Relative Mean Squared Error and Cost Considerations in Choosing Group Size for Group Testing to Estimate Infection Rates and Probabilities of Disease Transmission

William H. Swallow

Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203.

I thank G. G. Kennedy, J. W. Moyer, and Louise R. Romanow for their encouragement, and the senior editor and referees for helpful suggestions.

This research was supported by USDA-CSRS Competitive Grant 81-CRCR-1-0765. The work was partially completed while the author was a guest of the International Agriculture Center and the Department of Mathematics of the Agricultural University, Wageningen, the Netherlands.

Accepted for publication 31 March 1987 (submitted for electronic processing).

ABSTRACT

Swallow, W. H. 1987. Relative mean squared error and cost considerations in choosing group size for group testing to estimate infection rates and probabilities of disease transmission. Phytopathology 77:1376-1381.

Virus-vector research often has the goal of estimating some probability p, which may be an infection rate or a probability of disease transmission by a single vector. Statisticians have recommended group testing (multiple-vector-transfer designs) over single-vector-transfer designs for doing this efficiently. In group testing, vectors or other units are tested in groups, rather than individually. The statistical argument for group testing has been based on the superior properties of optimal (meaning minimum mean squared error) group-testing designs. However, in application, the optimal design cannot be used or even identified, because to do so requires knowing beforehand the unknown p, the very probability one is preparing to estimate. As a workable alternative, using group testing with a group size (vectors per test plant, pool size per test) presumed to be smaller than the optimal group size (for minimizing mean squared error with a fixed number of tests or test plants) has been recommended as a safe way to realize at least some (unknown) fraction of the benefits of the optimal group-testing design. A method for choosing that smaller group size has been suggested in earlier work. This paper provides more detailed information on the extent of benefits one can actually hope to realize with group testing using smaller-than-optimal group sizes. It shows that often a large fraction of the benefits of the optimal design can be attained with a much smaller group size. It further illustrates how, when costs are taken into account, using a smaller group size may actually be more cost effective than using the group size that is optimal for minimizing mean squared error.

Additional key words: aphid vectors, estimating plant resistance, insect vectors, maximum likelihood estimation, virus transmission.

Vector-transfer designs are widely used in studying the spread of insect-borne diseases, quantifying resistance factors in plants, and estimating proportions of infective vectors (3,4,6,7). For example, to quantify resistance in a particular cultivar, an experimenter might move k vectors (i.e, aphids or leafhoppers) from an infected source plant to each of N noninfected test plants of this cultivar and later record the fraction (H) of the N test plants remaining healthy or, alternatively, the fraction ($1 - H$) developing symptoms of disease. To compare several "treatments" (e.g., recipient test plant cultivars, virus sources, or vectors), this scheme could be carried out for each treatment in turn. When $k = 1$, we call the experimental design a single-vector-transfer design; when $k > 1$, we call it a multiple-vector-transfer or group-testing design (with k being the size of the group tested together).

Provided the k vectors on each of the N test plants can be assumed to behave independently, one can for each treatment estimate its

p = the probability of disease transmission by a *single* vector,

using the estimator $\hat{p}$ given by

$$\hat{p} = 1 - H^{1/k}.$$

This estimator has a long history of application (4,7), and its statistical properties have been discussed by a number of authors (1,2,5,6). As Swallow (5) emphasized, it is essential that treatments be compared through their $\hat{p}$s, not through their raw fractions of healthy or infected plants (Hs or $[1 - H]$s), which depend on the values of k used as well as on the treatments themselves.

Once the data have been collected, calculating $\hat{p}$ for a particular treatment is easy; all one needs to know is the fraction of test plants remaining healthy (H) and the number of vectors (k) placed on each of the N test plants. The estimator $\hat{p}$ does assume that the same value of k has been used for all N test plants that will be used to calculate a particular $\hat{p}$. However, different values of k (or N) may have been used for different treatments (different $\hat{p}$s) and perhaps should have been (5). In practice, one may even obtain several $\hat{p}$s for the same treatment, perhaps as replication over time where the amount of data one can collect at any one time is constrained by labor requirements, limited availability of vectors or test plants, or whatever. These $\hat{p}$s too can be from data having different values of N and k.

