Oikos 122: 1350–1356, 2013
doi: 10.1111/j.1600-0706.2013.00207.x

Subject Editor: Lonnie Aarssen. Accepted 16 January 2013

How consistent are trait data between sources? A quantitative assessment

Jay M. Fitzsimmons

J. M. Fitzsimmons (jayfitzsimmons@gmail.com), Canadian Wildlife Service, Environment Canada, 351 St. Joseph Blvd., 14th floor, Gatineau, QC, K1A 0H3, Canada.

The use of species' traits is increasing in ecological research. Many studies obtain trait data from a single source, implicitly assuming the accuracy of these data. I critically evaluate this assumption by measuring agreement among sources for trait data. I evaluate inter-source agreement for 22 traits (anatomical, behavioural, life-history and niche-related) among five authoritative data sources (two field guides, two atlases and one online resource) for 263 Canadian butterfly species. This represents the first quantitative comparison of trait data among field guides or atlases. Traits varied considerably in their agreement among sources. Some traits such as wingspan and overwinter stage were fairly consistent among sources, whereas other traits such as habitat breadth were remarkably inconsistent among sources. These findings call into question the reliability of research that relies on a single source for trait data. I offer several recommendations for how trait researchers can account for inter-source variation in trait data.

Species' traits, such as body size and dietary niche breadth, are related to most biological processes. While the investigation of species' traits has a long history (Statzner et al. 2001), their use as predictors of ecological responses to anthropogenic forces has increased rapidly in recent years. Species' traits have been found to be related to several features including invasiveness (van Kleunen et al. 2010), range shifts with climate change (Angert et al. 2011, Betzholtz et al. 2013), vulnerability to roads (Rytwinski and Fahrig 2012), and extinction risk (Lee and Jetz 2011).

Information on traits, or anything else, that comes from a single source may be unreliable. Analyses of consistency among raters has gained attention in several fields of ecology in which raters' judgements might differ including animal behaviour (Kaufman and Rosenthal 2009), citizen science (Fitzpatrick et al. 2009), and traditional ecological knowledge (Wong et al. 2011). As yet the consistency of trait information among sources has not been evaluated. Because many trait-based studies rely on a single source for trait data, their results are only as reliable as their data source.

I compare data for 22 traits among five authoritative sources (two field guides, two atlases and one online resource) for 263 Canadian butterfly species. Existing critical assessments of field guides and atlases are largely limited to book reviews, which often focus on layout, illustrations, and nomenclature (Coad 2012), and to one thorough review of mushroom field guides' descriptions of the edibility of *Amanita muscaria* (Rubel and Arora 2008). My study provides the first analysis of authoritative sources' agreement on species' trait information. I find variable inter-source agreement, suggesting caution should be exercised before relying on a single source's trait data.

Methods

Species and sources

I included all species, and some sub-species, listed as residents of Canada in The Butterflies of Canada atlas (Layberry et al. 1998). I excluded species described as migrant, mostly migrant, or likely migrant, and included partial migrants and potential migrants. This resulted in 263 butterfly taxa for analyses.

I chose five trait sources based on several criteria: coverage of North American butterflies, recent publication (less than 20 years old), traits included, availability, and authorship by respected authorities. Out of several sources meeting these criteria I chose two atlases (Bird et al. 1995, Layberry et al. 1998), two field guides (Brock and Kaufman 2003, Wagner 2005), and one online resource (Opler et al. 2011). My approach allows conclusions about inter-source variation but does not allow conclusions about the relative agreement between atlases and field guides, or continent-wide sources and regional sources. I classified all trait data available for all of the relevant 263 species for each source. Taxonomy and nomenclature follow Pelham (2008), which I also used to find species listed under synonyms.

Traits

I collected data on 22 traits related to anatomy (e.g. wingspan), behaviour (e.g. males' mate-finding strategies), life-history (e.g. overwintering stage), and niche (e.g. larval host plant breadth) (Table 1). Converting sources' descriptions into trait values sometimes required subjective judgements. I was the only person collecting data, so whatever biases I had were consistent across sources. For example, I chose to classify the habitat descriptor 'chaparral' under the 'dry habitat' category, whereas others might have chosen to classify it under 'shrub habitat', but I followed my categorization consistently across sources so my bias should not affect inter-source agreement analyses.

