A New Item Selection Procedure for Mixed Item Type in Computerized Classification Testing

C. Allen Lau
Harcourt Educational Measurement

Tianyou Wang
ACT

Paper presented at the 2000 AERA Annual Meeting in New Orleans, Louisiana, April 2000.
Please make correspondence to:
C. Allen Lau, Ph.D.
Project Director II (Psychometrician)
Harcourt Educational Measurement
555 Academic Court
San Antonio, Texas 78204
E-mail: allen_lau@harcourt.com

Introduction

Computerized Classification Testing

Mastery testing is used to classify the test takers into one of two categories: mastery (pass) or non-mastery (fail). Certification or licensure testing is a good example of it. When such tests are administered and scored in computer format, it is referred to as computerized classification testing (CCT) (Spray, Abdel-fattah, Huang, & Lau, 1997). The main objective of CCT is to make accurate mastery decisions with lowest possible cost, including the testing time and the number of items administered.

To implement an IRT-based CCT procedure, a cut-point on the ability scale (θ_c) must be established first. Two types of classification errors are considered: if the examinee is classified as a master but in fact his/her ability level (θ) is below θ_c, a false positive error (type I error) occurs; if the examinee is classified as a nonmaster but in fact his/her θ is at or above θ_c, a false negative error (type II error) occurs. The relative importance of these two types of error is situation dependent.

In CCT, sequential probability ratio testing (SPRT) procedure, first proposed by Wald in 1947, was found effective for mastery classification (Spray & Reckase, 1996; Kalohn & Spray, 1998; Lau, 1996; Lau & Wang, 1998, 1999).

Dichotomous, Polytomous, and Mixed Item Pool

In CCT, SPRT procedure works very well with dichotomous items. However, few if any research investigates how to apply polytomous models in computerized adaptive test (CAT) or CCT because of the difficulty of item scoring of the extended response items. Bennett, Steffen, Singley, Morley, and Jacquemin (1997) successfully adopted and scored open-ended format items in CAT that implies the feasibility of polytomous machine scoring in CCT in the future.

Lau and Wang (1998) found that SPRT procedure could be adapted with polytomous items in CCT. More specifically, they found that: (a) SPRT procedure could be adopted in three type of item pools (dichotomous, polytomous, & mixed); (b) SPRT procedure with polytomous item pool achieved better classification accuracy than that with dichotomous item; and (c) best classification accuracy and efficiency in terms of item consumption was gained when item selection was based on Fisher item information at the cutting point.

Lau and Wang (1998) also studied how to select item from item pools with mixed item types. They found that if a maximum information algorithm was used for item selection, polytomous items would have high priority being selected over dichotomous items. In order to control the proportions of each item types being selected, a fixed item selection scheme had to be used, such as selecting one polytomous item and then three dichotomous items.

Fisher and Kullback-Leibler (K-L) information

Eggen (1998) compared Fisher with Kullback-Leibler (K-L) information (Cover & Thomas, 1991) for item selection in the context of SPRT using dichotomous item pool. He found that the performance of the testing algorithms with K-L were sometimes better and never worse than that of Fisher information-based item selection. Lau and Wang (1999) extent this comparison with polytomous item pool and found that the classification accuracy and test efficiency in term of item consumption based on the two information algorithms were very similar.

Item Selection Based on Efficiency

This study focuses on the test efficiency and proposes a new item selection algorithm for mixed item types. By test efficiency, it means achieving the maximum classification accuracy with the least cost. Cost can be measured by the number of items or the testing time. When there is only one type of items (e.g., multiple-choice items), it is very appropriate to simply use the number of items as a measure of cost. When there are more than one item types (e.g., multiple-choice items and open-ended items), using the number of items to measure the cost can be misleading because the cost associated with an open-ended item is much more than that associated with a multiple-choice item in terms of item development, test administration, and scoring. In this case, it is considered more appropriate to use testing time as a measure of cost.

Research has shown that in SPRT, if the most informative items at the cutting point are administered, the best classification accuracy and the least item consumption are guaranteed. In other words, the more/better item information, the more accurate and the less items are needed to make the classifications. However, item information alone does not guarantee the best test efficiency in a mixed-item-type item pool. In order to achieve better test efficiency, an Information-Time index (IT), was created and used as a criterion for item selection in place of the Information index in this study. This index considers both item information and item response time simultaneously.

