alpha^* + \beta^*$, will be bounded by functions of α and β. By specifying power *a priori*, the classification error rate is controlled for $k = 1$. Likewise, it would appear that specifying power also controls the classification errors in the multiple-category situation. However, it is obvious from these results that, as k increases, the number of items required to meet the specified classification error rates also increases. In a practical testing situation, these large numbers of items may not be practical to administer. Thus, the multiple-category SPRT extension may have limited benefits beyond use with a relatively small number of decision points.

References

Mislevy, R. J., & Bock, R. D. (1984). *BILOG, maximum likelihood item analysis and test scoring: Logistic model.* Moorseville, IN: Scientific Software, Inc.

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.

Reckase, M.D. (1983). A procedure for decision making using tailored testing. In D.J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York: Academic Press.

Silvey, S. D. (1975). *Statistical inference.* London: Chapman and Hall.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (ACT Research Report Series No. 87-17). Iowa City, IA: American College Testing.

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

Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd ed.). London: Chapman and Hall.

Author Note

The author would like to thank Mark Reckase and Tim Davey for their comments and helpful suggestions.

TABLE 1

Expected CE Rate and ATL

k	MTL	E(CE Rate)	E(ATL)
1 largest indifference region	10 20 50	.077 .074 .073	1.883 1.879 1.880
1 medium indifference region	10 20 50	.044 .041 .040	3.319 3.435 3.464
1 smallest indifference region	10 20 50	.040 .029 .027	5.404 6.075 6.676
2	10 20 50	.127 .119 .126	4.580 4.758 5.012
3	10 20 50	.163 .110 .103	7.568 9.577 10.767
4	10 20 50	.209 .146 .127	9.546 15.048 22.546

Figure Captions

Figure 1. Classification Error Rate for k = 1, Smallest Indifference Region: (-.20,.30)

Figure 2. Classification Error Rate for k = 1, Medium Indifference Region: (-.45,.55)

Figure 3. Classification Error Rate for k = 1, Largest Indifference Region: (-.95,1.05)

Figure 4. ATL for k = 1, Smallest Indifference Region: (-.20,.30)

Figure 5. ATL for k = 1, Medium Indifference Region: (-.45,.55)

Figure 6. ATL for k = 1, Largest Indifference Region: (-.95,1.05)

Figure 7. Classification Error Rate for k = 2

Figure 8. ATL for k = 2

Figure 9. Classification Error Rate for k = 3

Figure 10. ATL for k = 3.

Figure 11. Classification Error Rate for k = 4

Figure 12. ATL for k = 4.

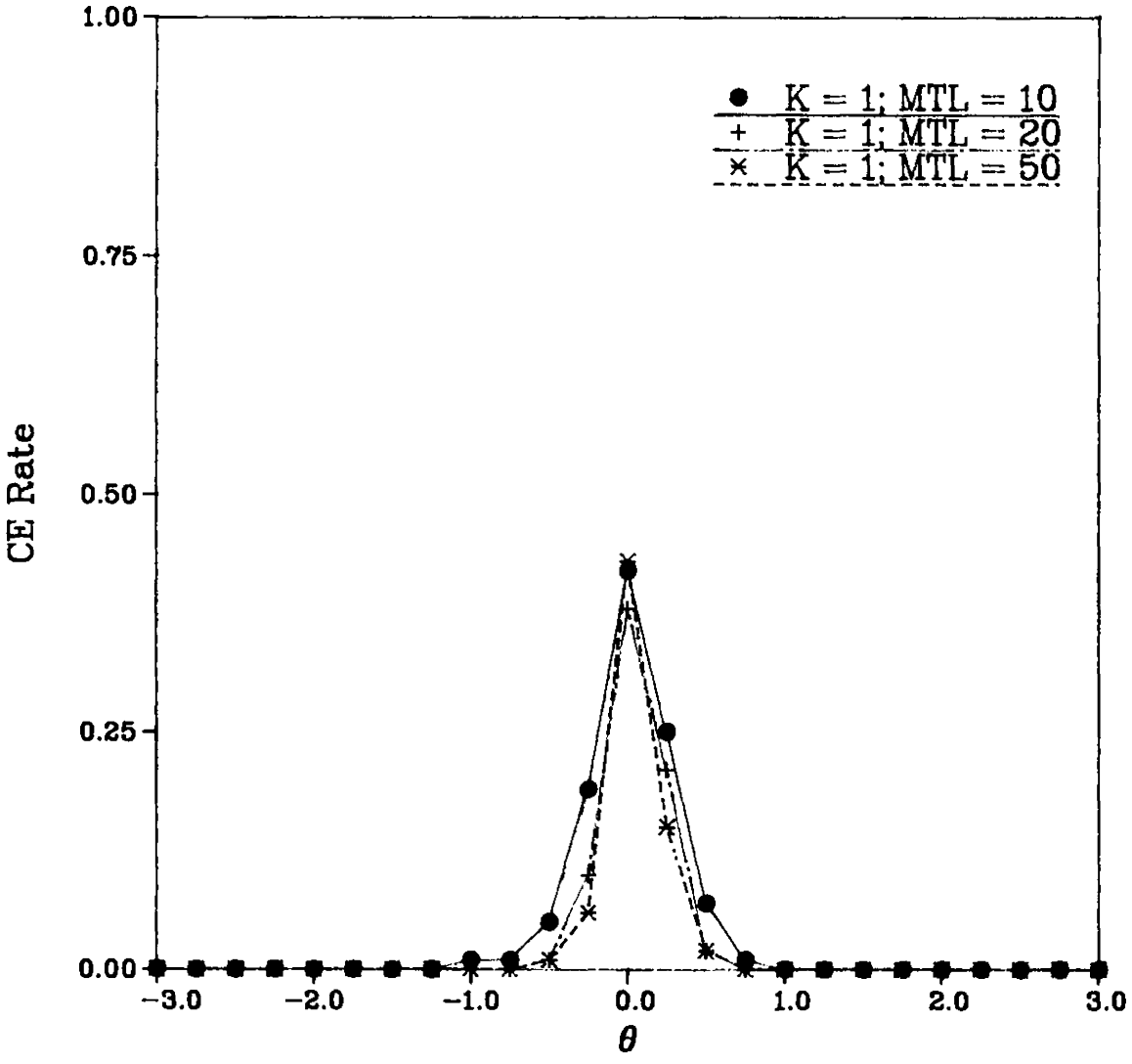
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 1

