THE RING 26, 1 (2004) DOI 10.2478/v10050-008-0058-2

REMARKS ON THE TAKING AND RECORDING OF BIOMETRIC MEASUREMENTS IN BIRD RINGING

John Howard Morgan

ABSTRACT

Morgan J.H. 2004. *Remarks on the taking and recording of biometric measurements in bird ringing.* Ring 26, 1: 71-78.

Ringing operations hold opportunities for introducing error into biometric recording. This situation needs to be addressed by field workers, data processors and archivists. Avoidable error may be systematic and/or random, and adds "noise" to random error from natural variation. Handling techniques and measuring equipment are responsible for introducing systematic errors in fieldwork. This aspect requires an increased level of professionalism among ringers to correct it. Analysis of data can induce further random error, *e.g.* when generating indices from measurements. Analysts also need to be aware of pitfalls inherent in field data, especially that collected historically.

J.H. Morgan, Chemin de Laval, Cabrespine, F-11160, France,
E-mail: john_howard_morgan@yahoo.co.uk

Key words: data analysis, error, morphometrics (biometrics), ringing equipment, ringing technique

INTRODUCTION

Technical papers incorporating morphometric measurements (biometrics) taken during bird ringing appear in journals with increasing frequency. A cursory glance at the last two volumes of *Ringing and Migration* suggests the proportion of papers to be approaching 40-50%; the biometrics they contain are either simply descriptive or part of a detailed analysis. Biometric studies of natural variation certainly help answer both theoretical and applied questions in avian ecology.

All biometrics are intrinsically prone to error. Those derived from bird ringing are now being subjected to computer archiving to await future analysis and publication by third parties. In all likelihood the latter will have no first hand knowledge of how the data was collected and the extent to which it holds measurement errors – *vide* the rebuttal of Maitav and Itzaki (1992) in Morgan and Shirihai (1997). To guide both those who gather the measurements and those who utilise them, a wide-ranging review of measurement techniques is needed.

DISCUSSION

Error will always occur in the gathering of ringing biometrics, entering the data at each of the taking, recording and archiving stages of the overall operation. When the measurements are bulked up for (statistical) analysis, errors will give additional variability – so-called "noise" – on the top of that occurring naturally. Because of the impossibility of separating the two after they have become compounded, noise will obscure natural variation and prevent its correct assessment.

Systematic *versus* random error

When a bird caught for ringing is measured, the errors arising can be either systematic, random, or both. The former – caused by the measurer, his tools and equipment – are repeated to the same extent each time the measurement is made; the latter happen through circumstances often outside the control of the investigator. Random errors are generally assumed to be "normally" (Gauss) distributed, varying above and below an average value which will be zero just as long as there is no systematic component. *Nota bene*, discussion of "explained" errors is not undertaken here.

An example of systematic error, very familiar to ringing trainers, is the consistently "short" maximum-chord wing length measured by trainee ringers. This error is caused by the need to acquire a degree of dexterity in straightening the wing. Its correction and elimination during training cannot be hurried.

Less obvious systematic errors arise if trainees attempt to accelerate the process (of standardising with their trainer) by trying to correct the shortfall. They can do this either by "fudging" their reading of the scale (*e.g.* always rounding up fractions to try to compensate for the amount assumed to be lacking) or by incorrectly positioning the bird relative to the ruler *e.g.* rotating the wing away from the line of the body to artificially produce a longer measurement as the carpal joint flexes. Unless corrected, such practice(s) will continue after their training is complete.

Authors frequently caution readers against uncritical application of biometric results presented in their papers, recognising the potential for systematic error. A classic example where this can happen is when using discriminatory indices or related procedures (see additional discussion below) that claim to allow separation of two sibling species, or of the sexes within a species. They are going to be almost valueless if derived from, or are used on, measurements subjected to systematic error, or if there are no estimates of upper and lower bounds to allow for random error.

For a fairly basic example, say one measures the wings of Cetti's Warblers (*Cettia cetti*) or Corn Buntings (*Milaria calandra*) by rotating the wing outwards and rounding up to the next millimetre. Follow this by weighing on a much used and carelessly-held spring balance that has never been calibrated. Sooner or later a female will be sexed as male when checked against the ageing and sexing guide *i.e.* Svensson (1992).

