New Phytologist

# Methods

# The measure and mismeasure of reciprocity in heterostylous flowers

**W. Scott Armbruster[1,2], Geir H. Bolstad[3], Thomas F. Hansen[4], Barbara Keller[5], Elena Conti[5] and Christophe Pélabon[6]**

[1]School of Biological Sciences, University of Portsmouth, Portsmouth, PO1 2DY, UK; [2]Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK 99775, USA; [3]Norwegian Institute for Nature Research (NINA), Trondheim NO-7485, Norway; [4]Department of Biology, CEES & Evogene, University of Oslo, PB1016, Oslo 0316, Norway; [5]Department of Systematic and Evolutionary Botany, University of Zürich, Zollikerstrasse 107, Zürich 8008, Switzerland; [6]Institute of Biology, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology (NTNU), Trondheim 7491, Norway

Author for correspondence:
*W. Scott Armbruster*
*Tel: +44 (0)1983 615361*
*Email: scott.armbruster@port.ac.uk*

Received: *23 November 2016*
Accepted: *31 March 2017*

*New Phytologist* (2017) **215:** 906–917
**doi**: 10.1111/nph.14604

**Key words:** adaptive accuracy, floral dimorphisms, heterostyly, maladaptation, measurement theory, phenotypic load, pollination, *Primula*.

## Summary

- The goal of biological measurement is to capture underlying biological phenomena in numerical form. The *reciprocity index* applied to heterostylous flowers is meant to measure the degree of correspondence between fertile parts of opposite sex on complementary (inter-compatible) morphs, reflecting the correspondence of locations of pollen placement on, and stigma contact with, pollinators. Pollen of typical heterostylous flowers can achieve unimpeded fertilization only on opposite-morph flowers. Thus, the implicit goal of this measurement is to assess the likelihood of 'legitimate' pollinations between compatible morphs, and hence reproductive fitness.
- Previous reciprocity metrics fall short of this goal on both empirical and theoretical grounds.
- We propose a new measure of reciprocity based on theory that relates floral morphology to reproductive fitness. This method establishes a scale based on *adaptive inaccuracy*, a measure of the fitness cost of the deviation of phenotypes in a population from the optimal phenotype. Inaccuracy allows the estimation of independent contributions of maladaptive bias (mean departure from optimum) and imprecision (within-population variance) to the phenotypic mismatch (inaccuracy) of heterostylous morphs on a common scale.
- We illustrate this measure using data from three species of *Primula* (Primulaceae).

## Introduction

Measurement is the process by which we assign numbers to entities so that the mathematical relationships among numbers capture relevant empirical relationships among the entities (Krantz *et al*., 1971; Hand, 2004). Measurement theory reminds us that we need to remain cognizant of the purpose of our measurements when we develop biological metrics (Houle *et al*., 2011). Inferences about numbers must be translated into inferences about the original entities, and the validity of this process depends on the empirical relational structure being clearly defined. Failure to do so will render uncertain the actual meaning of the measurement. Importantly, the empirical relational structure defines the scale type of the measurement, that is, the type of numerical relationships that are meaningful in representing the empirical relationships (Stevens, 1968). This means that rescaling and number manipulation should be performed in a way that reflects the empirical relationships and retains the meaning of the measurement. These general remarks underline the importance of having a precise theoretical description of the physical/biological processes that generate the empirical relational structure to be measured.

When the principles of measurement theory are ignored or violated, the result is numerical 'measurements' that are disconnected from, or misrepresent, the empirical relationship they are meant to capture. Examples of such pseudo-quantification are common in the biological literature, and may reflect a general absence of awareness of measurement theory in many areas of biology (reviewed in Houle *et al*., 2011). Numerous examples of this problem can be found in the proliferation of intuitive indices devised to capture various biological phenomena, but without any principled attempt at justifying the mapping from biology to numbers. For example, Armbruster *et al.* (2014) recently pointed out that a menagerie of indices of integration and modularity has been proposed largely without any explicit attempt at stating what exactly is being measured. In the fields with which we are familiar, there do not seem to be any established methods or demand for such justification, although a small literature pointing out and discussing the problem is beginning to emerge (e.g. Wolman, 2006; Hansen & Houle, 2008; Frank, 2009, 2014;

Mitteroecker & Huttegger, 2009; Schneider, 2009; Wagner, 2010; Chevin, 2011; Hansen *et al.*, 2011; Houle *et al.*, 2011; Hansen, 2015; Tarka *et al.*, 2015; Morrissey, 2016).

Heterostylous flowers have intrigued evolutionary biologists since Darwin (1877) used them as evidence of adaptation by natural selection. Heterostyly ('reciprocal herkogamy') occurs in 28 families of flowering plants, has evolved independently multiple times (Barrett, 1992; Naiki, 2012), and has implications for understanding the origins, maintenance and evolutionary dynamics of plant mating systems (cf. Charlesworth & Charlesworth, 1979; Lloyd & Webb, 1992a,b). The reciprocal positions of the anthers and stigmas across intercompatible morphs are thought to promote disassortative (among-morph) pollination (Darwin, 1877; Lloyd & Webb, 1992b), and recent empirical work has borne this out (Keller *et al.*, 2014; Zhou *et al.*, 2015).

Here, we discuss various reciprocity indices developed for heterostylous flowers as yet another example of theory-free indices associated with violations of basic measurement theoretical principles. After showing that existing reciprocity indices suffer from shortcomings that stem from the absence of an explicit theory or even a clear statement of what the index is supposed to represent, we propose a new reciprocity measure based on the concept of adaptive accuracy, with reproductive fitness as the underlying currency. Reproductive fitness of individual phenotypes may be either modelled or measured, as explained below. From this, we establish a scale that gives quantitative meaning to the values and variation in the values of the numerical measure. We illustrate the uses and advantages of our measure with data from 15 populations of three of the species of *Primula* that Darwin (1877), himself, first examined in his ground-breaking investigations into heterostyly.

Reciprocity indices are attempts to characterize numerically the degree of spatial correspondence of 'compatible' sexual organs in heterostylous flowers. Classically, in heterostylous flowers (in this example, distylous, i.e. two flower morphs), unimpeded fertilization can be achieved primarily by the pollen arriving from flowers of the opposite morph. Pollen from the L-morph flowers (long style and short stamens; also termed 'pin') is more capable of germination, tube growth and fertilization on S-morph stigmas (short style and long stamens; also termed 'thrum') than is pollen from S-morph flowers, and vice versa. Thus, the pollination target of L-morph pollen is S-morph stigmas, and the pollination target of S-morph pollen is L-morph stigmas (Barrett, 2002). It should be noted that the terminology of previous authors, and that followed herein, refers to L-morph flowers as having long (or tall) styles with stigmas in a high position in the flower and with short stamens with anthers in a low position. S-morph flowers have short styles with stigmas in a low position in the flower and long (or tall or high) stamens with anthers in a high position (see Fig. 1).

For most researchers, the goal of a reciprocity index seems to be to generate a measurement that captures, at least implicitly, the fitness or pollination consequences of a departure from perfect correspondence of the fertile parts of opposite sex between compatible morphs of heterostylous flowers. This has generally involved some measure of the correspondence of the positions of the high stigmas in long-styled flowers with the high anther positions in short-styled flowers, and the correspondence of the positions of the low anthers in long-styled flowers with the low stigma positions in short-styled flowers (Webb & Lloyd, 1986). This approach is taken because the positions of the anthers and stigmas in the flower are thought to represent the location on the pollinators' bodies where pollen is deposited and retrieved (Barrett, 2002; but see Keller *et al.*, 2014). Despite the concept of reciprocity having a long and venerable history, with continual development of new metrics (e.g. Richards & Koptur, 1993; Eckert & Barrett, 1994; Faivre & McDade, 2001; Lau & Bosque, 2003; Sánchez *et al.*, 2008, 2013; Zhou *et al.*, 2015), measures of reciprocity have, to date, lacked any explicit mathematical connection to models of pollination, selection or adaptation.

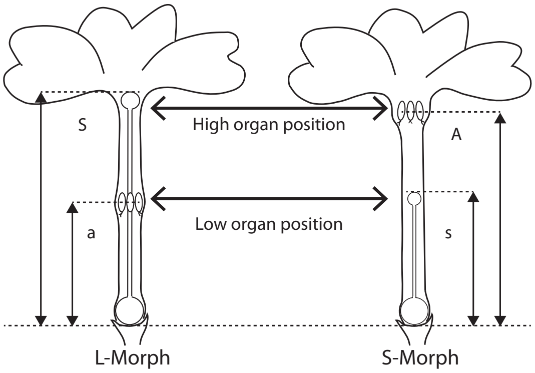

**Fig. 1** Diagram of distylous flowers (based on *Primula*) showing high anther (A), low anther (a), high stigma (S) and low stigma (s). Highest fitness is achieved when compatible pollen moves between organs at the same level, i.e. from A to S and from a to s. Figure modified, with permission, from Keller *et al.* (2012).

If the reciprocity index is meant to capture the ability to achieve disassortative pollinations and the connections of this ability to reproductive fitness, it can be measured as an accuracy around an optimum defined as the phenotype achieving the highest level of disassortative pollination. Assuming the pollinators are most efficient in transferring pollen to compatible stigmas when stigmas contact them in the same position as the pollen-donating anthers, the optimum is determined as matching positions of opposite-morph anthers and stigmas. Increasing deviation from perfect match can then be assumed to lower the probability of pollen transfer (Haller *et al.*, 2014), and thus seed set (Brys & Jacquemyn, 2015) and fitness.

