Journal of Statistical Planning and
Inference 136 (2006) 260–269

Improved bounds for group testing designs

P.S.S.N.V.P. Rao, S.B. Rao, Bikas K. Sinha*

Statistics-Mathematics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700108, India

Received 7 February 2003; accepted 4 May 2004
Available online 23 July 2004

Abstract

Group testing designs (GTDs), both adaptive and nonadaptive, are useful in reducing the number of tests needed to identify the defective items from a given set of at least six items. In this paper, we obtain improved bounds on the number of group tests necessary for both adaptive and nonadaptive GTDs. It is established that any nonadaptive GTD needs at least $2n$ group tests for identifying all the defective items from a group of 2^n items having at most 2 defective items. In the same context, an adaptive multistage GTD with a maximum of $2n$ group tests is presented here. It is further shown that under restrictions on group size, optimal nonadaptive GTDs can be constructed using Generalized Petersen Graphs. Also presented is the construction of a family of two-stage adaptive GTDs that are useful under certain conditions.

MSC: primary 62K05; secondary 05C50; 68R10

Keywords: Adaptive and non-adaptive designs; Petersen graphs; Generalized Petersen graphs; Two-stage and multistage designs

1. Introduction and preliminaries

Given a set S of p items, of which a few ($\leqslant d$) are suspected to be defective, the problem is to identify all the defective items. To test individually we need p tests which may be prohibitive if the tests are expensive. In view of this, group testing designs (GTDs) are proposed to reduce the number of tests. We refer to Du and Hwang (2000) for an account of theoretical results in this area of research.

* Corresponding author. Fax: +91-33-2577-3071.
E-mail address: bksinha@isical.ac.in (B.K. Sinha).

A GTD consists of a number of group tests $G_1, G_2, \ldots, G_g$ which are nonempty subsets of the original set S, so that the results of the tests on these g-groups collectively lead to the identification of all the defective items of S. A group test G will end in *either* a positive *or* a negative result. A negative result automatically labels all the constituent items as non-defective. On the other hand, a positive result labels *at least one* constituent item as defective.

Naturally, for a given number of items, we would like the number of group tests g to be minimum possible. However, the following aspects are also to be considered in trying to minimize g:

(a) As d is relatively much smaller than p, large size groups with one or two defective items may need a very sensitive test to result positive. So the sensitivity of the tests restricts the group size.

(b) Again, depending on the type of tests used, it may not be possible to include items in as many tests as we want and this restricts the number of groups in which any particular item can appear.

GTDs where all the groups are formed and tested simultaneously are called nonadaptive (or nonsequential) GTDs, whereas those where testing is done in stages, with the grouping at any stage being done based on the results of the tests of the previous stages, are called adaptive (or sequential) GTDs. Further, if nothing is known about the number of defective items then the situation is called Binomial one, while the other situation where a bound or an exact number of defective items is known, is called Hypergeometric situation. In this paper, we will deal with Hypergeometric group testing designs only.

2. Nonadaptive GTDs under hypergeometric situation: a review

Any nonadaptive GTD that uses g group tests to identify all the defective items from the given set S of p items may be represented by its design (or incidence) matrix A which is a boolean matrix (with elements 0 and 1) of order $g \times p$ such that $a_{ij} = 1$ if jth item appears in the ith group and $a_{ij} = 0$ otherwise, for $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, p$, where a_{ij} denotes the (i, j)th element of A.

Below we summarize some essential features of a GTD involving p items and g tests in terms of boolean vectors and matrices.

(a) A boolean vector D of order $p \times 1$ is said to be a vector of defective items (vod) if for $j = 1, 2, \ldots, p$, $D_j = 1$ implies jth item is defective and $D_j = 0$ implies jth item is not defective. (b) A boolean vector R of order $g \times 1$ is said to be a result vector if for $i = 1, 2, \ldots, g$, $R_i = 1$ implies the ith test resulted in positive and $R_i = 0$ implies the ith test resulted in negative. (c) For a given design matrix A and a given vod D, the result vector R is given by $R = AD$ where the product is boolean.

The weight of a boolean vector R is the number of its nonzero elements and is denoted by $w(R)$. The complement $\overline{X}$ of a vector X comprising of 0's and 1's, is the vector obtained by interchanging the 0's and 1's of X.

Definition 1.1. A GTD is said to have *d-detecting* power if it can detect the set of all the d defective items.

