American Journal of Primatology 77:727–740 (2015)

REVIEW ARTICLE

Taking the Aggravation Out of Data Aggregation: A Conceptual Guide to Dealing With Statistical Issues Related to the Pooling of Individual-Level Observational Data

THOMAS V. POLLET[1], GERT STULP[2], S. PETER HENZI[3,4], AND LOUISE BARRETT[4]
[1]*Department of Social and Organizational Psychology, VU University Amsterdam, The Netherlands*
[2]*Department of Population Health, London School of Hygiene and Tropical Medicine, United Kingdom*
[3]*Department of Psychology, University of Lethbridge, Canada*
[4]*Applied Behavioural Ecology and Ecosystems Research Unit, University of South Africa, South Africa*

Field data often include multiple observations taken from the same individual. In order to avoid pseudoreplication, it is commonplace to aggregate data, generating a mean score per individual, and then using these aggregated data in subsequent analyses. Aggregation, however, can generate problems of its own. Not only does it lead to a loss of information, it can also leave analyses vulnerable to the "ecological fallacy": the drawing of false inferences about individual behavior on the basis of population level ("ecological") data. It can also result in Simpson's paradox, where relationships seen at the individual level can be completely reversed when analyzed at the aggregate level. These phenomena have been documented widely in the medical and social sciences but tend to go unremarked in primatological studies that rely on observational data from the field. Here, we provide a conceptual guide that explains how and why aggregate data are vulnerable to the ecological fallacy and Simpson's paradox, illustrating these points using data on baboons. We then discuss one particular analytical approach, namely multi-level modeling, that can potentially eliminate these problems. By highlighting the issue of the ecological fallacy, and increasing awareness of how datasets are often organized into a number of different levels, we also highlight the manner in which researchers can more positively exploit the structure of their datasets, without any information loss. These analytical approaches may thus provide greater insight into behavior by permitting more thorough investigation of interactions and cross-level effects. Am. J. Primatol. 77:727–740, 2015.

Key words: ecological fallacy; Simpson's paradox; aggregation; linear mixed modeling observation methods; pseudoreplication

INTRODUCTION

Direct observation of individual animals' behavior forms the cornerstone of much research in ethology and behavioral ecology. A variety of systematic sampling methods are used to collect such data, including instantaneous scan sampling, focal animal sampling, and one-zero sampling [Altmann, 1974, 1984; Martin & Bateson, 1993]. While there has been vigorous debate regarding the best means to produce an unbiased record of behavioral activity [e.g., Altmann, 1974; Ary & Suen, 1983; Baulu & Redmond, 1978; Bernstein, 1991; Fragaszy et al., 1992; Rhine & Linville, 1980; Simpson & Simpson, 1977; Suen & Ary, 1984], there has been much greater consensus on how to deal with data once they have been collected. At one time, it was common for researchers to simply pool repeated samples of behavior taken from several individuals into a single dataset for analysis. As Machlis et al. [1985] pointed out, however, this form of data pooling violates a number of fundamental statistical assumptions such as the need for independent

Contract grant sponsor: The Netherlands Organisation for Scientific Research; contract grant number: 451.10.032; contract grant sponsor: The Netherlands Organisation for Scientific Research; contract grant sponsor: Natural Sciences and Engineering Research Council of Canada (NSERC); contract grant sponsor: National Research Foundation (NRF); contract grant sponsor: NSERC; contract grant sponsor: Canada Research Chairs Program

*Correspondence to: Department of Social and Organizational Psychology, VU University Amssterdam, Transitorium Building (1B17), Van der Boechorststraat 1, 1081BT Amsterdam, The Netherlands. E-mail: t.v.pollet@vu.nl

Received 8 January 2014; revised 18 February 2015; revision accepted 1 March 2015

DOI: 10.1002/ajp.22405
Published online 24 March 2015 in Wiley Online Library (wileyonlinelibrary.com).

observations, i.e., pooling of this kind introduces pseudoreplication [for a similar, more recent plea to avoid pseudoreplication, within the field of neuroscience, see Aarts et al. 2014]. To avoid this "pooling fallacy", Martin & Bateson [1993] recommended that data should be aggregated within individuals to yield a single data point for each subject, and that these values should then be subject to statistical analysis. Although aggregating data avoids issues of pseudoreplication, generating a mean, median, proportion or total score at the individual level can also create problems of its own.

First, as most of us are aware, data aggregation leads to a loss of information; something we can easily lose sight of, however, in our efforts to avoid sample size inflation. Second, aggregated data are vulnerable to the "ecological fallacy." This is a term used in the social sciences to capture a phenomenon where data at the level of the group or population (i.e., data at the so-called "ecological" level) are used to draw inferences about individual traits and activities. This is a fallacy because the relationships detected at the aggregate level do not necessarily translate to similar relationships at the individual level. We are all familiar with this kind of reasoning to at least some extent: it is obvious to us all that, when we are told that people in a population have 2.4 children on average, this does not mean that there will be individuals who have this exact number of offspring.

It could be argued that any problems related to aggregation (or the lack of it) are largely a thing of the past, and most primatologists now follow Janson's [2012] advice to use multi-level models that can deal with the issue of multiple observations per individual. However, such a view would be at odds with recent pleas from fields as distinct as, psychology [Kievit et al., 2013; Pollet et al., 2014], epidemiology [Tu et al., 2008], measurement science [D'Errico, 2014], evolutionary ecology [Scheiner et al., 2000], and those involved in meta-analyses [Cooper & Patall, 2009; Hanley & Thériault, 2000], all of which argue that the ecological fallacy and the Simpson's paradox are more than just statistical curiosities, and constitute seriously under-appreciated problems of analysis and interpretation. Given that behavioral primatology is, traditionally, a rather less mathematical and statistical discipline than many of the aforementioned fields, we strongly suspect that problems of aggregation may go unrecognized in the discipline of primatology. This is especially likely given that a widely used textbook in animal behavior by Martin & Bateson [1993] specifically recommends aggregation in order to avoid the pooling fallacy,

It is also worth pointing out that, even if unnecessary aggregation is rare, the fact that it persists at all will be problematic if it leads others to generate and test hypotheses based on inappropriate inferences; even a single paper can result in problems down the line if the findings of such a paper prove influential. The current review therefore aims to provide a non-technical introduction to issues relating to data aggregation, particularly the ecological fallacy and Simpson's paradox for those unfamiliar with such problems, as well as providing a useful summary of alternative analytical approaches for those who have already moved away from simple forms of data aggregation [see also Janson, 2012; Schielzeth & Forstmeier, 2009; Waller et al., 2013].

In what follows, we first provide a general reminder of the problems associated with data aggregation—namely loss of information, reliability, and power in statistical analyses—before discussing how aggregation can generate both the ecological fallacy and Simpson's paradox in behavioral data. We then provide a brief conceptual outline of analytical strategies that avoid aggregation, and so lower the likelihood of committing the ecology fallacy. Most importantly—and more positively—these techniques possess the added advantage of allowing field researchers to exploit their data to its maximum potential.

Loss of Information Through Aggregation and Implications for Reliability

Data aggregation inevitably results in a loss of valuable information whenever multiple observations for a given animal are reduced to a single value (proportions/means/ratios). Of course, if the question of interest concerns an aggregate measure then this is a perfectly appropriate way to test the relevant predictions (e.g., Does animal X groom substantially more than other animals in the group?). In many cases, however, ignoring variation means ignoring information. Most people who conduct behavioral studies fully recognize that attending only to the central tendency and ignoring the variance can generate a misleading impression of how consistently an animal performs a certain behavior: animal X might groom on average more than other group members, but perhaps only under certain circumstances. Thus, although the "aggregate" hypothesis might be about individual's X grooming behavior, we can gain further insight by investigating matters at the level of individual observations. To give the most extreme example, if animal X's average grooming time is entirely a consequence of a single event then this clearly suggests something quite different from a situation in which animal X consistently spends more time grooming than all other animals across a series of time points.

