Multivariate Behavioral Research, *27* (3), 435-458

The Quantitative Ethology of the Zebra Finch: A Study in Comparative Psychometrics

Aurelio Jose Figueredo and Donovan Michael Ross
University of Arizona

Lewis Petrinovich
University of California

A *quantitative ethogram* was developed for the Zebra finch, using one-zero focal animal sampling on an ethologically comprehensive checklist of 52 behavioral items, and was assessed for both interobserver reliability and construct validity. Interobserver reliabilities were highly acceptable (an eta-squared of .923 for aggregation periods of 5 minutes). Nine common factors (Singing & Parenting, Social Proximity, Social Contact, Social Submission, Social Aggression, Sex & Violence, Object Handling, Surface Foraging, and General Activity) produced highly acceptable convergent validities (high factor loadings for most behavioral items) and discriminant validities (low factor intercorrelations). Applying the quantitative methods of psychometrics thus permits the verification of ethological theory and the testing of diverse hypotheses with a high degree of sophistication.

The starting point for any proper ethological study of a species whose behavior is not well-understood is to construct an ethogram, or systematic

---

This research was supported in part by a Summer Stipend from the Social and Behavioral Sciences Research Institute (SBSRI), University of Arizona, Tucson, Arizona. Special thanks to Richard Gorsuch, Steve Reise, Keith Widaman, and the University of Arizona Evaluation Group for the Analysis of Data (EGAD) for helpful advice and consultation in the application of the complex quantitative methods, and to Stanley Mulaik for a helpful and insightful philosophical critique of the various metascientific limitations of these quantitative methods. Thanks also to Ramon Rhine and Roberta Cox for the written materials, verbal explanations and kind permission needed to adapt the UCR primate behavior sampling system for application to avian models. Thanks to Nancy Burley for providing a list of tested sexually neutral colors for Zebra finch identification bands. Our research assistants, Scott Biaggi, Andrea "Angel" Blinder, Andrea Bloom, Stephanie Cadwell, Stephanie Chiprin, Brian DeHaan, Ziya Dikman, Elizabeth Ely, Susan Ephraim, Lisa Gaudet, Sharon Harris, Plavilayil Kochumman "P.K." Jacob, Karen Jordan, Patricia Juerling, Dianne Kim, Adam Mann, Patrick McMurphy, Kristine Miller, Yolanda Nunez, Lynn Lackner, Jaime Platas, Amy Powell, Van Santiago, Natasha Smith, Cleo Strazdas, Despina "Michelle" Walter, and Arden Weitzman were exceptionally helpful with the data collection and entry, colony monitoring and maintenance. Thanks also to my wife, Maureen, and to our good friends, John Burling, Jim King, Tony Mascaro & Company, and Murray Victor for their help in the construction of the research aviary.

behavioral inventory. An ethogram is a detailed list of all the fundamental behavioral elements that a species exhibits in a given situation, along with clear descriptive guidelines for assignment to, and unequivocal discrimination between, elements. Such an ethogram can be converted into a quantified behavioral inventory, a *quantitative ethogram*, by enumerating occurrences of each element. It has been charged, however, that any human categorization of nonhuman animal behavior runs the risk of being idiosyncratic, ethnocentric (perhaps androcentric), and anthropomorphic. It has been argued that aspirations to attain descriptive objectivity only serve to obscure biases and "reify" these tainted constructs (Lewontin, Rose, & Kamin, 1984).

Arbitrary choices of responses to record for analysis can indeed produce quite variable results, as pointed out by Petrinovich & Patterson (1979). During the habituation of responses by the White-crowned sparrow (*Zonotrichia leucophrys*), a response decrement occurred to the repeated playback of recorded song for some behaviors, but not for others. Had only one or a few selected behaviors been chosen for study, it could have been concluded either that large differences in response level occurred as a function of reproductive condition, or that there were no such differences, and either that response level did not change over trials, or that it decreased significantly over trials.

Several approaches have been used to cope with such problems. The most obvious is to concentrate on those behaviors that appear most frequently, or which are most easily observable. This solution is not satisfactory because the more difficult to observe behaviors might be as important as the obvious ones, and their relatively infrequent occurrence might signal events of great biological significance. Thus, their relative importance might be underestimated or statistically swamped by those behaviors that occur with greater frequency. Another common method is to decide, on logical grounds, that certain behaviors are related to one another and that they can be represented along a unidimensional quantitative scale. Because all behaviors observed are not independent of one another, one can aggregate several behaviors that are recorded to obtain an index of the strength of behavioral expression. In fact, there is a long tradition of inference by animal behaviorists regarding hypothesized central states based on the patterning of temporal associations between observed behaviors. This tradition extends from Romanes (1882) through Tolman (1925), Tinbergen (1950), Brunswik (1952), Miller (1959), and Hinde (1970). The major problems with the method of postulating central states are how to decide, (a) which behaviors should be combined to provide an index of the different behavioral states, and (b) which behaviors, if any, should be given more weight than others as indices of the behavioral states of the organism.

Fortunately, sophisticated methods to detect and deal with potential biases in any quantitative instrument have been developed and used in the field of psychometrics. Although the traditional ethogram is purely qualitative, and thus not suitable for psychometric analysis, a quantitative ethogram can be evaluated psychometrically for reliability and validity as readily as any other behavioral test. This does not imply that any technique, psychometric or otherwise, represents a panacea for automatically preventing bias. Indeed, any method used may instead introduce an additional bias of its own (Mulaik, 1991). The judicious application of multivariate psychometric methods, however, permits the detection, estimation, and control of any prespecified bias by the systematic manipulation of the critical parameters that are presumably generative of such bias (in this case, the selection of behavioral indicators).

In the present study, the principal issues of methodological concern that must be dealt with are: (a) to establish acceptable levels of interobserver reliabilities, which address the problem of intersubjectivity of ethological description, and (b) to estimate construct validities (convergent and discriminant), which address the problem of theoretical interpretation of results (Campbell & Fiske, 1959). The central idea of construct validity is that an observation involves the measurement of some attribute that is not itself operationally defined: it is a postulated attribute, called a *latent construct*, assumed to be reflected in observed behaviors, called *manifest indicators*. The intended focus is on this postulated attribute rather than on the observed behaviors, scores on criterion variables, or results dependent on the particular method that is used (Ferketich, Figueredo, & Knapp, 1991; Figueredo, Ferketich, & Knapp, 1991; Widaman, 1985). The convergent validity of the different purported indicators of a given latent construct is the correlation of each indicator to that construct, or of these indicators to each other. Conversely, the discriminant validity of the purported indicators of different latent constructs is the lack of correlation of each indicator to the other constructs. Thus, a multivariate operationalization of each construct is obtained rather than an arbitrary univariate *operational definition*.

