

Behavioural syndromes, syndrome deviation and the within- and between-individual components of phenotypic correlations: when reality does not meet statistics

László Zsolt Garamszegi · Gábor Herczeg

Abstract Evolutionary mechanisms leading to correlations across different behaviours, called behavioural syndromes, are hard to study, mostly because behavioural syndromes are group/population level phenomena. Recently (Herczeg and Garamszegi *Behav Ecol Sociobiol* 66:161–169, 2012), we introduced the concept of syndrome deviation that allows the study of behavioural syndromes at the individual level by focusing on the individual deviation from the hypothetical perfect group-level behavioural correlation. Subsequently, Dingemanse et al. (*Behav Ecol Sociobiol* 66:1543–1548, 2012) emphasized that behavioural syndromes refer to the between-individual component of phenotypic correlations, and only this component is relevant for syndrome deviation. They also recommended mixed models to decompose the between- and within-individual correlations. We agree that separating these components is important, but the proposed approach is impractical to apply for functionally different behaviours because (1) the assumption of constant within-individual correlations is

unjustified and (2) different behaviours cannot be measured at the same time. Further, our simulations based on mixed models show that the statistical differentiation between the within- and between-individual components is inefficient when using realistic sample sizes. Until the separation of between- and within-individual correlations is resolved, we recommend alternative approaches for empirical behavioural syndrome research that consider the repeatability of the behaviours and the optimal balance between within- and between-individual sample sizes. Syndrome deviation calculated from phenotypic correlations of traits that are proven to be individual specific, or from the between-individual correlations if possible, is a meaningful metric to describe behavioural consistency and to explain its evolutionary significance.

Keywords Behavioural consistency · Behavioural syndrome · Behavioural type · Personality · Syndrome deviation · Temperament

Introduction

In the last decade, a growing number of studies have reported case evidence for behavioural consistency (for meta-analyses, see Smith and Blumstein 2008; Bell et al. 2009; Garamszegi et al. 2012). Behavioural consistency can be detected at several levels, for instance: (1) consistent individual differences in one behavioural trait across time or contexts; (2) consistent individual differences in the behavioural response to temporal or contextual changes; and (3) consistent individual differences (i.e. having similar ranks) in suites of functionally different behaviours (e.g. Dingemanse and Wolf 2010). Although these phenomena are often treated—rather confusingly—as synonymous (Dall et al. 2004; Réale et al. 2007; Dingemanse et al.

László Zsolt Garamszegi and Gábor Herczeg contributed equally to this paper.

L. Z. Garamszegi
Department of Evolutionary Ecology, Estación Biológica de Doñana–CSIC,
c/Americo Vespucio, s/n,
41092 Seville, Spain

G. Herczeg (✉)
Behavioural Ecology Group, Department of Systematic Zoology and Ecology, Eötvös Loránd University,
Pázmány Péter sétány 1/c,
1117 Budapest, Hungary
e-mail: gaborherczeg@caesar.elte.hu

2010; Sih et al. 2012), we distinctly classify forms of consistency in single behaviours (1) as animal personality and consistency in two or more functionally different behaviours (3) as behavioural syndrome (see Garamszegi et al. 2012), while we treat consistent behavioural responses (2) as a form of plasticity. The mainstream approach in studying behavioural consistency is to test for the repeatability of single behaviours to detect animal personality or for the correlation between multiple behaviours to detect behavioural syndromes. Upon obtaining significant results (i.e. proving behavioural consistency), researchers would go back and analyse single behaviours or the behavioural configuration (behavioural type, Bell 2007). However, if behavioural consistency (manifest as animal personality or behavioural syndrome) itself is supposed to be under selection, we should also study it directly, and thus we need a variable describing variation in consistency irrespective of the actual behavioural type. Further, as behavioural syndromes are group level traits (i.e. correlations), we cannot directly study their evolution and thus need individual level variables for this purpose (Herczeg and Garamszegi 2012). To solve these problems, we introduced a new concept, *syndrome deviation*, which is the individual deviation from the hypothetical perfect behavioural syndrome (Herczeg and Garamszegi 2012). We also provided two purposefully simple mathematical solutions for this concept.