This new index is calculated by the equation:

item information / item response time

where item response time is the average time (in seconds) the examinees used to respond to this item.

Purposes of the Study

The purposes of this study are (a) to propose a new Information-Time index as the basis for item selection in computerized classification testing (CCT), (b) to investigate how this new item selection algorithm can help improving test efficiency for item pools with mixed item types (i.e., dichotomous items and polytomous items), and (c) to investigate how practical constraints such as item exposure rate control, test difficulty (i.e., cut point), test length constraint, test time constraint affect the effectiveness of this new item selection algorithm.

Methods

Monte Carlo simulation technique was adopted to verify the decision criterion.

Design

The independent variables were manipulated:

1. Item selection algorithm:
   (1) maximum item information (based on Information index)
   (2) maximum information-time (based on Information-Time index)
2. Item information calibration:
   (1) Fisher
   (2) Kullback-Leibler
3. Randomesque item exposure control procedure
   (1) 3-pick-1
   (2) 5-pick-1

4. Test difficulty in terms of cutting theta:
   (1) $\theta_c = -1.0$
   (2) $\theta_c = 1.0$
5. Test length constraint (That is, the examinees must respond to a minimum number of items and not exceed a maximum number of items):
   (1) minimum = 10 items, maximum = 50 items
   (2) minimum = 1 item, maximum = 100 items
6. Time constraint (That is, the examinees cannot use more than the time limited):
   (1) 60 minutes
   (2) 120 minutes

This was a 2x2x2x2x2x2 crossed factorial design and there were 64 combinations of conditions totally. The evaluative criteria included (1) classification accuracy in terms of false positive and false negative error rates, (2) test efficiency: (a) number of items used, and (b) time used (in minutes) to make mastery decision, (3) item exposure rates, and (4) item utilization rate. In addition, the proportion of dichotomous and polytomous item types used, the percentage of the tests forced to terminated by either the test length constraints or time constraints in each condition were also recorded and compared.

Data

Item parameters from the 1996 NAEP Science assessment (O'Sullivan, Reese, & Mazzeo, 1997) were used to build the item pool. Combining three grades (4th, 8th and 12th) together, the assessment consists of 246 dichotomous items (48%) and 266 polytomous items (52%) (208 items of 3 categories, 47 items of 4 categories, 8 items of 5 categories, and 3 items of 6 categories) for the study. The mean and standard deviation of item difficulties were 0.714 and 1.536 respectively. The two types of items were calibrated on the same scale and the items for the three grades were also linked to the same scale. Item response data was generated and CCT was simulated on computer.

Response-Time Index Simulation

In order to create an Information-Time index, a response-time distribution was simulated. The average response time index for each item was generated from some normal distribution.

The means and standard deviations were different for different item types. For items with 2 categories (0/1), the mean was 60 seconds and standard deviation was 10 seconds; for items with 3, 4, 5, & 6 categories, the means were 120, 150, 180, & 210 seconds respectively and their standard deviations were 20 seconds. The simulated average response times were randomly assigned to items with different categories.

Results

The main effects of item selection method were listed in the first columns of Tables 1 and 2. The marginal interaction effects of item selection method (I or IT index) and other independent variables were summarized in the rest of tables.

Table 1 lists the classification accuracy, test efficiency, and the proportion of dichotomous and polytomous item types used. Across all conditions, the average type I, type II errors, item consumption rate, and time used rate (in minutes) were 0.020, 0.022, 14.610, and 25.112 respectively.

Table 2 contains the force-to-termination rates due to either test length constraints or time constraints, the item utilization rates, and the item exposure rates. Across all conditions, about 42% of items in the pool were utilized. No items were exposed over 0.4, which means no items were disclosed to more than 40% of the examinees. Only 2% of items were exposed over 30% but less than 40% of the examinees. Around 2.5% and 8.2% of the testing were forced to terminate and make mastery decisions due to test length constraints and the time constraints accordingly.

In this study, items were selected based on either Information (I) index or Information-Time (IT) index. The information indexes were computed based on either Fisher or Kullback-Leibler (K-L). So there were four combinations for item selection method totally: (1) FI is I based on Fisher information; (2) K-LI is I based on K-L information; (3) FIT is IT based on Fisher information; and (4) K-LIT is IT based on K-L information. It could be seen that the results of FI and K-LI; and that of FIT and K-LIT were almost identical. That means that information calibrations based on Fisher and Kullback-Leibler were very similar. For that, the discussion about item selection method only focuses on Information index and Information-Time index.