For a parallel example involving random error, consider changes in feather state, handling conditions, *etc.* brought about by variation in environmental humidity and temperature, and how this in turn might affect the maximum-chord wing length. The haphazard nature of meteorological conditions during a ringing season will be reflected as random variation (albeit tiny) in the wing chord measurement of a given bird. *Nota bene*, some rather larger changes in feather length, from abrasion and moult, can be "explained" by the date of capture and the age of the bird.

Errors in practice

What is the best approach when dealing with error? Because of the identifiable ways in which systematic errors arise, they can usually be eliminated. At worst some might be "pseudo-randomised" by using samples assembled with measurements from multiple sources. As for the other side of the picture, although random error can often be reduced, some noise from this source is inevitable. Overall, what is important is to be aware of the ease of propagation of error and to plan the gathering of measurements accordingly.

For an example of how one might approach the problem, consider the spring balance used to measure body mass. Initially, one can recognise the desirability of hanging a spring balance from a fixed point whenever possible, and shielding it from the wind. This should reduce random error arising from inaccuracies reading the scale caused by slight movements of the pointer.

Less obviously, failure to keep the balance absolutely vertical may cause internal parts to stick inside the barrel and lead to large, irregular errors. It is essential to allow a hand-held balance to find the vertical by holding it with the ring provided for this purpose. Another poorly-observed precaution is minimization of parallax by viewing the scale pointer at right angles. Systematic error results if the balance is habitually read from another position.

Finally it should be noted that all balances eventually show systematic error due to either sudden or gradual mechanical change. A balance needs to be regularly checked against standard mass(es), at the very least before starting a ringing session and at intervals during a long session too.

Exhaustively cataloguing the sources of error in this way could be done similarly for other measuring devices and techniques. It is then necessary to make sure the information is available for those who gather biometrics.

Error management

It would seem useful to try to set proper standards for taking ringing biometrics that will allow comparison and cross-application of research results. Lacking any satisfactory protocol, a result based on biometrics from more than one source becomes suspect. The actual degree to which replicated measurements in passerines hold errors due to bias, disagreement and imprecision has been thoroughly investigated by Gosler *et al.* (1995) and Gosler *et al.* (1998).

Though noise from random error is a fact of life, systematic error can make data containing it downright misleading. Working on Reed Warbler (*Acrocephalus scirpaceus*) body mass data gathered by multiple ringers at several (mainly UK) locations, the author found it impossible to tell whether or not one site had breeding birds averaging 0.5 g heavier than elsewhere, because there was a more parsimonious explanation – balance error (unpubl. data).

The common bird ringing biometrics are wing length and body mass, although there is increasing interest in having other measurements more widely taken. Many years ago the author heard even these two described as "taking gratuitous measurements", but the utility of these biometrics is now beyond doubt.

Systematic errors made by ringing trainees taking wing chord measurement has already been referred to, but it is also well documented that repeatability (reproducibility) of wing length even between experienced measurers is lower than for an individual measurer (Harper 1994). Each may be measuring shorter or longer as a matter of individual technique, but wings and the birds that carry them also differ in more obscure ways than size alone, *e.g.* flexibility, manageability, *etc.* These could interact differently with different measurers depending on the latter's dexterity. Indeed, repeatability could vary with species, both within and between measurers. It is likely it will depend also on the biometric being taken, *e.g.* between third primary and wing length.

Even when error from causes such as those above has been minimised there still remain problems of observation for the scale reading on the measuring device. One way to improve intra-observer repeatability, leading to reduced measurement error variability in data, could be to replicate readings. This would be particularly efficient when using a vernier scale such as that found on callipers, because making the measurement has to be completed in advance of reading the scale, and knowledge of prior readings cannot then affect the outcome of the current one. The mode of a set of three readings, or else the median if there is no mode, would be recorded; a very disparate reading in the set should prompt further re-measurement. Alternatively, all three readings can be recorded and used to estimate the actual extent of measurement error, eventually to be expressed as a fraction of the total variability as by Lougheed *et al.* (1991)

Ideas for measurement protocol aimed at error reduction, such as those above, need to be freely generated and widely discussed, so that universal standards can be implemented. Above all, it should be widely accepted that major improvement in error reduction will follow improvement in training.

Errors from measurement devices

The discussion so far has addressed only part of the problem. Equipment can be as much at fault as its user in generating error.

