AN OPTIMAL STRATEGY FOR SEQUENTIAL CLASSIFICATION ON PARTIALLY ORDERED SETS

Thomas S. Ferguson
Department of Mathematics
UCLA
Los Angeles, CA 90095, USA

Curtis Tatsuoka[1,2]
Department of Statistics
The George Washington University
Washington, DC 20052, USA

Abstract

A decision-theoretic framework is described for sequential classification when the parameter space is a finite partially ordered set. An example of an optimal strategy is then presented. This example establishes that an asymptotically optimal class of experiment selection rules is not necessarily optimal in the given decision-theoretic setting.

KEY WORDS: sequential selection of experiment, group testing, cognitively diagnostic educational testing, decision theory

1. Introduction.

Methods of choosing a sequence of experiments for sequential classification on finite sets in order to determine the true state at the maximum asymptotic rate were presented in Tatsuoka and Ferguson (2003). Application was made to the area of cognitively diagnostic educational testing, where the classification states form a partially ordered set (see

1 Corresponding author

2 This work was supported in part by NSF Grant SES-9810202, and NIMH Grant R01 MH65538.

Tatsuoka (2002)). Below, we treat the same problem in a decision-theoretic framework where there is a cost of observation. This requires that classification must be made in a finite time. It is shown that a class of rules, seen to be asymptotically optimal in Tatsuoka and Ferguson (2003) when the classification states form a lattice, satisfy a conceptually appealing property, here called *ordered experiment selection*. This property is related to the idea of playing the winner in bandit problems. Finally, an example is given for which we actually compute the optimal rule in the decision-theoretic sense, and it is seen not satisfy the ordered experiment selection property.

Partially ordered sets (posets) are useful in statistical applications such as the group testing problem originated by Dorfman (1943) (see also Ungar (1960), Sobel and Groll (1959) and (1966), Yao and Hwang (1990), and Gastwirth and Johnson (1994)). Another example of the use of partially ordered classification models is in cognitively diagnostic educational testing (e.g. Tatsuoka (2002), Falmagne and Doignon (1988)). These models may also be useful for neuropsychological assessment.

2. Statistical Formulation.

Let S denote the set of classification states, assumed to have at least two elements. Let $\mathcal{E}$ represent a collection of experiments. If an experiment $e \in \mathcal{E}$ is used, then a random variable, X, is observed whose distribution depends on e and on the true unknown state, denoted by $s \in S$. We assume that for $e \in \mathcal{E}$ and state $s \in S$, the corresponding conditional response distribution of X has some probability density $f(x|e, s)$ with respect to a σ-finite measure μ_e on a measurable space $(\mathcal{X}_e, \mathcal{B}_e)$. We also assume that the prior distribution of the true state is known, π_0, and consider the Bayes approach.

Let $\pi_0(j)$ denote the prior probability that $j \in S$ is the true state. At the first stage, an experiment $e_1 \in \mathcal{E}$ is chosen and a random variable X_1 having density $f(x|e_1, s)$ is observed. The posterior distribution on the parameter space, denoted by $\pi_1(j)$, is proportional to the prior times the likelihood, that is, $\pi_1(j) \propto \pi_0(j) f(x_1|e_1, j)$, where x_1 represents the

observed value of X_1. Inductively, at stage n for $n > 1$, conditionally on having chosen experiments e_1, e_2,..., e_{n-1}, and having observed $X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}$, an experiment $e_n \in \mathcal{E}$ is chosen and X_n with density $f(x|e_n, s)$ is observed. The posterior distribution then becomes

$$\pi_n(j) \propto \pi_0(j) \prod_{i=1}^{n} f(x_i|e_i, j).$$

The posterior probability distribution on S at stage n will be denoted by $\boldsymbol{\pi}_n$.

We assume that $(S, \leq)$ is a partially ordered set, and that experiments are identified with states in S as follows. If X represents the response random variable, then the density of X for a given experiment $e \in S$ and true state $s \in S$ is given by

$$f_X(x|e, s) = \begin{cases} f(x) & \text{if } e \leq s \\ g(x) & \text{otherwise.} \end{cases} \tag{1}$$

In the educational application, the interpretation given to (1) is as follows. For states j and k in S, $j \leq k$ means that a subject in state k has all the relevant knowledge that a subject in state j has. For each state $e \in S$, experiments can be designed so that subjects in any state $k \geq e$ have response distribution $f(x)$ while those in other states have response distribution $g(x)$.