Recently, Dingemanse et al. (2012) challenged our proposal by (1) highlighting the potential importance of the decomposition of within- and between-individual correlations when drawing inferences about behavioural syndromes from phenotypic correlations, (2) recommending the use of mixed models based on repeated measures of the same individuals for understanding the evolution of behavioural syndromes and (3) proposing an alternative mathematical solution for syndrome deviation. We welcome this commentary, as it leads to the clarification of issues with great theoretical and practical importance, and we were actually hoping that our new concept would somehow ‘stir the pot’ in behavioural syndrome research. Accordingly, we address the above main issues with the aim of advancing the field based on constructive discussion.

The decomposition of phenotypic correlations: theoretical implications for behavioural syndromes

Dingemanse et al. (2012) main concern with the way we approached syndrome deviation is that it would suffer from the confounding effect of within-individual correlations if it was based on phenotypic correlations. The authors derive a statistical formula showing that phenotypic correlations are composed of two components: the between- and within-individual correlations. The between-individual component

can be interpreted as being based on stable individual differences, while the within-individual component reflects that the traits are unstable within an individual, but their changes are not independent of each other. Dingemanse et al. (2012) suggest that only the between-individual correlation is relevant for behavioural syndromes, while within-individual correlation should reflect state-dependence (e.g. hunger level) of trait expression within an individual and can be seen as a confounding effect.

The importance of the separation of within-subject and between-subject effects is not new, but is a well-defined issue in the social and evolutionary disciplines (Davis et al. 1961; Kreft et al. 1995; Ives et al. 2007; Felsenstein 2008), with statistical solutions being available through mixed-effect or repeated-measure modelling (Laird et al. 1987; 1999; Snijders and Bosker 1999; Roy 2006). Moreover, the multilevel aggregation of behavioural data (i.e. within- and between-individual variance and covariance) has also been recognised, yet often remains inappropriately treated at the level of analysis (van de Pol and Wright 2009). Therefore, from the statistical point of view, Dingemanse and his co-authors correctly point to the problem posed by the multi-level aggregation of behavioural data for the study of behavioural syndromes. Correlations observed at the between-individual level might have different biological meaning than correlations observed at the within-individual level. The separation between these is important because it allows us to distinguish between alternative biological hypotheses, while inferences based on phenotypic correlations between behaviours measured only once incur the risk of erroneously attributing within-individual relationships to between-individual relationships or *vice versa*.

Yet, we infer that the treatment of the within-individual effect merely as a confounder is misleading. Within-individual correlations, albeit unrecognised, might represent phenomena that may be relevant for the study of behavioural consistency. For instance, if there was between-individual variation in how consistently different behaviours change within individuals (e.g. individuals may respond differently to hunger or stress levels, and demonstrate different degree of behavioural associations in different contexts), and this variation was heritable, than there is full potential for the emergence of behavioural correlations as a result of adaptive evolution affecting within-individual correlations alone. Behavioural correlations based on the within-individual component might be seen as another form of behavioural consistency (i.e. plasticity) and might beg for evolutionary answers.

In our paper (Herczeg and Garamszegi 2012), we defined behavioural syndrome as a ‘correlation between rank-order differences between individuals’ following Bell (2007). This definition implicitly implies that individuals can be discriminated and ranked based on their behaviours and that

individual ranks are stable, while we have not stated that behavioural syndromes are phenotypic correlations. This confusion may have arisen because we did not go into mathematical details regarding between- and individual-correlations, as it was not the goal of our paper. We aimed to propose an approach where behavioural consistency itself, as an individual-specific trait, is quantified and analysed besides behavioural type to better understand the evolution of behavioural consistency. We indeed welcome any working approach for separating the between- and within-individual correlations and emphasize that whenever such approach becomes available (but see our reservations below), our syndrome deviation concept can be easily applied for the pure between-individual correlations. Furthermore, the same concept can be extended for studying behavioural consistency in single behavioural traits, i.e. studying the evolution of animal personality where only the between-individual component is relevant.

The use of mixed models for separating different components

Dingemanse et al. (2012) recommend an approach in which one collects multiple behavioural data per individual for each trait via repeated trials and then applies mixed models. Although the decomposition of phenotypic correlations into within- and between-individual components via such repeated-measurement design seems promising, we have both conceptual and practical concerns regarding the usefulness of this approach specifically in the field of behavioural syndromes.