Habitat association is a family of eight traits, each a yes/no response to whether a species is associated with one of eight habitat types: disturbed, dry, open, rocky, shrub, wet, woods and woods' clearings. Habitat descriptors' classifications are listed in Supplementary material Appendix 1 Table A1. Thus I scored 'yes' for each habitat type described for each species, in each source.

Hilltopping is the tendency for butterflies to congregate on hilltops for the purposes of finding a mate (Shields 1967). Sources seldom described a species as non-hilltopping, thus the data file included only one category (yes) and many species with no mention of hilltopping. Some of these absences of information on hilltopping represent species that do not hilltop, and some represent species with inadequate information about their behaviour. Because an alternate (i.e. 'no') state is required for analysis of inter-source agreement, I classified those species without mention of hilltopping as 'non-hilltopping' if there was sufficient information provided about other aspects of adult behaviour that I could assume hilltopping would be noted if it were present. Specifically, I classified species as non-hilltopping when a source listed information on mate search habit (indicating knowledge of adults' behaviour) without mention of hilltopping.

Mate search habit is the tendency for males to search for females by perching (i.e. waiting for females then chasing them) or patrolling (i.e. scanning for females while flying) (Wiklund 2003). Occasionally a species was described as engaging in both perching and patrolling strategies, in which case I coded the species as whichever strategy was described as being more prevalent, and omitted the species if neither strategy was described as being more prevalent.

Mudpuddling describes the tendency of some adult butterflies to consume mud at puddles for nutrients. As with hilltopping, sources seldom indicated that a species does not mudpuddle, so I considered species to be non-mudpuddlers if a source included information on mate search habit (indicating knowledge of adult behaviour) but not information on mudpuddling.

Myrmecophily is the association of caterpillars with ants, such as providing nutritious secretions to ants in exchange for protection from predators, although other forms of ant association also exist (Oliver and Stein 2011, Pierce et al. 2002). Myrmecophily is confined to the family Lycaenidae. As with hilltopping and mudpuddling, sources seldom indicated that a species does not associate with ants, so I considered species to be non-myrmecophilous if they were lycaenids and the source provided information on the species' overwintering stage (indicating knowledge of the species' life history).

Overwintering stage describes the life-history stage (egg, larva, pupa or adult) in which butterflies overwinter. When more than one stage was listed for a species in a source (e.g. high-altitude species that require two years to develop and overwinter in different stages each winter) I recorded the earliest stage.

Wingspan is the length, in mm, between distal tips of forewings. Sources provided lower and upper boundaries of wingspan for species, so they are recorded as two separate continuous variables, along with a third continuous variable which is the wingspan range (upper minus lower boundary).

Flight period is the time of year when adult butterflies are flying. The start and end of the flight period are recorded as two separate ordinal variables, with numbers corresponding to months (e.g. eight represents August).

Generations per year is self-explanatory, ranging from 0.5 for species that take two years to develop to 3 representing species with three or more generations per year. Lower and upper boundaries of generations were provided in sources, and are kept as two distinct ordinal variables here.

Table 1. Analyzed traits, their sample sizes, and possible states. See Methods for trait descriptions.

Trait	Sources	Species	Records	Data type	Possible states
Habitat association (eight binary variables)	5	263	877	categorical	for each of eight habitat types: no, yes
Hilltopping	4	223	348	categorical	no, yes
Mate search habit	4	209	286	categorical	percher, patroler
Mudpuddling	3	112	141	categorical	no, yes
Myrmecophily	4	52	100	categorical	no, yes
Overwintering stage	5	221	408	categorical	egg, larva, pupa, adult
Wingspan (minimum)	3	263	539	continuous	N/A
Wingspan (maximum)	3	263	539	continuous	N/A
Wingspan (range)	3	263	539	continuous	N/A
Flight period (starting month)	3	263	651	ordinal	3–8
Flight period (ending month)	3	263	651	ordinal	3–12
Generations per year (minimum)	4	257	553	ordinal	0.5–3
Generations per year (maximum)	4	256	527	ordinal	1–3
Habitat breadth	5	263	877	ordinal	1–7
Larval host plant breadth	5	255	932	ordinal	1–5

Habitat breadth is the scaled sum of the number of habitat types classified as suitable for a species. Prior to scaling, habitat breadth could range from one to eight as an ordinal variable (but seven was the highest breadth for any species in my data set), with higher numbers representing greater breadth (habitat generalists). Because some sources tended to provide higher breadth scores than other sources for all species (i.e. some sources were more liberal than others), I scaled habitat breadth by subtracting the source's average breadth score from each species' breadth value. This allowed me to determine whether the relative scoring of breadth was consistent among sources after accounting for sources' tendencies to be conservative or liberal in habitat listings.