In the group testing problem, we are given a set of N objects each of which may be either defective or nondefective. The set of classification states consists of the 2^N possible subsets of objects. These subsets can be viewed as partially ordered through inclusion. A Bayesian formulation of the group testing problem can be treated using (1) by considering the true state, s, as the set of non-defective objects. An experiment, e is just a subset of the objects. The outcome has one distribution if $e \subseteq s$ (no defectives in e) and another if $e \not\subseteq s$ (at least one defective in e). By identifying $e \subseteq s$ with $e \leq s$, (1) follows.

If S has a bottom element $\hat{0}$ (i.e. $\hat{0} \leq j$, all $j \in S$), then the experiment $e = \hat{0}$ gives no information since all states in S will have the same response distribution f. Thus, when $\hat{0}$ exists, we take $\mathcal{E} = S \backslash \{\hat{0}\}$ as the set of experiments. Otherwise, let $\mathcal{E} = S$. We assume f and g are not identical distributions and not mutually singular.

3. A Class of Experiment Selection Procedures.

As described in Tatsuoka and Ferguson (2003), an appealing class of experiment selection rules can be described which depend on the quantity

$$m_n(e) = \sum_{j \geq e} \pi_n(j).$$

For example, the *halving algorithm* chooses that experiment e at stage $n+1$ for which $m_n(e)$ is closest to $1/2$. Such an experiment splits the state space into two subsets with different response distributions, which have as close as possible to equal probabilities.

We define a class of experiment selection procedures $\mathcal{U}$. Let experiment selection procedures in $\mathcal{U}$ be those that choose $e \in \mathcal{E}$ at stage n to maximize $U(m_n(e))$ for some continuous function U, defined on $[0,1]$, such that (i) $U(0) = U(1) = 0$, (ii) U strictly unimodal in (0,1), and (iii) there exist numbers $0 < k_0 < k_0' < \infty$ and $0 < k_1 < k_1' < \infty$ such that

$$k_0 x < U(x) < k_0' x \qquad \text{for all } x \text{ sufficiently close to 0}$$

$$k_1(1-x) < U(x) < k_1'(1-x) \qquad \text{for all } x \text{ sufficiently close to 1.}$$

We take $\mathcal{U}$ to be the class of all such selection procedures. The halving algorithm, for example, is a member of $\mathcal{U}$ associated with the function $U(x) = \min\{x, 1-x\}$. A procedure minimizing Shannon entropy one step ahead also satisfies the conditions for belonging to $\mathcal{U}$ (see Tatsuoka and Ferguson (2003)). This class of heuristics shares some nice properties. In particular, rules in $\mathcal{U}$ attain the optimal rate of convergence of the posterior probability of the true state to one when S is a lattice. Recall that a lattice is defined to be a poset such that any two elements have both a unique least upper bound and a unique greatest lower bound (cf. Davey and Priestly (2002)). Note that S is a lattice for the group testing problem. Also, these algorithms are computationally simple. However, use of this class of heuristics does not necessarily lead to optimal experiment selection in a decision-theoretic framework, as Example 1 of Section 6 demonstrates.

4. Decision-Theoretic Formulation.

We use a standard Bayesian decision-theoretic formulation (cf. Ferguson (1967)). The parameter space of classification states and the set of terminal actions are both taken to be the finite poset S. Action $j \in S$ denotes that the subject is classified into state j. The loss is taken to be 0-1 plus a constant cost of observation, and is written

$$L(s, j, n) = \begin{cases} 0 + cn & \text{if } j = s \\ 1 + cn & \text{otherwise,} \end{cases}$$

where s is the true state in S, j is the action of choosing state j, n is the number of observations and $c > 0$ is the cost per observation.

Rules that incorporate an experiment selection procedure as well as the stopping and terminal decision rules will be called strategies. When stopping occurs, the Bayes terminal decision rule that chooses the state with the largest posterior probability will be used. Let N denote the stopping time and let J denote the action taken after stopping. The risk of a strategy δ when s is the true state is given by

$$R(s, \delta) = E_s[L(s, J, N)|\delta].$$

The Bayesian decision-theoretic problem then is to find strategies that minimize the Bayes risk

$$r(\boldsymbol{\pi}_0, \delta) = \sum_{i \in S} R(i, \delta) \cdot \pi_0(i).$$

5. Ordered Experiment Selection.

In bandit problems, the property of play-the-winner says that under certain conditions if it is optimal to play an arm and it produces a winner, it is optimal to play the same arm again (see Berry and Fristedt (1985), for example). An analogous concept for this problem is defined below.

