Computerized Mastery Testing With Nonequivalent Testlets

Kathleen Sheehan and Charles Lewis
Educational Testing Service

A procedure for determining the effect of testlet nonequivalence on the operating characteristics of a testlet-based computerized mastery test (CMT) is introduced. The procedure involves estimating the CMT decision rule twice—once with testlets treated as equivalent and once with testlets treated as nonequivalent. In the equivalent testlet mode, the likelihood functions (LFs) estimated for specific number-correct (NC) scores are assumed to be constant across testlets and a single set of cutscores is estimated for all testlets. In the nonequivalent testlet mode, the LFs estimated for specific NC scores are allowed to vary from one testlet to another and a different set of cutscores is estimated for each permutation of testlet presentation order. Small differences between the estimated operating characteristics of the equivalent testlet decision rule and the nonequivalent testlet decision rule indicate that the assumption of equivalent testlets was warranted. This procedure is demonstrated with data from the Architect Registration Examination.
Index terms: Bayesian methods, computerized mastery testing, decision theory, item response theory, test equivalence, testlets.

Sequential testing techniques can be used to reduce the average test length of a computerized mastery test (CMT) while holding the number of misclassified examinees to an acceptable minimum. In Ferguson (1969a, 1969b), Reckase (1983), and Kingsbury and Weiss (1983), sequential mastery testing procedures are implemented at the item level. That is, the decision to classify or to continue testing is made after each item has been administered. The procedure proposed by Ferguson is an adaptation of Wald's (1947) sequential probability ratio test (SPRT) in which examinees' responses to items are treated as independent Bernoulli trials. The procedure proposed by Reckase is a modification of the SPRT in which the probability of a correct response to an item is allowed to vary from one item to the next. This probability is estimated using an item response theory (IRT) model. The procedure proposed by Kingsbury and Weiss also assumes an IRT model for item responses but differs from the Reckase procedure in that classification decisions are made using Bayesian confidence intervals.

A testlet-based sequential mastery testing procedure has been proposed by Lewis and Sheehan (1990). In this alternative approach, items are administered to examinees in randomly selected blocks called testlets. The decision rule is specified in terms of the cumulative number of correct responses obtained by the examinee at the completion of each testlet. Examinees with low cumulative number-correct (NC) scores are failed, those with high cumulative NC scores are passed, and those with scores indicating an intermediate level of mastery are required to respond to an additional testlet. The cutoff values to which examinees' scores are compared are determined using Bayesian decision theory. As in Reckase (1983) and Kingsbury and Weiss (1983), item responses are modeled using IRT. In a simulated application of this approach to a professional certification examination, it was shown that average test lengths could be reduced by half without sacrificing classification accuracy (Lewis & Sheehan, 1990).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 1, March 1992, pp. 65-76

0146-6216/92/010065-12$1.85

To implement the CMT, a pool of testlets must be created. In Lewis and Sheehan (1990), all testlets were assumed to be parallel. That is, the testlets were (1) composed of the same number of items, (2) equivalent with respect to content coverage, and (3) equivalent with respect to the likelihood of particular NC scores at proficiency levels located near the master/nonmaster cutoff value. There are several reasons for restricting the pool to parallel testlets:

1. Because the number of testlets administered to each examinee is variable, using parallel testlets insures that the tests administered to different examinees are equally difficult and cover the same content areas, even when they are not the same length.
2. Because there are only a finite number of testlets in the pool, use of equivalent testlets minimizes the impact of sampling without replacement.
3. The equivalent testlet design simplifies the computations needed to determine the optimal decision rule.

Although parallel testlets are desirable, they are not always readily available. In particular, the requirement of equivalent likelihoods for NC scores at proficiency levels located near the master/nonmaster cutoff value may be difficult to achieve. Thus, a procedure for evaluating the degree of testlet equivalence is needed. This paper introduces such a procedure and demonstrates its use with two different testlet pools from a professional certification examination.

