A State Exchanging Method for Parallel Test Forms

Koun-Tem Sun

Institute of Computer Science and Information Education
National Tainan Teachers College, Tainan, Taiwan, ROC
e-mail: ktsun@ipx.ntntc.edu.tw

Abstract

The construction of parallel test forms (i.e., tests with similar difficulty) is a very important task in educational measurement. In the past two decades, a variety of methods based on the item response theory (IRT) have been proposed to construct parallel test forms. The previously proposed methods can efficiently select items to construct a test that approximates the test information function of a target test. In this paper, we propose a more effective method based on the state exchange technique that can greatly reduce the error of the test information functions of parallel tests generated by other methods. Experimental results show that this method sharply reduces error with improvement ratios exceeding 96.9%. In addition, the computation complexity of our method is the same as that of other methods. This method should greatly aid in the construction of parallel test forms.

Keywords: *item response theory (IRT), parallel tests, test information function, improvement ratio, computation complex.*

1. Introduction

The design of parallel test forms is a very important task in educational measurement [14]. *Item response theory* (IRT) [11, 12, 24, 9] has served as the basis for test design for a variety of measurement purposes [21, 4, 1, 8]. However, the test construction problem (or item selection problem) can be formulated as a zero-one combinatorial optimization problem and therefore processing time increases exponentially with the number of items in the item bank. For this reasons, many mathematical programming methods and heuristic methods have been developed to facilitate test design. These methods commonly involve the construction of parallel test forms, in which the test information function varies as little as possible between forms. The greater the value of information of a test, the more precision the ability estimation. For example, if a scholarship test for an academic award is required to be constructed (see Figure 1), then items with greater information at the high ability levels would be selected from the item pool in order to screen students with high abilities. The test information function can be computed by calculating the sum of the item information function $I_i(\theta)$ for the items included on the test [5]:

$$I(\theta) = \sum_{i=1}^{m} I_i(\theta), \tag{1}$$

where m is the number of items on the test. Table 1 shows the information of items in the item bank.

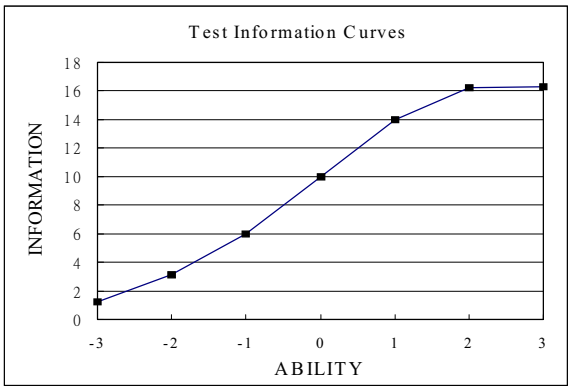

Figure 1. The curve of information for a scholarship test.

Building on concepts developed by Birnbaum [5], Lord [13] outlined the following procedure for test construction using the item information function:

1. Define the shape of the desired test information function (also called the target information function).
2. Select items with item information functions that will fill up the hard-to-fill areas under the target information function.
3. After each item is added to the test, calculate the test information function for all selected test items.
4. Continue selecting test items until the test information function approximates the target information function to a satisfactory degree.

Table 1. Item information function of the first 10 items of 320 items in the designed item bank.

Item	Ability Level						
	-3	-2	-1	0	1	2	3
1	0.242	0.193	0.078	0.024	0.007	0.002	0.001
2	0.024	0.052	0.083	0.093	0.075	0.049	0.027
3	0.000	0.003	0.035	0.171	0.250	0.130	0.041
4	0.000	0.002	0.059	0.466	0.293	0.052	0.007
5	0.031	0.191	0.364	0.215	0.068	0.017	0.004
6	0.001	0.012	0.075	0.199	0.192	0.090	0.030
7	0.000	0.001	0.010	0.062	0.169	0.164	0.076
8	0.062	0.147	0.164	0.101	0.044	0.017	0.006
9	0.000	0.000	0.000	0.000	0.001	0.102	0.715
10	0.006	0.024	0.069	0.117	0.114	0.072	0.035

