Available online at www.sciencedirect.com
ScienceDirect
Journal of Research in Personality 42 (2008) 821–838
JOURNAL OF RESEARCH IN PERSONALITY
www.elsevier.com/locate/jrp

# Personality assessment in the Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings

Jana Uher [a,b,*], Jens B. Asendorpf [a]

[a] *Humboldt-University Berlin, Institute for Psychology, Germany*
[b] *Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany*

Available online 4 December 2007

**Abstract**

Three methods of personality assessment (behavior measures, behavior ratings, adjective ratings) were compared in 20 zoo-housed Great Apes: bonobos (*Pan paniscus*), chimpanzees (*Pan troglodytes verus*), gorillas (*Gorilla gorilla gorilla*), and orangutans (*Pongo pygmaeus abelii*). To test a new bottom-up approach, the studied trait constructs were systematically generated from the species' behavioral repertoires. The assessments were reliable, temporally stable, and showed substantial cross-method coherence. In most traits, behavior ratings mediated the relations between adjective ratings and behavior measures. Results suggest that high predictability of manifest behavior is best achieved by behavior ratings, not by adjectives. Empirical evidence for trait constructs beyond current personality models points to the necessity of broad and systematic approaches for valid inferences on a species' personality structure.


*Keywords:* Behavior prediction; Bonobo; Bottom-up approach; Chimpanzee; Construct validation; Gorilla; Orangutan; Personality; Rating; Traits

## 1. Introduction

Personality research in animals can provide illuminating insights into the phylogenetic basis of personality—and into human uniqueness. The Great Apes are among the most interesting species to study because they are genetically more closely related to humans than any other of today's species. Extending personality research to animals entails two fundamental challenges. First, we need to determine for each non-human species the domains of the most important traits (which may differ across species). Second, we need reliable and valid assessment methods tailored to the specifics of non-human samples. Both questions are explored in this study.

* Corresponding author. Address: Institut für Psychologie, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany. Fax: +49 (0)721 151 257 159.
*E-mail address:* uher@primate-personality.net (J. Uher).

0092-6566/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jrp.2007.10.004

Research on animal personality is as old as human personality research. The first animal personality ratings (Crawford, 1938) and behavioral studies (Hebb, 1949; Pavlov, 1906) were done by contemporaries of William Stern and Gordon Allport. As in humans, ratings are frequently used to assess animal personality (Gosling, 1998, 2001; King & Figueredo, 1997; Weiss, King, & Perkins, 2006). But what do they measure in animals like hyenas, chimpanzees, or orangutans? Are they useful to predict real world outcomes in these species, or is anthropomorphism simply biasing human judgments?

The existence of consensual and discriminative personality traits in humans and animals has been questioned repeatedly (Kenrick & Funder, 1988; Yerkes, 1939). For example, not the targets' internal dispositions but the raters' implicit personality theories were argued to cause rater agreement. Ratings are no doubt susceptible to bias; but bias can be minimized.

Accurate personality judgment is a result of social-cognitive processes. The targets have to express trait-relevant behavior in relevant settings. This information must be available to the perceivers, who must be able to detect and use it correctly to form accurate judgments (Realistic Accuracy Model; Funder 1995; Funder 1999). Under these constraints, this complex task can be successfully mastered. Interestingly, interrater agreement for animal targets is similar to that for human targets (Gosling, 2001).

Compared to human personality research, animal personality research seems to address the standard criteria for personality assessments less systematically and on a much more limited data base. To take primate personality research as an example, only a few studies established test–retest reliability in personality judgments (Martau, Caine, & Candland, 1985; Stevenson-Hinde, Stillwell-Barnes, & Zunz, 1980a) and in manifest behavioral differences (Hebb, 1949; Stevenson-Hinde, Stillwell-Barnes, & Zunz, 1980b; Suomi, Novak, & Well, 1996; Uher, Asendorpf, & Call, 2008). Studies on cross-situational consistency and coherence in response in primate behavior are rare (Uher et al., 2008). But first results are surprisingly similar to those in humans: stability over time is high in judgments and in observed behavior (if behavior is sufficiently aggregated), whereas consistency across situations and coherence across responses are low to moderate.

What about the validity of animal ratings? To avoid anthropomorphic bias, personality ratings must reflect attributes of the targets. Primate personality research, for example, has already taken first steps towards validation. Some exploratory studies show low to moderate associations between personality factors extracted from ratings and a larger number of manifest single behaviors (Capitanio, 1999; McGuire, Raleigh, & Pollack, 1994; Pederson, King, & Landau, 2005). But multimethod studies that systematically validate the rating lists and the underlying personality constructs are still missing. To our knowledge, no non-human primate study used nomologic networks (Cronbach & Meehl, 1955) or even multitrait–multimethod validation (Campbell & Fiske, 1959) to establish construct validity through simultaneous convergence between different methods of assessment.

Perhaps the most fundamental and difficult question is which trait constructs we should try to validate in an animal species. Animals display a wide diversity of behaviors; their ecological and social systems differ greatly. Orangutans, for example, have an extended social system and are rarely seen to form groups in the wild. They have learned to survive in swampy rain forests where food is scarce (van Schaik, 2004). Which personality traits may such as species have?

The generation of trait constructs is a critical issue in trait psychology (Carver & Scheier, 2000). It is even more crucial in studies on non-human species to which we only have limited access as humans, that is, as non-conspecific outsiders. To obtain valid inferences on a species' personality structure, trait generation for non-human species must be careful and systematic, considering as much available information about the species as possible.