Similarly, if we calculate a mean score on some variable for animals X and Y and obtain a score of 50 for both, we have succeeded in making them "identical" for the purposes of further statistical analysis based on those means, but if animal X's score is based on four observations of 0, 0, 100, and

100, then, clearly, it is behaving very differently to animal Y who scores 50 across the board. That is, aggregating the data smooths out the idiosyncrasies in animals' behavior in ways that may obscure the variety of behavioral strategies present. Reporting a measure of variation (such as the standard deviation) in addition to the central tendency obviously alleviates this problem to some extent. Nevertheless, it remains the case that such measures of variation may not be included in particular kinds of analysis.

Conversely, ignoring within-individual variation can, in some cases, suggest that a phenomenon is more robust than is actually the case. For example, as shown in the schematic graph presented in Figure 1A, aggregate values can provide a very neat correlation, and one might be tempted to conclude that there is a strong relationship between two measures. An examination of the non-aggregated values, as in Figure 1B, however, generates a quite different impression: the variation observed within each animal far exceeds that between animals, and the relationship between the two variables is correspondingly weaker (see Fig. 1A, B). Pollet et al., [2014], using a human example, found that the correlation between two variables (parasite stress and personality) at the aggregate level (country) is close to 20 times stronger than the correlation between those two variables at the lower, individual level. Similarly the statistical association between two variables *within* individuals might be much weaker than the association between those two variables *between* individuals.

Aggregation can also lead to a loss of information with respect to the reliability of estimates. For example, discovering that animal X performs a certain behavior 30% of the time based on six observations is very different from discovering that animal Y also performs the behavior at an identical rate but that, in this case, the estimate is based on a sample of 600 observations. The degree of confidence

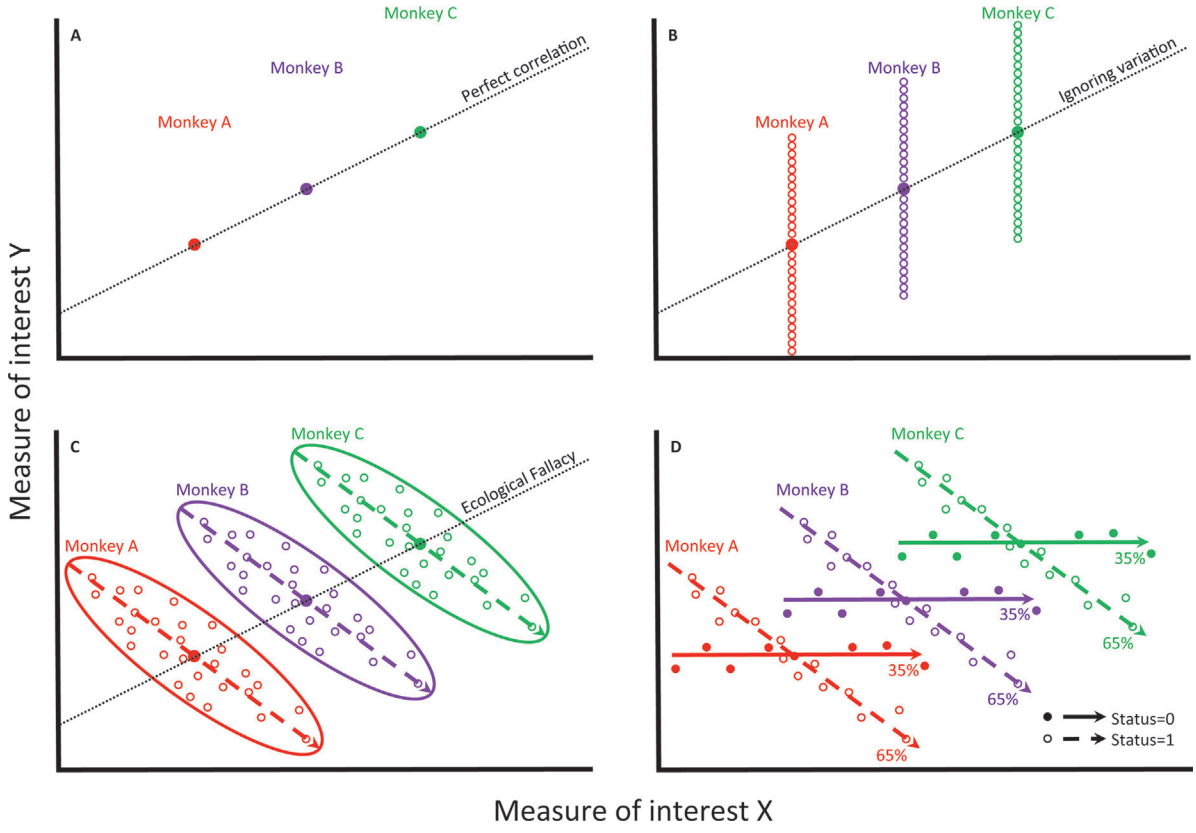

Fig. 1. A schematic overview of problems that may arise due to data aggregation. Simulated data for three monkeys (A, B, C; different colors); dotted lines reflect correlations through aggregated data; colored lines reflect individual level correlations. Aggregated data may (**A**) result in strong correlations, despite the fact that (**B**) individual variability is very high, and largely swamps the between-animal effect. The ecological fallacy arises when (**C**) aggregate data show a different relationship to data at the individual level, such that an inference about individual behavior based on aggregate data does not, in fact, correspond to actual patterns of individual behavior. The reversal in direction between population and individual levels of data shown here means that this example also represents Simpson's paradox. Using aggregate data may also prevent interaction effects from being (properly) tested. It is possible, for example, for the production of a certain behavior to depend on an animal's state (**D**). Calculating a single mean (or other aggregate) score for each individual (here, for instance, the percentage of time in a certain state; e.g., 15/23=65% measurements having Status 1) prevents the underlying variable to be associated with the other variables of interest, and hence there is no possibility of testing for an interaction effect. See text for further explanation; in the text we use grooming as our independent variable, aggressive behavior as the dependent variable, and reproductive state as the "Status" variable.

we can place in the inference that animal Y really does spend one third of its time performing the behavior compared to animal X is much higher in the case of the latter. Again, reporting additional information, such as sample size, can alert readers to the existence of inter-individual variability, but this does not always help if only mean values are then used in subsequent analysis. Given this, it is not difficult to see that aggregation has statistical implications: when data are aggregated, using only one [estimated] value per individual, all individuals are given equal weight in the analyses, despite the fact that there are much more reliable estimates available for some animals than others. One solution is simply to exclude data from animals with relatively few observations and attempt to increase reliability that way. The problem here is that the price of increasing the apparent reliability of the data is, again, the loss of potentially valuable information.

Finally, aggregating data may result in lower statistical power. If there are 10 observations for each of eight individuals, aggregating the data reduces the sample size for analysis to 8 data points, rather than the 80 that were actually collected. Observational field studies, particularly those on primates, often have quite small sample sizes to begin with [in terms of the number of animals and groups sampled], and aggregating data can lead to significance tests that are severely underpowered and, quite simply, unable to detect an effect even if one exists.

Given that differences in behavioral variability are highly relevant to particular research questions, especially in highly behaviorally flexible species like primates, aggregation can thus be a significant problem because it inevitably treats variability as though it were merely measurement error or statistical noise. The emphasis placed on avoiding pseudoreplication may have resulted, therefore, in the drawbacks of data aggregation being overlooked. Even worse perhaps, the emphasis on aggregation as a cure for the statistical errors generated by pseudoreplication has obscured the fact that aggregation in itself can lead to errors of statistical inference. Specifically, it increases the likelihood of falling prey to the ecological fallacy, and we now turn to explaining how and why this happens.

