THE EFFECT OF INCOMPLETE PENETRANCE ON THE ESTIMATION OF RECOMBINATION VALUES

R. W. ALLARD and H. L. ALDER
University of California, Davis, California

Received 1.v.60

1. INTRODUCTION

IN testcrosses or intercrosses involving double heterozygotes the most common disturbances to the expected 1:1:1:1 or 9:3:3:1 ratios are caused by: (1) linkage between the two loci; (2) failures of manifestation that lead to assigning a proportion of individuals to phenotypic classes inappropriate to their genotypes; and (3) upsets in the formation or function of gametes or zygotes, leading to differential viability. Linkage can be distinguished from the other two sources of disturbance because it upsets the joint distributions without affecting the single-gene ratios.

In experiments designed to estimate the linkage parameter, interest in the other parameters entering the expectations is generally limited to consideration of the disturbing effects they may have on estimation of the recombination fraction. There are two main methods of dealing with disturbances in linkage experiments. The first method is to set up conditions so that penetrance is complete and viability equivalent in all classes. This method cannot of course be used when there is no known way of eliminating difficulties of incomplete penetrance or differential viability. The second method is to develop statistical techniques that minimise the effects of the disturbing factors (Bailey, 1949, 1950; Sanchez-Monge, 1952, and Parsons, 1957). The present paper extends the methods developed by these investigators, and tests their efficiency by applying them to data from families in which disturbances attributable to penetrance are severe in one stage of development of the experimental organism, and negligible or absent in another stage.

2. TESTCROSSES

(a) *Undisturbed data*

When viability is unaffected and penetrance is complete, the expected proportions in the offspring of the testcross $ab \times \frac{AB}{ab}$ are as given in table 1 under $u = 1$. The likelihood of obtaining a given family with observed frequencies a, b, c and d is

$$\frac{n!}{a!b!c!d!} m_1^a m_2^b m_3^c m_4^d, \quad (1)$$

in which m_1, m_2, m_3 and m_4 represent the expected proportions corresponding to a, b, c and d . If L is the logarithm of this likelihood

expression, an efficient estimate of p , the linkage parameter, is obtained by maximising L with respect to p , giving

$$\hat{p} = \frac{b+c}{n}. \quad (2)$$

Strictly speaking, $\hat{p}$ or some similar notation should be used in formula (2) for p to distinguish the estimate of p from the parameter p . Such distinction will not be made in this paper for p or for other parameters since it should always be clear which is meant. The expected variance of p is given by

$$V_p = \frac{1}{-E\left(\frac{d^2L}{dp^2}\right)} = \frac{p(1-p)}{n}. \quad (3)$$

An estimate of this variance is given by

$$V_p = \frac{(a+d)(b+c)}{n^3}, \quad (4)$$

which is obtained by substituting in formula (3) the value of p obtained from formula (2).

TABLE 1

Observed and expected frequencies for backcross data

Phenotype	Observed frequencies	Expected frequencies $\times \frac{1}{n}$	
		$u = 1$	$u \neq 1$
<i>AB</i>	<i>a</i>	$m_1 = \frac{1}{2}(1-p)$	$m_5 = \frac{1}{2}(1-p)u$
<i>Ab</i>	<i>b</i>	$m_2 = \frac{1}{2}p$	$m_6 = \frac{1}{2}pu$
<i>aB</i>	<i>c</i>	$m_3 = \frac{1}{2}p$	$m_7 = \frac{1}{2}[p + (1-p)(1-u)]$
<i>ab</i>	<i>d</i>	$m_4 = \frac{1}{2}(1-p)$	$m_8 = \frac{1}{2}[(1-p) + p(1-u)]$
Total	<i>n</i>	1	1

The test for linkage in digenic backcrosses with undisturbed viability and complete penetrance is

$$\chi^2_{111} = \frac{(a-b-c+d)^2}{n}. \quad (5)$$

(b) *Incomplete penetrance*

For a coupling phase backcross in which viability is unaffected but a fraction $1-u$ of the *Aa* genotype is classified phenotypically as *aa*, expectations can be drawn up as in table 1 under $u \neq 1$. The likelihood of obtaining an observed family is

$$\frac{n!}{a!b!c!d!} m_5^a m_6^b m_7^c m_8^d. \quad (6)$$

Using maximum likelihood methods as given by Bailey (1950) and Sanchez-Monge (1952), we obtain the following estimates of p and u :

$$u = \frac{a}{a+c} + \frac{b}{b+d} \quad (7)$$

and

$$p = \frac{b(a+c)}{a(b+d) + b(a+c)}. \quad (8)$$

The expected variance of u is approximated by (see Fisher, 1954)

$$V_u = \Sigma \left[np \left(\frac{\delta u}{\delta a} \right)^2 \right] - n \left(\frac{\delta u}{\delta n} \right)^2 = \frac{2u}{n} [1 - u + 2pu - 2p^2u]. \quad (9)$$

This expression, which can also be written as

$$\frac{2u}{n} \left\{ 2p(1-p) + (1-u)[1 - 2p(1-p)] \right\},$$

is equivalent to Bailey's formula (3.5). It is a corrected version of Sanchez-Monge's formula (10).*

To find the variance of u obtained from information realised in a particular family, let

$$\alpha = \frac{a}{a+c}, \quad \beta = \frac{b}{b+d}.$$

Then

$$u = \alpha + \beta$$

and

$$V_u = V_{\alpha+\beta} = V_\alpha + V_\beta.$$

Since α and β can be presumed to be independent, the covariance term in $V_{\alpha+\beta}$ is zero. Now

$$V_\alpha = \frac{\alpha(1-\alpha)}{a+c} = \frac{ac}{(a+c)^3},$$

$$V_\beta = \frac{\beta(1-\beta)}{b+d} = \frac{bd}{(b+d)^3},$$

so that

$$V_u = \frac{ac}{(a+c)^3} + \frac{bd}{(b+d)^3}. \quad (10)$$

The estimate of V_u given by (9) for the values of u and p obtained from (7) and (8) will usually not be greatly different from the estimate of V_u given by (10), but these two estimates will be identical only when $a+c = b+d = \frac{n}{2}$.