Designing an experiment to obtain a $\hat{p}$ involves choosing appropriate values of N and k, the numbers of test plants and vectors per test plant, respectively. This is complicated by the fact that, when $k > 1$, $\hat{p}$ is a biased estimator ($\hat{p}$ overestimates p). The magnitude of the bias depends greatly on the values of N and k chosen and on p; the bias may be negligible or huge, depending on the design selected. Swallow (5) discussed this problem in some detail and showed the importance of comparing $\hat{p}$s from different possible designs (choices of N and k) through their mean squared errors (MSEs), not through their variances or standard errors as is usually and appropriately done with unbiased estimators. The MSE of the biased estimator $\hat{p}$ is

$$\text{MSE}(\hat{p}) = \text{Variance}(\hat{p}) + (\text{Bias}[\hat{p}])^2,$$

which incorporates measures of both the accuracy (bias) and precision (variance) of the estimator. To have a small MSE, the estimator must be both accurate (small bias) and precise (small variance). Work on experimental design for group testing, that is, on how to select appropriate values of N and k, has focused on how to choose designs with small MSE($\hat{p}$)s.

In considering the problem of design selection, Swallow (5) reached three important conclusions: First, in practice, a multiple-vector-transfer design ($k > 1$) almost always will be preferable to a

single-vector-transfer design ($k = 1$). For specified N, the optimal group-testing design (i.e., value of k that minimizes MSE[$\hat{p}$]) often provides an estimator of p with MSE manyfold smaller than would be realized with the same value of N but with $k = 1$. The advantage of group testing is most striking when p is very small, as it is in most applications. Second, the user must take care in selecting a group-testing design, as the penalties for a poor choice can be severe. Specifically, using too large a value of k (how large is too large depends on N and p) can lead to an estimator with large bias and thereby greatly inflated MSE. And third, in most practical applications it is easy to select a group-testing design that, though not optimal, is both safe (having acceptable, usually negligible, bias) and an improvement over a single-vector-transfer design. To determine the optimal value of k for a given N, one must know beforehand the value of the unknown parameter p that one is preparing to estimate—clearly an impossible requirement. The recommended approach, in practice, is to select a value of k using, in place of the unknown p, a value that one believes, perhaps based on preliminary or other data, to be an upper bound for p; two examples illustrating how a researcher might do this are provided in the following section, and further discussion and more extensive tables for use in this procedure can be found in (5). The value of k one selects in this way is expected to be smaller than the optimal value of k, providing insurance against excessive bias, but it has been assumed that it will still provide some reasonable (but unknown) fraction of the gains of the optimal group-testing design over a single-vector-transfer design.

Thus, although the argument for using group testing rather than single-vector-transfer designs in most applications has been based on the greater efficiency (i.e., smaller MSE) of the *optimal* group-testing designs (1,2,5,6), in application the optimal design cannot be used or even identified. Instead, using a value of k presumed to be smaller than the optimal value is recommended for insurance against excessive bias, and therefore only a portion of the potential benefits (MSE reduction) of the optimal group-testing design over the single-vector-transfer design will be realized. This raises important questions that need to be addressed. First, what fraction of the (often huge) MSE reduction of an optimal group-testing design over the single-vector-transfer design can one hope to realize in practice when using smaller-than-optimal values of k? And second, with a view of designing cost-efficient experiments, can savings in labor and other costs be taken into account as further reason to use values of k smaller than the values needed to minimize MSE($\hat{p}$)? This paper addresses these questions in the belief that the answers will enhance the appeal of group testing to potential users.

RELATIVE MEAN SQUARED ERRORS AND STANDARD ERRORS WHEN USING SMALLER-THAN-OPTIMAL VALUES OF k VERSUS THE OPTIMAL VALUE, k^*

Relative MSE (RMSE) is a convenient measure for comparing the MSEs of estimates ($\hat{p}$s) of the same p that could be obtained using different values of k with the same value of N, that is, using different experimental designs. We denote the MSE of $\hat{p}$ when k vectors per test plant are used as MSE($\hat{p};k$), and the MSE of $\hat{p}$ when k^*, the optimal group size, are used as MSE($\hat{p};k^*$). Expressed as a percent, the RMSE obtained with k versus k^* vectors per test plant is

$$\mathrm{RMSE}(k) = \left[\mathrm{MSE}(\hat{p};k)/\mathrm{MSE}(\hat{p};k)\right] \times 100. \qquad (1)$$

Calculated values of RMSE, like those of MSE, depend on N and p as well as on k, although, to simplify notation, N and p are not shown explicitly in equation 1. The notation k_{125}, for example, will be used to denote the value of k for which RMSE(k) = 125; that is, k_{125} is the value of k for which MSE($\hat{p};k$) is 25% larger than the minimum MSE realized with k^* for the same values of N and p.