Inappropriate Inferences: the Ecological Fallacy

The first formal demonstration of the Ecological Fallacy (EF) was by Robinson [1950], although it should be noted that others, most notably Thorndike [1939], had previously pointed to similar issues. Robinson [1950] was, however, the first to provide a detailed mathematical demonstration of how and why relationships at the individual level could differ in magnitude from those at different levels of aggregation [e.g., state level, country level] [see also: Menzel, 1950; Selvin, 1958; Subramanian et al., 2009; Te Grotenhuis et al., 2011; van de Pol & Wright, 2009]. In essence, the fallacy is a numerical phenomenon that arises because variability around the aggregate means is substantially different from the variability seen at the individual level [see Piantadosi et al., 1988]. The ecological fallacy has been demonstrated in a variety of fields within the social sciences (e.g., criminology: [Dutton, 1994]; political science: [Seligson, 2002]; psychology: [Yammarino & Markham, 1992]; educational sciences: [Connolly, 2006]) and medicine (e.g., Pearce, 2000; Portnov et al., 2006; Yip & Liu, 2006).

A special case of the fallacy is known as Simpson's paradox (SP) [Simpson, 1951], although Pearson et al. [1899] and most notably Yule [1900, 1903] pointed to this phenomenon earlier [Good & Mittal, 1987], hence it also known as the Yule–Simpson effect, the reversal paradox and the amalgamation paradox [see Pearl, 2014 for review]. Here, the direction of a relationship within a number of individual groups is reversed as a consequence of conducting the same analysis at the level of the population. Robinson's [1950] classic demonstration of the EF, where he showed how a positive relationship between literacy rates and immigration across US states (i.e., the more immigrants present, the higher the state's literacy level) was reversed at the level of the individual (i.e., an immigrant to the US was, as one might expect, actually less likely to be literate in English than an American citizen) was, therefore, also a demonstration of Simpson's paradox. In some instances, Simpson's paradox arises for the same numerical reasons as the more general form of the ecological fallacy, but it can also occur due to the presence of a confounding third variable (see also Cooper & Patall [2009] for discussion and for more technical information on the ecological fallacy and Simpson's paradox: see e.g., [Blyth, 1972; Freedman, 1999; Greenland & Robins, 1994; Pearl, 2014; Piantadosi et al., 1988]).

The ecological fallacy and Simpson's paradox are relevant to behavioral researchers because, when sample data are aggregated to give a single score for each individual, they become vulnerable to the problems identified above, and for the same reasons. These problems appear to be overlooked in the context of behavioral studies [although Simpson's paradox is more widely recognized in some areas of evolutionary biology [e.g., Allison & Goldberg, 2002; Nee et al., 1991,1996; Scheiner et al., 2000; van de Pol & Wright, 2009].

Intuitively, it might seem odd to suggest that behavioral data collected from individuals are vulnerable to the ecological fallacy. Given that most behavioral studies are focused at the individual level, rather than at the "ecological" level of the population, it would seem that, almost by definition,

the ecological fallacy should not apply: indeed, aggregating data points to give a single score per individual level seems both logical and statistically legitimate. A simple example can demonstrate that this impression is mistaken.

Consider a study of baboons, where the research question is concerned with the relationship between grooming and aggression. When each baboon is considered individually, there is a relationship found such that, when an animal receives more grooming from another monkey, it will display less aggression towards that monkey directly after the grooming bout. Thus, a longer bout of grooming decreases subsequent aggression. The monkeys in the group also differ, however, in the absolute amount of time that they devote to these activities: monkey A, on average, spends the least amount of time being groomed, and shows the lowest frequency of aggression after grooming, monkey B falls in the middle, and engages in a moderate amount of both grooming and aggression, while monkey C is groomed the most, and also displays the highest levels aggression (see Fig. 1C). Calculating a single mean score for each animal and then using this score for further analysis will produce a positive relationship between grooming and aggression across animals, as the line on the graph indicates (Fig. 1C). This, then, is an exact reversal of the relationship documented for each individual monkey. Concluding on the basis of the aggregate-level results that higher levels of grooming are associated with higher levels of aggression *within* individuals, would therefore be an instance of the ecological fallacy. Indeed, in this case, it would be particularly misleading because the relationship is reversed, and it is also an example of Simpson's paradox. In effect, the results of this analysis would describe precisely how our animals do not behave, rather than how they do.

It is important to note once again that we are not suggesting that aggregate analyses have no value or are inherently flawed. Rather, our point is that one needs to be careful about the kinds of conclusions one can draw from such data, given the nature of the research questions asked. In this baboon example, it would be appropriate to use the aggregate data to conclude that the overall duration with which baboons engage in these behaviors tends to rise in parallel: that is, the longer the duration of the grooming bout, on average, the more aggression is displayed towards other baboons, on average, even though we cannot infer how these are connected within individuals.

One potential criticism here is that the ecological fallacy as described above is a trivial phenomenon. Intuitively, one might expect there to be a high level of correspondence between aggregate and lower-level data, in which case there should be only a remote possibility that relationships at lower levels would be weakened, lost or reversed when these are explored using higher-level aggregate data. In other words, any bias should be minimal. There are two points to make here. First, Firebaugh [1978] has shown that, for regression models with aggregated data, a lack of bias is the exception rather than the rule: only in the exceptional case when the pooled mean of the independent variable (X) has no effect on the (non-aggregated, lower level) dependent variable Y, with the (non-aggregated lower level) variable X controlled in the analyses, can we expect there to be no bias in the regression model [also see Sheppard, 2003]. Second, although a close correspondence between different levels of analysis is theoretically plausible [see Openshaw, 1984], empirically, it need not always hold true. In many circumstances it may therefore be dangerous to make this assumption, given that it can lead to inaccurate inferences being drawn from the data.

The extent to which the ecological fallacy occurs, and whether it is problematic (because, for example, it leads to a reversal in the direction of a relationship), remains largely unknown ([Openshaw, 1984] but see [Brand et al., 2010; Brand & Bradley, 2012; Piantadosi et al., 1988]). Several cases have, however, been discussed in epidemiology [e.g., Berlin et al., 2002; Berhane et al., 2004; Greenland & Robins, 1994; Greenland, 2001; Pearce, 2000; Portnov et al., 2006; Richardson et al., 1987] and the social sciences [e.g., Connolly, 2006; Dutton, 1994; Selvin, 1958; Yip & Liu, 2006]. To get a sense of whether such effects are likely to be found in observational datasets common to behavioral studies, we conducted an exploratory analysis on data drawn from SPH and LB's long-term study of chacma baboons in South Africa. Specifically, we used data collected for a study on juvenile development, consisting of focal animal samples of 13 juvenile baboons from one of their study troops (VT). Data were collected on, among other things, the juvenile animals' activity and their proximity to other group members, and, using these variables, we investigated whether we could observe an example of the ecological fallacy. Basically, we asked a simple, straightforward question: are juvenile baboons more likely to be found playing when they are in the vicinity of an adult female or when they are at a distance from an individual in this age-sex class? When we aggregate the data in the standard way, our analysis reveals a strong negative correlation between the average distance of juveniles from adult females (log-transformed) and the proportion of playful behavior in which they engaged (versus resting behavior) (Fig. 2: black line and black dots: $r = -0.60$, $P = 0.032$, $N = 13$). This suggests, perhaps, that juveniles need to be more vigilant when they are further way from adults, and so they avoid activities like play, and instead spend more time resting, an activity during which they can remain vigilant.

When we run a logistic regression for each juvenile monkey separately, however, we find that 12 of the 13 slopes are now positive, and that six of

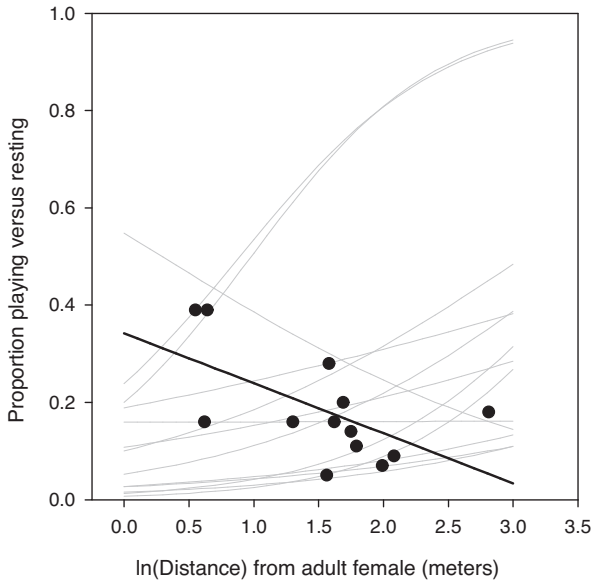

Fig. 2. The relationship between distance to an adult female group member (log-transformed) and the proportion of time spent playing (versus resting) for 13 juvenile baboons. Black dots reflect individual aggregated (mean) values; the black line is the OLS regression line through these points (Pearson $r=-0.60$; $P=0.032$; $N=13$). The grey lines are model predictions for each individual monkey based on logistic regression (see text); only one individual line had a negative slope.

these relationships are statistically significant (Fig. 2, grey lines). Only one individual displays a significant negative relationship. When we conduct our analyses at the individual level, then, we find that most juvenile baboons show an increased tendency to play when they are at a distance from adults, perhaps because adults are not particularly tolerant of playful juveniles, and move away from juvenile playgroups, or because adults use aggression to actively increase the distance between themselves and juveniles when the latter are playing. In other words, real-world data can show completely different patterns depending on the level at which the analysis is conducted, and can lead to inferences that directly oppose each other.