Concept

To claim their point, Dingemanse et al. (2012) derived an equation from Snijders and Bosker (1999), who published an equation that separates between-group and within-group correlations (see also Garamszegi et al. 2013). Snijders and Bosker (1999, page 31) assume that the within-group correlation is the same in each group. Therefore, the application of Eq. 1 from Dingemanse et al. (2012) to study behavioural syndromes will also assume that within-individual correlations are constant across individuals. However, as we argued above, individuals may be well expected to show differences in how they vary their behavioural profiles along different contexts, which can potentially result in between-individual differences in within-individual correlations. Such consistent individual variation in the within-individual component requires further development of Eq. 1 of Dingemanse et al. (2012). For separating between- and within-individual correlations properly, one may need to apply random-slope mixed models, which is a data hungry exercise as it necessitates

several repeats within subjects to capture individual-specific slopes. Our simulations that we detail below show that differentiating between phenotypic and between-individual correlations when assuming constant within-individual correlations demands large sample sizes already. The picture gets even more unrealistic when slopes within individual are also to be estimated. However, the assumption about constant within-individual correlation does not only call for practical issues, but also directly exaggerates the problem outlined below.

The calculation of Eq. 1 of Dingemanse et al. (2012) and the application of mixed models proposed by them also assume that the two (or more) traits can be measured simultaneously or at the very same individual state. However, this is rarely the case for a suite of behaviours (see, e.g. Garamszegi et al. 2012 showing that the time interval between measurements can vary up to several months). In fact, if behaviours are defined based on their functions and the corresponding ecological context (Réale et al. 2007), individuals can perform only a single behaviour in any given moment, e.g. they cannot explore an environment, fight against a territory intruder and show antipredatory behaviour at the very same time. In practice, researchers measure functionally different behaviours in separate trials. Dingemanse et al. (2012) point to Wilson et al. (2011) as the example study applying a mixed model, but this study was able to estimate within-individual correlations between different behavioural variables that actually describe the same behavioural phenomenon (aggression) and were measured within the same trial (e.g. latency to attack, number of attacks and number of flees).

The non-simultaneous measurement of traits imposes a fallacy for the separation of between- and within-individual correlations for the following reason. Let us assume that an observer scores two functionally different behaviours based on the repeated measures design as suggested by Dingemanse et al. (2012). For example, s/he measures behaviours b_1 and b_2 on two occasions. However, as the two behaviours cannot be measured at the very same time, if b_1 is assayed at t_1 and t_2 , b_2 can only be tested assayed at t'_1 and t'_2 , such that $t_1 \neq t'_1$ and $t_2 \neq t'_2$. Even if $|t_1 - t_2|$ or $|t'_1 - t'_2| \gg |t_1 - t'_1|$ or $|t_2 - t'_2|$, and the difference between t and t' occasions seems small, such differences can be important for behaviours. Behavioural traits can change considerably within very short time, even from one moment to another (for example, singing in male collared flycatchers, *Ficedula albicollis*; Garamszegi et al. 2007). Therefore, it is possible that the measured b_2 trait would be completely different in t and t' occasions. However, when the investigator calculates within-individual correlation based on the observed variance between t_1 and t_2 (or t'_1 and t'_2), s/he would inherently assume that the estimate of b_2 at t' occasions reflects what could have been estimated on t occasions (otherwise b_1 and b_2 estimates would not correspond to each other), and thus that s/he is supposed to deal with individual-specific stable trait values and neglect within-individual

correlations within the $t-t'$ interval. This would be theoretically incompatible with Dingemans et al. (2012) aiming to estimate between-individual correlations while controlling for the within-individual components. Variations within the $t-t'$ interval can represent not only errors but also biases, if such variations occur at an individual-specific manner and are also mediated by properties that belong to behavioural consistency.

Furthermore, the measurement process of b_1 at t_1 can directly affect b_2 measured at t_1' if $|t_1-t_1'|$ is small (e.g. an exposure to an aggressive conspecific or predator at t_1 will affect behaviour at t_1' if the animal becomes stressed) making a considerable pause between t_1 and t_1' necessary for re-acclimation. Such a problem due to contextual overlap calls for increasing the time lag between measuring the different behaviours, while an attempt to deal with the previous problem of neglecting within-individual variation between subsequent measurements of different behaviours would require extremely short time lags. Hence, addressing these two problems in parallel is challenging if not impossible. Note that these issues are irrelevant when stable morphological traits are correlated with each other (i.e. traits within the t and t' interval can be stable), or when higher-level within-group and between-group correlations (e.g. within- and between-

school) are considered as originally exemplified by Snijders and Bosker (1999). Hence, the problem posed by the temporal arrangements of measurements is specific to the study of behavioural syndromes.

