Computerized Test Construction

Bernard P Veldkamp, University of Twente, Twente, The Netherlands

This article is a revision of the previous edition article by W.J. van der Linden, volume 4, pp. 2477–2480, © 2001, Elsevier Ltd.

Abstract

This article starts with a brief introduction to the area of computerized test construction. In the first section, the problem is formally stated as an optimization problem, with an objective function, for example, that maximizes the amount of information in the test or minimizes the amount of items in a test. In the second section, an overview is given of the methods that are available for computerized test construction. In sections computerized construction of linear test forms, construction of computerized adaptive tests, and construction of multistage tests, some specific issues related to the three most important test forms are discussed. The article ends with a discussion on one of the recent issues in computerized test construction, the problem of uncertainty in some of the parameters during optimization.

Test Construction as a Combinatorial Optimization Problem

Combinatorial optimization problems are those where mathematical techniques are applied to find optimal solutions within a finite set of possible solutions. The set of possible solutions is generally defined by a set of restrictions, and the set is too large for exhaustive search. A well-known example is the knapsack problem, where the value of the goods carried in the knapsack has to be maximized, while the weight of the goods that can be carried is limited. A second example is the traveling salesman problem, where the total traveling distance has to be minimized while each client is visited exactly once. The computerized test construction problem is also an example of a combinatorial optimization problem. From a finite item bank, a group of items has to be selected that is optimal with respect to the goal of testing, while the resulting test has to meet all specifications.

To formulate the test construction problem as a combinatorial optimization problem, a decision variable x_i is introduced that denotes whether item $i = 1,\ldots, I$, is selected ($x_i = 1$) or not ($x_i = 0$), an objective function has to be formulated to be optimized, and restrictions have to be identified.

Objective Functions in Computerized Test Construction

An objective function in test construction has to reflect the goal of testing that has been set in the first step of Birnbaum's approach. How to translate the objectives for test construction to psychometric properties depends on the measurement framework that is used. When the classical test theory (CTT) is applied to relate the responses of the candidate to a score representing the ability (Lord and Novick, 1968), the goals of testing are generally formulated in terms of reliability (approximated by Cronbach's alpha) or predictive validity of the test. Maximizing the reliability can be formulated in terms of the decision variable x_i as

$$\max \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{I}\sigma_i^2 x_i}{\left(\sum_{i=1}^{I}\sigma_i\rho_{iX}x_i\right)^2}\right], \qquad [1]$$

where n denotes the test length, σ_i the item variance, and ρ_{iX} the item discrimination. Maximizing the predictive validity can be formulated as

$$\max \frac{\sum_{i=1}^{I}\sigma_i\rho_{iY}x_i}{\sum_{i=1}^{I}\sigma_i\rho_{iX}x_i}, \qquad [2]$$

where ρ_{iY} is the item validity.

Within an item response theory (IRT) framework, the relationship between the responses of a candidate and the ability are modeled by an item response function, for example, the two-parameter logistic model (2PLM, Lord, 1980):

$$P_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1+e^{a_i(\theta-b_i)}}, \qquad [3]$$

where $P_i(\theta)$ denotes the probability that someone with ability θ will provide a correct answer to item i, and (a_i, b_i), also referred to as item parameters, denote the discrimination and the difficulty parameter of item i. Other examples of IRT models are the Rasch model, the three-parameter logistic model, or the normal ogive models (Lord, 1980). Within an IRT framework, the objective functions in test construction are generally formulated in terms of Fisher information. Fisher information has some favorable features when it comes to test construction. Fisher information for the whole test is equal to the sum of Fisher information of the items, and the inverse of Fisher information is asymptotically equal to the variance of the ability estimate. In other words, high information implies small uncertainty in the resulting scores. For the 2PLM, Fisher information for the whole test is given by

$$I(\theta) = \sum_{i=1}^{I}\frac{\partial P_i(\theta)/\partial\theta}{P_i(\theta)[1-P_i(\theta)]}x_i. \qquad [4]$$

 http://dx.doi.org/10.1016/B978-0-08-097086-8.43019-9

This test information function (TIF) can be maximized for a single ability value or for a range of ability values, or different targets can be set for various values of the TIF.

For an overview of objective functions, both in a CTT and in an IRT framework, see van der Linden (2005, Chapter 5).