Simpson's paradox

As described above, Simpson's paradox (SP) is a special case of the ecological fallacy that occurs when the direction of a relationship is reversed completely when data are analyzed at the individual level compared the ecological level. As noted above, although the reversal of the statistical relationship in Simpson's Paradox may occur due to differences in values across groups (as was the case in Fig. 1C), another common reason why it occurs is due to the presence of a third variable that has a differential influence on relationships at the aggregate versus the individual level [see Ameringer et al., 2009; Armistead, 2014; Blyth, 1972; Hintzman, 1980; Pearl, 2014 for further description and analyses].

To illustrate this in concrete terms, imagine a study of nut-cracking by chimpanzees, where the central question of interest is the frequency with which different kinds of hammer (made of stone or wood) are used to crack nuts. Data are therefore collected on individual nut-cracking attempts, and mean values for each hammer type are calculated (Table 1). In addition, we also note whether the chimpanzees have cracked a small or a large nut (Table 1). When we examine the cracking of small nuts, we find that stone hammers have a higher success rate than wooden hammers (0.97 vs. 0.89). Similarly, larger nuts are also more successfully cracked by stone hammers (0.27 vs. 0.10). If the overall success rate of nut-cracking is calculated, however, stone hammers actually have a lower success rate than wooden hammers (0.76 vs. 0.81). In other words, the incidence of a given behavior shown by individuals in each of a number of groups can be consistently higher than those found in second series of groups but, once aggregated (i.e., in this case, combining nut size across hammer types to give a single value for each hammer type), the overall proportion observed in the combined grouping is opposite to that found in the individual groups (and vice versa, of course). To put this in terms of how this might arise in the field, imagine that some of the results are collected by Researcher A, who works on several chimpanzee communities at a study site containing only small nuts, while Researcher B conducts her work on a number of chimpanzee communities at a study site containing only large nuts. In each *local* ecology, success is greater with stone hammers. However, looking *across* ecologies, and combining the results from Researchers A and B, the chimpanzees perform better with wooden hammers. This is counter-intuitive (or paradoxical) at first glance, because the higher success rates for stone hammers compared to wooden hammers for both small and large nuts would suggest that when the data are combined the *overall* success rate would also be higher for stone hammers than for wooden ones Had we concluded that wooden hammers are a more effective tool on the basis of the ecological level proportions, we would have committed an ecological fallacy.

So, why exactly does this reversal occur? First, notice that the sizes of the groups for each combination are very different, but such information is

TABLE 1. A Hypothetical Case of Simpson's Paradox, Using Success Rate of Nut-Cracking by Chimpanzees

	Stone hammer		Wooden hammer	
Small nuts	340/350	0.97	400/450	0.89
Large nuts	40/150	0.27	5/50	0.10
Overall success	380/500	0.76	405/500	0.81

ignored when fractions are calculated (Table 1). We have larger sample sizes for small nuts being cracked than for large nuts being cracked (800 vs. 200). Moreover, relatively more small nuts are cracked with a wooden hammer than with a stone hammer (450 vs. 350), whereas the opposite is true for large nuts, where relatively more large nuts are cracked with stone hammers than with wooden hammer (150 vs. 50). Second, the third variable, nut size, has a large effect on the success of nut cracking: success is more strongly related to what kind of nut is being cracked, than to what kind of hammer is being used. Indeed, it appears that chimpanzees mostly use wooden tools for small nuts and stone tools for large nuts: wooden hammers are rarely used for the larger and less common nuts (which probably is due to the low success rate of the endeavour), whereas wooden hammers are often used for small nuts (perhaps because wooden hammers are more easily found, or perhaps they require less effort to use than stone hammers, and are thus slightly more energy-efficient than stone hammers for this job) (Table 1). Either way, the "paradox" arises because the stone hammers are more often used to crack difficult nuts with low success rates, whereas wooden hammers are more often used for easily-cracked nuts that have high success rates. Aggregating the data to overall success rate, results in a loss of this information, and the relatively high occurrence of the stone hammers and large nuts (and their attendant low success rate) drives the overall success rate of stone hammers down, whereas the relatively high occurrence of wooden hammers and small nuts (with their associated high success rate) will drive the overall success rate of wooden hammers up.

Thus, Simpson's Paradox applies to data presented in contingency tables, like the above chimpanzee example [and indeed it was originally identified in tabular data presented in this way, e.g., in studies on the gender admission bias at Berkeley [Bickel et al., 1975] but see [Wagner, 1982]]. Yet, it also applies to other types of data as seen in our baboon play example (Fig. 2) [for further explanation of Simpson's paradox and confounding variables, see [Ameringer et al., 2009; Armistead, 2014; Julious & Mullee, 1994; Pearce, 2000; Pearl, 2014; Samuels, 1993]; see also Figure 1C and [Kievit et al., 2013] for some examples using human psychological studies].

As Kievit et al. [2013] argue, Simpson's paradox, although treated as a "rare statistical curiosity" in the human psychological literature, may occur far more frequently than we realize, at least partly because we tend to be very bad at detecting the paradox when we observe it. Generally speaking, humans are not adept at reasoning with respect to more than two variables simultaneously, nor are we very good at recognizing conditional contingencies; it is one of our "inferential blindspots" [e.g., Dawid, 1979; Fiedler et al., 2003; Krämer & Gigerenzer, 2005; Kievit et al., 2013]. An experimental psychological study by Fiedler et al. [2003], for instance, on the understanding of Simpson's paradox among students, found that participants still drew incorrect conclusions based on the aggregated level in experimental situations where all relevant factors were made salient with varying degrees of explicitness, and even when all information was made entirely explicit (lower level percentages, higher level percentages, how the differences arise, e.g., in our case, telling them of the differential success rate of wooden and stone tools depending on nut size). Consequently, we can often draw incorrect causal inferences from aggregate data, even under circumstances when we have good knowledge of the relationships that exist at the lower levels [Kievit et al., 2013] (as could be the case with the hypothetical example of the chimpanzees above).

Lower Level Relationships Between Variables

One final way in which data aggregation creates as many problems as it solves is that it obscures our ability to detect more complex relationships, like statistical interaction effects. Assume, for instance, that the hypothetical relationship between grooming and aggression described above for baboons depends critically on a female's current reproductive state. More specifically, imagine that the negative relationship plotted between aggression and grooming is found only if the female is in the fertile stage of her menstrual cycle, and has a swollen perineal skin, whereas the relationship is non-existent for females that are not swollen (see Fig. 1D for schematic overview of such a statistical relationship). Calculating just the aggregate measures would not allow any investigation of this statistical interaction. For instance, aggregating the mean amount of time spent in each possible reproductive state (swollen, pregnant, lactating, and cycling) would produce the percentage of time that a given female spends in a swollen state, but that is not the information sought here. Moreover, sampling each female at a similar rate in both reproductive states (each female was observed in a fertile state X number of times and in a non-fertile state for Y number of times), would result in the aggregate measure being similar for each monkey (X/(X+Y)), again preventing any assessment of the relevant interaction. Put simply, once data are aggregated, relationships that exist only at the lower level cannot be tested; it is no longer possible to see them in the data, and so they cannot be pulled out of a statistical analysis. One straightforward way to deal with this problem is to follow the strategy used in the real-life baboon example above: calculate separate regressions for each individual. It would be preferable, however, to use *all* the data simultaneously and run an analysis that can account for both the

differences between individuals and the difference in reproductive status within individuals. One highly effective way to do this is by using multi-level modeling techniques and, in the next section, we provide a brief conceptual guide to their use.