The procedure for evaluating testlet equivalence introduced here assumes that the first two equivalence criteria have been met and considers testlet nonequivalence with respect to the third criterion only. The procedure involves estimating the CMT decision rule twice, once with testlet likelihoods treated as equivalent and once with testlet likelihoods treated as nonequivalent. In the equivalent mode, the likelihood functions (LFs) estimated for specific NC scores are assumed to be constant across testlets and a single set of cutscores is estimated for all testlets. In the nonequivalent mode, the LFs estimated for specific NC scores are allowed to vary from one testlet to another and a different set of cutscores is estimated for each permutation of testlet presentation order. Cutscores associated with more difficult testlets will be somewhat lower, and those for easier testlets will be somewhat higher. Passing and failing cutscores will be further apart for less discriminating testlets and closer together for more discriminating testlets.

When these two decision rules are applied to a single set of simulated data, differences in their operating characteristics can easily be determined. The test operating characteristics or outcome properties that generally will be of interest include the average test length, the expected number of false positive decisions, the expected number of false negative decisions, and the overall pass rate. Small differences between the test outcome properties of the equivalent testlet CMT and those of the nonequivalent testlet CMT will indicate that the assumption of equivalent testlets was warranted.

Determining an Optimal CMT Decision Rule

In the Lewis and Sheehan (1990) model for mastery testing, items are assumed to follow a unidimensional IRT model with known item parameters. The master/nonmaster cutscore is defined as a point θ_c on the latent proficiency scale. Because it may not always be feasible to specify θ_c precisely, two additional mastery values are considered: θ_n is the highest level at which an examinee will be considered a nonmaster, and θ_m is the lowest level at which an examinee will be considered a master. The region of the θ scale between θ_n and θ_m is referred to as the region of indifference (Reckase, 1983).

Discriminating examinees at θ_m from those at θ_n is treated as a problem in Bayesian decision theory by Lewis and Sheehan (1990). At each stage of testing, decisions are made by selecting the option (pass, fail, or continue testing) that minimizes posterior expected loss. For simplicity, a threshold loss function is used. This loss function is specified in terms of three parameters: A, the loss associated

with passing a nonmaster; B, the loss associated with failing a master; and C, the cost of administering a single testlet.

The prior distribution is specified in terms of two probabilities—P_m, the prior probability of being a master, and $P_n = 1 - P_m$, the prior probability of being a nonmaster. Note that P_m can be interpreted as the prior probability of a candidate having a proficiency level that is greater than or equal to θ_m, and P_n can be interpreted as the prior probability of a candidate having a proficiency level that is less than or equal to θ_n.

In a departure from some IRT-based Bayesian procedures, posterior probabilities are calculated conditional on the observed NC score rather than the entire vector of observed item responses. To simplify the notation let

$$P_{m|i} = \mathrm{P}(\Theta = \theta_m | X_1, X_2, \ldots, X_i) \quad , \tag{1}$$

where X_i is the score observed for the ith testlet administered ($i = 1, \ldots, k$). This probability is calculated iteratively as follows:

$$P_{m|i} = \frac{P(X_i | \Theta = \theta_m) P_{m|i-1}}{P(X_i | \Theta = \theta_m) P_{m|i-1} + P(X_i | \Theta = \theta_n) P_{n|i-1}} \quad , \tag{2}$$

where $P(X_i|\Theta = \theta_m)$ and $P(X_i|\Theta = \theta_n)$ refer to the conditional probability of observing a NC score of X_i, given a proficiency level of θ_m or θ_n, respectively. Procedures for calculating these conditional probabilities differ depending on whether the testlet equivalence assumption is in effect.