The main obstacle to the application of multivariate psychometric methods of construct validation, such as confirmatory factor analysis, to ethology is produced by the nature of the sampling designs most often used. Most psychometric techniques were developed for *mass testing*, which is sampling a large number of individuals on a single occasion using a moderate number of variables. With human subjects, it is relatively easy to collect a great deal of data on large numbers of subjects in a short period of time, especially if pencil and pencil tests are used. Idiographic applications were also developed for sampling a single individual on a large number of occasions, again using a moderate number of variables. Most animal behavior research designs conform to neither of these models, and instead sample a moderate number of individuals on a moderate number of occasions using a moderate number of variables.

In the situations typically encountered in animal behavior, it is not possible to increase the number of subjects in an observational sample because natural breeding populations are often relatively small and stable (such as the captive Zebra finch population studied here). In many species, furthermore, some individuals are present for multiple breeding seasons, while others disappear and are replaced. This makes it difficult to obtain either independent or mutually exclusive samples, even if the observer spends many years studying the breeding population. Furthermore, control over important variables is not possible when the animals are sampled from populations composed of several distinct subpopulations, which differ in terms of breeding experience, quality of territory, age, and characteristics of available mates.

It would be truly unfortunate if animal behaviorists were permanently prevented from applying multivariate analytical procedures to describe and explain observational data by the realities of nature that constrain the data base. The intent of this article is, therefore, to adapt certain psychometric procedures to the situations typically encountered in the study of the natural behavior of animals, such as that of the Zebra finch. Thus, standard animal behavior sampling designs will be used which employ substantive theoretical criteria to select an optimal compromise between what have been called "extensive" (multiple individuals) versus "intensive" (multiple occasions) research designs (Kraemer, 1978, 1979). In addition, the practical constraints that exist for many animal studies, such as the natural sizes and distributions of wild social groups, the cost of housing and maintaining captive colonies, and the animal welfare concerns regarding the numbers of subjects necessary, can be accepted without violating critical data requirements if the appropriate statistical procedures are applied.

A validated but infrequently applied psychometric procedure exists for blending the information gathered from multiple individuals on multiple occasions. This procedure will be described and documented in the statistical methods section. As in repeated measures ANOVA, the degrees of freedom from multiple measurement occasions are pooled across multiple individuals by statistically controlling for individual differences between means. An additional benefit of this procedure in the present context is that, unlike traditional psychometrics, but more like classical ethology, species-typical patterns are extracted from the welter of individual differences. Such individual differences can and should be studied also (e.g., Cox, 1989), but perhaps can be most profitably viewed from the perspective of a well-characterized species-typical baseline.

We decided to apply these analyses to the behavior of the Zebra finch, *Poephila (Taeniopygia) guttata* (Passeriformes: Estrildidae). This species of Australian desert grassfinch has been called the “white mouse” of birds because of its many desirable properties as an avian model for: (a) social development; (b) sexual development; (c) vocal development; and (d) substratal neuroanatomy, neurophysiology and behavioral endocrinology (for reviews of the various separate bodies of research literature on the Zebra finch, see Burley, 1986; DeVoogd, 1986; Konishi, et al., 1989; Slater, Eales, & Clayton, 1988; ten Cate, 1989). The present methodological study provides a quantitative framework on which to base a program of integration between all these diverse aspects of Zebra finch psychobiology.

The application of confirmatory factor analysis permitted us to test ethological hypotheses concerning specific theoretical interpretations, rather than rely solely on the data-driven explorations of traditional factor analysis. The hypotheses on which we based our common factor model were quite simple and guided by fairly traditional ethological ideas. Because we wanted to factor analyze all the behaviors in the ethogram, we hypothesized that all behaviors could be characterized primarily as either social or individual in nature. Furthermore, the social behaviors could be roughly divided into affiliative and agonistic categories. The affiliative behaviors would be roughly reciprocal, but could be graded by the closeness of the relationship, and the agonistic behaviors would be more directional, perhaps reflecting relative dominance, but could be graded by the intensity of the conflict. In addition, there would be separate categories for special sexual and parental behaviors. The individual behaviors could be divided into general undirected activity and more specific goal-directed activities, such as foraging for either food or nonfood items. Vocalizations were assigned to the common factors representing the behavioral states that they were believed to signal. Because our guidance from ethological theory was so qualitative and general, some exploratory factor analyses were performed on a pilot sample at the initial stages of theory construction. The results of these data explorations were not accepted uncritically, but were used as an aid in model development for the specification of behaviors not unequivocally assigned to common factors by the above considerations. The hypotheses thereby generated were held to the more demanding standards of subsequent cross-validation at the final stage of theory verification.

Methods

Subjects

The subjects were the adult members of a healthy and vigorous captive colony of Zebra finches. An original stock of 14 breeding adults was introduced in June of 1988. The first 43 offspring produced (all those fledged in 1988) were allowed to remain and mature in the parental colony, yielding a maximal population of 57 resident adults. After that time, the monthly removal of new juveniles became necessary in order to keep the study population manageable for detailed behavioral observation. The subject population therefore varied systematically through time and is described in more detail under the sampling procedures.

Apparatus

A walk-in aviary was constructed 7.25 km northwest of Tucson, Arizona. This aviary measured 3 × 3 × 2.5 meters, and was provided with an inner 1 × 1 × 2.5 meter double-door antechamber. The top was roofed over at an angle, guttered and rainproof, but the 1.25 cm galvanized wire mesh sides were otherwise open to the elements. The interior was provided with multiple wooden nest boxes, placed about 0.6 meters apart, and several multitiered wooden perches (with multiple crossbars at staggered right angles) hanging at different heights, each about 1.2 meters tall and 1 meter in diameter.

Electronic "clickers" were constructed to pulse regularly every 30 seconds. These were connected to commercially available amplified speaker units to time the onset and offset of each observation interval. Data were recorded on standardized checklists. The Size C numbered metallic and colored plastic bands used to identify individual birds were purchased from A.C. Hughes, Ltd., Middlesex, United Kingdom. The colors used were selected to have neither positive nor negative effects upon Zebra finch sexual attractiveness to conspecifics (N. Burley, personal communication).