A Potential Solution: Multi-level Modeling

Multi-level modeling is often used as a solution to the ecological fallacy in sociology and epidemiology [e.g., Goldstein, 2011; Hox & Kreft, 1994]. These statistical techniques are also known as mixed models, random coefficient models, random effect models, hierarchical models, and nested models. In the following, we will use the term multi-level modeling, and we will refer to these as GLMM (Generalized Linear Mixed Models). These are models that can accommodate nested data and deal with several distinct error distributions. Readers should, however, be aware that, within this class of methods, models can vary widely, for example in how they estimate parameters (e.g., Maximum Likelihood, Restricted Maximum Likelihood, Iterative Generalized Least Squares). It is also important to stress that GLMM's are not a 'magic bullet' or a panacea for all statistical ills: they might not accommodate all sampling schemes, for example. If a given sampling scheme is of poor quality, and produces inconsistent, noisy data, then it will simply be a matter of "garbage in, garbage out", regardless of the sophistication of the GLMM approach (or any other statistical technique) used. Similarly, if a sequential, temporal pattern in the data is disregarded, then invalid inferences may be drawn when such temporal structures are not taken into account even if observations are nested within individuals in a GLMM. GLMMs can also be complex to interpret and come with several assumptions of their own which should be borne in mind [for reviews see Bolker et al., 2009; Zuur et al., 2009 for example].

These caveats aside, multi-level models have achieved greater prominence in recent years [Janson, 2012], and an increasing number of primatological studies now use them [to give just a few examples: obeservational data Clarke et al., 2009; Gomes et al., 2009; Engelhardt et al., 2012; Koyama et al., 2012; Henzi et al., 2013, experimental field data: Wheeler, 2010; Ducheminsky et al., 2014; Price & Fischer, 2014], not least because they are an ideal way to combat potential problems of pseudoreplication. Waller et al. [2013] explicitly recommend the use of multi-level models (GLMM) as a way to combat pseudoreplication in experimental studies of primate communication. Multilevel models can do much more, however, than simply combat issues related to pseudoreplication.

First, multilevel models are well suited to deal with problems of the ecological fallacy (and Simpson's Paradox), by controlling for possible clusters within datasets due to individual variability and/or the influence of third variables (provided that these have been measured). Second, these powerful techniques allow researchers to extract the maximum amount of information from their data and, as indicated above, investigate more complex relationships than is possible with data based on aggregate values. It would be a mistake to think that multi-level models are useful solely because they simply avoid particular kinds of analytical errors. Rather, they can be exploited in a number of positive and creative ways that permit deeper and more thorough interrogations of data.

Our intention here, then, is to explain how and why multi-level modeling deals effectively with ecological fallacies and the other drawbacks of data aggregation, and then to highlight a few of the positive advantages that multi-level models offer to the analysis of behavioral datasets, once we recognize their hierarchical nature. Our description of these models is—quite intentionally—more of a "sales pitch" than a "how-to" guide or instruction manual. There are a number of excellent books and papers available, all of which do a much better job than we can offer here, and we urge interested readers to seek out these resources and follow the guidance offered therein [e.g., Bolker et al., 2009; Gelman & Hill, 2006; Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999, 2012; van de Pol & Wright, 2009; Zuur et al., 2009].

One major reason why multilevel modeling has gained ground in behavioral research is because there are now a number of statistical packages that allow one to implement multilevel techniques easily and efficiently, including MLWin [Rasbash et al., 2000], SAS [Singer, 1998; Littell, 2006], SPSS mixed package [IBM SPSS Statistics, 2011], Stata [Stata Press, 2005], and R [R Development Core Team, 2008] lme package, and later versions such as lme4 [e.g., Pinheiro et al., 2007] and MCMCglmm [Hadfield, 2010]. Computing power is no longer a limiting factor on behavioral researchers' ability to conduct sophisticated statistical analyses, and some software packages, like R [R Development Core Team, 2008], are open-source and free to access. There is no reason not to take advantage of the more powerful statistical techniques now on offer. To drive this point home, we also want to highlight the way that these models can use the full range of data that a researcher has available, rather than having to rely on aggregate statistics, which is a key advantage in our view. Modern statistical packages therefore yield substantially more statistical power as a consequence [for reviews on power and multilevel models see: Maas & Hox, 2005; Snijders, 2005; Scherbaum & Ferreter, 2009].

So, how do multi-level models allow us to combat the problems created by data aggregation? To take a

simple case, they do so by nesting data points within individual animals, which takes into account individual variability without any loss of data. In the simplest case, a "random-intercept" model is used, where the data for each animal is fitted with its own intercept. In such cases, the statistical model "knows" that the different data points belong to the same individual thereby accounting for the non-independence in the data [although some argue that it achieves this only partially: Schielzeth & Forstmeier, 2009]. Random intercept models recognize that individuals will show variation with respect to, say, the overall frequency with which they perform a given behavior, but they make the simplifying assumption that the direction and strength of the relationship is identical for each individual. That is, they take the form shown in Figure 1C: if we were to extend the line drawn through each individual back to the y-axis, they would all hit the axis at a different point (which is related to the predicted mean frequency with which the behavior is displayed by that animal when the X-variable is set to zero), but nevertheless all the relationships would be negative, and the lines would be parallel because the slopes would be similar. Of course, as we saw in our real-life example (Fig. 2), the relationships shown by different individuals need not be at all similar: most of our juvenile baboons displayed a positive relationship, but these differed quite substantially in magnitude, plus there was one odd individual that showed a negative relationship. With a dataset like this, an intercept-only model, which constrains all the data points to lie on the same slope, would not give a very good fit [see also Schielzeth & Forstmeier, 2009].

Mixed models, however, can easily be extended to incorporate a random slope as well as a random intercept, so that the nature of the relationship is allowed to vary for each individual, thus accounting for a second kind of variability. If we apply such an analysis to the juvenile baboon data (Fig. 2), there is a significant positive overall effect of distance on the likelihood of playing when using a random intercept-only model (binomial mixed model parameter estimate ($\pm$ SE) = 0.90 ($\pm$ 0.063); $P < 0.0001$, AIC = 2697.9). When the data are allowed to vary not only in their intercept, but also in their slope, there is an overall significant positive effect (0.61 ($\pm$ 0.017); $P = 0.0002$, AIC = 2660.6), albeit weaker, but the analysis also reveals that there is large variation in the slopes between the individuals. That is, including a random slope improves the model fit relative to the intercept-only model, which we can see by considering the difference in the Akaike Information Criterion (AIC) values. Here, ΔAIC = 37.3. The AIC value is a measure of model fit, based on the loglikelihood of the model where, in this case, lower values indicate a model with a better fit to the data [Akaike, 1974; Burnham & Anderson, 2002, 2004; Bolker et al., 2009; Symonds & Moussalli, 2011; Zuur et al., 2009]. Such an analysis is highly informative, because it allows one to conclude that, overall, across individuals, the relationship between X and Y is positive (juveniles who are further away from adult females tend to display more play behavior), but it also points to the high variability seen between the juveniles, which may be of interest in its own right. It is important to note, however, that this variation might change if we were control for additional individual factors, such as juvenile sex, age, maternal dominance rank, or season. Here, we performed a rather quick-and-dirty analysis to demonstrate that a multi-level model produces results that correspond more closely with the series of individual regressions, rather than with the negative relationship found using aggregate data. Indeed, a thorough analysis would require the data to be explored more fully prior to beginning any analysis and would consider, for instance, the distribution of the dependent variable and correct for any possible overdispersion [e.g., Bolker et al., 2009; Molenberghs et al., 2007; Zuur et al., 2009]. It would also be advisable to check the residuals of any analyses (for the possibility of heteroscedasticity, for example). It is also possible to bootstrap the results or employ another technique to check if the results are upheld consistently [e.g., MCMCglmm: Hadfield, 2010]. Other authors have outlined the steps that can be followed to check the results obtained from each model run [Bolker et al., 2009; Zuur et al., 2009].