In the equivalent testlet mode, the probability of observing a NC score of s at the ith stage of testing is obtained by averaging over all testlets in the pool as follows:

$$P(X_i = s | \Theta = \theta_m) = \exp\left\{ 1/\mathrm{T} \sum_{t=1}^{T} \ln[P(X_i = s | \Theta = \theta_m, t)] \right\} \quad , \tag{3}$$

where t refers to the particular testlet administered at stage i. The testlet-specific probabilities in Equation 3 are obtained as follows:

$$P(X_i = s | \Theta = \theta_m, t) = \sum \prod_{j=1}^{n} P_{jt}(\theta_m)^{x_j} [1 - P_{jt}(\theta_m)]^{1 - x_j} \quad , \tag{4}$$

where the summation is taken over all response patterns such that the NC score is s, x_j is 1 or 0 depending on whether the response pattern considered is defined with a correct or incorrect response to the jth item, and $P_{jt}(\theta_m)$ is the conditional probability of a correct response to the jth item in testlet t by an examinee with proficiency level θ_m (as given by the assumed IRT model). The conditional probability for an examinee at θ_n is defined analogously.

In the nonequivalent testlet mode, testlet likelihoods are not averaged. Instead, score probabilities are calculated conditional on the particular testlet administered, as indicated in Equation 4. This difference in the manner of calculating score probabilities results in differences in the way that expected losses are calculated. For example, consider the calculations needed to determine the expected loss of the continue testing option. The losses associated with all possible outcomes of future testlet administrations are calculated, and then a weighted average of those losses is computed with weights corresponding to the predictive probabilities.

In the equivalent testlet mode, the number of possible outcomes of future testlet administrations is simply $2N$, where N is the number of possible future stages of the test. In the nonequivalent testlet mode, the number of possible outcomes of future testlet administrations depends on T (the size of the testlet pool) and on N, because each combination of testlets that could possibly be administered

at future stages of the test must be considered as a separate test outcome. Thus, the number of possible test outcomes can very quickly become quite large. The procedure introduced here handles this combinatorial problem by calculating expected losses conditional on specific permutations of testlet presentation order.

To illustrate this procedure, consider a test with a maximum of four stages and a testlet pool containing 10 testlets labeled 0 through 9. This scenario allows for a total of 5,040 ($10 \times 9 \times 8 \times 7$) permutations of testlet presentation order. Associated with each permutation is a set of eight possible test outcomes. The set of all possible outcomes associated with the permutation (4,3,7,8) can be enumerated as follows:

1. P_4
2. F_4
3. $C_4\ P_3$
4. $C_4\ F_3$
5. $C_4\ C_3\ P_7$
6. $C_4\ C_3\ F_7$
7. $C_4\ C_3\ C_7\ P_8$
8. $C_4\ C_3\ C_7\ F_8$

where P_i indicates that the examinee's score on Testlet i was in the "pass immediately" range, F_i indicates that the examinee's score on Testlet i was in the "fail immediately" range, and C_i indicates that the examinee's score on Testlet i was in the "continue testing" range. The expected losses associated with each of these eight outcomes can easily be determined from the equations given in Lewis and Sheehan (1990), with the specific score probabilities for Testlets 4, 3, 7, and 8 substituted for the single pool-wide average probability considered previously. That is, testlet likelihoods are calculated using Equation 4 instead of Equation 3.

This change in the method of determining cutscores requires a compensating change in the method of administering testlets—instead of randomly selecting the next testlet to be administered after each completed testlet (as is done in the equivalent testlet design), a random permutation of testlet presentation order is selected at the start of each examinee's testing session. This permutation determines the testlets that will be administered to the examinee at all future stages of the test and the cutscores to be used for making the pass/fail/continue testing decision at each stage of testing. Not all permutations of testlet presentation order will necessarily lead to a distinct set of cutscores. For example, in the extreme case of identically equivalent testlets, all permutations of testlet presentation order would yield the exact same set of cutscores.

The final step of the nonequivalent testlet procedure is to apply the equivalent testlet decision rule and the nonequivalent testlet decision rule to a single set of simulated data. Testlet equivalence can then be evaluated by comparing the test-outcome properties of the equivalent testlet decision rule with those of the nonequivalent testlet decision rule.