Procedure

The basic behavioral sampling system used was an adaptation of one developed for use with cercopithecine monkeys. This system has been developed and validated over 18 years of continuous use with a captive colony of macaques at the University of California, Riverside, as well as upon wild troops of baboons in a field site at Mikumi, Tanzania (R.J. Rhine, personal communication). The system is a form of One-Zero Focal Animal Sampling, which is useful for the simultaneous sampling of multiple behaviors on a single individual (the *focal animal*) at a time. Because one or more occurrences of any specific behavior within a sampling interval are scored as a 1, and none as a 0, making no distinction between single and multiple occurrences of that behavior within sampling intervals, the observer is freed, for the remainder of that interval, to concentrate on recording other behaviors. One-Zero Sampling has been shown, both with primate data and through Monte Carlo simulations, to have extremely high multiple correlations (typically exceeding .95 with sampling intervals as long as 120 seconds) with the weighted sum of behavioral frequencies and durations (Rhine & Ender, 1983; Rhine & Flanigon, 1978; Rhine & Linville, 1980), thus providing a single composite index of both frequencies and durations of interaction. No such comprehensive system of behavioral recording has previously been applied to study social interactions in colonial birds.

To develop a system for use with birds, One-Zero Focal Animal Sampling (at 30-second intervals) was used to collect data on colorbanded individuals using an ethologically comprehensive checklist (the quantitative ethogram) of 52 specific behavioral items (Table 1, next page). Birds were sampled randomly (with replacement) as focal animals for observation. Focal animals were followed continuously during observation sessions of 10 (later reduced to 5) minutes each. Twenty-four trained observers were used throughout this project in different pairwise combinations. A total of 26 interobserver reliability studies, of 35 observation sessions each, were done.

There were two major data subsets, utilizing different sampling designs, which were used for different analytical purposes. The first 2 interobserver reliability studies constituted the Pilot Sample: both were performed during Summer 1988. In this sample, totaling 70 observation sessions, 5 adults were sampled every day of the study as focal animals, thus yielding a balanced sampling design. This data set was used only for the initial reliability studies and exploratory factor analyses. The next 24 interobserver reliability studies constituted the Confirmatory Sample. These were performed from Fall 1988 to Fall 1990, as follows: 3 during Fall 1988, 2 during Spring 1989, 4 during Fall 1989, 5 during Spring 1990, and 10 during Fall 1990. In this sample, totaling 840 additional observation sessions, although it was not possible to sample every subject every day, all colony adults were available to be sampled as focal animals. Of 57 adults in the colony, 50 were selected randomly at least once, and, of these, 45 were selected by chance for more than one observation session. This latter data set was used for both reliability and optimal aggregation studies, as well as for confirmatory factor modeling.

Table 1
Definitions of Specific Behavioral Items.

FA	Focal animal (subject under observation)
NFA	Nonfocal animal (not under observation)

SINGS (DIRECTED)	FA directs full adult song to NFA.
BEGS (FOOD)	FA directs begging call towards NFA.
IS BEGGED FROM	NFA directs begging call towards FA.
FOOD IS STOLEN	NFA appropriates food item from FA.
ALLOFEEDS	FA regurgitates food for NFA.
SINGS (UNDIRECTED)	FA produces undirected full song.
BABBLES (SUBSONG)	FA produces juvenile subsong.
APPROACHES	FA comes within 3-4 inches (8-10 cm) of NFA.
IS APPROACHED	NFA comes within 3-4 inches (8-10 cm) of FA.
REMAINS NEAR ("STILL")	FA remains within 3-4 inches (8-10 cm) of NFA without contact for at least 5 seconds.
IS SUNG TO	NFA directs full adult song to FA.
ALLOPREENS	FA preens NFA.
IS ALLOPREENED	NFA preens FA.
HUDDLES	Extensive body contact without preening.
CONTACTS	Physical contact not otherwise indicated.
IS SQUAWKED AT	NFA directs aggressive call towards FA.
IS THREATENED	NFA, with head up and beak agape, tilts body towards and visually "tracks" FA.
IS CHARGED	NFA rapidly approaches FA.
IS CHASED	NFA sustains rapid pursuit of FA.
FLEES (NO PURSUIT)	FA escapes NFA without chase.
TAIL IS PULLED	NFA vigorously pulls FA tail.
FEATHERS ARE PLUCKED	NFA removes FA feathers.
SQUAWKS (AGGRESSIVE)	FA directs aggressive call towards NFA.
THREATENS	FA, with head up and beak agape, tilts body towards and visually "tracks" NFA .
CHARGES	FA rapidly approaches NFA.
CHASES	FA sustains rapid pursuit of NFA.
BEAK FENCES	FA exchanges facial pecks with NFA.
PULLS TAIL	FA vigorously pulls NFA tail.
PLUCKS FEATHERS	FA removes NFA feathers.

FA	Focal animal (subject under observation)
NFA	Nonfocal animal (not under observation)
WHINES (PRECOPULATORY)	FA directs courtship call towards NFA.
FIGHTS (ESCALATED)	FA bites or grapples with NFA.
SOLICITS/COURTS	FA directs lateral darting movements with rapid tail fluttering towards NFA.
MOUNTS	FA sexually mounts NFA.
IS MOUNTED	NFA sexually mounts FA.
CONFISCATES NONFOOD	FA appropriates nonfood item from NFA.
NONFOOD IS CONFISCATED	NFA appropriates nonfood item from FA.
INVESTIGATES NONFOOD/ NONNEST	FA closely examines nonfood item which is not typical nesting material.
MANIPULATES NONFOOD/ NONNEST	FA pushes, pulls or transports nonfood item, not directed towards nest.
GATHERS NESTING MATERIALS	FA collects typical nesting materials and transports them to nest site.
BUILDS/REARRANGES NEST	FA manipulates materials within nest.
WALKS	FA moves more than 3-4 inches (8-10 cm) on foot.
STAYS	FA, with eyes mostly open, does not move for the entire 30 seconds.
RESTS	FA, with eyes mostly closed, does not move for the entire 30 seconds.
EATS	FA consumes food or grit.
STEALS FOOD	FA appropriates food item from NFA.
IS ALLOFED	NFA regurgitates food for FA.
CONTACT CALLS	FA produces short, "beeping" call.
CHORUS CALLS	FA joins in long contact call with NFAs.
AUTOPREENS	FA preens, grooms, scratches self, wipes beak, bathes, dries body, etc.
FLIES	FA moves more than 3-4 inches (8-10 cm) on wing.
DRINKS WATER	FA imbibes water from any source.
BROODS (NEST/EGGS/ CHICKS)	FA sits upon nest, eggs or nestlings.

