

David L. Roberts^{1, 2, 3*} &
Charles R. Marshall^{1, 2, 4}

¹Department of Organismic and
Evolutionary Biology, Harvard
University, 26 Oxford Street,
Cambridge, MA 02138, USA

²Museum of Comparative
Zoology, Harvard University, 26
Oxford Street, Cambridge, MA
02138, USA

³Royal Botanic Gardens, Kew,
Richmond, Surrey, TW9 3AB, UK

⁴Department of Earth and
Planetary Sciences, Harvard
University, 20 Oxford Street,
Cambridge, MA 02138, USA

submitted July 2008

accepted May 2009

Perspective

Are higher taxa described earlier or later than expected by chance?

Abstract The record of species descriptions within a taxonomic group represents the product of a sampling process. How useful such a record is in inferences about biodiversity and evolutionary patterns can depend on the nature of this sampling process. Here we describe a test for two potential biases: the novelty bias, the preferential description of species from higher taxa with relatively few previously described species; and the familiarity bias, the preferential description of species from already described higher taxa. At the heart of the test is the determination of whether the description of the higher taxa proceeded at a rate faster (the novelty bias) or slower (the familiarity bias) than expected by chance given the total number of species described for each higher taxon. Ambiguity may arise if there is uncertainty in the exact order in which new species and higher taxa were described. We apply the test to description records for eight groups of orchids. A novelty bias is detected in two to three cases and familiarity bias may be present in one case. The results are discussed in relation to perception of morphological complexity and the potential for human vision-based bias in biodiversity assessments.

Key words biodiversity, date of description, morphology, Orchidaceae, taxonomic sampling, rarefaction

Introduction

The total number of species that have been formally described and are considered to be both biologically and nomenclaturally valid is generally estimated at between 1.5 and 1.8 million (Cracraft, 2002). Although these descriptions provide invaluable information, they remain only a sample of the Earth's biological diversity. How this information is used to address broader issues depends on the way in which this sample was collected (Funk & Richardson, 2002). For example, there is interest in understanding the way in which species numbers are distributed across genera and how this may shed light on macroevolution (e.g. Burlando, 1990, 1993; Solé & Bascompte, 1996). The assessment of the goodness of fit of specific models of this distribution assumes that species are sampled at random with respect to genus (Solow *et al.*, 2003). If this assumption is violated, then the result of such an assessment may be invalid. As a second example of the use of description records, the total number of species in a group is sometimes estimated by extrapolating the way in which species descriptions have accumulated with time or effort (e.g. Dolphin & Quicke, 2001). The historical pattern of species accumulation –

and therefore its extrapolation into the future – can depend on the sampling process.

There is growing interest in factors favouring the description of some species over others (see Collen *et al.*, 2004 for a review). Much of the work in this area has focused on geographical or behavioural traits. Few studies have, however, examined the role of morphological traits; the only morphological trait having been considered thus far being body size (e.g., Gaston & Blackburn, 1994; Reed & Boback, 2002; Cabrero-Sañudo & Lobo, 2003). This lack of interest in other morphological traits is curious, since most macro-organisms are described based on a morphological (Cronquist, 1978) rather than a biological species concept (Mayr, 2000).

Here we focus on two particular types of non-randomness that may occur in species' description: the preferential description of species from groups with few previously described species, the novelty bias; or, the preferential description of species from already described higher taxa, the familiarity bias (in the discussion below we restrict ourselves to genera). The existence of these biases has practical implications. For example, in the presence of a novelty bias, the observed distribution of relative genus size will tend to be biased toward uniformity. To see this, consider the case in which no new species descriptions are made in previously described genera. In this extreme case, the