An Application

The data available for this study were derived from the Architect Registration Examination (ARE), a professional certification examination administered annually by the National Council of Architectural Registration Boards. This examination covers eight skill areas, known as divisions, that are administered as independently scored subtests. Two divisions were selected for consideration in this study: Division E, a test of the skill area known as Structural Technology—Lateral Forces, and Division D/F, a test of the skill area known as Structural Technology—General and Long Span. Each of these divisions had previously been administered in a paper-and-pencil (P&P) format.

The test specifications for the P&P ARE examinations had been developed to insure adequate content coverage and acceptable test reliability. The P&P tests consisted of 60 items and 125 items for Division E and D/F, respectively. In order to satisfy all content specifications at the testlet level, testlet lengths of 10 and 25 items were selected for the Division E and D/F CMTs, respectively. The Division E CMT had a minimum test length of two testlets (20 items) and a maximum test length of six testlets (60 items); the Division D/F CMT had a minimum test length of two testlets (50 items) and a maximum test length of five testlets (125 items).

Both the Division E items and the Division D/F items were adequately modeled by a three-parameter logistic IRT model. For each division, a pool of equivalent testlets was constructed by assigning items to testlets so that (1) all testlets were composed of the same number of items, (2) all testlets satisfied the content specifications, (3) all testlets had similar mean difficulty and discrimination levels, and (4) none of the testlets had large variances of difficulty or discrimination. This method of constructing testlets was expected to lead to similar testlet LFs in the region of the cutscore. The Division E item pool yielded a total of six testlets; the Division D/F item pool yielded a total of 10 testlets.

The LFs estimated for the six 10-item Division E testlets are plotted in Figure 1a, which shows one set of LFs for nonmasters and one set for masters. The plot shows considerable agreement: For each testlet, a maximally competent nonmaster is most likely to score 4 out of 10, whereas a minimally competent master is most likely to score 7 out of 10. The LFs that were estimated for the 10 25-item Division D/F testlets are given in Figure 1b. Note that the Division D/F testlets show slightly less agreement.

The Simulation Technique

Two sets of simulated data were constructed, one for each of the divisions. Each simulated dataset contained 100 simulated examinees at each of 40 different proficiency levels. The proficiency levels selected for each division corresponded to the 20 NC score points located immediately above and below the cutscore defined for the current operational P&P test. In order to provide weighted simulation results reflecting the expected proportion of examinees at each of the selected NC score points, a case weight was generated for each simulated examinee. Case weights were defined so that the weighted distribution of NC scores in the simulated dataset matched the observed distribution of NC scores in a recent administration of the P&P test.

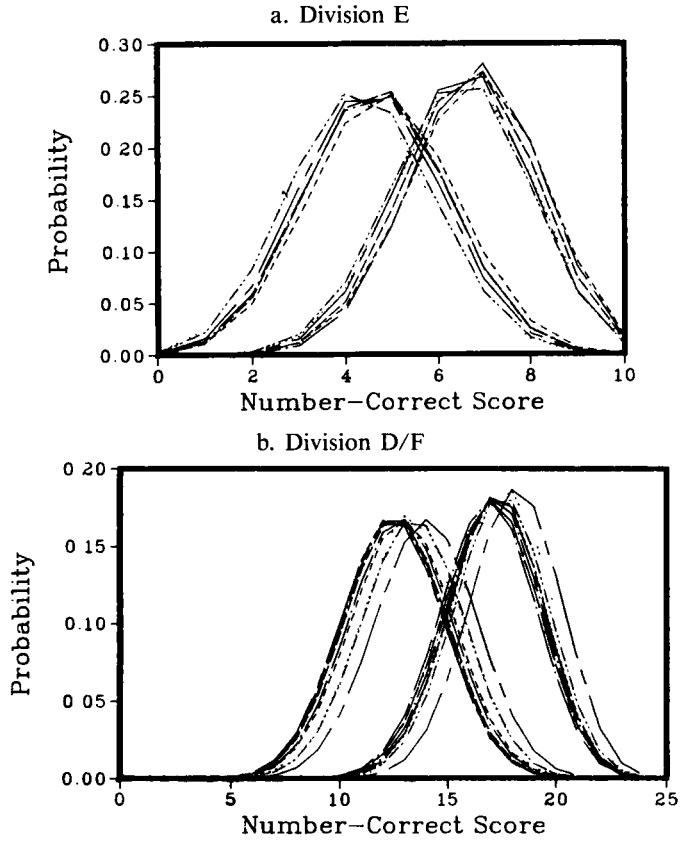
For each simulated examinee, item responses were generated according to the three-parameter logistic IRT model (using actual ARE item parameters) for the items that appeared on the P&P test and the items that appeared in the CMT testlet pools. There was no overlap between the items included on the P&P test and the items included in the CMT testlet pools.