Practice

For a powerful statistical discrimination between within- and between-individual correlations, an appropriate sampling is required at both levels, which might also set up practical constraints for the study of behavioural syndromes. To investigate the effect of sample size on the effectiveness of the combined use of repeated-measurement design and the mixed model approach to separate within- and between-individual components, we performed a simulation (Fig. 1). As a start, we modelled a realistic situation when a researcher works with behaviours that have moderate repeatability ($R=0.5$, see Bell et al. 2009 and Garamszegi et al. 2012) and show modest phenotypic correlations ($r_p=0.3$; see Garamszegi et al. 2012). According to Eq. 1 of Dingemans et al., this can occur if, for example, the between-individual correlation is $r_{ind}=0.5$ and the within-individual correlation is $r_e=0.1$, which may imply that even a weak within-individual correlation can inflate the interpretations one

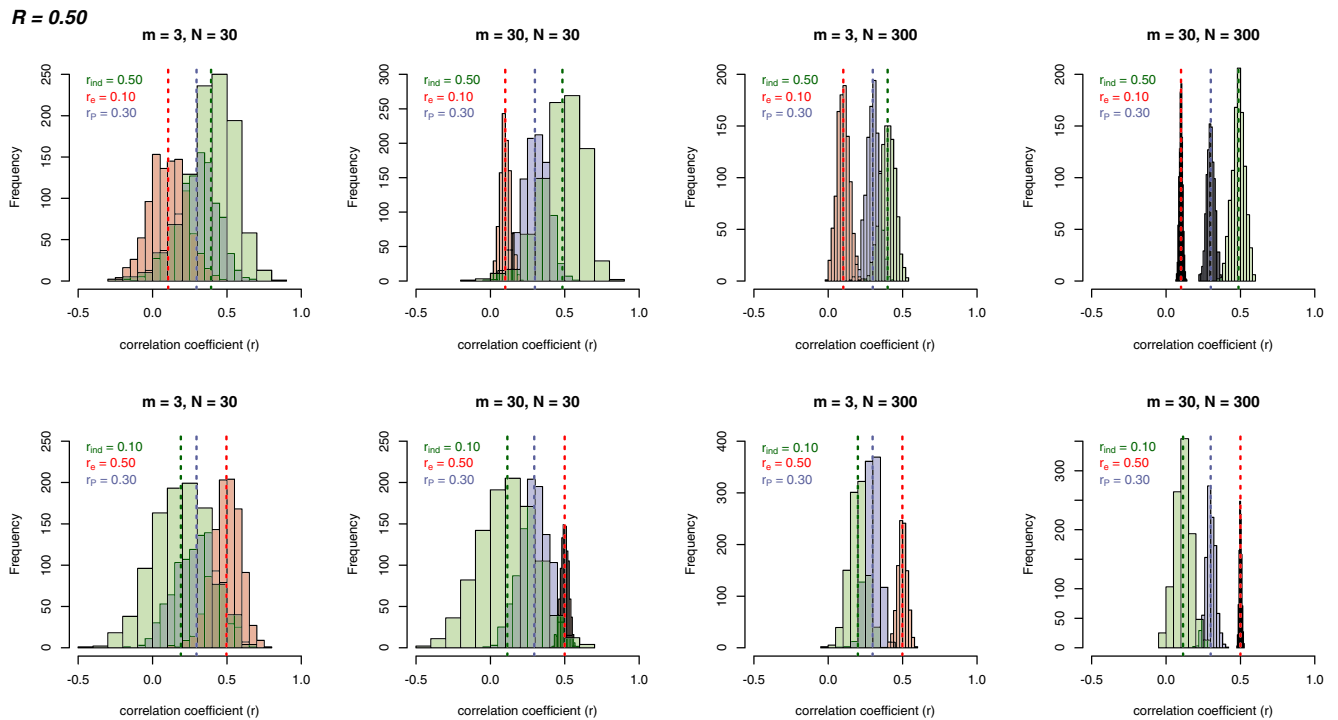


Fig. 1 Histograms showing estimates of different correlations (*green* between-individual correlation, r_{ind} ; *red* within-individual correlation, r_e ; *blue* phenotypic correlations, r_p) between two behavioural traits that have been simulated 1000 times with true correlations of $r_{ind}=0.5$ and $r_e=0.1$ (*upper panels*) and of $r_{ind}=0.1$ and $r_e=0.5$ (*lower panels*) at a repeatability of $R=0.5$ under different sample size scenarios. The expected phenotypic correlation (r_p) is calculated based on Eq. 1 of