In contrast to human personality research, a lexical approach fails as a systematic starting point for research on animal personality. The underlying "sedimentation hypothesis" assumes that humans perceive their most important personality traits in social interactions, and that these traits are coded in human language (Allport & Odbert, 1936; Goldberg, 1990). The validity of the lexicon's trait-related words for systematic studies on personality is thus limited to humans. There is no comparable reason to assume that humans have developed an equally systematic body of trait-related words to describe personality traits in other species with which they generally interact only little or not at all. Therefore it is also unclear whether lexically derived models of human personality are valid starting points for studies on non-human personality. Similar challenges arise in expert nominations. Experts may be more likely to nominate those traits that are salient to

human observers or that match their implicit theories of human personality and they may pay less attention to other traits. Such biases in trait generation can distort empirical outcomes. The importance of this problem is illustrated by the following empirical example from primate personality research.

Stevenson-Hinde and Zunz (1978) generated an adjective list for rhesus macaques (*Macaca mulatta*) by expert nomination revealing three underlying factors (confident, excitable, sociable). These three factors could be replicated in the same species (Stevenson-Hinde et al., 1980a), as well as in stump-tailed macaques (*Macaca arctoides*; Figueredo, Cox, & Rhine, 1995), pig-tailed macaques (*Macaca nemestrina*; Caine, Earle, & Reite, 1983), and chimpanzees (*Pan troglodytes*; Murray, 1998). Other studies—in gorillas (*Gorilla gorilla*; Gold & Maple, 1994) and again in rhesus macaques (Bolig, Price, O'Neill, & Suomi, 1992; Capitanio, 1999; Capitanio & Widaman, 2005)—yielded four factors including two or all three factors shown in the other studies. These findings seem to suggest that major personality constructs (for example sociability and excitability) are relatively universal across these species (Gosling, 2001; Gosling & John, 1999).

Other item lists, however, lead to different results. Whereas the Stevenson-Hinde and Zunz (1978) adjective list showed only three factors in chimpanzees (*Pan troglodytes*; Murray, 1998), an adjective list based on the human Five Factor Inventory repeatedly yielded six factors in this species (King & Figueredo, 1997; King & Landau, 2003; King, Weiss, & Farmer, 2005; Pederson et al., 2005; Weiss, King, & Figueredo, 2000). This six-factorial structure covered all three factors yielded by the former list. Thus, in the same species one item list yielded only half of the factor structure yielded by another item list. These differences suggest that the item lists differed in scope; the older list was incomprehensive to identify more factors. In other words, the empirical differences were not caused by the species' personality differences but only by the trait domains covered by the items. Of course, empirical conclusions should be determined as much as possible by the studied phenomena, not by methodology. Clearly, trait generation can have a significant impact on the validity of empirical results that in turn affects comparisons between species. Consequently, trait generation should be based on a broad and systematic footing (Uher, submitted for publication).

But how to generate traits systematically in animals? Research in humans can heavily rely on self-reports to identify and measure traits; but when participants are unable to provide self-reports or when social desirability bias may be too strong, traits are inferred from manifest behavior. In animals, we can only rely on trait inferences from manifest behavior. Behavior is therefore a good starting point for trait generation.

Trait theory assumes that personality traits create stable relations between situations and the individual's reactions (Funder, 2004; John & Gosling, 2000; Mischel, Shoda, & Mendoza-Denton, 2002). A trait construct thus comprises specific behavioral tendencies that are related to specific types of situations. In animals, such behavior–situation units could be derived from the species' behavioral repertoire by merging universal behaviors with typical situations members of the species encounter. Systematic trait generation should then be based upon a representative selection of a species' behaviors and relevant situational features. These ideas form the basis of the behavioral repertoire approach, a new bottom-up approach that Uher recently proposed for systematic studies on animal personality (Uher, 2005, submitted for publication).

A first step is to review a species' behavioral repertoire systematically for universal behaviors and typical related situations. A broad and systematic review of biological publications on a species' behavior and its social and ecological system constitute the foundation for such a review. Research that is explicitly not dealing with personality is particularly useful because it describes the species' general behavior independently of any previous personality research; it therefore constitutes unbiased raw material. In the review, one lists broad and universal rather than specific behaviors, and general features of typical situations rather than particular situations because increasing detail requires greater empirical efforts (especially larger samples) to identify the underlying structure. In a matrix, the listed behaviors and situational features are then merged systematically into behavior–situation units that are conceived as potential trait constructs. Data on interindividual variability and its temporal stability show which of these theoretically generated constructs are in fact traits in the studied species. Trait generation from bottom-up implies that these constructs can only be assumed to be mono-polar until empirical analyses clarify their underlying factorial structure. Personality factors that result empirically from this bottom-up procedure are most representative for the species' behavioral variability, are ecologically valid, and permit inferences on its personality structure with high validity.

The behavioural repertoire approach has been applied to the Great Ape species (Uher, 2005, submitted for publication). Extensive biological publications about these species' behaviors, their social systems and

ecologies in the wild and in captivity were reviewed, and universal behaviors and typical associated situational features were listed separately for each species. For example, ape researchers repeatedly describe responses to the animate and inanimate (non-conspecific) environment like encounters with unfamiliar objects, strange environments, uncertain situations, and threats (Gold, 1992; Lukas, Hoff, & Maple, 2003; Maple, 1980; Meder, 1993; Rijksen, 1978; Schaller, 1963; Susman, 1984; van Lawick-Goodall, 1968).