Adaptive inaccuracy provides a scale in units of expected fitness cost or 'phenotypic load' (i.e. maladaptation) resulting from the departure of sampled phenotypes in a population from the optimal phenotype for that population (Armbruster *et al.*, 2004, 2009; Hansen *et al.*, 2006; Pélabon & Hansen, 2008; Pélabon *et al.*, 2012; Opedal *et al.*, 2016). Except when based on empirical fitness surfaces, adaptive inaccuracy is not a direct measure of fitness, but rather provides a scale whereby different traits or populations can be compared in units of the difference in their relative fitness or load if they were under quadratic stabilizing selection of the same strength. It should be noted that we refer to the general concept and mathematical approach as 'adaptive accuracy', but the measurements themselves are 'inaccuracies', that is, deviation from the optimum.

## Description

### A critical review of reciprocity measures

The concept of reciprocity begins with Darwin. He devoted two papers (1862, 1864) and a book (1877) to describing the biology of heterostylous flowers. Darwin suggested that the reciprocal arrangement of anthers and stigmas of complementary morphs mechanically promoted compatible ('legitimate') pollinations and thereby enhanced both female and male reproductive fitness (because intra-morph pollinations produce few or no seeds in most systems). Darwin (1862, p. 92; 1877, p. 33) defined reciprocity of sexual organs qualitatively by the similarity of heights of reciprocal organs. Implicit in Darwin's presentation is the idea that maladaptation is captured by the degree of deviation between heights of correspondingly placed reciprocal organs in opposite morphs. Darwin's argument was based on observations that the height of the anther (as determined by the stamen length) establishes where pollen is placed on a (dead) bumble bee whose proboscis was inserted into the floral tube of *Primula* flowers (Darwin, 1862, 1877). This has recently been confirmed in detail with living bees visiting *Primula* (Keller *et al.*, 2014). Various studies have supported this model, and thereby the functional significance and adaptive origins of reciprocity (see reviews in Vuilleumier, 1967; Ganders, 1979; Barrett, 1990, 2002; Barrett *et al.*, 2000).

The first attempt at the quantification of reciprocity appears to be that of Richards & Koptur (1993), who published a difference-based index based on unpublished work by J. H. Richards, D. G. Lloyd and S. C. H. Barrett. They examined departure of organs from reciprocity (equal heights; presumably maximum pollination fitness) and, in order to compare species of Rubiaceae with different-sized flowers, they scaled the difference in reciprocal organ heights by the sum of the means of the reciprocal organs. This gave two separate, but comparable, reciprocity measures (*R*) for the tall (= high) and short (= low) organs:

$$R_{\text{tall}} = \frac{(\overline{A} - \overline{S})}{(\overline{A} + \overline{S})}, \qquad \text{Eqn 1}$$

$$R_{\text{short}} = \frac{(\overline{a} - \overline{s})}{(\overline{a} + \overline{s})}, \qquad \text{Eqn 2}$$

where $\overline{A}$ is the population mean height of anthers on tall stamens, $\overline{S}$ is the mean height of stigmas on tall pistils, $\overline{a}$ is the mean height of anthers on short stamens and $\overline{s}$ is the mean height of stigmas on short pistils (as illustrated for *Primula* in Fig. 1). With these indices, perfect reciprocity is 0, that is when the anthers and stigmas of the reciprocal morphs are of exactly the same mean height. Because this index is calculated on a proportional scale, a 1 mm change in tall organs results in a smaller change in reciprocity than does a 1 mm change in short organs. Except for 'facilitating' interspecific comparisons, no explicit justification was given for this choice of scale. One could perhaps imagine a probabilistic model of pollen transfer and argue that the probability of pollen transfer also scales with organ size. The main problem in terms of measurement protocol is that Richards & Koptur (1993) did not specify what the index is supposed to measure quantitatively, and did not relate their choice to pollination rates, fitness or any other biologically relevant scale. Furthermore, as pointed out by Sánchez *et al.* (2008, 2013), the Richards & Koptur index does not account for the influence of phenotypic variation among flowers in the population on pollen transfer.

In the following year, Eckert & Barrett (1994) presented a single measure of reciprocity that combines the reciprocities of short and tall organs:

$$R = \frac{(\overline{A} - \overline{a})}{(\overline{S} - \overline{s})}, \qquad \text{Eqn 3}$$

where $\overline{A}$, $\overline{a}$, $\overline{S}$ and $\overline{s}$ are as above. Perfect reciprocity was indicated by $R = 1$, that is when the difference between the high and low anthers is equal to the difference between the high and low stigmas. This index has some intuitive shortcomings, however, including showing high reciprocity even when the positions of the high and low anthers do not match the positions of the high and low stigmas, but the difference between anthers equals the difference between stigmas. Eckert & Barrett (1994) also did not specify exactly what the index was meant to measure. Without a model of the relationship between the underlying biological entities and the index, it is not possible to judge the metric or to specify where the intuitive shortcomings come from. However, Eckert & Barrett (1994) did recognize the importance of flower variation within the population, and they proposed a separate precision index based on averaging the coefficients of variation (CVs) of the individual morphs. For two morphs together, this is

$$PI = (\mathrm{CV}_L + \mathrm{CV}_S)/2. \qquad \text{Eqn 4}$$

This is a mean-scaled measure of variation, but not strictly on the same scale as their reciprocity index. How one is to combine or compare *R* and PI is not clear. Furthermore, the averaging operation was not justified and is problematic because CVs are not expected to combine additively. Although it could have made sense to average variances, which are additive when their arguments are independent, we see no obvious case for averaging CVs.

More recently, Sánchez *et al.* (2008) proposed to incorporate variation in the reciprocity index by including all inter-individual relationships in the sample population:

$$r_y = \frac{1}{nm}\sum_{i}^{n}\sum_{j}^{m}\left(\frac{|A_i - S_j|}{\overline{X}}\right), \qquad \text{Eqn 5}$$

where $r_y$ (termed $r_a$ in the original paper) is the mean level of reciprocity at level *y* (high or low), $A_i$ and $S_j$ are the heights of the anthers and stigmas of opposite morphs for individual flowers *i* and *j*, $\overline{X}$ is the mean of all organ lengths, with one observation or mean taken per flower (one stigma height or the mean and one anther height or the mean per flower), and $n$ is the number of anther-height values and $m$ the number of stigma-height values included. It should be noted that this index is on a proportional scale, but the scaling is by the joint mean of all traits. The authors explain this choice in that it allows comparisons across both tall and short organs. However, there is no explicit link of the reciprocity measure to fitness, pollination rates or anything that could provide it with a biologically meaningful scale.

In the second step, Sánchez *et al.* (2008) estimated an overall reciprocity by calculating the Euclidian distance from zero of the two reciprocity indices:

$$r = \sqrt{r_L{}^2 + r_S{}^2}. \qquad \text{Eqn 6}$$

The use of the Euclidian distance to combine the two reciprocity indices for the short, $r_S$, and long, $r_L$, organs was not given a theoretical justification and is questionable in our opinion. Indeed, considering that deviation from reciprocity has a negative effect on fitness, one can ask why a decrease in fitness generated on the short and long organs would be additive on a square scale and not directly on the original scale. If, for example, the imperfect reciprocity in the short organ represents a decrease of two seeds on average and the imperfect reciprocity in the long organ represents a decrease of three seeds, the final costs estimated by the index from Sánchez *et al.* will not be five seeds but, instead, 3.6 ($\sqrt{2^2 + 3^2}$). Of course, the imperfect reciprocity may not have been intended to translate into number of seeds lost, but the choice of the Euclidian distance in order to combine the effects of imperfect reciprocity on the short and long organs remains to be justified.

In the third step, Sánchez *et al.* (2008) introduced the standard deviation of $r$ as a way to account for the phenotypic variation among individuals. For each level (short and long organs), they estimated the standard deviation as:

$$\mathrm{SD}_{r_y} = \sqrt{\frac{1}{nm}\sum_i^n \sum_j^m \left(\frac{|A_i - S_j|}{\overline{X}} - r_y\right)^2}, \qquad \text{Eqn 7}$$

and they calculated an average standard deviation for the short and long organs combined as:

$$\mathrm{SD}_r = (\mathrm{SD}_{r_L} + \mathrm{SD}_{r_S})/2. \qquad \text{Eqn 8}$$

Using the arithmetic mean for the calculation of the average of the two standard deviations implies that standard deviations are additive, which is rarely the case, in contrast with variances, as mentioned above. Once again, a justification for the mathematical operation is simply missing.

In the final step, the total reciprocity, $R$, was obtained by multiplying the arithmetic mean of the standard deviations for long and short organs ($\mathrm{SD}_r$) by the reciprocity index $r$:

$$R = r \times \mathrm{SD}_r. \qquad \text{Eqn 9}$$

The use of the multiplication is arbitrary here. Multiplying $r$ by the average standard deviation implies that the consequences of a deviation from perfect reciprocity of 2 mm, for example, should be twice as big when the standard deviation is twice as large. Conversely, even a large deviation from perfect reciprocity will have almost no effect on the total reciprocity ($R$) if the standard deviation is close to zero. It is also important to note that measures of variation are incorporated into the metric twice: (1) by deriving an initial metric using iterative calculations based on individual measurements (reflecting the distribution of differences); and (2) by multiplying this metric by its standard deviation.