Dingemans et al. (2012). Particular estimates of between- and within-individual correlations originate from mixed models based on within-subject centring and by using individuals as random factors. Phenotypic correlations are approximated by randomly taking one measurement from each individual. *Dashed lines* are the corresponding means of the distributions. On the printed version, the colours may appear as different shades of grey (r_{ind} : light grey; r_p : medium grey; r_e : dark grey)

would make from phenotypic correlations. As a contrast, we choose an opposite setup, where $r_{\text{ind}}=0.1$ and $r_e=0.5$ to model the situation when the within-individual component is considerable and introduces an upward bias. We also used a parameterization based on $r_{\text{ind}}=0.3$ and $r_e=0.3$.

We commenced with a reasonable sampling scheme, in which the investigator is allowed to take 90 measurements per behaviour, i.e. to assess the two behaviours in $N=30$ individuals at $m=3$ occasions. Subsequently, we augmented these sample sizes to $N=300$ and $m=30$ sequentially, resulting in four sample size combinations (Fig. 1). Based on these parameter setups, we simulated data in the following fashion. First, we created N individual-specific values from a normal distribution with a zero mean and unit variance for the two variables that were forced to take the specified between-individual correlation (r_{ind}). Then, around each individual-specific datum as an individual mean, we simulated m within-individual measurements with $(1-R)/R$ variance and with a correlation structure between behaviours as specified by the within-individual correlation (r_e). Such a procedure was performed 1000 times within a set, and was repeated at different sample size, repeatability ($R=0.2$ and $R=0.8$ in addition to $R=0.5$) and correlation scenarios.

To analyse the simulated data, we used mixed models relying on within-subject centring to separate between slopes that correspond to the within-individual and between-individual levels (Snijders and Bosker 1999; van de Pol and Wright 2009). Such an approach requires that one behavioural variable is arbitrarily handled as a predictor and the other as a response, while such discrimination is not warranted in association with behavioural syndromes that deal with correlations. However, we repeated each procedure in both possible combinations of the two behaviours being predictor or response and the results were basically identical from the parallel runs. We included ‘individual’ as a random effect term to capture the hierarchical structure of the data. From each model, we extracted the between- and within-individual slope estimates, and based on the corresponding t and df values we converted them into effect sizes in the form of Pearson correlation coefficient (Nakagawa and Cuthill 2007). We also used alternative approaches to derive the within- and between-individual correlations (such as based on the correlation of the Best Linear Unbiased Predictors and residuals from models that included a random effect and an intercept only), and these also gave very similar results to those we report here (data not shown). Finally, we estimated phenotypic correlations from the generated data by randomly taking a single measurement of each behaviour from each individual. The simulation and the mixed-effect modelling were performed in the R statistical environment (R Development Core Team 2007).

Given the *statistically* modest sample sizes and the sampling variance, we expected that the estimated parameters would show some deviations across the 1000 repeats. We

inspected the frequency distributions of different assessments of correlations through each simulation set and compared them across different parameter scenarios. The results that were based on $r_{\text{ind}}=0.5$ and $r_e=0.1$ as well as on $r_{\text{ind}}=0.1$ and $r_e=0.5$ while considering $R=0.5$ are presented in Fig. 1 (upper and lower panel, respectively). The outputs from other scenarios are given in the Appendix (Figs. S1–3).

The simulations generally suggest that using realistic sample sizes ($N=30$, $m=3$) the mixed effect modelling is rather inefficient in discriminating between-individual correlations from phenotypic correlations, as their distribution is largely overlap. Due to the sampling variance, there are fair chances to detect a broad range of estimates around the true between-individual correlation even when mixed-models are applied on multiple measurements. Increasing sample size in terms of the number of individuals results in narrowed intervals, but at a low within-individual sample size the estimate remains a poor approximation of reality, and phenotypic and between-individuals correlations are still hard to distinguish. The simulated data can reproduce the specified correlations with small errors at extreme sample sizes only ($N=300$, $m=30$ signifying 9000 measurements on two behaviours). When repeatability is high (see Appendix), the effect of within-individual sample size is not that pronounced, as estimates from small within-individual samples are more reliable. However, between-individual and phenotypic correlations cannot be well discriminated, because r_{ind} approximates r_P (while r_e can be neglected) when repeatability is high (Eq. 1 of Dingemans et al. 2012). In the opposite situation, when repeatability is low, the within-individual component dominates the phenotypic correlation. In such case, it is more important to have large within-individual sample sizes. Note that in the above simulations, we systematically assumed that the within-individual components are constant across individuals. However, if the within-individual component is also allowed to vary to be realistic, and thus random-slope mixed models are to be used, we suspect that we would detect an even stronger role for sample sizes affecting the appropriate estimation of within-individual effects.