*Corresponding author. Email: d.roberts@kew.org

Genus	Location	No. species	No. sections	Novelty				Familiarity			
				Conservative		Liberal		Conservative		Liberal	
				D_{\max}	$P =$	D_{\max}	$P =$	D_{\max}	$P =$	D_{\max}	$P =$
<i>Caladenia</i> ¹	Global	245	6	1.3	0.218	2.3	0.004	0.0	0.571	0.0	0.571
<i>Coelogyne</i> ²	Global	190	22	3.3	0.217	3.6	0.127	-0.9	0.278	-1.9	0.268
<i>Dendrobium</i> ³	Java	52	17	0.8	0.485	1.9	0.306	-2.2	0.231	-2.2	0.231
<i>Dendrobium</i> ⁴	Sarawak	114	17	4.4	0.003	4.4	0.003	0.0	0.474	0.0	0.474
<i>Dendrobium</i> ⁵	Sumatra	100	17	3.6	0.034	3.7	0.031	-0.8	0.521	-0.8	0.521
<i>Dendrochilum</i> ⁶	Borneo	81	7	1.7	0.143	1.9	0.119	-1.5	0.259	-1.5	0.259
<i>Disa</i> ⁷	Southern Africa	131	14	0.8	0.469	1.4	0.440	-2.5	0.113	-3.1	0.037
<i>Paphiopedilum</i> ⁸	Global	69	7	0.3	0.476	0.3	0.476	-1.8	0.108	-1.8	0.108

Table 1 Statistical tests for novelty and familiarity effects in the description of sections of orchids from eight datasets. See text for description of conservative and liberal tests.

References: ¹Hopper and Brown (2004); ²Clayton (2002); ³Comber (1990); ⁴Beaman *et al.* (2001); ⁵Comber (2001); ⁶Wood (2002); ⁷Linder and Kurzweil (1999); ⁸Cribb (1998).

observed size distribution of genera will be exactly uniform. Also, in the presence of a novelty bias, descriptions of genera will tend to accumulate at a more rapid rate than under random sampling, leading to over-extrapolation of genus number. Conversely, under familiarity bias, the observed distribution of relative genus size will tend to be more left-skewed than expected, and descriptions of genera will tend to accumulate less rapidly than under random sampling, leading to under-extrapolation of genus number. For these and other reasons, it may be important to know whether a novelty or familiarity bias is present. Here we describe and illustrate a simple test for these biases.

A test for a novelty and familiarity bias

Consider a set of n species descriptions. Let g_1, g_2, \dots, g_n be the corresponding genera (or other higher taxonomic level) in the temporal order in which they were described and let $g(m)$ be the number of genera represented in the first m of these species descriptions with $m \leq n$. The sequence $g(1), g(2), \dots, g(n)$ is non-decreasing and increases from 1 to k , where k is the number of genera represented in the full set of descriptions. We will refer to a plot of $g(j)$ against j for $j = 1, 2, \dots, k$ as a genus accumulation plot. We define the random variable $G(m)$ as the number of genera in the first m elements of a random permutation of g_1, g_2, \dots, g_n . The expected value of this random variable is given by:

$$E(G(m)) = k - \sum_{j=1}^k \binom{n-n_j}{m} / \binom{n}{m} \quad (1)$$

where n_j is the number of descriptions of species contained in genus j and $\binom{a}{b} = 0$ if $a < b$. This is equivalent to a rarefaction (see Hurlbert, 1971; Simberloff, 1972).

Interest here centres on testing the null hypothesis that the order of description is independent of genus against the alternative hypotheses that there is a tendency to describe species from genera with either few previous descriptions or many previous descriptions than expected by chance alone. Under these alternative hypotheses, the number of genera will accumulate with the number of species descriptions at a higher rate than expected under the null hypothesis (the novelty bias) or at a lower rate than expected under the null hypothesis (the familiarity bias). Let:

$$D(m) = g(m) - E(G(m)) \quad (2)$$

be the difference between the observed number of genera represented in the first m species descriptions and its expected value under randomness. In the absence of a specific model of the description process, a natural omnibus test for a novelty or familiarity bias can be based on:

$$D_{\max} = \max D(m) \quad (3)$$

where the maximum is taken over $1 \leq m \leq n$. The significance level (or P value) can be estimated by the proportion of random permutations of the sequence g_1, g_2, \dots, g_n for which the value of D_{\max} exceeds its observed value. The P values were estimated from 1000 randomisations. The hypothesis of a novelty bias finds support if D_{\max} is unusually large and positive; the hypothesis of a familiarity bias finds support if D_{\max} is unusually large and negative.

An application to the Orchidaceae

Here the method outlined above is applied to species descriptions to eight examples from the Orchidaceae, representing three of the five subfamilies (Table 1). In this analysis, sections play the role of genera.