Definition. *An procedure is said to have the ordered experiment selection property if, for any stage n, given that $e_n = e$ and $f(X_n) > g(X_n)$ (respectively $f(X_n) < g(X_n)$), then $e_{n+1} \not< e$ (respectively $e \not< e_{n+1}$).*

We expect good rules to have this property. Suppose it is good to use e at stage n. If we use e, then an observation X_n with $f(X_n) > g(X_n)$ lends support to the hypothesis that the student has at least the relevant knowledge of state e. Why then would we use an experiment to see if the student has the relevant knowledge of some state $j < e$? Similarly, an observation X_n with $f(X_n) < g(X_n)$ lends support to the hypothesis that the student doesn't have the relevant knowledge of state e. Why then would it be good to test to see if the student has even more knowledge?

In the group testing problem, a related idea is that of *nested* in Sobel and Groll (1959) (see also Yao and Hwang (1990)). At stage n, suppose that f and g are mutually singular, and that a defective is present in an experiment e (that is, suppose that $0 = f(X_n) < g(X_n)$). An experiment selection rule that is nested would require that the next stage experiment consist of a subset of objects from those pooled at stage n (i.e. that $e_{n+1} < e$). The ordered experiment selection property only requires $e_{n+1} \not> e$, and so it is not as restrictive.

We first show that procedures in $\mathcal{U}$ have the ordered experiment selection property. Because of the possibility of ties in achieving the maximum of $U(m_n(e))$, there may exist a choice of experiments for a given $U \in \mathcal{U}$ that does not satisfy the property. We can say, however, that for a given U there exists an experiment selection that does satisfy the property. We prefer, instead, to add a condition satisfied in the main applications under which all procedures in $\mathcal{U}$ have the ordered experiment selection property. This condition is equivalent to the condition that $\pi_n(j) > 0$ a.s. for all $j \in S$ and all $n = 0, 1, \ldots$.

Theorem 1. *If $\pi_0(j) > 0$ for all $j \in S$ and $0 < f(x)/g(x) < \infty$ for almost all x $(d\mu)$, then every procedure in $\mathcal{U}$ has the ordered experiment selection property.*

6. An Example of an Optimal Strategy.

Recall the group-testing problem, where few optimality results exist. In this section, an optimal strategy is characterized. Surprisingly, the optimal rule may not always satisfy the ordered experiment selection property. Consider the following counterexample.

Example 1. Consider the poset of Figure 1, and assume that f and g are densities for Bernoulli random variables, with respective parameters $p_u = .9$ and $p_l = .1$. Let $c = .04$. As the prior, we take $\pi_0(A) = .75/(1+\epsilon)$, $\pi_0(\hat{1}) = .25/(1+\epsilon)$, $\pi_0(B) = \epsilon/(1+\epsilon)$, and $\pi_0(\hat{0}) = 0$, where ϵ is a small positive number. Note that the only reasonable choices of experiment are ones associated with $\hat{1}$ or B, since the only separation of interest is between A and $\hat{1}$. Define the strategy $\delta_{\hat{1}B}$ to be the strategy that alternates forever between experiments $\hat{1}$ and B starting with $\hat{1}$, except that it stops and classifies to A if there is a failure using $\hat{1}$ and stops and classifies to $\hat{1}$ if there a success using B. In the appendix it is shown that $\delta_{\hat{1}B}$ is the unique optimal strategy for all ϵ sufficiently small. However, $\delta_{\hat{1}B}$ does not have the ordered experiment selection property. If an experiment $\hat{1}$ is a success, it is followed by experiment B, and $B < \hat{1}$. Hence an optimal rule does not necessarily have this property!

