Euphytica **70**: 225–235, 1993.

Component analysis of complex characters in plant breeding

I. *Proposed method for quantifying the relative contribution of individual components to variation of the complex character*

L.D. Sparnaaij & I. Bos[1]
Formerly: Centre for Plant Breeding and Reproduction Research, Wageningen, The Netherlands; [1] Dept. of Plant Breeding, Agric. University, P.O. Box 386, 6700 AJ Wageningen, The Netherlands*

Received 5 February 1993; accepted 2 September 1993

Key words: chrysanthemum, coefficient of determination, complex characters, *Dendranthema grandiflora,* plant breeding methods, recombinative heterosis, sequential component analysis

Summary

A method is described by which the contribution of individual components to the variation of a complex character is quantified. The method is adapted for use in plant breeding from the sequential component analysis, developed for agronomic experiments by Eaton & Kyte (1978). It applies to a situation in which the complex character y is the product of n components (x_1,. . .x_n). The components are ratios of observed primary characters, introduced in a logical sequence. The proposed method differs from that of Eaton & Kyte in that: (1) the data are not log-transformed, and (2) the complementary determinations of y by its components are obtained directly from differences between the coefficients of determination of y by the successive primary characters rather than from a stepwise multiple regression of ln y on its log-transformed components.

A comparison of the two methods suggests that the differences in results are caused mainly by the log-transformation which Eaton & Kyte apply to transform the relationship between y and its components into a linear one. The proposed alternative procedure avoids the transformation of the data; the multiplicative relationship between y and its components is retained and so is the assumed additive inheritance of the components. These two features of component analysis allow an active exploitation of specific combining ability based on recombinative heterosis.

Introduction

Many of the characters plant breeders seek to improve, are physiologically and genetically complex. Commercial breeders in particular are continuously engaged in the improvement by selection of complex characters as yield, growth rate, rate of propagation, vase life, time of flowering, suitability for forcing (e.g. of flower bulbs), partial resistance to diseases and pests, etc. In contrast, present day plant breeding research, in particular molecular genetics, is more concerned with monogenic qualitative characters, as these present fewer problems in the development and application of modern techniques. However, complex characters are too important to remain on the sideline for long. To make them more amenable to improvement by conventional and perhaps also modern biotechnological breeding methods, a detailed analysis of their components is required. Identification of the major

* Present address: Giltaylaan 12, 6703 GB Wageningen, The Netherlands.

components and determination of their relative contribution to the variation of the complex character is the first objective of such an analysis. This is, in fact, the main topic of this paper.

Component analyses have already been done for different characters (mainly yield characters) in a wide range of crops. In the majority of cases the objective was modest, i.e. to simplify selection on a complex character by replacing it by selection on one of its major components. Provided the correlation with the complex character is high and the selection can be done more efficiently, i.e. earlier and/or at less cost, this is a very sound procedure. It is of practical value for the selection of potential new cultivars in crops with a long juvenile phase.

In the present paper a form of component analysis is presented with more ambitious objectives. A careful choice of components, placed in a logical order, allows the estimation of their mutually independent, relative contributions to the variance of the complex character. The most important components serve as criteria for the selection of breeding parents and parent combinations.

Concepts and definitions

Complex characters. They are characters for which variation is determined by variation of a number of component traits. We use the term 'component traits' to refer to a group of two or more traits which, together, completely determine the complex character: their product equals the complex character, whatever the number of components (x_1,. . .x_n) and whatever the value of y:

$$x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n = y.$$

However, the mean of y (e.g. over a number of cultivars) is not equal to the product of the means of the components. Thus, when crossing two cultivars, the product of the mid-parent values of the components is not equal to the mid-parent value of y.

In this paper additive inheritance is taken as the basic model for the genetic control of the component traits. This implies that we assume components to be *simple* quantitative traits, i.e. traits that are not a product of subcomponents. Components often do not inherit additively. This may be due to:

- a complex nature of the component trait. Subdivision into two or more sub-components may yield additively inheriting components.
- interdependence of components. Correction for the commonly observed influence of preceding components is an essential feature of our procedure.
- dominance, in particular when the number of loci of the component trait is low or when the component is a qualitative trait. Discontinuous variation in the progeny would draw attention to this.

If our model applies, the F_1 values of the components will approximate their mid-parent value. Those of y will deviate from it. This will be the case in particular when the components are complementary, i.e. when a relatively high value of one component in parent P_i is complemented by a high value for another component in parent P_j. This situation is, in our view, a major cause of the non-additive inheritance of complex characters, known as specific combining ability (= SCA). It implies that there is scope for active exploitation of SCA by crossing parents with complementary component traits. The more contrasting the parents are for these traits, the more the complex character in the progeny deviates from its mid-parent value. This deviation is named 'recombinative heterosis', see Part II.