Table 1 gives k^*, MSE($\hat{p};k^*$), k_{125}, k_{150}, k_{200}, k_{400}, k_{900}, and RMSE(1) for a wide range of combinations of N and p. RMSE(1) compares the MSE of the single-vector-transfer design ($k = 1$) with the MSE of the optimal group-testing design. The table is based on mean squares calculated for $k = 1$ to 25 by 1, and 25 to 50 by 5. Thus, recorded values of k^* greater than 25 are correct to within 5, except that $k^* = 50$ means only $k^* \geq 50$, since 50 was the largest value considered. RMSEs and the values of k associated with specified RMSEs are therefore approximate for $k^* \geq 25$, but are fine for all practical purposes.

Table 1 shows, for example, that for $p = 0.04$ and $N = 25$, the optimal value of k is $k^* = 22$ for which MSE($\hat{p};k^*$) = 0.000131, the minimum MSE against which one can compare MSEs of $\hat{p}$s obtained with other values of k. If one uses $k = 8$ (shown as k_{200} in Table 1), then the MSE of $\hat{p}$, MSE($\hat{p}$;8), will be about twice the minimum MSE (by the definition of k_{200}). With $k = 4$ (k_{400}), MSE($\hat{p}$;4) will be four times the minimum MSE, MSE($\hat{p};k^*$), but to attain that minimum one would require a group size, k^*, which is $(k^*/k) = (22/4) = 5.5$ times the value $k = 4$. For the same example, Table 1 gives RMSE(1) = 1,175, meaning that MSE($\hat{p}$;1) from the single-vector-transfer design would be 11.75 times MSE($\hat{p};k^*$) from the optimal group-testing design.

Table 1 can be used to support the method suggested by Swallow (5) and Thompson (6) for choosing a group size (k) in practice when N is already determined, perhaps by availability of screen cages or glasshouse space. They recommended using a group size that would be optimal for a value of p thought to be a reasonable upper bound for the (unknown) p that was to be estimated. For example, when $N = 25$ and the value of the true p is 0.05, the optimal group size is $k^* = 18$ (from Table 1), but to determine $k^* = 18$ one needs to know that the true $p = 0.05$. If one could just say, "I'm quite certain that p is not larger than 0.10," then one could enter Table 1 at $p_e = 0.10$ (p_e for "entry p") and $N = 25$, and find $k = 9$ (shown there as k^*). Returning to the section of Table 1 for $p = 0.05$ (the true value of p) and $N = 25$, one finds that $k = 9$ is the value of k_{150}. That is, RMSE($\hat{p}$;9) = 150. Specifying the upper bound for p as $p_e = 0.10$ leads to using $k = 9$ for which MSE($\hat{p}$;9) is 50% larger than the minimum MSE, MSE($\hat{p};k^*$). For comparison, RMSE(1) = 951, so using $k = 9$ as in this example provides most of the potential gains of group testing over the single-vector-transfer design.

For a second example, suppose one has decided to use $N = 25$, and the true $p = 0.03$. The optimum group size is given as $k^* = 30$. Using $p_e = 0.05$ would give $k = 18$, which is k_{125} under $p = 0.03$. Using $p_e = 0.10$ would give $k = 9$, which is approximately k_{200} under $p = 0.03$. And using $p_e = 0.20$ would give $k = 5$, which is k_{400} under $p = 0.03$. Again for comparison, RMSE(1) = 1,539. The closer p_e is to the true value of p, that is, the more tightly one can bound p, the closer the resulting design (choice of k) will be to the optimal (minimum MSE) design. But even using an extremely conservative (safe) upper bound, $p_e = 0.20$, yields a design (value of k) for which MSE($\hat{p};k$) is a quarter of that for the single-vector-transfer design (RMSE = 400 vs. 1,539). Of course, one must understand that the comparisons of MSEs made in the preceding examples cannot be made in a particular application, since they required knowing p. Still, they add credibility to the approach recommended by Swallow (5) and Thompson (6) for selecting group size in practice.

More generally, however, perusing the values of RMSE(1) and k^* shown in Table 1 leads to two basic observations. First, for fixed N and p, and especially when p is small, RMSE(1) is likely to be large. That is, the optimal group-testing design is usually far superior to the single-vector-transfer design, judged by comparison of MSE($\hat{p}$)s. This has been noted by many authors (1,2,5,6) and has been the basis for their recommending group testing. Second, again especially when p is small, the optimal group size, k^*, may be very large. Thus, the potential gains (MSE reduction) with group testing are large, but a very large group size may be needed to realize group testing's full potential. In many biological applications, the thought of using such large values of k may raise serious questions of cost, practicality, and whether the assumption (for $\hat{p}$) that the vectors operate independently can be justified (5). This suggests investigating the merits of designs that, although they may realize only a fraction of the benefits of the optimal group-testing design, will require a group size that is only a fraction of the group size, k^*, required for that optimal design.