Determining the Decision Rules

For each division, three decision rules were generated—one estimated from the simulated P&P data, and two estimated from the simulated CMT data. The P&P decision rule was estimated as a baseline from which to gauge the performance of the two CMT decision rules.

For each division, the proficiency estimate corresponding to a maximally competent nonmaster was taken to be the θ value located 1.5 standard errors of measurement below the cutscore. (A standard error of measurement at the cutscore was calculated for each division from data that had been collected in a recent administration of the P&P test.) Similarly, the proficiency estimate corresponding to a minimally competent master was taken to be the θ value located 1.5 standard errors of measurement above the cutscore. The loss function parameters were taken to be $A = 40$ for misclassifying a nonmaster as a master, and $B = 20$ for misclassifying a master as a nonmaster, on a

Figure 1
Testlet-Specific Likelihood Functions for the Number-Correct Score on a Division E Testlet and a Division D/F Testlet (Different Line Types Represent Different Testlets)



scale in which 1 corresponds to the loss associated with administering an additional testlet. This loss function is referred to as a 40/20 loss function. These values were selected based on results obtained in the simulation studies reported in Lewis and Sheehan (1990).

The decision rule estimated for the simulated P&P data was determined by considering each P&P form as a single testlet. Thus, the 60-item Division E P&P test was considered to be a single 60-item testlet, and the 125-item Division D/F P&P test was considered to be a single 125-item testlet. For each division, a maximum test length of one testlet was specified, causing the decision rule estimation program to terminate at the completion of the first testlet. In this way, the decision rule that minimized posterior expected loss was estimated as the first stage cutscore such that, for examinees with NC scores below the cutscore, the posterior expected loss of a fail decision was less than that of a pass decision; conversely, for examinees with NC scores above the cutscore, the posterior expected loss of a pass decision was less than that of a fail decision. Note that, because each pool contained just one testlet, this approach did not require the assumption of equivalent testlets. The cutscores determined in this manner were 34 for the 60-item Division E test and 77 for the 125-item Division D/F test.

For the simulated CMT data, two decision rules were estimated—one that incorporated the

assumption of equivalent testlets and one that did not. To determine the decision rule that incorporated the assumption of equivalent testlets, the LFs given in Figure 1 were averaged using Equation 3. The pool-wide average LFs obtained then were used to determine the posterior expected loss of the various test outcomes at each stage of testing. The resulting decision rules are listed by stage in Table 1. (Stage 1 NC cutscores were not estimated because the CMT test specifications called for a minimum of two testlets per examinee.)