Our recommendations

The simulations unanimously indicate that when working with sample sizes and within-individual replicates with which behavioural ecologists operate, it is almost impossible to obtain considerable improvement from mixed models over approaches based on phenotypic correlations if one aims to estimate the between-individual correlation. Although the separation of different correlations is important statistically, practical limitations lead us to come up with the following recommendations.

First, it might be useful to perform a pilot study based on a fewer number of individuals, on which the behaviours are

assessed repeatedly. Such a study would focus on within-individual variations, and thus it would require a balanced within-individual sampling at a modest between-individual sample size that allows the estimation of repeatability of traits and also the within-individual correlation (see also Harmon and Losos 2005 for an analogous problem). A subsequent study could focus on between-individual variations and could be designed according to the findings of the pilot study. If repeatability is high or within-individual correlation is generally low, given the mathematical equation describing the link between different components, more effort can be invested in the collection of data across individuals rather than continuing sampling within individuals. However, if repeatability is modest and the covariance of traits within individuals is not negligible (especially if there is an evidence for individually varying within-individual effects), it remains important to collect multiple data from the same individuals. In this latter case, improving the sample size in terms of the number of individuals at the cost of lowering within-individual sample size would only result in biased estimates even in cases when mixed models are used to handle few repeats within individuals.

If the above design is not feasible due to logistic constraints embedded in the study system, one is left with a more general approach. Namely, first estimate repeatability of the studied behaviours, then work with individual-specific estimates of traits and subsequently assume that the detected phenotypic correlations will represent between-individual correlations. Although the use of such ‘phenotypic correlation of individual

means’ may not offer a statistically perfect solution to the elimination of the within-individual component (see Appendix of Dingemanse et al. 2012), it may still be the best proxy of between-individual correlations if (1) the repeatability of the behavioural traits within individuals is established or (2) within-individual correlation is shown to be negligible.

In general, we know very little about how within-individual correlations inflate phenotypic correlations in nature. In our meta-analysis (Garamszegi et al. 2012), we have shown that the strength of the detected phenotypic correlation is positively related to the geometric mean of the repeatability of traits. According to Eq. 1 in Dingemanse et al. (2012), such a positive relationship can occur under certain assumptions if the between-individual component is dominant, which may provide some empirical justification for neglecting the within-individual component. However, to appropriately estimate the error that the ignorance of the within-individual component causes, more empirical studies are warranted.

Estimating syndrome deviation

Let us assume that repeatability estimates support that individuals can be consistently ranked along the same behaviour (i.e. the most aggressive individual is the most aggressive at each consecutive trials), thus they can be discriminated based on their average trait expression. We measure syndrome deviation as the inconsistencies of these ranks within the same individual

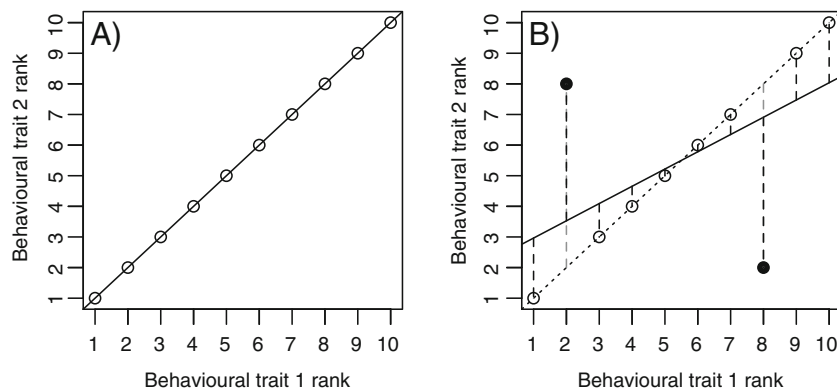


Fig. 2 **a** The correlation between the ranks of two behavioural traits, when each individual has the same rank in both behaviours, as it would be expected from the null hypothesis of perfect behavioural syndrome. **b** The correlation between the ranks of two behavioural traits, when each individual has the same rank in both behaviours except two individuals that changed ranks along the second behaviour (filled circles). Dashed lines show syndrome deviation as could be calculated from the *observed* (black dashes, as suggested by Dingemanse et al. 2012) correlation and from the *expected* correlation based on perfect syndrome structure (grey dashes, as suggested by Herczeg and Garamszegi 2012). Individuals that are ranked the same way along the two behaviours (open circles), in theory, do not deviate from the syndrome, but receive different syndrome deviation values, if these are calculated based on the *observed* correlation (solid black line). In fact,

