

A Generalizability Analysis of Subjective Personality Assessments in the Stumptail Macaque and the Zebra Finch

Aurelio José Figueredo

University of Arizona

Roberta Lea Cox and Ramon J. Rhine

University of California, Riverside

Psychometric findings are reported from two studies concerning the construct validity, temporal stability, and interrater reliability of the latent common factors underlying subjective assessments by human raters of personality traits in two nonhuman animal species: (a) the Stumptail macaque (*Macaca arctoides*), a cercopithecine monkey; and (b) the Zebra finch (*Poephila guttata*), an estrildid songbird. Because most theories of animal personality have historically implied that certain personality constructs should be relatively universal across taxa, parallel analyses of similar data are reported for two phylogenetically distant species of subject using the same psychometric methods. Each of the samples was drawn from a socially-housed colony of the same species: that of macaques consisted of 5 mature adult females and 8 of their adult offspring and that of finches consisted of 5 adult individuals. A modified version of the 1978 Stevenson-Hinde and Zunz (SHZ) list of personality items was applied to the macaques at various times during the eight years from 1980-1988 and to the finches during 1992. This study also used the three SHZ scales — Confident, Excitable, and Sociable — originally derived from principal components. Generalizability analyses were used to assess the construct validity, temporal stability, and interrater reliability of the hypothesized factors. Both Stumptail macaques and Zebra finches manifest measurable personality factors that are highly valid across multiple items, stable across multiple years, and reliable across multiple raters. The same model fits both species, as predicted by theory. The construct validity of the factors is slightly higher for the finches than for the macaques, although the interrater reliability is somewhat lower. This study illustrates how generalizability analysis can be used to test prespecified confirmatory factor models when the number of individual subjects is quite small.

We wish to thank Jim King, Lewis Petrinovich, Lorna Roney, Anne Scott, Lee Sechrest, and Brad Smith for their many helpful comments on the manuscript. We also thank David Funder for helping us navigate through the bewildering maze of literature on the person-situation debate. The many students who assisted with the Stumptail macaque research over the 21 years of the existence of the Riverside, California, Stumptail macaque colony left an irreplaceable record of Stumptail behavior and personality, and for this we are forever in their debt. We also thank the many students who assisted with the Zebra finch research over the 5 years of the existence of the Tucson, Arizona, Zebra finch colony for their help in colony monitoring, maintenance, and data collection.

This article reports the psychometric findings from two studies concerning the construct validity, temporal stability, and interrater reliability of the latent common factors underlying subjective assessments by human raters of personality traits in two nonhuman animal species: (a) the Stumptail macaque (*Macaca arctoides*), a cercopithecine monkey; and (b) the Zebra finch (*Poephila guttata*), an estrildid songbird. Theories that have historically motivated the study of animal personality, such as Plutchik's (1962, 1980) psychoevolutionary theory of emotion, imply that certain personality constructs should be relatively universal across taxa. These constructive replications were done to explore the generality of both the substantive theory and the psychometric methods. This article reports the results of parallel analyses of similar data for two phylogenetically distant species of subject using the same psychometric techniques.

By way of introduction, we first review the status of personality as a concept in both human and nonhuman animals. We then present a brief history of the theories of both personality and emotion as specifically applied to nonhuman animals and describe the prior empirical work that has been done in this area. Because no previous work has been done on personality assessment in avian species, this review is limited to the literature on nonhuman primates. Finally, we provide a methodological critique of the statistical models used in much of that prior work and propose alternative solutions based on creative applications of Generalizability Theory (GT). In the process, we propose a general rationale for the use of generalizability analysis for common factor modeling with small samples. This may represent a methodological innovation of more general interest.

The Scientific Status of Personality Constructs

Personality is a term that has historically been difficult to define, with some theorists denying it a place in scientific inquiry (e.g., Skinner, 1953). A conceptual definition of personality was given by Allport (1961) as follows: "Personality is the dynamic organization within the individual of those psychophysical systems that determine his characteristic behavior and thought (p. 28)." This statement reflects the position that personality: (a) exists as a definable construct; (b) represents a fluid property of a constantly changing and adapting organism; and (c) characterizes the individual, rather than being constructed by the observer. This formulation also does the following: (a) stresses the importance of both psychological and biological factors in personality, offering a holistic approach that may be applied to animals; (b) implies that personality as a construct has a measurable effect on behavior; (c) presents personality as a construct that is consistent over

time and across situations; and (d) implies that personality may express itself in all aspects of life. Since consistency allows one to predict future behavior, knowledge of personality aids in such prediction.

The sustained data collection effort for the present studies spanned the years from 1980 to 1992. This empirical research program was initiated and directed by Rhine, continued by Cox, and extended by Figueredo. During that period of time, which included the entire graduate training of the latter two coauthors, there have been major developments in personality theory. For example, there has emerged a widespread consensus around what have come to be known as the "Big Five" personality factors (cf., Digman, 1989; Goldberg, 1990, 1992, 1993), although the five factor model itself had been proposed much earlier (Fiske, 1949). Although this consensus is still far from universal (cf., Ozer & Reise, 1994; Waller & Ben-Porath, 1987), no such broad convergence was evident at the inception of these studies. Furthermore, the focus of debate has shifted from the resolution of the so-called "person-situation debate" (e.g., Bem, 1972; Houts, Cook, & Shadish, 1986; Mischel, 1968, 1973) to a variety of more current concerns (see Ozer & Reise, 1994, for an excellent review of contemporary thinking in the field). Although some convergence seems to have been reached over the ontological reality and practical utility of both personological and situational factors in the prediction of behavior (cf., Funder, 1983a, 1991; Funder & Colvin, 1991; Mischel, 1983, 1984; Mischel & Peake, 1982), including the substantial influence of interactions between persons and situations (e.g., Caspi, Bem, & Elder, 1989), there has remained some residual controversy over the use of subjective assessments of personality (Funder, 1983b).

That subjective judgments of personality are effectively used by people to predict the behavior of other individuals over time and across different situations is evident in everyday life. However, the polemical experience of the person-situation debate has served to refine a set of more stringent criteria for the evaluation of such assessments (Kenrick & Funder, 1988). Although numerous potentially biasing factors have been identified in the subjective assessment of personality, which limit the circumstances under which they might be valid, certain converging forms of empirical evidence have been proposed to mitigate against the premature conclusion that such judgments are composed of *nothing but* rater bias. These criteria include, but are not limited to, the following: (a) interrater reliability, (b) interitem consistency, (c) cross-situational consistency, and (d) cross-temporal stability.

What is thought to occur in subjective personality assessment is that a rater makes a prediction of the average behavior of an individual in a sample of events, based on their exposure to that individual's behavior over a

previous sample of events (cf., Epstein, 1977; Funder & Dobroth, 1987; Funder & Sneed, 1993). For such ratings to have adequate predictive validity, Kenrick and Funder (1988, p.31) have drawn lessons from the history of the person-situation debate to recommend that the following critical design features be used: "(a) raters who are thoroughly familiar with the person being rated; (b) multiple behavioral observations; (c) multiple observers; (d) dimensions that are publicly observable; and (e) behaviors that are relevant to the dimensions in question."