Data Encoding & Aggregation

The 30-second interval one-zero "hits" (i.e., sampling intervals with scores of 1) were summed across each series of consecutive intervals (totaling either 5 or 10 minutes, as specified below), for each separate behavioral item, yielding 52 numerical item scores, called "Hansen frequencies" (Kraemer, 1979a), per observation session. For common factor modeling, raw item scores were averaged across the two observers, theoretically producing a mean score that is more reliable than either score reported separately by either of the two observers.

Statistical Analyses

Statistical analyses were performed using the SAS (SAS Institute, 1985) and EQS (Bentler, 1989) software packages. Analyses of variance were performed using the SAS ANOVA and GLM procedures; univariate means and bivariate covariance matrices were obtained using the SAS MEANS and CORR procedures. Exploratory factor analyses were performed using the SAS FACTOR procedure; confirmatory factor analyses were performed, and related factor analytic structural equation models were developed using the EQS causal modeling program.

The covariance matrices used for factor analysis were computed on the residuals of a General Linear Model (GLM) that predicted item scores from individual birds alone. As with repeated measures ANOVA, this GLM procedure: (a) statistically controls for individual differences in mean item scores; (b) yields $N - S$ residual degrees of freedom, where N is the total number of sessions for all birds and S is the total number of subjects observed; and (c) assumes no significant treatment by subject interaction (K. F. Widaman, personal communication).

Factor analyzing such a covariance matrix of subject residuals is mathematically equivalent to factor analyzing a weighted mean covariance matrix for the whole group. In this alternative procedure (Gorsuch, 1965, 1983), separate covariance matrices are first computed for each individual subject, collapsing across repeated observation sessions; a single matrix for the whole group is obtained by computing a weighted mean covariance matrix, collapsing across individuals and weighting by the variable number of sessions per subject. For either of these procedures, however, the assumption of "no treatment by subject interaction" (which here translates into statistically equivalent item covariances across subjects), is not critical if the results are applied and interpreted as an ethological description of the proposed species-typical pattern (R.L. Gorsuch, personal communication), the proper unit of analysis here being the "occasion" rather than the "individual" (Gorsuch, 1983).

Confirmatory factor models were evaluated by the use of the statistics chi-squared, CFI (the Bentler-Bonnett Comparative Fit Index), and NNFI (the Bentler-Bonnett NonNormed Fit Index). Chi-squared measures the statistical goodness-of-fit of the covariance matrix observed to that reproduced by the factor model. A significant chi-squared is therefore grounds for rejection of the factor model specified, and a non-significant chi-squared is grounds for its tentative acceptance. The Bentler-Bonnett Comparative and NonNormed Fit Indices are measures of *practical* goodness-of-fit for large sample sizes, such as those used in this study. With such large samples, a small effect will result in a statistically significant lack of fit. However, with such large samples, the CFI and NNFI values should be greater than 0.90 to be considered satisfactory levels of practical goodness-of-fit, even if significant chi-squared values are obtained (Bentler, 1989; Bentler & Bonnett, 1980).

Results

Interrater Reliability

Complex Rater Effects

Classical Test Theory (CT) defines reliability as the proportion of total variance which is *true score* variance. The total variance observed is thus partitioned into two complementary components: *true score* variance and *error* variance. In the estimation of interobserver reliability (Caro, et al., 1980), the *error* variance is defined as the variance attributable to the different raters. Applying the logic of Generalizability Theory (GT), however, permits us to further partition both the true score and error variances into more elementary components. These multiple sources of variance are called GT "facets" and can be modeled as multiple factors in ANOVA (Shavelson, Webb, & Rowley, 1989).

When a true score is multifaceted, the nonrater effects are composed of multiple sources of true score variance. In the present interobserver reliability study, true score is composed of all the main effects and the interactions of the three Nonrater facets: (*I*) Items (specific behaviors); (*S*) Subjects (individual birds); and (*D*) Days (measurement occasions). It follows that the rater effects may also be multifaceted because multiple sources of true score variance make possible multiple corresponding sources of error. In this study, the different raters represent another GT facet and corresponding ANOVA factor. Multifaceted rater effects can thus be modeled as multiple interactions between the *Rater* and the multiple *Nonrater* GT facets represented as ANOVA factors. Thus, in this model, error is composed of the main effect and the interactions (with all of the above effects) of a single Rater facet: (R) Raters (human observers).

Table 2 contains the results of a GT-inspired ANOVA for the Summer 1988 Pilot Sample. Significance tests are not reported because GT treats all variance components as *real* and relies exclusively on parameter estimation for evaluations of relative magnitude. Eta-squared (the ANOVA equivalent of Multiple R^2 – indicating proportions of variance accounted for) was used to hierarchically partition the incremental proportion of variance attributable to each of the different sources: the semipartial eta-squared indicates the proportion of the total variance accounted for by any given source of variance; the partial eta-squared indicates the proportion of either the true score variance (Nonrater effects) or the error variance (Rater effects), respectively, accounted for by any given source of variance (Cohen & Cohen, 1983).

The first major result is that the overall CT reliability is very high. This reliability is represented by a total semipartial eta-squared of .964 for the combined Nonrater effects. Because this number represents the proportion of the total variance which is attributable to true score variance, it may be accepted as the best estimate of CT reliability. The second major result is that both the semipartial and the partial eta-squareds (proportions of .036) of the main effect of Raters and all of the two-way interactions of Raters with Items, Subjects, and Days are extremely close to zero. Thus, most of the rater errors were highly unsystematic. This means that rater errors were not produced by any particular human observers, specific behavioral items, individual birds, or different measurement occasions.

It is therefore unnecessary to report separate estimates of interobserver reliability for any of the different levels of the three Nonrater facets, because the differences between such specific estimates of reliability would be mostly produced by chance. The least systematic Rater effect is the four-way interaction, the closest thing we have to a *pure error* term, and this alone accounts for .673 of the already small combined Rater effects of .036. This term would overwhelm any smaller effects in a formal test of significance. Thus, we may report the single CT estimate of interobserver reliability, applicable across all three GT facets.