Table 1
Number-Correct Cutscores Calculated Under the Assumption of Equivalent Testlets

Division and Stage	Number of Items	Maximum Fail Score	Minimum Pass Score
Division E			
2	20	9	15
3	30	15	20
4	40	21	26
5	50	27	31
6	60	34	35
Division D/F			
2	50	28	35
3	75	43	50
4	100	59	65
5	124	77	78

The cutscores estimated for the nonequivalent testlet CMT were based on the testlet-specific LFs given in Figure 1, rather than the pool-wide average LFs given by Equation 3. For the Division E pool, testlet-specific cutscores were estimated for each of the $6! = 720$ possible permutations of testlet presentation order. A frequency distribution for the unique cutscores obtained at each stage of testing is given in Table 2. The table shows, for example, that of the 720 permutations of testlet presentation order analyzed, 72% resulted in a Stage 2 maximum fail score of 9, and 28% resulted in a Stage 2 maximum fail score of 10. In general, there were only two or three unique pass or fail cutscores obtained at each stage of testing. Table 2 also indicates which cutscores correspond to the cutscores obtained under the equivalent testlet model. At each stage of testing, the testlet-specific cutscores with the highest frequency are those that correspond to the equivalent testlet model.

A slightly different procedure was employed to obtain testlet-specific cutscores for the Division D/F pool. Specifically, because the Division D/F pool contained 10 testlets and the test specifications called for a maximum of five testlets per examinee, full calibration of the entire set of testlet presentation permutations would have involved 30,240 ($10 \times 9 \times 8 \times 7 \times 6$) runs of the cutscore estimation program. Because the comparison of test outcome properties could be performed on a random subset of permutations, a random subset of 1,000 permutations was selected for calibration. The results also are summarized in Table 2. At each stage, the calibration yielded between five and six unique sets of cutscores. Note that the equivalent testlet cutscores always appear at the center of the distribution and generally have the highest frequency.

Simulation Results

The classification performance of the P&P test, the equivalent testlet CMT (EQ CMT), and the nonequivalent testlet CMT (NE CMT) were estimated by applying the decision rules noted above to the

Table 2
Testlet Specific Number-Correct Cutscores Calculated for the Division E and Division D/F Pools Without Assuming Equivalent Testlets, and Proportion of Times Obtained in Analysis of Testlet Presentation Permutations

Stage and Maximum Fail Score	Proportion	Minimum Pass Score	Proportion
Division E			
Stage 2			
9*	.72	14	.22
10	.28	15*	.78
Stage 3			
14	.07	19	.01
15*	.84	20*	.70
16	.09	21	.29
Stage 4			
20	.16	25	.29
21*	.84	26*	.71
Stage 5			
26	.11	30	.07
27*	.89	31*	.93
Stage 6			
33	.01	34	.01
34*	.99	35*	.99
Division D/F			
Stage 2			
26	.10	33	.09
27	.25	34	.30
28*	.42	35*	.38
29	.22	36	.22
30	.01	37	.01
Stage 3			
41	.06	48	.10
42	.23	49	.27
43*	.36	50*	.36
44	.29	51	.23
45	.06	52	.04
Stage 4			
57	.11	62	.01
58	.30	63	.15
59*	.34	64	.31
60	.21	65*	.32
61	.04	66	.19
		67	.02
Stage 5			
74	.03	75	.03
75	.15	76	.15
76	.33	77	.33
77*	.31	78*	.31
78	.17	79	.17
79	.01	80	.01

*Cutscore for the equivalent testlet design.

Table 3
Comparison of Overall Pass Rates for the Two Divisions

Test	Division E	Division D/F
P&P	.57	.40
EQ CMT	.56	.39
NE CMT	.55	.40
True Pass Rate	.62	.42

simulated response vectors described previously. Overall passing rates calculated from the weighted data are given in Table 3. For both divisions, the pass rates calculated for the two CMTs were almost identical. This indicates that the deviations from equivalence evidenced in the testlet-specific LFs shown previously were not sufficient to have a differential impact on the overall CMT pass rates. Table 3 also provides the "true" pass rate determined from the generating θs. For both divisions, the "true" pass rate was higher than the pass rates calculated from the simulated data because the 40/20 loss function (which was built into all the simulated decision rules) assigned a larger loss to a false positive decision, thus reducing the number of examinees classified as masters.