The empirical studies reported in this article are based on ratings by observers who had a minimum experience of several months to several years in the systematic behavioral observation of the individuals in question. The cumulative number of behavioral observations, on which the subjective ratings were presumably based, numbered literally in the thousands. These behavioral observations were ethologically comprehensive (Figueredo, Petrivovich, & Ross, 1992; Rhine, 1972; Rhine & Kronenwetter, 1972), in that they included virtually all of the social behaviors that the animals spontaneously engaged in. The number of raters used in each study was 22 for the Stumptail macaques and 7 for the Zebra finches. The dimensions used were explicitly tied to overt behavioral manifestations with which the raters had direct observational experience. These latent dimensions, however, were more than just mechanical translations of molecular behaviors into summary labels, but subjective judgments of the complex interrelationships between behavior of the individual and the social environment in which it is expressed and to which it contributes.

One of the two empirical studies reported in this article attempted to present a more natural situation for the inference of the consistency of personality by utilizing personality ratings taken over a span of eight adult years, which is a significant proportion of the Stumptail macaque's adult lifespan. Furthermore, the personality ratings for the present studies were based on observations of individual animals interacting normally within the colonies. Thus, these studies avoided the potentially artifactual nature of many contrived laboratory settings, which may not generalize well to natural behavior. These studies therefore make use of types of personality data that have proven successful to demonstrate personality consistency and continuity for humans (e.g., Block, 1977; Caspi et al., 1989), yet which are often very difficult to gather, namely, observer evaluations of individuals leading natural lives.

While we do not expect that the emerging consensus standards for subjective personality assessment be artificially lowered for the study of nonhuman animals, neither do we expect that they be arbitrarily raised. Thus, we propose that any empirical evidence for the validity of such

judgments in nonhuman animals be given the same scientific status as comparable evidence for those in humans. It might be argued that rating an individual's personality might require knowledge of its private thoughts, feelings, motivations, and attitudes, and not just its behavior. Whereas such insights might or might not be deemed necessary by different thinkers, we are not convinced that self reports provide any more direct access to these things than do rater reports (cf., Digman, 1989; Funder, 1991; Funder & Drobth, 1987; Funder & Sneed, 1993; Kenrick & Funder, 1988). While we admit that we are necessarily lacking self-report measures for these nonhuman animals, we do not see that omission as automatically fatal to our enterprise.

The Study of Emotion and Personality in Nonhuman Animals

Darwin (1872/1965) first stressed the evolutionary continuity of behavior patterns which are interpreted as emotion within each species, and tested his conclusions regarding the evolution of emotion by the comparative method. While emotion does not equal personality, Darwin provided the base upon which all nonhuman animal personality research was built.

Hebb (1946) wished to objectify the intuitive categorization of emotion to make this a method more suitable for scientific research. To do this Hebb started out with intuitive emotion categories, and through a retracing of the process of intuitive judgment determined the behaviors on which they were based. He concluded that the classification of emotional behavior is based upon a complex set of cues which are of varying degrees of importance depending on the specific judgment in question. Hebb's study was one of the earliest attempts to objectify the intuitive categorization of emotion in nonhuman primates. He discussed the importance of the large number of complex cues which allow the observer to make personality judgments. Where Darwin investigated the evolutionary continuity of emotional expression Hebb emphasized the importance of continuity of individual behavioral and emotional patterns over time.

To further objectify the study of expressive movements Van Hoof (1962) steered away from Hebb (1946) and Darwin's (1872/1965) style of anthropomorphic labeling of emotion, and anecdotal method, toward the comparative ethological method of objective description of displays. Van Hoof concluded that there are two sets of compound expressions in primates, each grouped according to the motivational states surrounding their performance: (a) Agonistic, a balance between the tendencies to flee and to attack; and (b) Attraction-Repulsion, a conflict between social attraction and the tendency to flee. Van Hoof described in detail the physical movements

and their associated social situations, thus emphasizing the importance of species-specific knowledge when attempting to categorize emotional states. Due to his extreme objectification, and in line with his ethological method, Van Hoof presents what is simply a descriptive list of facial expressions and their associated behavioral situations. While these descriptions are important in understanding the behavioral *repertoire* of a species, which, as demonstrated by Hebb and later Van Lawick-Goodall (1968), is necessary for an understanding of emotion and personality in each particular species, it fails to answer questions of continuity and consistency in individual behavior. Van Hoof described how the primate behaves in specific situations, but lacked a unifying theoretical interpretation of behavior.

Within her comprehensive monograph of chimpanzee behavior Van Lawick-Goodall (1968) investigated chimpanzee expressive, communicative movements. Goodall based her descriptions and interpretations on extensive observational field data. The behaviors were described on a molar level, rather than on the more molecular level typical of the ethologists. The contexts of the behaviors were described in terms of their emotional as well as physical content. Anthropomorphic terms such as "frustration," "tolerance," and "reassurance" were used to describe the expressive movements and their functions. Goodall presented a comprehensible categorization of expressive movements based upon anthropomorphic terminology and anecdotal methodology. She revealed the complexity of the large number of communicative cues surrounding each expressive behavior, and showed the importance of long term observation of known individuals in a species-typical social environment.

Chamove, Eysenck and Harlow (1972) conducted one of the first studies which utilized psychometric methodology and relatively complex statistical analyses of social interaction data in order to uncover stable personality traits in nonhuman primates. Data were obtained from observations of social behaviors, such as social exploration, social play, and hostile contact. Factor analyses of these behaviors revealed three strong, almost independent factors: hostile, fearful, and affiliative/sociable. Chamove, Eysenck and Harlow noted that these factors are very similar to the extroversion, psychoticism and emotionality factors often found in humans. Physiological, anatomical, and socio-behavioral similarities between humans and nonhuman primates were cited as reasons for the similar findings. The researchers concluded that "This observational study demonstrates marked individual differences between monkeys in their social behavior. These differences are apparently highly reliable, and characteristic of the animals concerned, and may thus be regarded as aspects of the 'personality' (p. 502-503)." Chamove, Eysenck and Harlow brought together Darwin's

(1872/1965) evolutionary theory and comparative method with Van Hoof's (1962) call for objectification and presented experimental evidence for individual behavioral continuity over time (i.e., personality) for nonhuman primates, as well as for interspecific continuity of personality types.

Plutchik's (1962) psychoevolutionary theory of emotion represents one of the earliest attempts to integrate evolutionary and psychological theories to explain and describe emotions across all species. Based on this theory, Plutchik and others have developed rating scales of emotions for humans which have been modified and used for a variety of species. Plutchik's basic premise is that the concept of emotion should be considered relevant to the entire evolutionary scale, applying equally well to human and nonhuman animals. Plutchik postulated that emotions should be identified in terms of overall behavior, which may be observed across species. This functionalist approach assumes that emotions serve an adaptive role in helping organisms interact with their environment in ways which insure continuity of their genes over future generations. The eight primary emotions hypothesized by Plutchik (1965) are Destruction vs. Protection, Incorporation vs. Rejection, Reproduction vs. Deprivation, and Exploration vs. Orientation.

The Subjective Assessment of Personality in Nonhuman Primates

Based on Plutchik's (1962) psychoevolutionary theory of emotion, Kellerman and Plutchik (1968) and Plutchik and Kellerman (1974) developed the Emotions Profile Index (EPI). The EPI is a personality test for humans consisting of 12 personality trait terms that are paired in all possible combinations (66 pairs). Judges are asked to choose one word from each pair of words that most closely resembles the individual in question. The forced choices are then scored in terms of Plutchik's eight primary emotional dimensions. A mean rating for each judge and for each primary emotion is thus collected for each subject, thereby forming an emotion profile for each individual.