The third major result relates to that of true score variance: the single largest proportion of true score variance is attributable to the main effect of Items, whereas both the semipartial and the partial eta-squareds of the main effects and interaction of Subjects and Days are extremely close to zero. This means that there are substantial species-typical variations in rates of occurrence between specific behaviors, but negligible variations between individual

Table 2
Nonrater and Rater Effects (Summer 1988)

	ETA-SQUARED	
NONRATER EFFECTS	SEMIPARTIAL	PARTIAL
ITEMS (*I*)	.445	.462
SUBJECTS (*S*)	.002	.002
$I \times S$	.055	.057
DAYS (*D*)	.001	.001
$I \times D$	.087	.090
$S \times D$	.008	.008
$I \times S \times D$	.366	.380
COMBINED NONRATER EFFECTS	.964	1.000
RATER EFFECTS	SEMIPARTIAL	PARTIAL
RATERS (*R*)	.000	.001
$R \times I$	.000	.010
$R \times S$	.000	.001
$R \times I \times S$	.003	.069
$R \times D$	.000	.002
$R \times I \times D$	.008	.229
$R \times S \times D$	.000	.014
$R \times I \times S \times D$	.025	.673
COMBINED RATER EFFECTS	.036	1.000
TOTAL EFFECTS	1.000	--

subjects (*S*), days (*D*), or individual subjects by days ($S \times D$) in the overall rates of observed behavior.

Because the three true-score facets (Subjects, Items, and Days) are essentially equivalent to Cattell's three dimensions of the data matrix (individuals, variables, and occasions), and thus the three modes of factor analysis (Gorsuch, 1983), these results generally support the assumptions made above in common factor modeling. These models were based on two procedures: (a) the statistical removal of systematic individual differences; and

(b) the treatment of repeated measurement occasions as replications. Although there is some evidence for small variations in the rates of specific behavioral items between individual subjects ($I \times S$) as well as days ($I \times D$), the overall results justify both procedures because the variance attributable to each of the affected GT facets was trivial in magnitude. The most unsystematic Nonrater effect, the three-way interaction, is the closest thing we have to an estimate of random behavioral variability. This effect is interpretable as the degree to which, according to both raters, different birds happened to be doing different things on different occasions of measurement, which is hardly surprising because each bird can be temporarily occupied with a different behavior at random without indicating any persistent differences between either subjects or items over time. The only reason that the effect appears as systematic in the model is that there is substantial interrater agreement on what any bird happens to be doing at any given time, regardless of the portent of the act. This effect accounts for .380 of the combined Nonrater effects of .964, the second largest proportion of true score variance, and would overwhelm any smaller effects in a formal test of significance.

Optimal Aggregation Periods

The problem of optimal data aggregation over time was encountered in the present study when it was decided, at first, to uncritically adopt a 10 minute observation session (data aggregation period), that had been used previously with relatively sedentary primates, for use with this hyperactive avian species. To investigate the optimal period for aggregation, interobserver reliablity analyses of the Fall 1988 to Fall 1989 data were performed separately for the first 5 minutes of each session and for the whole 10 minutes of each session. Table 3 contains the CT reliabilities (semipartial eta-squareds for combined Nonrater effects) for these and other time samples by the length of aggregation period used. As predicted by contemporary aggregation theorists (e.g., Epstein, 1979, 1980, 1983), the reliabilities did decrease with shorter aggregation periods, but the relative loss of interobserver reliability was of no practical importance for these data, as indicated by the negligible differences between the CT reliabilities obtained for the first 5-minute versus the whole 10-minute sample of the Fall 1988 to Fall 1989 data.

After Fall 1989, each observation session was reduced to 5 minutes and the Confirmatory Sample was aggregated over only the first 5 minutes of each session, regardless of whether or not the whole 10 minutes of data were originally collected. This was done even at the cost of omitting the last 5 minutes of the data already collected. It was deemed more cost-effective, in the long run, in terms of optimal time allocation in relation to statistical power,

Table 3
Nonrater Effects (True Score Variance)

TIME SAMPLES	ETA-SQUARED 10 MINUTE	ETA-SQUARED 5 MINUTE
SUMMER 1988 PILOT	.964	--
FALL 1988 TO FALL 1989	.971	.959
SPRING 1990 TO FALL 1990	--	.899
FALL 1988 TO FALL 1990	--	.923

to sample the behavioral states of different individuals at nearly twice the previous rate than preserve what would have been relatively marginal gains in interobserver reliability. Table 3 contains the CT reliabilities (semipartial eta-squareds for combined Nonrater effects) for the Spring 1990 to Fall 1990 data, including all subsequently collected data (using the abbreviated observation sessions), and combined for the Fall 1988 to Fall 1990 data, including the entire Confirmatory Sample (using the reduced aggregation periods). The further small reduction in reliability observed after Fall 1989 was deemed acceptable in exchange for the gain in statistical power achieved by doubling the effective rate of data collection. Thus, the CT estimate of interobserver reliability applicable to the 5-minute aggregation of the entire Confirmatory Sample is .923 – a highly acceptable level.

Construct Validity

Multivariate Models

Convergent and discriminant validities were assessed by multivariate analysis. Multivariate models were used to first identify (using exploratory factor analysis) and then cross-validate (using confirmatory factor analysis) the common factors or latent dimensions underlying the observed behaviors. As noted above, the Pilot Sample was used for the exploratory stage of factor modeling. To maximize comparability with later confirmatory factor modeling procedures, the covariance matrix was analyzed using: (a) a principal factors model, (b) a promax rotation, and (c) a scree test for the number of factors retained. The results of that exploratory stage are not reported here because they were used for initial theory construction and should not count as evidence for subsequent theory verification. For the Confirmatory Sample, the Factor Loading Pattern (Lambda) was used to obtain the convergent validities between the specific behavioral items, and the Factor Intercorrelation Matrix (Phi) was used to obtain the discriminant validities between the latent common factors (Widaman, 1985).

The confirmatory factor model fit indices and nested model comparisons, which indicate both the absolute and relative, statistical (chi-squared) and practical (CFI and NNFI) goodness-of-fit of each of three alternative models, are presented in Table 4; the factor intercorrelations (bivariate correlation coefficients), which indicate the discriminant validities between the latent constructs, are presented in Table 5; and the factor loadings (standardized regression coefficients), which indicate the convergent validities among the manifest indicators, are presented in Table 6. Factor intercorrelations and factor loadings that were set equal to zero are not shown to avoid visual clutter.