The expected variance of p is obtained as

$$V_p = \frac{2}{n} \frac{p(1-p)(1-2pu+2p^2u)}{u}. \quad (11)$$

* Many of the formulæ in this paper replace formulæ given incorrectly by Sanchez-Monge. The errors are being corrected here with the approval of Mr Sanchez-Monge, whose cooperation is herewith gratefully acknowledged.

This expression, which can also be written as

$$\frac{2p(1-p)}{n} \left[\frac{1}{u} - 2p(1-p) \right],$$

is equivalent to Bailey's formula (3.6), and corrects formula (12) of Sanchez-Monge.

To find the variance of p obtained from information realised in a particular family, we rewrite (8) as

$$p = \frac{\beta}{\alpha + \beta}.$$

Then

$$V_p = \left(\frac{\partial p}{\partial \alpha} \right)^2 V_\alpha + \left(\frac{\partial p}{\partial \beta} \right)^2 V_\beta + 2 \frac{\partial p}{\partial \alpha} \cdot \frac{\partial p}{\partial \beta} \cdot V_{\alpha, \beta}$$

or, if we assume that α is uncorrelated with β ,

$$\begin{aligned} V_p &= \frac{\beta^2 V_\alpha + \alpha^2 V_\beta}{(\alpha + \beta)^4} \\ &= \frac{(a+c)(b+d)[ab^2c(b+d) + a^2bd(a+c)]}{[a(b+d) + b(a+c)]^4}. \end{aligned} \quad (12)$$

The estimate of V_p given by equation (11) for the values of u and p obtained from (7) and (8) will ordinarily not differ greatly from the estimate of V_p given by (12), but the two estimates will be identical only when $a = d$ and $b = c$.

To test the significance of the departure of u from unity, correcting for linkage, we assume $u = 1$, for which case the maximum likelihood estimate of p is

$$p = \frac{b+c}{n},$$

so the expected frequencies are

$$\frac{1}{2}(a+d), \frac{1}{2}(b+c), \frac{1}{2}(b+c), \frac{1}{2}(a+d).$$

Comparing these expectations with the observed frequencies a, b, c and d , we obtain

$$\chi^2_{[12]} = \frac{(a-d)^2}{a+d} + \frac{(b-c)^2}{b+c}, \quad (13)$$

which is based on two degrees of freedom, since, of the three originally available, one has been used to estimate p . The value of $\chi^2_{[12]}$ given by (13) can be considered as the sum of two values of χ^2 , each based on one degree of freedom, in which the first tests for the departure of the Bb segregation from the 1 : 1 ratio, *i.e.*,

$$\chi^2_{[11]} = \frac{[(a+c) - (b+d)]^2}{n}, \quad (14)$$

and the second tests for the departure of u from unity, so that

$$\begin{aligned}\chi^2_{[1]} &= \frac{(a-d)^2}{a+d} + \frac{(b-c)^2}{b+c} - \frac{[(a+c)-(b+d)]^2}{n} \\ &= \frac{4(ab-cd)^2}{n(a+d)(b+c)}.\end{aligned}\quad (15)$$

Formula (15) is identical with Bailey's formula (3.7).

To test the departure of p from 0.5, correcting for incomplete penetrance at the Aa locus, we first find the maximum likelihood estimate of u for $p = 0.5$, which is

$$u = \frac{2(a+b)}{n}.$$

The expectations are then

$$\frac{1}{2}(a+b), \frac{1}{2}(a+b), \frac{1}{2}(c+d), \frac{1}{2}(c+d),$$

which, compared with the observed frequencies a, b, c and d , gives

$$\chi^2_{[2]} = \frac{(a-b)^2}{a+b} + \frac{(c-d)^2}{c+d}. \quad (16)$$

This χ^2 can also be considered as the sum of two values of χ^2 , each based on one degree of freedom. The first is identical with (14) and the second tests for the departure of p from 0.5, so that

$$\chi^2_{[1]} = \frac{(a-b)^2}{a+b} + \frac{(c-d)^2}{c+d} - \frac{[(a+c)-(b+d)]^2}{n}.$$

This can be rewritten as

$$\chi^2_{[1]} = \frac{4(bc-ad)^2}{n(a+b)(c+d)}, \quad (17)$$

which is identical to Bailey's formula (3.2).

As shown by Bailey, the following transpositions of symbols are necessary for repulsion crosses and for the opposite type of misclassification, *i.e.*, where a proportion of aa individuals is classified with Aa :

		(a)	(b)	(c)	(d)
(1) $AB/ab \times ab$	. .	$A \rightarrow a$	AB	Ab	aB
(2) $Ab/aB \times ab$	. .	$A \rightarrow a$	Ab	AB	ab
(3) $AB/ab \times ab$	. .	$a \rightarrow A$	ab	aB	Ab
(4) $Ab/aB \times ab$	. .	$a \rightarrow A$	aB	ab	AB

(c) Consistency of estimation under incomplete penetrance

Suppose that penetrance is complete at maturity (stage 2) but that larger families can be accommodated at an earlier stage of development (stage 1) when classification is complicated by incomplete

manifestation at one of two loci whose linkage relations are under investigation. The following questions arise: (1) What effect does incomplete penetrance have on the value of p calculated by the usual estimator (formula 2)? (2) Will the estimator that corrects for incomplete penetrance (formula 8) give a consistent estimate of p ? (3) Will stage 1 or stage 2 data allow more economical accumulation of information about p , *i.e.* which will produce the smaller standard error for given effort? The first two questions are considered in this section, and the third in the next section.