The EPI and Plutchik's (1962) psychoevolutionary theory of emotions have been used as the basis for the study of personality in nonhuman animals such as Dolphins (*Tursiops truncatus*) (Kellerman, 1966), Olive baboons (*Papio anubis*) (Buirski et al., 1973), Chimpanzees (*Pan troglodytes schweinfurthi*) (Buirski, Plutchik & Kellerman, 1978), Hamadryas baboons (*Papio hamadryas*), Japanese macaques (*Macaca fuscata*) (Martau, Caine & Candland, 1985), and two species of Squirrel monkey (*Saimiri* spp.) (Martau et al., 1985).

Buirski et al. (1973) and Buirski et al. (1978) developed an emotion rating instrument comprised of a list of behaviorally defined adjectives. The

instrument and its interpretation are based on Plutchik's psychoevolutionary theory of emotion (Plutchik, 1980). The application of this instrument to Olive baboons (*Papio anubis*) (Buirski et al., 1973) and Chimpanzees (*Pan troglodytes schweinfurthi*) (Buirski et al., 1978) revealed gender, dominance, and personality differences in the subject's emotions profile, and demonstrated a relationship between behavior, dominance rank and personality. Using the EPI on Hamadryas baboons (*Papio hamadryas*), Japanese macaques (*Macaca fuscata*), and two species of Squirrel monkey (*Saimiri* spp.), Martau et al. (1985) were able to demonstrate the stability of personality over a one year period for the macaques. They also discussed the need to separate the degree of familiarity with the subject animals from the degree of collaboration among raters in order to discover the relative importance of these related variables. The stability of personality over time was assessed by EPI ratings of two different sets of five student observers taken one year apart for the macaques. Ratings remained stable for all of the 14 animals but one.

The second major assessment instrument for nonhuman primate personality was developed by Stevenson-Hinde and Zunz (1978). Observers were asked to rate individual rhesus on a seven-point scale using a behaviorally defined list of adjectives. The original list of adjectives was based on the descriptions obtained from observers who were routinely recording Rhesus macaque behavior in colony groups in 1972. The ratings were organized and interpreted by means of principal component analysis. Studies using Plutchik's (1962) theory and EPI also used adjectives that were behaviorally defined, but their interpretation rested heavily on Plutchik's psychoevolutionary theory. The method used by Stevenson-Hinde and Zunz circumvents this problem by basing their organization and interpretation of data on the results of factor analyses. As Stevenson-Hinde and Zunz noted, principal component analysis is independent of any personality theory, and has been previously used atheoretically to assess individual differences in the behaviors of Rhesus macaques (Chamove et al., 1972).

Stevenson-Hinde and Zunz (1978) and Stevenson-Hinde, Stillwell-Barnes, and Zunz (1980) studied the personality profiles of six single-male, multi-female groups of captive Rhesus macaques over a four year period (1974-1977). Three principal components were identified using this method: (a) confident to fearful, (b) active to slow, and (c) sociable to solitary. In an extension of this study, Stevenson-Hinde et al. (1980) reported that: (a) some features of the macaque's personalities remained stable across four years, (b) a number of gender and age related differences in personality occurred, (c) some measures of social behavior correlated

significantly with personality scores, and (d) correlations between mother and offspring personality items were relatively high.

Caine, Earle, & Reite (1983) used a modified version of Stevenson-Hinde and Zunz's (1978) procedure to rate the personalities of adolescent Pigtailed macaques (*Macaca nemestrina*). The rating process was simplified by adopting a three-point rather than a seven-point scale, and the results were analyzed with simple correlations rather than the more complex statistical procedure of principal component analysis. Caine et al. reported correlations between dominance rank and personality, and supported the Stevenson-Hinde and Zunz categorization of the three personality components with dominance measures.

Like Buirski et al. (1973, 1978) the present study uses Plutchik's (1962) psychoevolutionary theory as its philosophical foundation, but unlike Buirski et al., it does not rely on Plutchik's theory for a methodological and interpretive framework. Like Stevenson-Hinde and Zunz (1978) and Stevenson-Hinde et al. (1980), the present study utilizes a simple list of traits, which are then organized according to their naturally occurring relationships. The present study also utilizes the results of the Stevenson-Hinde and Zunz and Stevenson-Hinde et al. analyses for the organization of personality items into personality scales. To test the generality of these factor scales across different taxa, the same organization of items was used for personality assessments of the Zebra finches as for the Stumptail macaques.

The Application of Generalizability Theory to Common Factor Modeling

Unfortunately, the use of exploratory factor models, such as principal components analysis, for small samples of subjects — such as those typically studied in this research area — remains controversial (cf., Gorsuch, 1983; Nunnally, 1978). Moreover, exploratory factor models, such as principal components, produce latent constructs that are atheoretically derived from sample data and capitalize on chance associations among variables. These threats to factor interpretation are further aggravated when the samples of subjects are small. This article describes the application of generalizability theory to confirmatory common factor modeling where the number of cases is small. Furthermore, generalizability analysis permits us to assess these personality factors simultaneously for temporal stability and interrater reliability, as well as for construct validity.

The application of generalizability theory to common factor modeling was originally suggested by Cronbach et al. (1972) and reviewed in a more recent discussion by Shavelson, Webb, & Rowley (1989), on the

generalizability of multivariate profiles. According to this perspective, convergent validity (cf., Campbell & Fiske, 1959) is no more than the generalizability of a latent variable score across a variety of alternative indicator variable scores that are, presumably, at least partially intersubstitutable measures of the same construct. This reduces much of psychometrics, including the study of validity as well as that of reliability, to different *facets* of generalizability. To accomplish such a grand unification, however, one must possess the appropriate algorithm for estimating common factor scores from the manifest indicators. This presupposes a prior factor analysis. Nevertheless, this article considers the analysis of two data sets for which there is no possibility of a traditional R-type factor analysis: For the same list of 21 personality items, one data set contains only 13 subjects and the other contains only 5 subjects.

The fundamental reason that factor analysis requires so many cases is that it requires a large number of parameter estimates. In exploratory factor analysis, the number of parameter estimates is at least equal to the number of indicators *times* the number of common factors, because every factor is allowed a loading on every indicator. In confirmatory factor analysis, even assuming perfect factorial simplicity, the number of parameter estimates is at least equal to the number of indicators, corresponding to one factor loading for each indicator. Both models are thought to require a minimum number of cases equal to some multiple of the number of parameter estimates required. The lowest multiple that has been seriously proposed is probably that of 5 cases per parameter estimate for confirmatory factor modeling (Bentler, 1989). To our knowledge, no one has proposed a multiple of less than 1.0, such as our ratios of 0.62 and 0.24, representing many fewer cases than indicators. On the other hand, it is reasonable to question the practical need for so many parameter estimates. Numerous empirical studies, as well as Monte Carlo simulations, have shown that *unit weighted* factor scores, in which all significant indicators are weighted equally (i.e., 1.0), possess the following desirable characteristics: (a) they are typically correlated about 0.95 to *differentially weighted* factor scores, (b) they are more generalizable across independent samples, and (c) they are considerably easier to calculate (Gorsuch, 1983). Once the factor model has been correctly specified (i.e., once the significant indicators have been identified), the principal function of differentially estimated factor loadings appears to be the estimation of differentially weighted factor scores. Thus, one may question why they are needed at all.

The reason for this doubt is that, even with what are normally considered *adequate* sample sizes, the standard errors for the different factor loadings are typically so large (i.e., the estimates are so *unstable* or sample-specific)

that it is seldom possible to discriminate, with any satisfactory degree of confidence, any more than between the *large* ones and the *small* ones, also called the “salient” and the “hyperplane” loadings (Gorsuch, 1983). As sample sizes get smaller, and standard errors progressively larger, the correlation between estimates provided by *unit weighting* and *differential weighting* becomes progressively higher, ultimately due to the increasing hopelessness of discriminating between differentially estimated factor loadings. It therefore stands to reason that one should, perhaps, not even attempt to estimate any differential loadings with smaller samples. If this logic is valid, it implies that factor analysis with small samples (and, perhaps, with many samples formerly considered sufficiently large) need only support a *single* parameter estimate, assumed to be equal across all prespecified indicators. This model simplification, occasioned by *inadequate* sample sizes, reduces the need for data dramatically.