In a later paper, Sánchez *et al.* (2013) modified their index arithmetically to make its variation more intuitive, so that large values mean greater reciprocity rather than lower:

$$R_2 = 1 - (R \times 10), \qquad \text{Eqn 10}$$

where $R$ is the index of reciprocity of Sánchez *et al.* (2008). However, despite a possible heuristic value, this arithmetic manipulation was also not given a theoretical justification.

Another approach to the quantification of reciprocity was developed by Lau & Bosque (2003) and used by Keller *et al.* (2012) and Zhou *et al.* (2015). This method quantifies the overlap of the distributions of anther and stigma positions of reciprocal morphs using an index of distributional overlap. Although this approach captures some aspects of both bias and imprecision, it has no explicit theoretical relationship to reproductive fitness and applies no penalty for imprecision. The index fails by deviating from any implicit concept of pollination fitness whenever the distributions are broad (low precision). In this situation, the index will show high 'reciprocity' (distributions of reciprocal organs largely overlap) even though the average distance between reciprocal structures is very large.

The common thread in all these attempts is that insufficient attention has been paid to the relationship between the behaviour of the numbers and the properties they are meant to represent. In the next section we develop an example of how this can be done.

## Reciprocity as adaptive accuracy

### Application of the adaptive accuracy concept to reciprocity

Reciprocal herkogamy (morph reciprocity) can be viewed as an adaptation promoting compatible pollination and reproductive fitness, as Darwin and most authors since have argued (see, for example, Simón-Porcar *et al.*, 2015; Zhou *et al.*, 2015). This means that the reproductive fitness of individuals with any particular anther position is determined by the distribution of stigma positions among its potential mates, weighted by its fitness in relation to each, and vice versa for stigma positions. As individuals of any given morph or genotype vary in their exact anther/stigma position, we also have to consider the fitness consequences of this variation and not just the mean positions. In this situation, we can use adaptive inaccuracy, which is designed to measure the degree of maladaptation of a morph or genotype on a fitness scale that accounts for both the mean and variance of the phenotypic values of the morph (Armbruster *et al.*, 2004; Hansen *et al.*, 2006). This was expanded later to also include variation in the optimum (Armbruster *et al.*, 2009) and more general fitness functions (Pélabon *et al.*, 2012). If we assume, for the moment, a quadratic form of the fitness function,

$$\frac{W(z;\theta)}{W(\theta;\theta)} = 1 - s(z-\theta)^2, \qquad \text{Eqn 11}$$

where $\frac{W(z;\theta)}{W(\theta;\theta)}$ is the fitness of a phenotype $z$ relative to the fitness, $W(\theta;\theta)$, at an optimum $\theta$, and $s$ is the strength of stabilizing selection around the optimum (Fig. 2), the adaptive inaccuracy is:

$$\text{Inaccuracy} = \mathrm{E}[(z-\theta)^2] = (\mathrm{E}[z] - \mathrm{E}[\theta])^2 + V_z + V_\theta, \qquad \text{Eqn 12}$$

where $\mathrm{E}[z] - \mathrm{E}[\theta]$ is the bias in adaptation, defined as the difference between the expected morph value, $\mathrm{E}[z]$, and the expected optimal value, $\mathrm{E}[\theta]$ (e.g. the difference between mean anther position and mean stigma position), $V_z$ is the variance in the trait (e.g. anther position) and $V_\theta$ is the variance in the target optimum (e.g. stigma position).

In this form, the inaccuracy is on a squared distance scale in units of trait units squared. To make this meaningful as a measure of maladaptation, we can use the assumption of a quadratic fitness function to map inaccuracy to fitness (or load) relative to maximum fitness. For a phenotype, $z$, the load, $L$, is defined as:

$$L(z;\theta) = \frac{W(\theta;\theta) - W(z;\theta)}{W(\theta;\theta)}, \qquad \text{Eqn 13}$$

from which it follows that the inaccuracy is directly proportional to the load:

$$\text{Inaccuracy} = \mathrm{E}[(z-\theta)^2] = \frac{1}{s}\mathrm{E}[L(z;\theta)], \qquad \text{Eqn 14}$$

and a doubling of the inaccuracy implies a doubling of the load regardless of $s$. This establishes a scale for comparisons of inaccuracies in terms of fitness. This scale also allows a counterfactual interpretation of inaccuracy as the load that would ensue if the trait were under quadratic stabilizing selection of strength $s$. A value of $s = 1$ trait units squared means that the inaccuracy equals the load. It should be noted that $s$ is not equal to the usual quadratic selection gradient, $\gamma$, defined as the expected value of the second derivative of fitness relative to the mean with respect to the trait. When the true fitness function is as given by Eqn 11, the two are related as:

$$|\gamma| = 2\frac{W(\theta;\theta)}{\mathrm{E}[W(z;\theta)]}|s| = 2\frac{|s|}{1-\mathrm{E}[L(z;\theta)]}, \qquad \text{Eqn 15}$$

which can be used to compute the load predicted from a given stabilizing selection gradient and level of inaccuracy. As we show below, this 'load' scale can be extended to specified general fitness functions.

In distylous populations comprising L-morph and S-morph plants, seeds are produced by crosses between flowers of the two morphs, but with reduced or zero fertility by crosses between flowers of the same morph. Let us assume that the length of the stamen, or corolla plus stamen in epipetalous flowers, determines the height of the anther above the reward or other relevant landmark, and this height, in turn, determines where pollen is placed on the pollinator (see Keller *et al.*, 2014). Similarly, the length of the pistil determines the height of the stigma, which, in turn, determines where the stigma touches the pollinator to pick up pollen. Under these assumptions, we can estimate four adaptive inaccuracies by the use of Eqn 12:

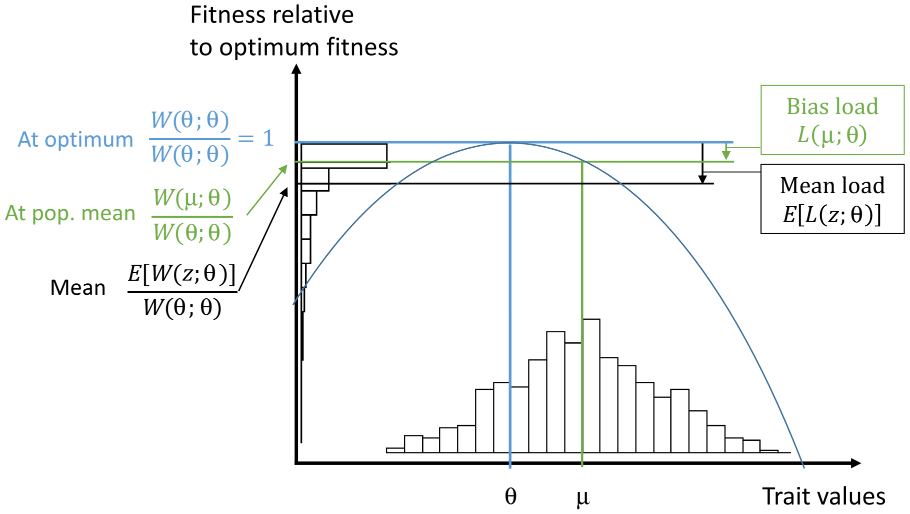
Fitness relative to optimum fitness
At optimum $\frac{W(\theta;\theta)}{W(\theta;\theta)} = 1$
At pop. mean $\frac{W(\mu;\theta)}{W(\theta;\theta)}$
Mean $\frac{E[W(z;\theta)]}{W(\theta;\theta)}$
Bias load $L(\mu;\theta)$
Mean load $E[L(z;\theta)]$
θ
μ
Trait values

**Fig. 2** Relationship between trait values, fitness and load assuming the quadratic fitness function $\frac{W(z;\theta)}{W(\theta;\theta)} = 1 - s(z-\theta)^2$ in blue. The distribution of trait values (horizontal histogram), with mean given by $\mu$, is transformed into a distribution of fitness values (vertical histogram) using the quadratic fitness function with an optimum at trait value $\theta$. The green arrow labelled 'At pop. mean' refers to the fitness accrued at the population mean. $\frac{W(z;\theta)}{W(\theta;\theta)}$ is the fitness of a phenotype $z$ relative to the fitness, $W(\theta;\theta)$, at an optimum $\theta$.

L-morph inaccuracies:

$$\text{Male Inaccuracy}_{\text{L-morph}} = (\bar{a} - \bar{s})^2 + V_a + V_s, \qquad \text{Eqn 16}$$

$$\text{Female Inaccuracy}_{\text{L-morph}} = (\bar{S} - \bar{A})^2 + V_S + V_A. \qquad \text{Eqn 17}$$

S-morph inaccuracies:

$$\text{Male Inaccuracy}_{\text{S-morph}} = (\bar{A} - \bar{S})^2 + V_A + V_S, \qquad \text{Eqn 18}$$

$$\text{Female Inaccuracy}_{\text{S-morph}} = (\bar{s} - \bar{a})^2 + V_s + V_a, \qquad \text{Eqn 19}$$

where $A$ is the height of high anthers on tall stamens, $S$ is the height of high stigmas on tall pistils, $a$ is the height of low anthers on short stamens, $s$ is the height of low stigmas on short pistils, letters with bars are the corresponding population means and $V$ represents the corresponding variances.