With undisturbed (stage 2) data the estimate of p obtained by the standard estimator is $\frac{b+c}{n}$, whose expectation is p . With incomplete penetrance (stage 1 data) the corresponding expectation is

$$\frac{1}{2}pu + \frac{1}{2}[p + (1-p)(1-u)].$$

Since whenever $u < 1$,

$$p < \frac{1}{2}pu + \frac{1}{2}[p + (1-p)(1-u)],$$

stage 1 data can be expected to underestimate the intensity of linkage if incomplete penetrance is ignored.

To evaluate the consistency of the estimate obtained from formula (8), let u_1 be the value of u occurring in the phenotype AB , and u_2 that occurring in the phenotype Ab , so that u_1 and u_2 are two sample values of the variable u . Then we have the following situations at stages 1 and 2:

Phenotype	Stage 1	Stage 2
AB	a	$\frac{a}{u_1}$
Ab	b	$\frac{b}{u_2}$
aB	c	$c - \left(\frac{a}{u_1} - a\right)$
ab	d	$d - \left(\frac{b}{u_2} - b\right)$

The value of p given by stage 2 data is found, in accordance with formula (2), as

$$\begin{aligned} p = p_2 &= \frac{\frac{b}{u_2} + c - \left(\frac{a}{u_1} - a\right)}{n} \\ &= \frac{\frac{b}{u_2} - \frac{a}{u_1} + (c+a)}{n} \end{aligned} \quad (18)$$

and that given by stage 1 data is found, in accordance with formula (8) as

$$p = p_1 = \frac{b(a+c)}{a(b+d) + b(a+c)}.$$

To compare p_1 and p_2 , let us define ϵ_1 by means of the following equation:

$$u_1 = \frac{a(b+d) + b(a+c)}{(a+c)(b+d) + \epsilon_1}.$$

Note that if $\epsilon_1 = 0$, then $u_1 = u$, so that ϵ_1 is a measure of the departure of u_1 from its maximum likelihood estimate, u . Similarly, we define ϵ_2 by the equation

$$u_2 = \frac{a(b+d) + b(a+c)}{(a+c)(b+d) + \epsilon_2}.$$

Substituting these values of u_1 and u_2 into (18), we obtain

$$p_2 = p_1 + \frac{b\epsilon_2 - a\epsilon_1}{n[a(b+d) + b(a+c)]}.$$

Therefore, the difference between the estimates of p obtained from stages 1 and 2 is

$$p_1 - p_2 = \frac{a\epsilon_1 - b\epsilon_2}{n[a(b+d) + b(a+c)]}.$$

It follows that:

if $a\epsilon_1 - b\epsilon_2 = 0$ or $\frac{\epsilon_1}{b} = \frac{\epsilon_2}{a}$, then $p_2 = p_1$;

if $a\epsilon_1 - b\epsilon_2 > 0$ or $\frac{\epsilon_1}{b} > \frac{\epsilon_2}{a}$, then $p_1 > p_2$;

if $a\epsilon_1 - b\epsilon_2 < 0$ or $\frac{\epsilon_1}{b} < \frac{\epsilon_2}{a}$, then $p_1 < p_2$.

In particular, it follows that:

if $\epsilon_1 = \epsilon_2 = 0$, then $p_2 = p_1$;

if ϵ_1 is positive and ϵ_2 is negative, i.e. $u_1 < u$ and $u_2 > u$, then $p_1 > p_2$;

if ϵ_1 is negative and ϵ_2 is positive, i.e. $u_1 > u$ and $u_2 < u$, then $p_1 < p_2$.

Both ϵ_1 and ϵ_2 can be assumed to be so distributed that their mean is zero, since $\bar{u}_1 = \bar{u}_2 = u$, for which case $\epsilon = 0$. Consequently, if we accept the value of p obtained from stage 2 data as the "true" value of p , we expect that p will be overestimated half of the time and underestimated half of the time from stage 1 data and that, over many trials, $p_1 \rightarrow p_2$. Since p_2 is a consistent estimator of p , then p_1 is also a consistent estimator of p so long as penetrance $A \rightarrow a$ is the same for all genetic phases of the Bb locus.

(d) *Efficiency of estimation with incomplete penetrance*

When p is estimated from undisturbed data, the expected variance of p is given by

$$V_{p_2} = \frac{p(1-p)}{n} \text{ (formula 3).}$$

The corresponding expression when correction is made for u is

$$V_{p_1} = \frac{2}{n} \frac{p(1-p)(1-2pu+2p^2u)}{u} \text{ (formula 11).}$$

The estimates of p obtained by calculation from formulas (3) and (11) are, in general, not the same. We note, however, that if the values

FIG. 1.—The relationship between information expected for undisturbed contrasted with disturbed testcross data, expressed as the ratio I_{p_1}/I_{p_2} .

of p_1 and p_2 do not differ markedly, say by less than 0.10, the corresponding values of the expression $p(1-p)$ will differ little. Hence, in comparing (3) and (11) we write

$$V_{p_1} \cong \frac{2(1-2pu+2p^2u)}{u} V_{p_2}$$

from which it can be seen that $V_{p1} = V_{p2}$ only when $p = 0.5$ and $u = 1$. As p changes from 0.5 to 0, V_{p1} changes from $\frac{2-u}{u} V_{p2}$ to $\frac{2}{u} V_{p2}$. Thus, when u is estimated to increase the accuracy of estimation of p , there is a decrease in the amount of information, $I_p = \frac{1}{V_p}$. The ratio I_{p1}/I_{p2} is depicted graphically in fig. 1, from which it is seen that p_1 is only slightly less efficient than p_2 as an estimator of p when penetrance is nearly complete and linkage is loose, but drastically less efficient with tight linkage and severe failure of manifestation.