Only two conditions need be met for this: (a) the factor model must be theoretically prespecified (i.e., confirmatory), and (b) the indicators must either be originally expressed in (as in these personality ratings), or subsequently transformed into, a common metric (e.g., Z-scores). If these two basic conditions are met, the estimated factor score can be reduced to the simple arithmetic mean of the specified indicator scores. This means that common factors can be constructed as *grouping* variables in an analysis of variance, within which the indicators are *nested*, like “subjects” within experimental “groups”. If desired, an *F*-ratio can be constructed to test this model for significance, comparing the variance *between* common factors to the variance *within* common factors (i.e., between different indicators of the same construct). The proportion of variance between indicators which is accounted for by common factors, as grouping variables, becomes our single estimate of commonality (i.e., of *common factor variance*), the residual (i.e., the *error* term) becomes our single estimate of the unique variance of the indicators, and the square root of this commonality becomes our single factor loading. The eta-squared representing the commonality is thus comparable to the generalizability of common factor scores across manifest indicators, completing the equivalence between generalizability theory and common factor modeling. As an added bonus, the greater generalizability of *unit weighted* factor scores, mentioned above, also dovetails quite fortuitously with the demands of generalizability analysis across other dimensions.

In essence, this procedure is identical to those described for scale construction in classical texts (e.g., Wiggins, 1973). The major difference is that the procedures we describe are completely confirmatory. For example, no items were eliminated empirically after the inspection of the item-total

correlations. Instead of traditional exploratory correlational techniques, GT-based criteria are proposed for evaluating the statistical fit of the theoretically-specified model to the data as an a priori hypothesis. This represents another functional analogy of this analytical strategy to that of confirmatory factor analysis. As in the case of confirmatory factor analysis, empirical respecifications of the model are always possible, but these are fraught with potential dangers on which much has been written elsewhere. In the case of the smaller sample sizes presently at issue, empirical selection of items should probably not be recommended.

Of course, this analytical strategy is only an approximation for smaller sample sizes and presumably should be abandoned when more precise parameter estimation becomes possible. In this article, however, we will examine the application of this small-sample strategy to personality data collected on: (a) 13 subjects over 6 years by 22 different raters; and (b) 5 subjects over 1 year by 7 different raters. Both of these studies used a list of 21 items assigned to 3 common factors on the basis of an a priori factor model (Cox, 1989). In this study, the hypothesized personality factors we constructed accounted for over 60% of the item variance. For a merely approximate solution, this is a respectable performance. Presumably, a differentially weighted factor model, based on sample-specific parameter estimates, would have fit the data even better. Of course, this better-fitting model might have produced lower estimates of generalizability across alternative dimensions.

Methods

Study 1: The Stumptail Macaque

Subjects

A Stumptailed macaque colony at the University of California, Riverside, was established in 1968 (Rhine, 1972). The Stumptails were imported from Thailand and were sexually mature upon arrival (Harvey & Rhine, 1983). The original groups, their housing, and their social behavior were described by Rhine (1972) and by Rhine and Kronenwetter (1972). Over the 21 years of its existence, the Riverside colony has varied in size due to births, deaths, and the culling of some adults and many immatures to prevent overcrowding (Harvey & Rhine, 1983). At the time this study was completed, the colony consisted of five of the original females and eight of their adult offspring: four mature adult males, two mature adult females, one young adult female, and one third generation young adult female.

Procedures

A modified version of the 1978 Stevenson-Hinde and Zunz list of personality traits was used to rate all adult macaques in the colony. Ratings were taken in 1980, 1981, 1982, 1983, 1987, and 1988. The number of raters varied across the years, totalling 22 raters over all six years. Prior to rating the animals on personality traits, each observer was given the list of traits with definitions. The instructions stated that each animal and behavior were to be scored independently of each other, and each animal was to be given a 1-7 score with 1 being a rating of extreme antithesis to the behavior and 7 the extreme manifestation of the behavior, with the intervening ratings indicating a sequentially stronger or weaker personality item rating. Raters independently rated each of the colony members, basing their decisions on observational experience with each subject. Raters had several months to years of experience doing observations of Stumptail macaque behavior.

The 21 items were assigned a priori to three common factors. Factor 1, "Confident", was measured by the items *aggressive* (+), *apprehensive* (-), *confident* (+), *effective* (+), *fearful* (-), *insecure* (-), *popular* (+), *strong* (+), *subordinate* (+), and *tense* (-). Factor 2, "Excitable", was measured by the items *active* (+), *curious* (+), *equable* (-), *excitable* (+), and *slow* (-). Factor 3, "Sociable", was measured by the items *eccentric* (-), *opportunistic* (+), *playful* (+), *protective* (+), *sociable* (+), and *solitary* (-). The negatively loaded items were reverse coded for this analysis. Table 1 (next page) lists the verbal definitions of each item given to the raters.

Statistical Analyses

A hierarchical general linear model was constructed to orthogonalize statistically the correlated facets of the unbalanced sampling design (using SAS, 1990, PROC GLM, TYPE I SS). The following sequential order was rationally derived: F = 3 Factors, I(F) = 21 Items Within Factors, S = 13 Subjects, Y = 6 Years, R = 22 Raters.

Factors were treated as fixed facets, and all others as random facets. The interaction of Subjects with Factors, representing individual differences in factor scores, was the focal effect in this study. This is somewhat unusual in that the focal effect in most generalizability studies are the main effects of variables rather than the interactions. The main effect of Factors, however, would be averaged across all subjects and, thus, estimate a parameter of little substantive importance in the study of individual differences. The interactions of {Items, Years, and Raters} with {Subjects, Factors, and Subjects \times Factors} were estimated as alternative *error* terms for the relative

Table 1
Definitions for Modified Stevenson-Hinde and Zunz (1978) Personality Ratings

<i>Active</i>	Moves about a lot.
<i>Aggressive</i>	Causes harm or potential harm.
<i>Apprehensive</i>	Seems to be anxious about everything; fears and avoids any kind of risk.
<i>Confident</i>	Behaves in a positive, assured manner, not restrained or tentative.
<i>Curious</i>	Readily explores new situations.
<i>Eccentric</i>	Shows stereotypes or unusual mannerism.
<i>Effective</i>	Gets own way; can control others.
<i>Equable</i>	Reacts to others in an even, calm way; is not easily disturbed.
<i>Excitable</i>	Over-reacts to any change.
<i>Fearful</i>	Fear grins; retreats readily from others or from outside disturbances.
<i>Insecure</i>	Hesitates to act alone; seeks reassurance from others.
<i>Opportunistic</i>	Seizes a chance as soon as it arises.
<i>Playful</i>	Initiates play and joins in when play is solicited.
<i>Popular</i>	Is sought out as a companion by others.
<i>Protective</i>	Prevents harm or possible harm to others.
<i>Slow</i>	Moves and sits in a relaxed manner; moves slowly and deliberately, not easily hurried.
<i>Sociable</i>	Seeks the company of others.
<i>Solitary</i>	Spends time alone.
<i>Strong</i>	Depends upon sturdiness and muscular strength.
<i>Subordinate</i>	Gives in readily to others; submits easily.
<i>Tense</i>	Shows restraint in posture and movement; carries the body stiffly, which suggests a shrinking tendency, as if trying to pull back and be less conspicuous.

generalizabilities of this focal effect. It made sense to test the validities of the items within common factors *first*, the stabilities of the factors over the years *second*, and the reliabilities of the factors over different raters *third*. This is because: (a) if the hypothesized factors did not capture a significant proportion of the item variance, there would be no sense in proceeding much

further with the analysis, and (b) the possibility of objective (*true score*) developmental changes in individual personalities over the years should take theoretical precedence over that of any subjective (*error*) effects of changing raters.