Because both trait and target variances are included (Armbruster *et al.*, 2009), the male inaccuracy of the L-morph and the female inaccuracy of the S-morph are mathematically identical, as are the female inaccuracy of the L-morph and the male inaccuracy of the S-morph. Because male and female components of fitness contribute equally to population mean fitness, these inaccuracy terms should be weighted by 0.5 and then added to obtain the joint (male + female) inaccuracy. The sum of the male and female inaccuracies can then be used to estimate separately the joint inaccuracy of the high (L-morph stigmas and S-morph anthers) and low (L-morph anthers and S-morph stigmas) organs.

$$\text{Inaccuracy}_{\text{high organs}} = (\bar{A} - \bar{S})^2 + V_A + V_S, \qquad \text{Eqn 20}$$

$$\text{Inaccuracy}_{\text{low organs}} = (\bar{a} - \bar{s})^2 + V_a + V_s. \qquad \text{Eqn 21}$$

Importantly, this measure brings the effects of mean deviation from the optimum and variance of organ position onto the same scale, so that their relative effects can be compared and combined. Although high- and low-organ inaccuracies are additive, whether and how they should be combined for the estimation of overall population inaccuracy depends on morph frequencies and the questions being addressed (see discussion below).

An important consideration in using these measures is whether and how to standardize the traits. The unit of the inaccuracy is trait units squared. The unit can be adjusted or eliminated by a variety of standardization procedures. These include proportional scales, obtained through mean standardization or log transformation, and ‘variance’ scales, obtained by standardizing with measures of trait variation. The latter is problematic in this case, because we want to capture the effects of different levels of variation (precision), which would be lost if variance standardization were employed. The choice between an absolute (unstandardized) and a proportional scale is more difficult. The correct choice in scaling is also influenced by the choice of fitness function and by whether fitness declines quadratically with absolute or proportional deviation of the trait from the optimum.

This choice becomes particularly pertinent when comparing the high and low organs. When using a proportional scale (e.g. by dividing the index with the overall trait mean or the mean of each organ type), one assumes that a percentage difference in organ position would mean the same in terms of the fitness decrease for high and low organs, whereas, using an absolute scale, one assumes that a 1 mm difference, for example, would mean the same in terms of fitness for high and low organs. The former might be a better choice if the pollinators or their behaviours scale with organ height, so that the fitness surface is less downwardly curved per millimetre difference for high organs than for low organs. The latter might be a better choice if interacting pollinators and their behaviours are the same for both high and low organs. For the comparison of organs of different heights within a population, it might be better to use an absolute scale. For the comparison of populations or species, it may be more appropriate to mean standardize by the average organ height. We leave the choice of scale open, but emphasize that this choice is not just a matter of removing units or statistical convenience; it entails biological assumptions, and these assumptions need to be made explicit.

**Reciprocity as a fitness surface** Improved measures of reciprocity could be obtained if there are empirical or theoretical grounds to further specify the fitness function. As discussed above, this could include biological reasons for choice of trait scale or strength of stabilizing selection. More generally, Pélabon *et al.* (2012) developed a measure of inaccuracy for an arbitrarily specified fitness function that could be adapted for reciprocity. The basis for this is to compute the fitness load ($L$) of a morph with respect to an optimal state, as defined in Eqn 13, where $W(z; \theta)$ is now an arbitrary fitness function for a trait $z$, assuming an optimal value at $z = \theta$ (where maximum fitness is $W(\theta; \theta)$). Applying this to a high anther with length $A$ relative to a given high stigma of length $S$, the load is

$$L(A; S) = \frac{W(S; S) - W(A; S)}{W(S; S)}, \qquad \text{Eqn 22}$$

where we have assumed that a perfect match, $A = S$, is optimal. To develop a measure of reciprocity, we need to take account of the fact that, in addition to variation in the focal organs, there is variation in the target organs, thus presenting a variable optimum. Pélabon *et al.* (2012) proposed to compute the inaccuracy as $\mathrm{E}[L(z; \theta)]$, where the expectation is taken over both the trait, $z$, and the optimum, $\theta$. For the high anthers, this can be broken down as:

$$\begin{aligned}\mathrm{E}[L(A; S)] = & L(\bar{A}; \bar{S})\\ & + \mathrm{E}_A[\mathrm{L}(A; \bar{S})] - \mathrm{L}(\bar{A}; \bar{S})\\ & + \mathrm{E}_S[\mathrm{L}(\bar{A}; S)] - \mathrm{L}(\bar{A}; \bar{S})\\ & + \mathrm{E}_A\mathrm{E}_S[\mathrm{L}(A; S)] - (\mathrm{E}_A[\mathrm{L}(A; \bar{S})] - \mathrm{L}(\bar{A}; \bar{S}))\\ & - (\mathrm{E}_S[\mathrm{L}(\bar{A}; S)] - \mathrm{L}(\bar{A}; \bar{S})) - \mathrm{L}(\bar{A}; \bar{S}),\end{aligned} \qquad \text{Eqn 23}$$

where the first line is the maladaptive bias as a result of a mismatch of the means of the anther and stigma. The second line is the adaptive imprecision as a result of variation in the anther position. The third line is the adaptive imprecision as a result of variation in the target stigma position, and the last two lines represent the result of interactions between the anther and stigma positions of mating individuals (this interaction term will vanish if between-morph mating is random with respect to trait position and the fitness function is quadratic). This equation is symmetric with respect to $A$ and $S$, and hence gives the inaccuracy for both anthers and stigma. It can therefore be used as a measure of the reciprocity of high organs in general. The same argument applies to low organs simply by replacing upper case $A$ with lower case $a$ and upper case $S$ with lower case $s$.

To use this measure, it is necessary to specify a fitness function, $\mathrm{W}(z; \theta)$, that describes the fitness of any combination of anther and stigma positions. This could be based on functional arguments derived from pollination mechanics or from empirical measurements. It should be noted that the inaccuracy in this case is measured in units of fitness load.

**Inaccuracy at the level of individuals** Thus far, we have treated inaccuracy as a population property, but, as discussed in Hansen *et al.* (2006), it can also be applied to individuals or genotypes for which the level of adaptation can be assessed in terms of imprecision and bias in their realized phenotypes relative to an adaptive optimum. Hansen *et al.* (2006) used this to assess the effects of developmental stability measured as fluctuating asymmetry on individual- and population-level adaptive imprecision in animals (see also Pélabon & Hansen, 2008). Individual plants with multiple flowers provide a good system to assess individual-level imprecision. On the quadratic fitness scale, the individual-level imprecision contributes additively to population-level imprecision, and hence to inaccuracy. It will therefore often be feasible to decompose population-level imprecision into within- and among-individual contributions, where the former stem from developmental instability and plasticity, and the latter from genetic and environmental variation across individuals (as illustrated in Pélabon *et al.*, 2012).

In the case of heterostyly, within-individual imprecision resulting from developmental instability and microenvironmental effects may often be an important contributor to population-level imprecision. This effect can be measured by computing the variance in anther and stigma positions across flowers within single plants.

## Results and Discussion

### An empirical example: accuracy of reciprocity in *Primula*

As a heuristic example of the accuracy measure, we reanalysed the data published in Keller *et al.* (2012). These data are from five populations of each of three species of *Primula* (*P. veris, P. elatior* and *P. vulgaris*) in which the heights of both high and low anthers and stigmas were measured (Fig. 1; Table 1). To calculate the different measures of adaptive inaccuracy, we used Eqns 20 and 21. In addition to presenting the unstandardized inaccuracies, we also calculated and present the inaccuracies standardized by the squared mean of all anther and stigma heights in each population to facilitate comparison across populations and species (Table 2). To obtain 95% confidence intervals, we bootstrapped 1000 times at the level of the individual plant.

In Table 2, we present the bias, imprecision and inaccuracy values for each population broken down by organ type. The overall levels of inaccuracy vary both among species and among populations, ranging from *c.* 3 to 8 $\mathrm{mm}^2$ on a metric scale and 2 to 9% on a mean-standardized scale. Interpreted as loads (Eqn 14), these values indicate that the fitness is reduced by 3–8% assuming stabilizing selection of strength $s = 0.01\ \mathrm{mm}^{-2}$, or by 2–9% assuming that the mean-scaled stabilizing selection is $s_\mu = 1$.