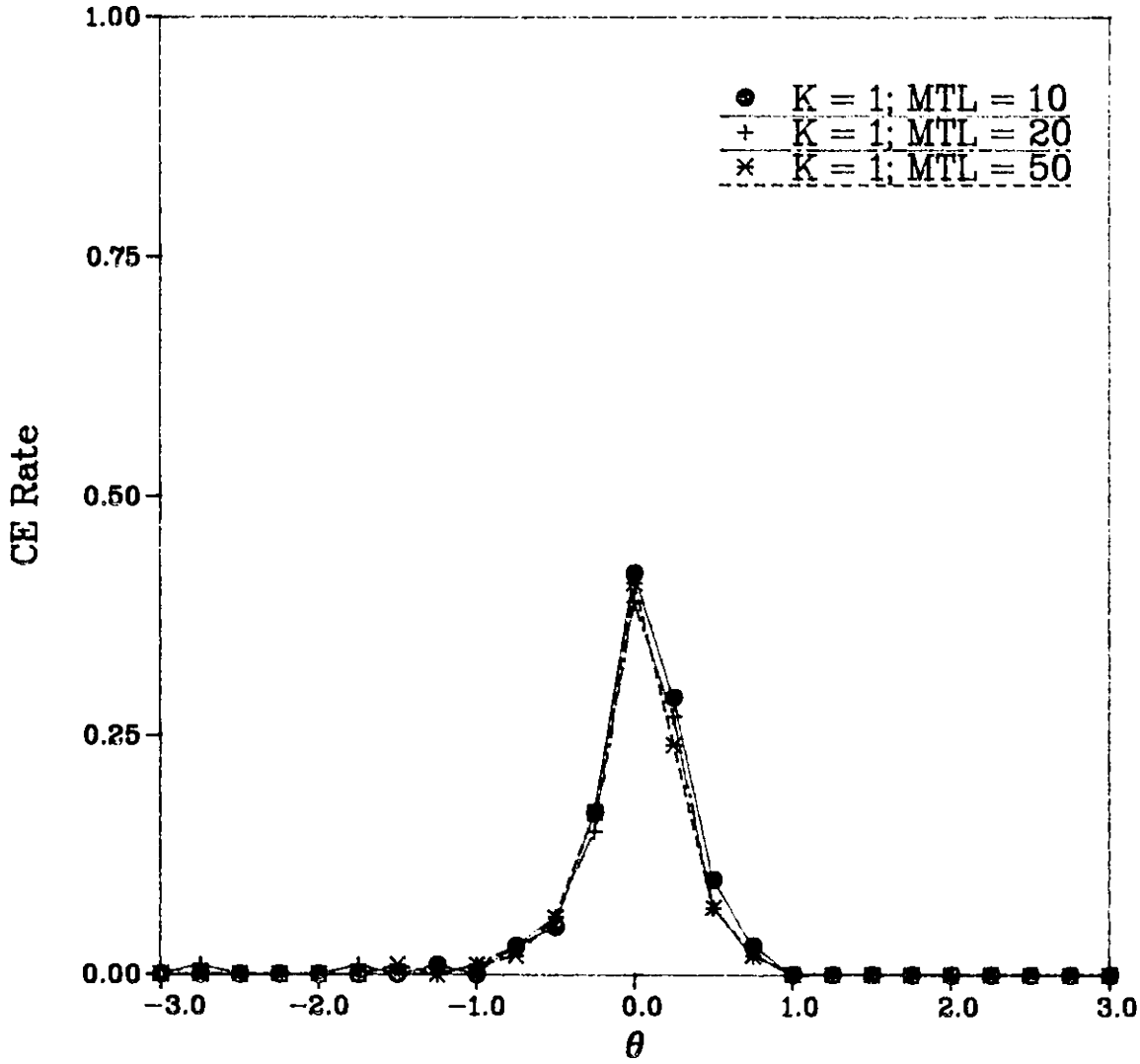
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 2

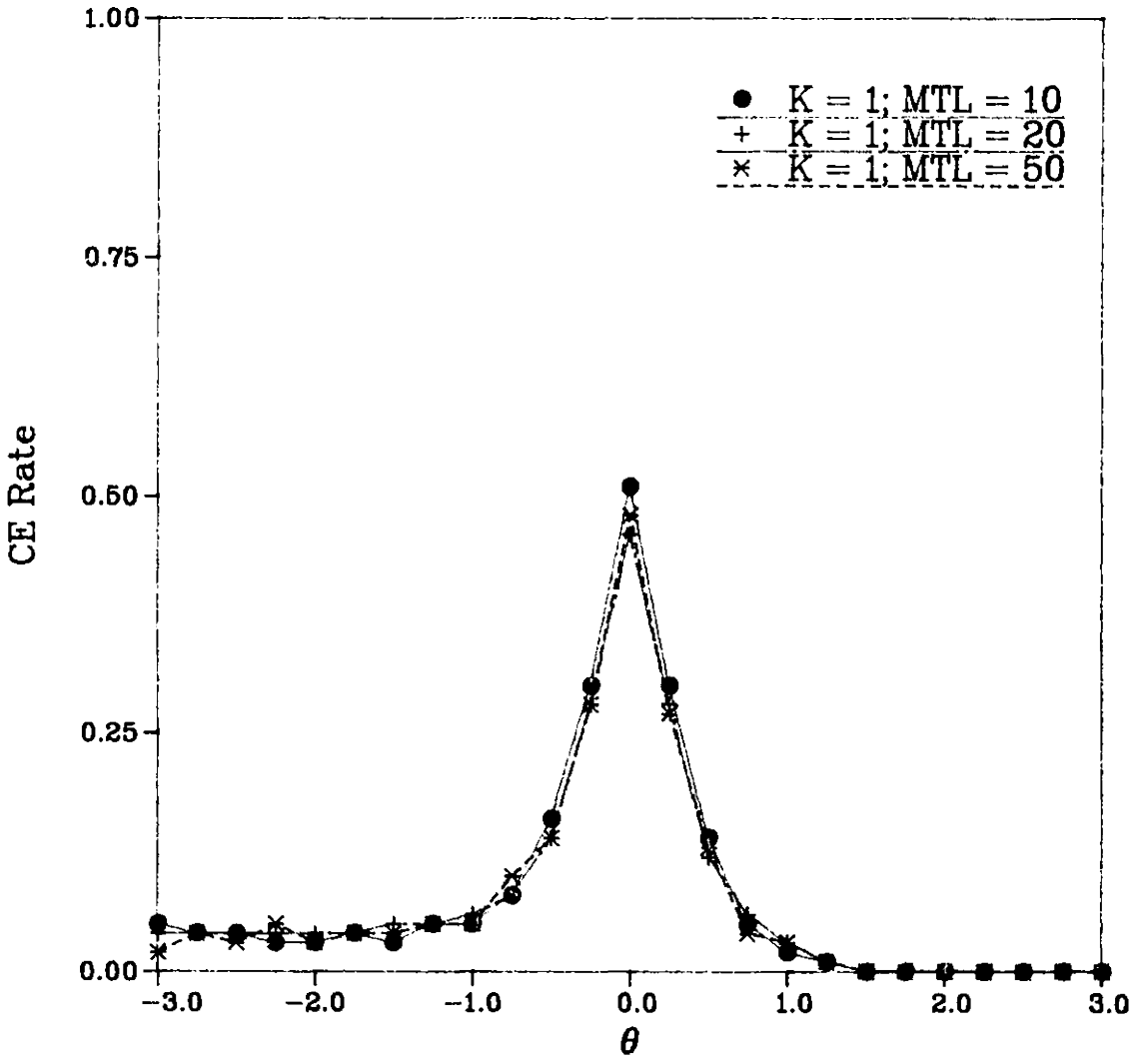
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 3