Because it was not possible to estimate all of the multi-way interactions between three of the random facets of this sampling design, the full $21(3) \times 13 \times 6 \times 22$ factorial analysis of variance was not performed. Specifically, the various interactions of {Items, Years, and Raters} with *each other* were all assumed to be statistically nonsignificant. Thus, they were not separately estimated in the model, but pooled into a common residual. This assumption implies, for example, that the item interrater reliabilities were assumed not to vary significantly between years. Any such variability would be interpreted in this model as purely random error. In contrast, the various interactions of Subjects with the other facets, representing systematic individual differences, were hypotheses of major substantive importance. Thus, the reduced model that was run only estimated the interactions that were essential to answering our three major hypotheses regarding the relative strength of generalizations on the hypothesized personality factors, respectively, across the following alternative dimensions: (a) Items, (b) Years, and (c) Raters.

Unlike Classical Test Theory, which compares a theoretically unitary *true score* to a homogeneous *error* term, Generalizability Theory (GT) recognizes *multiple* facets across which one may wish to generalize, called “random” facets, requiring alternative hypothesis tests and corresponding parameter estimates for the relative strengths of any generalizations across these multiple random facets. GT generalizability coefficients, based on estimated variance components, were obtained by the following equations (Shavelson et al., 1989), where *f* is the “focal” facet and *r* is the “random” facet:

$$E^2_{\text{rel}} = \sigma^2_f / (\sigma^2_f + \sigma^2_{\text{rel}*})$$

* if *r* is *nested* within *f*: $\sigma^2_{\text{rel}} = \sigma^2_{r(f)}$
 * if *r* is *crossed* with *f*: $\sigma^2_{\text{rel}} = \sigma^2_{r \times f}$

For example, the classical repeated measures analysis of variance can be viewed as a special case of a random effects generalizability model for generalizing the effects of an experimental treatment on each subject across a random sample of subjects. In the traditional within-subjects design, *treatment* is typically modeled as a fixed effect and *subjects* as a random effect. The treatment effect is therefore tested against the *treatment* ×

subjects interaction, representing the variance in treatment effect across multiple subjects. The generalizability of the treatment effect across subjects would thus represent the proportion of treatment effect that was *invariant* between subjects, although this fascinating parameter estimate is not normally computed by traditional users of repeated measures analysis of variance. More mathematical treatments of this subject are presented in the textbooks, involving the *intrusions* of the various *expected mean squares* generated by the crossing of random facets with fixed. In contrast, GT theorists eschew tests of significance in favor of the estimation of variance component parameters. This is an eminently reasonable position because failing to reject the null hypothesis more often betokens a lack of sufficient statistical power than the absence of an effect in the real world (cf., Cohen, 1990; Meehl, 1978).

Study 2: The Zebra Finch

Subjects

A research colony of captive Zebra finches was established in 1988, 7.25 km northwest of Tucson, Arizona. Their housing and social behavior were described by Figueredo et al. (1992). Over the 4 years of its existence, the research colony has varied in size due to births, deaths, and periodic removal of new fledglings to prevent overcrowding. By rater consensus, 5 subjects (2 adult males and 3 adult females) were selected from the colony for study, on which all the raters felt they had sufficient observational experience (at least 15 hours) to perform subjective assessments of their personalities.

Procedures

As in Study 1, a modified version of the 1978 Stevenson-Hinde and Zunz list of personality traits was used to rate all subjects in the study. Raters independently rated each of the 5 selected colony members, basing their decisions on observational experience with each subject. The 7 raters that were used had several months to several years of experience doing observations of Zebra finch behavior. As in Study 1, for the Stumptail macaques, the 21 items were assigned a priori to the same three common factors. Thus, both the standard instructions and the verbal definitions listed on Table 1 for the Stumptail macaque ratings were also used for the Zebra finch ratings. All ratings of Zebra finch personalities were done in 1992. None of the raters who participated in Study 2 were those who had previously participated in Study 1.

Statistical Analyses

A hierarchical general linear model was constructed using SAS (1990) PROC GLM (TYPE I SS). Although this design was otherwise fully balanced, the uneven nesting of items within factors violated strict orthogonality. Thus, the same rationally-derived sequential order used in the analysis of the Stumptail macaque data was followed for entering variables into the equation: F = 3 Factors, I(F) = 21 Items Within Factors, S = 5 Subjects, R = 7 Raters.

As in the previous study, factors were treated as fixed facets, and all others as random facets (Shavelson et al., 1989). As before, the interaction of Subjects with Factors, representing individual differences in factor scores, was the focal effect in this study; the interactions of Items and Raters with Subjects \times Factors were estimated as alternative *error* terms for the relative generalizabilities of this focal effect.

Results

Study 1: The Stumptail Macaque

Table 2 (next page) is a full breakdown of the proportions of variance estimated in the model, both as *semipartials* of the total and as *partials* of either rater or nonrater effects, respectively. This table is designed to show the breakdown of interrater reliability across multiple *true score* facets. Note that a conservative estimate of the Classical Test Theory reliability can be obtained by summing the semipartial eta-squareds for the estimated nonrater effects, indicating that at least .503 of the total is *true score* variance, which is a respectable proportion in the field of personality assessment.

Table 3 displays the estimated variance components (obtained through SAS, 1990, PROC VARCOMP METHOD=TYPE1). Negative variance estimates are statistical artifacts that may be obtained when the true values of the parameters are very close to zero (Shavelson et al., 1989). Alternative estimation procedures have been developed to avoid such statistical artifacts, including reweighted maximum likelihood (METHOD=REML), but these alternatives do not accommodate an unbalanced sampling design.

The four critical variance components are, $\sigma^2(F \times S)$, $\sigma^2[I(F) \times S]$, $\sigma^2(F \times S \times Y)$, and $\sigma^2(F \times S \times R)$. These estimated variance components can, in turn, be used to construct our GT generalizability coefficients (Shavelson

Table 2

Hierarchical General Linear Model for Generalizability Analysis of Stumptail Macaque Personality Factors

Source	<i>DF</i>	$\text{ETA}^2_{\text{SEMIPARTIAL}}$	$\text{ETA}^2_{\text{PARTIAL}}$
F	2	.001	.002
I(F)	18	.002	.004
S	12	.231	.459
S × F	24	.115	.229
S × I(F)	216	.075	.149
Y	5	.001	.002
F × Y	10	.001	.002
S × Y	52	.041	.082
S × F × Y	104	.036	.072
<hr/>			
SUBTOTAL _{NONRATER}	443	.503	1.000
<hr/>			
R	20	.001	.009
F × R	40	.001	.009
S × R	213	.048	.432
S × F × R	426	.061	.550
<hr/>			
SUBTOTAL _{RATER}	699	.111	1.000
<hr/>			
MODEL _{RATER+NONRATER}	1142	.614	
RESIDUAL	6690	.386	
TOTAL	7832	1.000	

Table 3

Estimated Variance Components for Stumptail Macaques

Component	Estimate
$\sigma^2[I(F)]$	-0.015
$\sigma^2(S)$	0.538
$\sigma^2(F \times S)$	0.518
$\sigma^2[I(F) \times S]$	0.245
$\sigma^2(Y)$	-0.002
$\sigma^2(F \times Y)$	-0.013
$\sigma^2(S \times Y)$	0.077
$\sigma^2(F \times S \times Y)$	0.152
$\sigma^2(R)$	-0.006
$\sigma^2(F \times R)$	-0.026
$\sigma^2(S \times R)$	0.065
$\sigma^2(F \times S \times R)$	0.249
$\sigma^2(ERROR)$	1.397

et al., 1989). For the focal effect in this study (Subjects \times Factors), across each of the following random facets, these GT coefficients were as shown in Table 4.