The less the deviation there is between the target test information function and the constructed test information function, the more satisfactory the test is. Therefore, a test designer selects items which allow the information function of the constructed test to most closely approach the target test information function. Since the item selection problem is a combinatory optimization problem, the number of combinations increases exponentially with the number of items. For this reason, the designers must use the weak methods (heuristic algorithms) which are capable only of finding "good" solutions but not "optimal" solutions. For example, linear programming (LP) techniques are the most commonly used for the test construction [6, 7, 4, 22, 20, 23]. In linear programming techniques, items are selected in order to optimize objectives within the given constraints. Good solutions can be produced by a variety of heuristic methods, such as the branch-and-bound method [2], the revised simplex method which use the relaxed 0-1 linear programming model of Adema [3], the weighted deviation model [20], the neural network technique [18], and the greedy approach [19]. Test construction problems commonly involve a list of objective functions with various purposes [22], but the test information function is the common objective of all test design problems. Therefore, in this paper, we will only consider how to select items in order to meet the requirements of the test information function. The difficulty of this problem, however, is not reduced by eliminating consideration of the content attributes. Here, we propose a very effective method for item selection based on the discrete updating method (DUM) of Sun [15]. The DUM method has successfully solved many optimization problems by using binary state variables to represent the state of problem and exchange them between two sets [16, 17]. Now, the item selection problem is solved by the state exchange approach. In order to evaluate the performance of this method, two hundred test information functions for parallel test forms of two types (one-peak and two-peak functions) were randomly generated. Results show that the new method greatly reduces the error of test information functions generated by other methods with improvement ratios exceeding 96.9% (i.e., the errors of test information functions are reduced more than one order of magnitude). The new method produces tests in which the test information functions very closely approach the target test information functions. In addition, the computation complexity [10] of our method does not exceed that of other methods. In other words, the new method can be incorporated into other methods to produce better results without increasing the computation complexity. Because this method is such an effective tool for test construction, it should prove very useful in educational measurement.

2. Test Construction by the State Exchange Process

The proposed approach is based on the discrete updating method (DUM) [15] which exchanges the states of items (an item's inclusion on or exclusion from a test) to reduce the value of the energy function. For implementing the exchange process, the error between the information functions of the target test and the constructed test is represented by an energy function. When a test is constructed, the information function $O(\theta_j)$ for the constructed test can be determined. Then, the error between this function and the target test information function is squared and a sum is taken, as in Equation (2):

$$E_I = \sum_{j=1}^{s} (d_j - O_j)^2 \cdot \quad (2)$$

When an item p is removed from the test and an item q is added to the test, the energy of function E_I may either increase or decrease. The change in E_I is calculated as follows:

$$\Delta E_{I,pq} = E_{I,pq} - E_I \ , \quad (3)$$

where

$$E_{I,pq} = \sum_{j=1}^{s} (d_j - O_{j,pq})^2 , \quad (4)$$

and

$$O_{j,pq} = \sum_{k=1}^{n} w_{kj} x_k - w_{pj} + w_{qj} = O_j - w_{pj} + w_{qj} . \quad (5)$$

After exchanging the state of the item p with item q, the updated value of the energy function becomes

$$\Delta E_{I,pq} = \sum_{j=1}^{s}(d_j - O_{j,pq})^2 - \sum_{j=1}^{s}(d_j - O_j)^2$$
$$= \sum_{j=1}^{s}(d_j^2 - 2d_jO_{j,pq} + O_{j,pq}^2 - d_j^2 + 2d_jO_j - O_j^2)$$
$$= \sum_{j=1}^{s}(2d_jw_{pj} - 2d_jw_{qj} - 2O_jw_{pj} + 2O_jw_{qj} - 2w_{pj}w_{qj} + w_{pj}^2 + w_{qj}^2) \quad (6)$$
$$= \sum_{j=1}^{s}2d_j(w_{pj} - w_{qj}) - 2O_j(w_{pj} - w_{qj}) + (w_{pj} - w_{qj})^2$$
$$= \sum_{j=1}^{s}(w_{pj} - w_{qj})(2d_j - 2O_j + w_{pj} - w_{qj})$$

After exchanging the states of item p and item q, if the updated value of energy function E_I, $\Delta E_{I,pq}$, is less than zero, then the new test information function more closely approximates the desired information function. If this condition is true, we remove the item p from the test, and added item q to the test. The exchange approach is designed to select items with the greatest information to fill the gap between the information function of the test under construction and that of the target test, and it can effectively construct a test which closely approximates the target test information function. The detailed operations of the proposed approach are described as follows.