When using a stopped ruler to measure wing chords, primary tips seldom fall exactly on an engraved scale marking. Observers must arbitrarily decide on whether

or not a particular wing should round up or round down, or if a fractional amount should be recorded. Furthermore, wing and pin rulers supplied to ringers often have a "stop" that does not coincide with the scale's zero point. A gap of as much as 0.3 mm has been found (unpubl. data) so that two measurers using different rulers may have their measurements differ. Both of these potential sources of error arise from the way the ruler is made, reflecting the fact that these are relatively *ad hoc* devices and not dedicated designs.

Vernier callipers can also have defects in zeroing, and which may differ between the three functions that they can perform. The part of the device designed to measure internal diameter is particularly difficult to calibrate. Because callipers are mainly used to a precision of 0.1 mm, any zero error should ideally be less than half the difference between successive rounded digits (less than 0.005 mm, not discernable by eye).

Obtaining body mass by spring balance is an application of an existing tool, and although one can identify fewer "hard-wired" shortcomings than for rulers (see *Errors in practice*), improved technology is certainly available. Top pan digital balances small enough for field use exist and are already being used by ringers. Even so, one should expect to regularly calibrate these with standard masses.

Error in continuous variables

Eliminating systematic error and reducing random measurement errors by correctly using the most suitable equipment is not the only noise reduction possible. Random errors can be reduced by increased measurement precision. This happens because of an error that is always generated when values of a continuous variable are rounded off.

For example, if one were taking wing length to the nearest millimetre and thereby assigned, quite properly, a value of 60 mm to a female Cetti's Warbler whose "true" wing length was 59.51 mm then there is a relative (fractional) error of 0.82%. This seems almost negligible. However, in the context of wing length range, perhaps a more relevant comparison, it acquires greater significance – around 4-5% in the smaller passerines. Taking the measurement with 0.5 mm precision reduces errors of this kind by half.

Unfortunately, systematic "rounding errors" can be made with rulers marked in half millimetres because whole millimetre scale markings are more prominently inscribed than halves. Also, centimetre and half centimetre marks are invariably singled out by labelling and/or enlargement that can draw the eye towards them. This error again appears when spring balances with gram/half-gram calibrations are used to measure masses to the nearest half gram. The effect can be seen in bulked data as multiple modes, leading to a saw-toothed histogram when graphed (unpubl. data and Busse pers. comm.)

The simple expedient of "guesstimating" to the nearest 0.1 mm or gram could alleviate such errors. Measurements taken in 0.1 increments can afterwards be easily rounded to 0.5 precision, if desired, by applying a simple function to the database

in the computer. *Nota bene*, reading a scale to 0.2 units might appear easier but it cannot be used if rounding to 0.5 might take place. Any point reading taken from the scale actually represents an interval, with limits which follow from the chosen level (of precision). Each limit after rounding must coincide with one of the pre-rounding limits.

Closer attention required for additional precision could also help reduce an error due to observer distraction; it seems the "glaring errors" of Lougheed *et al.* (1991) may be of this kind. One can mistakenly read on a wing ruler, say, 79 mm as 89 mm because of the proximity of the wing tip to the inscription 80 (unpubl. data).

Sokal and Rohlf (1993) imply in chapter 2.3 that additional precision in a measurement may lead to improved accuracy of measurement. It cannot cause loss of accuracy and costs nothing, apart from a little extra attention to the task in hand. The best solution of all, however, would be to design dedicated measuring devices that try to address the underlying problems.

Errors arising from rounding

It is pertinent to mention here two, erroneous, rounding methods for continuous variables that some statistics texts advocate for a terminal five (..xx5). Rounding on/up is clearly wrong as the interval represented is 450.. to 550.. of which the portion 450.. to 500.. should round down. It cannot be argued that rounding back/down zero compensates for this, because a terminal zero effectively represents intervals already rounded up (950.. to 000..) and rounded down (000.. to 050..).

Another method quoted is to round up (or round down) if the digit leading the terminal five is odd, to do the opposite if it is even. This should lead in the long run to half rounded up and half rounded down. However, a data set consisting of, say, 4.5, 5.5, 5.5 and 6.5 rounded to the nearest unit will yield 4, 6, 6, 6 (or 5, 5, 5, 7). Analysis focusing on the location of modes within the data will be perverted (unpubl. data).

Rounding "5"s should always be done with the aid of random numbers, a simple and rapid process once the raw data has been entered into a computer data base. This ensures that, on average, half will round up and half will round down, without any reference to the datum value. If it is necessary to work "by hand", a reasonable alternative to random rounding is to round every other "5" in the unsorted data, perhaps even as it is being collected, or to round unwanted figures after classing the data whereupon "5"s can be divided half-and-half between classes.