choosing the *observed* correlation as reference creates a rank-dependent bias because very low or high ranked individuals receive higher false deviation scores than individuals with middle ranks. Note that the *observed* correlation always depends on the individuals at hand, thus deviation scores of the ‘non-deviant’ individuals are always conditional to the proportion of individuals that deviate from the syndrome structure with different degree. However, when using the *expected* correlation as a reference (dotted black line), only the deviant individuals receive non-zero scores. Lines are the regression lines and shown merely for illustrative purposes. When the correlation is based on ranks that have the same range (e.g. 1–10), $s_{\text{trait1}} = s_{\text{trait2}}$ that yields $\beta = r$ given that $r = \beta \times s_{\text{trait2}} / s_{\text{trait1}}$, where s is the standard deviation, β is the slope, and r is the correlation coefficient

across different behaviours (Herczeg and Garamszegi 2012). Therefore, such an approach naturally targets the between-individual component of the correlation, and we continue our discussion based on this scenario.

We wholeheartedly agree that the statistical tools to estimate syndrome deviation can/should be improved, and in fact welcome such progress. Dingemanse et al. (2012) also suggest a possible solution for our syndrome deviation concept. However, they emphasize using the *observed* correlation as the baseline for the calculation of syndrome deviation instead of the *expected* perfect rank correlation ($r_s=1$) advocated by us, labelling the latter as arbitrary. We think that this would lead to misinterpretations. First, in a perfect rank correlation (rank correlation is used to define behavioural syndromes, see Bell 2007) represented by $r_s=1$ (or -1) the slope of a fitted regression line can also be only 1 (or -1), unlike parametric correlations where in the case of $r=1$ (or -1) the slope could be anything but zero (i.e. in a rank correlation only the slope of 1 or -1 represent cases where all data points are on the straight line, while in a parametric correlation any slope but 0 could represent cases with all points in a straight line). So using the perfect rank correlation ($r_s=1$) as the null model of a perfect behavioural syndrome is a logical choice when behaviours are ranked and brought into the same scale (Herczeg and Garamszegi 2012), as it represents the null scenario that individuals are ranked in the same way along both behavioural axes (Fig. 2a). On the other hand, using the *observed* correlation of ranks as a null would be biased, because the *observed* correlation is already a result of individuals differing in their consistency (besides measurement error), so identifying individual deviation based on this correlation would be practically meaningless (Fig. 2b). Taken together, the baseline (null model) for the calculations should be the theoretical perfect syndrome and not the *observed* one, and only in rank correlations do we know how perfect syndromes would look.

Summary

We anticipate the theoretical suggestions of Dingemanse et al. (2012) about the importance of different components of phenotypic correlations. We infer that shortcomings in association with the use of phenotypic correlations for making inferences for behavioural syndromes do not directly undermine our concept about syndrome deviation (Herczeg and Garamszegi 2012), because it can be easily applied to between-individual correlations. We find that the emerging discussion is useful, because it points to a neglected issue, and the whole field of behavioural syndrome research would benefit from differentiating between the different components of phenotypic behavioural correlations, which might be one of the main goals in the future. However, before dismissing field studies based on phenotypic correlations (i.e. most studies), we

recommend that more consensus and research are needed to appropriately deal with the problem. Such a scientific progress should consider the constraints of the study of animal behaviour in terms of sample size and also address assumptions about variation in within-individual correlations and the temporal/contextual overlap between behavioural measurements. Until the problems with separating between- and within-individual behavioural correlations are convincingly solved, we suggest focusing on the phenotypic correlations of individual-specific estimates of traits (or their ranks) together with the repeatability of behaviours, or if feasible, running a pilot study to determine the balance between sample size within and between individuals. The field would also benefit from more empirical evidence about the magnitude by which within-individual correlations affect the phenotypic correlation, and about the biological significance of within-individual correlations and their potential variation among individuals. If a potential behavioural correlation (phenotypic or pure between-individual) is biologically meaningful, the calculation of syndrome deviation becomes meaningful as well (and even makes perfect sense in the absence of a significant correlation) and appropriate evolutionary questions and research designs can be tailored. We think that by adapting the concept of syndrome deviation, studying the evolution (heritability, fitness consequences, genetic background) of behavioural consistency observed at different levels became easier, since this approach provides an individual-level trait that can be subject to selection.