In our opinion, a quantitative character must be treated as a complex one when there is a high degree of independent variation of two or more components. The complex nature of a character can only be brought to light and exploited via a component analysis of a range of genotypes.

When components are used as criteria for the selection of parents and parent combinations, the data collected should *not* be log-transformed. This is best explained by a simple example.

Let us consider an actual situation of a cross between parents P_1 and P_2 in which the components x_1 and x_2 of complex character y are complementary and inherit additively and independently:

x_1 x_2 y
P_1: $6 \cdot 9 = 54$
P_2: $10 \cdot 5 = 50$; mid-parent value of y is 52
F_1: $8 \cdot 7 = 56$ (positive heterosis)

Log-transformation changes this to:

	x_1		x_2		y
P_1:	ln 6	+	ln 9	=	ln 54
P_2:	ln 10	+	ln 5	=	ln 50
Actual F_1 values (log scale) :	ln 8	+	ln 7	=	ln 56
mid-parent values :	$^1/_2$ln 60	+	$^1/_2$ln 45	=	$^1/_2$ln 2700
or :	ln 7.75	+	ln 6.71	=	ln 51.96

The observed heterotic effects in the components are artefacts. They remain unexplained and unpredictable.

We conclude that log-transformation obscures additive inheritance of the components. It is to be avoided in genetic analyses in which the hypothesis of additive inheritance of quantitative traits is tested. This is the case, for example, in the analysis of variance of a diallel crossing scheme.

Variation within a cultivar – between individual plants and/or between plots – also leads to deviations between the mean of y and the product of the component means. When the cultivar is a clone or a pure line such deviation is caused by non-genetical variation (environment, plant quality, plant age, plant health). It can be minimized by controlled environmental conditions and by adequate replication and randomization.

We treat the complex character as the end point of a process of which the successive stages are represented by the observed primary characters. As an example, let us consider the process leading to the production of a certain seed yield (y).

The first stage observed is the number of stems (a). The number of stems per plant is the first component x_1. When the data for the primary characters are averages over the number of plants observed, x_1 equals a.

The second stage is the stage of flowering, resulting in a number of flowers (b). The step to the second stage is controlled by genes determining the number of flowers per stem: $b/a = x_2$.

The third stage is that of seed production, leading to a number of seeds (c). It is the result of the activity of genes controlling the fertility of the flowers as expressed by the number of seeds per flower: $c/b = x_3$.

The final stage, the seed yield in grams (y), involves the action of genes controlling the growth of the seeds, expressed by the average number of grams per seed: $y/c = x_4$. In summary:

$x_1 \cdot x_2 \cdot x_3 \cdot x_4 = y$, or:
$a \cdot b/a \cdot c/b \cdot y/c = y$, or:

$$no.\ stems \cdot \frac{no.\ flowers}{no.\ stems} \cdot \frac{no.\ seeds}{no.\ flowers} \cdot \frac{seed\ yield}{no.\ seeds} = seed\ yield$$

Component analysis. In our definition, this is the analysis of the variation (in plant breeding: variation across a number of genotypes) of the complex character via the variation of its components. To achieve this, the components should be arranged in the correct ontogenetical order. This allows determination of the mutually independent contributions of the components to the variation of the complex character. In the next section a method to arrive at such determination is proposed.

All components, excepting the first one, are ratios of observed primary characters. It depends on the complex character to be analysed whether the analysis starts with an observed quantity or a ratio of two observed quantities:

– When the complex character is an observed quantity, e.g. number of seeds per plant or per m^2, the first component is also an observed quantity, e.g. number of stems per plant or per m^2. In such a case the number of plants or the plot area is not a primary character and thus not part of a component. The above seed yield analysis is an example of this situation.

– When the complex character is a ratio, e.g. relative growth rate (= growth rate per unit plant weight), the first component is also a ratio of two primary characters, e.g. leaf area per unit plant weight (= LA/W):

$$\frac{LA}{W} \cdot \frac{dW/dt}{LA} = \frac{dW/dt}{W} = RGR.$$

Any quantity observed in the course of the process under consideration can be used as the start of the component analysis. The later in the process the start is made, the more complex the first component will be and the more likely its mode of inheritance will be non-additive.

In a component analysis applied to a set of varieties, we assume the components to be individual genetic traits. They have their own variation and their own genetic control. In this respect the components differ from the primary characters. The latter are under indirect genetic control; they are not considered as individual genetic traits but rather as the product of the components, i.e. the resultant of the action of several genetic traits.