Larval host-plant breadth represents the phylogenetic diversity of host plants listed in a source. This trait was coded as an ordinal variable from one to five with higher numbers representing greater breadth as follows: one host-plant species 1); more than one host-plant species within one genus 2); more than one host-plant genus within one family 3); more than one host-plant family within one order 4); host-plants in multiple orders 5). Plant species' taxonomic associations were obtained from Wikipedia, which has been found to be highly accurate in its taxonomic information (Page 2010). Because some sources' average breadth scores were higher than others (i.e. some sources were more liberal than others), I scaled larval host-plant breadth by subtracting the source's average breadth score from each species' breadth value. Other researchers employ measures of host-plant breadth similar to the one I have used, such as Komonen et al. (2004) and Burke et al. (2011) who classified butterflies within the first three categories (i.e. ignoring variation above family level). It is logical to classify dietary breadth with taxonomic hierarchies since host-plant breadth is often limited by foliage chemistry (Ehrlich and Raven 1964, Janz and Nylin 1998), and closely-related plants often have similar chemistry (Ricklefs 2008). For this reason researchers sometimes calculate host-plant breadth as mean phylogenetic distance among host-plants (Pellissier et al. 2012). Host-plant breadth is likely confounded with sampling effort, so species with numerous host plant records (e.g. common species) are more likely to be found on non-preferred plant species leading to broader niche classification whereas butterflies with few host-plant records are more likely to only be found on their preferred plant species and be classified as specialists (Beck et al. 2006). This bias should not affect inter-source agreement analyses since the bias is consistent across sources. A review of host-plant records for butterflies of eastern North America is provided by Tallamy and Shropshire (2009).

Analyses

I analyzed inter-source agreement for each trait, using different methods for continuous, categorical and ordinal traits. All methods of analysis allowed more than two raters and missing data (i.e. unbalanced designs). All analyses were conducted with AgreeStat ver. 2011-1 and 2011-3 in MS Excel 2007.

Continuous traits were analyzed with the intraclass correlation coefficient. Specifically, I used two-way random ANOVA models without interactions. The intraclass correlation coefficient is a popular method to calculate repeatability in behavioural and evolutionary biology, and is based on values found within the *F* table of an ANOVA (Nakagawa and Schielzeth 2010) so it is more intuitive than more complicated methods.

Categorical and ordinal traits were analyzed with six methods that differ mathematically but have the shared purpose of calculating inter-source agreement: percent agreement, Gwet's AC1, the Brennan–Prediger method, Conger's kappa (K), Fleiss' K, and Krippendorff's alpha (α). The mathematical differences among methods are beyond the scope of this paper (Banerjee et al. 1999, Berry et al. 2008, Gwet 2008, 2010, Warrens 2010); of greater relevance to ecologists are the functional differences among methods. Thus I analyzed each categorical and ordinal trait with each method and present all results, including confidence intervals, in Supplementary material Appendix 1 Table A2.

Percent agreement among sources is the most intuitive measure of inter-source agreement, but its failure to correct for agreement expected from chance can result in overestimation of agreement. When one response is much more common than other responses (e.g. non-association with ants was more common than association with ants in my data set), random distribution of ratings will result in high percent agreement (Kaufman and Rosenthal 2009). I include percent agreement for comparison purposes, but rely on results from the five chance-corrected measures of agreement.

All six methods of analysis for non-continuous data permit unweighted analysis for categorical traits and weighted analysis for ordinal traits. Thus any discrepancy between sources is equally-weighted for categorical traits (e.g. egg vs pupa and egg vs adult both count as non-agreement) but discrepancies are weighted by the extent of difference for ordinal traits (e.g. August vs August is full agreement, August vs September is partial agreement, and August vs April is no agreement). Weight values for each ordinal trait's analysis are provided in Supplementary material Appendix 2.