The Exchange Approach to Item Selection

In the proposed exchange approach, each item x_p on the test is evaluated to see the effect of its exchange with one of the items in the item bank. For each item x_q from the item bank which is selected for exchange with x_p, the updated value of $E_{I,pq}$, $1 \le q \le$n, can be determined in parallel. However, only one item x_p on the test can be exchanged with one item, x_q, in the bank in each exchange. If more than one item in the test or bank being exchanged, then the energy function E_I may be over-updated, causing the error to increase rather than decrease. In this case, the energy function will not converge to a stable state. So, the states of only one pair of items, x_q and x_p, are updated in each iteration, and the energy function E_I is reduced monotonically until it reaches a minimum state. Figure 2 shows the exchange process for evaluating the exchange of item 1 on the test with one item in the item bank.

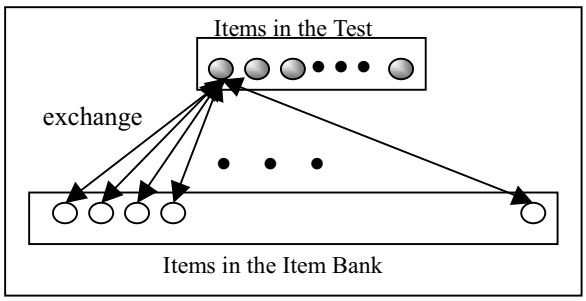

Figure 2. The exchange process for evaluating the replacement of the first item on the test by one of the items in the item bank. The exchange process proceeds in the same way for the other items on the test.

The detailed operations for this approach are stated as follows.

1. Set the initial values of all variables.
   Each variable d_j, j=1 ~ s, is set to match the target test information function specified by the test designer; the state of item i is given by x_i, i =1 ~ n, and initially is set to zero (i.e., initially, no items are included on the test). The value of w_{ij}, i = 1~ n and j = 1 ~ s, is equal to the amount of information for item i at ability level j. The initial time, t, (the iteration index) is set to zero.
2. Randomly construct a test with m items as the initial test and determine the information function $O_j(t)$ of the test at each ability level j.
$$O_j(t) = \sum_{i=1}^{n} w_{ij}x_i(t), \ \forall j = 1 \sim s. \quad (7)$$
3. For each item p on the test, complete the following item exchange process.
   3.1. Remove item p from the test and replace it with one item q from the item bank, until all items in the item bank have been evaluated. In each case, calculate the updated value of the energy function, E_I, as follows:
$$\Delta E_{I,pq}(t) == \sum_{j=1}^{s}(w_{pj} - w_{qj})(2d_j - 2O_j(t) + w_{pj} - w_{qj}), \quad (8)$$
   $1 \le q \le n$.
   3.2. Find the smallest negative value $\Delta E_{I,pq*}(t)$ from Equation (16) for all q where $x_q(t)$ is not equal to one (i.e., for those items not included on the test). Then,
$$\Delta E_{I,pq*}(t) = Min\{\Delta E_{I,pq}(t), \forall\, \Delta E_{I,pq}(t) < 0 \text{ and } x_q(t) \neq 1, q = 1 \sim n\} \quad (9)$$
   3.3. If the value $\Delta E_{I,pq*}(t)$ is less than zero, exchange the states of x_p and x_{q^*}
$$x_p(t) = 0, \text{and } x_{q^*}(t) = 1, \text{ when } \Delta E_{I,pq^*}(t) < 0. \quad (10)$$
   3.4. Compute the new value of the test information function $O_j(t+1)$.
$$O_j(t+1) = \sum_{k=1}^{n} w_{kj}x_k - w_{pj} + w_{qj} = O_j(t) - w_{pj} + w_{qj} \quad (11)$$
   3.5. Go to Step 3.1 to perform the exchange process on the next item in the test, and increase the time index by one.
$$t \leftarrow t+1. \quad (12)$$
4. Stop.
   At the end of computation, variables with x_is = 1 are selected for inclusion on the final version of the test.