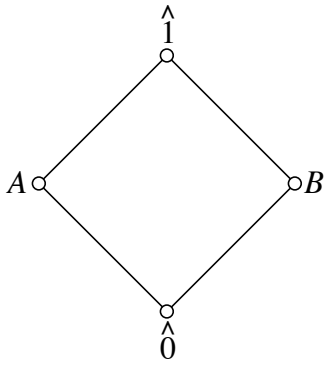

Figure 1

This example is counter-intuitive to the strategy of trying to build the largest finishing posterior probability value, which would require that the experiment choices be reversed (alternating between B and $\hat{1}$ starting with B). This shows how the cost of observation can

override this consideration. In general, however, it is expected that ordered experiment selection can perform quite reasonably.

7. Proofs.

Proof of Theorem 1. Consider an experiment selection rule in $\mathcal{U}$, let U denote its associated function and suppose $e \in \mathcal{E}$ maximizes $U(m_n(e))$. Then, experiment $e_{n+1} = e$ is used at stage $n+1$ and X_{n+1} is observed. Suppose $X_{n+1} = x$.

First consider the case $f(x) > g(x)$. We are to show that $U(m_{n+1}(j))$ is not maximized by any $j < e$. We do this by showing $U(m_{n+1}(j)) < U(m_{n+1}(e))$ for all $j < e$. Since $f(x) > g(x)$, we have $\pi_{n+1}(j) > \pi_n(j)$ for all $j \geq e$, and $\pi_{n+1}(j) < \pi_n(j)$ for all other j. In particular, $m_{n+1}(e) > m_n(e)$. Moreover, for all $j < e$ we have $m_{n+1}(j) \geq m_n(j)$, because $1 - m_{n+1}(j) = \sum_{i \not\geq j} \pi_{n+1}(i) \leq \sum_{i \not\geq j} \pi_n(i) = 1 - m_n(j)$. In addition, since we are assuming that $\pi_n(j) > 0$ for all n and j, we have for $j < e$ that $m_n(j) > m_n(e)$ and $m_{n+1}(j) > m_{n+1}(e)$.

Since e gives the maximum value of $U(m_n(e))$ and $m_n(j) > m_n(e)$, $m_n(j)$ must be on the downward sloping part of U. If $m_{n+1}(e)$ is also on the downward sloping part of U, we automatically have $U(m_{n+1}(j)) < U(m_{n+1}(e))$ for all $j < e$ since $m_{n+1}(e) < m_{n+1}(j)$. If $m_{n+1}(e)$ is on the upward sloping part of U, then $U(m_{n+1}(e)) > U(m_n(e)) > U(m_n(j)) \leq U(m_{n+1}(j))$. In all cases, $U(m_{n+1}(j)) < U(m_{n+1}(e))$ for all $j < e$, as was to be proved.

A similar argument shows that for the case $f(x) < g(x)$, $U(m_{n+1}(e)) > U(m_{n+1}(j))$ for all $j > e$.

Proof of Example 1. Let us denote the prior probability as $\boldsymbol{\pi}_0 = (\pi_0(A), \pi_0(\hat{1}), \pi_0(B))$.

Lemma 1. *There exists a number, ϵ_0 ($\epsilon_0 = 1/288$ will do), such that for all $0 < \epsilon < \epsilon_0$,*

(1) if $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$ or $\boldsymbol{\pi}_0 \propto (.25, .75, \epsilon)$, then it is optimal to take at least one observation,

(2) if $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$ *and a failure on* $\hat{1}$ *or* B *occurs, then it is optimal to stop and classify to* A*, and*

(3) if $\boldsymbol{\pi}_0 \propto (.25, .75, \epsilon)$ *and a success on* $\hat{1}$ *or* B *occurs, then it is optimal to stop and classify to* $\hat{1}$*.*

Proof. (1) follows by considering the one-stage look-ahead rule. Consider (2) when experiment $\hat{1}$ is used (the other statements are proved similarly). If failure is observed, the posterior probability is proportional to $(.75(.9), (.25)(.1), \epsilon(.1)) = (.675, .025, .9\epsilon)$. Stopping and classifying to A incurs only the misclassification cost, $(.025+.9\epsilon)/(.7+.9\epsilon)$. Continuing costs at least $c = .04$. The former is less than the latter if $\epsilon < 1/288$.