The number of components is at least two. In the above example:

$$no.\ seeds \cdot \frac{seed\ yield}{no.\ seeds} = seed\ yield$$

The number of components of seed yield can be increased in several ways:
(1) by choosing an earlier observable trait as first component, e.g. no. flowers, giving three components, or no. stems giving four components (see above example);
(2) by inserting additional primary characters, e.g. the number of inflorescences, extending the ratio

$$\frac{no.\ flowers}{no.\ stems} \text{ to the product } \frac{no.\ infl.}{no.\ stems} \cdot \frac{no.\ flowers}{no.\ infl.}$$

(3) by introducing a parallel analysis of the same complex character using different parameters. In the seed yield example two parallel processes are involved:
(a) a morphological process, expressed in numbers, e.g. of stems, inflorescences, flowers, seeds:

$$no.\ stems \cdot \frac{no.\ infl.}{no.\ stems} \cdot \frac{no.\ flowers}{no.\ infl.} \cdot \frac{no.\ seeds}{no.\ flowers} = no.\ seeds$$

(b) a physiological process expressed in weights, e.g. seed yield, which can be related to plant weight at planting W_0, and to plant weight at harvest W_t:

$$W_0 \cdot \frac{W_t}{W_0} \cdot \frac{seed\ yld}{W_t} = seed\ yld$$

Both the component analysis of seed number and that of seed yield are complete. They may, however, be combined by adding a further component, 'average seed weight':

$$no.\ stems \cdot \frac{no.\ infl.}{no.\ stems} \cdot \frac{no.\ flowers}{no.\ infl.} \cdot \frac{no.\ seeds}{no.\ flowers} \cdot \frac{seed\ yld}{no.\ seeds} = seed\ yld$$

Proposed method for quantifying the mutually independent contributions of the components to the variance of the complex character

Coefficients of correlation (r) or determination (r^2) do not quantify the contribution of individual components to the variance of the complex character because y does not have a linear relation with its components. It is, however, possible to calculate the coefficients of determination of y by the successive primary characters. These primary characters represent the successive ontogenetical or chronological stages in the process leading to y. Therefore, a linear relation with y is a good approximation of the actual relation, at least within the normal range of values. The coefficients of determination (r^2) of y by the successive primary characters thus measures the proportion of the variance of y determined at the successive stages.

The essence of our approach is the argument that an increase in the r^2 value from one stage to the next represents the influence of the intervening component. This increase we have named the 'complementary determination' of y by component x_i, designated as $cd(x_i,y)$. For example, the increase from $r^2(b,y)$ to $r^2(c,y)$ is caused by the contribution of component c/b = x_3, so that $cd(x_3,y) = r^2(c,y) - r^2(b,y)$.

It is advisable to verify for the actual range of values for y and some primary character whether regression through the origin (Snedecor & Cochran, 1967; p. 166) gives rise to a smaller residual sum of squares. If this is the case the part of the variance of y due to the primary character is not equal to r^2, but to the sum of squares due to regression divided by the total sum of squares. In the following it is assumed that in the normal range of values, the regression does not go through the origin.

The basic equation in a component analysis is one with two components: $y = x_1 \cdot x_2$, or: $y = a \cdot y/a$. Component x_1 coincides with the first primary character a. There are no data on earlier stages affecting a, so a is assumed to vary independently. In our approach this means that $cd(x_1,y) = r^2(a,y)$. The two components x_1 and x_2 together fully determine y and we assume that the sum of their complementary determinations is equal to 1. Thus, the contribution of x_2

to the variation in y is the fraction of the variation in y not due to x_1. It is equal to $1-r^2(a,y)$ and designated as $cd(x_2,y)$.

The analysis may be extended, e.g. to four components, i.e.

$$x_1 \cdot x_2 \cdot x_3 \cdot x_4 = y, \text{ or}$$
$$a \cdot b/a \cdot c/b \cdot y/c = y.$$

In this case the calculation of complimentary determinations is not really different. The expression is partitioned into three basic equations:

$$\begin{aligned} a \cdot y/a &= x_1 \cdot (x_2 \cdot x_3 \cdot x_4) = y \\ b \cdot y/b &= (x_1 \cdot x_2) \cdot (x_3 \cdot x_4) = y \\ c \cdot y/c &= (x_1 \cdot x_2 \cdot x_3) \cdot x_4 = y \end{aligned}$$

In each successive line a different, more advanced primary character is used as the starting point. The contribution of a more advanced primary character to the variation in y, measured by r^2, is larger than that of the previous one. The difference between the r^2 values of two consecutive primary characters is taken to be the complementary determination of variation in y by variation in the intervening component:

$$\begin{aligned} cd(x_1,y) &= r^2(x_1,y) &&= r^2(a,y) \\ cd(x_2,y) &= r^2(x_1 \cdot x_2,y) - r^2(x_1,y) &&= r^2(b,y) - r^2(a,y) \\ cd(x_3,y) &= r^2(x_1 \cdot x_2 \cdot x_3,y) - r_2(x_1 \cdot x_2,y) &&= r^2(c,y) - r^2(b,y) \\ cd(x_4,y) &= (r^2(x_1 \cdot x_2 \cdot x_3 \cdot x_4,y) - r^2(x_1 \cdot x_2 \cdot x_3,y) &&= 1 - r^2(c,y) \end{aligned}$$

In this way, one can establish which components contribute most to the variation in y and deserve most attention when choosing and combining breeding parents.