Surprisingly, all universal behaviors and situational features yielded in the literature review were strongly similar across the Great Ape species; they were therefore pooled at the end of the review. This similarity may reflect the species' phylogenetic relatedness; but it may also be due to selecting broad and universal rather than more specific behaviors and situations. Species-specific behaviors can be considered in the trait operationalizations, of course. The listed Great Ape behaviors and situations were organized into six broad domains that reflect biological behavior taxonomies: activity/ranging patterns, solitary behavior, feeding behavior, social behavior, sexual behavior, and breeding behavior. Interestingly, these broad behavior domains can be related to adaptive problems that are frequently discussed in evolutionary psychology (Buss, 1999).

By merging behaviors with situational features into behavior–situation units potential trait constructs were generated (Uher, 2005, submitted for publication). For example, detecting unfamiliar objects in the environment could be related to a trait construct labeled vigilance. Merging further behavioral responses to unfamiliar objects yielded potential trait constructs like curiosity (approach, investigation), playfulness (playing), arousability (excitement when spotting it in terms of pilo erection or scratching), anxiousness (fearful reactions like screaming or fleeing), and aggressiveness (attacking, destroying). This procedure was applied to all behaviors and situations in all behavior domains (except for behaviors that only occur in the wild like traveling or territoriality). More fine grained subtraits that involve more specific behaviors and situations (for example, arousability in social versus non-social situations) were then subsumed into broader trait constructs (arousability) to facilitate first empirical tests in a small sample. Nineteen qualitatively distinct (mono-polar) potential trait constructs emerged from the review material. Their definitions and operationalizations can be based on the behavior–situation units from which they were derived. For example, vigilance could be defined as the tendency to quickly detect novel objects, hidden food, or potential dangers; latencies and success rate in finding hidden food could be possible behavioral operationalizations in captive Great Apes.

First data on these trait constructs were obtained in a behavioral study on 20 zoo-housed Great Apes. It found substantial stability in over 70 single behavior–situation units and in aggregated trait-indices that operationalized the 19 bottom-up generated trait constructs. Cross-situational consistency and coherence in response were also highly similar to those found in humans (Uher et al., 2008).

These first results have several important implications. They show that the behavioral bottom-up approach generates trait constructs that are real and measurable in manifest behavior. This is important because the traits were operationalized by behaviors that were easily visible for trained observers and therefore minimally affected by implicit personality theories of human observers. Furthermore, situations played an important role in measuring the traits. Behavior was not measured in *ad hoc* situations, but in specific situations that were *a priori* selected to be ecologically valid and considered as trait-relevant.

Perhaps the most interesting result was that the bottom-up approach yielded traits (for example, food orientation or sexual activity) that are relevant to these species but not well represented in factor models of human personality. These findings argue for the importance of ecologically valid trait generation procedures. They also show that if trait constructs are representative for the species, then their operationalizations in manifest behavior successfully meet the standards established in human personality psychology.

Whereas behavioral measures can be directly based on the species' behavioral repertoire, operationalizations by rating items must additionally rely on the repertoire of human language. Animal personality may however vary on dimensions that are quite different from those in humans. Descriptors of human personality that suggest more human-centered interpretations of behavior or that require high inference from perceivable behavior—like adjectives—may induce rater biases, such as halo effects or anthropomorphic bias.

Adjectives are assumed to code important human traits; they are thus valid descriptors of human personality. But what do adjective ratings measure in animals?

In primate personality research, for example, many studies have tried to adapt adjectives from human personality inventories by adding a "clarifying definition" consisting of some further adjectives (McGuire et al.,

1994), or behavioral descriptions in the context of the species' behavior (e.g. Buirski, Plutchik, & Kellerman, 1978; Figueredo et al., 1995; Gold & Maple, 1994; King & Figueredo, 1997; King et al., 2005; Pederson et al., 2005; Stevenson-Hinde & Zunz, 1978; Stevenson-Hinde et al., 1980a), or both (Weiss et al., 2006). The definitions' implicit connotations may however mask those of the adjectives.

Whether adjectives or behavioral descriptions are more valid can be decided only with regard to a third criterion, and manifest behavior in trait-relevant situations seems to be the best available criterion for non-human species. Therefore, the present study compared two different kinds of items—single adjectives and behavior-descriptive verbs—in terms of their validity for the trait constructs generated by Uher (2005, submitted for publication). Both kinds of items were rated in the same sample of Great Apes studied by Uher et al. (2008) and therefore could be correlated with the trait-relevant behavior measures. Together, these three methods constitute a nomologic network around each trait construct. This design allowed us to compare the reliability and temporal stability resulting from the different methods, and to explore the empirical relations between the three methods of trait assessment.