In the initial stage of development of this topic of research, emphasis was given on d-completeness property (see Bush et al., 1980; Saha et al., 1982). However, it turned out that d-detecting power makes more sense (see Saha and Sinha, 1980). In order that a GTD has *d-detecting* power, the *result* vector of order $g \times 1$ must have *one-to-one correspondence* with the set of defective items. For $d = 1$, this means that the column vectors of A should be distinct. For $d = 2$, we also need the condition that the boolean sums of any pair of column vectors of A should be different from that of any other pair. These were first laid down in Saha and Sinha (1981), who observed that the number of tests is the same as the number of items p, whenever $p \leqslant 5$. Subsequently, Hwang and Sós (1987), Vakil et al. (1990) and Saha (2002) restated these features of group testing designs for $d = 2$.

Below we state a few results involving GTDs.

Theorem 1.1. *A necessary and sufficient condition for a matrix $A_{g\times p}$ to be the design matrix of a GTD that has* d-detecting *power is that*

$$Ax = Ay \Longrightarrow x = y$$

for all boolean vectors x, y of order $p \times 1$ and of weight $\leqslant d$.

The following result for the case $d = 1$, is well established in the literature:

Theorem 1.2. *Suppose $2^{n-1} \leqslant p < 2^n$ and let $A_{n\times p}$ be the matrix with jth column as the n-bit binary representation of the number j, for $j = 1, 2, \ldots, p$. Then the GTD with A as the design matrix has 1-detecting power involving n items and p tests.*

Corollary 1.1. *If it is known that there is exactly one defective the above design can also work for $p = 2^n$, with the design matrix obtained by appending a null column to the above matrix A.*

Observe that in the above case the maximum group size (maximum row weight) of A is $\lceil p/2 \rceil$ and the maximum frequency of an item (i.e. appearance in number of tests) is n, assuming $p = 2^n - 1$.

Theorem 1.3. *For the case of $p = 2^n$ and $d = 2$, any nonadaptive GTD needs at least $2n$ tests.*

Proof. Any nonadaptive GTD needs at least $1 + p + \binom{p}{2}$ different configurations of result vectors (g-tuples) to identify the cases of no defective item and all possible cases of 1 and 2 defective items.

Therefore we have, the total number of distinct g-tuples,

$$2^g \geqslant \binom{p}{2} + p + 1 = \binom{2^n}{2} + 2^n + 1$$

which is equivalent to the condition $g \geqslant 2n$. □

3. GTDs based on generalized Petersen graphs

Available GTDs for $d = 2$ are based on the use of designs such as BIBDs, GDDs, etc. In this section, we propose to examine the use of graphs for constructing GTDs when $d = 2$. Du and Hwang (2000) briefly mention about it.

A graph $G = (V, E)$ without multiple edges is an ordered pair where V is a nonempty set called the set of vertices of G, and E is a set of unordered pairs of distinct vertices of G called the set of edges of G. Let $v_1, v_2, \ldots, v_n$ be the vertices and $e_1, e_2, \ldots, e_m$ be the edges of the graph G. Let m be positive. The incidence matrix $A(G)$ is the matrix of order $n \times m$ where the (i, j)th element is 1 or 0 according as the edge e_j is incident at the vertex v_i or not. Note that each column of A is of weight 2 and the columns of A are distinct. An *n-cycle* is a sequence of n distinct vertices $(a_1, a_2, \ldots, a_n)$ such that for every i, $1 \leqslant i \leqslant n$, (a_i, a_{i+1}) is an edge of G where $a_{n+1} = a_1$.

Theorem 2.1. *The incidence matrix A of any graph G with no multiple edges and containing neither* 3*-cycles nor* 4*-cycles is a design matrix of a GTD for* $d = 2$.

Proof. We will base the proof on Theorem 1.1. First, it is easy to see that if G has a 3-cycle or a 4-cycle then there exists a pair of distinct boolean vectors x and y of weight 2 such that $Ax = Ay$. Conversely, let there exist a pair of distinct boolean vectors x and y of weight at most 2 such that $Ax = Ay$. As G has no multiple edges, each of x and y is of weight 2. Consequently, $Ax(=Ay)$ is of weight 3 or 4. In case the weight of Ax is 3, A contains, after necessary row and column permutations, the following submatrix of order 3×3:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