The procedure we propose has been developed from a method of analysis presented by Eaton & Kyte (1978). Their method involves a stepwise multiple regression analysis of the log-transformed data for yield and its components: $\ln x_1 + \ln x_2 + \ln x_3 + \ln x_4 = \ln y$. Though both methods start from the same premisses, there are important differences in the processing of the data.

We purposely avoid multiple regression and so also the need for log-transformation. Instead, independent contributions of components to the variation of the complex character y are directly calculated from differences between coefficients of determination (r^2) of y by the primary characters. Eaton & Kyte, on the other hand, apply a stepwise multiple regression analysis to calculate, for each log-transformed component, residuals that are independent of all preceding components. These residuals are entered in a final multiple regression analysis with ln y as the dependent trait. This results in a complete partitioning of the determination of ln y over the mutually independent residuals of the log-transformed components.

Numerical example

In the following numerical example, dealing with leaf miner susceptibility of chrysanthemums, Eaton's method and ours are compared. The data in this example are used primarily to illustrate, explain and compare the two methods of component analysis. They do not serve to provide additional information on the genetic variation in leaf miner susceptibility. Therefore, in order not to distract the reader from the main subject, we have not considered the possible need for tests of linearity or weighted least square regression analyses. These procedures may improve the reliability of the calculated regression function. It is, of course, advisable to apply such procedures, where appropriate.

De Jong & Rademaker (1991) compared 12 genotypes of chrysanthemum (*Dendranthema grandiflora*) for the complex character 'susceptibility to the leaf miner *Liriomyza trifolii*' (y). Per genotype ten plants were raised in 15 cm pots on top of which a 14 cm cage was fitted. In each cage two male and two female adult leaf miners were placed. The following primary characters were observed per plant:
a: number of feeding punctures (fp)
b: number of visible eggs (observed microscopically in boiled leaves)
c: number of larvae (= number of mines)
y: number of pupae emerging from the leaves.
Characters a and b were observed on three of the ten plants (observation of character b is destructive) and characters c and y on the remaining seven plants. This has caused some aberrant results (more larvae than eggs) which were corrected for the pur-

poses of this example. In the most severely affected cultivars the number of mines could not be counted accurately. This resulted in some cases in a lower average value for number of larvae than for number of pupae. This too was corrected. The number of pupae emerging from the leaves (y) was taken as a measure of susceptibility to the leaf miner.

For a component analysis of 'number of pupae' the components are:

$$x_1 = no.\ fp.,\ x_2 = \frac{no.\ eggs}{no.\ fp.},\ x_3 = \frac{no.\ larvae}{no.\ eggs},\ x_4 = \frac{no.\ pupae}{no.\ larvae}$$

so that

$$no.\ fp. \cdot \frac{no.\ eggs}{no.\ fp.} \cdot \frac{no.\ larvae}{no.\ eggs} \cdot \frac{no.\ pupae}{no.\ larvae} = no.\ pupae.$$

The mean values per genotype of the primary characters and the calculated ratios are shown in Table 1. The correlations between number of pupae and the primary characters are shown in the fifth line of the correlation matrix in Table 2. The correlations between each component and its preceding primary character (in bold figures) illustrate how the components and (the product) of the preceding components are related.

Applying the method of calculation described before, the complementary determinations (bottom line of Table 2) indicate that the two most important components are x_2 and x_3, explaining 32% and 37%, respectively, of the variation of y. The components x_1 and x_4 have less influence, explaining 15% and 16%, respectively.

In the analysis according to Eaton & Kyte (1978) the multiplicative function $x_1 \cdot x_2 \cdot x_3 \cdot x_4 = y$ is log-transformed into the linear function $\ln x_1 + \ln x_2 + \ln x_3 + \ln x_4 = \ln y$.

– First $\ln x_2$ is regressed on $\ln x_1$. This yields the constant and regression coefficient in the first line of Table 3(a). The resulting residual (per genotype) represents the (log-transformed) number of eggs per feeding puncture as far as independent from number of feeding punctures. These residuals are indicated as r_{lnx2}.