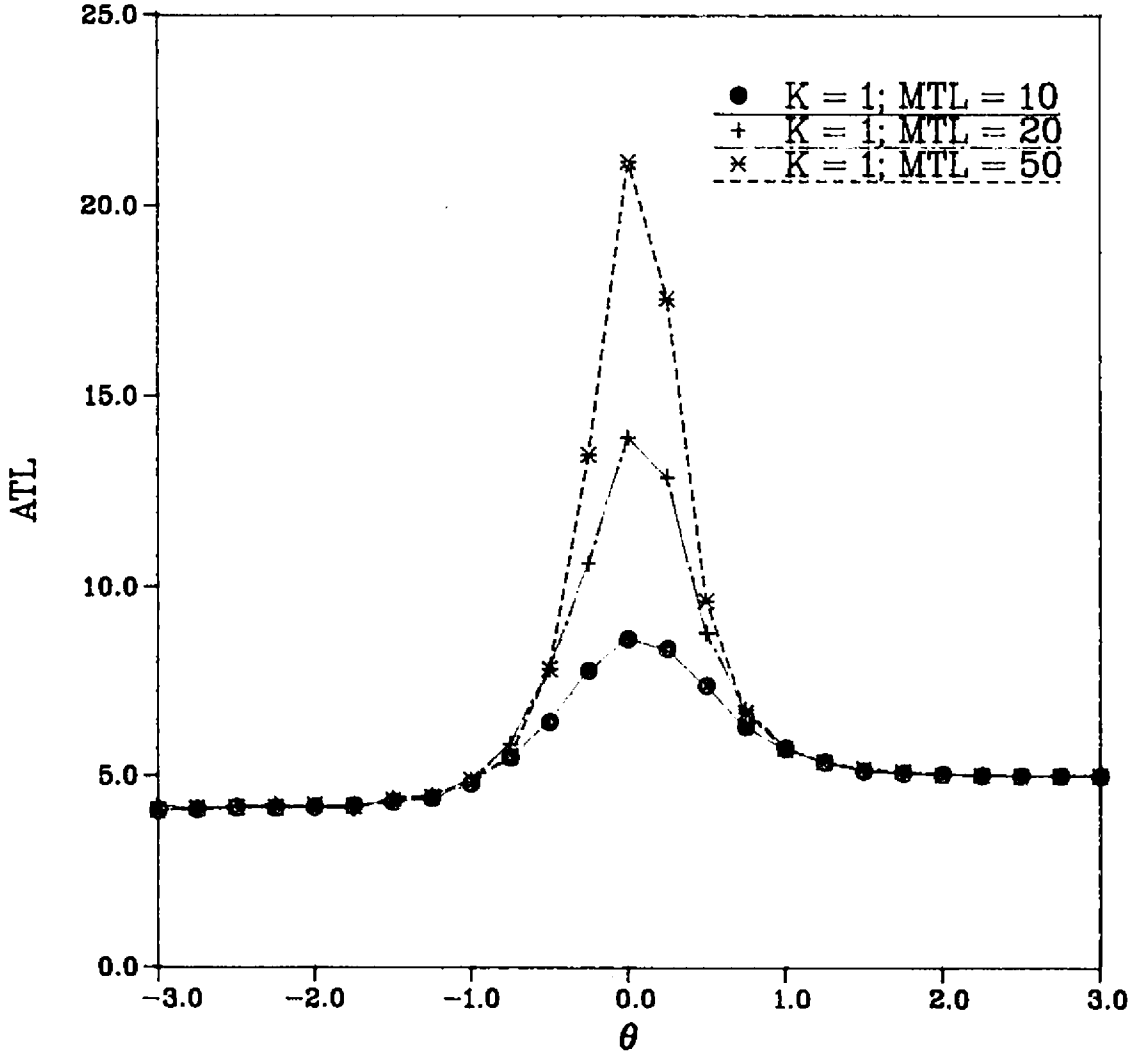
25.0
20.0
15.0
10.0
5.0
0.0
ATL
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 4