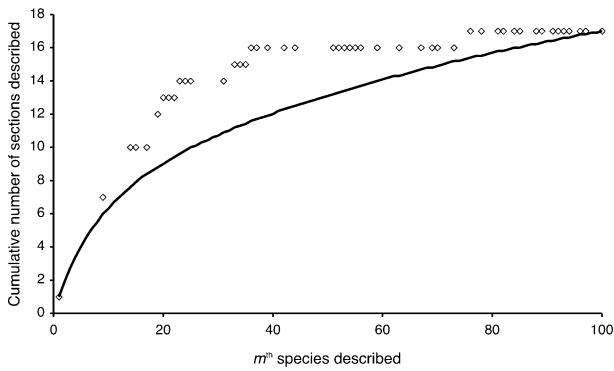


Figure 1 Section accumulation plot for the orchid genus *Dendrobium* from Sumatra (diamonds) strongly supports the novelty effect, where sections were described preferentially (see Table 1 for statistics). The solid curve shows the section accumulation curve expected under random sampling of sections.

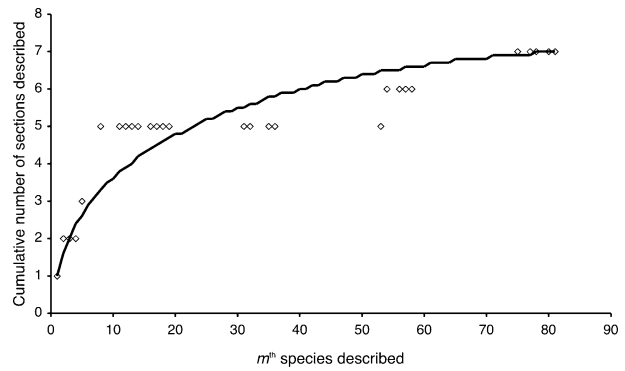


Figure 3 Section accumulation plot for the orchid genus *Dendrochilium* for Borneo (diamonds), showing no support for either the novelty or familiarity effects; sections appear to have been described randomly with respect to the species descriptions (see Table 1 for statistics). The solid curve shows the section accumulation curve expected under random sampling.

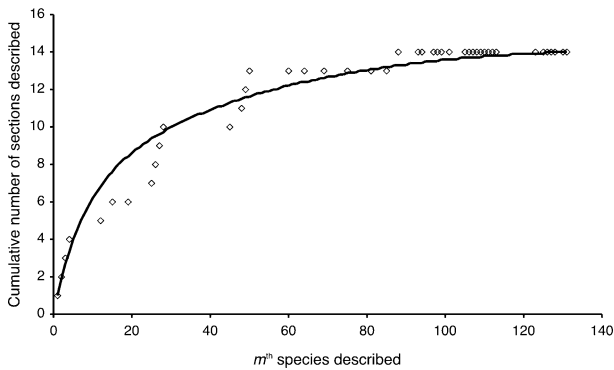


Figure 2 Section accumulation plot for the orchid genus *Disa* for southern Africa (diamonds), which may support the familiarity effect, where sections were described more slowly than expected by chance, depending on the (unknown) order in which sections were described in a single year (see Table 1 for statistics and text for discussion). The solid curve shows the section accumulation curve expected under random sampling.

Section accumulation plots are shown for three of the datasets (Figs 1–3) along with the corresponding rarefaction curves, that is, the expected rate of accumulation of sections as species are described under the null hypothesis of random species description with respect to section description. In many cases, several species and at least one section were described in a year, so there is ambiguity as to the exact shape of the section accumulation curve (i.e. the values of the section accumulation plot can sometimes only be calculated for certain values of m). We deal with this ambiguity in one of two ways. First, we assume that if in a given year x new species and y new sections were described (where $x > y$), that the first y species descriptions corresponded to the new sections. In the second approach we assume that the last y species descriptions led to the y new section descriptions. If the section accumulation curve lies above the rarefaction curve, then there is the possibility of a novelty bias (see Fig. 1), in which case the first approach is termed ‘liberal’ because it maximises the chance of rejecting

the null hypothesis (it pushes the section accumulation curve as far above the rarefaction curve as the data will allow), while the second approach is ‘conservative’ because it minimises the chance of rejecting the null hypothesis. If, on the other hand, the section accumulation occurs beneath the rarefaction curve, then the data suggest the possibility of a familiarity bias (Fig. 2), and now the first approach is conservative (minimising the probability of rejecting the null hypothesis by keeping the section accumulation curve as close to the rarefaction curve as possible), while the second approach is liberal (maximising the probability of rejecting the null hypothesis).