Indeed, it begs the question, "Even if we could know k^*, would we want to use such a large group size?"

As rough approximations, $k_{125} = (2/3)k^*$, $k_{150} = (1/2)k^*$, $k_{200} = (1/3)k^*$, and $k_{400} = (1/6)k^*$ in Table 1. So, for example, if one reduces the group size by one third, from k^* to $(2/3)k^*$, one increases MSE($\hat{p}$) by only 25%. The proportional reduction in group size exceeds the proportional increase in MSE($\hat{p}$).

In considering the effects of increased sample size, we often focus on the standard error (SE) of an estimator, the square root of its variance. It's natural to do so, because differences in standard errors determine differences in widths of confidence intervals, least significant differences, and so forth, for many familiar estimators. In the previous section of this paper, I argued that the biased estimator $\hat{p}$ should be evaluated through its MSE, not its variance, and, accordingly, Table 1 is based on calculated MSEs. In fact, however, for the optimal group-testing designs of Table 1, the MSE is approximately 1% squared bias and 99% variance (see Table 1 of Swallow [5]). Designs using smaller group sizes than that of the optimal design (for the same N and p) have even smaller biases. The group-testing designs that were disqualified for having serious bias (and, thereby, large MSE), and that therefore do not appear in Table 1, all use larger-than-optimal group size. Thus, the MSE($\hat{p}$)s in Table 1 are essentially the variances of those $\hat{p}$s, and the square roots of those MSE($\hat{p}$)s are approximate standard errors.

By analogy to equation 1, relative standard error (RSE) is taken to be the ratio of the standard error of $\hat{p}$ based on group size k (SE[$\hat{p};k$]) to the standard error of $\hat{p}$ from the optimal group-testing design (SE[$\hat{p};k^*$]) for the same N and p and is expressed as a percent; to obtain the RSE one can take the square root of RMSE/100 and multiply the result by 100. Converting from RMSEs to RSEs shows that the group sizes designated k_{125}, k_{150}, k_{200}, k_{400}, and k_{900} are those that have relative standard errors approximately equal to 110, 120, 140, 200, and 300, respectively. If the approximations given two paragraphs above are stated in terms of relative standard errors, rather than RMSEs, they indicate that reducing group size by one third from k^* to $(2/3)k^*$ increases the standard error of $\hat{p}$ by approximately 10%, reducing from k^* to $(1/2)k^*$ increases the SE by 20%, reducing from k^* to $(1/3)k^*$ increases the SE by 40%, and reducing from k^* to $(1/6)k^*$ increases the SE by 100% (that is, doubles it). Thus, when the quality of $\hat{p}$ is judged by its approximate standard error rather than its MSE, it is even more striking that proportional reduction in group size from k^* exceeds the proportional loss in quality of $\hat{p}$. This too makes it tempting to consider group-testing designs using smaller-than-optimal group size. And the case for using smaller-

TABLE 1. Optimal group size or number of vectors per plant (k^*) to use in estimating infection rate or probability (p) of disease transmission by a single vector for specified numbers of tests or test plants (N), the mean squared error [MSE($\hat{p};k^*$)] of the estimator of p using k^*, the relative mean squared error (RMSE[1] = 100 × MSE[$\hat{p}$; 1]/ MSE[$\hat{p};k^*$]) using $k = 1$ versus k^*, and values of k [k_{125}, k_{150}, k_{200}, k_{400}, k_{900}] having relative mean squared errors 125, 150, 200, 400, and 900% of MSE($\hat{p};k^*$)