All three alternative common factor models described in Table 4 were acceptable by both of the practical (CFI and NNFI) measures of fit. Nested model comparisons using the stricter statistical (chi-squared) measures of fit revealed that: (a) Model A, a *fully oblique* model, freely estimating all of the possible factor intercorrelations, was significantly better than Model C, a *fully orthogonal* model, freely estimating none of the possible factor intercorrelations; (b) Model A, the *fully oblique* model, however, was not significantly better than Model B, a *partially oblique* model, freely estimating only four of the possible factor intercorrelations; and (c) Model B, the *partially oblique* model, was still significantly better than Model C, the *fully orthogonal* model. After the *omnibus null* hypothesis for full orthogonality was rejected by comparing Model C to Model A (as partial *protection* against chance selection of individual correlations), the four factor intercorrelations to be freely estimated in the partially oblique model were empirically specified by the use of diagnostic Lagrange Multiplier Tests on the fully orthogonal model (Bentler, 1989). Model B was therefore accepted as the best of the three alternatives, specifying the smallest number of factor intercorrelations consistent with the data.

Tables 5 and 6 contain the parameter estimates obtained for Model B. Although not all of the hypothesized factor loadings achieved statistical significance, many others were quite high, indicating high convergent validity between these behavioral measures. The use of confirmatory factor analysis permitted the imposition of factorial simplicity for all the items. This condition, of no more than one significant factor loading per item, renders the factorial interpretation of each item unequivocal, though perhaps fallible, as an

Table 4
Nine Common Factor Model Comparisons (Fall 1988 To Fall 1990)

Codes	Alternative Models	Chi-Squared	*df*	*P*(Ho)	CFI	NNFI
A	Fully Oblique	2548.973	1192	.001	.984	.982
B	Partially Oblique	2580.506	1224	.001	.984	.983
C	Fully Orthogonal	2652.533	1228	.001	.983	.982
Codes	**Nested Comparisons**	**Chi-Squared**	***df***	***P*(Ho)**	**CFI**	**NNFI**
C-A	-36 Common Factor	103.558	36	.001	-.001	.000
--	Intercorrelations					
B-A	-32 Common Factor	31.533	32	.490	.000	.001
--	Intercorrelations					
C-B	-4 Common Factor	72.027	4	.001	-.001	-.001
--	Intercorrelations					

Table 5
Nine Common Factor Intercorrelations (Fall 1988 To Fall 1990)

Factors\Factors	F1	F2	F3	F4	F5	F6	F7	F8	F9
Singing & Parenting	1.000								
Social Proximity	.245*	1.000							
Social Contact	.	.	1.000						
Social Submission	.	.	.	1.000					
Social Aggression	.	.	.	.	1.000				
Sex & Violence	.	.	.	.	.	1.000			
Object Handling	.	.	.	.	.	.	1.000		
Surface Foraging	.	.300*	.	.	-.190*	.	.	1.000	
General Activity	.	-.125*	.	.	.	.	.	.	1.000

* *P*(Ho) < .05

indicator of a single latent construct. This means that although each such indicator contains unique, or *error*, variance not associated with its common factor, it does not contain confounding, or *extraneous*, variance from other common factors which would complicate interpretation. The fact that the final model remains rejectable statistically by the strict chi-squared criterion,

Table 6
Nine Common Factor Pattern (Fall 1988 To Fall 1990)

Items\Common Factors	F1	F2	F3	F4	F5	F6	F7	F8	F9
Sings (Directed)	.099*	.	.	.	.	.	.	.	.
Begs (Food)	.	.	.	.	.	.	.	.	.
Is Begged From	.684*	.	.	.	.	.	.	.	.
Food IsStolen	.188*	.	.	.	.	.	.	.	.
Allofeeds	1.000*	.	.	.	.	.	.	.	.
Sings (Undirected)	.	-.275*	.	.	.	.	.	.	.
Babbles (Subsong)	.	-.043	.	.	.	.	.	.	.
Approaches	.	.643*	.	.	.	.	.	.	.
Is Approached	.	.439*	.	.	.	.	.	.	.
Remains Near ("Still")	.	.646*	.	.	.	.	.	.	.
Is Sung To	.	.	.103*	.	.	.	.	.	.
Allopreens	.	.	1.000*	.	.	.	.	.	.
Is Allopreened	.	.	.197*	.	.	.	.	.	.
Huddles	.	.	.018	.	.	.	.	.	.
Contacts	.	.	.118*	.	.	.	.	.	.
Is Squawked At	.	.	.	.089*	.	.	.	.	.
Is Threatened	.	.	.	.389*	.	.	.	.	.
Is Charged	.	.	.	.427*	.	.	.	.	.
Is Chased	.	.	.	.026	.	.	.	.	.
Flees (No Pursuit)	.	.	.	1.000*	.	.	.	.	.
Tail Is Pulled	.	.	.	-.034	.	.	.	.	.
Feathers Are Plucked	.	.	.	-.037	.	.	.	.	.
Squawks (Aggressive)	.	.	.	.	.289*	.	.	.	.
Threatens	.	.	.	.	.611*	.	.	.	.
Charges	.	.	.	.	.473*	.	.	.	.
Chases	.	.	.	.	.189*	.	.	.	.
Beak Fences	.	.	.	.	.241*	.	.	.	.
Pulls Tail	.	.	.	.	.253*	.	.	.	.
Plucks Feathers	.	.	.	.	.111*	.	.	.	.
Whines (Precopulatory)	.	.	.	.	.	-.009	.	.	.
Fights (Escalated)	.	.	.	.	.	.542*	.	.	.
Solicits/Courts	.	.	.	.	.	.756*	.	.	.
Mounts	.	.	.	.	.	.518*	.	.	.
Is Mounted	.	.	.	.	.	.500*	.	.	.
Confiscates Non Food	.	.	.	.	.	.	.268*	.	.
Non Food Is Confiscated	.	.	.	.	.	.	.606*	.	.
Investigates Non Food/ Nest	.	.	.	.	.	.	-.022	.	.