A mean-scaled $s_\mu = 1$ means that a load of 2% would result from an individual phenotype being shifted 14% away from the

**Table 1** Descriptive statistics: sample size for the two morphs (long and short), mean organ height for each type of organ (high stigmas *S*; high anthers *A*; low stigmas *s*; low anthers *a*), mean organ height across all organ types and the variance (Var) of each organ type

| Species | Locality | *n* L-morph | *n* S-morph | Mean *S* (mm) | Mean *A* (mm) | Mean *s* (mm) | Mean *a* (mm) | Average organ height (mm) | Var(*S*) ($\mathrm{mm}^2$) | Var(*A*) ($\mathrm{mm}^2$) | Var(*s*) ($\mathrm{mm}^2$) | Var(*a*) ($\mathrm{mm}^2$) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Primula elatior* | Küsnacht | 18 | 17 | 11.84 | 12.79 | 6.00 | 6.02 | 9.16 | 0.866 | 1.41 | 0.266 | 0.234 |
| | Kollbrunn | 30 | 26 | 12.00 | 13.00 | 6.15 | 6.52 | 9.41 | 0.840 | 0.481 | 0.328 | 0.173 |
| | Zurich 1 | 29 | 28 | 13.00 | 14.29 | 6.86 | 7.26 | 10.35 | 1.90 | 1.58 | 0.594 | 0.710 |
| | Zurich 2 | 22 | 19 | 13.41 | 12.65 | 6.07 | 6.87 | 9.78 | 4.49 | 1.96 | 0.205 | 0.498 |
| | Thörigen | 34 | 28 | 12.40 | 12.50 | 5.47 | 6.86 | 9.34 | 1.67 | 1.48 | 0.711 | 0.280 |
| | Average | | | 12.53 | 13.05 | 6.11 | 6.70 | 9.61 | 1.95 | 1.38 | 0.421 | 0.379 |
| *Primula veris* | Seewis | 30 | 26 | 14.11 | 14.53 | 8.78 | 9.25 | 11.67 | 1.18 | 0.807 | 1.07 | 0.532 |
| | Montreux | 31 | 25 | 14.73 | 14.82 | 8.77 | 9.12 | 11.87 | 0.694 | 0.642 | 0.487 | 0.525 |
| | Kollbrunn | 28 | 31 | 13.28 | 13.80 | 8.21 | 8.91 | 11.05 | 0.903 | 1.87 | 0.772 | 0.388 |
| | Pfungen | 30 | 30 | 14.46 | 14.55 | 7.89 | 9.31 | 11.55 | 1.32 | 1.73 | 0.407 | 0.234 |
| | Glarus | 29 | 28 | 14.87 | 14.89 | 8.16 | 10.10 | 12.01 | 0.928 | 0.393 | 0.225 | 0.380 |
| | Average | | | 14.29 | 14.52 | 8.36 | 9.34 | 11.63 | 1.00 | 1.09 | 0.592 | 0.412 |
| *Primula vulgaris* | Pompagles | 15 | 9 | 16.30 | 16.22 | 9.10 | 10.07 | 12.99 | 1.32 | 2.05 | 0.104 | 0.446 |
| | Arogno | 26 | 27 | 14.97 | 16.41 | 8.53 | 9.02 | 12.24 | 1.04 | 1.59 | 0.429 | 0.615 |
| | Vaglio | 27 | 29 | 16.31 | 17.58 | 8.75 | 9.47 | 13.03 | 1.05 | 2.35 | 0.354 | 0.336 |
| | Collonges | 27 | 29 | 15.48 | 16.05 | 8.58 | 9.47 | 12.39 | 0.806 | 3.26 | 0.527 | 1.16 |
| | Lausanne | 28 | 28 | 16.10 | 17.21 | 9.23 | 9.16 | 12.93 | 1.78 | 3.50 | 0.613 | 0.883 |
| | Average | | | 15.83 | 16.69 | 8.84 | 9.44 | 12.72 | 1.20 | 2.55 | 0.405 | 0.689 |

**Table 2** Estimates of inaccuracy and its different components across species and populations (95% confidence intervals in parentheses)

| Locality | Organ type | Inaccuracy | Maladaptive bias$^2$ | Variance anther | Variance stigma | Total inaccuracy | Mean$^2$-standardized total inaccuracy |
|---|---|---|---|---|---|---|---|
| *Primula elatior* | | | | | | | |
| Küsnacht | High | 86 (72, 92)% | 24 (0, 66)% | 38 (9, 65)% | 23 (7, 49)% | 3.7 (1.7, 6.0) mm$^2$ | 4.4 (1.9, 9.3)% |
| | Low | 14 (8, 28)% | 0 (0, 10)% | 6 (3, 12)% | 7 (2, 13)% | | |
| Kollbrunn | High | 78 (59, 88)% | 34 (6, 61)% | 16 (6, 31)% | 28 (10, 44)% | 3.0 (1.5, 4.9) mm$^2$ | 3.3 (1.7, 6.4)% |
| | Low | 22 (12, 41)% | 4 (0, 22)% | 6 (3, 10)% | 11 (2, 26)% | | |
| Zurich 1 | High | 78 (59, 88)% | 25 (1, 56)% | 24 (8, 40)% | 28 (14, 43)% | 6.6 (3.8, 9.9) mm$^2$ | 6.2 (3.7, 12.2)% |
| | Low | 22 (12, 41)% | 3 (0, 20)% | 11 (4, 19)% | 8 (2, 15)% | | |
| Zurich 2 | High | 84 (73, 92)% | 7 (0, 53)% | 23 (5, 38)% | 53 (15, 77)% | 8.4 (4.9, 12.6) mm$^2$ | 8.8 (4.1, 15.9)% |
| | Low | 16 (8, 27)% | 8 (1, 19)% | 6 (2, 10)% | 2 (1, 5)% | | |
| Thörigen | High | 52 (36, 71)% | 0 (0, 18)% | 24 (12, 33)% | 27 (11, 45)% | 6.1 (3.7, 8.8) mm$^2$ | 7.0 (3.8, 10.7)% |
| | Low | 48 (29, 64)% | 32 (17, 50)% | 5 (2, 9)% | 11 (4, 17)% | | |
| Average* | High | 75% | 15% | 25% | 35% | 5.5 mm$^2$ | 5.9% |
| | Low | 25% | 10% | 7% | 7% | | |
| *Primula veris* | | | | | | | |
| Seewis | High | 54 (39, 69)% | 4 (0, 27)% | 20 (9, 28)% | 30 (16, 39)% | 4.0 (3.0, 5.2) mm$^2$ | 2.9 (2.0, 4.0)% |
| | Low | 46 (31, 61)% | 6 (0, 32)% | 13 (7, 18)% | 27 (7, 38)% | | |
| Montreux | High | 54 (41, 71)% | 0 (0, 19)% | 26 (11, 40)% | 28 (16, 40)% | 2.5 (1.7, 3.4) mm$^2$ | 1.8 (1.1, 2.5)% |
| | Low | 46 (29, 59)% | 5 (0, 27)% | 21 (9, 31)% | 20 (8, 32)% | | |
| Kollbrunn | High | 65 (44, 79)% | 6 (0, 35)% | 40 (20, 50)% | 19 (7, 32)% | 4.7 (3.1, 6.6) mm$^2$ | 3.8 (2.5, 5.8)% |
| | Low | 35 (21, 56)% | 11 (0, 36)% | 8 (3, 15)% | 16 (8, 22)% | | |
| Pfungen | High | 54 (35, 75)% | 0 (0, 20)% | 30 (6, 46)% | 23 (12, 32)% | 5.7 (3.9, 7.6) mm$^2$ | 4.3 (2.6, 6.0)% |
| | Low | 46 (25, 65)% | 35 (17, 55)% | 4 (2, 6)% | 7 (2, 12)% | | |
| Glarus | High | 23 (15, 35)% | 0 (0, 7)% | 7 (3, 11)% | 16 (9, 25)% | 5.7 (3.9, 7.3) mm$^2$ | 3.9 (2.5, 4.9)% |
| | Low | 77 (65, 85)% | 66 (55, 77)% | 7 (3, 12)% | 3 (2, 7)% | | |
| Average* | High | 48% | 2% | 24% | 22% | 4.51 mm$^2$ | 3.4% |
| | Low | 52% | 29% | 9% | 13% | | |
| *Primula vulgaris* | | | | | | | |
| Pompagles | High | 69 (52, 86)% | 0 (0, 32)% | 42 (14, 60)% | 27 (5, 38)% | 4.9 (3.3, 6.3) mm$^2$ | 2.9 (1.9, 3.9)% |
| | Low | 31 (14, 48)% | 19 (4, 40)% | 9 (2, 16)% | 2 (1, 3)% | | |
| Arogno | High | 79 (60, 89)% | 35 (6, 65)% | 27 (11, 43)% | 17 (6, 27%) | 6.0 (3.9, 8.7) mm$^2$ | 4.0 (2.6, 7.8)% |
| | Low | 21 (11, 40)% | 4 (0, 21)% | 10 (6, 14)% | 7 (2, 13)% | | |
| Vaglio | High | 81 (59, 92)% | 26 (3, 55)% | 38 (22, 50)% | 17 (7, 28)% | 6.2 (3. 7, 9.8) mm$^2$ | 3.7 (2.1, 6.4)% |
| | Low | 19 (8, 41)% | 8 (8, 24)% | 5 (2, 9)% | 6 (2, 12)% | | |
| Collonges | High | 64 (41, 84)% | 5 (0, 34)% | 47 (26, 64)% | 12 (5, 20)% | 6.9 (4.6, 9.5) mm$^2$ | 4.5 (3.0, 7.0)% |
| | Low | 36 (16, 59)% | 11 (1, 33)% | 17 (5.9, 26)% | 8 (3, 13)% | | |
| Lausanne | High | 81 (68, 91)% | 15 (0, 51)% | 44 (21, 59)% | 22 (6, 37)% | 8.0 (5.1, 11.1) mm$^2$ | 4.8 (3.0, 8.2)% |
| | Low | 19 (9, 31)% | 0 (0, 8)% | 11 (3, 18)% | 8 (3, 13)% | | |
| Average* | High | 75% | 16% | 40% | 19% | 6.4 mm$^2$ | 4.0% |
| | Low | 25% | 8% | 11% | 6% | | |

*These are the percentages of the averages, as measured in mm$^2$ (not the average of the percentages); average total inaccuracy is in units of mm$^2$ or in percentages of trait means.