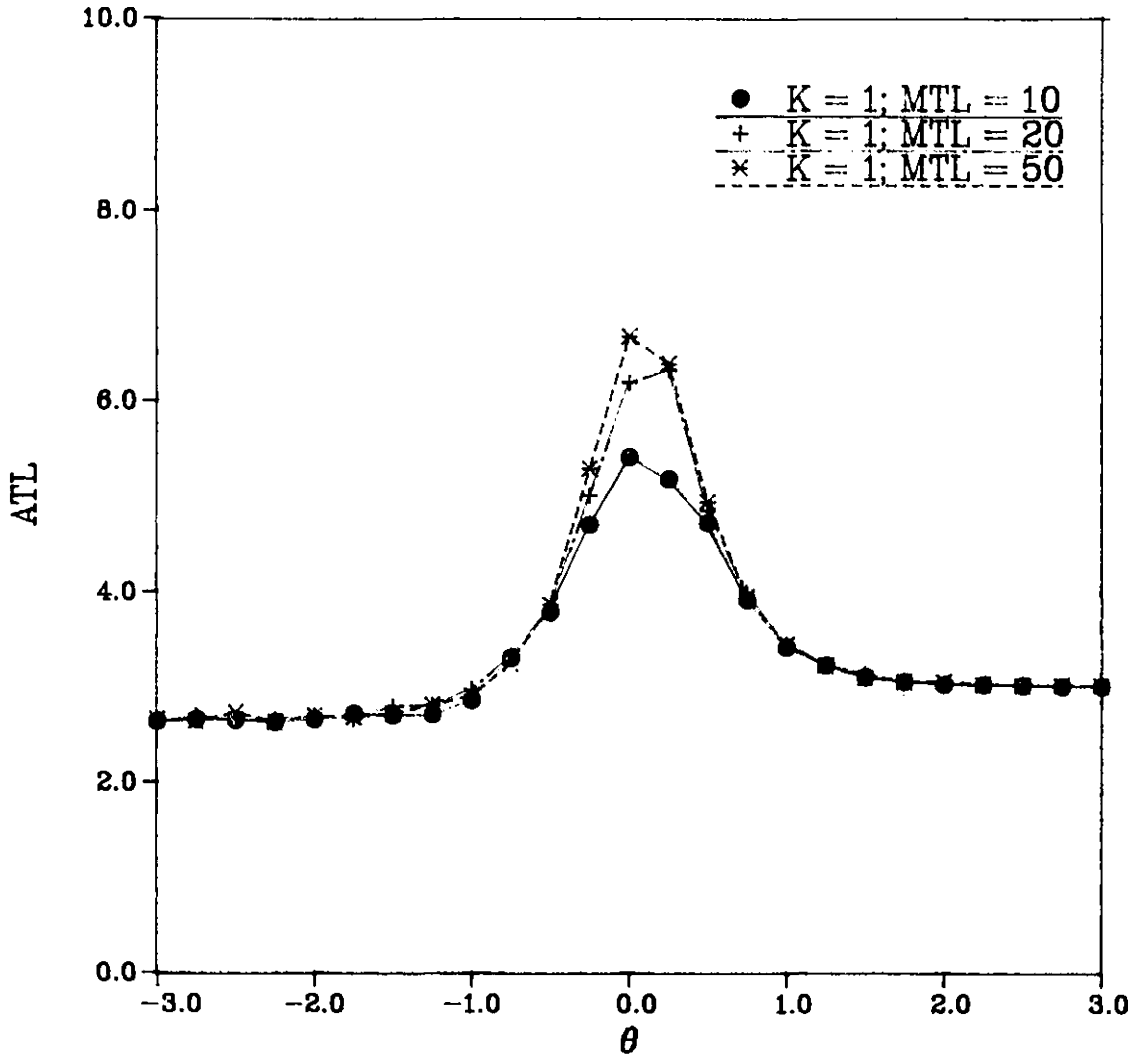
10.0
8.0
6.0
4.0
2.0
0.0
ATL
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 5

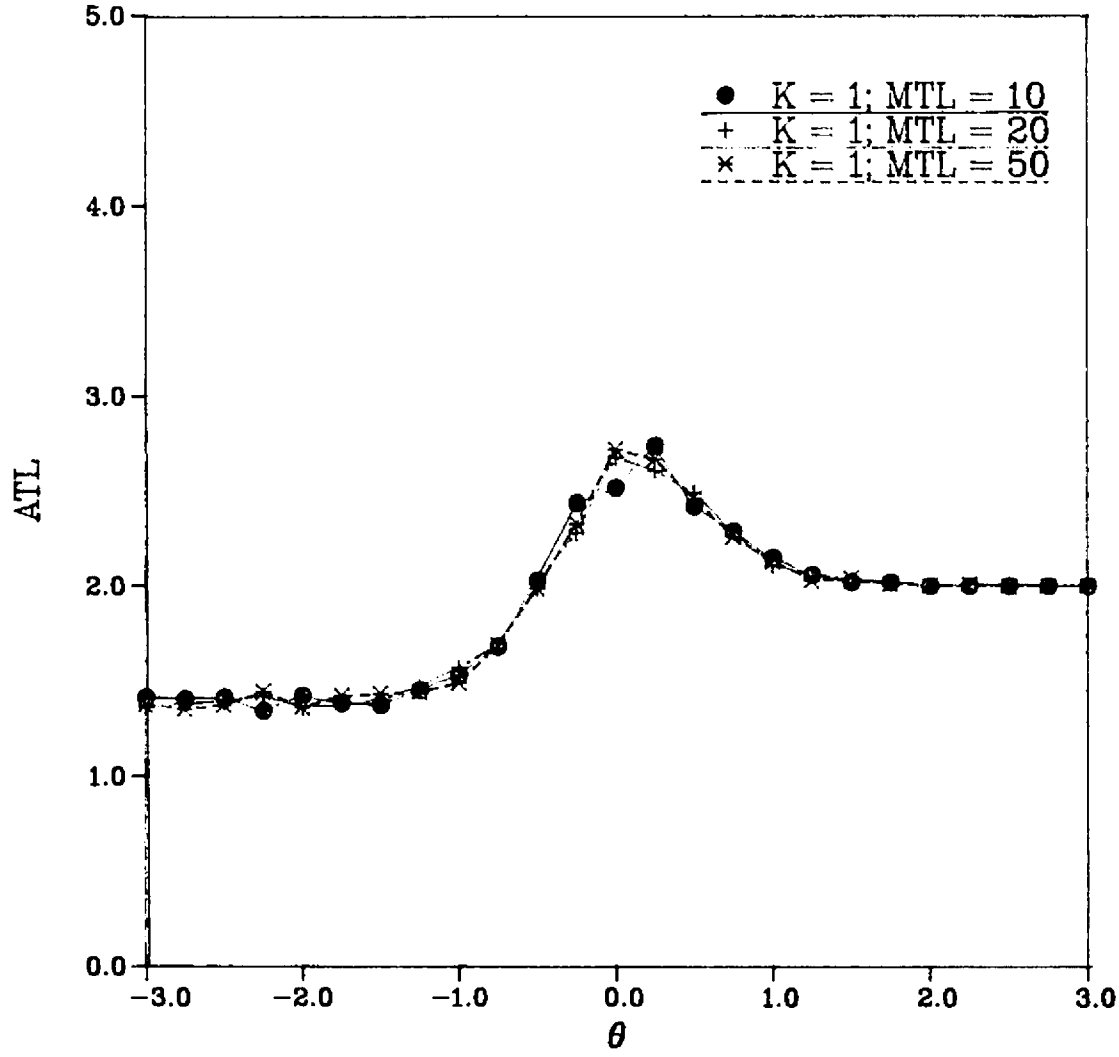
5.0
4.0
3.0
2.0
1.0
0.0
ATL
K = 1; MTL = 10
K = 1; MTL = 20
K = 1; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 6