Generally speaking, then, while the random intercept controls for systematic error variance as a consequence of group membership, random slopes model systematic error variance as a consequence of the attributes associated with belonging to a particular group (or individual, if we have observations nested in individuals). In addition, models that combine both a random slope and a random intercept can control for non-independence even more effectively than intercept-only models, and are argued to be preferable for this reason [Schielzeth & Forstmeier, 2009]. For examples of random effects models and more detailed explanation, we refer readers to the existing literature [e.g., Bolker et al., 2009; Gelman & Hill, 2006; Gillies et al., 2006; McCulloch & Neuhaus, 2001; McCulloch, 2006; Pinheiro & Bates, 2000; Raudenbush, 1994; Snijders & Bosker, 1999, 2012; Raudenbush & Bryk, 2002; Verbeke & Molenberghs, 2009; Zuur et al., 2009].

The advantages of multi-level models for observational data should be readily apparent from the above explanation. The same is also true for experimental designs. In what can be viewed as an extension of Waller et al.'s [2013] paper (even though it was published four years earlier), van de Pol & Wright [2009] demonstrated that mixed models can be used very effectively to separate between- and within-individual effects in experimental designs, at

least in part by solving the problems associated with the ecological fallacy. Van de Pol & Wright [2009] suggest that the use of such models is particularly appropriate for studies relating to reproductive timing, sex allocation, and anti-predator behavior because, for each of these topics, different predictions arise from one's hypotheses depending on whether they are focused on the within-subjects level or the between-subjects level. Van de Pol and Wright's [2009] paper therefore demonstrates the positive contribution that multi-level modelling can make to the analysis of behavioral data, and not just the "negative" contribution of ensuring any erroneous inferences are kept under control. For example, applying this logic and method to the example shown in Figure 1C, one could simultaneously assess the relationship between the amount of grooming an individual receives and the amount of aggression displayed (the within-individual effect), as well as the relationship between the average amount of grooming received and the amount of aggression displayed (the between-individual effect).

"Scaling up": Adding Further Levels and Longitudinal Data

Another positive—and very useful—aspect of modern multi-level techniques is that they allow for the classification of many different levels of groups. For example, the behavior of the juvenile baboons in the sample analyzed above not only can be nested within individuals, but also within social cliques, which could then be further nested into different groups (and if they were gelada or hamadryas baboons, rather than these rather unexciting chacma baboons, we could further nest the data within even larger groupings, such as clans and herds).

Investigating membership in more than one group at a time

Multi-level models can also accommodate the social reality that an individual is often a member of many different groups, which may or may not overlap in membership (known as "cross-classified" models). For example, in a school context, a student may attend a biology class with a certain group of people, but also attend a chemistry class with another group of people, who do not all take the biology class. If it was necessary to predict this student's grades in biology based on some aspect of the teacher's instructional methods then, because the teacher instructs some of the student's fellow pupils but not others, cross-classified models could be used to examine the independent influence of the teacher on the student's grades, as well as account for the influence of overlap in the composition of the class. It is also possible to specify multiple groups with non-overlapping membership, where each level might differentially help to predict grades. At first sight, it might seem that multiple group membership is rare and hence not relevant to non-human animals. In fact, there are many cases where multiple group membership occurs and would be useful to model. One obvious example is those species that show fission-fusion dynamics, where animals display flexible sub-group membership across time. Cross-classified models can help tease apart the ways in which sub-group membership influences foraging and social strategies based on the presence or absence of particular individuals or particular age-sex classes.

Even within stable social systems, multiple group membership may occur: an animal may form an alliance with other more powerful individuals in order to increase the chances of mating success, but forage close to those that are more subordinate in order to increase its foraging success. Thus, even among non-human animals, "group" membership does not necessarily overlap completely, and multi-level techniques can allow us to more clearly differentiate the effects of membership in one group versus another in ways that are would not be possible using simple correlational techniques based on aggregate data.

Modeling who collected the data

Another, much more pragmatic, application of multi-level modeling that fieldworkers in particular might find useful is to specify the identity of the field assistants who collect the data, and nest observations within those individuals. In field settings, standard inter-rater reliability procedures (e.g. Cohen's kappa, Krippendorf alpha [Hayes & Krippendorff, 2007] often are not feasible because field assistants rarely collect exactly the same data on the same animals at the same time (i.e., they usually do not overlap in their data sampling). For example, over the course of the long-term De Hoop baboon project, data were collected by at least 12 different people over a period of 10 years. Modeling non-independence using multi-level techniques can be very useful in such circumstances. For instance, a multi-level modeling approach allows one to disentangle the effects of field assistant attributes (e.g., perhaps some field assistants and researchers consistently over- or underestimate distances) from other factors potentially influencing observations (e.g., perhaps weather conditions affect estimates of distance). Aggregating the data obscures the potential effect of influences like these and, as with our discussion of interaction effects above, prevents us from considering whether factors, like researcher identity, have any influence on our results, and whether these interact with other influences, such as climatic factors.

Working with sequential data

One final advantage of multi-level models that deserves a brief mention is that they allow us to

retain the temporal sequencing of repeated measures within longitudinal data. Repeated measures are a commonly recognized source of non-independence (and often result in autocorrelation). Such repeated measures represent one of the main reasons why multi-level models are used; they allow for the construction of random effects that take account of several data points for each individual over a period of time. On occasion, however, we might well be interested in the nature of the relationships that exist between these repeated measures (i.e., we may want to know something about whether a certain behavior occurs every day, or whether it tends to happen in bursts over several days followed by a period when it does not occur at all). If we were to aggregate to a mean rate of occurrence per unit time for each individual, we would lose this sequencing information, plus we again become vulnerable to the ecological fallacy.

The simplest way to deal with this issue is to include "time" as a fixed factor in a multi-level model, and use this to account for patterns in our data. Alternatively, we could model time via additional random effects in a multi-level model [see Goldstein et al., 1994 for example] or we could use multi-level time-series models [for example, random effects Cox regression, frailty models, or multi-level Cox regression models; Mills, 2011]. All these models help to ensure we draw the correct inference when a time component is present in the data, because they retain the actual frequency of behavior as it was distributed across a given time period. These models thus allow us to make inferences over the long-term, while taking into account variation over that period, without data loss. In some cases, we can even make future predictions based on an observed time series [forecasting models, see Box et al., 2013].

CONCLUSION

Although problems relating to data pooling, like pseudoreplication, are well known in the behavioral literature, and researchers now make concerted efforts to avoid such statistical errors (but see Waller et al., 2013 who discuss this problem in primate communication research), one of the solutions proposed—data aggregation—ironically creates statistical problems of its own. Despite the fact that aggregation seems both intuitive and straightforward when analyzing data, any procedure that reduces individual measures to a single data point results in both an overall loss of available information and the rather more insidious problems of the ecological fallacy and Simpson's paradox, both of which can result in false inferences being drawn.

Although this survey has been necessarily brief, we hope that by highlighting the problems that can arise via data aggregation and pointing out some of the positive advantages of multi-level models, we will both convince more people to begin using such models, as well as encourage those who already do so to exploit them to even greater advantage. At the same time, it is important to acknowledge that the multilevel approach is not the holy grail of data analysis. In some cases, an aggregated value may be precisely what is required to test a specific prediction, while in others small sample sizes or insufficient data at each level may preclude their use (though some options to combat these also exist such as MCMCglmm [Hadfield, 2010]). In many cases, however, aggregation may prove to be unnecessary and the research questions asked would benefit from a multilevel approach. Indeed, our survey of problems and solutions relating to the ecological fallacy suggests that, in many cases of behavioral observational studies, our understanding of the behavior would be further advanced by the use multi-level modeling. These techniques therefore offer genuine promise for more sophisticated and creative forms of hypothesis testing. The speed and ease with which such models can be constructed and run, using dedicated software, more than compensates for the time needed to learn such techniques. Given the immense amount of time and energy required to collect good-quality field data, the use of techniques that allow field researchers to exploit all their data to its fullest potential are undoubtedly preferable to less powerful methods that require some very hard-earned data simply to be discarded.