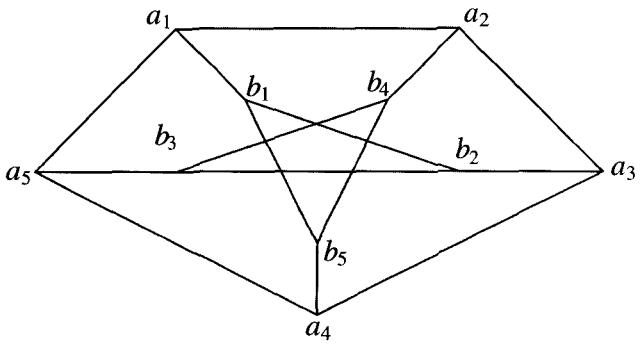
Now, it is clear that the edges corresponding to the above three columns of A form a 3-cycle. Similarly, in case of the weight of Ax is 4, it can be verified that A contains, after necessary row and column permutations, the following submatrix of order 4×4:

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

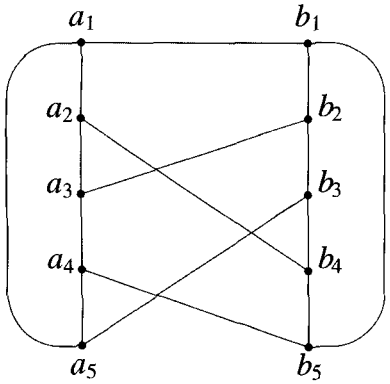
In this case the edges corresponding to these 4 columns of A form a 4-cycle. This completes the proof. □

Graphs satisfying the properties stated in Theorem 2.1 will be referred to as GTD graphs. Below we introduce generalized Petersen graphs and verify that such graphs are indeed GTD graphs. We also provide an alternative solution to a problem considered by Vakil (1990), using generalized Petersen graphs.

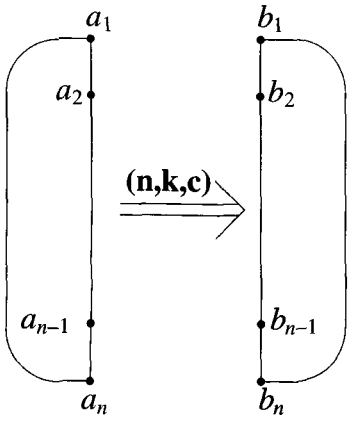
The following graph $P(10, 15)$ with 10 vertices and 15 edges is known as Petersen graph (see Harary, 1988).



This graph has no 3-cycle and no 4-cycle, and each vertex has degree 3. Writing it as



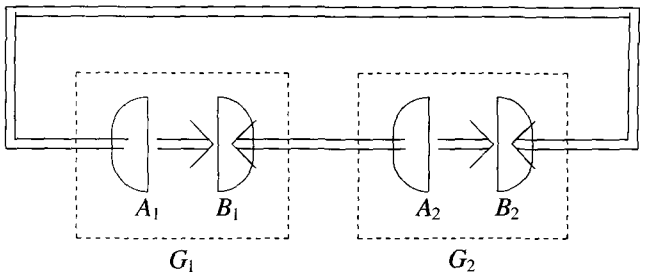
we can see that we can also construct generalized Petersen graphs $P(2n, 3n)$ with $2n$ vertices and $3n$ edges for $n \geqslant 5$, except for $n = 6$ as there is no such graph for $n = 6$. These can be constructed as follows:



Consider a graph consisting of two disjoint n-cycles $A : (a_1, a_2, \ldots, a_n)$ and $B : (b_1, b_2, \ldots, b_n)$. Connect a_i and b_j if $j \equiv (ik + c) \pmod{n}$ for $i = 1, 2, \ldots, n$, where k is any fixed integer such that $2 \leqslant k \leqslant n - 2$ and k is relatively prime to n and the value of the constant c can be chosen to be any integer between 0 and $n - 1$ where the modulo n is taken in the set $\{1, 2, \ldots, n\}$. Adding edge set from A to B as above is denoted as $A \overset{(n,k,c)}{\Longrightarrow} B$. For example, the Petersen graph given above is $A \overset{(5,3,3)}{\Longrightarrow} B$.

Again, it is easy to see that this resultant graph has neither a 3-cycle nor a 4-cycle.

Further, consider two $P(2n, 3n)$ graphs G_1 and G_2 where edge sets from A_1 to B_1 and from A_2 to B_2 are added as above, that is, $A_1 \overset{(n,k,c)}{\Longrightarrow} B_1$ and $A_2 \overset{(n,k,c)}{\Longrightarrow} B_2$.