## 2. Methods

### 2.1. Subjects

Twenty adolescent or adult Great Apes representing five each of bonobos (*Pan paniscus*), chimpanzees (*Pan troglodytes verus*), gorillas (*Gorilla gorilla gorilla*), and orangutans (*Pongo pygmaeus abilii*) were studied at the Wolfgang Köhler Primate Research Center (WKPRC) in the Leipzig Zoo, Germany, from January to March 2005. The 13 females and 7 males were 7–31 years old ($Mdn = 16$; $M = 17.3$; $SD = 8.6$; for details see Table 1). They were housed in social groups of 5–18 animals in large, naturally designed indoor and outdoor enclosures,

Table 1
Species, sex, age, and rearing history of the subjects

| Species | | | |
|---|---|---|---|
| Subject | Sex | Age (years) | Rearing history |
| Bonobo | | | |
| Joey | M | 22 | Nursery |
| Kuno | M | 8 | Nursery |
| Limbuko | M | 9 | Nursery |
| Ulindi[a] | F | 11 | Mother |
| Yasa | F | 7 | Mother |
| Chimpanzee | | | |
| Dorien[a] | F | 24 | Nursery |
| Fraukje | F | 28 | Nursery |
| Frodo | M | 11 | Mother |
| Robert | M | 29 | Nursery |
| Sandra | F | 11 | Mother |
| Gorilla | | | |
| Bebe | F | 25 | Mother/peer |
| Gorgo[a] | M | 23 | Nursery |
| Ndiki | F | 27 | Mother/peer |
| Ruby | F | 7 | Mother |
| Viringika | F | 9 | Mother/peer |
| Orangutan | | | |
| Bimbo[a] | M | 24 | Nursery |
| Dokana | F | 16 | Mother |
| Dunja | F | 31 | Nursery |
| Padana | F | 7 | Mother |
| Pini | F | 16 | Mother |

[a] Subjects dropped from data collection in the series of laboratory tests. F, female; M, male.

and in special sleeping and testing rooms each with several interlinked cages. The apes were always treated in accordance with ethical guidelines for the use of animals in research; all apes received their complete daily diets of various fresh foods and permanent access to water.

### 2.2. Study design

The trait constructs were generated bottom-up from the species' behavioral repertoires as described above (Uher, 2005, submitted for publication). To generate a nomologic network for construct validation (Cronbach & Meehl, 1955), each trait was operationalized equivalently with three different methods: with behavior measures (M), behavior ratings (B), and adjective ratings (A). Raters and observers were mutually blind to each others' assessments. Data collection was repeated completely after a break of some weeks within 50 days to analyze temporal stability.

### 2.3. Behavior measures: Tests and observations

Operationalizations were based upon the behavior–situation units that were derived bottom-up from the species' behavioral repertoires (Uher, 2005, submitted for publication). To elicit different trait facets and to study cross-situational consistency, several similar test situations were set up for most traits; curiosity for example, was measured in a novel food and a novel object test. In each situation, multiple behaviors were recorded wherever feasible to study coherence in response and to consider possible species-typicalities. The behavior categories were chosen to be applicable to all four species; only some arousal expressions tended to be species-specific. In the novel food test, for example, several curiosity-related behaviors were measured like latencies to touch novel (colored) versus normal food pieces, durations of exploring them, and the number of novel food pieces that were rejected (some apes threw them immediately out of their cage, whereas others did not).

This scheme of operationalization was applied to all traits and resulted in 76 different behavior-situation variables that were measured in 14 different laboratory based tests and in two different situations in the apes' social groups (see Table 2 for an overview of the tests and observations with situational descriptions and an index which trait constructs were measured; more details are reported in Uher et al., 2008).

To reduce the impact of behavioral fluctuations on the data, all behavioral tests were repeated 2–3 times within 15 consecutive days (except for two tests with a previously unknown potential to elicit fear). The tests were administered in a reasonably random sequence to avoid exposing the apes to similar situational features more than once a day. After a break of about a fortnight, the data collection was repeated completely following the same scheme of repetitions and randomization. Thus, most tests were ultimately repeated 4–6 times. The group observations were also repeated multiple times; within each data collection period, prefeeding behavior was observed at 5 days, the afternoon observations were done at 12 days.

Data on all apes were collected for both types of observation; four subjects per species could be tested in the laboratory. Testing and behavior recording were highly standardized in all respects. All behavioral tests and the prefeeding observation were videotaped and later coded in detail using the coding software INTERACT (Rel. 7.2.4., Mangold, 2006); only the afternoon observation was recorded online with observation sheets. In total, more than 120 h of video records were coded in detail and 240 h were additionally observed online.

### 2.4. Behavior ratings and adjective ratings: Instruments

The bottom-up generated trait constructs were operationalized with two different inventories.[1] Ratings were indicated on a 5-point Likert scale from (1) *strongly disagree* to (5) *strongly agree*. Because behavioral data for competitiveness and self-care could only be measured post hoc from video, these two traits were not included in the ratings.

[1] The German items and English translations can be obtained from Jana Uher.

Table 2
Behavioral tests and observations: Situational descriptions and measured trait constructs