Results

Agreement among sources varied dramatically between traits (Table 2). Maximum and minimum wingspan were quite consistent among sources ($R_A = 0.87$, 0.88), whereas the third continuous trait (wingspan range) was less consistent ($R_A = 0.56$). The relative lack of consistency for wingspan range reflects certain sources consistently reporting broader ranges than other sources. Comparing the average wingspan range for the 40 species reported in all three sources of wingspan data (the remaining two sources did not provide wingspan values) reveals Opler et al. (2011) to be the most liberal (12.6 mm), Bird et al. (1995) to be the most conservative (8.2 mm), and Layberry et al. (1998) to be intermediate (10.4 mm). The categorical traits ranged from Fleiss' K = 0.20 (woods' clearing habitat type) to K = 0.93 (overwintering stage). Ordinal traits ranged from Fleiss' K = 0.07 (habitat breadth) to K = 0.62 (maximum generations per year). Results of all analysis methods for all traits are available in Supplementary material Appendix 1 Table A2.

Table 2. Among-source agreement for traits. Each trait was ranked relative to others of its data type (continuous, ordinal, or categorical) within each measure of agreement. Mean ± standard deviation (SD) ranks across the five chance-corrected agreement measures are presented for ordinal and categorical traits (a sixth measure, percent agreement, was excluded because it is not chance-corrected). Only one measure of agreement is available for continuous traits. One measure's agreement coefficient is shown (intraclass correlation coefficient for continuous traits; weighted Fleiss' K for ordinal traits; unweighted Fleiss' K for categorical traits). Coefficients with confidence intervals for all measures of agreement are available online (Supplementary material Appendix 1 Table A2). Traits are sorted from greatest to lowest agreement within data types.

Trait	Rank within data type mean ± SD	Agreement coefficient (ICC or Fleiss' K)
Continuous traits		
Wingspan (min)	1	0.88
Wingspan (max)	2	0.87
Wingspan (range)	3	0.56
Ordinal traits		
Generations (max)	1.8 ± 1.1	0.62
Generations (min)	1.8 ± 0.8	0.55
Flight period (end)	3.2 ± 1.1	0.48
Larval host plant breadth	3.6 ± 1.3	0.57
Flight period (start)	4.6 ± 0.5	0.40
Habitat breadth	6.0 ± 0.0	0.07
Categorical traits		
Overwintering stage	1.0 ± 0.0	0.93
Myrmecophily	2.4 ± 0.5	0.58
Mate search habit	3.2 ± 1.6	0.71
Habitat type: dry	5.2 ± 1.6	0.39
Habitat type: disturbed	5.4 ± 1.5	0.40
Habitat type: rocky	7.2 ± 3.7	0.30
Habitat type: wet	7.4 ± 1.3	0.39
Hilltopping	8.4 ± 3.0	−0.01
Habitat type: open	8.8 ± 2.9	0.38
Habitat type: shrub	9.2 ± 3.6	0.26
Habitat type: woods	10.2 ± 0.8	0.27
Mudpuddling	11.2 ± 1.8	0.22
Habitat type: woods' clearing	11.4 ± 1.1	0.20

The six methods of analysis for categorical and ordinal traits produced results that were qualitatively similar. Methods' coefficients differed in consistent ways from each other, with percent agreement and Gwet's AC1 consistently producing the highest and second-highest agreement coefficients, respectively, and Krippendorff's α often producing the lowest coefficients (Table 3). Visual examination of results suggests the analytical methods produced similar rankings of traits' inter-source agreement (Supplementary material Appendix 1 Table A2). Despite the overall similarity among methods, some notable differences arose (e.g. rocky habitat type had the third-highest agreement among categorical traits according to Gwet's AC1 but twelfth-highest agreement according to Conger's K).

Table 3. Comparison of the average ranking of six measures of agreement relative to each other. For each trait the six measures' coefficients were ranked from greatest to lowest values (i.e. one is the highest coefficient value, six is the lowest). Mean ± standard deviation (SD) rank positions for all 13 categorical (unweighted analysis) traits and 6 ordinal (weighted analysis) traits are presented.