In the proposed item exchange approach, only one test item x_p at a time is taken through the exchange process (Steps 3.1 to 3.5). Therefore, the maximum

number of iterations for a test with m items is m, and the magnitude of computations [10] for each iteration is O(n) for computing n exchanging operations on $\Delta E_{I,pq*}(t)$. Thus, the total computation complexity of the proposed method is O(mn) = $m \times$ O(n), which is the same as that of other methods [18]. However, the results of our method are much better than that of other methods, as will be discussed in the next section.

3. Performance Evaluation

We used a real item bank (see Table 1) based on the three-parameter model of IRT to compare the performance of our method with that of other methods. The amount of information on the target test varied within the ranges shown in Table 2 (for a one-peak shape information curve) and Table 3 (for a two-peak shape information curve). Following the limitations of the information quantities defined in these two tables, one hundred target test information functions were randomly generated for each of the two shapes. The initial test for Step 2 of the proposed item exchange approach was generated by other methods, and then Steps 3 and 4 were executed in order to improve the results. Figures 3 through 10 show the effect of applying the exchange process to test construction, using initial tests generated by different methods (greedy [19], neural network [18], Swanson & Stocking [20], and Wang & Ackerman [23] methods) for both one-peak and two-peak test cases. The average sum of the squared error between the information function of the target tests and that of the constructed tests is shown in Table 4. We see that the proposed exchange approach greatly reduces error, with improvement ratios greater than 96.9%.

Table 2. The ranges of test information used to randomly generate 100 target test information functions (one-peak shape).

	Index of Ability Level				
	1	2	3	4	5
Ability Level	-2.0	-1.0	0.0	1.0	2.0
Test Information	4 ~ 5	6 ~ 8	18 ~ 21	6 ~ 8	4 ~ 5

The initial test used in the exchange process can be also randomly constructed. After five or six runs of the program (the result of the first run becomes the initial test of the second run, the result of the second run becomes the initial test of the third run, and so on), the error of the energy function reaches a stable state. The final results (Table 5) closely approximate those obtained in Table 4.

Table 3. The ranges of test information used to randomly generate 100 target test information functions (two-peak shape).

	Index of Ability Level				
	1	2	3	4	5
Ability Level	-2.0	-1.0	0.0	1.0	2.0
Test Information	5 ~ 6	11 ~ 13	7 ~ 9	11 ~ 13	5 ~ 6

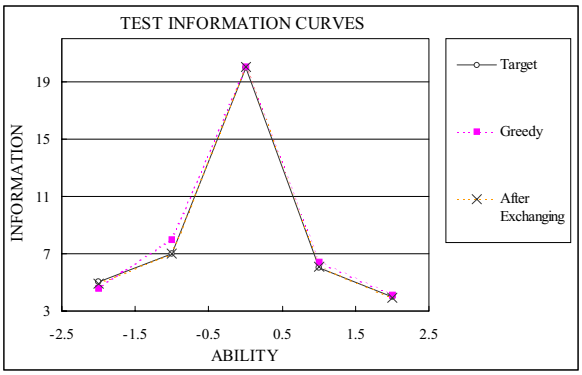

Figure 3. Test information curves (one-peak shape) for a target test, a test produced by the greedy approach, and the same test improved by the exchange approach.