| Tests and observations | Situational description | Measured trait constructs |
|---|---|---|
| Button box test | When pressing 20 buttons of a specially constructed box, rewards failed to materialize in contrast to preceding training trials | Persistency |
| Cage intruder test | The experimenter entered one of the cages next to the ape's cage and offered raisins for a limited time | Aggressiveness to humans<br>Friendliness to humans<br>Anxiousness |
| Food box test | Food of different preference and quantity was placed successively in a small transparent box, which the ape could open | Food orientation |
| Blocked food box test | The same transparent box was baited with highly preferred food, still looked the same but was blocked by a screw | Impulsiveness |
| Food competition test | Two apes faced each other across neighboring cages from where both could reach into a transparent Plexiglas tunnel, which was centrally baited with one banana piece | Dominance<br>Competitiveness |
| Hidden food test | The ape entered the test room in which small food pieces were hidden on the rims of the cage's frame or stuck with honey to the variegated walls | Vigilance<br>Physical activity |
| Honey grid test | The ape could reach through the mesh for honey spread on a Plexiglas panel outside the cage, while the experimenter knocked continuously on another panel | Distractibility |
| Keeper interaction test | A familiar keeper sat in front of the ape's cage and encouraged him to approach and to play, and fed him for a limited time | Aggressiveness to humans<br>Friendliness to humans |
| Masked human test | Disguised as a masked human, the experimenter entered the test room silently, and offered continuously food in gloved hands; she thereby stuck the right's glove stiffed fingers through the mesh into the ape's cage | Aggressiveness to humans<br>Friendliness to humans<br>Anxiousness<br>Arousability |
| Novel food test | The ape received in turn apple slices and novel (colored) food pieces | Curiosity |
| Novel object test | The ape found a small novel object inside his cage and was given time to explore it | Curiosity<br>Sexual activity |
| Pile of food test | The experimenter was cutting food in full view of the ape into small pieces, which were piled up in a bowl | Arousability<br>Impulsiveness |
| Food out of reach test | An amount of food was placed in front of the ape but still out of his reach on a table, while the experimenter either sat next to it doing nothing or left the room | Impulsiveness |
| Sudden noise test | A foreign news program was suddenly played back to the ape in moderate volume independent of the experimenter's activities inside the test room for half a minute | Aggressiveness<br>Anxiousness<br>Arousability |
| Prefeeding observation | The apes could hear and see the keepers approaching right before the afternoon feeding | Arousability<br>Sexual activity |
| Afternoon observation | The apes were in their social groups in the spacious and naturally designed indoor enclosures equipped with enrichment boxes | Friendliness to conspecifics<br>Friendliness to youngsters<br>Food orientation<br>Gregariousness<br>Persistency<br>Physical activity<br>Playfulness |

*Note*. More details are reported in Uher et al. (2008).

### 2.4.1. Great Ape Personality Inventory-Behaviors (GAPI-B)

For behavior ratings, traits were operationalized with descriptions of observable, trait-indicating behaviors in circumscribed situations using verbs only. These items require a lower degree of inference from perceivable behavior than adjectives. The descriptions corresponded to some of the behaviors measured either in the group observations or in the laboratory tests. For example, curiosity was operationalized with "[Animal's name] often touches new objects (e.g. enrichment items) at great length.", which corresponded to the duration

of exploring in the novel object test, and with "Confronted with novel food, [animal's name] mostly ignores it.", which corresponded to the number of rejected novel food pieces in the novel food test. To limit the number of items to an acceptable size (because each keeper rated 5–15 animals twice), most traits were operationalized with two items. In total, 34 behavior-descriptive items were constructed; 10 items were reversed in their meaning to reduce the possibility of response sets.

#### 2.4.2. Great Ape Personality Inventory-Adjectives (GAPI-A)

For adjective ratings, each trait was operationalized by one single adjective that best described the trait construct in the everyday language of the raters. For example, curiosity was operationalized with "[Animal's name] is very curious", or food orientation with "[Animal's name] is very gluttonous". The resulting 17 statements constitute a pool of items that require a higher degree of inference from the animals' perceivable behavior. No item was reversed in its meaning.

### 2.5. Raters

Ten keepers, two women and eight men who had been working with the target apes for 2.5–30 years ($M = 3.9$ years; $SD = 3.3$), provided the ratings. They assessed their own familiarity with these individual apes as well ($M = 4.0$; $SD = 0.7$) on a five point rating scale from (1) *not at all* to (5) *very well*. Due to specializations in their work, four keepers rated animals of just one species ($n = 5$), another four keepers rated two species ($n = 10$), and two more keepers rated three species ($n = 15$). In total, four to five raters were available for each animal.

### 2.6. Rating procedure

The keepers rated the animals twice with an interval of about five weeks ($M = 36.4$ days; $SD = 7.7$). Each time, they were asked to assess how the apes are *currently behaving*. The keepers were briefed individually about the rating procedure and cautioned not to discuss their ratings with the others. The set of all 51 items (34 behavior items and 17 adjective items) was presented on a computer screen in a fixed randomized order. The order in which the target individuals were rated was randomized between raters within each species. Ratings were scheduled in parallel to behavior recording that the keepers did not attend however (except for single keepers assisting in the keeper interaction test or the food competition test).

## 3. Results

### 3.1. Reliability

#### 3.1.1. Behavior measures

Reliability was analyzed as agreement in absolute raw data between different observers. Therefore, a second person (JK) recorded in each species all 71 raw variables in 20% of the sessions in all tests and observations independently from the first author (JU) who recorded all behavioral data. Agreement per session and subject was computed with Cronbach's $\alpha$. The reliability of the 71 single behavior-situation measures was high; $\alpha$ ranged from .71 to 1.00 with a mean and median of $\alpha = .96$ ($n = 71$; for aggregations over time, zero-one coded data were treated as metric variables).

#### 3.1.2. Behavior ratings and adjective ratings

Reliability of the ratings was analyzed as agreement between independent raters. Because each keeper rated only a part of the sample, standard $\alpha$ could not be computed because it requires non-missing scores. Instead, we first computed the intercorrelations between the four to five raters per animal, calculated mean intercorrelations (using Fisher's $r$-to-$Z$ transformation), and then estimated interrater reliability by means of the Spearman–Brown formula. We also calculated intraclass correlation coefficients for direct comparison with other studies; the reliability of single ratings is indicated by ICC(3, 1), and the mean reliability of the $k$ ratings per ape is indicated by ICC(3, $k$) (Shrout & Fleiss, 1979). All reliability scores were calculated separately for each data collection period.