These two can be connected by adding an edge set $A_2 \overset{(n,k,c)}{\Longrightarrow} B_1$ from A_2 to B_1 and also an edge set $A_1 \overset{(n,k,c_0)}{\Longrightarrow} B_2$ from A_1 to B_2, $c \neq c_0$. By choosing the values of the constants k, c, c_0 as above it can be shown that the graph has neither 3-cycles nor 4-cycles.This process of adding edges between two $P(2n, 3n)$graphs G_1 and G_2 is denoted by $G_1 \overset{(n,k,c,c_0)}{\Longleftrightarrow} G_2$. Observe that the degree of each vertex is 4 in the resultant graph.

Moreover, consider any $P(2n, 3n)$ graph G. Replace each vertex of G with a $P(2m, 3m)$ graph and each edge of G by edge set $\overset{(m,k,c_1,c_2)}{\Longleftrightarrow}$. The resulting graph G^* with $4mn$ vertices and $12mn$ edges, also does not contain any 3-cycle or 4-cycle. In fact, one can start with any GTD graph (graph without 3- and 4-cycles), replace each vertex by a $P(2m, 3m)$ graph and replace edges by edge sets $\overset{(m,k,c_1,c_2)}{\Longleftrightarrow}$ as above resulting in a GTD graph.

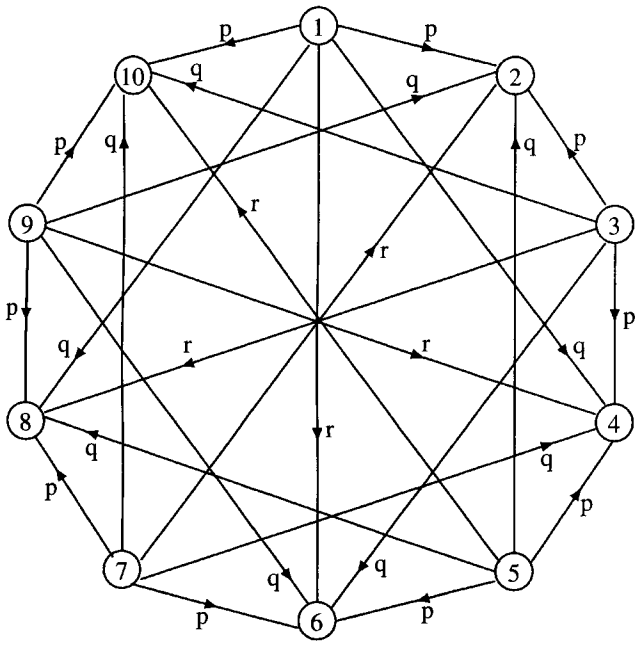
Also, graphs obtained by removing any subset of vertices and the edges incident on them from any of the above graphs result in GTD graphs.

The following rules can be used in identifying the defective items from the results of a GTD based on graphs.

As edges represent items and vertices represent groups in GTDs based on graphs, we refer the edges corresponding to defective items as defective edges and the vertices corresponding to groups that tested positive as positive vertices. Then the problem is to identify the defective edges based on the given set of positive vertices. Observe that no positive vertex implies no defective edge. There cannot be just a single vertex that is defective as every edge is incident on two vertices. In case of only two positive vertices the edge connecting them is the single defective edge which means that there is a single defective item. However, the case of two defective items may result in 3 or 4 positive vertices. In case of 3 positive vertices, the two edges connecting these vertices are the defective edges (as there are no

3-cycles in the graph there cannot be more than two edges connecting these 3 vertices). In case of 4 positive vertices there can be either 2 or 3 edges in the subgraph of these 4 vertices (as there are no 3-cycles and 4-cycles in the graph). If there are only 2 edges in this subgraph these two edges are the defective edges and in case of 3 edges, these 3 edges must form a chain of length 3 and it is easy to see that the two end edges of this chain are the defective edges. There cannot be more than 4 positive vertices as $d = 2$.

Vakil (1990) used design-based techniques to construct optimal GTDs in certain cases which include a GTD for 175 units and 50 tests for the case $d = 2$. We construct below a GTD for the same parameter values using graph-theoretic techniques.



Consider a graph consisting of 10 pairwise disjoint 5-cycles $C_1, C_2, \ldots, C_{10}$. In the figure above, circled i represents the 5-cycle C_i for $i = 1, 2, \ldots, 10$ and the arrows joining them represent the edge sets added as follows:

For $i = 1, 3, 5, 7, 9$, add edge sets from C_i to $C_{i\pm1}$, $C_{i\pm3}$ and C_{i+5} as below:

$$C_i \overset{p}{\Longrightarrow} C_{i\pm1},$$
$$C_i \overset{q}{\Longrightarrow} C_{i\pm3},$$
$$C_i \overset{r}{\Longrightarrow} C_{i+5},$$

where $p = (5, 2, 0)$, $q = (5, 2, 2)$ and $r = (5, 2, 1)$ and these are referred to as p, q and r connections. Observe that if there is a x $(=(5, 2, c))$ connection from A to B then the connection from B to A is $\overline{x}$ $(=(5, 3, 2c))$.