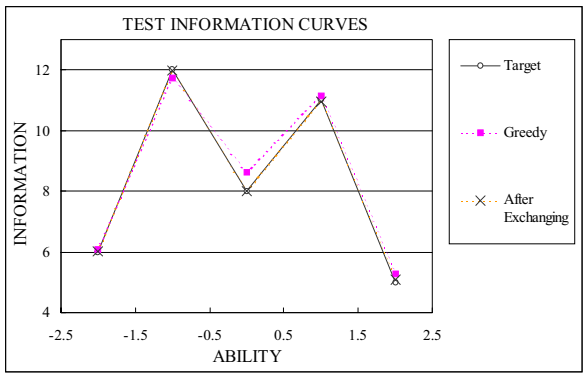

Figure 4. Test information curves (two-peak shape) for a target test, a test produced by the greedy approach, and the same test improved by the exchange approach.

We see that the proposed item selection method is a very flexible and effective tool which can be applied to a variety of initial tests with excellent results. It should prove be very useful to test designers who are constructing parallel test forms or desired tests for a variety of assessment purposes.

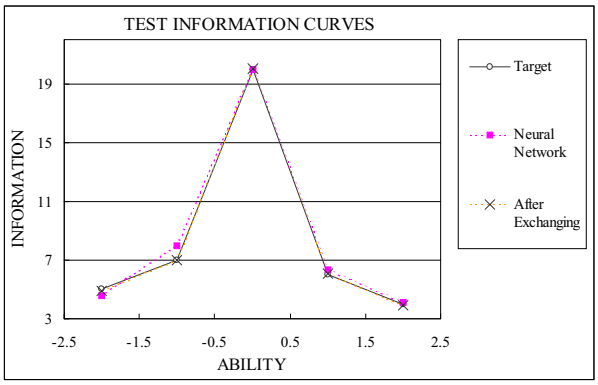

Figure 5. Test information curves (one-peak shape) for a target test, a test produced by the neural network approach, and the same test improved by the exchange approach.

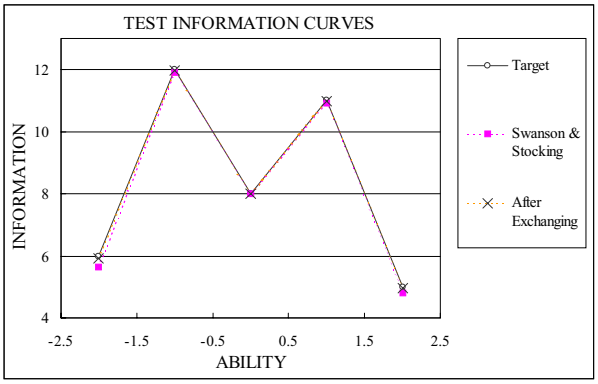

Figure 8. Test information curves (two-peak shape) for a target test, a test produced by Swanson & Stocking's method, and the same test improved by the exchange approach.

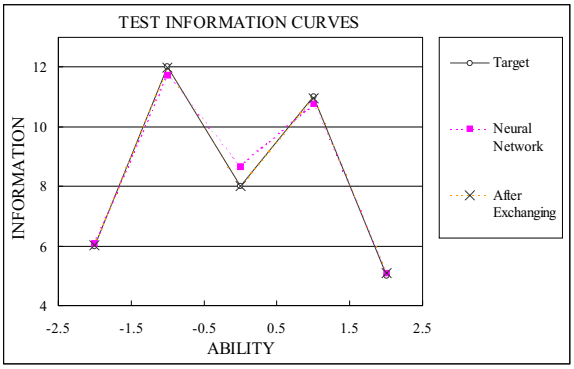

Figure 6. Test information curves (two-peak shape) for a target test, a test produced by the neural network approach, and the same test improved by the exchange approach.

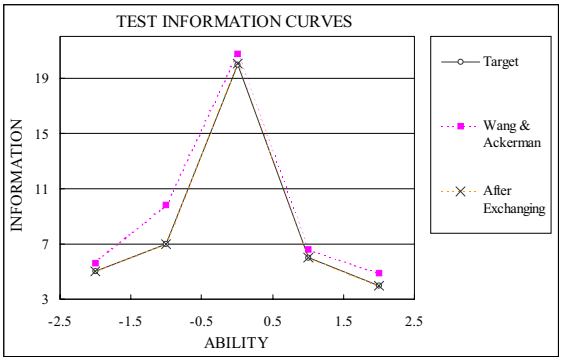

Figure 9. Test information curves (one-peak shape) for a target test, a test produced by Wang & Ackerman's method, and the same test improved by the exchange approach.