The independent raters agreed substantially both on the pattern of the item means over all animals, and on the relative order of the animals on the single items; mean reliability of item means was $\alpha = .90$ for adjective ratings, and $\alpha = .92$ for behavior ratings, mean reliability of the single items was $\alpha = .83$ for adjective ratings, and $\alpha = .81$ for behavior ratings (computed with Fisher's $r$-to-$Z$ transformation). Interrater reliability 5 weeks later was virtually identical. A few items were unreliable in one of the two data collections; but because no item was unreliable at both times and because the animal sample was very small, all items were included in the subsequent analyses. Table 3 shows all reliability scores for the item means and the single items (range and mean) broken down by period of data collection and inventory.

To examine processes of personality judgment according to the Realistic Accuracy Model (Funder, 1995; Funder, 1999), we analyzed the relation between the traits' reliability scores (Cronbach's $\alpha$) and their rated availability at the Research Center with Pearson correlations. For the adjective ratings, reliability and availability correlated $r = .63$, $p < .05$ ($t_2$: $r = .55$, $p < .05$; $N = 17$); for the behavior ratings they correlated $r = .67$, $p < .001$ ($t_2$: $r = .45$, $p < .05$; $N = 34$). This shows that the keepers based their agreement in both types of ratings to a great extent on behavior observations and not only on communication about the animals.

### 3.2. Temporal stability

#### 3.2.1. Behavior measures

The temporal stability of the behavior measures between the first and the second data collection period was analyzed on different levels of aggregation with Cronbach's $\alpha$ and Pearson correlation $r$. We first aggregated the data over occasions to increase reliability (Epstein, 1979; Epstein, 1980), and analyzed the stability of all 76 single behavior-situation measures and of individual response profiles within situations. Subsequently, we aggregated the $z$-scored behavior-situation measures (some were reversed to share the same meaning) into trait-indices within situations to analyze the stability of individual situational profiles. Finally, we aggregated all $z$-scored behavior-situation measures related to a trait across situations, and analyzed the temporal stability of these aggregated trait-indices. In this study, we focus on the stabilities of single behavior-situation measures and of aggregated trait-indices; and refer to Uher et al. (2008) for individual profile stabilities and further data on coherence in response and cross-situational consistency.

Table 3
Reliability and temporal stability of behavior ratings and adjective ratings[1]

| Method | Reliability | | | | | | Temporal stability | |
|---|---|---|---|---|---|---|---|---|
| | $t_1$ | | | $t_2$ | | | $\alpha$ | $r$ |
| | $\alpha^a$ | ICC(3,1) | ICC(3,$k^b$) | $\alpha^a$ | ICC(3,1) | ICC(3,$k^b$) | | |
| *GAPI-Behaviors* ($n = 34$) | | | | | | | | |
| Item means | .92 | .74 | .93 | .93 | .73 | .93 | .98 | .97 |
| *Single items* | | | | | | | | |
| Mean | .81 | .37 | .72 | .78 | .40 | .74 | .93 | .88 |
| Min | .22 | −.41 | −.69 | .25 | −.01 | −.50 | .33 | .20 |
| Max | .98 | .90 | .98 | .94 | .81 | .96 | .99 | .98 |
| *GAPI-Adjectives* ($n = 17$) | | | | | | | | |
| Item means | .90 | .72 | .93 | .91 | .70 | .92 | .99 | .98 |
| *Single items* | | | | | | | | |
| Mean | .83 | .44 | .79 | .81 | .46 | .79 | .94 | .88 |
| Min | .25 | −.04 | −.28 | .59 | .21 | .35 | .80 | .67 |
| Max | .98 | .89 | .98 | .94 | .81 | .95 | .99 | .98 |

[a] As none of the keepers rated all animals, $\alpha$ scores were estimated with the Spearman–Brown formula from mean Pearson intercorrelations between all raters per item (using Fisher's $r$-to-$Z$ transformation). ICC(3,1) single rater reliability, ICC(3,$k$) mean reliability of $k$ ratings (Shrout & Fleiss, 1979).

[b] $k = 4$ in chimpanzees and gorillas, $k = 5$ in bonobos and orangutans. $t_1$ first, $t_2$ second data collection period. Temporal stabilities were computed using mean rating scores aggregated over keepers within each period.

Test–retest stability of the behavior measures was high at different levels of aggregation. The 76 single behavior-situation measures had a mean stability of $\alpha = .86$ (ranging from .07 to 1.00) and $r = .78$ (ranging from .03 to 1.00). The temporal stability for all 19 aggregated trait-indices was also high; mean $\alpha$ was .87 (ranging from .40 to .98); mean $r$ was .77 (ranging from .29 to .97). All means were calculated with Fisher's $r$-to-$Z$ transformation. Stability scores of all single behavior-situation measures and of the aggregated trait-indices are reported in Uher et al. (2008).

### 3.2.2. Behavior ratings and adjective ratings

The stabilities of behavior ratings and adjective ratings were computed with Cronbach's $\alpha$ and with Pearson correlation $r$ on the four to five raters' mean ratings. The temporal stability of the personality judgments was high; the mean stability of the pattern of item means over all animals was $\alpha = .98$ for behavior ratings, and $\alpha = .99$ for adjective ratings; the mean stability of the between-animal differences for each item was $\alpha = .93$ for behavior ratings, and $\alpha = .94$ for adjective ratings. All means were computed with Fisher's $r$-to-$Z$ transformation. Table 3 shows means and ranges of all stability scores for the item means and the single items broken down by period of data collection and inventory.