		N										
P		10	15	20	25	30	40	50	60	80	100	200
0.01	MSE($\hat{p};k^*$)	0.000046	0.000021	0.000014	0.000011	0.000009	0.000007	0.000005	0.000004	0.000003	0.000003	0.000001
	k^*	35	50	50	50	50	50	50	50	50	50	50
	k_{125}	23	35	40	40	40	40	40	40	40	40	40
	k_{150}	18	30	30	30	30	30	30	30	30	30	30
	k_{200}	13	19	21	21	21	21	22	22	22	22	22
	k_{400}	7	9	10	10	10	10	10	10	10	11	11
	k_{900}	3	4	5	5	5	5	5	5	5	5	5
	RMSE(1)	2,148	3,165	3,450	3,538	3,594	3,664	3,705	3,733	3,767	3,787	3,828
0.02	MSE($\hat{p};k^*$)	0.000162	0.000078	0.000048	0.000035	0.000027	0.000019	0.000015	0.000012	0.000009	0.000007	0.000003
	k^*	19	30	35	40	45	50	50	50	50	50	50
	k_{125}	13	19	23	30	30	35	35	35	35	35	35
	k_{150}	11	15	18	20	22	24	24	24	25	25	25
	k_{200}	8	11	13	14	15	16	17	17	17	17	18
	k_{400}	4	5	6	7	7	8	8	8	8	8	8
	k_{900}	2	2	3	3	3	4	4	4	4	4	4
	RMSE(1)	1,210	1,674	2,036	2,264	2,434	2,623	2,686	2,727	2,777	2,806	2,865
0.03	MSE($\hat{p};k^*$)	0,000337	0.000164	0.000104	0.000076	0.000059	0.000041	0.000031	0.000025	0.000018	0.000015	0.000007
	k^*	14	20	25	30	30	35	40	45	45	45	50
	k_{125}	10	13	16	18	19	21	23	23	24	24	25
	k_{150}	8	11	13	14	15	16	17	18	18	18	19
	k_{200}	6	8	9	10	10	11	12	12	12	13	13
	k_{400}	3	4	4	5	5	5	6	6	6	6	6
	k_{900}	1	2	2	2	2	3	3	3	3	3	3
	RMSE (1)	863	1,181	1,396	1,539	1,635	1,774	1,862	1,912	1,970	2,003	2,080
0.04	MSE($\hat{p};k^*$)	0.000565	0.000281	0.000180	0.000131	0.000102	0.000072	0.000055	0.000045	0.000032	0.000026	0.000012
	k^*	11	16	19	22	25	30	30	35	35	35	35
	k_{125}	8	10	12	14	15	16	17	17	18	18	19
	k_{150}	6	8	10	11	11	12	13	13	14	14	14
	k_{200}	4	6	7	8	8	9	9	9	9	10	10
	k_{400}	2	3	3	4	4	4	4	4	4	5	5
	k_{900}	1	2	2	2	2	2	2	2	2	2	2
	RMSE(1)	680	911	1,068	1,175	1,252	1,342	1,400	1,432	1,479	1,506	1,558
0.05	MSE($\hat{p};k^*$)	0.000842	0.000424	0.000274	0.000200	0.000157	0.000110	0.000084	0.000069	0.000050	0.000040	0.000019
	k^*	9	13	16	18	20	23	25	25	25	30	30
	k_{125}	6	9	10	11	12	13	14	14	14	15	15
	k_{150}	5	7	8	9	9	10	10	11	11	11	12
	k_{200}	4	5	6	6	7	7	7	7	8	8	8
	k_{400}	2	2	3	3	3	3	4	4	4	4	4
	k_{900}	1	1	1	2	2	2	2	2	2	2	2
	RMSE(1)	564	747	868	951	1,009	1,083	1,125	1,150	1,179	1,202	1,249

(CONTINUED)

than-optimal group size may be even stronger when costs are taken into account, as illustrated in the following section.

MINIMIZING COST PER UNIT OF INFORMATION

Although the literature on multiple-vector-transfer designs has focused on choosing designs to minimize MSE($\hat{p}$), one might wish to explicitly introduce cost considerations, too, and redefine what one means by "optimal" design. In that case, one's goal might shift from minimizing MSE($\hat{p}$) to minimizing cost per unit of information or, equivalently, maximizing information per unit of cost.

In experimental design, "information" is often defined to be the reciprocal of variance. As noted in the preceding section, for the minimum-MSE group-testing designs (using k^*) of Table 1 or designs using values of k smaller than k^* with the same N and p, the MSEs shown are essentially the variances of $\hat{p}$s from those designs. Thus, for any of the designs in Table 1, the reciprocal of the MSE is approximately the reciprocal of the variance, and will be used as our measure of information. For those designs,

$$\text{cost per unit of information} = \text{cost}/(1/\text{MSE}) = \text{cost} \times \text{MSE}$$

is a reasonable approximation.

One might view the cost of a vector-transfer design as coming from two components: C_1, the cost of each of the N tests, and C_2, the cost of each of the $N \times k$ vectors required. The kinds of costs entering into C_1 might include costs of screen cage or glasshouse space for test plants during the experiment, costs of growing or obtaining the test plants, and costs of laboratory tests to finally classify test plants as healthy or infected. C_1 absorbs all costs that are independent of k. Costs contributing to C_2 might include costs of rearing or otherwise obtaining vectors, and costs of collecting, handling, and transferring vectors, including labor. The total cost would then be

$$\text{cost} = N \times C_1 + N \times k \times C_2.$$

Table 1 cont.