In this study, human raters were partially confounded with the years that they participated in the project. Because certain raters participated for multiple years, the apparent temporal stability of subject ratings might have reflected the consistency of those raters rather than the stable personalities of the subjects. To control for this possibility, the model was respecified to

Table 4

Generalizability Coefficients for Stumptail Macaques

Facet	E^2_{rel}
Items	.679
Years	.772
Raters	.675

assign hierarchical causal priority to the main effects and interactions of Raters with respect to those of Years. In the interests of brevity, the full breakdown of observed proportions of variance and estimated variance components for this alternative model are not shown. Instead, the alternative GT generalizability coefficients are reported for comparison. For the focal effect in this study (Subjects \times Factors), across each of the following random facets, these GT coefficients were as shown in Table 5.

These results indicate that statistically controlling for rater effects actually *enhances*, rather than reduces, the estimated temporal stability of the personality ratings by removing a source of extraneous noise in the data. Thus, the occasional repeated participation of some of the human raters over consecutive years does not represent a threat to the basic finding of temporal stability of personality factors by inflating this estimate.

For purposes of illustration, Table 6 presents the mean personality factor scores of all 13 monkeys, averaged across all raters.

Study 2: The Zebra Finch

Table 7 (see page 188) is a full breakdown of the proportions of variance estimated in the model, both as *semipartials* of the total and as *partials* of either rater or nonrater effects, respectively. As in Study 1, this table is designed to show the partitioning of interrater reliability across multiple *true score* facets. As in Study 1, a conservative estimate of the Classical Test Theory reliability can be obtained by summing the semipartial eta-squareds for the estimated nonrater effects, indicating that at least .419 of the total is *true score* variance. Also as in Study 1, this is a respectable proportion in the field of personality assessment. Note that the Rater effects — the consistent differences between the raters — are slightly smaller than the Subject effects — the systematic differences between the individual birds. Also, the residual term is small, indicating that most of the variance in the ratings is dependable.

Table 5
Alternative Generalizability Coefficients for Stumptail Macaques

Facet	E^2_{rel}
Items	.682
Years	.648
Raters	.824

Table 6

Mean Personality Factor Scores for Individual Stumptail Macaques

Subject	Sex	Confident	Excitable	Sociable
DIANE	♀	3.792	4.523	4.321
EMMA	♀	4.747	4.038	4.417
GAIL	♀	3.563	3.874	4.204
HEATHER	♀	3.381	3.325	3.859
IVAN	♂	3.991	3.019	3.542
JOAN	♀	2.663	3.694	3.052
LOIS	♀	1.625	4.450	2.286
MARIA	♀	5.238	3.506	4.380
NED	♂	3.266	4.519	3.531
PAUL	♂	6.350	4.600	4.974
QUEEN	♀	3.188	4.250	3.979
SAM	♂	5.634	3.688	4.198
ZARIA	♀	5.431	5.723	5.397

Table 8 (see page 189) displays the estimated variance components (obtained through SAS, 1990, PROC VARCOMP METHOD=TYPE1). Although the present sampling design was balanced for ratings across individuals, the previous Stumptail macaque study was not. Thus, identical estimation procedures were used to assure comparability between constructive replications.

The three critical variance components are, again, $\sigma^2(F \times S)$, $\sigma^2[I(F) \times S]$, and $\sigma^2(F \times S \times R)$. These estimated variance components can, in turn, be used to construct our GT generalizability coefficients (Shavelson et al., 1989). For the focal effect in this study (Subjects \times Factors), across each of the following random facets, these GT coefficients were as shown in Table 9 (following Table 8).

Table 7

Hierarchical General Linear Model for Generalizability Analysis of Zebra Finch Personality Factors

Source	<i>DF</i>	$\text{ETA}^2_{\text{SEMIPARTIAL}}$	$\text{ETA}^2_{\text{PARTIAL}}$
F	2	.006	.014
I(F)	18	.059	.141
S	4	.244	.582
S × F	8	.058	.138
S × I(F)	72	.052	.124
SUBTOTAL _{NONRATER}	104	.419	1.000
R	6	.032	.086
F × R	12	.016	.043
I(F) × R	108	.155	.420
S × R	24	.115	.312
S × F × R	48	.051	.138
SUBTOTAL _{RATER}	198	.369	1.000
MODEL _{RATER+NONRATER}	302	.789	
RESIDUAL	432	.211	
TOTAL	734	1.000	

Table 8
 Estimated Variance Components for Zebra Finches

Component	Estimate
$\sigma^2[I(F)]$	0.121
$\sigma^2(S)$	0.874
$\sigma^2(F \times S)$	0.333
$\sigma^2[I(F) \times S]$	0.087
$\sigma^2(R)$	0.010
$\sigma^2(F \times R)$	-0.055
$\sigma^2[I(F) \times S]$	0.497
$\sigma^2(S \times R)$	0.453
$\sigma^2(F \times S \times R)$	0.227
$\sigma^2(ERROR)$	1.272

Table 9
 Generalizability Coefficients for Zebra Finches

FACET	E^2_{rel}
ITEMS	.793
RATERS	.595

For purposes of illustration, Table 10 presents the mean personality factor scores of all 5 birds, averaged across all raters.

Discussion

These generalizability coefficients are quite high. Although they might seem low compared to the interobserver reliabilities typically obtained for directly observed behaviors, they are unusually high for subjective ratings of indirectly inferred latent traits. As predicted by theorists like Epstein (1979, 1980, 1983), the relative strength of these numbers probably resides in the degree of data aggregation possible with this kind of longitudinal data base. This technique might, therefore, be most effective where the number of *cases*, or individual organisms, is small, but the number of actual observations, including repeated measures, is quite high, that is, in “intensive” rather than “extensive” research designs (Kraemer, 1978).

The implications of these results can be classified into two categories: (a) ethological, and (b) methodological. The first of the ethological implications is that Stumptail macaques do, indeed, manifest measurable personality factors that are highly valid across multiple items, stable across multiple years, and reliable across multiple raters. This ethological implication, although well-supported by these statistical results, cannot be proven by statistical results alone. For example, the *validity* investigated here is convergent validity, which is only one form of that concept. A more detailed treatment of the substantive and design considerations needed to support such claims is presented in the original work (Cox, 1989). The second ethological implication is that Zebra finches, like Stumptail macaques, manifest measurable personality factors that are highly valid

Table 10
Mean Personality Factor Scores for Individual Zebra Finches

Subject	Sex	Confident	Excitable	Sociable
EAGLE	♂	6.214	4.943	4.905
SEAL	♂	5.100	4.743	5.000
MIESKEIT	♀	2.471	4.314	2.595
SCHMUTZ	♀	3.386	4.229	3.672
FLASH	♀	4.757	4.714	4.595

across multiple items. Indeed, they appear to be the very same factors, as predicted by our theory. If anything, the hypothesized factors work slightly *better* for the Zebra finches than for the Stumptail macaques. On the other hand, the individual factor scores for Zebra finches are *much* less reliable across multiple raters than those for the Stumptail macaques.