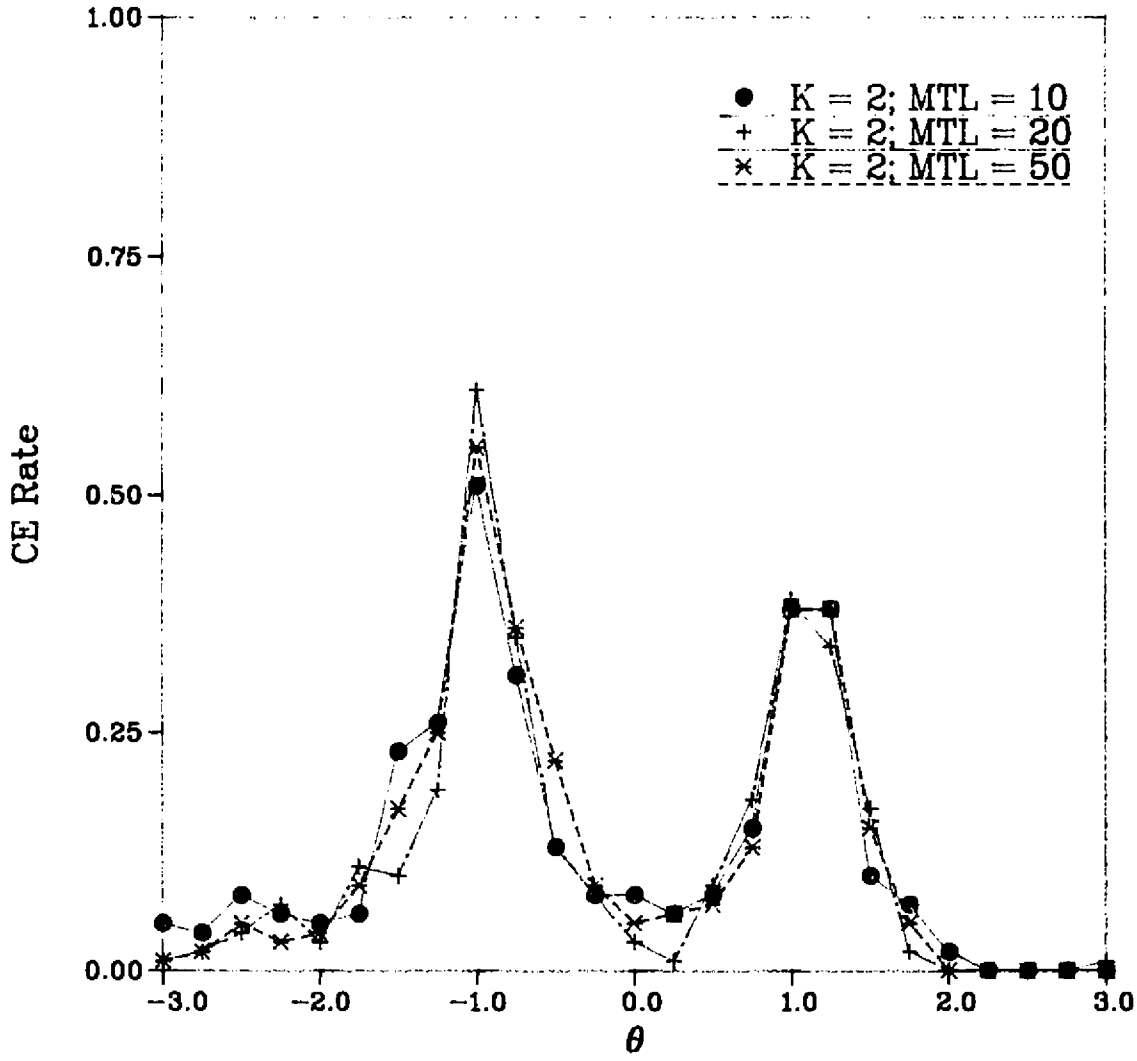
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 2; MTL = 10
K = 2; MTL = 20
K = 2; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 7