The inaccuracies of the high and low organ types are presented as a percentage of total inaccuracy, so that they sum to 100%. The inaccuracies of the high and low organ types are further decomposed into maladaptive bias$^2$ (the square of the departure of the trait mean from the optimum), variance (= imprecision) of the anthers and variance (= imprecision) of the stigmas, and these three components sum to the inaccuracy of each respective organ type. The six components for each population sum, in turn, to 100%. Total inaccuracy for each population is given as the absolute value (in units of mm$^2$) in column 7 and in percentage of the mean$^2$ in column 8.

optimum, and a load of 9% would require a shift of 30% (because $0.02 \approx 0.14^2$ and $0.09 \approx 0.30^2$). Whether this relatively strong selection is reasonable for the system is hard to assess in view of the lack of good quantitative data on selection on reciprocity in heterostylous flowers and, indeed, on stabilizing selection in general (Stinchcombe *et al.*, 2008; Morrissey, 2015). If the stabilizing selection were an order of magnitude less ($s_\mu = 0.1$), the loads from our observed inaccuracies would range from 0.2% to 0.9%. This may still be strong enough to keep the trait reasonably accurate if this is variationally possible. Hence, it is at least possible to hypothesize that *P. elatior*, with an average inaccuracy of 6%, has experienced weaker or more variable net selection in the past than the other species, which average 3–4% inaccuracies.

Examination of the contribution of high vs low organs to total inaccuracy reveals striking differences among species and populations. For example, total inaccuracy and imprecision in *P. veris* were affected by high and low organs to similar extents. By contrast, in *P. elatior* and *P. vulgaris*, most of the inaccuracy and imprecision was generated by the high organs alone (Table 2; Fig. 3). Interestingly, the high sexual organs of *P. elatior* and *P. vulgaris* contribute more strongly than the low sexual organs to limiting pollen transfer between the two species (Keller *et al.*, 2016). These differences between species are captured by our measure of reciprocal inaccuracy, but would not be obvious from other reciprocity indices (Table 3), either because they mix the properties of short and tall organs (Eckert & Barrett, 1994; Sánchez *et al.*, 2013) or because the calculations fail to reveal this property of the data (Richards & Koptur, 1993; Table 3).

As seen in Table 4, the Sánchez index was strongly correlated with the mean-scaled inaccuracy across these populations and species. This is driven by the fact that the factor $r_y$ of the Sánchez index in Eqn 5 equals the expected square root of the individual-level inaccuracy on the corresponding level. In addition, when there is little bias, traits are normally distributed, and trait variances are similar across levels (as in most of our populations when mean scaled); then, the Sánchez $r$ in Eqn 6 becomes proportional to the square root of the imprecision. Consequently, $R = r \times \mathrm{SD}_r$ is approximately proportional to inaccuracy under these conditions. However, such a strong relationship is not a general expectation. It should also be noted that only inaccuracy provides a numerical connection to a model of fitness and hence a means for quantitative interpretation of the data. Previous indices lack this property, and the numbers they produce, as well as the differences between populations or species provided by these indices, remain largely devoid of biological meaning.

Imprecision in floral sexual organs may often result from developmental variation, that is, within- and among-individual variation in phenotypes resulting from developmental noise generated by environmental and/or genetic factors (see discussions in Hansen *et al.*, 2006). Such developmental variation is expected to affect the imprecision of organs proportionally (see Eckert & Barrett, 1994), just as variation of biological size measurements usually scales with the mean. Consistent with this expectation, across all organs, populations and species, the unstandardized imprecision of organs scaled with the square of the means of the respective organ ($b = 0.86 \pm 0.11$; $r^2 = 50\%$; Fig. 3a). A similar relationship was also evident as a weak trend among populations within species (Fig. 3b).

The effect of developmental variation on imprecision provides a possible explanation for the different pattern observed in *P. veris*, where low organs contributed more strongly to floral imprecision (means of 27.5–37.1% of total population imprecision in *P. veris* vs 17.7–21.5% in *P. elatior* and 21.3–25.3% in *P. vulgaris*; calculated from Table 2). Inspection of Table 1

**Table 3** Comparisons of several previous reciprocity indices calculated for the *Primula* study populations

| Species | Population | Sánchez *et al.* (2013) $R_2$ | Eckert & Barrett (1994) $R$ | Richards & Koptur (1993) $R_{\mathrm{tall}}$ | Richards & Koptur (1993) $R_{\mathrm{short}}$ |
|---|---|---|---|---|---|
| *P. elatior* | Küsnacht | 0.87 | 0.38 | 0.038 | 0.001 |
| *P. elatior* | Kollbrunn | 0.90 | 0.36 | 0.040 | 0.029 |
| *P. elatior* | Zurich 1 | 0.81 | 0.35 | 0.047 | 0.029 |
| *P. elatior* | Zurich 2 | 0.75 | 0.30 | −0.029 | 0.062 |
| *P. elatior* | Thöringen | 0.77 | 0.32 | 0.004 | 0.113 |
| *P. veris* | Seewis | 0.91 | 0.23 | 0.014 | 0.026 |
| *P. veris* | Montreux | 0.94 | 0.24 | 0.003 | 0.020 |
| *P. veris* | Kollbrunn | 0.88 | 0.23 | 0.019 | 0.041 |
| *P. veris* | Pfungen | 0.87 | 0.24 | 0.003 | 0.082 |
| *P. veris* | Glarus | 0.90 | 0.21 | 0.001 | 0.106 |
| *P. vulgaris* | Pompagles | 0.92 | 0.24 | −0.002 | 0.051 |
| *P. vulgaris* | Arogno | 0.88 | 0.32 | 0.046 | 0.028 |
| *P. vulgaris* | Vaglio | 0.89 | 0.32 | 0.037 | 0.040 |
| *P. vulgaris* | Collonges | 0.86 | 0.27 | 0.018 | 0.049 |
| *P. vulgaris* | Lausanne | 0.85 | 0.32 | 0.033 | −0.004 |

Sánchez *et al.* (2013) $R_2$ refers to the modification of the Sánchez *et al.* (2008) index $R$ as proposed in Sánchez *et al.* (2013).

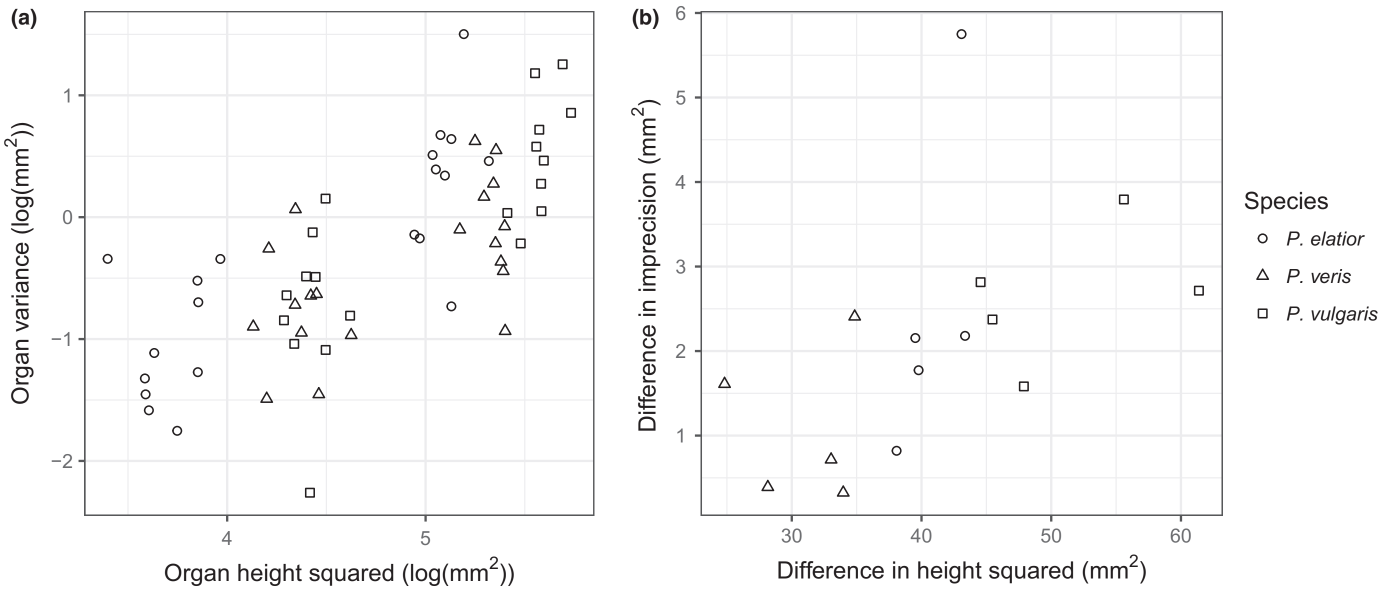

**Fig. 3** (a) The relationship between the log of squared organ height and the log of organ variance ($b = 0.86 \pm 0.11$; $r^2 = 50\%$) across the three *Primula* study species. (b) The relationship between difference in imprecision (imprecision of tall organs minus imprecision of short organs) and squared difference in mean organ height of low and high organs ($b = 0.08 \pm 0.03$; $r^2 = 28\%$).