Lemma 2. *If* $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$*, then for the posterior* $\boldsymbol{\pi}_1$ *we have*

$$\boldsymbol{\pi}_1 \propto (.25, .75, \epsilon/3) \quad \textit{given success on } \hat{1}$$

$$\boldsymbol{\pi}_1 \propto (.25, .75, 3\epsilon) \quad \textit{given success on } B$$

$$\boldsymbol{\pi}_1 \propto (.75, .25, \epsilon/9) \quad \textit{given success on } A$$

$$\boldsymbol{\pi}_1 \propto (.75, .25, 9\epsilon) \quad \textit{given failure on } A$$

If $\boldsymbol{\pi}_0 \propto (.25, .75, \epsilon)$*, then for the posterior* $\boldsymbol{\pi}_1$ *we have*

$$\boldsymbol{\pi}_1 \propto (.75, .25, 3\epsilon) \quad \textit{given failure on } \hat{1}$$

$$\boldsymbol{\pi}_1 \propto (.75, .25, \epsilon/3) \quad \textit{given failure on } B$$

$$\boldsymbol{\pi}_1 \propto (.25, .75, \epsilon/9) \quad \textit{given success on } A$$

$$\boldsymbol{\pi}_1 \propto (.25, .75, 9\epsilon) \quad \textit{given failure on } A$$

It is sufficient to restrict attention to strategies that, given the selection rules, employ the optimal Bayes stopping and classification rules. From these two lemmas we may conclude that if $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$ with $\epsilon < \epsilon_0/9^M$, an optimal strategy will take at least M observations unless we get a failure on $\hat{1}$ or B when $\pi_n(A) \approx .75$ (resp. success on $\hat{1}$ or B when $\pi_n(\hat{1}) \approx .75$), in which case we stop and classify to A (resp. $\hat{1}$). Let us denote this class of strategies by $\mathcal{D}_M$.

Lemma 3. *For $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$, and for all $\delta \in \mathcal{D}_M$,*

$$r(\boldsymbol{\pi}_0, \delta) \geq \frac{c}{1+\epsilon}\left[\frac{(1-(.3)^M)}{.7} + \epsilon\frac{(1-(.1)^M)}{.9}\right].$$

Proof. The Bayes risk is at least $c \cdot E(\min\{N, M\})$ where N is the stopping time of the strategy. This is minimized by stopping as soon as possible, which is achieved using only experiments $\hat{1}$ and B. Conditional on A or $\hat{1}$ being true, N has a geometric distribution with success probability $(3/4).9 + (1/4).1 = .7$, and $E(\min\{N, M\}) = (1-(.3)^M)/.7$ is an easy calculation. If B is true, N is stochastically smallest if $\hat{1}$ and B are used repeatedly in that order, in which case N has a geometric distribution with success probability .9, and $E_B(\min\{N, M\}) = (1-(.1)^M)/.9$. Combining these gives the result.

Lemma 4. *If $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$, then for all ϵ sufficiently small and any strategy δ that begins with experiment A, $r(\boldsymbol{\pi}_0, \delta) > r(\boldsymbol{\pi}_0, \delta_{\hat{1}B})$.*

Proof. A straight-forward computation gives

$$r(\boldsymbol{\pi}_0, \delta_{\hat{1}B}) = \frac{(1/28)+\epsilon}{1+\epsilon} + \frac{c}{1+\epsilon}\left[\frac{1}{.7} + \epsilon\frac{1}{.9}\right].$$

For ϵ sufficiently small and any strategy that starts with A, there is an immediate cost of c and a subsequent cost of almost $c/.7$. Since $1/28 = .035 < c$, the strategy $\delta_{\hat{1}B}$ is better.

In fact, we can conclude that there is an M_0 such that if $\epsilon < \epsilon_0/9^{M+M_0}$, then $\delta_{\hat{1}B}$ is better than any strategy that uses experiment A in any of the first M stages. Let $\mathcal{D}^*_{M+M_0}$ denote the subset of $\mathcal{D}_{M+M_0}$ that uses only $\hat{1}$ and B as experiments in the first M stages. Note that $\delta_{\hat{1}B}$ is optimal if $\epsilon = 0$ (It is a SPRT though not uniquely). Therefore, for any other strategy δ,

$$R(A, \delta) \cdot \pi_0(A) + R(\hat{1}, \delta) \cdot \pi_0(\hat{1}) \geq R(A, \delta_{\hat{1}B}) \cdot \pi_0(A) + R(\hat{1}, \delta_{\hat{1}B}) \cdot \pi_0(\hat{1}),$$

since this is true when dividing through by $\pi_0(A) + \pi_0(\hat{1}) = 1 - \epsilon$. So in comparing strategies in $\mathcal{D}^*_{M+M_0}$, we need only compare $R(B, \delta)$.