Measure	Categorical rank mean ± SD	Ordinal rank mean ± SD
Percent agreement	1.0 ± 0.0	1.0 ± 0.0
Gwet's AC1	2.1 ± 0.3	2.0 ± 0.0
Brennan–Prediger	3.2 ± 0.4	3.0 ± 0.0
Conger's K	4.3 ± 1.2	4.5 ± 0.8
Fleiss' K	5.2 ± 0.8	5.2 ± 0.8
Krippendorff's α	5.3 ± 0.6	5.3 ± 0.8

Discussion

I found that sources were consistent in their information for some traits, but inconsistent for others. In general subjective traits (e.g. habitat association) seemed less consistent than objective traits (e.g. overwinter stage). This suggests researchers should exercise caution in relying on a single source for trait data.

Comparing measures of agreement

I found the five chance-corrected measures of inter-source agreement produced similar results of which traits were more consistent than others. Consistency in inter-rater agreement methods' rankings, with consistent differences in coefficient values among methods (i.e. some methods consistently produce larger coefficients than others), is consistent with the statistical literature (Warrens 2010). Regardless of method, researchers should report weighting values whenever weighted analyses are used, to allow replication.

It is difficult to compare rates of inter-source agreement among traits of different types of data (categorical, ordinal, and continuous). Traits of each type of data were analyzed in the way most appropriate for that data type, producing results that defy easy comparison. It is possible that ordinal traits, having characteristics in between categorical and continuous data types, could serve as a bridge allowing comparison among all three data types. Both categorical type (unweighted chance-corrected agreement measures such as Fleiss' K) and continuous type (intraclass correlation) measures of agreement can be calculated for ordinal traits. It might thus be possible to determine equivalent values of categorical and continuous agreement measures using ordinal traits as a bridge. However, analyzing data of a type unsuited for a method of analysis would violate statistical assumptions. The influence of such violations on the accuracy of agreement measures is not known, which is why I have avoided such analyses. Given the appeal of comparing agreement levels between data types, future biostatistics research could explore whether, and under what circumstances, it is acceptable to analyze data using methods targeted to other data types.

Agreement thresholds

How much agreement among sources is necessary? The answer depends on the research methods and goals, which is why I do not suggest agreement thresholds. Lintott et al. (2011) highlight the tradeoff between type I and type II errors associated with inter-rater agreement thresholds in their analysis of Galaxy Zoo data. Galaxy Zoo was an astronomy citizen science project in which volunteers classified pictures of galaxies based on their morphology. Only using galaxies with very high agreement among raters resulted in a small sample size of very reliable data, while low agreement thresholds produced large sample sizes of less reliable data. Classifications of 'merger galaxies' were accurate if as few as 40% of raters indicated the galaxy was this type, while other analyses with the same data set required more stringent agreement thresholds (Lintott et al. 2011), illustrating the perils of 'one size fits all' agreement thresholds.

Why did sources differ?

Why did sources sometimes differ in butterfly trait information? I provide four potential, non-exclusive reasons. First, differences among sources could reflect real intraspecific variation. I would expect such variation to be especially evident when comparing regional sources separated by vast geographic and climatic distances (e.g. southern populations have more generations per year than northern populations). Future studies could attempt to tease apart regional intraspecific variation from other sources of inter-source variation by comparing sources for the same target region, perhaps using birds for which there are numerous sources available for many regions. Second, differences among sources might reflect changes in our knowledge of butterflies over time. Newer sources might include more potential host-plants or behaviours, for example, than earlier sources. Ongoing changes in our understanding of biology is why there is a demand for updated editions of atlases and guides. Third, errors (Cole-Fletcher et al. 2011) and differences in interpretation among sources' authors would lead to inconsistency among sources. I would expect subjective traits to be more variable among sources than easily-quantified traits, an expectation consistent with my finding of habitat association having lower inter-source agreement than less ambiguous traits such as overwinter stage. I would also expect authors to provide more consistent information on common than rare species as a result of judgements being based on more experience, an expectation I could not evaluate with my data set as species' abundances differed among sources' focal regions. Fourth, choice of wording can result in different scoring among sources. For example, although the habitat breadth values for *Megisto cymela* differed substantially between Brock and Kaufman (2003; breadth = 1) and Wagner (2005; breadth = 4), their written descriptions are somewhat similar. Brock and Kaufman (2003, p. 230): 'Very common in woodland edges and clearings …', Wagner (2005, p. 132): 'Woods and forest edges, wooded swamps, and brushy fields …'. According to my habitat scoring system (Supplementary material Appendix 1 Table A1) Brock and Kaufman's description indicated only one habitat category (woods' clearing) while Wagner's description included four categories (woods, woods' clearing, wet and shrub). If sources shared a consistent categorization system (e.g. checkboxes for several habitat types) it would reduce this type of discrepancy both within and among sources. Such a system should not come at the expense of additional written descriptions of micro-habitat preferences. Consistent categorization of habitats is challenging but possible (Raciti et al. 2012, Ratnam et al. 2011). van Swaay's et al. (2006) survey of European butterfly experts provides one potential framework to develop consistent habitat terminology among numerous experts.