It is an interesting property of generalizability analysis that the validity of a construct can thus appear to exceed the reliability. That is because the GT validity is computed using the aggregated *mean* scores of multiple raters, in which individual rater effects tend to cancel each other out. It is the validity of the *aggregate* rating that exceeds the reliability of any *single* rating, but not the correspondingly higher reliability of the aggregate itself. This is the same result that would be predicted using the Spearman-Brown formula of Classical Test Theory. Another interesting outcome of these analyses is that the basic findings of Stevenson-Hinde and Zunz (1978) for the factor structure of subjective personality assessments in nonhuman animals were essentially cross-validated by converging operations on two independent samples of subjects from two very distantly related species. This is in spite of the joint limitations of their low sample size and their questionable application of exploratory principal components. Although such a replication might have happened by fortuitous coincidence, this practical confirmation of their basic model might instead be construed to mitigate against any premature or blanket condemnation of such applications.

This study was aimed primarily at demonstrating the usefulness of latent personality constructs in nonhuman animals. Because this field of research is relatively new and unexplored, we do not suggest that the particular constructs tested in this article are the ultimate dimensions of personality in either monkeys, birds, or other animals, with any delusions of finality. We merely propose that these factors possess sufficient verisimilitude to warrant further scientific study and that they have shown sufficiently acceptable psychometric properties of reliability and validity to lay some tentative claim to entitivity. We fully expect that future research in this area will further evolve and refine these concepts over time, as it manifestly has in the long history of human personality research. Furthermore, we have made no claims regarding any direct correspondence of these constructs to the better-known personality factors in humans. Given the degree of cross-species generality found between the widely disparate taxa presently reported, however, it is reasonable to ask in what relation these animal factors might stand to the better-understood and accepted dimensions of human personality.

To explore any such possible phylogenetic continuities, we used our existing list of animal trait adjectives to construct unit-weighting vectors for the "Big Five" human personality factors as well as for the three animal factors here reported. This was done by assigning a unit weighting of either +1, 0, or -1 to these adjectives according to their expected theoretical relation to the "Big Five" human personality factors (cf., Goldberg, 1990, 1992; Hofstee, de Raad, & Goldberg, 1992). The unit-weighting vectors thus obtained for the five human factors were then correlated with those created for the three animal factors using the same list of adjectives. All eight unit-weighting vectors obtained by this procedure are defined in Table 11. Because our trait adjectives were not originally selected to measure human personality constructs, the "Big Five" were not equally well sampled by this admittedly post hoc procedure. Nevertheless, certain rough correspondences emerged from this simple analysis. Our first animal factor, labelled "Confidence", was significantly positively correlated ($r = .679, p = .0007$) with the fourth of the human factors, labelled either positively as "Emotional Stability" or negatively as "Neuroticism", and with none of the others. Somewhat surprisingly, our second animal factor, labelled "Excitability", was not significantly correlated with any of the "Big Five" human personality factors. Our third animal factor, labelled "Sociability", was significantly positively correlated ($r = .457, p = .0372$) with the first of the human factors, labelled either "Surgency" or "Extraversion", and with none of the others.

We acknowledge that there was a danger of capitalization on chance with this procedure due to the number of bivariate significance tests performed. Nevertheless, these associations made conceptual sense to us, albeit in hindsight. Thus, it is conceivable that two of our three animal personality factors may be at least analogous to two of the "Big Five" human personality factors with some degree of specificity. We are currently conducting similar research with Chimpanzees in which we address the suggested correspondences more directly by using an expanded list of trait adjectives to better sample the human personality domain (King & Figueredo, 1994). Using our closest living relative among nonhuman animals will enable us to better explore the hypothesized phylogenetic continuities.

Finally, the principal methodological implications are that generalizability analysis can be used to test prespecified confirmatory factor models when the number of individual cases is quite small. Generalizability coefficients can be estimated that express the common factor variance among the items specified as indicators as a proportion of the total variance between those items. Our straightforward operationalization of the

Table 11
Unit-Weighting Vectors for the Stevenson-Hinde and Zunz (SHZ1 to SHZ3)
and the Human "Big Five" (HBF1 to HBF5) Personality Factors

Items	SHZ1	SHZ2	SHZ3	HBF1	HBF2	HBF3	HBF4	HBF5
<i>Aggressive</i>	+1	0	0	0	-1	0	0	0
<i>Apprehensive</i>	-1	0	0	0	0	0	-1	0
<i>Confident</i>	+1	0	0	0	0	0	+1	0
<i>Effective</i>	+1	0	0	0	0	+1	0	0
<i>Fearful</i>	-1	0	0	0	0	0	-1	0
<i>Insecure</i>	-1	0	0	0	0	0	-1	0
<i>Popular</i>	+1	0	0	+1	0	0	0	0
<i>Strong</i>	+1	0	0	0	0	0	+1	0
<i>Subordinate</i>	-1	0	0	-1	0	0	0	0
<i>Tense</i>	-1	0	0	0	0	0	-1	0
<i>Active</i>	0	+1	0	+1	0	0	0	0
<i>Curious</i>	0	+1	0	0	0	0	0	+1
<i>Equable</i>	0	-1	0	0	0	0	+1	0
<i>Excitable</i>	0	+1	0	0	0	0	-1	0
<i>Slow</i>	0	-1	0	-1	0	0	0	0
<i>Eccentric</i>	0	0	-1	0	0	-1	0	0
<i>Opportunistic</i>	0	0	+1	0	0	0	0	+1
<i>Playful</i>	0	0	+1	+1	0	0	0	0
<i>Protective</i>	0	0	+1	0	+1	0	0	0
<i>Sociable</i>	0	0	+1	+1	0	0	0	0
<i>Solitary</i>	0	0	-1	-1	0	0	0	0

Stevenson-Hinde and Zunz (1978) personality traits was greatly facilitated by their relatively simple factor structure. However, it is possible that this method can also be extended to factor structures with greater factorial complexity. The representation of the common factors as mutually exclusive *class* variables would have to be abandoned, but non-orthogonally unit-weighted vectors could be constructed to code for factors that share certain items as indicators. The statistical software that was used for these analyses (SAS, 1990, PROC VARCOMP) only accepts categorical variables for variance component estimation and, thus, does not accept such sets of numerical vectors in their place the way that more generalized linear

modeling programs do (e.g., SAS, 1990, PROC GLM). More sophisticated or dedicated statistical software packages for generalizability analysis should, in principle, be able to implement the more complex analytical designs that would be needed to estimate such models.

On the other hand, it is possible that what we are actually studying here is not the dimensionality of animal personality across different species, but merely common artifacts of the human mind. The threat of common rater biases in human personality ratings, such as those potentially attributable to either the shared semantic features of trait adjectives or to implicit theories of personality, are discussed by Carlson and Mulaik (1993). Invariances, however impressive, may be due to species-typical ways in which we *humans* tend to construct animal personalities and not in how the animals themselves are constituted. Nevertheless, the proposed existence of such generalizable anthropomorphic patterns of personality construction, per se, does not disprove the actual existence of animal personalities, but merely suggests greater caution in the interpretation of our results. Common rater artifacts might either *contribute to* or actually *detract from* the personality ratings, either inflating or deflating them depending on the relative direction of the bias, and may not be sufficient to explain the whole phenomenon. This equivocality of influence was illustrated by the paradoxical relationship between controlling for rater effects and the temporal stability of personality factors in the Stumptail macaque data. Nevertheless, these are methodological concerns that also merit further exploration.