Restrictions in Computerized Test Construction

Restrictions in combinatorial optimization refer to specifications that have to be met for groups of items that are selected. In computerized test construction, three types of specifications can be distinguished: categorical, quantitative, and logical specification. Categorical specifications refer to item attributes, such as content, item type, or format, that categorize the items in different subsets. For each of these subsets, the number of items to be included in the test can be specified. Quantitative specifications refer to attributes like response times, word counts, or statistical attributes. The sum of the quantitative attributes of all items selected for the test, for example, the total word count, is restricted by these constraints. Logical specifications deal with relationships between items. Items that belong to an enemy set cannot be selected for the same test. Therefore, only one of the items from this subset can be selected. Van der Linden (2005, Chapter 3) presents an extensive overview of various types of constraints, and of its applications to test construction. All specifications can also be formulated in terms of the decision variables. A complete combinatorial optimization model for computerized test construction is formulated in the next section.

Computerized Test Construction Model

A very general model for computerized test construction can be formulated as

$$\max \sum_{i=1}^{I} I_i(\theta_j)x_i \quad \text{(objective function)} \qquad [5]$$

subject to

$$\sum_{i \in V_c} x_i \leq b_c \quad \forall c \quad \text{(categorical constraints)} \qquad [6]$$

$$\sum_{i=1}^{I} q_i x_i \leq b_q \quad \forall q \quad \text{(quantitative constraints)} \qquad [7]$$

$$\sum_{i \in V_l} x_i \leq b_l \quad \forall l \quad \text{(logical constraints)} \qquad [8]$$

$$\sum_{i=1}^{I} x_i = n, \qquad [9]$$

$$x_i \in \{0,1\} \quad i = 1,\ldots,I \quad \text{(decision variables)}. \qquad [10]$$

Where I is the number of items in the bank, c, q, l are indices for the categorical, quantitative, and logical constraints, (b_c, b_q, b_l) denote the bounds for the constraints, and V_c and V_l denote sets of items affected by constraints c or l. The objective in this model is to maximize the TIF for ability level θ_j. This type of objective function is generally applied for mastery testing, where most information is needed around the cut-off point θ_j. Depending on the goals of testing, other objective functions can be applied.

Algorithms for Computerized Test Construction

Two classes of algorithms have been proposed for computerized test construction. First of all, mixed integer programming (MIP) solvers based on branch-and-bound algorithms can be applied. These algorithms guarantee that an optimal solution of the test construction problem is found, if one exists. If no solution to the test construction problem can be found, i.e., no test can be selected from the item bank that meets all the test specifications, these solvers can provide feedback on sets of constraints that together or by themselves cause the infeasibility problems (Huitzing et al., 2005). The commercial software package CPLEX (IBM, 2010) is one of the most powerful programs to solve test construction problems, but alternative solvers like the LpSolveAPI in the R-package are also available. Over the years, MIP solvers have proved to solve most computerized test construction problems within minutes or even seconds. Second, various heuristics for solving test construction problems have been proposed in the literature. These heuristics are often tailor-made for the test construction problem at hand. They are often easy to implement, but they approximate the optimal solution, so it cannot be guaranteed that the optimal test is constructed. Besides, finding the optimal settings of a heuristic might be rather time consuming. Some well-known heuristics for computerized test construction are the weighted deviation model (Stocking and Swanson, 1993), the normalized weighted average deviation heuristic (Luecht and Hirsch, 1992), network programming (Armstrong et al., 2005), simulated annealing (Veldkamp, 2002), genetic algorithms (Verschoor, 2007), and the MCMC (Markov Chain Monte Carlo)-based test assembly algorithm (Belov and Armstrong, 2005).

Computerized Construction of Linear Test Forms

Linear tests can be characterized as a fixed set of items presented in a fixed order. Paper-and-pencil tests are in this category, but they are also administered more and more often on a computer. The model in Eqns [5]–[10] can be applied to assemble a linear test form suitable for making mastery/nonmastery decisions around a cut-off point θ_j. If the goal of the test is to measure for a broader ability range, targets are generally formulated for Fisher information for these ability values, and the objective function for computerized test construction is to minimize the distance between the TIF and its target. Instead of minimizing the difference for all ability values, it often suffices to minimize the difference for a number of points θ_k, $k = 1,\ldots, K$, spread over the ability range of interest. This objective can be modeled as

$$\min_{k=1,\ldots,K} \left| \sum_{i=1}^{I} I_i(\theta_k)x_i - T(\theta_k) \right|, \qquad [11]$$