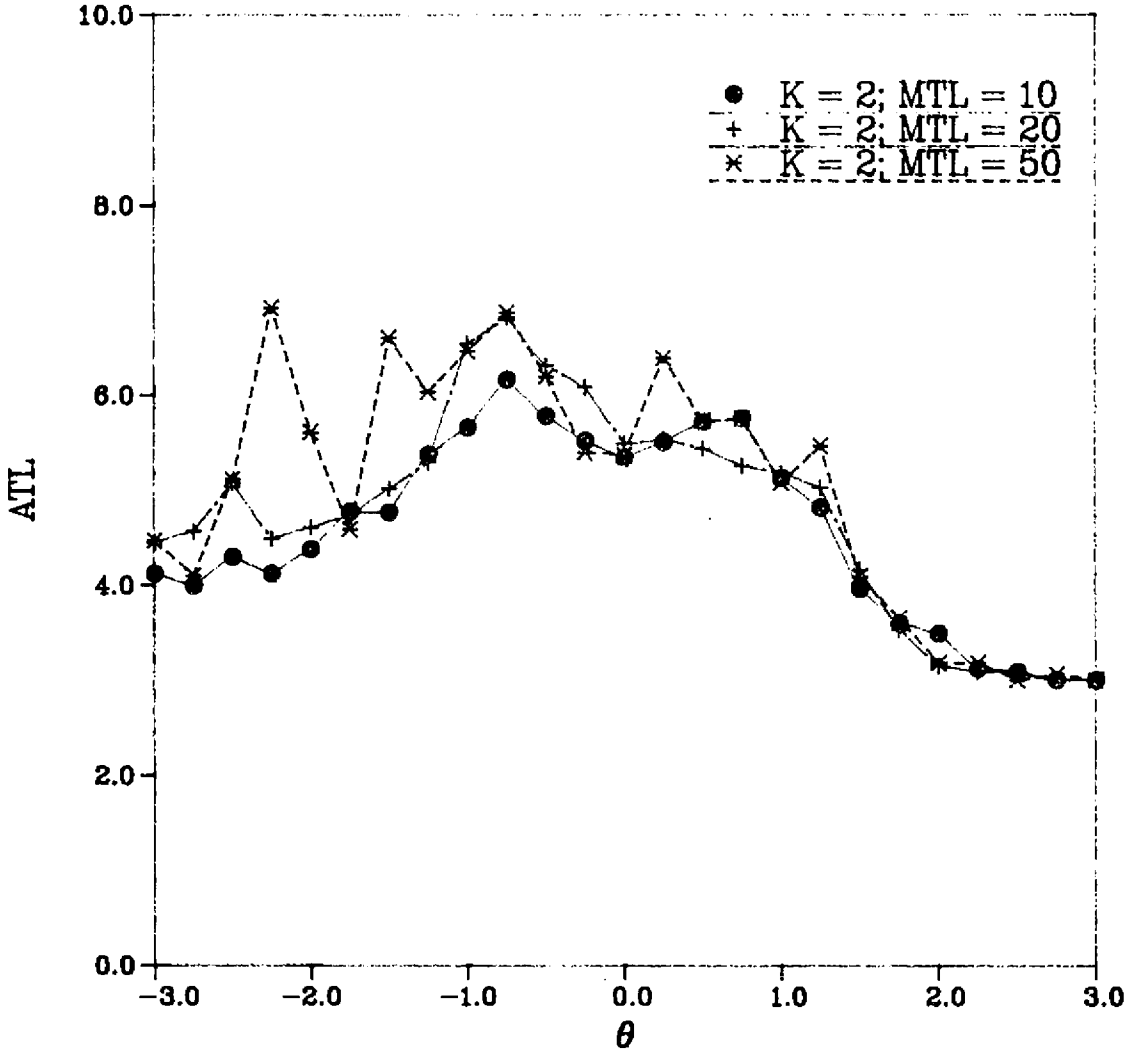
10.0
8.0
6.0
4.0
2.0
0.0
ATL
K = 2; MTL = 10
K = 2; MTL = 20
K = 2; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 8

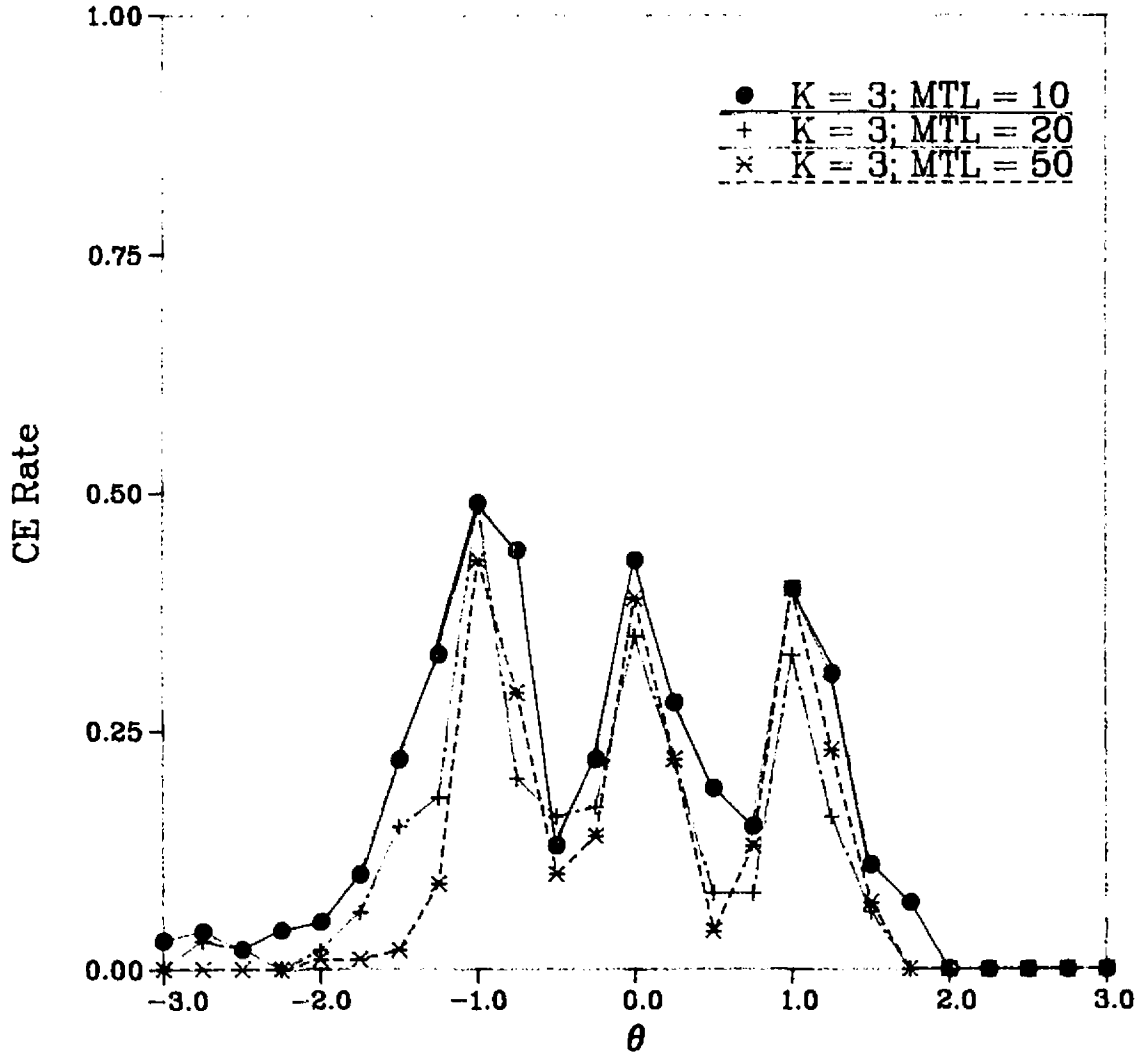
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 3; MTL = 10
K = 3; MTL = 20
K = 3; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 9