2. INTERCROSSES

(a) Undisturbed data

When viability is unaffected and penetrance is complete, the expected gametic output of an intercross $AaBb \times AaBb$ is:

	<i>AB</i>	<i>Ab</i>	<i>aB</i>	<i>ab</i>
♂	$\frac{1}{2}(1-p_m)$	$\frac{1}{2}p_m$	$\frac{1}{2}p_m$	$\frac{1}{2}(1-p_m)$
♀	$\frac{1}{2}(1-p_f)$	$\frac{1}{2}p_f$	$\frac{1}{2}p_f$	$\frac{1}{2}(1-p_f)$

in which p_m and p_f are the recombination values in male and female gametogenesis. These two parameters enter the expectations in

TABLE 2
Observed and expected frequencies for intercross data

Phenotype	Observed frequencies	Expected frequencies $\times \frac{1}{n}$	
		$u = 1$	$u \neq 1$
<i>AB</i>	<i>a</i>	$m_9 = \frac{1}{4}(2+P)$	$m_{13} = \frac{1}{4}(2+P)u$
<i>Ab</i>	<i>b</i>	$m_{10} = \frac{1}{4}(1-P)$	$m_{14} = \frac{1}{4}(1-P)u$
<i>aB</i>	<i>c</i>	$m_{11} = \frac{1}{4}(1-P)$	$m_{15} = \frac{1}{4}[(1-P) + (2+P)(1-u)]$
<i>ab</i>	<i>d</i>	$m_{12} = \frac{1}{4}P$	$m_{16} = \frac{1}{4}[P + (1-P)(1-u)]$
Total	<i>n</i>	1	1

different ways for coupling and repulsion data, so that in principle p_m and p_f can be estimated individually when both sorts of data are available. In practice, however, it is convenient to assume $p_m = p_f$ and estimate p from $\sqrt{P}$. Writing $P = p_m p_f$ for matings in repulsion and $P = (1-p_m)(1-p_f)$ for matings in coupling, the expected proportions of the four distinguishable phenotypic classes are as given in table 2 under $u = 1$, and the likelihood of obtaining a given family with observed frequencies a, b, c and d is

$$\frac{n!}{a!b!c!d!} m_9^a m_{10}^b m_{11}^c m_{12}^d. \quad (19)$$

Differentiating the logarithm of this likelihood with respect to P and equating it to zero, we obtain P as the positive solution of

$$a\left(\frac{1}{2+P}\right) + (b+c)\left(\frac{-1}{1-P}\right) + d\left(\frac{1}{P}\right) = 0 \quad (20)$$

or
$$nP^2 - (a-2b-2c-d)P - 2d = 0.$$

The expected variance of P is

$$V_P = \frac{2P(2+P)(1-P)}{n(1+2P)}. \quad (21)$$

The expected variance of p for matings in repulsion is

$$V_p = \frac{V_P}{\left(\frac{dP}{dp}\right)^2} = \frac{(2+p^2)(1-p^2)}{2n(1+2p^2)} \quad (22)$$

and for matings in coupling

$$V_p = \frac{V_P}{\left(\frac{dP}{dp}\right)^2} = \frac{(3-2p+p^2)(2p-p^2)}{2n(3-4p+2p^2)}. \quad (23)$$

The usual test for linkage in the absence of disturbing influences is

$$\chi^2_{[1]} = \frac{(a-3b-3c+9d)^2}{9n}. \quad (24)$$

(b) *Incomplete penetrance*

For an intercross in which the penetrance of gene A is incomplete, so that a fraction $1-u$ of AA and Aa genotypes is classified phenotypically as aa , the expected frequencies of the four phenotypes in the offspring are as given in table 2 under $u \neq 1$. The likelihood of obtaining an observed family is now

$$\frac{n!}{a!b!c!d!} m_{13}^a m_{14}^b m_{15}^c m_{16}^d. \quad (25)$$

Using maximum likelihood methods as given by Sanchez-Monge, we obtain estimates of u and P as follows:

$$u = \frac{a}{a+c} + \frac{b}{3(b+d)} \quad (26)$$

and

$$P = \frac{ab+3ad-2bc}{4ab+3ad+bc}, \quad (27)$$

from which we obtain an estimate of p for matings in repulsion by

$$P = p^2$$

and for matings in coupling by

$$P = (1-p)^2.$$

The expected variance of u is approximated by

$$V_u = \Sigma \left[np \left(\frac{\delta u}{\delta a} \right)^2 \right] - n \left(\frac{\delta u}{\delta n} \right)^2 \\ = \frac{4u}{27n} [9 - 7u + 2Pu + 4P^2u], \quad (28)$$

which is a correction of formula (18) of Sanchez-Monge.

The variance of u actually realised in a given family is

$$V_u = \frac{ac}{(a+c)^3} + \frac{bd}{9(b+d)^3}. \quad (29)$$

The estimates of V_u given by (28) and (29) will usually not differ greatly.

The expected variance of P is

$$V_P = \frac{4}{n} \frac{(1-P)(2+P)(9-8u+4Pu+4P^2u)}{27u}, \quad (30)$$

which is a correction of formula (20) of Sanchez-Monge. The expected variance of p for matings in repulsion is given by

$$V_p = \frac{V_P}{\left(\frac{dP}{dp} \right)^2} = \frac{(1-p^2)(2+p^2)(9-8u+4p^2u+4p^4u)}{27p^2un}, \quad (31)$$

and for matings in coupling by

$$V_p = \frac{V_P}{\left(\frac{dP}{dp} \right)^2} \\ = \frac{(2p-p^2)(3-2p+p^2)[9-8u+4(1-p)^2u+4(1-p)^4u]}{27(1-p)^2un}, \quad (32)$$

which is a correction of formula (21) of Sanchez-Monge.