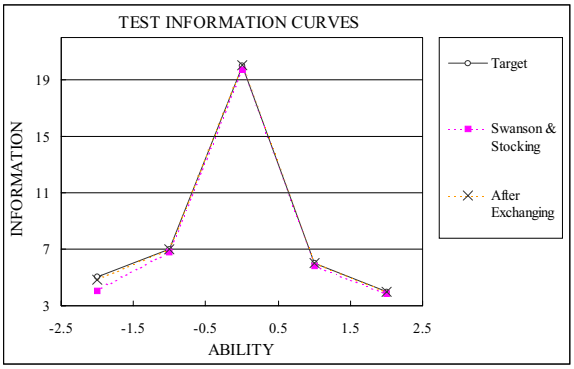

Figure 7. Test information curves (one-peak shape) for a target test, a test produced by Swanson & Stocking's method, and the same test improved by the exchange approach.

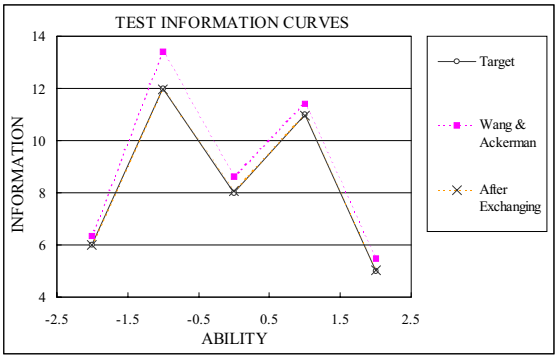

Figure 10. Test information curves (two-peak shape) for a target test, a test produced by Wang & Ackerman's method, and the same test improved by the exchange approach.

Table 4. The average sum of squared error, before and after applying the exchange process, between the information function of the target test and the information function of the constructed test. The improvement ratios of the average error are also shown.

Conditions \ Errors \ Methods		Greedy Approach	Neural Network	Swanson & Stocking	Wang & Ackerman
Before applying the exchange process ($error_B$)	Table 2	0.6986	0.6945	0.9715	12.8340
	Table 3	0.7251	0.7416	0.2755	2.9536
	Average Errors	0.7119	0.7181	0.6245	7.8938
After applying the exchange process ($error_A$)	Table 2	1.9293 E-2	2.0046E-2	2.6717E-2	1.1610E-2
	Table 3	6.1564E-3	6.2491E-3	1.1781E-2	1.1421E-2
	Average Errors	1.2725E-2	1.3148E-2	1.9249E-2	1.3761E-2
Improvement Ratio* (%)		98.2125	98.1228	96.9177	99.8257

*Improvement Ratio (%) = $(error_B - error_A) / error_B \times 100$
$error_B$: the error generated before applying the exchange process
$error_A$: the error generated after applying the exchange process

Table 5. The average sum of the squared error between the information function of a target test and that of a test constructed from a randomly produced initial test and subsequently improved by the exchange process.

	Number of runs by applying the exchange process					
	1	2	3	4	5	6
Table 2	0.7779	8.7624E-2	5.8080E-2	5.1869E-2	4.8380E-2	4.7904E-2 (stable)
Table 3	1.6444E-2	6.9345E-3	6.3417E-3	6.2951E-3	6.2148E-3 (stable)	6.2148E-3

4. Conclusions

In this paper, an effective method, based on the concept of exchanging the state of items between two sets, is proposed to construct a desired test from an item bank. The proposed method can effectively construct parallel test forms or a test whose test information function closely approximates that of a target test. A real item pool was used to evaluate the performance of our method. The experimental results show that the proposed approach is able to reduce the error generated by other methods more than 96.9%. In addition, the computational complexity of our method is O(*mn*), equivalent to that of other methods. In other words, the proposed method significantly reduces the error between the test information functions of parallel tests while maintaining the efficiency of computation time.

Acknowledgments

This research was supported by the National Science Council of Taiwan, ROC, under the grant NSC 89-2520-S-024-001-.