ACKNOWLEDGMENTS

There is no known conflict of interest. TVP was supported by The Netherlands Organisation for Scientific Research (Veni scheme; 451.10.032). GS is supported by The Netherlands Organisation for Scientific Research (Rubicon Fellowship). SPH is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant and the National Research Foundation (NRF) South Africa. LB is supported by a NSERC Discovery Grant and the Canada Research Chairs Program. We thank Peter Dekker for statistical advice, and Willem Frankenhuis, Ian Rickard, Joshua Tybur for fruitful discussion of some of the issues outlined in this paper. In particular, we thank Tim Fawcett for his extensive and very helpful comments on an earlier draft, and the comments of Tony Di Fiore and three anonymous referees, which greatly improved the manuscript.

REFERENCES

Aarts E, Verhage M, Veenvliet JV, Dolan CV, van der Sluis S. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. Nature Neuroscience 17: 491–496.

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19:716–723.
Allison VJ, Goldberg DE. 2002. Species-level versus community-level patterns of mycorrhizal dependence on phosphorus: an example of Simpson's paradox. Functional Ecology 16:346–352.
Altmann J. 1974. Observational study of behavior: sampling methods. Behaviour 49:227–266.
Altmann J. 1984. Observational sampling methods for insect behavioral ecology. Florida Entomologist 67:50–56.
Ameringer S, Serlin RC, Ward S. 2009. Simpson's Paradox and Experimental Research. Nursing Research 58:123–127.
Armistead TW. 2014. Resurrecting the Third Variable: A Critique of Pearl's Causal Analysis of Simpson's Paradox. The American Statistician 68:1–7.
Ary D, Suen HK. 1983. The use of momentary time sampling to assess both frequency and duration of behavior. Journal of Psychopathology and Behavioral Assessment 5:143–150.
Baulu J, Redmond DE. 1978. Some sampling considerations in the quantitation of monkey behavior under field and captive conditions. Primates 19:391–399.
Berhane K, Gauderman WJ, Stram DO, Thomas DC. 2004. Statistical issues in studies of the long-term effects of air pollution: The Southern California Children's Health Study. Statistical Science 19:414–449.
Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. 2002. Individual patient-versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in Medicine 21:371–387.
Bernstein IS. 1991. An empirical comparison of focal and ad libitum scoring with commentary on instantaneous scans, all occurrence and one-zero techniques. Animal Behaviour 42:721–728.
Bickel PJ, Hammel EA, O'Connell JW. 1975. Sex Bias in Graduate Admissions: Data from Berkeley. Science 187:398–404.
Blyth CR. 1972. On Simpson's paradox and the sure-thing principle. Journal of the American Statistical Association 67:364–366.
Bolker BM, Brooks ME, Clark CJ, et al. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.
Box GEP, Jenkins GM, Reinsel GC. 2013. Time series analysis: forecasting and control. New York, NY: Wiley.
Brand A, Bradley MT. 2012. More voodoo correlations: When average-based measures inflate correlations. The Journal of General Psychology 139:260–272.
Brand A, Bradley M, Best L, Stoica G. 2010. Multiple Trials May Yield Exaggerated Effect Size Estimates. The Journal of General Psychology 138:1–11.
Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York, NY: Springer.
Burnham KP, Anderson DR. 2004. Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33:261–304.
Clarke PMR, Henzi SP, Barrett L. 2009. Sexual conflict in chacma baboons, Papio hamadryas ursinus: absent males select for proactive females. Animal Behaviour 77:1217–1225.
Connolly P. 2006. Summary statistics, educational achievement gaps and the ecological fallacy. Oxford Review of Education 32:235–252.
Cooper H, Patall EA. 2009. The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. Psychological Methods 14:165–176.
Dawid AP. 1979. Conditional independence in statistical theory. Journal of the Royal Statistical Society. Series B (Methodological) 1–31.
D'Errico GE. 2014. Aggregation of comparisons data and reversal phenomena of metrological interest. Measurement 50:319–323.
Ducheminsky N, Barrett L, Henzi SP. 2014. Responses of vervet monkeys in large troops to land and aerial predator alarm calls. Behavioral Ecology. 25:1474–1484.
Dutton DG. 1994. Patriarchy and wife assault: The ecological fallacy. Violence and Victims 9:167–182.
Engelhardt A, Fischer J, Neumann C, Pfeifer J-B, Heistermann M. 2012. Information content of female copulation calls in wild long-tailed macaques (Macaca fascicularis). Behavioral Ecology and Sociobiology 66:121–134.
Fiedler K, Walther E, Freytag P, Nickel S. 2003. Inductive reasoning and judgment interference: Experiments on Simpson's paradox. Personality and Social Psychology Bulletin 29:14–27.
Firebaugh G1978. A Rule for inferring individual-level relationships from aggregate data. American Sociological Review 43:557–572.
Fragaszy DM, Boinski S, Whipple J. 1992. Behavioral sampling in the field: comparison of individual and group sampling methods. American Journal of Primatology 26:259–275.
Freedman DA. 1999. Ecological inference and the ecological fallacy. International Encyclopedia of the Social and Behavioral sciences 6:4027–4030.
Gelman A, Hill J. 2006. Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press.
Gillies CS, Hebblewhite M, Nielsen SE. et al. 2006. Application of random effects to the study of resource selection by animals. Journal of Animal Ecology 75:887–898.
Goldstein H2011. Multilevel statistical models. New York, NY: Wiley.
Goldstein H, Healy MJR, Rasbash J. 1994. Multilevel time series models with applications to repeated measures data. Statistics in Medicine 13:1643–1655.
Gomes CM, Mundry R, Boesch C. 2009. Long-term reciprocation of grooming in wild West African chimpanzees. Proceedings of the Royal Society B: Biological Sciences 276:699–706.
Good IJ, Mittal Y. 1987. The Amalgamation and Geometry of Two-by-Two Contingency Tables. The Annals of Statistics 15:694–711.
Greenland S. 2001. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. International Journal of Epidemiology 30:1343–1350.
Greenland S, Robins J. 1994. Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. American Journal of Epidemiology 139:747–760.
Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. Journal of Statistical Software 33:1–22.
Hanley JA, Thériault G. 2000. Simpson's paradox in meta-analysis. Epidemiology 11:613.
Hayes AF, Krippendorff K. 2007. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 1:77–89.
Henzi SP, Forshaw N, Boner R, Barrett L, Lusseau D. 2013. Scalar social dynamics in female vervet monkey cohorts. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368:20120351.
Hintzman DL. 1980. Simpson's paradox and the analysis of memory retrieval. Psychological Review 87:398.
Hox JJ. 2010. Multilevel analysis: Techniques and applications. London, UK: Taylor & Francis.
Hox JJ, Kreft IGG. 1994. Multilevel analysis methods. Sociological Methods & Research 22:283–299.
IBM SPSS Statistics 2011. IBM SPSS Statistics 20.0. Chicago, IL: SPSS Inc.