The variance of P actually realised in a given family is obtained as

$$V_P = \frac{81(a+c)(b+d)[ab^2c(b+d)+a^2bd(a+c)]}{[3a(b+d)+b(a+c)]^4}. \quad (33)$$

The variance of p realised from matings in repulsion is

$$V_p = \frac{V_P}{\left(\frac{dP}{dp} \right)^2} = \frac{81(a+c)(b+d)[ab^2c(b+d)+a^2bd(a+c)]}{4p^2[3a(b+d)+b(a+c)]^4} \quad (34)$$

and from matings in coupling

$$V_p = \frac{V_P}{\left(\frac{dP}{dp} \right)^2} = \frac{81(a+c)(b+d)[ab^2c(b+d)+a^2bd(a+c)]}{4(1-p)^2[3a(b+d)+b(a+c)]^4}. \quad (35)$$

To test the departure of u from unity, correcting for linkage, we assume $u = 1$, for which case the maximum likelihood estimate of P is obtained as the positive solution of equation (20). The expected frequencies are then found by substituting this estimate of P in the column labelled $u = 1$ of table 2. No simple explicit formulas for these expectations can be given, since there is no simple explicit formula for P . Comparing the expectations with the observed frequencies a , b , c and d , we obtain a value of χ^2 , denoted by $\chi^2_{[1]}$, based on two degrees of freedom since, of the three originally available, one has been used to estimate P . This value of χ^2 can be considered as the sum of two values of χ^2 , each based on one degree of freedom, where the first tests for the departure of the Bb segregation from the theoretical 3 : 1 ratio, *i.e.*

$$\chi^2_{[1]} = \frac{[(a+c) - 3(b+d)]^2}{3n}, \quad (36)$$

and the second tests for the departure of u from unity, which is then given by

$$\chi^2_{[1]} = \chi^2_{[2]} - \frac{[(a+c) - 3(b+d)]^2}{3n}. \quad (37)$$

To detect the existence of linkage, correcting for incomplete penetrance at the Aa locus, we first find the maximum likelihood estimate of u for $P = \frac{1}{4}$, ($p_m = p_f = \frac{1}{2}$), which is $u = \frac{4(a+b)}{3n}$.

The expectations are then $\frac{3}{4}(a+b)$, $\frac{1}{4}(a+b)$, $\frac{3}{4}(c+d)$, $\frac{1}{4}(c+d)$, which, compared with the observed frequencies a , b , c and d , give

$$\chi^2_{[2]} = \frac{(3b-a)^2}{3(a+b)} + \frac{(3d-c)^2}{3(c+d)}.$$

This χ^2 is the sum of two values of χ^2 , each based on one degree of freedom. The first is identical to (36) and the second tests for the departure of P from $\frac{1}{4}$ so that

$$\chi^2_{[1]} = \frac{(3b-a)^2}{3(a+b)} + \frac{(3d-c)^2}{3(c+d)} - \frac{[(a+c) - 3(b+d)]^2}{3n}.$$

This can be rewritten as

$$\chi^2_{[1]} = \frac{16(ad-bc)^2}{3n(a+b)(c+d)}. \quad (38)$$

(c) Efficiency of estimation from disturbed intercross data

When p is estimated from undisturbed intercross data, the amount of information expected is

$$I_{p^2} = \frac{2n(1+2p^2)}{(2+p^2)(1-p^2)}.$$

Information expected from disturbed data is given by

$$I_{p1} = \frac{27p^2un}{(1-p^2)(2+p^2)(9-8u+4p^2u+4p^4u)}.$$

The values of p will not be exactly the same in these expressions, but, since the two estimates of p can be expected to differ little, we can use

FIG. 2.—The relationship between information expected for undisturbed contrasted with disturbed intercross data, expressed as the ratio I_{p1}/I_{p2} .

the ratio I_{p1}/I_{p2} as a measure of efficiency of estimation for various values of p and u . This ratio is depicted graphically in fig. 2, from which it is seen that p_1 is only slightly less efficient than p_2 as an estimator of p when linkage is loose and penetrance is nearly complete. When linkage is tight, incomplete penetrance causes only moderate losses in efficiency for coupling crosses but severe losses in efficiency for repulsion crosses.

3. APPLICATIONS TO TESTCROSS AND INTERCROSS DATA

(a) *Experimental materials*

Indeterminant *v.* determinant habit of growth in the lima bean, *Phaseolus lunatus* L., is governed by a single Mendelian gene pair, *D-d* (Allard, 1953a). Red *v.* green hypocotyl colour in seedling plants is also governed by a single Mendelian gene pair, *R-r*, which

TABLE 3

Observed frequencies for the testcross * $\frac{RD}{rd} \times rd$, and the coupling and repulsion intercrossoes $RrDd \times RrDd$ in seedling and at mature plant stages of development †

Family	Phenotypic class							
	<i>a</i>		<i>b</i>		<i>c</i>		<i>d</i>	
	Seedling	Mature	Seedling	Mature	Seedling	Mature	Seedling	Mature
Testcrosses								
1	81	107	48	55	89	63	120	113
2	47	112	40	72	138	73	157	125
3	76	131	36	60	99	44	149	125
4	97	124	53	67	110	83	112	98
Intercrosses								
5 (C)	165	200	45	57	84	49	42	30
6 (C)	206	274	60	71	132	64	56	45
7 (R)	207	293	75	107	205	119	67	35
8 (C)	92	180	25	50	134	46	61	36
9 (R)	139	188	58	74	125	76	31	15
10 (R)	140	202	50	84	120	58	46	12
11 (R)	135	182	52	78	120	73	44	18

* Testcrosses were made with the aid of a male-sterility gene.