Assumptions and biases in trait-based research

Inconsistent trait information among sources has implications for the rapidly-growing field of trait-based biology. Many comparative studies rely on a single source for trait data. Trait data from a single source may be inaccurate, which can compromise the validity of results based on those data. For example, habitat breadth is used as a predictor variable in numerous studies that relied on a single source for scoring this trait (Dennis et al. 2004, Kotiaho et al. 2005), but according to my analyses habitat breadth data are inconsistent among sources. How sensitive the results of trait-based studies are to different sources being used for trait data is not yet known; research on this topic would help trait-based researchers estimate how robust results are when based on a single source for data.

My study contributes to a growing field critically evaluating the assumptions, biases, and justifications for trait-based research. Recent work in three areas is particularly noteworthy. First, researchers have found that not only are certain types of species more likely than others to have trait data available (Tyler et al. 2012), but also this bias in trait data 'missingness' can influence results of comparative analyses (González-Suárez et al. 2012). Second, assigning a single trait value to each species instead of integrating intraspecific variation can influence results of comparative analyses (Garamszegi and Møller 2010, Harmon and Losos 2005). Third, while biological conservation is often provided as a justification for comparative studies (e.g. on the traits associated with extinction risk), applied conservation workers do not rate trait-based comparative studies as particularly useful (Cardillo and Meijaard 2012). Thus the assumptions, biases, and motivations of trait-based research are under critical examination. Such examinations allow us to improve trait-based approaches. In my opinion trait-based research offers promise to reveal large-scale biological patterns and link diverse fields of research through their common connection to natural history (descriptive trait information for species).

How to account for inter-source variation in trait-based research

I recommend five non-exclusive strategies for trait-based comparative studies to minimize bias from reliance on a single source. First, researchers can include analyses of inter-source agreement for their traits; high agreement among sources suggests results are credible (but what agreement scores qualify as 'high' is subjective). Second, researchers could repeat their analyses using data from different sources; if results are similar regardless of data source then results are more credible. Third, researchers could incorporate multiple sources' data within a single analysis that accounts for intraspecific variation (inter-source variation) and phylogenetic autocorrelation. Multiple sources' estimates of trait means surely underestimate actual intraspecific variation, so interpretation of the 'intraspecific variation' component should be made with caution. Development of statistical methods that incorporate intra- and inter-specific variation provides a promising approach to comparative studies (Hadfield and Nakagawa 2010, Hansen and Bartoszek 2012, Revell and Reynolds 2012). Fourth, if researchers compile trait data from multiple sources into a single data set they should consider accounting for systematic biases in sources to be liberal or conservative in their trait data. For example, if I were to use wingspan range data from Opler et al. (2011), with missing species' wingspan ranges obtained from Bird et al. (1995), I should add 4 mm to all data coming from Bird et al. (1995) to account for its tendency among shared species to report ranges 4 mm narrower than those in Opler et al. (2011). While it is common for researchers to compile trait data from multiple sources, often employing a hierarchical strategy whereby trait data from alternative sources are used to fill in the gaps of a primary 'trump' source (Martin and Husband 2009, Betzholtz et al. 2013), it is uncommon for researchers to re-calibrate trait data to account systematic differences among sources. Finally, if a researcher must rely on a single source for data then they should explicitly mention this as a potential source of error.