Manipulates Non Food/ Nest	.	.	.	.	.	.	.036	.	.
Gathers Nesting Materials	.	.	.	.	.	.	.217*	.	.
Builds/Rearranges Nest	.	.	.	.	.	.	-.093	.	.
Walks	.	.	.	.	.	.	.	.508*	.
Stays	.	.	.	.	.	.	.	-.673*	.
Rests	.	.	.	.	.	.	.	-.251*	.
Eats	.	.	.	.	.	.	.	.936*	.
Steals Food	.	.	.	.	.	.	.	.069	.
Is Allofed	.	.	.	.	.	.	.	.000	.
Contact Calls	.	.	.	.	.	.	.	.	.148*
Chorus Calls	.	.	.	.	.	.	.	.	.025
Autopreens	.	.	.	.	.	.	.	.	.910*
Flies	.	.	.	.	.	.	.	.	.249*
Drinks Water	.	.	.	.	.	.	.	.	.126*
Broods (Nest/Eggs/ Chicks)	.	.	.	.	.	.	.	.	-1.000*

* $P(\mathrm{Ho}) < .05$

however, may indicate that such factorial simplicity is not perfectly consistent with the data. Nevertheless, the magnitude of the *practical* indices of fit indicate that any violation of this pattern must be relatively minor. Furthermore, diagnostic Lagrange Multiplier Tests for any item factorial complexity that remained undetected suggested no model respecification that appeared to us to be either theoretically interpretable or intuitively reasonable. The 4 largest, and 11 of the 20 largest, standardized residuals were variance rather than covariance terms, indicating that much of the lack of statistical model fit (as indicated by chi-squared) was due to large item unique variances rather than unexplained item covariances.

Buttressing this argument is the high discriminant validity obtained between the common factors, because incorrectly forcing factorial simplicity would have had the effect of increasing factor intercorrelations. Indeed, *oblique* factor rotations are often used to achieve more *simple structure* at the expense of sacrificing independence between common factors (Gorsuch, 1983). Although four of the factor intercorrelations were statistically significant, they were quite small, none higher than .300, indicating very good model discrimination even between common factors that were found to be significantly correlated.

Common Factors

Nine common factors were obtained: three Affiliative factors, (F1) Singing & Parenting, (F2) Social Proximity, and (F3) Social Contact; three Agonistic factors, (F4) Social Submission, (F5) Social Aggression, and (F6) Sex & Violence; and three Individual factors, (F7) Object Handling, (F8) Surface Foraging, and (F9) General Activity. Social Proximity (F2) was positively correlated to both Singing & Parenting (F1) and Surface Foraging (F8), but negatively correlated to General Activity (F9). Social Aggression (F5) was negatively correlated to Surface Foraging (F8).

The Affiliative Factors

Singing & Parenting (F1), involves only two of the major vocal behaviors because certain specific vocal signals were more highly correlated with other common factors. F1 includes a vocal signal commonly directed by adults at juveniles, "directed singing", a vocal signal and a related behavior directed by juveniles at adults, "being begged from" and "food being stolen", and the most typical adult response, "allofeeding". This pattern of factor loadings suggests a parent-offspring context for this affiliative interaction. Both Social Proximity (F2) and Social Contact (F3) are also affiliative factors, distinguished by spatial congregation, such as mutual "approaching" and "remaining near", versus actually physically touching, and such as mutual "allopreening" and "contacting", which suggests increasing degrees of closeness in an affiliative relationship. In addition, whereas "undirected singing" is negatively correlated to Social Proximity (F2), "being sung to" is positively correlated with Social Contact (F3), which suggests very different social contexts for what appear to be the same vocalizations, modulated only by "directedness" as a metacommunicative signal.

The Agonistic Factors

Both Social Submission (F4) and Social Aggression (F5) are agonistic factors, primarily distinguished by the direction of the agonistic act, such as "squawking", "threatening", and "charging", which suggest degrees of relative dominance in an agonistic relationship. "Fleeing without pursuit" correlates with Social Submission (F4), however, indicating that this factor is not merely defined by the passive receipt of aggressive behavior but is characterized by an active avoidance response (and one apparently accepted by the aggressor, who does not offer pursuit). Similarly, the more perseverating agonistic behaviors, such as "chasing", "beak fencing", "tail pulling", and "feather plucking" are significantly correlated with Social Aggression (F5), but not reflexively with Social Submission (F4).

Sex & Violence (F6) is also a partially agonistic factor, distinguished from Social Aggression (F5) by the severity of the attack, which suggests there is an escalation of the conflict. Interestingly, several moderately injurious agonistic behaviors, such as "beak fencing", "tail pulling", and "feather plucking" correlate with Social Aggression (F5) rather than with Sex & Violence (F6), suggesting that they represent more of a highly escalated threat than any serious combat. Overtly sexual behaviors, such as "soliciting/courting" and mutual "mounting", also correlate highly with Sex & Violence (F6), perhaps reflecting a state of hormonal activation that underlies both sexual motivation and the escalation of violence. Social Submission (F4) seems to have no escalated counterpart, which suggests that unrelenting violence (nonacceptance of submission) forces defensive retaliation rather than more vigorous appeasement.

The Individual Factors

Object Handling (F7) involves manipulation of inedible items, such as "gathering nesting materials". It is interesting to note that the mutual "confiscating" of such items correlates with this factor rather than with the social ones. In primates such confiscations are interpreted as being laden with ulterior social significance; in these birds, however, such transfers seem motivated instead by direct attraction to the object rather than the indirect opportunity to display social status.

Surface Foraging (F8), which involves "walking" and "eating", also correlates negatively with the two most passive states, "staying" and "resting". General Activity (F9), which involves "contact calling", "autopreening", "flying", and "drinking water" also correlates negatively with the next most passive state, "brooding". These results suggest that subtle differences between such passive states may help indicate not only what the bird is not doing within the given 30-second interval, but also what the bird is least likely to be doing within the rest of that session.

Discussion

A *quantitative ethogram* was developed for the Zebra finch, using a system of one-zero focal animal sampling on an ethologically comprehensive checklist of 52 specific behavioral items, and was assessed for both interobserver reliability and construct validity. Interobserver reliabilities were found highly acceptable regardless of the data aggregation period used. Using an aggregation period of 5 minutes, an average interobserver reliability of .923 was obtained. Nine common factors were identified which produced highly acceptable convergent validities (high common factor loadings) for the majority of behavioral items, and excellent discriminant validities (low common factor intercorrelations) between the common factors.

These last two points merit some further comment. The simultaneous achievement of both structural simplicity and low intercorrelations of the common factors is desirable but unusual. Most factor analyses reported in the existing literature, however, are applied to human data. It could well be the case that such results are more typical for nonhuman animal data, but this remains to be determined. Nevertheless, such a simple behavioral factor structure, as is here reported for the Zebra finch, permits efficient testing of diverse hypotheses. It should be clarified that once this multivariate groundwork has been adequately laid, such factor solutions can be applied with great simplicity in the service of any future quantitative study that is more focused in scope. Researchers desiring to use these results are not required to perform needless repetitions of the complex modeling procedures detailed here, but can better analyze their data by computing validated and meaningful factor scores that are justified on the basis of this study. "Unit weightings" (Gorsuch, 1983) of standardized scores, or simple averages of raw scores (because all the indicators are expressed in a common metric), can be computed across all the significant indicators of each common factor.