– Next $\ln x_3$ is regressed on $\ln x_1$ and r_{lnx2}, yielding residuals r_{lnx3}.

– Regression of $\ln x_4$ on $\ln x_1$, r_{lnx2} and r_{lnx3} yields residuals r_{lnx4}.

– Finally, see Table 3(b), $\ln y$ is regressed on the independent components $\ln x_1$, r_{lnx2}, r_{lnx3} and r_{lnx4}. The successive increases in R^2 represent the coefficients of determination of $\ln y$ by individual components. This partitioning shows a major share of 0.56 for r_{lnx2}. This means that the variation of $\ln x_2$, in as far as independent from $\ln x_1$, explains most of the variation of $\ln y$. The influence of $\ln x_1$ on $\ln y$ is negligible. The coefficients of determination of the last two components are 0.23 and 0.21.

These results deviate considerably from those obtained by the method we propose (Table 2): they indicate a different component as the most influen-

Table 1. Observations (averages per plant) on the primary characters: a = number of feeding punctures; b = number of eggs; c = number of larvae and y = number of pupae of the leaf miner *Liriomyza trifolii* on twelve cultivars of chrysanthemum.
Components: x_1 = a; x_2 = b/a; x_3 = c/b; x_4 = y/c. Ranking orders in small figures.

Cultivar	x_1 = a	x_2	b	x_3	c	x_4	y
4131	2207 $_{11}$	0.0281 $_{2}$	62	0.468 $_{2}$	29	0.0207 $_{1}$	0.6 $_{1}$
Penny Lane	1756 $_{8}$	0.0165 $_{1}$	29	0.759 $_{8}$	22	0.0727 $_{2}$	1.6 $_{2}$
Reyellow	1436 $_{4}$	0.0327 $_{5}$	47	0.936 $_{10}$	44	0.0795 $_{3}$	3.5 $_{3}$
Statesman	1311 $_{1}$	0.0351 $_{6}$	46	0.674 $_{6}$	31	0.1258 $_{4}$	3.9 $_{4}$
Redemine	1333 $_{2}$	0.0300 $_{3}$	40	0.825 $_{9}$	33	0.3121 $_{5}$	10.3 $_{5}$
Delta	1791 $_{9}$	0.0396 $_{7}$	71	0.746 $_{7}$	53	0.3604 $_{6}$	19.1 $_{7}$
Refine	1744 $_{7}$	0.0614 $_{12}$	107	0.402 $_{1}$	43	0.4233 $_{7}$	18.2 $_{6}$
White Spider	1620 $_{6}$	0.0451 $_{9}$	73	0.548 $_{3}$	40	0.5400 $_{9}$	21.6 $_{9}$
Circus	1369 $_{3}$	0.0321 $_{4}$	44	0.977 $_{12}$	43	0.4581 $_{8}$	19.7 $_{8}$
Renine	1845 $_{10}$	0.0515 $_{11}$	95	0.653 $_{5}$	62	0.9806 $_{12}$	60.8 $_{10}$
Pink Pompon	2414 $_{12}$	0.0489 $_{10}$	118	0.602 $_{4}$	71	0.9296 $_{10}$	66.0 $_{12}$
Refla	1567 $_{5}$	0.0440 $_{8}$	69	0.942 $_{11}$	65	0.9523 $_{11}$	61.9 $_{11}$

Table 2. Coefficients of correlation (r) between the components (x_1,. . .,x_4) of the complex character y and the primary characters (a,. . .,y). Complementary determination (cd), derived from the r^2 (y, a,. . .,y) values as described in the text.

	a	b	c	y
x_1 = a	1.00**	0.65	0.36	0.39
x_2 = b/a	**0.23**	0.89**	0.67*	0.62
x_3 = c/b	− 0.58	− **0.58**	0.12	0.04
x_4 = y/c	0.28	0.68*	**0.88****	0.98**
y	0.39	0.69*	0.92**	1.00**
r^2(y,a,. . .,y)	0.15	0.47	0.84	1.00
cd(y,x_1,. . .,x_4)	0.15	0.32	0.37	0.16

r^2-values indicate fractions of determination of y by stages a, b, c and y. cd-values indicate increases in determination of y from stage to stage, attributable to intervening components x_1,. . .,x_4.

tial one. The two methods are thus not interchangeable and a choice must be made. This choice can only be based on an understanding of the causes of the difference in results.

There are two major differences between our approach and Eaton & Kyte's:

(1) Eaton applies log-transformation to enable him to carry out multiple linear regression. We do not transform and do not apply multiple linear regression.