When the test is applied to the eight datasets, two show strong evidence of the novelty effect, while an additional dataset shows evidence of the novelty effect when the liberal approach is applied to the data. Another of the datasets shows a familiarity bias if the data are interpreted liberally (Table 1).

Discussion

The novelty bias arises from a tendency to describe preferentially species that are morphologically dissimilar from previously described species. Conversely, if a taxonomic specialist is focused on a particular taxonomic group, this might result in the familiarity bias. Among the eight data sets examined (Table 1), we found two unequivocal cases of the novelty bias (see Fig. 1 for one of these) with the possibility of a third, one possible case of familiarity bias (Fig. 2), as well as four cases where species description was random with respect to taxonomic group membership (see Fig. 3 for one of these). Note that a familiarity bias might not be detected even if a taxonomist tends to focus on one or a few higher taxa if he or she is most interested in the less speciose groups first (perhaps for reasons of tractability).

We note that even if a taxonomist does not have a novelty or familiarity bias, that even before he or she begins to describe the specimens, a bias may have occurred in the initial

collection of the samples. For example, consider a savannah in Africa studded with a few species of *Acacia* tree. The first species collected may well be a grass as this is the most abundant group. The second species to be collected one might also expect to be a grass, based on the high abundance and species richness of grasses. However, early in the sequence of discoveries a species of *Acacia* is likely to be collected, not because it is abundant, but because of its size and its morphological dissimilarity from the grasses previously collected.

The presence or absence of a novelty bias may be related to the morphological complexity of the organisms in question (Valentine, 2003). The potential biases that differences in morphological complexity may introduce in biodiversity and evolutionary studies has long been recognised in the palaeontological community. For example, Vaughan (1924), in his presidential address to the Paleontological Society stated, 'It has been pointed out by many paleontologists that the more complex the structural features of an organism are, the more restricted will be its stratigraphic range.' Arkell (1956) reiterated this point by noting 'It is therefore futile to draw up comparative tables of faunas in different formations and different parts of the world on the basis of 'numbers of species'.' Similarly, simulations have shown that for a tenfold increase in morphological complexity, there is a tenfold increase in apparent evolutionary rate (Schopf *et al.*, 1975). That is, the greater the number of available traits, the more readily species can be discriminated. Hence, when comparing the durations of taxa through geological time, there may be a bias such that morphologically complex taxa might have artificially shorter measured durations (Schopf, 1982).

In general, the greater the morphological disparity, the greater the opportunity for the novelty bias to become manifest. We may be seeing this effect in the orchid datasets analysed here. For example, of the three examples where there is evidence of the novelty effect, two are from the genus *Dendrobium*. This may not be surprising since it is not only extremely variable in floral morphology, but also vegetatively; it exhibits the most disparity of all the genera we examined. The third example, *Caladenia*, only came to light with the more liberal analysis, and has less disparity than *Dendrobium*.

Although we have found stronger evidence of the novelty bias than the familiarity bias in the species descriptions of a number of the orchid groups examined, it is clearly premature to generalise to other taxa. The fact remains, however, that inferences drawn from description records can be sensitive to the underlying sampling process and the method described here is designed to detect biases in this process. We note that this method can be modified to test for other types of non-random sampling, and we emphasise that explicit knowledge of any kind of non-randomness in species descriptions may provide useful information in the proper use of these records.

Implications for measuring biodiversity

Following the 2002 World Summit in Johannesburg, the Convention on Biological Diversity has called for a decrease in

the rate of biodiversity loss by 2010. However, a 2003 UK Royal Society report *Measuring Biodiversity for Conservation* discussed the unavailability of satisfactory measures of biodiversity (Royal Society, 2003). Currently, conservation priorities are guided more by raw estimates of diversity (i.e. species lists) rather than by any measure of how quickly that diversity is being lost (e.g. Myers *et al.*, 2000).

In the race to document the Earth's biodiversity, scientists use numerous techniques of sampling. Kunzig (2000) eloquently described the possible effects of such biased sampling on our knowledge of biodiversity.