More general implications also follow. First of all, it becomes meaningful to discuss the psychometric properties of an ethogram. The accepted and demanding psychometric standards of construct validation can be applied to ethological theory. Second, an already existing but underappreciated quantitative method has been successfully applied which blends data from both *intensive* and *extensive* research strategies, and permits optimal allocation of research effort relatively unconstrained by the technical requirements of traditional statistical models. Third, it should be noted that such applications are not limited to nonhuman animals, but may be extended to human behavior where similar research problems are encountered.

The generality of the mode of approach detailed here has been demonstrated in a series of studies of the social behaviors of different species in a variety of contexts. For example, these factoring methods were successfully applied to compare the ethograms of a captive colony of Stumptail macaques at the University of California, Riverside, California, and a semi-free ranging population of Stumptail macaques on Tanaxpillo Island, Catemaco, Mexico. The method was also used profitably to study the relationships between the verbal self-reports and physiological responses involved in human emotional expression. Although all the methodological issues and philosophical implications brought up in the introduction could not be properly addressed by the present study alone, due to limitations inherent in its design, this is but the first in a series of articles, currently in preparation, representing an ongoing research program in comparative psychometrics with the purpose of adapting and applying multivariate quantitative methods in ethology. These results were discussed by Figueredo and Petrinovich (1991).

References

Bentler, P. M. (1989). *EQS structural equations program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Burley, N. (1986). Sexual selection for aesthetic traits in species with biparental care. *American Naturalist, 137,* 415-445.

Brunswik, E. (1952). The conceptual framework of psychology. In *International encyclopaedia of unified science* (Vol. 1). Chicago: University of Chicago Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Caro, T. M., Roper, R., Young, M., & Dank, G. R. (1980). Interobserver reliability. *Behavior, 69,* 303-315.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cox, R. L. (1989). *Personality profiles of mother and offspring stumptailed macaques (Macaca arctoides): Behavior, personality, and dominance rank.* Unpublished doctoral dissertation, University of California, Riverside.

DeVoogd, T. J. (1986). Steroid interactions with structure and function of avian song control regions. *Journal of Neurobiology, 17,* 177-201.

Epstein, S. (1979). The stability of behavior, I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097-1126.

Epstein, S. (1980). The stability of behavior, II. Implications for psychological research. *American Psychologist, 35,* 790-806.

Epstein, S. (1983). Aggregation and beyond, Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360-391.

Ferketich, S. L., Figueredo, A. J., & Knapp, T. R. (1991). Focus on psychometrics: The multitrait-multimethod approach to construct validity. *Research in Nursing and Health, 14,* 315-320.

Figueredo, A. J., Ferketich, S. L., & Knapp, T. R. (1991). More on MTMM: The role of confirmatory factor analysis. *Research in Nursing and Health, 14,* 387-391.

Figueredo, A. J., & Petrinovich, L. (1991). *Multivariate methods in ethology.* Symposium conducted at the Western Psychological Association Conference, Burlingame, California. Edited symposium monograph in preparation.

Gorsuch, R. L. (1965). *The clarification of some superego factors.* Unpublished doctoral dissertation, University of Illinois.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale, NJ: Lawrence Erlbaum.

Hinde, R. A. (1970). *Animal behaviour: A synthesis of ethology and comparative psychology.* New York: McGraw Hill.

Konishi, M., Emlen, S. T., Ricklefs, R. E., & Wingfield, J. C. (1989). Contributions of bird studies to biology. *Science, 246*, 465-472.

Kraemer, H. C. (1978). Empirical choice of sampling procedures for optimal research design in the longitudinal study of primate behavior. *Primates, 18*, 825-833.

Kraemer, H. C. (1979a). One-zero sampling in the study of primate behavior. *Primates, 20*, 237-244.

Kraemer, H. C. (1979b). A study of reliability and its hierarchical structure in observed chimpanzee behavior. *Primates, 20*, 553-561.

Lewontin, R. C., Rose, S., & Kamin, L. J. (1984). *Not in our genes*. New York: Pantheon.

Miller, N. E. (1959). Liberalization of basic S-R concepts. In Koch, S. (Ed.), *Psychology, a study of a science: Study I*. New York: McGraw Hill.

Mulaik, S. A. (1991). Factor analysis, information-transforming instruments, and objectivity: A reply and discussion. *British Journal for the Philosophy of Science, 42*, 87-100.

Petrinovich, L., & Patterson, T. L. (1979). Field studies of habituation: I. the effect of reproductive condition, number of trials, and different delay intervals on the responses of the white-crowned sparrow. *Journal of Comparative and Physiological Psychology, 93*, 337-350.

Rhine, R. J., & Flanigon, M. (1978). An empirical comparison of one-zero, focal-animal, and instantaneous methods of sampling spontaneous primate social behavior. *Primates, 19*, 353-361.

Rhine, R. J., & Linville, A. K. (1980). Properties of one-zero scores in observational studies of primate social behavior, the effect of assumptions on empirical analyses. *Primates, 21*, 111-122.

Rhine, R. J., & Ender, P. B. (1983). Comparability of methods used in the sampling of primate behavior. *American Journal of Primatology, 5*, 1-15.

Romanes, G. J. (1882). *Animal intelligence*. London: Paul Kegan.

SAS Institute (1985). *SAS user's guide, Statistics, Version 5*. Cary, NC: SAS Institute.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Slater, P. J. B., Eales, L. A., & Clayton, N. S. (1988). Song learning in Zebra finches (*Taeniopygia guttata*): Progress and prospects. *Advances in the Study of Behavior, 18*, 1-34.

ten Cate, C. (1989). Behavioral development: Toward understanding process. In Bateson, P. P. G., *Perspectives in ethology, 8*, 243-269. New York: Plenum.

Tinbergen, N. (1950). The hierarchical organization of nervous mechanisms underlying instinctive behavior. *Symposia of the Society for Experimental Biology, 4*, 302-312.

Tolman, E. C. (1925). Behaviorism and purpose. *Journal of Philosophy, 22*, 36-41.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1-26.

Accepted February 7, 1992.