† In these families, mortality averaged 0.8 per cent. after seedling classification and before mature-plant readings could be taken. Plants scored in the seedling stage that failed to survive until seed-coat colour could be determined were not included in the seedling-stage data.

often fails to manifest (Allard, 1952). *R-r* also governs red *v.* dark-red seed-coat colour, a mature-plant character with complete manifestation (Allard, 1953b). Studies of mature plants have established that recombination between these two loci is about 40 per cent. (Allard, 1956; Allard and Clement, 1959). Among many families studied in field plantings over several years, eleven were classified in both seedling and mature-plant stages of development so as to maintain the identity of individual plants in both stages. In this

section, data from these families are used for experimental evaluation of the effectiveness of the statistics of Sections 1 and 2 in estimation of p when the data are disturbed by incomplete penetrance.

(b) *Detection and estimation of incomplete penetrance*

Among the three degrees of freedom available for each backcross, or F_2 , family, one can be used to calculate a χ^2 for the deviation of the $D-d$ segregation from the expected 1 : 1 or 3 : 1 ratios. The observed deviation was significant for only one of the eleven families listed in table 3 (Family 4). This result, combined with extensive previous experience that the $D-d$ segregation gives good fits to expectancies based on monofactorial Mendelian segregation, suggests that the single significant deviation observed here was a chance event. Hence it is concluded that the $D-d$ segregation is undisturbed in these families.

Another degree of freedom can be used to test goodness of fit for the $R-r$ segregation. Only Family 10 deviated significantly from expectation ($P=0.02$) for the mature-plant data. The deviation was toward an excess of dominant phenotypes. Thus, for mature-plant data, there is no reason to suspect difficulties of manifestation. In seedling-stage plants the $R-r$ segregation showed a significant excess of individuals with green hypocotyls in all eleven families (table 4, column 1). Since this segregation gave a good fit to Mendelian expectations for the mature-plant data, viability difficulties can be eliminated as the cause of the deviation and the deviation is thus most reasonably attributed to failure of a portion of the genotypically coloured plants to develop pigmentation. A measure of the failure to manifest in seedling stage is given by the estimated values of u in column 7 of table 4. It should be noted that the deviation of u from unity is less than 2 standard errors for the mature-plant data (excepting Family 10) and uniformly greater than two standard errors for the same families in seedling stage.

In summary, the values presented in columns 1, 2, 7 and 8 of table 4 establish that the $R-r$ gene segregates according to Mendelian expectations at maturity, but that penetrance is incomplete, $R \rightarrow r$, in seedling stage, in degrees varying with the environments encountered by the seedling plants.

(c) *Detection of linkage*

The standard χ^2 tests for linkage (formulae 5, 24) indicate linkage in 9 of the 11 families when they were classified at maturity (table 4, column 4) but in only 7 of the families when classified as seedlings, (table 4, column 3). It will also be noted that the χ^2 values were, with one exception, smaller for seedling-stage than for mature-stage classification. Formulae 17 and 38 take incomplete penetrance into account and therefore minimise the bias resulting from incomplete manifestation. For these particular families the correction for incomplete penetrance did little to improve the sensitivity of the χ^2

test. This was the case both for mature-plant data, where penetrance was complete (compare columns 4 and 6, table 4), and also for seedling data (compare columns 3 and 4, table 4), where penetrance was incomplete. The differences between the linkage χ^2 s calculated in the two ways were negligible in the present families. Nevertheless, the corrected χ^2 s should be used routinely when there is significant disturbance from incomplete penetrance, since it is easy to visualise situations in which the standard test could be misleading. The formula that takes incomplete penetrance into account should not

TABLE 4—*Detection and estimation of*

Family	Departure of u from unity		Departure of p from 0.5, ignoring incomplete penetrance		Departure of p from 0.5, correcting for incomplete penetrance	
	Seedling	Mature	Seedling	Mature	Seedling	Mature
	1	2	3	4	5	6
	$\chi^2_{[1]}$	$\chi^2_{[1]}$	$\chi^2_{[1]}$	$\chi^2_{[1]}$	$\chi^2_{[1]}$	$\chi^2_{[1]}$
1	19.83	0.69	12.12	30.78	13.03	30.88
2	112.89	0.34	1.77	22.16	1.41	21.98
3	52.81	2.32	22.50	64.18	24.09	64.94
4	16.27	0.01	5.69	13.94	8.18	13.51
5	28.17	0.50	8.05	7.64	5.95	8.04
6	65.21	0.31	4.39	18.37	3.07	18.94
7	171.74	2.35	0.18	0.98	0.28	0.96
8	233.08	0.10	9.58	16.62	3.83	15.84
9	69.91	0.12	5.40	5.08	4.25	4.98
10	88.47	5.02	0.60	4.20	0.09	4.48
11	88.11	0.26	0.07	3.76	0.04	3.75

be used for undisturbed data since it reduces the amount of information realised, particularly when the recombination value is small.

(d) Estimation of linkage

Estimates of recombination fractions, calculated from both seedling and mature-plant data, with and without correction for incomplete penetrance, are given in columns 9-12 of table 4. The values calculated from mature-plant data from formulæ 2 or 20 correspond to the usual estimates made from undisturbed data. These values are here accepted as "true" values for comparison.

Correcting the mature-plant data for non-existent incomplete penetrance (column 12) led to values of p that differed trivially from the standard estimates (column 10). This was expected since if $u \cong 1$, formula (8) reduces to formula (2) and formula (27) reduces to formula (20). There was a slight tendency for the standard error of p to be larger when correction was made for incomplete penetrance.

This is also expected, because, for $u \cong 1$, efficiency should be reduced only slightly when linkage is loose (figs. 1 and 2).

The estimates of p calculated from seedling-stage data, ignoring incomplete penetrance, are given in column 9 of table 4. Comparison with the "true" values calculated from mature-plant data (column 10) shows that, for these families, incomplete penetrance caused the "true" intensity of the linkage to be consistently underestimated. Again this is expected, for reasons discussed in section 2.