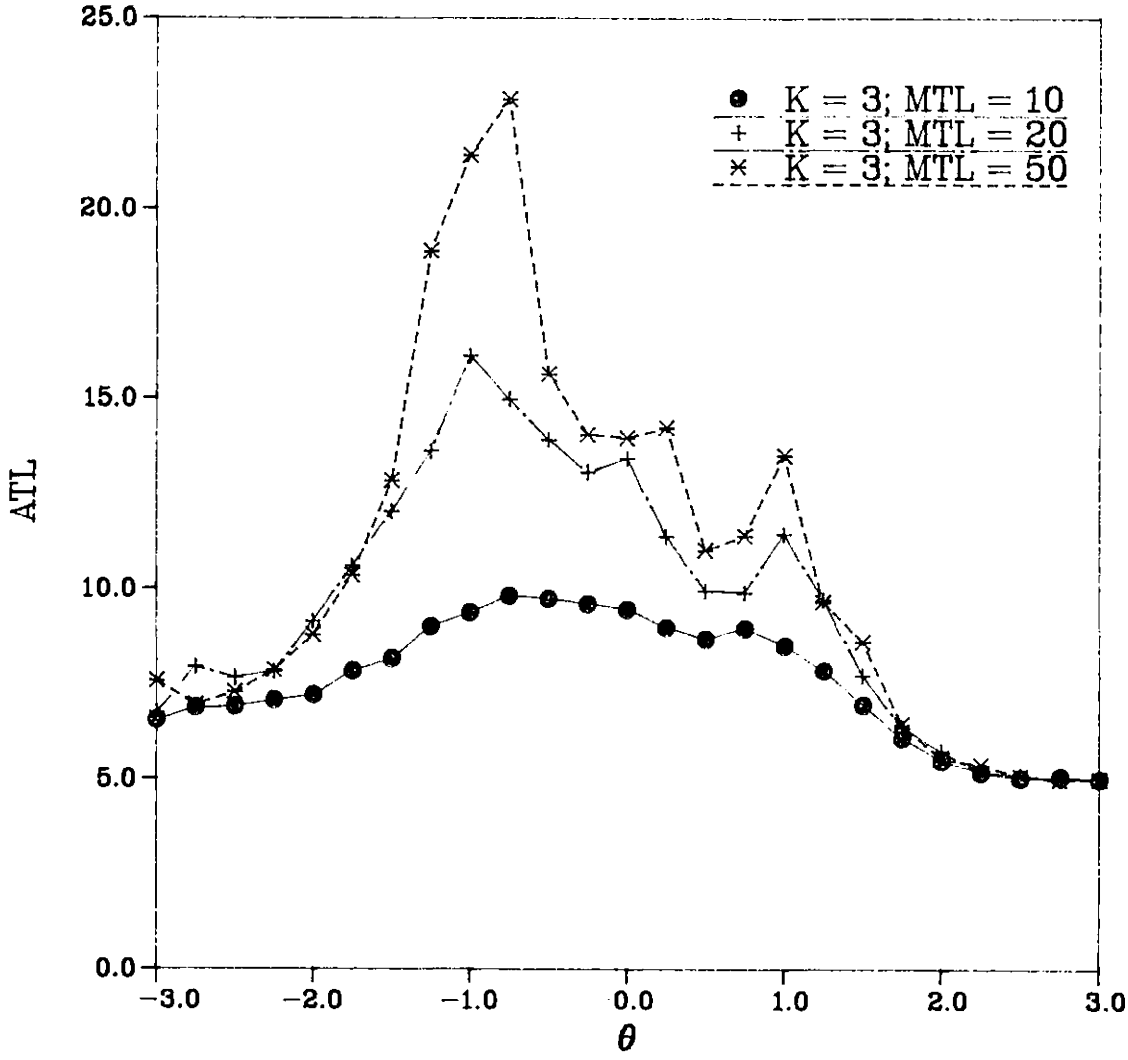
25.0
20.0
15.0
10.0
5.0
0.0
ATL
K = 3; MTL = 10
K = 3; MTL = 20
K = 3; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 10