(2) Eaton uses all preceding components successively as independent predictors on which the components are regressed. The residuals are entered in the stepwise multiple regression from which the coefficients of determination are obtained. Our approach is different: we consider the successive primary characters as the products of an increasing number of components. The contributions of successive primary characters to the variation of y (r^2-values) thus represent the combined contribution of the components involved. We calculate complementary determinations of y by individual components as differences between the r^2-values of the successive primary characters with respect to y.

Application of the Eaton & Kyte procedure without log-transformation, yields results (Table 4) much closer to ours (Table 2). This shows that the results from Eaton's method are not fundamentally different from ours when log-transformation is omitted. Thus the large differences between the two methods must be attributed mainly to the log-transformation. The fact that the coefficients of determination do not explain 100% of the variation of y confirms that multiple linear regression is not capable of achieving complete explanation of y when y is a multiplicative function of the components.

Table 3. (a) Coefficients of regression obtained when regressing the log-transformed components ln x_2, ln x_3 and ln x_4 on ln x_1 and on the residuals obtained after regression on preceding components (according to Eaton & Kyte, 1978). (b) Coefficients of regression obtained when regressing ln y on the residuals obtained from the regression functions in Table 3(a). Differences between the coefficients of determination (R^2) in the last column yield the partitioning of R^2 in the last line

(a) Dependent variable	Constant	Regression coefficient for			
		ln x_1	$r_{\ln x2}$	$r_{\ln x3}$	
ln x_2	− 5.739	0.329			
ln x_3	− 5.791	− 0.831	− 0.231		
ln x_4	− 0.799	− 0.069	2.527**	2.332	

(b) Dependent variable	Constant	Regression coefficient for				R^2
		ln x_1	$r_{\ln x2}$	$r_{\ln x3}$	$r_{\ln x4}$	
ln y	− 0.747	0.429				0.003
ln y	− 0.747	0.429	3.295**			0.558
ln y	− 0.747	0.429	3.295**	3.332*		0.792
ln y	− 0.747	0.429	3.295**	3.332*	1.000	1.000
Partitioning of R^2		0.00	0.56	0.23	0.21	1.000

Differences in the results of the two versions of Eaton's method, with and without transformation, are disturbingly large (Tables 3 and 4). As the components are to be used as criteria for parent selection, it is questionable whether the relative importance of the components, as indicated by the analysis of the log-transformed data, is relevant.

We prefer our method as it processes the data in the scale in which they are to be used and as it fully explains the variation of y without log-transformation. The multiplicative relationship between y and its components and the assumed additive inheritance of the components (or their residuals) are retained. These are essential features of a component analysis that should not be transformed away. Plant breeders can make active use of these two features to predict progeny performance and to achieve maximum recombinative heterosis; see Part II (Bos & Sparnaaij, 1993).

Uses of component analysis

In plant breeding, component analysis is mostly used to find selection criteria for yield that can be measured earlier and/or at less cost than yield itself. When this is the only objective, there is no need to pay attention to the nature and the sequence of the components. Any plant trait that can be measured accurately may be used as a selection criterion, provided it is sufficiently correlated with yield. The fact that a trait is correlated with yield does not imply that it is a component of yield.

When the objective is to select parent genotypes for breeding purposes, selection criteria must be chosen among the components on the basis of their mutually independent effects on yield, i.e. on the basis of their cd-values. These cd-values are obtained from a sample of genotypes representative for the available germplasm. A set of potential parents for a yield improvement programme would consist of genotypes of above average yield level with a high value for at least one of the influential components. This enables the breeder to produce hybrids in such a way that maximum recombinative heterosis is obtained. This is treated in Part II.

By including different agronomic treatments it is possible to obtain information about environmental effects on the relative contribution of individual components. Eaton et al. (1986) have extended their method in this direction.

Discussion

In our approach to the analysis of complex characters, the terms 'component' and 'component analysis' are used in a stricter sense than is customary in plant breeding literature. We consider a component to be a constituent part of a complex character, not just any character or factor affecting it. It will be generally accepted that the use of the word component for plant age, or for external factors affecting the complex character, is incorrect and confusing. It may be less obvious that, in our definition, the susceptibility to a certain disease or the sensitivity to drought is not a component of yield. Strictly speaking, number of flowers per plant is not a yield com-

Table 4. Coefficients of regression obtained when regressing y on x_1 and on the residuals obtained as illustrated in Table 3(a). Partitioning of R^2, comparable to the data in Table 3(b), but in this case without prior log-transformation.

Dependent variable	Constant	Regression coefficients for				R^2
		x_1	r_{x2}	r_{x3}	r_{x4}	
y	−23.92	0.0282				0.153
y	−23.92	0.0282	1131.6			0.446
y	−23.92	0.0282	1131.6*	100.9**		0.783
y	−23.92	0.0282	1131.6***	100.9***	61.19***	0.983
Partitioning of R^2		0.15	0.29	0.34	0.20	0.98

ponent either. It represents an earlier stage in the process of seed production and is the product of the same components as for number of seeds, apart from the final one (no. seeds/no. flowers).