The magnitude of this bias can be determined by comparing the

incomplete penetrance and linkage

Estimated values of u , and standard errors		Estimated values of p , and standard errors, ignoring incomplete penetrance		Estimated values of p , and standard errors, correcting for incomplete penetrance	
Seedling	Mature	Seedling	Mature	Seedling	Mature
7	8	9	10	11	12
0.7622 $\pm$ 0.0518	0.9568 $\pm$ 0.0518	0.4053 $\pm$ 0.0267	0.3491 $\pm$ 0.0259	0.3749 $\pm$ 0.0342	0.3422 $\pm$ 0.0282
0.4571 $\pm$ 0.0430	0.9709 $\pm$ 0.0497	0.4660 $\pm$ 0.0255	0.3796 $\pm$ 0.0248	0.4442 $\pm$ 0.0348	0.3764 $\pm$ 0.0262
0.6289 $\pm$ 0.0474	1.0729 $\pm$ 0.0475	0.3750 $\pm$ 0.0255	0.2889 $\pm$ 0.0239	0.3094 $\pm$ 0.0372	0.3023 $\pm$ 0.0245
0.7898 $\pm$ 0.0502	1.0051 $\pm$ 0.0512	0.4382 $\pm$ 0.0257	0.4032 $\pm$ 0.0254	0.4067 $\pm$ 0.0319	0.4040 $\pm$ 0.0265
0.8351 $\pm$ 0.0396	1.0216 $\pm$ 0.0374	0.4186 $\pm$ 0.0371	0.3961 $\pm$ 0.0360	0.3831 $\pm$ 0.0457	0.4011 $\pm$ 0.0358
0.7819 $\pm$ 0.0336	1.0147 $\pm$ 0.0333	0.4505 $\pm$ 0.0332	0.3656 $\pm$ 0.0296	0.4182 $\pm$ 0.0445	0.3701 $\pm$ 0.0303
0.6785 $\pm$ 0.0292	0.9624 $\pm$ 0.0271	0.4918 $\pm$ 0.0322	0.4703 $\pm$ 0.0329	0.4706 $\pm$ 0.0576	0.4658 $\pm$ 0.0359
0.5040 $\pm$ 0.0387	0.9903 $\pm$ 0.0409	0.4376 $\pm$ 0.0394	0.3600 $\pm$ 0.0354	0.3495 $\pm$ 0.0694	0.3574 $\pm$ 0.0373
0.7437 $\pm$ 0.0354	0.9893 $\pm$ 0.0321	0.4366 $\pm$ 0.0426	0.5966 $\pm$ 0.0440	0.3519 $\pm$ 0.0867	0.3994 $\pm$ 0.0469
0.7121 $\pm$ 0.0369	1.0686 $\pm$ 0.0302	0.5187 $\pm$ 0.0389	0.3991 $\pm$ 0.0440	0.5183 $\pm$ 0.0621	0.4257 $\pm$ 0.0362
0.7100 $\pm$ 0.0368	0.9845 $\pm$ 0.0327	0.5065 $\pm$ 0.0397	0.4233 $\pm$ 0.0433	0.4869 $\pm$ 0.0667	0.4181 $\pm$ 0.0461

value of p obtained from the combined mature-plant data and the value obtained from the combined seedling data. Since the scoring method (see Mather, 1951; Rao, 1952) is convenient when several sets of data are to be combined in estimating a parameter, it was adopted here. The estimated value of p for the combined mature-plant data was 0.3770 and p for the combined seedling data was 0.4477. Thus the effect on the usual estimator was an upward bias of $0.4477 - 0.3770 = 0.0707$ in the estimated value of p .

The estimates of p obtained, taking incomplete penetrance into account, are given in column 11 of table 4. In all 4 backcross families these values were intermediate between the "true" values (column 10) and the uncorrected estimates (column 9). Thus, the correction was in the right direction but was uniformly too small. Among the seven F_2 families, four under- and three overcorrections were observed. It will be recalled from Section 1 that undercorrection occurs if $u_1 > u_2$, and overcorrection with the reverse situation. The preponderance

of undercorrections in these eleven families does not appear to be significant, but present data do not rule out the possibility that there is a tendency for manifestation to be less complete in D than in d genotypes.

To obtain a combined estimate of p from all eleven families, taking the differences in degree of manifestation into account, we note that the expectations involve the two parameters u and p and that we are particularly concerned with the estimation and significance of p . If L is the logarithm of the likelihood, an efficient score for $p = p_1$ is

$$S(p_1) = \left. \frac{\partial L}{\partial p} \right|_{p=p_1},$$

where maximum-likelihood estimates of p and u , given $p = p_1$, have been substituted. Thus, we can find the combined maximum-likelihood estimate of p by substituting u into the following equations and solving for p by arithmetic trial and interpolation:

Backcross:

$$S(p) = \frac{a}{p-1} + \frac{b}{p} + \frac{cu}{1+u(p-1)} + \frac{du}{pu-1} \quad (39)$$

Intercross (repulsion):

$$S(p) = 2p \left\{ \frac{a}{2+p^2} + \frac{b}{p^2-1} - \frac{cu}{3-u(2+p^2)} + \frac{du}{1-u(1-p^2)} \right\} \quad (40)$$

Intercross (coupling):

$$S(p) = 2(p-1) \left\{ \frac{a}{3-2p+p^2} + \frac{b}{p^2-2p} - \frac{cu}{3-u(3-2p+p^2)} + \frac{du}{1-u(2p-p^2)} \right\} \quad (41)$$

The information about p realised can be calculated as

$$I_p \cong \frac{S(p_1) - S(p_2)}{p_2 - p_1}, \quad (42)$$

in which p_1 and p_2 represent adjacent values of p for which the scores take opposite signs.

Fisher (1954) has shown that the sum of the squares of the scores for the pooled families, divided by the information provided by each family, satisfies a χ^2 distribution. Thus,

$$\chi^2 = \sum_1^N \frac{[S(p)]^2}{I_p}$$

provides a heterogeneity test with $N-1$ degrees of freedom, in which N is the number of pooled families.