### 3.2.3. Comparison of stability scores across methods

The present design permitted direct comparisons between stabilities of different methods because they were obtained simultaneously, on the same traits constructs, and in the same sample. We compared stability scores between single behavior-situation measures, aggregated behavior indices, behavior ratings, and adjective ratings. A one-way ANOVA on $r$-to-$Z$-transformed stability scores showed significant differences, $F(3, 142) = 6.15$, $p < .001$. Bonferroni tests revealed that the stabilities of assessments differed significantly between method classes (manifest behavior measures versus ratings), $p < .05$, but not within, $p > .60$. We calculated the magnitude of these differences with Cohen's effect size $d$ on pooled standard deviations (Cohen, 1988). The effect sizes of stability differences between behavior ratings and single behavior-situation measures was $d = .73$, between adjective ratings and single behavior-situation measures $d = .87$, between behavior

Table 4
Coherence between behavior measures (M), behavior ratings (B), and adjective ratings (A); and mediation analyses of the behavior ratings' effect (B) on the relation between behavior measures (M) and adjective ratings (A) on the trait level

| Trait construct | Coherence[a] | | | Mediation analyses[b] | |
|---|---|---|---|---|---|
| | B–A | M–B | M–A | (B)M–A | (M)B–A |
| Aggressiveness (to humans) | .71** | .76** | .30 | −.61* | 1.20** |
| Anxiousness | .21 | .69** | .12 | −.20 | .46 |
| Arousability (general) | .11 | .28 | .18 | .17 | .06 |
| Prefeeding context | — | .60** | .07 | — | — |
| Strange situations | — | −.22 | .42# | — | — |
| Curiosity | .68** | .72** | .55* | .31 | .34 |
| Distractibility | .68** | .41 | .40 | .20 | .49# |
| Dominance | .88** | .73** | .61* | −.10 | .97** |
| Food orientation | .87** | .69** | .54* | −.11 | .94** |
| Friendliness to youngsters | .88** | .76** | .67** | −.01 | .88** |
| Friendliness to conspecifics | .00 | .63** | −.06 | −.10 | .06 |
| Friendliness to humans | .29 | .56* | .00 | −.11 | .19 |
| Gregariousness | .94** | .44# | .34 | −.09 | .98** |
| Impulsiveness | .42# | .51* | .33 | .07 | .51# |
| Persistence | .97** | .48** | .47** | .01 | .97** |
| Physical activity | .95** | .63** | .58** | −.03 | .97** |
| Playfulness | .63** | .57** | .43# | .10 | .57* |
| Sexual activity | .90** | .62** | .48** | −.13 | .98** |
| Vigilance | .69** | .33 | −.15 | −.42* | .80** |

*Note.* [a] Pearson correlations $r$; [b]standardized regression coefficients $\beta$ in multiple regression equations; $^{**}p < .01$, $^{*}p < .05$, $^{\#}p < .10$. (M) B–A: regression of adjective ratings (A) on behavior ratings (B) controlling for behavior measures (M); (B) M–A: regression of adjective ratings (A) on behavior measures (M) controlling for behavior ratings (B).

ratings and aggregated behavior indices $d = .74$, and between adjective ratings and aggregated behavior indices $d = .91$. These effect sizes are substantial and show that ratings, in particular adjective ratings, yield higher stability scores than manifest behavior measures.

### 3.3. Cross-method coherence

Coherence between behavior measures (M), behavior ratings (B), and adjective ratings (A) establishes a nomologic network around each trait construct. To increase reliability (Epstein, 1979; Epstein, 1980), data were first aggregated over time within each method: the *z*-scored trait-indices of the behavior measures were aggregated over time; the scores of the single behavior items were first *z*-scored over animals (some were reversed to share the same polarity), then aggregated within traits and over time; and the adjective item scores were also *z*-scored and aggregated over time. This procedure yielded three distribution patterns of the animals' relative order on each trait, one for each method. We analyzed coherence in these rankings across the three methods with Pearson correlations *r*.

The correlations between these assessments in 17 trait constructs were significantly different from zero in one-sample *t*-tests, $t_{M–B}(16) = 16.03$, $p < .001$; $t_{M–A}(16) = 5.76$, $p < .001$; $t_{B–A}(16) = 8.32$, $p < .001$. Table 4 lists the coherence scores for all traits constructs.

More detailed analyses on coherence between arousability assessments revealed that behavior ratings were more strongly associated with arousability in prefeeding situations (prefeeding observation, pile of food test; $r = .60$, $p < .01$), whereas adjective ratings were more strongly associated with arousability in strange situations (sudden noise test, masked human test; $r = .42$, $p < .10$; see Table 4). Given that the behavior items operationalized only prefeeding arousability, their correlation to manifest prefeeding behavior is obvious. By contrast, the correlation between the adjective item and a specific type of situation is surprising. Recall that the adjectives were presented without any additional behavioral or situational context. This suggests that in animal ratings an adjective can have an implicit connotation for raters that is not obvious from its general meaning.

The strength of coherence differed significantly between methods; coherence between behavior measures and adjective ratings (M–A) was significantly lower than both the coherence between the two ratings (B–A), $t_{MA–BA}(16) = 5.46$, $p < .001$, and the coherence between behavior measures and behavior ratings (M–B), $t_{MA–MB}(16) = 4.44$, $p < .001$. Coherence between behavior measures and behavior ratings was however not different from coherence between behavior ratings and adjective ratings, $t_{MB–BA}(16) < 1$. Fig. 1 shows the empirical relations between the three methods in mean coherence scores across 17 trait constructs (computed with Fisher's *r*-to-*Z* transformation).