References

[1] Ackerman, T., *An alternative methodology for creating parallel test forms using the IRT information function*, The Annual Meeting of the National Council for Measurement in Education, San Francisco, 1989.

[2] Adema, J. J., *Implementations of the branch-and-bound method for test construction problems*, Research Report 89-6. Enschede: Department of Education, University of Twente, The Netherlands, 1989.

[3] Adema, J. J., *A revised simplex method for test construction problems*, Research Report 90-5. Enschede: Department of Education, University of Twente, The Netherlands, 1990.

[4] Baker, F. B., Cohen, A. S., & Barmish, B. R., "Item characteristics of tests constructed by linear programming," *Applied Psychological Measurement*, vol. 12, pp. 189-199, 1988.

[5] Birnbaum, A., Some latent trait models and their use in inferring an examinee's ability, In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley, 1968.

[6] Boekkooi-Timminga, E., "Simultaneous test construction by zero-one programming," *Methodika*, vol. 1, pp. 101-112, 1987.

[7] Boekkooi-Timminga, E., *Models for computerized test construction*. The Netherlands: Academisch Boeken Centrum, 1989.

[8] De Gruijter, D. N. M., "Test construction by means of linear programming," *Applied Psychological Measurement*, vol. 14, pp. 175-181, 1990.

[9] Hambleton, R. K., & Swaminathan, H., *Item Response Theory-- Principles and Applications*, Netherlands: Kluwer Academic Publishers Group, 1985.

[10] Horowitz, E., Sahni, S., and Rajasekaran, S., *Computer Algorithms--Pseudocode*, New York: Computer Science Press, Inc., 1997.

[11] Lord, F. M., "A theory of test scores," *Psychometric Monograph*, vol. 7, 1952.

[12] Lord, F. M., & Novick, M. R., *Statistical theories of mental test scores*, Reading, MA: Addison-Wesley, 1968.

[13] Lord, F. M., "Practical applications of item characteristic curve theory," *Journal of Educational Measurement*, vol. 14, pp. 117-138, 1977.

[14] Lord, F. M., *Applications of item response theory to practical testing problems*, Hillsdale, NJ: Erlbaum, 1980.

[15] Sun, K. T., *A study of optimization problems by neural networks*, Unpublished doctoral dissertation, National Chiao Tung University, Taiwan, ROC, 1992.

[16] Sun, K. T., & Fu, H. C., "A neural network implementation for traffic control problem on crossbar switch networks," *International Journal Neural Systems*, vol. 3, no. 2, pp. 209-218, 1992.

[17] Sun, K. T., & Fu, H. C., "A neural network approach to the traffic control problem on reverse baseline networks," *Circuits, Systems and Signal Processing*, vol. 12, no. 2, pp. 247-261, 1993.

[18] Sun, K. T. & Chen, S. F., "A study of applying the artificial intelligent technique to select test items," *Psychological Testing*, vol. 46, no. 1, pp. 75-88, 1999.

[19] Sun, K. T., "A greedy approach to test construction problems," Submitted to *The Proceedings of the National Science Council (Part D): Mathematics, Science, and Technology Education, 2000.*

[20] Swanson, L., & Stocking, M. L., "A model and heuristic for solving very large item selection problems," *Applied Psychological Measurement*, vol. 17, no. 2, pp. 151-166, 1993.

[21] Theunissen, T. J. J. M., "Binary programming and test design," *Psychometrika*, vol. 50, pp. 411-420, 1985.

[22] Van der Linden, W. J., & Boekkooi-Timminga, E., "A maximum model for test design with practical constraints," *Psychometrika*, vol. 54, pp. 237-247, 1989.

[23] Wang, C. S., & Ackerman, T., "Two item selection algorithms for creating weakly parallel test forms using the IRT information functions," *Psychological Testing*, vol. 44, no. 2, pp. 123-140, 1997.

[24] Weiss, D. J., "Improving measurement quality and efficiency with adaptive testing," *Applied Psychological Measurement*, vol. 6, pp. 379-396, 1982.