where $T(\theta_k)$ represents the target for Fisher information at θ_k. Boekkooi-Timminga and van der Linden (1989) developed a maximin model for this problem, in which a new variable y is introduced, which serves as an upper bound for the

absolute difference between the TIF and the actual information function. In this way the optimization problem in Eqn [10] is rewritten as an MIP problem, and standard MIP solvers or heuristics can be applied:

$$\min y \qquad [12]$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \leq T(\theta_k) + y \qquad [13]$$

$$\sum_{i=1}^{I} I_i(\theta_k)x_i \geq T(\theta_k) - y \qquad [14]$$

In many testing programs, more than one test has to be constructed. It could be that several parallel test forms have to be administered at the same time, that the test is administered periodically, or that the test is used in a quasiexperimental pre-posttest design (e.g., Cook and Campbell, 1979). To assemble multiple test forms concurrently, the test construction model in Eqns [5]–[10] is slightly modified by using decision variable x_{it} that indicate whether item i is selected ($x_{it} = 1$) for test t or not ($x_{it} = 0$), instead of decision variable x_i. When overlap between tests is prohibited or limited, the following logical constraint can be added to the model:

$$\sum_{t=1}^{T} x_{it} \leq n_i \quad i = 1, \ldots, I \qquad [15]$$

where n_i denotes the number of times item i can be selected concurrently. Since both the number of decision variables and the number of constraints grow linearly in the number of tests to be selected, 0–1 MIP algorithms might be very time consuming. Especially for the problem of concurrent construction of multiple partly overlapping tests, genetic algorithms and MCMC-based test assembly have demonstrated to be very efficient.

Construction of Computerized Adaptive Tests

Computerized adaptive tests (CATs) are individualized tests where the difficulty of the items is adapted to the ability of the candidate. In CAT, items are selected sequentially. After administration of each item, the ability of the candidate is estimated based on the responses he or she provided to the previous items. Then the item is selected and administered that maximizes Fisher information at the estimated ability level. The procedure is repeated until a stopping criterion, e.g., a fixed number of items or a minimum level of measurement precision, has been met. The first item can be selected randomly, or based on an initial guess of the ability level of the candidate. A major advantage of CAT is that test length is reduced up to 40%, and tests can be administered individually when a candidate prefers. For a thorough introduction to CAT, the reader is referred to Wainer (2000).

In case a set of specifications has to be met, the shadow test approach (van der Linden, 2005, Chapter 9) could be applied. This two-stage procedure for item selection constructs a 'shadow test' in every iteration of CAT from which the most informative item is chosen to be administered. The 'shadow test' consists of a group of most informative not-administered items which in combination with the items that have been administered already fulfill the requirements of the specifications. The following model can be applied to construct a 'shadow test' for selecting the g-th item:

$$\max \sum_{i=1}^{I} I_i(\hat{\theta}_{g-1})x_i \qquad [16]$$

subject to

$$\sum_{i=1}^{I} x_i = n \quad \text{(test length)} \qquad [17]$$

$$\sum_{i \in V_c} x_i \leq b_c \quad \forall c \quad \text{(categorical constraints)} \qquad [18]$$

$$\sum_{i=1}^{I} q_i x_i \leq b_q \quad \forall q \quad \text{(quantitative constraints)} \qquad [19]$$

$$\sum_{i \in V_l} x_i \leq b_l \quad \forall l \quad \text{(logical constraints)} \qquad [20]$$

$$\sum_{i \in S_{g-1}} x_i = g - 1 \quad \text{(previous items)} \qquad [21]$$

$$x_i \in \{0, 1\} \quad i = 1, \ldots, I \quad \text{(decision variables)}, \qquad [22]$$

where $\hat{\theta}_{g-1}$ denotes the ability to estimate after administering $(g-1)$ items and S_{g-1} denotes the set of items that has been administered so far. By application of this model, the same kind of specifications can be applied in CAT as in linear testing. An alternative procedure for dealing with test specification in CAT would be to apply the weighted deviation model (Stocking and Swanson, 1993).

In CAT, some items are selected more often than others, since they are more informative. As a result, their content might be compromised. To prevent these security issues and to increase item bank usage, exposure control methods can be applied. These methods prevent the selection of popular items either by conducting a chance experiment after selection of each item where the probabilities of selection are inversely related to the popularity (Sympson and Hetter, 1985) or by temporarily excluding them when they are administered too often (van der Linden and Veldkamp, 2004).