Acknowledgements LZG was supported by the “Plan Nacional” program of the Spanish government (ref. nos. CGL2009-10652 and CGL2009-09439). GH received financial support from the Hungarian Scientific Research Fund (OTKA-K 105517) and was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. We are highly indebted to Mike Webster for his comments and improving the English language. We are grateful to Roger Mundry and Holger Schielzeth for stimulating discussions. We thank MANU-STAT Central for statistical consultation. Comments from two anonymous reviewers also helped to improve the paper.

References

- Bell AM (2007) Future directions in behavioural syndromes research. *Proc R Soc Lond B* 274:755–761
- Bell AM, Hankison SJ, Laskowski KL (2009) The repeatability of behaviour: a meta-analysis. *Anim Behav* 77:771–783
- Dall SRX, Houston AI, McNamara JM (2004) The behavioural ecology of personality: consistent individual differences from an adaptive perspective. *Ecol Lett* 7:734–739
- Davis JA, Spaeth JL, Huson C (1961) A technique for analyzing effects of group composition. *Am Sociol Rev* 26:215–255
- Dingemanse NJ, Wolf M (2010) A review of recent models for adaptive personality differences. *Philos T Roy Soc B* 365:3947–3958
- Dingemanse NJ, Kazem AJN, Reale D, Wright J (2010) Behavioural reaction norms: animal personality meets individual plasticity. *Trends Ecol Evol* 25:81–89

- Dingemanse NJ, Dochtermann NA, Nakagawa S (2012) Defining behavioural syndromes and the role of “syndrome deviation” in understanding their evolution. *Behav Ecol Sociobiol* 66 (11):1543–1548
- Felsenstein J (2008) Comparative methods with sampling error and within-species variation: contrasts revisited and revised. *Am Nat* 171:713–725
- Garamszegi LZ, Török J, Hegyi G, Szöllösi E, Rosivall B, Eens M (2007) Age-dependent expression of song in the collared flycatcher, *Ficedula albicollis*. *Ethology* 113:246–256
- Garamszegi LZ, Markó G, Herczeg G (2012) A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. *Evol Ecol* 26:1213–1235
- Garamszegi LZ, Markó G, Herczeg G (2013) Erratum to: A meta-analysis of correlated behaviours with implications for behavioural syndromes: mean effect size, publication bias, phylogenetic effects and the role of mediator variables. *Evol Ecol* (in press)
- Harmon LJ, Losos JB (2005) The effect of intraspecific sample size on type I and type II error rates in comparative studies. *Evolution* 59:2705–2710
- Herczeg G, Garamszegi LZ (2012) Individual deviation from behavioural correlations: a simple approach to study the evolution of behavioural syndromes. *Behav Ecol Sociobiol* 66:161–169
- Ives AR, Midford PE, Garland T (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Syst Biol* 56:252–270
- Kreft IGG, de Leeuw J, Aiken LS (1995) The effect of different forms of centering in hierarchical linear models. *Multivar Behav Res* 30:1–21
- Laird NM, Lange N, Stram DO (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. *J Am Stat Assoc* 82:97–105
- Lam M, Webb KA, O’Donnell DE (1999) Correlation between two variables in repeated measures. *Joint Statistical Meeting (JSM)*, Baltimore, pp 213–218
- Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev* 82:591–605
- R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org>
- Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ (2007) Integrating animal temperament within ecology and evolution. *Biol Rev* 82:291–318
- Roy A (2006) Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biom J* 48:286–301
- Sih A, Cote J, Evans M, Fogarty S, Pruitt J (2012) Ecological implications of behavioural syndromes. *Ecol Lett* 15:278–289
- Smith BR, Blumstein DT (2008) Fitness consequences of personality: a meta-analysis. *Behav Ecol* 19:448–455
- Snijders TAB, Bosker RJ (1999) Multilevel analysis—an introduction to basic and advanced multilevel modelling. Sage, London
- van de Pol MV, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. *Anim Behav* 77:753–758
- Wilson AJ, de Boer M, Arnott G, Grimmer A (2011) Integrating animal personality research and animal context theory: aggressiveness in the green swordtail *Xiphophorus helleri*. *PLoS One* 6: e28024