**Table 4** Pearson correlations between scaled and unscaled inaccuracies and previous reciprocity indices for the *Primula* study populations ($n = 15$)

| | Unstandardized inaccuracy (mm$^2$) | Sánchez *et al.* (2008) *R* | Eckert & Barrett (1994) *R* | Richards & Koptur (1993) *R* (high organs) | Richards & Koptur (1993**)** *R* (low organs) |
|---|---|---|---|---|---|
| Mean$^2$ standardized inaccuracy | 0.73 | 0.99 | 0.36 | 0.54 | 0.90 |
| Unstandardized inaccuracy (mm$^2$) | | 0.71 | 0.11 | 0.57 | 0.83 |
| Sánchez *et al*. (2008) *R* | | | 0.38 | – | – |

Correlations with Sánchez *et al.* (2008) R are presented here; correlations with Sánchez *et al.* (2013) $R_2$ are identical, but with opposite sign. Richards & Koptur (1993) reciprocities were converted from signed values to absolute values. They could be correlated only with the inaccuracy measures because only the latter provide measurements for high and low organs separately, as does the Richards & Koptur (1993) index.

reveals that the difference between high and low organ heights in *P. veris* is smaller than in the other two species. Taken together, these observations suggest that the part of the inaccuracy resulting from variation in floral organ height reflects developmental imprecision of rather similar magnitude in the different populations and species. We can further speculate that greater precision is either not developmentally possible or selection for it is not strong enough to overcome genetically correlated costs. Indeed, greater realized imprecision caused by pollinator movement and variation in pollinator orientation could weaken selection for floral precision (see Armbruster, 2014; Keller *et al.*, 2014).

## General discussion and conclusions

The most salient criticism made by Sánchez *et al.* (2008) of earlier reciprocity indices was that those indices failed to incorporate the within-population variation into a single reciprocity measure. This parallels criticisms by Orzack & Sober (1994a,b) and Hansen *et al.* (2006) of optimality studies, most of which fail to include within-population variation as a component of maladaptation. Indeed, the total departure from reciprocity in a population is clearly affected by variation in the population, as well as by deviation of the mean from the optimum. Sánchez *et al.* (2008) dealt with this problem by incorporating variation into their reciprocity metric. Although the reciprocity indices of Sánchez *et al.* yield results that correlate surprisingly closely with our inaccuracy metric across the populations in the *Primula* dataset (Table 4), we cannot recommend the former approach because of its lack of connection to theory and its use of *ad hoc* arithmetic manipulations. The high correlation in our example is case specific and not general. There will be cases where the two diverge and where the index of Sánchez *et al.* gives counterintuitive results. For example, if a trait has near-zero imprecision, the Sánchez index will indicate perfect reciprocity even when there is substantial maladaptive bias. The inaccuracy index, by contrast, will correctly capture the non-zero fitness load in these cases.

In addition to establishing a meaningful scale in terms of pollination probability or fitness load, adaptive inaccuracy also has the advantage of distinguishing the relative contribution of 'maladaptive bias' (departure of the population mean from the optimum, which corresponds, in this case, to departure from perfect reciprocity) and 'imprecision' (variation around the population mean) to the overall phenotypic load. Although we are not the first to recognize that both bias and imprecision contribute to inaccuracy in heterostylous pollen transfer (e.g. Eckert & Barrett, 1994; Sánchez *et al.*, 2008, 2013), the measures we propose are the first to express these contributions on a common scale, thereby allowing direct comparison of the respective contributions of these two components to the decrease in fitness.

The estimation and comparison of the relative importance of the bias and imprecision components of inaccuracy, as we have shown here, provide valuable insights into how adaptive improvements in accuracy are likely to occur. The opportunity for evolution of the mean is greater if maladaptive bias is the major contributor to adaptive inaccuracy ('selection on the mean'). By contrast, increased precision (e.g. through canalization) will be the only possible evolutionary response if maladaptive bias is not an important contributor to adaptive inaccuracy.

Adaptive accuracy is also flexible in that it allows generalization to any form of optimizing selection (Pélabon *et al.*, 2012). There are, indeed, two possible ways to relate reciprocity to fitness. When no specific information about the fitness function is available, we can use the measure based on a quadratic fitness function to set a scale. In this case, the absolute values of the inaccuracy index can only be interpreted counterfactually, but the relative contributions of bias, precision and target variance can be interpreted as relative effects on the fitness load under quadratic selection. Similarly, the relative values of traits or populations can be interpreted as their relative loads if they are subject to the same levels of weak (hence quadratic) stabilizing selection. When an empirical fitness function is available, this can be used to give exact interpretations of the inaccuracy values as fitness loads, as explained above and in Pélabon *et al.* (2012). This is the closest one can get to understanding the actual selection for reciprocity.

The advantage of using a flexible fitness model for the assessment of the adaptive significance of reciprocity is well illustrated by the case of *Linum suffruticosum* (Linaceae), a heterostylous perennial of the western Mediterranean. In this system, pollen placement and retrieval operate in three dimensions. Reciprocity occurs on a plane rather than on a line as normally modelled (Armbruster *et al.*, 2006). As a result, standard measures of reciprocity would lead one to expect inefficient inter-morph transfer of pollen (e.g. *A* and *S* differ greatly), when, in fact, this arrangement appears to work well in generating inter-morph (disassortative) pollen flow (Armbruster *et al.*, 2006; see also discussion in Eckert & Barrett, 1994). This efficiency can be captured by an adaptive-accuracy measure relating directly to the mechanics of pollinator contact with fertile parts (Armbruster *et al.*, 2009). An important next step will be to use phenotypic selection analysis to test the fitness consequences of the departure of individual flowers from accuracy, in terms of both arrival of compatible pollen and seed set.

Here, we have illustrated the utility of adaptive-accuracy metrics by examining likelihoods of compatible pollinations as revealed by reciprocity of heterostylous morphs; however, this approach has much broader application. It is a useful framework of analysis whenever variation in morphological, physiological or behavioural traits (see, for example, Dvorak & Gvozdik, 2010) is thought to influence biological function and, ultimately, reproductive fitness. For example, expected pollen-flow rates between compatible morphs of tristylous plants (Darwin, 1877), enantiostylous plants (Barrett, 2002; Vallejo-Marin *et al.*, 2013), flexistylous and heterodichogamous plants (Li *et al.*, 2001a,b; Renner, 2001), and inversostylous plants (Pauw, 2005), and between staminate and pistillate flowers in plants with unisexual flowers (e.g. Armbruster *et al.*, 2009), can be modelled in the fashion described above for heterostylous plants. Flower-part movements also make adaptive sense in light of precision and accuracy (Li *et al.*, 2001a; Armbruster *et al.*, 2004, 2014). In addition, the adaptive nature of floral polymorphisms, such as stigma height dimorphisms, and heterostylous flowers that are too widely open to work in a linear fashion as classically described (Darwin, 1877; Barrett, 2002) can be interpreted using adaptive accuracy. All that is required for the adaptive-accuracy model is a floral landmark that constrains the position of the pollinator (e.g. nectary or corolla throat) and measurements that capture where pollen is likely to be placed on the pollinators and where stigmas are likely to contact the pollinators when they are collecting the reward.

There are also general lessons to be learned from the botanical story recounted here, with applications to all areas of biology. We biologists have been largely ignorant of measurement-theoretical problems, and have been lax in demanding mathematical and biological justification when developing numerical indices to capture ecological and functional properties of organisms. Regardless of the utility of measuring reciprocity as an accuracy, the future development and evaluation of measures of reciprocity should adhere to the principles and procedures described herein to ensure an appropriate quantitative connection between numbers and biology.

## Acknowledgements

Research by W.S.A. was supported by the Royal Society (UK), and by C.P. was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme, project no. 223257. We thank Rocío Pérez-Barrales and three anonymous reviewers for discussion and/or comments on the manuscript.

## Author contributions

W.S.A. developed the initial idea. W.S.A., G.H.B., T.F.H. and C.P. refined the idea and further developed the method. B.K. and E.C. provided example data. G.H.B. analysed the data. W.S.A., T.F.H. and C.P. wrote the first draft of the manuscript, and all authors contributed to further manuscript revision.

## References

**Armbruster WS. 2014.** Floral specialization and angiosperm diversity: phenotypic divergence, fitness trade-offs and realized pollination accuracy. *AoB Plants* **6**: plu003.

**Armbruster WS, Hansen TF, Bolstad GH, Pélabon C. 2014.** Integrated phenotypes: understanding trait covariation in plants and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**: 20130245.

**Armbruster WS, Hansen TF, Pélabon C, Pérez-Barrales R, Maad J. 2009.** The adaptive accuracy of flowers: measurement and microevolutionary patterns. *Annals of Botany* **103**: 1529–1545.

**Armbruster WS, Pélabon C, Hansen TF, Mulder CPH. 2004.** Floral integration and modularity: distinguishing complex adaptations from genetic constraints. In: Pigliucci M, Preston KA, eds. *The evolutionary biology of complex phenotypes.* Oxford, UK: Oxford University Press, 23–49.

**Armbruster WS, Pérez-Barrales R, Arroyo J, Edwards ME, Vargas P. 2006.** Three-dimensional reciprocity of floral morphs in wild flax (*Linum suffruticosum*): a new twist on heterostyly. *New Phytologist* **171**: 581–590.

**Barrett SCH. 1990.** The evolution and adaptive significance of heterostyly. *Trends in Ecology and Evolution* **5**: 144–148.

**Barrett SCH. 1992.** *Evolution and function of heterostyly.* Berlin, Germany: Springer-Verlag.

**Barrett SCH. 2002.** The evolution of plant sexual diversity. *Nature Reviews Genetics* **3**: 274–284.

**Barrett SCH, Jesson LK, Baker AM. 2000.** The evolution and function of stylar polymorphisms in flowering plants. *Annals of Botany* **85**: 253–265.