		N										
P		10	15	20	25	30	40	50	60	80	100	200
0.06	MSE($\hat{p};k^*$)	0.001158	0.000592	0.000385	0.000282	0.000222	0.000156	0.000120	0.000098	0.000071	0.000056	0.000027
	k^*	8	11	13	15	17	19	21	22	23	23	24
	k_{125}	6	7	9	10	10	11	11	12	12	12	13
	k_{150}	5	6	7	7	8	8	9	9	9	9	10
	k_{200}	3	4	5	5	6	6	6	6	6	6	7
	k_{400}	2	2	2	3	3	3	3	3	3	3	3
	k_{900}	1	1	1	1	1	2	2	2	2	2	2
	RMSE(1)	487	635	732	799	847	905	940	961	989	1,006	1,041
0.08	MSE($\hat{p};k^*$)	0.001922	0.001002	0.000656	0.000484	0.000382	0.000269	0.000208	0.000170	0.000124	0.000097	0.000047
	k^*	6	9	10	12	13	15	16	16	17	17	18
	k_{125}	4	6	7	7	8	8	9	9	9	9	10
	k_{150}	4	5	5	6	6	6	7	7	7	7	7
	k_{200}	3	3	4	4	4	5	5	5	5	5	5
	k_{400}	1	2	2	2	2	2	2	2	2	2	3
	k_{900}	...	1	1	1	1	1	1	1	1	1	1
	RMSE(1)	383	490	561	609	642	683	707	724	744	756	782
0.10	MSE($\hat{p};k^*$)	0.002807	0.001482	0.000987	0.000732	0.000579	0.000409	0.000317	0.000258	0.000189	0.000149	0.000072
	k^*	5	7	8	9	10	12	12	13	13	14	14
	k_{125}	4	5	5	6	6	7	7	7	7	7	8
	k_{150}	3	4	4	5	5	5	5	5	6	6	6
	k_{200}	2	3	3	3	3	4	4	4	4	4	4
	k_{400}	1	2	2	2	2	2	2	2	2	2	2
	k_{900}	...	1	1	1	1	1	1	1	1	1	1
	RMSE(1)	321	405	456	492	518	550	568	581	596	606	626
0.15	MSE($\hat{p};k^*$)	0.005409	0.002976	0.002014	0.001516	0.001202	0.000858	0.000665	0.000544	0.000398	0.000314	0.000152
	k^*	4	5	6	6	7	8	8	8	9	9	9
	k_{125}	3	3	4	4	4	5	5	5	5	5	5
	k_{150}	2	3	3	3	3	4	4	4	4	4	4
	k_{200}	2	2	2	2	2	3	3	3	3	3	3
	k_{400}	1	1	1	1	1	1	1	1	2	2	2
	k_{900}	...	...	...	...	...	...	...	...	1	1	1
	RMSE(1)	236	286	317	337	353	372	383	390	400	407	419
0.20	MSE($\hat{p};k^*$)	0.008356	0.004764	0.003284	0.002459	0.001975	0.001416	0.001100	0.000901	0.000662	0.000523	0.000253
	k^*	3	4	4	5	5	6	6	6	6	7	7
	k_{125}	2	3	3	3	3	3	4	4	4	4	4
	k_{150}	2	2	2	3	3	3	3	3	3	3	3
	k_{200}	1	2	2	2	2	2	2	2	2	2	2
	k_{400}	...	1	1	1	1	1	1	1	1	1	1
	k_{900}	...	...	...	...	...	...	...	...	...	...	...
	RMSE(1)	192	224	244	260	270	283	291	296	302	306	316
0.25	MSE($\hat{p};k^*$)	0.012089	0.006652	0.004735	0.003514	0.002831	0.002056	0.001597	0.001306	0.000959	0.000757	0.00370
	k^*	3	3	3	4	4	4	5	5	5	5	5
	k_{125}	2	2	2	3	3	3	3	3	3	3	3
	k_{150}	2	2	2	2	2	2	2	2	2	2	2
	k_{200}	1	1	1	2	2	2	2	2	2	2	2
	k_{400}	...	...	...	1	1	1	1	1	1	1	1
	k_{900}	...	...	...	...	...	...	...	...	...	...	...
	RMSE(1)	155	188	198	214	221	228	235	239	244	248	254

One could consider more complicated ways to account for costs, but that isn't necessary for what follows below. What matters is that one incorporate some measure of the relative costs of a test (plant) and of a vector.