A definition of a component corresponding with ours was already given by Thomas & Grafius (1976): 'strictly those characters which when multiplied together give yield exactly'. They also note that 'the term is often used loosely to include both "components" and "contributors to yield"' and that the main components 'may be further and logically subdivided, retaining the multiplicative principle'.

Similarly, in our view, component analysis is not the bringing together of a number of more or less relevant plant characters, in the hope of finding one that is closely correlated with the complex character. It is, instead, the subdivision of the complex character into two or more components, representing chronological or ontogenetical steps in the process of which the complex character is the final stage. Our aim is to find the component(s) that have the strongest influence on the genetic variation of the complex character.

The genetic control of components is bound to be simpler than that of complex characters. In our opinion, a component trait is more likely to be inherited additively. Components should therefore provide a better basis for parent selection and parent combination. In particular the pursuit of recombinative heterosis is more likely to be successful when we know the most important components and their mutual relationships. It was, in fact, the heterosis observed in diallel crosses showing specific combining ability, which led us to look for more adequate statistical procedures to determine the effects of individual components. The sequential yield component analysis, presented by Eaton & Kyte (1978), appeared the most suitable for our purpose. They used their method mainly in agronomic research in various crops. We have considered its application in plant breeding programmes but have eventually given preference to a simpler, more direct method for reasons given earlier.

The component analysis described in this paper requires that the components are introduced in the correct order. It is not always self-evident what is the correct order. In the case of relative growth rate, for example, crop physiologists tend to treat the net assimilation rate (= NAR) as the first component and the leaf area ratio (= LAR) as the second:

$$RGR = NAR \cdot LAR = \frac{dW/dt}{LA} \cdot \frac{LA}{W}$$

For a component analysis, however, LAR should be the first, independent variable and NAR the second, dependent one, because dW/dt depends on LA and should come after LA in the sequence of primary characters.

Another possible source of error in the calculated contributions is an inadequate sample underlying the analysis. The sample should involve an adequate number of representative genotypes. Per genotype a 'sufficient' number of plants should be observed.

The proposed method of analysis may show lack of correspondence between the coefficients of correlation (between individual components and the complex trait) and the corresponding complementary determination. In an analysis of RGR, for example, one may find a higher correlation (e.g. $r = 0.65$) with LAR than with NAR ($r = 0.20$) and conclude that RGR is largely determined by LAR (De Jong & Jansen, 1992). The complementary determination, however, takes into account the negative correlation between LAR and NAR. It shows that LAR is responsible for 42% of the variation of RGR and NAR for the remaining 58%, thus reversing the conclusion.

Yield component analysis has been applied in crop research in many different ways. In a comprehensive review Fraser & Eaton (1983) list and evaluate 15 different methods.

A simple analysis of covariance is the most commonly applied procedure (e.g. Goldy, 1988). However, as Fraser & Eaton observe correctly, (partial) correlation coefficients provide incomplete information about the nature of the relation between yield and its components because the components are mutually correlated.

Path coefficient analysis (Li, 1975) is a more ambitious method. It seeks to partition the correlation coefficients into their direct and indirect effects. Provided it is applied in a sensible manner, based on

knowledge of the causal relationships between the variables, it provides valuable information on the relative importance of the variables involved in a given process (e.g. Ottaviano & Camussi, 1981). For our form of component analysis, the choice of variables as used for a path coefficient analysis is not sufficiently delimited.

Sequential component analysis as applied to the yield of cranberries (Eaton & McPherson, 1975; Eaton & Kyte, 1978) and to the growth of beans (Joliffe et al., 1982) is a method that is directly applicable to our situation, even though it has not been applied in plant breeding sofar. As discussed, its major disadvantage is that it requires log-transformation.

Thomas & Grafius (1976) use component analysis 'for predicting offspring yield – surely the central problem of plant breeding' and thus go much further than Eaton and co-authors. Thomas & Grafius also make the components mutually independent but, as in our approach, they do not apply log-transformation. The variation for a particular component, calculated for orthogonalized data, is divided by the variation for untransformed data in order to obtain 'true relative genetic variances'. The present authors fear that, as a consequence of this method, a higher number of preceding components will lead to a lower independent variance of any given component. Thus, if the first component were to be subdivided into two components, the 'true relative genetic variance' of all other components would be lowered. We consider this incorrect as, in our view, the subdivision of any component, including the first one, should not affect the independent variance of the other components in any way.