Construction of Multistage Tests

Multistage testing is a hybrid of linear testing and CAT. Instead of administering a single item, a small linear test, also referred to as a module (Luecht and Nungester, 1998), is administered before the ability estimate is updated. After estimating the ability, the candidate is routed to an easier or more difficult module. A multistage test design typically consists of several stages in which a module is administered. Each path a candidate might follow through the test has to meet all test specifications. An example of a multistage test design with three stages, six modules, and four paths is given in **Figure 1**.

To develop a multistage test, several decisions have to be made about the number of stages, routing rules, number of items in a module, and psychometric properties of the modules

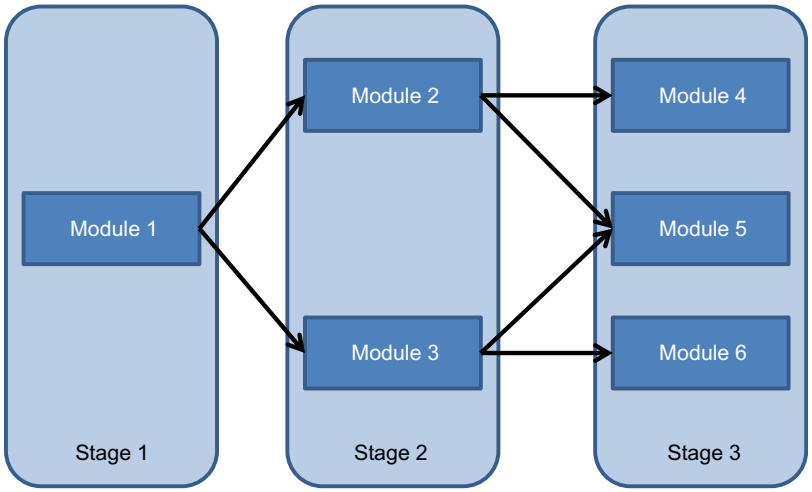

Figure 1 Multistage test design.

(Zenisky et al., 2010). The model for constructing the modules is equivalent to the model for the construction of a number of (non)overlapping small linear tests, while an additional set of specifications for each combination of tests that belong to a path has to be met:

$$\sum_{t \in P_p} \sum_{i \in V_c} x_{it} \leq b_c \quad \forall c, \forall p, \qquad [23]$$

$$\sum_{t \in P_p} \sum_{i=1}^{I} q_i x_{it} \leq b_q \quad \forall q, \forall p, \qquad [24]$$

$$\sum_{p \in P_p} \sum_{i \in V_l} x_{it} \leq b_l \quad \forall l, \forall p, \qquad [25]$$

where p is an index denoting various paths in the design. Zenisky et al. (2010) pointed out that it should be considered whether specifications should be achieved within stages or across the whole test. Meeting the constraints over the whole test provides greater flexibility, at the cost of imposing the additional constraints in Eqns [23]–[25] to the model.

Robust Computerized Test Construction

In all these test construction models, it is assumed that the item parameters and item attributes are known. Fixed values are used to compute the contribution of each item to both the objective function and the constraints. Unfortunately, many of them have been estimated. IRT parameters, for example, have been estimated when the items were pretested, and they do have uncertainty in them that is reflected by the standard error of estimation. This uncertainty is generally not taken into account. Hambleton and Jones (1994) already warned about the consequences. When the objective of test construction is to maximize the information in the test, and the uncertainty in the item parameters is not taken into account, the amount of information in the resulting test will be seriously overestimated, especially when the item parameters have been estimated with small sample sizes or when relatively few items are selected. Since the amount of information in the test is inversely related to the variance of the ability estimate, this implies that overestimation of the information in the test has serious consequences for the measurement precision and the reliability of the test.

De Jong et al. (2009) proposed a robust test construction algorithm to deal with these problems. Inspired by the work of Soyster (1973), they subtracted one time the standard of estimation from the parameter estimates, and formulated a robust test construction model as

$$\max \sum_{i=1}^{I} I_i(\theta_j, \zeta) x_i \quad \text{(objective function)} \qquad [26]$$

where ζ denotes the level of uncertainty, and

$$I_i(\theta_j, \zeta) = I_i(\theta_j) - SE(I_i(\theta_j)). \qquad [27]$$

For small values of the standard error of estimation, the results for the test construction problem in Eqns [5]–[10] and its robust counterpart are almost the same, but for large values, a considerable difference will be obtained.