The unweighted classification results are displayed in Figure 2, which presents the percent of examinees classified as masters by the two CMTs, plotted as a function of θ. The fact that the two curves for each division are nearly indistinguishable indicates that the nonequivalent testlet CMT did not provide significantly improved classification accuracy over the equivalent testlet CMT.

Table 4 presents error rates calculated from the weighted classification results for true nonmasters and true masters in each division. In both divisions, the error rates listed for the equivalent testlet CMT and the nonequivalent testlet CMT are very similar. For example, in Division E, both CMTs had a false positive error rate of 6% (6% of true nonmasters were falsely classified as masters) and a false negative error rate between 13% and 14% (between 13% and 14% of true masters were falsely classified as nonmasters). The similar error rates obtained for the two CMTs provide further evidence that the equivalent testlet CMT and the nonequivalent testlet CMT did not differ in their classification accuracy. Table 4 also shows slightly lower classification accuracy for the P&P test, as compared to either one of the two CMTs. This difference can be attributed to the fact that the CMTs were created to be peaked at the cutscore, whereas the P&P test was not.

Table 4
Percent Classified as Nonmaster (PCNM) and Percent Classified as Master (PCM) for True Nonmasters and True Masters, By Division

Division and Test	True Nonmasters PCNM	True Nonmasters PCM	True Masters PCNM	True Masters PCM
Division E (2,462 True Nonmasters and 3,963 True Masters)				
P&P	89	11	14	86
EQ CMT	94	6	13	87
NE CMT	94	6	14	86
Division D/F (4,957 True Nonmasters and 3,601 True Masters)				
P&P	92	8	16	84
EQ CMT	94	6	16	84
NE CMT	93	7	14	86

Figure 2
Percent Classified as Master for the Equivalent Testlet CMT (Solid Line) and the Nonequivalent Testlet CMT (Dashed Line) for Division E and Division D/F Testlets

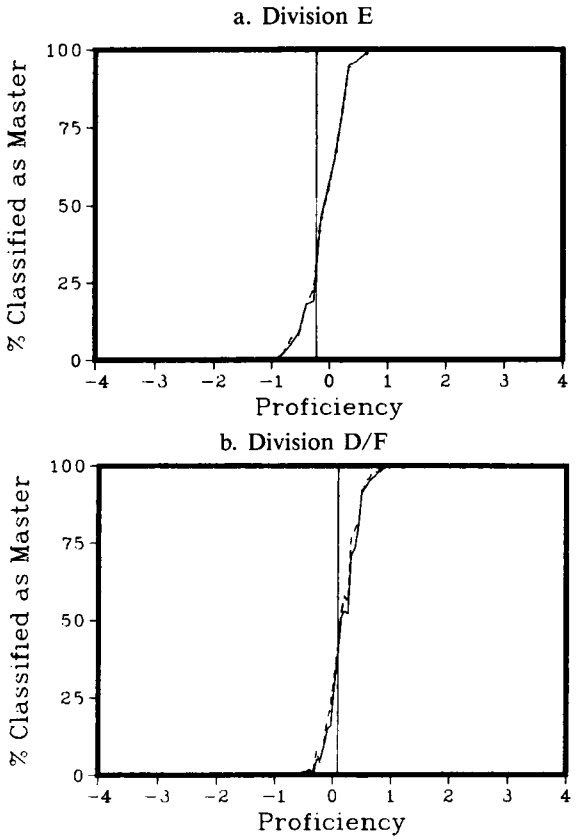


Table 5 summarizes the simulation results in terms of the minimum, maximum, and mean test length observed for each test. Table 5 shows no significant differences between the two CMTs with respect to these statistics. An alternative view of the test length data is provided in Figure 3, which depicts average test length plotted as a function of θ. Because 100 response vectors were generated at each θ value, each point is an average of 100 values. For both divisions, the plots show minimal differences.