References

- Allport, G. (1961). *Pattern and growth in personality*. New York, NY: Holt Rinehart & Winston.
- Bem, D. J. (1972). Constructing cross-situational consistencies in behavior: Some thoughts on Alker's critique of Mischel. *Journal of Personality*, 40, 17-26.
- Bentler, P. M. (1989). *EQS structural equations program manual*. Los Angeles, CA: BMDP Statistical Software.
- Block, J. (1977). Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 37-63). Hillsdale, NJ: Lawrence Erlbaum.
- Buirski, P., Kellerman, H., Plutchik, R., Weinger, R., & Buirski, N. (1973). A field study of emotions, dominance, and social behavior in a group of baboons (*Papio anubis*). *Primates*, 14(1), 67-78.
- Buirski, P., Plutchik, R., & Kellerman, H. (1978). Sex differences, dominance, and personality in the chimpanzees. *Animal Behaviour*, 26, 123-129.
- Caine, N. G., Earle, H., & Reite, M. (1983). Personality traits of adolescent pigtailed monkeys (*Macaca nemestrina*): An analysis of social rank and early separation experience. *American Journal of Primatology*, 4, 253-260.

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.
- Carlson, M. & Mulaik, S.A. (1993). Trait ratings from descriptions of behavior as mediated by components of meaning. *Multivariate Behavioral Research*, *28*, 111-159.
- Caspi, A., Bem, D. J., & Elder, G. H. (1989). Continuities and consequences of interactional styles across the life course. *Journal of Personality*, *57*(2), 375-406.
- Chamove, A. S., Eysenck, H. J., & Harlow, H. F. (1972). Personality in monkeys: Factor analysis of rhesus social behavior. *Quarterly Journal of Experimental Psychology*, *24*, 496-504.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.
- 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.
- Cronbach, L. J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York, NY: Wiley.
- Darwin, C. (1872/1965). *The expression of emotions in man and animals*. Chicago, IL: The University of Chicago Press.
- Digman, J. M. (1989). Five robust trait dimensions: Development, stability, and utility. *Journal of Personality*, *57*(2), 195-214.
- Epstein, S. (1977). Traits are alive and well. In D. Magnusson & N.S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 83-98). Hillsdale, NJ: Lawrence Erlbaum.
- 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.
- Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329-344.
- Figueredo, A. J., Petrinvich, L., & Ross, D. M. (1992). The quantitative ethology of the Zebra finch: A study in comparative psychometrics. *Multivariate Behavioral Research*, *27*(3), 413-436.
- Funder, D. C. (1983a). The "consistency" controversy and the accuracy of personality judgments. *Journal of Personality*, *51*(3), 346-359.
- Funder, D. C. (1983b). Three issues in predicting more of the people: A reply to Mischel and Peake. *Psychological Review*, *90*(3), 283-289.
- Funder, D. C. (1991). Global traits: A Neo-Allportian approach to personality. *Psychological Science*, *2*(1), 31-39.
- Funder, D. C., & Colvin, C. R. (1991). Explorations in behavioral consistency: Properties of persons, situations, and behaviors. *Journal of Personality and Social Psychology*, *60*(5), 773-794.
- Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*(2), 409-418.
- Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology*, *64*(3), 479-490.

A. Figueredo, R. Cox, and R. Rhine

- Goldberg, L. R. (1990). An alternative "Description of Personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59, 1216-1229.
- Goldberg, L. R. (1992). The development of markers of the Big-Five factor structure. *Psychological Assessment*, 4, 26-42.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48(1), 26-34.
- Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Lawrence Erlbaum.
- Harvey, N. C. & Rhine, R. J. (1983). Some reproductive parameters of Stumptailed macaques (*Macaca arctoides*). *Primates*, 24, 530-536.
- Hebb, D. O. (1946). Emotion in man and animal: An analysis of the intuitive processes of recognition. *Psychological Review*, 53(2), 88-106.
- Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and Circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, 63(1), 146-163.
- Houts, A. C., Cook, T. D., & Shadish, W. R. (1986). The person-situation debate: A critical multiplist perspective. *Journal of Personality*, 54, 52-105.
- Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, 43(1), 23-34.
- King, J. E., & Figueredo, A. J. (April, 1994). Human personality traits in zoo Chimpanzees? Paper presented at the Western Psychological Association Conference, Kona, Hawaii. Manuscript submitted for publication.
- Kraemer, H. C. (1978). Empirical choice of sampling procedures for optimal research design in the longitudinal study of primate behavior. *Primates*, 18, 825-833.
- Kellerman, H. (1966). The emotional behavior of dolphins, *Tursiops truncatus*: Implication for psychoanalysis. *International Mental Health Letter*, 8(1), 1-7.
- Kellerman, H. & Plutchik, R. (1968). Emotion-trait interrelations and the measurement of personality. *Psychological Report*, 23, 1107-1114.
- Martau, P. A., Caine, N. G., & Candland, D. K. (1985). Reliability of the emotions profile index, primate form, with *Papio hamadryas*, *Macaca fuscata*, and two *Saimiri* species. *Primates*, 26(4), 501-505.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Mischel, W. (1973). Toward a cognitive social learning conceptualization of personality. *Psychological Review*, 80, 252-238.
- Mischel, W. (1983). Alternatives in the pursuit of the predictability and consistency of persons: Stable data that yield unstable interpretations. *Journal of Personality*, 51(3), 578-604.
- Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist*, 39(4), 351-364.
- Mischel, W. & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review*, 89, 730-755.
- Nunnally, J. C. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.
- Ozer, D. J., & Reise, S. P. (1994). Personality assessment. *Annual Review of Psychology*, 45, 357-388.
- Plutchik, R. (1962). *The emotions: Facts, theories, and a new model*. New York, NY: Random House.
- Plutchik, R. (1965). What is emotion? *The Journal of Psychology*, 61, 295-303.

- Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotions: Theory, research, and experience* (pp. 3-33). New York, NY: Academic Press.
- Plutchik, R. & Kellerman, H. (1974). *Emotions profile index*. Los Angeles, CA: Western Psychological Services.
- Rhine, R. J. (1972). Changes in the social structure of two groups of Stumptail macaques (*Macaca arctoides*). *Primates*, 13(2), 181-194.
- Rhine, R. J., & Kronenwetter, C. (1972). Interaction patterns of two newly formed groups of Stumptail macaques (*Macaca arctoides*). *Primates*, 13(4), 19-33.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922-932.
- Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Macmillan.
- Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M. (1980). Subjective assessment of rhesus monkeys over four successive years. *Primates*, 21(1), 66-82.
- Stevenson-Hinde, J., & Zunz, M. (1978). Subjective assessment of individual rhesus monkeys. *Primates*, 19(3), 473-482.
- Van Hoof, J. A. R. M. (1962). Facial expressions in higher primates. *Symposium of the Zoological Society of London*, 8, 97-125.
- Van Lawick-Goodall, J. (1968). The behavior of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behavior Monograph*, 1(3), 161-311.
- Waller, N. G., & Ben-Porath, Y. S. (1987). Is it time for clinical psychology to embrace the five-factor model of personality? *American Psychologist*, 42, 887-889.
- Wiggins, J. S. (1973). *Personality and predictors: Principles of personality assessment*. Reading, MA: Addison-Wesley Publishing Company.

Accepted June, 1994.