**Brys R, Jacquemyn H. 2015.** Disruption of the distylous syndrome in *Primula veris. Annals of Botany* **115**: 27–39.

**Charlesworth D, Charlesworth B. 1979.** A model for the evolution of distyly. *American Naturalist* **114**: 467–498.

**Chevin LM. 2011.** On measuring selection in experimental evolution. *Biology Letters* **7**: 210–213.

**Darwin C. 1862.** On the two forms, or dimorphic condition in the species of *Primula* and on their remarkable sexual relations. *Proceedings of the Linnean Society (Botany)* **6**: 77–96.

**Darwin C. 1864.** On the existence of two forms, and on their reciprocal sexual relation, in several species of the genus *Linum. Proceedings of the Linnean Society (Botany)* **7**: 69–83.

**Darwin C. 1877.** *The different forms of flowers on plants of the same species.* London, UK: Murray.

**Dvorak J, Gvozdik L. 2010.** Adaptive accuracy of temperature oviposition preferences in newts. *Evolutionary Ecology* **24**: 1115–1127.

**Eckert CG, Barrett SCH. 1994.** Tristyly, self-compatibility and floral variation in *Decodon verticillatus* (Lythraceae). *Biological Journal of the Linnean Society* **53**: 1–30.

**Faivre AE, McDade LA. 2001.** Population-level variation in the expression of heterostyly in three species of Rubiaceae: does reciprocal placement of anthers and stigmas characterize heterostyly? *American Journal of Botany* **88**: 841–853.

**Frank SA. 2009.** The common patterns of nature. *Journal of Evolutionary Biology* **22**: 1563–1585.

**Frank SA. 2014.** Generative models versus underlying symmetries to explain biological pattern. *Journal of Evolutionary Biology* **27**: 1172–1178.

**Ganders FR. 1979.** The biology of heterostyly. *New Zealand Journal of Botany* **17**: 607–635.

**Haller BC, de Vos JM, Keller B, Hendry AP, Conti E. 2014.** A tale of two morphs: modeling plant–pollinator interactions, reproductive isolation, and local adaptation in parapatry. *PLoS ONE* **9**: e106512.

**Hand DJ. 2004.** *Measurement theory and practice: the world through quantification.* London, UK: Arnold.

**Hansen TF. 2015.** Measuring gene interactions. In: Moore JH, Williams SM, eds. *Epistasis: methods and protocols.* New York, NY, USA: Humana Press, 115–143.

**Hansen TF, Carter AJR, Pélabon C. 2006.** On adaptive accuracy and precision in natural populations. *American Naturalist* **168**: 168–181.

**Hansen TF, Houle D. 2008.** Measuring and comparing evolvability and constraint in multivariate characters. *Journal of Evolutionary Biology* **21**: 1201–1219.

**Hansen TF, Pélabon C, Houle D. 2011.** Heritability is not evolvability. *Evolutionary Biology* **38**: 258–277.

**Houle D, Pelabon C, Wagner GP, Hansen TF. 2011.** Measurement and meaning in biology. *Quarterly Review of Biology* **86**: 3–34.

**Keller B, deVos JM, Conti E. 2012.** Decrease of sexual organ reciprocity between heterostylous primrose species, with possible functional and evolutionary implications. *Annals of Botany* **110**: 1233–1244.

**Keller B, deVos JM, Schmidt-Lebuhn A, Thomson JD, Conti E. 2016.** Both morph- and species-dependent asymmetries affect reproductive barriers between heterostylous primroses. *Ecology and Evolution* **6**: 6223–6244.

**Keller B, Thomson JD, Conti E. 2014.** Heterostyly promotes disassortative pollination and reduces sexual interference in Darwin's primroses: evidence from experimental studies. *Functional Ecology* **28**: 1413–1425.

**Krantz DH, Luce RD, Suppes P, Tversk A. 1971.** *Foundations of measurement, volume I: additive and polynomial representations.* New York, NY, USA: Academic Press.

**Lau P, Bosque C. 2003.** Pollen flow in the distylous *Palicourea fendleri* (Rubiaceae): an experimental test of the Disassortative Pollen Flow Hypothesis. *Oecologia* **135**: 593–600.

**Li QJ, Xu ZF, Kress WJ, Xia YM, Zhang L, Deng XB, Gao JY, Bai ZL. 2001a.** Flexible style that encourages outcrossing. *Nature* **410**: 432.

**Li Q-J, Xu ZF, Xia YM, Zhang L, Deng XB, Gao JY. 2001b.** Study on the flexistyly pollination mechanism in *Alpinia* plants (Zingiberaceae). *Acta Botanica Sinica* **43**: 364–369.

**Lloyd DG, Webb CJ. 1992a.** The evolution of heterostyly. In: Barrett SCH, ed. *Evolution and function of heterostyly.* Berlin, Germany: Springer Verlag, 151–178.

**Lloyd DG, Webb CJ. 1992b.** The selection of heterostyly. In: Barrett SCH, ed. *Evolution and function of heterostyly.* Berlin, Germany: Springer Verlag, 179–208.

**Mitteroecker P, Huttegger SM. 2009.** The concept of morphospaces in evolutionary and developmental biology: mathematics and metaphors. *Biological Theory* **4**: 54–67.

**Morrissey MB. 2015.** Evolutionary quantitative genetics of non-linear developmental systems. *Evolution* **69**: 2050–2066.

**Morrissey MB. 2016.** Meta-analysis of magnitudes, differences, and variation in evolutionary parameters. *Journal of Evolutionary Biology* **29**: 1882–1904.

**Naiki A. 2012.** Heterostyly and the possibility of its breakdown by polyploidization. *Plant Species Biology* **27**: 3–29.

**Opedal ØH, Listemann J, Albertsen E, Armbruster WS, Pélabon C. 2016.** Multiple effects of drought on pollination and mating-system traits in *Dalechampia scandens. International Journal of Plant Sciences* **177**: 682–693.

**Orzack SH, Sober E. 1994a.** Optimality models and the test of adaptationism. *American Naturalist* **143**: 361–380.

**Orzack SH, Sober E. 1994b.** How (not) to test an optimality model. *Trends in Ecology and Evolution* **9**: 265–267.

**Pauw A. 2005.** Inversostyly: a new stylar polymorphism in an oil-secreting plant, *Hemimeris racemosa* (Scrophulariaceae). *American Journal of Botany* **92**: 1878–1886.

**Pélabon C, Armbruster WS, Hansen TF, Bolstad GH, Pérez-Barrales R. 2012.** Adaptive accuracy and the adaptive landscape. In: Svensson E, Calsbeek R, eds. *The adaptive landscape in evolutionary biology.* Oxford, UK: Oxford University Press, 150–168.

**Pélabon C, Hansen TF. 2008.** On the adaptive accuracy of directional asymmetry in insect wing size. *Evolution* **62**: 2855–2867.

**Renner SS. 2001.** How common is heterodichogamy? *Trends in Ecology & Evolution* **16**: 595–597.

**Richards JH, Koptur S. 1993.** Floral variation and distyly in *Guetarda scabra* (Rubiaceae). *American Journal of Botany* **80**: 31–40.

**Sánchez JM, Ferrero V, Navarro L. 2008.** A new approach to the quantification of degree of reciprocity in distylous (*sensu lato*) plant populations. *Annals of Botany* **102**: 463–472.

**Sánchez JM, Ferrero V, Navarro L. 2013.** Quantifying reciprocity in distylous and tristylous plant populations. *Plant Biology* **15**: 616–620.

**Schneider DC. 2009.** *Quantitative ecology: measurement, models, and scaling, 2$^{nd}$ edn.* London, UK: Academic Press.

**Simón-Porcar VI, Meagher TR, Arroyo J. 2015.** Disassortative mating prevails in style-dimorphic *Narcissus papyraceus* despite low reciprocity and compatibility of morphs. *Evolution* **69**: 2276–2288.

**Stevens SS. 1968.** Measurement, statistics, and the schemapiric view. *Science* **161**: 849–856.

**Stinchcombe JR, Agrawal AF, Hohenlohe PA, Arnold SJ, Blows MW. 2008.** Estimating nonlinear selection gradients using quadratic regression coefficients: double or nothing? *Evolution* **62**: 2435–2440.

**Tarka M, Bolstad GH, Wacker S, Räsänen K, Hansen TF, Pélabon C. 2015.** Did natural selection make the Dutch taller? A cautionary note on the importance of quantification in understanding evolution. *Evolution* **69**: 3204–3206.

**Vallejo-Marin M, Solis-Montero L, Vilaros DS, Lee MYQ. 2013.** Mating system in Mexican populations of the annual herb *Solanum rostratum* Dunal (Solanaceae). *Plant Biology* **15**: 948–954.

**Vuilleumier BS. 1967.** The origin and evolutionary development of heterostyly in the angiosperms. *Evolution* **21**: 210–226.

**Wagner GP. 2010.** The measurement theory of fitness. *Evolution* **64**: 1358–1376.

**Webb CJ, Lloyd DG. 1986.** The avoidance of interference between the presentation of pollen and stigmas in angiosperms. 2. Hercogamy. *New Zealand Journal of Botany* **24**: 163–178.

**Wolman AG. 2006.** Measurement and meaningfulness in conservation science. *Conservation Biology* **20**: 1626–1634.

**Zhou W, Barrett SCH, Wang H, Li D-Z. 2015.** Reciprocal herkogamy promotes disassortative mating in a distylous species with intramorph compatibility. *New Phytologist* **206**: 1503–1512.