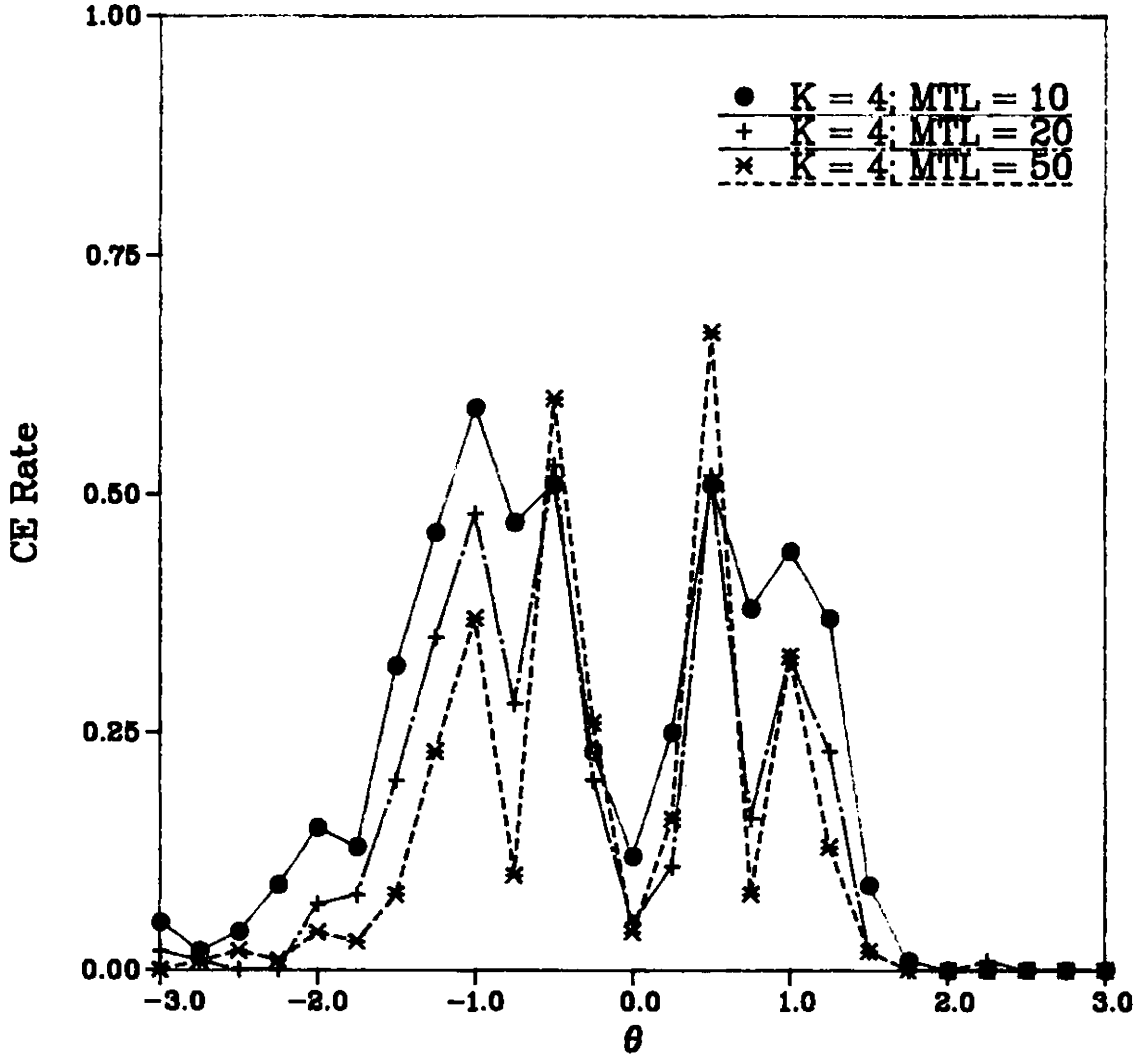
1.00
0.75
0.50
0.25
0.00
CE Rate
K = 4; MTL = 10
K = 4; MTL = 20
K = 4; MTL = 50
−3.0
−2.0
−1.0
0.0
1.0
2.0
3.0
θ

Figure 11

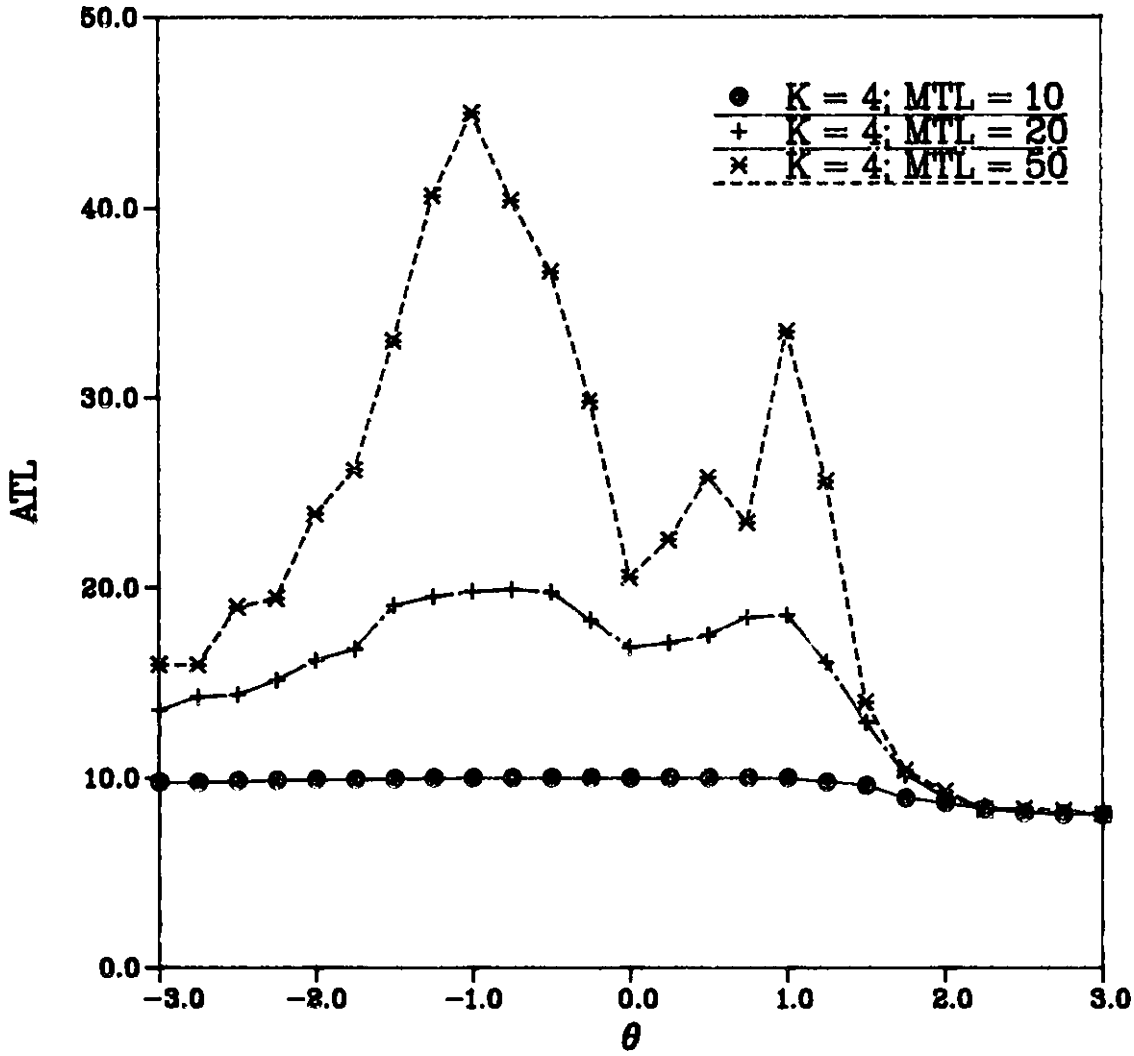
50.0
40.0
30.0
20.0
10.0
0.0
ATL
K = 4; MTL = 10
K = 4; MTL = 20
K = 4; MTL = 50
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
θ

Figure 12