Information and Information-Time Index

Based on type I and II error rates, The average classification accuracy of applying item selection based on I index and IT index were identical.

Within every condition, item consumption was consistently less when applying I index and time consumption was consistently less when applying IT index. In average, applying I index reduced about 12% of item consumption while applying IT index reduced about 14% of time consumption.

Besides, item selection based on I index tends to use more polytomous items while on IT index tends to use more dichotomous items within all conditions. For I index, about 23% of dichotomous items, and 77% of polytomous items were used correspondingly. For IT index, about 54% of dichotomous items, and 46% of polytomous items were used respectively.

For the item utilization rates, item selection based on IT index performed slightly better than I index: about 5% more items in the pool were utilized. Force-to-termination rates of I index and IT index were similar.

Randomesque Method

Randomesque method (Kingsbury & Zara, 1989) was adapted for item exposure rate control in this study. Two operations, 3-pick-1 and 5-pick-1 were adopted. Both operations performed satisfactorily. For classification precision and test efficiency, 3-pick-1 performed better. The average type I, II error rates were 0.019 and 0.021 for 3-pick-1, and .022 and 0.023 for 5-pick-1. However, for the item exposure rate control, 5-pick-1 procedure performed better. No items were exposed over 30% and only 3.1% of items were exposed over 20% for the 5-pick-1 operation. Within 3-pick-1or 5-pick-1 operation, the main effects of I index and IT index were the same. No interaction effects between item selection method and Randomesque method were found.

Test Difficulty

Two cutting thetas, -1.0 and 1.0 were adopted to set up the degree of test difficulty. Test difficulty was again found influencing the classification accuracy and test efficiency in this study. As the cutting level increased, the total error and item utilization rate decreased. The average type I, type II error rate, item consumption rate, time used rate (in minutes) were 0.020, 0.028, 17.297, and 29.822 respectively for the cutting theta = -1.0 and 0.020, 0.015, 11.923, and

20.401 for the cutting theta = 1.0. The average number of item and time used for the cutting theta = -1.0 was 45% and 46% more than that of the cutting theta = 1.0.

These results were reasonable because the average item difficulty of the mixed item pool was 0.714. In theory, if the average item difficulty in the pool matches the cutting theta, better classification accuracy and test efficiency could be achieved. Within cutting theta = -1.0 or cutting theta = 1.0 condition, the main effects of I index and IT index were found the same. No interaction effects between item selection method and test difficulty were found.

Test Length and Time Constraints

Four combinations of test length and time constraints were set up: (1) minimum=10 & maximum=50 and 60 minutes, (2) minimum=10 & maximum=50 and 120 minutes, (3) minimum=1 & maximum=100 and 60 minutes, and (4) minimum=1 & maximum=100 and 120 minutes. Combinations 1 and 4 represented the most restricted and the most relaxed constraint respectively.

In terms of classification accuracy, combination 4, the most relaxed constraint, performed the best and the other 3 combinations were similar. In terms of efficiency (item & time consumption), combination 3 was the best. For the proportion of using dichotomous and polytomous items, all four combinations showed the similar patterns. Besides, different causes to force-to-termination were found: in combination 2, force-to-termination was only due to test length constraint; and in other combinations (1, 3, & 4), force-to-termination was only due to time constraint.

Based on that, it could be concluded that different combinations of test length and time constraints could affect classification accuracy, test efficiency, and the cause to force-to-termination. No interaction effects between item selection method and test length and time constraints were found.

Discussion

In this study, the Information-Time index was found feasible and effective in computerized classification testing with mixed item pool. In CCT, item selection is always based on item information, plus some item exposure control mechanism because it guarantees the classification accuracy and test efficiency. In fact, it works well with dichotomous item pool. Lau and Wang (1998) successfully extent CCT with polytomous item pool and mixed item pool and found that

item selection based on information would cause more polytomous items were chosen than dichotomous items with a mixed item pool. It is because polytomous items usually have better information than dichotomous items in average. Information-Time index offers an alternative for item selection: both item information and testing-time consumption could be considered simultaneously. It suggests another aspect of test efficiency: testing-time consumption in addition to item consumption alone when administering a test.