The appropriate computations are set out below:

Family	Value of u	Trial Value of p		Information I_p	$\chi^2_{[1]} = \frac{[S(p)]^2}{I_p}$ at $p = 0.3967$
		$0.3967 S(p)$	$0.3966 S(p)$		
1 (BC)	0.7622	-18.78554	-18.70075	847.9	0.4162
2 (BC)	0.4571	+22.36566	+22.41440	487.4	1.0263
3 (BC)	0.6289	-59.75428	-59.68989	643.9	5.5452
4 (BC)	0.7898	+9.95353	+10.05371	1001.8	0.0989
5 (F_2 C)	0.8351	-6.60505	-6.55814	469.1	0.0930
6 (F_2 C)	0.7819	+11.71894	+11.77645	575.1	0.2388
7 (F_2 R)	0.6785	+17.88877	+17.90641	176.4	1.8141
8 (F_2 C)	0.5040	-9.07086	-9.05454	163.2	0.5042
9 (F_2 R)	0.7437	-7.34582	-7.32701	188.1	0.2869
10 (F_2 R)	0.7121	+23.06830	+23.07978	114.8	4.6354
11 (F_2 R)	0.7100	+16.34402	+16.35675	127.3	2.0984
Totals		-0.22233	+0.25717	4795.0	$\chi^2_{[10]} = 16.7574$

Interpolating gives $p = 0.3967$ with a standard error of $4795.0^{-\frac{1}{2}}$. Therefore $p = 0.3967 \pm 0.0144$. The non-significant heterogeneity χ^2 indicates that the families are consistent in supporting the joint estimate of the recombination value.

The "true" value, estimated from the undisturbed mature-plant data, was $p = 0.3770 \pm 0.0095$. Hence the estimate from the disturbed seedling data, although 0.0197 too high, represents a substantially better estimate than the value of 0.4477 obtained when incomplete penetrance was ignored.

Information realised, respectively, from the mature-plant data and from seedling data, taking incomplete penetrance into account, was $11,088.0$ and 4795.0 units. The efficiency of the seedling data relative to the mature-plant data was therefore 43.25 per cent. Since seedling data are obtained with less than one-tenth as much effort and expense as mature-plant data, it is clear that information about linkage will, in this case, be accumulated more rapidly from seedling data.

These results indicate that consistent estimates of the linkage parameter can be obtained from data in which incomplete penetrance is a disturbing factor, and further, that loss in efficiency is not drastic unless failure of manifestation is severe. Estimators that take incomplete penetrance into account should therefore be used routinely (a) when the experimenter cannot set up conditions where penetrance is complete or (b) when the effort or expense of achieving complete penetrance exceeds the loss in efficiency that accompanies addition of the penetrance parameter to the estimator.

4. SUMMARY

The consistency and efficiency of linkage estimators that take incomplete penetrance into account are examined theoretically.

Introduction of a penetrance parameter, u , does not affect consistency of estimation of p (the linkage parameter) so long as incomplete manifestation $A \rightarrow a$ or $a \rightarrow A$ is equivalent in all genetic phases at the second locus in two point tests. Efficiency of estimation depends on both p and u when penetrance is taken into account. Efficiency is equivalent to that of the usual estimator when $u = 1$ and $p = 0.5$, but decreases as p and u tend toward zero.

The effectiveness of the estimators that take penetrance into account was examined experimentally by applying them to data from four backcross and seven F_2 families in which manifestation was incomplete in seedlings but complete in mature plants. Ignoring incomplete penetrance led to severely biased estimates of linkage intensity for the seedling data. This bias was removed when correction was made for varying degrees of manifestation, but at the expense of information. Nevertheless, the policy of using seedling data was clearly advantageous since the loss in efficiency was more than compensated by the greater expense and effort required to grow plants to maturity.

The estimators that take incomplete penetrance into account should be particularly useful in linkage experiments involving genes for disease resistance, and in other similar situations where disturbances resulting from failures of manifestation are common and no certain method of achieving complete penetrance is known.

Acknowledgments.—We are pleased to acknowledge the assistance of Mr S. K. Jain and Mrs Francis Jones who performed many of the computations.

5. REFERENCES

- ALLARD, R. W. 1952. Inheritance of hypocotyl color in lima beans. *Proc. Amer. Soc. Hort. Sci.*, 60, 387-390.
- ALLARD, R. W. 1953a. Inheritance of some seed coat colors and patterns in lima beans. *Hilgardia*, 22, 167-177.
- ALLARD, R. W. 1953b. Inheritance of four morphological characters in lima beans. *Hilgardia*, 22, 383-389.
- ALLARD, R. W. 1956. Formulas and tables to facilitate the calculation of recombination values of heredity. *Hilgardia*, 24, 235-278.
- ALLARD, R. W., AND W. M. CLEMENT. 1959. Linkage in lima beans. *J. Heredity*, 50, 63-67.
- BAILEY, N. T. J. 1949. The estimation of linkage with differential viability. *Heredity*, 3, 220-228.
- BAILEY, N. T. J. 1950. The influence of partial manifestation on the detection of linkage. *Heredity*, 4, 327-336.
- FISHER, R. A. 1954. *Statistical Methods for Research Workers*. 12th ed. Oliver and Boyd, Edinburgh.
- MATHER, K. 1951. *The Measurement of Linkage in Heredity*. Methuen & Co., London.
- PARSONS, P. A. 1957. Partial manifestation of a gene in complete three and higher point backcross data. *Heredity*, 11, 217-222.
- RAO, C. R. 1952. *Advanced Statistical Methods in Biological Research*. John Wiley & Sons, New York.
- SANCHEZ-MONGE, E. 1952. The estimation of linkage with incomplete penetrance. *Heredity*, 6, 121-125.