Imagine a race of space aliens discovers Earth and, being scientifically inclined, sets out to understand the life on its surface . . . For want of a better idea they decide to blindly drag a large, sturdily framed net from their spacecraft as it cruises safely above the clouds.

The net touches down one Friday evening in Tyler, Texas, where it first clips a flag pole off the top of the courthouse. Next it bounces through a playground in Bergfield Park, scattering children and parents without snaring one; collects a dog that is studying a soup bone on College Street; and swoops into a backyard a few blocks later, picking up an azalea bush, a clothesline bearing assorted lacy underwear, and a patch of lettuce with associated rabbit. Finally it nearly comes to grief in a dark corner of the Sears parking lot, where the weight of a 1979 Chevy and the teenagers in its back seat cause the net cable to groan audibly. The aliens reel in their catch: one carnivore; one herbivore with food; assorted mysterious shell fragments; and a large metal crusted animal (the Nova) whose source food remains obscure, but which seems to be serviced by remarkably sophisticated intestinal endosymbionts.

By far the most important tool for detecting species is the human senses, in particular vision. Geographic range has repeatedly been found to be an important explanatory variable, with a number of areas in Africa, for example, being poorly collected (e.g. Guinean montane forests, north-western Congolian lowland forests and southern Albertine Rift montane forests) (Küper *et al.*, 2006). However, once within a region other factors such as morphology and colour may affect species discovery. It is already known from psychological studies that human vision is biased and follows strict patterns of colour and shape recognition, and categorisation bias (Palmer, 1999). Any attempt to use species descriptions to draw inferences about biodiversity needs to have an understanding of the process by which species are discovered. For example, the relative paucity of specimens of some taxa may be related to the relative recency of their time of discovery, rather than because they are rare. When sampling effort is consistently lower for recently identified taxa, there will be a tendency to underestimate their status such as size ranges (Solow & Roberts, 2006). This begs the question as to why taxa are discovered when they are. Do certain morphological, geographical and ecological characters (conspicuousness) increase the probability of a species being discovered or recorded? Do conservation and biodiversity prioritisation reflect a level of conspicuousness and accumulation of knowledge? From this, perhaps even more importantly is whether what we are collecting is representative of true biodiversity. This is particularly important given the time and money that is currently being spent on 'rapid biodiversity assessment'. Are rapid biodiversity assessments significantly biased by human cognitive biases? Here we offer a test for some of those potential biases.

Acknowledgements

The authors thank Andrew Solow for extensive statistical advice, especially in relation to testing the null hypothesis. We thank Jeffery Wood for helping assign species to sections for the Sarawak dataset, and three referees for their very helpful comments; Richard Bateman, Torsten Dikow and Donald Quicke. This work was conducted while DLR held the Sarah and Daniel Hrdy Fellowship in Conservation Biology at Harvard University.