In practice, costs generally increase with increased k for a given N. This, together with the diminishing gains (MSE or SE reduction) with larger k noted in the previous section, means that values of k smaller than k^* may be most cost effective, that is, may minimize cost per unit of information. Of course, for any given N and p, values of k larger than k^* cannot be cost effective and need not be considered; they will yield less information (larger MSE) than will k^* at greater cost.

Figure 1 illustrates the relationship between costs and cost-effective group size for the case wherein $N = 25$ and $p = 0.05$. For this case, $k^* = 18$ (from Table 1), so only group sizes (values of k) to 18 are of interest and shown on the horizontal axis. Cost per unit of information, calculated as

$$\text{cost per unit of information} = (N \times C_1 + N \times k \times C_2) \times \text{MSE}, \quad (2)$$

is shown on the vertical axis. The value of MSE in equation 2 depends on N, k, and p. For specified N ($N = 25$ here), finding the value of k that minimizes cost per unit of information of equation 2 is equivalent to finding the value that minimizes

$$(C_1 + k \times C_2) \times \text{MSE},$$

or

$$(1 + k \times [C_2/C_1]) \times \text{MSE}.$$

In other words, the value of k that minimizes cost per unit of information is determined by C_1 and C_2 only through the ratio C_2/C_1. Thus, one may as well set $C_1 = 1$ and let $C_2 = C_2/C_1$ take on several different values for comparison. In that case, taking $C_2 = C_2/C_1 = 0.05$, say, means that the cost of a vector is taken to be 5% of the cost of a test; other values of C_2 can be interpreted similarly. Illustrating the calculation of cost per unit of information: for $N = 25$, $p = 0.05$, $C_1 = 1$, and $C_2 = C_2/C_1 = 0.05$, equation 2 gives the cost per unit of information when using groups of size $k^* = 18$ to be $([25 \times]) + [25 \times 18 \times 0.05]) \times 0.000200 = 0.0095$, as plotted on Figure 1. The value MSE = 0.000200 is given in Table 1 as $\text{MSE}(\hat{p};k^*)$ for $N = 25$ and $p = 0.05$. If, instead, $C_2 = 0.10$ and one uses $k = 6$, the cost per unit of information will be approximately $([25 \times 1] \times [25 \times 6 \times 0.10]) \times 0.000400 = 0.016$. The value MSE = 0.000400 in this calculation is obtained from Table 1 by noting that the value $k = 6$ is shown as k_{200}, and thus $\text{MSE}(\hat{p};6) = 2 \times \text{MSE}(\hat{p};k^*) = 2 \times 0.000200 = 0.000400$ (approximately). The values of cost per unit of information actually used in plotting Figure 1 are exact, being based on more complete tables than Table 1.

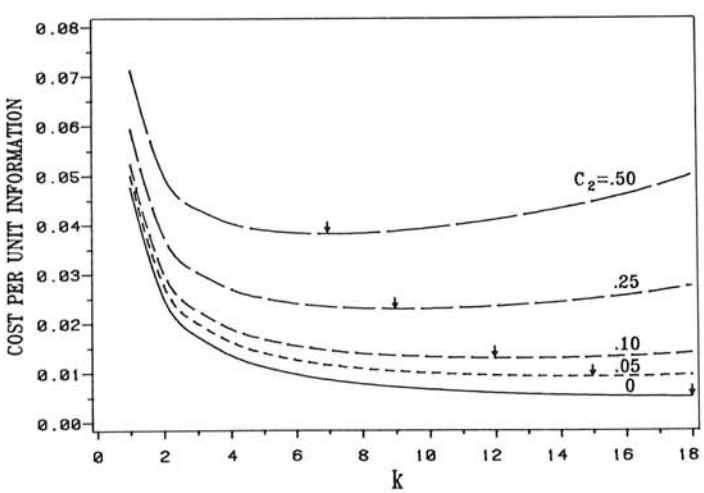

Fig 1. Cost per unit of information in estimating an infection rate or probability (p) of disease transmission by a single vector, versus the number of vectors (k) used on each of $N = 25$ test plants, when $p = 0.05$, the cost per test plant is $C_1 = 1$, and the cost per vector is $C_2 = 0, 0.05, 0.10, 0.25$, or 0.50.