Lemma 5. *For all strategies δ in $\mathcal{D}^*_{M+M_0}$,*

$$R(B,\delta) \geq 1-(.9)^{M+1} + c(1-(.1)^M)/.9.$$

Proof. Let N denote the stopping time of δ. Then, $R(B,\delta) = P_B(\text{misclassification}) + cE_B(N) \geq P_B(N \leq M+1) + cE_B \min\{N, M\}$, since if we stop before stage $M+1$ we classify to A or $\hat{1}$. The probability $P_B(N \leq M+1)$ is minimized by stopping as slowly as possible, and is achieved using B and $\hat{1}$ repeatedly in that order, so that $P_B(N \leq M+1) \geq 1-(.9)^{M+1}$. The expectation is minimized by stopping as quickly as possible, achieved using $\hat{1}$ and B repeatedly in that order, so that as in Lemma 3, $E\min\{N, M\} \geq (1-(.1)^M)/.9$.

Lemma 6. *If $\boldsymbol{\pi}_0 \propto (.75, .25, \epsilon)$, then $\delta_{\hat{1}B}$ is optimal for all sufficiently small ϵ.*

Proof. Since $R(B, \delta_{\hat{1}B}) = 1 + c/.9$, take M very large so that the result of lemma 5 is within some ϵ' of $R(B, \delta_{\hat{1}B})$. If δ in $\mathcal{D}^*_{M+M_0}$ starts with experiment B, then $R(B,\delta) \geq c + .1 + .9(1 + c/.9 - \epsilon') > R(B, \delta_{\hat{1}B})$, and hence $r(\boldsymbol{\pi}_0, \delta) > r(\boldsymbol{\pi}_0, \delta_{\hat{1}B})$. Thus for ϵ sufficiently small, $r(\boldsymbol{\pi}_0, \delta) > r(\boldsymbol{\pi}_0, \delta_{\hat{1}B})$ for any strategy δ that starts with A or B. Similarly, if δ begins with $\hat{1}\hat{1}$ or $\hat{1}A$, it is not as good as $\delta_{\hat{1}B}$. Now note that use of $\delta_{\hat{1}B}$ reduces ϵ at every step of continuation. If any strategy δ differs from $\delta_{\hat{1}B}$ there is a first time it differs, but then ϵ is even smaller so that $\delta_{\hat{1}B}$ is better.

References.

Berry, D. A. and Fristedt, B. (1985), *Bandit Problems*, London: Chapman & Hall.

Davey, B. A. and Priestley, H. A. (2002), *Introduction to Lattices and Order, Second Edition*, Cambridge: Cambridge University Press.

Dorfman, R. (1943), "The Detection of Defective Members of a Large Population", *Annals of Mathematical Statistics*, 14, 436-440.

Falmagne, J.-C. and Doignon, J.-P. (1988), "A Class of Stochastic Procedures for the Assessment of Knowledge", *British Journal of Mathematical Psychology*, 41, 1-23.

Ferguson, T. (1967), *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press.

Gastwirth, J. L. and Johnson, W. O. (1994), " Quality Control for Screening Tests: Potential Applications to HIV and Drug Use Detection", *Journal of the American Statistical Association*, 89, 972-981.

Sobel, M. and Groll, P. (1959), "Group Testing to Eliminate Efficiently All Defectives in a Binomial Sample", *Bell System Technical Journal*, 38, 1179-1252.

Sobel, M. and Groll, P. (1966), "Binomial Group-Testing with an Unknown Proportion of Defectives", *Technometrics*, 8, 631-656.

Tatsuoka, C. and Ferguson, T. (2003), "Sequential Classification on Partially Ordered Sets", *Journal of the Royal Statistical Society, Series B*, 65, 143-157.

Tatsuoka, C. (2002), "Data Analytic Methods for Latent Partially Ordered Classification Models", *Applied Statistics*, 51, 337-350.

Ungar, P. (1960), "The Cutoff Point for Group Testing", *Communications in Pure and Applied Mathematics*, 13, 49-54.

Yao, Y. C. and Hwang, F. K. (1990), "On Optimal Nested Group Testing Algorithms", *Journal of Statistical Planning and Inference*, 24, 167-178.