References

- ARKELL, W.J. 1956. Species and species, In: SYLVESTER-BRADLEY P.G., Ed., *The Species Concept in Palaeontology*. Systematics Association, London, pp. 97–99.
- BEAMAN, T.E., WOOD, J.J., BEAMAN, R.S. & BEAMAN, J.H. 2001. *Orchids of Sarawak*. Royal Botanic Gardens, Kew.
- BURLANDO, B. 1990. The fractal dimension of taxonomic systems. *Journal of Theoretical Biology* **146**, 99–144.
- BURLANDO, B. 1993. The fractal geometry of evolution. *Journal of Theoretical Biology* **163**, 161–172.
- CABRERO-SAÑUDO, F.J. & LOBO, J.M. 2003. Estimating the number of species not yet described and their characteristics: the case of Western Palaearctic dung beetle species (Coleoptera, Scarabaeoidea). *Biodiversity Conservation* **12**, 147–166.
- CLAYTON, D. 2002. *The Genus Coelogyne: a Synopsis*. Natural History Publications (Borneo), Kota Kinabalu.
- COLLEN, B., PURVIS, A. & GITTLEMAN, J.L. 2004. Biological correlates of description date in carnivores and primates. *Global Ecology and Biogeography* **13**, 459–467.
- COMBER, J.B. 1990. *Orchids of Java*. Royal Botanic Gardens, Kew.
- COMBER, J.B. 2001. *Orchids of Sumatra*. Natural History Publications (Borneo), Kota Kinabalu.
- CRACRAFT, J. 2002. The seven great questions of systematic biology: An essential foundation for conservation and the sustainable use of biodiversity. *Annals of the Missouri Botanical Garden* **89**, 127–144.
- CRIBB, P. 1998. *The Genus Paphiopedilum*. Second Edition. Royal Botanic Gardens, Kew.
- CRONQUIST, A. 1978. Once again, what is a species? In: KNUTSON L.V., Ed., *Biosystematics in Agriculture*. Allenheld Osmun, Montclair, NJ, pp. 3–10.
- DOLPHIN, K., & QUICKE, D.L.J. 2001. Estimating the global species richness of an incompletely described taxon: an example using parasitoid wasps (Hymenoptera: Braconidae). *Biological Journal of the Linnean Society* **73**, 279–286.
- FUNK, V.A. & RICHARDSON, K.S. 2002. Systematic data in biodiversity studies: Use it or lose it. *Systematic Biology* **51**, 303–316.
- GASTON, K.J. & BLACKBURN, T.M. 1994. Are newly described bird species small-bodied? *Biodiversity Letters* **2**, 16–20.
- HOPPER, S.D. & BROWN, A.P. 2004. Robert Brown's *Caladenia* revisited, including a revision of its sister genera *Cyanicula*, *Ericksonella* and *Pheladenia* (Caladeniinae: Orchidaceae). *Australian Systematic Botany* **17**, 171–240.
- HURLBERT, S.H. 1971. The non-concept of species diversity: a critique and alternative parameters. *Ecology* **52**, 577–586.
- KUNZIG, R. 2000. *Mapping the Deep: The Extraordinary Story of Ocean Science*. Sort of Books, UK.
- KÜPER, W, SOMMER, J.H., LOVETT, J.C. & BARTHLOTT, W. 2006. Deficiency in African plant distribution data – missing pieces of the puzzle. *Botanical Journal of the Linnean Society* **150**, 355–368.
- LINDER, H.P. & KURZWELL, H. 1999. *Orchids of Southern Africa*. A.A. Balkema, Rotterdam.
- MAYR, E. 2000. A critique from the biological species concept perspective: what is a species, and what is not? In: WHEELER, Q.D. & MEIER, R., Eds., *Species Concepts and Phylogenetic Theory. A Debate*. Columbia University Press, New York.
- MYERS, N., MITTERMEIER, R.A., MITTERMEIER, C.G., DA FONSECA, G.A.B. & KENT, J. 2000. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858.
- PALMER, S.E. 1999. *Vision Science: Photons to Phenomenology*. MIT Press, Boston.
- REED, R.N. & BOBACK, S.M. 2002. Does body size predict dates of species description among North American and Australian reptiles and amphibians? *Global Ecology and Biogeography* **11**, 41–47.
- Royal Society 2003. *Measuring Biodiversity for Conservation*. Document 11/03.
- SCHOPF, T.J.M. 1982. A critical assessment of punctuated equilibria. I. Duration of taxa. *Evolution* **36**, 1144–1157.
- SCHOPF, T.J.M., RAUP, D.M., GOULD, S.J. & SIMBERLOFF, D.S. 1975. Genomic versus morphologic rates of evolution: influence of morphologic complexity. *Paleobiology* **1**, 63–70.
- SIMBERLOFF, D.S. 1972. Properties of the rarefaction diversity measurement. *American Naturalist* **106**, 414–418.
- SOLÉ, R.V. & BASCOMPTE, J. 1996. Are critical phenomena relevant to large-scale evolution? *Proceedings of the Royal Society B* **263**, 161–168.
- SOLOW, A.R., COSTELLO, C.J. & WARD, M. 2003. Testing the power law model for discrete size data. *American Naturalist* **162**, 685–689.
- SOLOW, A.R. & ROBERTS, D.L. 2006. Museum collections, species distribution, and rarefaction. *Diversity and Distribution* **12**, 423–424.
- VALENTINE, J.W. 2003. Architectures of biological complexity. *Integrative and Comparative Biology* **43**, 99–103.
- VAUGHAN, T.W. 1924. Criteria and status of correlation and classification of Tertiary deposits. *Geological Society of America Bulletin* **35**, 677–742.
- WOOD, J.J. 2002. *Dendrochilum of Borneo*. Royal Botanic Gardens, Kew.