It is easy to see that there are no 3-cycles in the resultant graph. To show that there are no 4-cycles in this we proceed as follows:

Observe that any 4-cycle in the graph should involve four 5-cycles, say C_a, C_b, C_c and C_d. Without loss of generality we assume that a is odd which implies c is also odd and b and d are even. Then the 4-cycle must be having one of the following types of connections(or its rotation):

$$q\bar{r}q\bar{r},\ p\bar{q}q\bar{q}, p\bar{q}q\bar{r}, p\bar{q}r\bar{q}, p\bar{r}q\bar{q},\ p\bar{p}q\bar{r}, p\bar{p}r\bar{q}, p\bar{q}p\bar{r}, p\bar{q}p\bar{q}, p\bar{r}p\bar{r} \text{ and } p\bar{p}p\bar{q}$$

(these are 4-cycles with no p, one p, two p and three p connections) and it is easy to check that starting with any fixed vertex i of C_a none of these connections will connect back to the same vertex. Hence there cannot be any 4-cycle in the graph.

4. Adaptive GTDs

In this section we consider adaptive group testing designs and present two theorems in this direction.

Theorem 3.1. *For the case $d=2$ and for p of the form x^y, there exists a two-stage adaptive GTD with xy tests in the first stage and at most $y-1$ tests in the second stage.*

Proof. We number the $p(=x^y)$ items serially from 0 to x^y-1 and represent them as y-digit numbers with base x. For $i=1, 2, \ldots, y$ and $j=0, 1, 2, \ldots, x-1$ form groups $G_i^{(j)}$ where $G_i^{(j)}$ contains all those items with j as the ith digit in their representation. Observe that for each i, $G_i^{(0)}, G_i^{(1)}, \ldots, G_i^{(x-1)}$ form a partition of the x^y items, each containing x^{y-1} items.

In the first stage we test all the $x \cdot y$ groups $G_i^{(j)}$. If no test shows positive, we conclude that there is no defective item. Otherwise for each i either one or two of the values of j, $G_i^{(j)}$ results in positive. However if for all i only for one value of j, $G_i^{(j)}$ is positive, then there exists exactly one defective item and it can be identified easily. If there is at least one i for which there are two values of j such that $G_i^{(j)}$ tests positive, then there are two defective items and we go to second stage.

Observe that for each i there can be at most two values of j, for which $G_i^{(j)}$ is positive. Let i_1 and i_2 denote these values of j, if there are two values; and for some i, if there is a single value of j for which $G_i^{(j)}$ is positive, we will give both i_1 and i_2 the same value j. This gives us $2y$ digits having two digits (need not be distinct) for each of the y digit positions giving a maximum of 2^y numbers that represent the 2^y items that contain the two defective items.

Without loss of generality let i_1 and i_2 be different for $i=1$. From the above 2^y items form the group G with those items having numbers with first digit as 1_1. Observe that this group contains 2^{y-1} items of which exactly one item is defective and that can be identified using $y-1$ tests, by Corollary 1.1. This in turn identifies the y digits that represent its number. Then the remaining y digits give the number of the other defective item. Thus we need $y-1$ tests at the second stage making the total number of tests needed $xy+y-1$. This completes the proof. □

In passing, we may mention that Das and Roy Choudhury (1987) suggested a two-stage adaptive GTD which needs at most $(k+3)$ tests for $k(k+1)/2$ items.

Theorem 3.2. *There exists a multistage adaptive GTD which needs a maximum of* $2n$ *group tests to identify all the defective items from a given set of* $p=2^n$ *items containing a maximum of* 2 *defective items.*

Proof. Number the $p(=2^n)$ items serially from 0 to 2^n-1 and represent them as *n-bit* binary numbers.

Form groups $G_1^{(0)}$ and $G_1^{(1)}$, where $G_1^{(j)}$, for $j=0, 1$, contains all those items with j as the first bit in their binary representations. Observe that $G_1^{(0)}$ and $G_1^{(1)}$ form a partition of 2^n items each containing 2^{n-1} items.