In this study, both item selection algorithms, I index and IT index were found beneficial to test efficiency but in different aspects: item selection based on Information index (FI or K-LI) could reduce 12% item consumption; and item selection based on item Information-Time index (FIT or K-LIT) could enhance 14% time consumption. Besides, applying different item selection algorithms affected the type of items chosen: when item selection was based on I index, the ratio of dichotomous items and polytomous item consumption was about 23:77; when it was based on IT index, the proportion was about 54:46. It means that applying IT index for item selection could balance the type of items chosen in a mixed item pool. With IT index, dichotomous items could have equal opportunity to be chosen as polytomous items.

Information-Time index for item selection worked well with other testing variables like item exposure control, test difficulty, and test constraints. Within each variable, Information-Time index achieved almost identical classification accuracy as Information index; and consistently reduced test time consumption.

This new index might also provide a better base for setting time constraint in CCT. In addition to test length constraint, time constraint was usually adopted in testing situation. In this study, combination 4 gained the best classification accuracy and combination 3 achieved the best test efficiency. Depended on the purpose, a comparatively better result can be obtained by arranging different combinations of time and test length constraints. It means the test users can have more power to control the testing condition.

Two things were confirmed again in this study: (1) information calibration based on Fisher and Kullback-Leibler produced very similar results, consistent with previous studies (e.g., Eggen, 1998; Lau & Wang, 1999); and (2) sequential probability ratio testing was again found to be a robust and user-friendly procedure in computerized classification testing (Spray & Reckase, 1996, Lau, 1996, Lau & Wang, 1998, 1999). Even about 11% of the testing was forced to cease

and make classification decision, reasonable classification accuracy and testing efficient were still achieved in different conditions and combinations of conditions in this study.

According to this study, it is concluded that Information-Time index could provide more flexibility in CCT, especially with a mixed item pool. Depending on the practical testing situation, the test users could choose item efficiency or time efficiency, and which item type should be used more by using Information-Time index.

References

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*, 162-176.

Eggen, T. J. H. M. (1998). *Item selection in adaptive testing with the sequential probability ratio test*. Measurement and Research Department Report, 98-1. Arnhem: Cito.

Ercikan, K., Burket, G., Julian, M., Link, V., Schwarz, R., & Weber, M. (1996). *Calibration and scoring of tests with multiple-choice and constructed response item types*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

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 American Educational Research Association, San Diego.

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.

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

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.

Luecht, R. M. (1998). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the annual meeting of the American Educational Research Association, San Diego.

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

O'Sullivan, C. Y., Reese, C. M., & Mazzeo. J. (1997). NAEP 1996 science report card for the nation and the states, Washington, DC: National Center for Educational Statistics.

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.

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

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

Spray, J., Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test* (ACT Research Report Series 87-1). Iowa City, IA: American College Testing.

Stocking, M. L & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analyses of two tests. *Journal of Educational Measurement, 31*, 113-123.

Wald, A. (1947). *Sequential Analysis*. New York: Dover Publications, Inc.

Wilson, M., & Wang, W. (1995). Complex composites: Issues that arise in combining different modes of assessment. *Applied Psychological Measurement, 19*, 51-71.

Table 1. Average Error Rates, Test Length, Dichotomous & Polytomous Item Type Consumption, and Time Used of the Independent Variables