The problem of 'compensation', a term often used to indicate the influence of the variation of one component on that of another, is discussed in detail by Adams (1967), Rasmusson & Cannell (1970), Lee & Kaltsikes (1972) and Grafius et al. (1976).

Hardwick & Andrews (1980), though quoting Thomas & Grafius (1976) in their paper, express the opinion that 'it is still not possible to quantify the degree of dependence between components'. They propose an index W, which is a function of the variances and covariances of the log-transformed yield components in such a way that:
$W = 0$ when there is complete compensation (negative correlation),
$W = 1$ when there is complete additivity (positive correlation,
$W = 0.5$ when there is full independence.

The specific problem we try to solve is the subject of a paper by Brown & Alexander (1991) on the analysis of variance and covariance of products. They discuss, amongst other cases, a 'situation where a response variable can be expressed as a product of random variables and where there is a need to assess which variables contribute most to the variation of the overall response'. Their treatment has certain features in common with ours but does not lead to all the practical answers we require. It does nevertheless merit a comparative study.

There appears to be little doubt that an efficient and reliable method of component analysis is an essential prerequisite for a successful genetic improvement of complex traits, whether by traditional or biotechnological methods. It is therefore of considerable interest to compare the scientific and practical merits of the various methods proposed in the literature. This goes beyond the scope of the present paper which is primarily concerned with the principle of component analysis and its application in plant breeding. We do not claim that the statistical method we have adopted is the best one available, but we found it to be simpler to use and easier to interpret in biological terms than alternative methods.

Acknowledgements

The authors wish to express their gratitude to Dr. J. de Jong (CPRO-DLO) for allowing us to use data from his research on the leaf miner in chrysanthemum. Dr. L.R. Verdooren and Mr. A. Otten (Agric. University, Wageningen) are thanked for their critical reading of the manuscript.

References

Adams, M.W., 1967. Basis of yield component compensation in crop plants with special reference to the field bean, *Phaseolus vulgaris*. Crop Science 7: 505–510.

Bos, I. & L.D. Sparnaaij, 1993. Component analysis of complex characters in plant breeding. II. The pursuit of heterosis. Euphytica 70: 237–245.

Brown, D. & N. Alexander, 1991. The analysis of variance and covariance of products. Biometrics 47: 429–444.

Eaton, G.W., P.A. Bowen & P.A. Joliffe, 1986. Two-dimensional partitioning of yield variation. HortScience 21: 1052–1053.

Eaton, G.W. & E.A. Mac Pherson, 1978. Morphological components of yield in cranberry. Hort. Res. 17: 73–82.

Eaton, G.W. & T.R. Kyte, 1978. Yield component analysis in the cranberry. Am. Soc. Hort. Sci. 103: 578–583.

Fraser, J. & G.W. Eaton, 1983. Application of yield component analysis to crop research. Field Crop Abstracts 36: 787–797.

Goldy, R.G., 1988. Variation in some yield determining components in muscadine grapes and their correlation to yield. Euphytica 39: 39–42.

Grafius, J.E., R.L. Thomas & J. Barnard, 1976. Effect of parental component complementation on yield and components of yield in barley. Crop Science 16: 673–677.

Hardwick, R.C. & D.J. Andrews, 1980. Genotypic and environmental variation in crop yield. A method of estimating the interdependence of the components of yield. Euphytica 29: 177–188.

Joliffe, P.A., G.W. Eaton & J. Lovett Doust, 1982. Sequential analysis of plant growth. New Phytologist 92: 287–296.

Jong, J. de & W. Rademaker, 1991. Life history studies of the leaf miner *Liriomyza trifolii* on susceptible and resistant cultivars of *Dendranthema grandiflora*. Euphytica 56: 47–53.

Jong, J. de & J. Jansen, 1992. Genetic differences in relative growth rate and partitioning growth components in *Chrysanthemum morifolium*. Scientia Hort. 49: 267–275.

Lee, J. & P.J. Kaltsikes, 1972. Diallel analysis of correlated sequential characters in durum wheat. Crop Science 12: 770–772.

Li, C.C., 1975. Path Analysis – A primer. Boxwood Press, 347 pp.

Ottaviano, E. & A. Camussi, 1981. Phenotypic and genotypic relationships between yield components in maize. Euphytica 30: 601–609.

Rasmusson, D.C. & R.Q. Cannell, 1970. Selection for grain yield and components of yield in barley. Crop Science 10: 51–54.

Snedecor, G.W. & W.G. Cochran, 1967. Statistical methods. 6th Ed. Iowa State Univ. Press, Ames, Iowa, pp. 593.

Thomas, R.L. & J.E. Grafius, 1976. Prediction of heterosis levels from parental information. Proc. Seventh Congress of Eucarpia; 173–180.