See also: Classical (Psychometric) Test Theory; Psychometrics: Classical Test Theory; Psychometrics; Reliability: Measurement.

Bibliography

Armstrong, R.D., Jones, D.H., Wang, Z., 1995. Network optimization in constrained standardized test construction. In: Lawrence, K.D. (Ed.), Applications of Management Science: Network Optimization Applications, vol. 8. JAI, Greenwich, CT, pp. 189–212.

Belov, D.I., Armstrong, D.H., 2005. Monte Carlo test assembly for item pool analysis and extension. Applied Psychological Measurement 29, 239–261. http://dx.doi.org/10.1177/0146621605275413.

Birnbaum, A., 1968. Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M., Novick, M.R. (Eds.), Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA, pp. 397–479.

Boekkooi-Timminga, E., van der Linden, W.J., 1989. A maximin model for test design with practical constraints. Psychometrika 54, 237–247. http://dx.doi.org/10.1007/BF02294518.

Cook, T.D., Campbell, D.T., 1979. Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin Company, Boston.

De Jong, M.G., Steenkamp, J.-B.G.M., Veldkamp, B.P., 2009. A model for the construction of country-specific yet internationally comparable short-form marketing scales. Marketing Science 28, 674–689. http://dx.doi.org/10.1287/mksc.1080.0439.

Hambleton, R.H., Jones, R.W., 1994. Item parameter estimation errors and their influence on test information functions. Applied Measurement in Education 7, 171–186. http://dx.doi.org/10.1207/s15324818ame0703_1.

Huitzing, H.A., Veldkamp, B.P., Verschoor, A.J., 2005. Infeasibility in automated test assembly models: a comparison study of different methods. Journal of Educational Measurement 42, 223–243. http://dx.doi.org/10.1111/j.1745-3984.2005.00012.x.

IBM, 2010. CPLEX. IBM Corporation Software Group, Somers, NY.

Lord, F.M., 1980. Applications of Item Response Theory to Practical Testing Problems. Erlbaum, Hillsdale, NJ.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Welsley, Reading MA.

Luecht, R.M., Hirsch, T.M., 1992. Item selection using an average growth approximation of target information functions. Applied Psychological Measurement 16, 41–51. http://dx.doi.org/10.1177/014662169201600104.

Luecht, R.M., Nungester, R.J., 1998. Some practical examples of computer-adaptive sequential testing. Journal of Educational Measurement 35, 229–249. http://dx.doi.org/10.1111/j.1745–3984.1998.tb00537.x.

Nemhauser, G.L., Wolsey, L.A., 1988. Integer and Combinatorial Optimization. Wiley, New York.

Soyster, A.L., 1973. Convex programming with set-inclusive constraints and applications to inexact linear programming. Operations Research 21, 1154–1157.

Stocking, M.L., Swanson, L., 1993. A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement 17, 277–292. http://dx.doi.org/10.1177/014662169301700308.

Sympson, J.B., Hetter, R.D., 1985. Controlling item-exposure rates in computerized adaptive testing. In: Proceedings of the 27th Annual Meeting of the Military Testing Association. Navy Personnel Research and Development Center, San Diego, CA, pp. 973–977.

van der Linden, W.J., 2005. Linear Models of Optimal Test Design. Springer, New York.

van der Linden, W.J., Veldkamp, B.P., 2004. Constraining item exposure rates in computerized adaptive testing with shadow tests. Journal of Educational and Behavioral Statistics 29, 273–291. http://dx.doi.org/10.3102/10769986029003273.

Veldkamp, B.P., 2002. Multidimensional constrained test assembly. Applied Psychological Measurement 26, 133–146. http://dx.doi.org/10.1177/01421602026002002.

Verschoor, A.J., 2007. Genetic Algorithms for Automated Test Assembly, Unpublished Doctoral Dissertation. Enschede, The Netherlands: University of Twente.

Wainer, H., 2000. Computerized Adaptive Testing: A Primer, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.

Zenisky, A., Hambleton, R.K., Luecht, R.M., 2010. Multistage testing: issues, design and research. In: van der Linden, W.J., Glas, C.A.W. (Eds.), Elements of Adaptive Testing. Springer, New York, pp. 355–372. http://dx.doi.org/10.1007/978-0-387-85461-8_18.