Figure 1 shows cost per unit of information plotted against k for $C_1 = 1$ and $C_2 = 0, 0.05, 0.10, 0.25$, and 0.50. Each value of $C_2 = C_2/C_1$ determines a different curve, and the value of k that minimizes cost per unit of information is indicated along that curve by a small arrow. Provided vectors are free ($C_2 = 0$), k^* is always most cost-effective ($k^* = 18$ here). But as vectors become more costly, smaller and smaller values of k become preferable. For $C_2 = 0.05$ (the cost of a vector being 5% of the cost of a test plant), $k = 15$ minimizes cost per unit of information. For $C_2 = 0.10$, $k = 12$ minimizes; for $C_2 = 0.25$, $k = 9$ minimizes; and for $C_2 = 0.50$, $k = 7$ minimizes cost per unit of information. As a practical matter, even smaller values of k nearly minimize cost per unit of information. For $C_2 = 0$, $k^* = 18$, and $k = 10$ differ little in cost per unit of information. For $C_2 = 0.05$, the curve is quite flat down to $k = 8$, and for $C_2 = 0.10$, down to $k = 6$.

The case wherein $N = 25$ and $p = 0.05$ is shown (Fig. 1) is a typical example. Similar figures (not shown) were drawn for all nine combinations of $N = 10, 25$, or 50 with $p = 0.01, 0.05$, or 0.10. The general impression from each of these figures is the same, although the details depend on N and p. For fixed N, the smaller the value of p, the more exaggerated is the flatness evident in the curves (as discussed in the final three sentences of the previous paragraph). For fixed p, as N is made larger, the minima (marked by arrows in Fig. 1) become smaller fractions of k^*, and the curves rise more steeply to the right of those minima.

The fact that, in practice, it may be difficult to state costs precisely doesn't void the value of the message in Figure 1 or similar figures. The point is that, for given N and p, using a value of k smaller than the k^* from the optimal (meaning minimum MSE) design may make sense from the standpoint of cost effectiveness. The value k^* is optimal for achieving the statistical goal of MSE minimization with specified N but without regard for cost considerations.

DISCUSSION

Although this paper has discussed group testing largely in the context of vector-transfer designs, the topic is much broader. Group testing is often useful, for example, in batch serological testing where samples from k plants (or blood from k individuals, or whatever) can be homogenized and tested together. The infection rate in a field (or other population) can then be estimated from N laboratory tests, each done on a pool of size k. The problem is the same—how to choose a suitable value for k. The points made here and in (5) are as applicable to batch testing as to vector-transfer designs.

None of the above resolves the fundamental difficulty in choosing a group-testing design, namely, that to choose an optimal design one must know in advance the true value of the p that the proposed experiment is to estimate. For the case where N is taken to be fixed and the goal is to choose k to minimize $\text{MSE}(\hat{p})$, this dilemma led Swallow (5) and Thompson (6) to recommend the approach to design selection discussed and illustrated above. Their procedure is expected to lead to using a group size (k) smaller than the value k^* that is optimal in the sense that it minimizes the mean squared error of $\hat{p}$. However, values of k larger than k^* are avoided, which is essential because using values larger than k^* can lead to badly biased estimates. And using smaller group sizes makes the important assumption that vectors operate independently more plausible in many applications.

This paper provides additional support for the method recommended by Swallow (5) and Thompson (6) for choosing k. It demonstrates that most of the advantage (in MSE reduction) of group-testing or multiple-vector-transfer designs over single-vector-transfer designs can indeed be realized with group sizes substantially smaller than k^*. It also illustrates how, when the cost of increasing group size is not negligible, using group sizes smaller than k^* may even be more cost-effective than using k^* itself. It is hoped that, with this information, prospective users will be further encouraged to use group testing rather than much less powerful single-vector-transfer designs whenever group testing is appropriate and feasible.

LITERATURE CITED

1. Gibbs, A. J., and Gower, J. C. 1960. The use of a multiple-transfer method in plant virus transmission studies—Some statistical points arising in the analysis of results. Ann. Appl. Biol. 48:75-83.
2. Loyer, M. W. 1983. Bad probability, good statistics, and group testing for binomial estimation. Am. Stat. 37:57-59.
3. Romanow, L. R., Moyer, J. W., and Kennedy, G. G. 1986. Alteration of efficiencies of acquisition and inoculation of watermelon mosaic virus 2 by plant resistance to the virus and to an aphid vector. Phytopathology 76:1276-1281.
4. Smith, K. M., and Lea, D. E. 1946. The transmission of plant viruses by aphides. Parasitology 37:25-37.
5. Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75:882-889.
6. Thompson, K.H. 1962. Estimation of the proportion of vectors in a natural population of insect. Biometrics 18:568-578.
7. Watson, M. A. 1936. Factors affecting the amount of infection obtained by aphis transmission of the virus Hy. III. Philos. Trans. Roy. Soc. London, Ser. B 226:457-489.