Conduct the first pair of tests on $G_1^{(0)}$ and $G_1^{(1)}$. If both result in negative, we can conclude that there is no defective. If both test results are positive then there are two defective items. However, if only one result, say $G_1^{(j_1)}$, is positive then there could be one or two defective items, in which case we conduct tests on the next pair $G_2^{(0)}$ and $G_2^{(1)}$ where $G_2^{(j)}$, for $j=0, 1$, contains all those items with j_1 as the first bit and j as the second bit in their binary representations. Thus we continue testing pairs as long as exactly one of them results in positive. Let $(r+1)$th be the first pair of tests that results in both positive. At this stage, we have the following situation.

The first r pairs of tests resulted in $G_1^{(j_1)}, G_2^{(j_2)}, \ldots, G_r^{(j_r)}$ positive and at the $(r+1)$st stage both $G_{r+1}^{(0)}$ and $G_{r+1}^{(1)}$ resulted in positive. Now, consider the two groups, the first group containing all those items having representation with the first $(r+1)$ bits as $j_1 j_2 \ldots j_r 0$ and the second group containing all those items having the first $(r+1)$ bits as $j_1 j_2 \cdots j_r 1$. Observe that each of these groups contains exactly $2^{(n-r-1)}$ items with exactly one defective which can be identified using $(n-r-1)$ tests by Corollary 1.1. Thus at this last stage we need $2(n-r-1)$ tests to identify both the defective items. Hence, as $2(r+1)$ tests are used in the first $(r+1)$ stages, a total of $2n$ tests are enough to identify all the defective items.

5. Concluding remarks

In this paper, we have basically demonstrated the use of Petersen graphs and their Generalizations towards finding solutions to the GTDs. We have confined our attention to the cases wherein at most 2 defective units are to be found in the population. In the literature, however, there are some results available for the general case. We did not enter into any discussion along that line. We refer to the results on Tactical configurations including Steiner Triple Systems, in Hanani (1961), Raghavarao (1971) and Wideman and Raghavarao (1987a, b) for the interested readers. Neither did we discuss any results related to binomial testing which can be found in, for example, Kumar and Sobel (1971) and the references therein.

Acknowledgement

The authors are thankful to Prof. G.M. Saha for some initial discussions during the preparation of this paper. Thanks are also due to the referees for their constructive and insightful comments.

References

Bush, K.A., Federer, W.T., Pesotan, H., Raghavarao, D., 1980. New combinatorial designs and their applications to group testing. Ars Combin

Das, M.N., Roy Choudhury, D., 1987. On problems of search using group testing. Sankhya B 49, 137–147.

Du, D.-Z., Hwang, F.K., 2000. Combinatorial Group Testing and Its Applications. World Scientific, Singapore.

Hanani, H., 1961. The existence and construction of balanced incomplete block designs. Ann. Math. Statist. 32, 361–386.

Harary, F., 1988. Graph Theory, Narosa.

Hwang, F.K., Sós, V.T., 1987. Non-adaptive hypergeometric group testing. Studia Sci. Math. Hung. 257–263.

Kumar, S., Sobel, M., 1971. Finding a single defective in binomial group testing. J. Amer. Statist. Assoc. 66, 824–828.

Raghavarao, D., 1971. Constructions and Combinatorial Problems in Design of Experiments. John Wiley, New York.

Saha, G.M., 2002. Designs for Group Testing Experiments. Indian Science Congress Association, Manuscript.

Saha, G.M., Sinha, B.K., 1980. Some combinatorial aspects of designs useful in group testing experiments, unpublished manuscript.

Saha, G.M., Sinha, B.K., 1981. Some combinatorial aspects of designs useful in group testing experiments—an addendum. Tech. Report no. 11/81, stat–Math Unit, Indian Statistical Institute.

Saha, G.M., Pesotan, H., Raktoe, B.L., 1982. Some results on t-complete designs. Ars Combin. 13, 195–201.

Vakil, F., 1990. Non-adaptive group testing procedures for hypergeometric problem. Ph. D. Dissertation, Temple University.

Vakil, F., Parnes, M., Raghavarao, D., 1990. Group testing with at most two defectives when every item is included in exactly two group tests. Utilitas Math. 38, 161–164.

Weideman, C.A., Raghavarao, D., 1987a. Some optimum non-adaptive hypergeometric group testing designs for identifying two defectives. J. Statist. Plann. Inference 16, 55–61.

Weideman, C.A., Raghavarao, D, 1987b. Nonadaptive hypergeometric group testing designs for identifying at most two defectives. Commun. Statist. 16A, 2991–3006.