Error compounding in indices

Fractional errors become additive when one calculates ratios, indices, *etc.* by multiplying or dividing the original measurements. An example will serve to illustrate this point.

In the far west and south-west of Europe the occurrence rate of the Marsh Warbler (*Acrocephalus palustris*) relative to the number of Reed Warblers caught is very low. Few ringers there have the opportunity to familiarise themselves with the

former species. An index from three measurements is proposed by Svennson (1992) as an aid to separating individuals of these two.

To calculate the index, one must initially multiply the width of the tarsus above the fitted ring by the width of the bill at the nares. Both are taken to the nearest 0.1 mm so the rounding error is 0.05 mm in each. The measurements themselves typically come out around 1.8 mm and 4.1 mm, respectively, so the fractional errors are approximately 2.8% (0.05/1.8) and 1.2% (0.05/4.1). These add to give a 4% fractional error for this segment of the index. This is an absolute error equal to 4% of 1.8×4.1, more or less 0.3. *Nota bene*, for the final index, measurement units are unassignable and thus irrelevant.

To complete the calculation of the index, the product (of the "widths") must be subtracted from the length of the bill. This latter measurement is subject to an absolute error of 0.05 mm, yielding a final index with error limits of ± 0.35. The most critical values of the index for identification purposes lie around 8.0-8.5; a fractional error of *ca* 4.24% (0.35/8.25).

These error limits apply regardless of the skill, repeatability, *etc.* with which the observer makes the measurement. Because of problems with repeatability they will be minimum values. Results given in Lougheed *et al.* (1991) and Gosler *et al.* (1998) for calliper measurements on passerines tend to indicate at least as much again. Svennson's (1992) quoted overlap range of 8.0-8.5 should at the very least be more than doubled to 7.65-8.85, because the amount of rounding error associated with measurements from birds falling in-between – 6 out of 959 according to the ringer, G. Walinder (Svennson 1992) – is unknown and unknowable. However, to be realistic (*i.e.* expecting repeatability errors) the safe range for overlap should be taken as 7.3-9.2.

It must be pointed out that for the above example, error is apparently much less if an index were constructed by adding rather than multiplying the two width measures. Also, dividing the bill length by the "widths" rather than subtracting their product would lead to an index that shows a slightly greater error. The reason this happens is methodological. A rigorous treatment of the data – providing a suitable transformation to normality exists – would use means and variances to find an overlap range, within which a given proportion of each species will fall. This would have the additional advantage of incorporating variability due to observer repeatability errors. *Nota bene*, it is likely necessary for any index treated this way to have dimensional consistence, and Svennson's (1992) index is not so constructed.

CONCLUSION

If biometric measurement in bird ringing is to produce reliable results from increasingly sophisticated and detailed analyses, then close attention must be paid to the processes generating the raw data and to its subsequent treatment. Necessary training of field personnel by itself is not sufficient. The measuring equipment used requires careful choice and/or design and those who interpret the collected data need a first-class knowledge of statistical processing.

REFERENCES

Gosler A.G., Greenwood J.J.D., Baker J.K., King J.R. 1995. *A comparison of wing length and primary length as size measures for small passerines: a report to the British Ringing Committee.* Ring. & Migr. 16: 65-78.

Gosler A.G., Greenwood J.J.D., Baker J.K., Davidson N.C. 1998. *The field determination of body size and condition in passerines: a report to the British Ringing Committee.* Bird Study 45: 92-103.

Harper D.G.C. 1994. *Some comments on the repeatability of measurements.* Ring. & Migr. 15: 84-90.

Lougheed S.C., Arnold T.W., Bailey R.C. 1991. *Measurement error of external and skeletal variables in birds and its effect on Principal Component Analysis.* Auk 108: 432-436.

Maitav A., Izhaki I. 1994. *Stopover and fat deposition by Blackcap Sylvia atricapilla following spring migration over the Sahara.* Ostrich 65: 160-166.

Morgan J.H., Shirihai H. 1997. *Blackcap.* In: Morgan J.H., Shirihai H., Yosef R. (Eds). *Passerines and Passerine Migration at Eilat.* IBCE Tech. Publ. 6, 1: 39-40.

Sokal R.R., Rohlf, F.J. 1995. *Biometry.* Freeman, New York.

Svensson L. 1992 *Identification Guide to European Passerines*. Stockholm.