I have revealed variable consistency in trait information among authoritative sources. Because mine is the first quantitative investigation of inter-source agreement for trait data, other studies are required to reveal the generality of my findings to sources for other taxa and regions. If inconsistency in sources' trait information is common, it has the potential to influence all trait-based research.

Acknowledgements – I thank Kilem Gwet for software help and Lauren Fitzsimmons for helpful comments on the manuscript.

References

Angert, A. L. et al. 2011. Do species' traits predict recent shifts at expanding range edges? – Ecol. Lett. 14: 677–689.

Banerjee, M. et al. 1999. Beyond kappa: a review of interrater agreement measures. – Can. J. Stat. 27: 3–23.

Beck, J. et al. 2006. Diet breadth and host plant relationships of southeast-Asian sphingid caterpillars. – Ecotropica 12: 1–13.

Berry, K. J. et al. 2008. Weighted kappa for multiple raters. – Perceptual Motor Skills 107: 837–848.

Betzholtz, P.-E. et al. 2013. With that diet, you will go far: trait-based analysis reveals a link between rapid range expansion and a nitrogen-favoured diet. – Proc. R. Soc. B 280: 20122305.

Bird, C. D. et al. 1995. Alberta butterflies. – Provincial Mus. Alberta.

Brock, J. P. and Kaufman, K. 2003. Kaufman field guide to butterflies of North America. – Houghton Mifflin Company.

Burke, R. J. et al. 2011. A mobility index for Canadian butterfly species based on naturalists' knowledge. – Biodivers. Conserv. 20: 2273–2295.

Cardillo, M. and Meijaard, E. 2012. Are comparative studies of extinction risk useful for conservation? – Trends Ecol. Evol. 27: 167–171.

Coad, B. W. 2012. The ROM field guide to freshwater fishes of Ontario, by E. Nicholas et al. 2009. [book review]. – Can. Field-Nat. 126: 173–174.

Cole-Fletcher, S. et al. 2011. Errors in length-weight parameters at FishBase.org. – Nature Precedings: 2011.5927.1.

Dennis, R. L. H. et al. 2004. Host plants and butterfly biology. Do host-plant strategies drive butterfly status? – Ecol. Entomol. 49: 12–26.

Ehrlich, P. R. and Raven, P. H. 1964. Butterflies and plants: a study in coevolution. – Evolution 18: 586–608.

Fitzpatrick, M. C. et al. 2009. Observer bias and the detection of low-density populations. – Ecol. Appl. 19: 1673–1679.

Garamszegi, L. Z. and Møller, A. P. 2010. Effects of sample size and intraspecific variation in phylogenetic comparative studies: a meta-analytic review. – Biol. Rev. 85: 797–805.

González-Suárez, M. et al. 2012. Biases in comparative analyses of extinction risk: mind the gap. – J. Anim. Ecol. 81: 1211–1222.

Gwet, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. – Brit. J. Math. Stat. Psych. 61: 29–48.

Gwet, K. L. 2010. Handbook of inter-rater reliability. – Advanced Analytics.

Hadfield, J. D. and Nakagawa, S. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. – J. Evol. Biol. 23: 494–508.

Hansen, T. F. and Bartoszek, K. 2012. Interpreting the evolutionary regression: the interplay between observational and biological errors in phylogenetic comparative studies. – Syst. Biol. 61: 413–425.

Harmon, L. J. and Losos, J. B. 2005. The effect of intraspecific sample size on type I and type II error rates in comparative studies. – Evolution 59: 2705–2710.

Janz, N. and Nylin, S. 1998. Butterflies and plants: a phylogenetic study. – Evolution 52: 486–502.

Kaufman, A. B. and Rosenthal, R. 2009. Can you believe my eyes? The importance of interobserver reliability statistics in observations of animal behaviour. – Anim. Behav. 78: 1487–1491.

Komonen, A. et al. 2004. The role of niche breadth, resource availability and range position on the life history of butterflies. – Oikos 105: 41–54.

Kotiaho, J. S. et al. 2005. Predicting the risk of extinction from shared ecological characteristics. – Proc. Natl Acad. Sci. USA 102: 1963–1967.

Layberry, R. A. et al. 1998. The butterflies of Canada. – Univ. of Toronto Press.