Independent Variable	Type I Error	Type II Error	Test Length	Time Used	DI Item	PO Item
Main Effect						
Information (I) index	0.020	0.022	13.679	27.058	0.228	0.772
Information/Time (IT) index	0.020	0.022	15.541	23.166	0.538	0.463
Item Selection						
I index						
FI	0.020	0.022	13.706	26.992	0.236	0.764
K-LI	0.020	0.022	13.652	27.124	0.220	0.780
IT index						
FIT	0.020	0.022	15.650	23.167	0.549	0.451
K-LIT	0.020	0.022	15.432	23.165	0.526	0.474
Randomesque						
3-pick-1						
I index	0.019	0.021	12.799	26.156	0.206	0.794
IT index	0.019	0.021	14.794	21.689	0.556	0.444
5-pick-1						
I index	0.022	0.023	14.559	27.960	0.250	0.750
IT index	0.021	0.023	16.288	24.643	0.520	0.480
Cutting Theta						
θ_c = -1.0						
I index	0.020	0.029	16.097	31.925	0.212	0.788
IT index	0.020	0.028	18.496	27.720	0.511	0.489
θ_c = 1.0						
I index	0.020	0.015	11.261	22.190	0.244	0.756
IT index	0.020	0.015	12.586	18.613	0.564	0.436
Length & Time Constraint						
(1) L: 10,50 T: 60 min						
I index	0.021	0.024	14.177	28.367	0.214	0.786
IT index	0.021	0.023	15.855	23.411	0.543	0.457
(2) L: 10,50 T: 120 min						
I index	0.020	0.021	16.098	31.637	0.231	0.769
IT index	0.020	0.022	16.952	25.230	0.535	0.465
(3) L: 1,100 T: 60 min						
I index	0.021	0.024	10.470	21.206	0.213	0.787
IT index	0.020	0.023	13.063	19.317	0.549	0.451
(4) L: 1,100 T: 120 min						
I index	0.019	0.020	13.983	27.020	0.254	0.746
IT index	0.019	0.019	16.295	24.706	0.523	0.477

Note: FI & K-LI are the Fisher and Kullback-Leibler information respectively. FIT & K-LIT are the Information-Time index, in which the information was calibrated based on Fisher and K-L respectively. Randomesque is the item exposure control method. DI & PO items are the proportions of using dichotomous and polytomous items. Time Used was calculated in minutes. In Length & Time Constraint, L indicates the minimum & maximum test length by the first & second figures; T shows the time constraint in minute.

Table 2. Force-to-Termination Rates and Item Exposure Rates of the Independent Variables

			Item Exposure Rate					
Independent Variable	**LCon**	**TCon**	**r=0**	**0<r<.1**	**.1≤r<.2**	**.2≤r<.3**	**.3≤r<.4**	**.4≤r**
Main Effect								
Information (I) index	0.024	0.084	0.607	0.322	0.036	0.018	0.019	0.000
Information/Time (IT) index	0.026	0.080	0.558	0.359	0.044	0.020	0.020	0.000
Item Selection								
I index								
FI	0.024	0.084	0.605	0.323	0.036	0.017	0.019	0.000
K-LI	0.024	0.084	0.608	0.320	0.035	0.019	0.019	0.000
IT index								
FIT	0.026	0.080	0.558	0.357	0.045	0.020	0.020	0.000
K-LIT	0.025	0.080	0.557	0.360	0.043	0.020	0.020	0.000
Randomesque								
3-pick-1								
I index	0.020	0.072	0.716	0.220	0.020	0.006	0.038	0.000
IT index	0.022	0.066	0.665	0.259	0.028	0.008	0.040	0.000
5-pick-1								
I index	0.028	0.097	0.497	0.423	0.051	0.030	0.000	0.000
IT index	0.030	0.093	0.451	0.458	0.060	0.032	0.000	0.000
Cutting Theta								
θ_c = -1.0								
I index	0.033	0.118	0.606	0.310	0.045	0.019	0.021	0.000
IT index	0.036	0.112	0.560	0.337	0.060	0.021	0.022	0.000
θ_c = 1.0								
I index	0.014	0.051	0.606	0.333	0.026	0.017	0.018	0.000
IT index	0.015	0.048	0.555	0.380	0.029	0.019	0.018	0.000
Length & Time Constraint								
(1) L: 10,50 T: 60 min								
I index	0.000	0.131	0.709	0.207	0.031	0.024	0.030	0.000
IT index	0.000	0.123	0.639	0.271	0.037	0.025	0.030	0.000
(2) L: 10,50 T: 120 min								
I index	0.095	0.000	0.610	0.307	0.029	0.026	0.030	0.000
IT index	0.102	0.000	0.610	0.299	0.039	0.023	0.030	0.000
(3) L: 1,100 T: 60 min								
I index	0.000	0.128	0.708	0.231	0.042	0.011	0.009	0.000
IT index	0.000	0.120	0.640	0.284	0.050	0.016	0.011	0.000
(4) L: 1,100 T: 120 min								
I index	0.000	0.078	0.399	0.541	0.041	0.011	0.009	0.000
IT index	0.000	0.076	0.343	0.581	0.050	0.016	0.011	0.000

Note: LCon and TCon are the percentages of the testing being forced to stop and make classification decisions because of the test length constraint and time constraint respectively. r=0 is the percentage of unused items.