Janson CH. 2012. Reconciling Rigor and Range: Observations, Experiments, and Quasi-experiments in Field Primatology. International Journal of Primatology 33:520–541.
Julious SA, Mullee MA. 1994. Confounding and Simpson's paradox. BMJ 309:1480–1481.
Kievit RA, Frankenhuis WE, Waldorp LJ, Borsboom D. 2013. Simpson's paradox in psychological science: a practical guide. Frontiers in Psychology 4:513.
Koyama NF, Caws C, Aureli F. 2012. Supply and demand predict male grooming of swollen females in captive chimpanzees. Pan troglodytes. Animal Behaviour 48:1419–1425.
Krämer W, Gigerenzer G. 2005. How to Confuse with Statistics or: The Use and Misuse of Conditional Probabilities. Statistical Science 20:223–230.
Littell RC. 2006. SAS for mixed models. Cary, NC: SAS institute.
Maas CJM, Hox JJ. 2005. Sufficient sample sizes for multilevel modeling. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 1:86–92.
Machlis L, Dodd PWD, Fentress JC. 1985. The pooling fallacy: problems arising when individuals contribute more than one observation to the data set. Zeitschrift für Tierpsychologie 68:201–214.
Martin PR, Bateson P. 1993. Measuring behaviour: an introductory guide. Cambridge, UK: Cambridge University Press.
McCulloch CE. 2006. Generalized linear mixed models. New York, NY: Wiley Online Library.
McCulloch CE, Neuhaus JM. 2001. Generalized linear mixed models. New York, NY: Wiley Online Library.
Menzel H. 1950. Comment on Robinson's "Ecological correlations and the behavior of individuals". American Sociological Review 15:674.
Mills M. 2011. Introducing survival and event history analysis. London, UK: Sage.
Molenberghs G, Verbeke G, Demétrio CGB. 2007. An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime data analysis 13:513–531.
Nee S, Read AF, Greenwood JJD, Harvey PH. 1991. The relationship between abundance and body size in British birds. Nature 351:312–313.
Nee S, Read AF, Harvey PH. 1996. Why phylogenies are necessary for comparative analysis. In: Martins EP, editor. Phylogenies and the comparative method in animal behavior. Oxford, UK: Oxford University Press. p 399–411.
Openshaw S. 1984. Ecological fallacies and the analysis of areal census data. Environment and Planning A 16:17–31.
Pearce N. 2000. The ecological fallacy strikes back. Journal of Epidemiology and Community Health 54:326–327.
Pearl J. 2014. Comment: Understanding Simpson's Paradox. The American Statistician 68:8–13.
Pearson K, Lee A, Bramley-Moore L. 1899. Mathematical Contributions to the Theory of Evolution. VI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. Philosophical. Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 192 257–330.
Piantadosi S, Byar DP, Green SB. 1988. The ecological fallacy. American Journal of Epidemiology 127:893–904.
Pinheiro JC, Bates DM. 2000. Linear mixed-effects models: basic concepts and examples. New York, NY: Springer.
Pinheiro J, Bates D, DebRoy S. 2007. Linear and nonlinear mixed effects models. R package version 3:57.
Pollet TV, Tybur JM, Frankenhuis WE, Rickard IJ. 2014. What can cross-cultural correlations teach us about human nature? Human Nature (Hawthorne, N.Y.) 25:410–429.
Portnov BA, Dubnov J, Barchana M. 2006. On ecological fallacy, assessment errors stemming from misguided variable selection, and the effect of aggregation on the outcome of epidemiological study. Journal of Exposure Science and Environmental Epidemiology 17:106–121.
Price T, Fischer J. 2014. Meaning attribution in the West African green monkey: influence of call type and context. Animal cognition 17:277–286.
R Development Core Team 2008. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
Rasbash J, Browne W, Goldstein H. et al. 2000. A user's guide to MLwiN. London, UK: University of London, Institute of Education, Centre for Multilevel Modelling.
Raudenbush SW. 1994. Random effects models. In: Cooper H, Hedges L, editors. The Handbook of Research Synthesis. New York, NY: Russell Sage Foundation. p 301–321.
Raudenbush SW, Bryk AS. 2002. Hierarchical linear models: Applications and data analysis methods. London: Sage Publications.
Rhine RJ, Linville AK. 1980. Properties of one-zero scores in observational studies of primate social behavior: The effect of assumptions on empirical analyses. Primates 21: 111–122.
Richardson S, Stücker I, Hémon D. 1987. Comparison of Relative Risks Obtained in Ecological and Individual Studies: Some Methodological Considerations. International Journal of Epidemiology 16:111–120.
Robinson WS. 1950. Ecological correlations and the behavior of individuals. American Sociological Review 15:351–357.
Samuels ML1993. Simpson's Paradox and Related Phenomena. Journal of the American Statistical Association 88: 81–88.
Scheiner SM, Cox SB, Willig M. et al. 2000. Species richness, species-area curves and Simpson's paradox. Evolutionary Ecology Research 2:791–802.
Scherbaum CA, Ferreter JM. 2009. Estimating statistical power and required sample sizes for organizational research using multilevel modeling. Organizational Research Methods 12:347–367.
Schielzeth H, Forstmeier W. 2009. Conclusions beyond support: overconfident estimates in mixed models. Behavioral Ecology 20:416–420.
Seligson MA. 2002. The Renaissance of Political Culture or the Renaissance of the Ecological Fallacy?. Comparative Politics 34:273–292.
Selvin HC 1958 Durkheim's Suicide and Problems of Empirical Research. American Journal of Sociology 63: 607–619.
Sheppard L. 2003. Insights on bias and information in group-level studies. Biostatistics 4:265–278.
Simpson EH. 1951. The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society. Series B (Methodological) 13:238–241.
Simpson MJA, Simpson AE. 1977. One-zero and scan methods for sampling behaviour. Animal Behaviour 25:726–731.
Singer JD. 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models,;1; and individual growth models. Journal of Educational and Behavioral Statistics 23:323–355.
Snijders TAB. 2005. Power and sample size in multilevel linear models. Encyclopedia of statistics in behavioral science 3:1570–1573.
Snijders TAB, Bosker RJ. 1999. Multilevel analysis: An introduction to basic and advanced multilevel modeling. London, UK: Sage Publications.
Snijders TAB, Bosker RJ. 2012. Multilevel analysis: An introduction to basic and advanced multilevel modeling. 2nd ed. London, UK: Sage Publications Limited.
Stata Press. 2005. Stata reference manual. College Station, TX: Stata Press.
Subramanian SV, Jones K, Kaddour A, Krieger N. 2009. Revisiting Robinson: the perils of individualistic and ecologic fallacy. International Journal of Epidemiology 38:342–360.

Suen HK, Ary D. 1984. Variables influencing one-zero and instantaneous time sampling outcomes. Primates 25:89–94.
Symonds MRE, Moussalli A. 2011. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behavioral Ecology and Sociobiology 65:13–21.
Te Grotenhuis M, Eisinga R, Subramanian SV. 2011. Robinson's Ecological Correlations and the Behavior of Individuals: methodological corrections. International Journal of Epidemiology 40:1123–1125.
Thorndike EL. 1939. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. The American Journal of Psychology 52:122–124.
Tu Y-K, Gunnell D, Gilthorpe MS. 2008. Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon-the reversal paradox. Emerging Themes in Epidemiology 5:2.
Van de Pol M, Wright J. 2009. A simple method for distinguishing within-versus between-subject effects using mixed models. Animal Behaviour 77:753–758.
Verbeke G, Molenberghs G. 2009. Linear mixed models for longitudinal data. New York, NY: Springer.
Wagner CH. 1982. Simpson's paradox in real life. The American Statistician 36:46–48.
Waller BM, Warmelink L, Liebal K, Micheletta J, Slocombe KE. 2013. Pseudoreplication: a widespread problem in primate communication research. Animal Behaviour 86:483–488.
Wheeler BC. 2010. Production and perception of situationally variable alarm calls in wild tufted capuchin monkeys (Cebus apella nigritus). Behavioral Ecology and Sociobiology 64:989–1000.
Yammarino FJ, Markham SE. 1992. On the application of within and between analysis: Are absence and affect really group-based phenomena? Journal of Applied Psychology 77:168–176.
Yip PSF, Liu KY. 2006. The ecological fallacy and the gender ratio of suicide in China. The British Journal of Psychiatry 189:465–466.
Yule GU. 1900. On the Association of Attributes in Statistics: With Illustrations from the Material of the Childhood Society, &c. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 194:257–319.
Yule GU. 1903. Notes on the theory of association of attributes in statistics. Biometrika 2:121–134.
Zuur A, Ieno EN, Walker N, Saveliev AA, Smith GM. 2009. Mixed effects models and extensions in ecology with R. New York, NY: Springer.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web-site.