Lee, T. M. and Jetz, W. 2011. Unravelling the structure of species extinction risk for predictive conservation science. – Proc. R. Soc. B 278: 1329–1338.

Lintott, C. et al. 2011. Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. – Mon. Not. R. Astron. Soc. 410: 166–178.

Martin, S. L. and Husband, B. C. 2009. Influence of phylogeny and ploidy on species ranges of North American angiosperms. – J. Ecol. 97: 913–922.

Nakagawa, S. and Schielzeth, H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. – Biol. Rev. 85: 935–956.

Oliver, J. C. and Stein, L. R. 2011. Evolution of influence: signaling in a lycaenid-ant interaction. – Evol. Ecol. 25: 1205–1216.

Opler, P. A. et al. 2011. Butterflies and moths of North America. <www.butterfliesandmoths.org/> ver. June 2011. – Big Sky Inst.

Page, R. D. M. 2010. Wikipedia as an encyclopaedia of life. – Org. Divers. Evol. 10: 343–349.

Pelham, J. P. 2008. A catalogue of the butterflies of the United States and Canada. – J. Res. Lepidopt. 40: 1–658.

Pellissier, L. et al. 2012. Shifts in species richness, herbivore specialization, and plant resistance along elevation gradients. – Ecol. Evol. 2: 1818–1825.

Pierce, N. E. et al. 2002. The ecology and evolution of ant associations in the Lycaenidae (Lepidoptera). – Annu. Rev. Entomol. 47: 733–771.

Raciti, S. M. et al. 2012. Inconsistent definitions of "urban" result in different conclusions about the size of urban carbon and nitrogen stocks. – Ecol. Appl. 22: 1015–1035.

Ratnam, J. et al. 2011. When is a 'forest' a savanna, and why does it matter? – Global Ecol. Biogeogr. 20: 653–660.

Revell, L. J. and Reynolds, R. G. 2012. A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. – Evolution 66: 2697–2707.

Ricklefs, R. E. 2008. Foliage chemistry and the distribution of Lepidoptera larvae on broad-leaved trees in southern Ontario. – Oecologia 157: 53–67.

Rubel, W. and Arora, D. 2008. A study of cultural bias in field guide determinations of mushroom edibility using the iconic mushroom, *Amanita muscaria*, as an example. – Econ. Bot. 62: 223–243.

Rytwinski, T. and Fahrig, L. 2012. Do species life history traits explain population responses to roads? A meta-analysis. – Biol. Conserv. 147: 87–98.

Shields, O. 1967. Hilltopping: an ecological study of summit congregation behavior of butterflies on a southern California hill. – J. Res. Lepidopt. 6: 69–178.

Statzner, B. et al. 2001. Species traits and environmental constraints: entomological research and the history of ecological theory. – Annu. Rev. Entomol. 46: 291–316.

Tallamy, D. W. and Shropshire, K. J. 2009. Ranking lepidopteran use of native versus introduced plants. – Conserv. Biol. 23: 941–947.

Tyler, E. H. M. et al. 2012. Extensive gaps and biases in our knowledge of a well-known fauna: implications for integrating biological traits into macroecology. – Global Ecol. Biogeogr. 21: 922–934.

van Kleunen, M. et al. 2010. A meta-analysis of trait differences between invasive and non-invasive plant species. – Ecol. Lett. 13: 235–245.

van Swaay, C. et al. 2006. Biotope use and trends of European butterflies. – J. Insect Conserv. 10: 189–209 [with erratum on pages 305–306].

Wagner, D. L. 2005. Caterpillars of eastern North America. – Princeton Univ. Press.

Warrens, M. J. 2010. Inequalities between multi-rater kappas. – Adv. Data Anal. Classification 4: 271–286.

Wiklund, C. 2003. Sexual selection and the evolution of butterfly mating systems. – In: Boggs, C. L. et al. (eds), Butterflies: ecology and evolution taking flight. Univ. of Chicago Press, pp. 67–90.

Wong, P. B. Y. et al. 2011. Interpretations of polar bear (*Ursus maritimus*) tracks by Inuit hunters: inter-rater reliability and inferences concerning accuracy. – Can. Field-Nat. 125: 140–153.

Supplementary material (available online as Appendix oik-00207 at <www.oikosoffice.lu.se/appendix>). Appendix 1–2.