Discussion

The results of the CMT procedure showed that it was feasible (although costly in terms of computer processing time) to construct the testlet-specific sets of cutscores needed for the nonequivalent testlet CMT design and that, for the two testlet pools studied, dropping the assumption that all testlets are equivalent had negligible impact on classification performance. These results can be interpreted as a validation of the decision to score the ARE testlet pools using the equivalent testlet design.

A drawback of the method proposed here is that, although the different sets of cutscores are specifically constructed to insure that all examinees are treated equally with respect to the measurement properties of their particular set of testlets, the fact that different cutscores are used for

Table 5
Minimum, Mean, and Maximum Number of Items Administered to True Masters and True Nonmasters, By Division

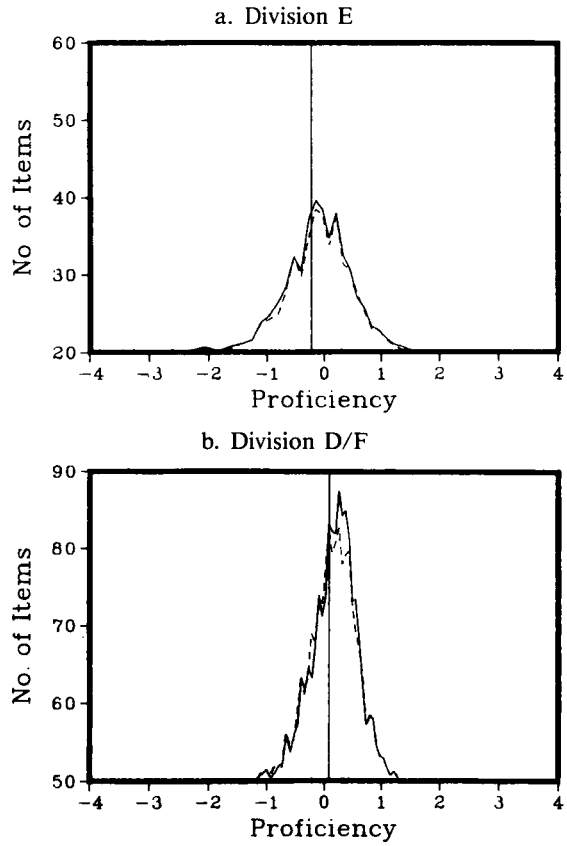
Division and Test	True Nonmasters			True Masters		
	Min	Mean	Max	Min	Mean	Max
Division E						
EQ CMT	20	27	60	20	29	60
NE CMT	20	26	60	20	29	60
Division D/F						
EQ CMT	50	60	125	50	68	125
NE CMT	50	61	125	50	67	125

different individuals may be objectionable to some examinees. This objection does not hold, however, for the equivalent testlet model in which a single set of cutscores is used for all examinees.

The implications of the research described in this paper include the following:

1. It is now possible to employ testlet-based CMTs when testlets are not constructed to be equivalent. This capability may prove useful for adaptive testing applications.

Figure 3
Average Test Length for the Equivalent Testlet CMT (Solid Line) and the Nonequivalent Testlet CMT (Dashed Line) for Division E and Division D/F Testlets



2. The methods and techniques used in the simulation provide a validation procedure that can be applied to any testlet pool developed to contain equivalent testlets. It is recommended that such validation be performed, on a routine basis, as new equivalent testlet pools are developed.

References

Ferguson, R. L. (1969a). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh PA: University of Pittsburgh Learning and Research Development Center. (ERIC Documentation Reproduction Service No. ED 037 089)

Ferguson, R. L. (1969b). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction.* Unpublished doctoral dissertation, University of Pittsburgh. (University Microfilms No. 70-4530)

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). New York: Academic Press.

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

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.

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

Acknowledgments

This research was supported by Educational Testing Service through the Program Research Planning Council.

Author's Address

Send requests for reprints or further information to Kathleen Sheehan, Educational Testing Service, Princeton NJ 08541, U.S.A.