### 3.4. Are adjective ratings mediated by behavior ratings?

To analyze how raters may have formed their adjective assessments, we tested whether adjective ratings are directly based on a broad range of perceivable behaviors like the behavior measures, or whether specific, trait-indicating behaviors like those described in the behavior items served as a mediator to assess the adjectives. Partial mediation would be evidenced if the manifest behavior measures would still directly affect adjective

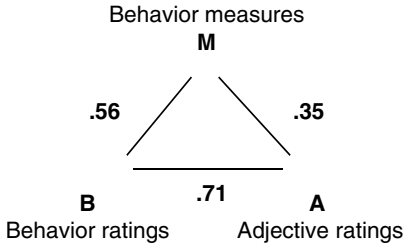


Fig. 1. Mean coherence between behavior measures (M), behavior ratings (B), and adjective ratings (A) across 17 trait constructs. *Note.* Computed with Fisher's *r*-to-*Z* transformation.

ratings when controlling for behavior ratings; complete mediation would be evidenced if the manifest behaviors would no longer affect adjective ratings when controlling for behavior ratings.

We estimated and tested this model by multiple regression analyses following Baron and Kenny (1986). The behavior measures as predictors correlated significantly ($p < .05$) both with the adjective ratings as criteria (M–A) and with the behavior ratings as potential mediators (M–B) in 7 out of 17 traits (see Table 4). In 6 of these traits, multiple regressions of adjective ratings on both behavior measures and behavior ratings (a) showed a significant impact of the mediator (behavior ratings) on the criterion (adjective ratings), and (b) rendered the manifest behavior measures' effect on the adjective ratings non-significant, showing a full mediation of the effects of manifest behavior on the adjective ratings by the behavior ratings.

Because of the small sample of animals ($N = 16$–$20$), only correlations above .47 were significant, and less than half of the 17 traits met the requirements for mediation analysis. Therefore, we relaxed the criterion in an additional analysis to $p < .10$. According to this criterion, one further trait (playfulness) met the requirements for mediation analysis, and showed a significant effect of the mediator on the criterion and full mediation (see Table 4).

## 4. Discussion

We compared three different methods of personality assessment in Great Apes: behavior measures, behavior ratings, and adjective ratings. All three methods yielded reliable and stable assessments that converged notably within the traits' nomological networks. The cross-method coherence is remarkable given that rating data were collected independently from behavior measures. These results establish strong empirical evidence for the construct validity of personality traits that were generated in a bottom-up approach from the species' behavioral repertoires (Uher, 2005, submitted for publication). These findings have several important implications.

When tailored to the specifics of non-human species, standard methods from human personality psychology, manifest behavior measures and ratings, provided equally reliable and valid personality assessments in animals, and yield results that are surprisingly consistent with those from human samples (Funder, 2004; Mischel & Shoda, 1995). For the Great Apes, it is thus not a question of human versus non-human species to meet the standard criteria of personality assessments (reliability, stability, and validity), but rather it is a question of appropriately adapting the principles of trait generation and trait operationalization to the specifics of animals.

### *4.1. Bottom-up trait generation from the species' behavioral repertoire*

Generating trait constructs and items for an empirical study is an important step; it determines the scope of identifiable traits and can therefore influence empirical results. For example, in chimpanzee personality ratings, one item list (Murray, 1998) yielded only half of the factors repeatedly shown by another item list (King & Figueredo, 1997). When different generation procedures can lead to such different results, it becomes obvious that trait generation can strongly influence the validity of species comparisons.

From our perspective as non-conspecific outsiders it is difficult to decide which personality traits may be important in other species because animal personality may vary on other dimensions than human personality. Anthropocentric bias is difficult to control in expert nominations, and the validity of a lexical approach is inherently confined to humans. It was therefore suggested to base trait generation systematically on a species' behavioral repertoire (Uher, 2005, submitted for publication). The core ideas of this bottom-up approach are grounded in trait theory. When traits as internal behavior regulating mechanisms are expressed in behavior, then it should be possible to infer the personality structure of a species from its behavioral variability. And because traits are assumed to create stable relations between situations and an individual's behavior, trait constructs can be generated by behavior–situation units that are derived from a species' known behavioral repertoire and the typical situations its members encounter.

This bottom-up approach was applied to the Great Apes for which 19 mono-polar trait constructs were generated for a first empirical test in a small captive sample (Uher, 2005, submitted for publication). The present data on these traits' construct validity suggest that the behavioral repertoire approach generates valid

personality traits. Though further steps of empirical analyses in larger samples are needed to identify the underlying factor structure, the convergent assessments from three independent methods already establish strong empirical evidence for trait domains beyond those covered by current models of human and non-human personality, for example sexual activity or food orientation.

This important finding highlights that personality models valid in one species may have only limited validity for another species because the trait domains' scope or even the trait constructs themselves may be confined to the first species. For this reason, personality models should be adapted to other species with great caution.

Whereas in the Great Apes, trait generation could be combined for all species due to strong similarities in their universal behaviors, bottom-up approaches in taxonomically more distant species will very probably generate different sets of trait constructs. In cross-species comparisons, the bottom-up approach is analogous to the emic approach in human cross-cultural research (Church, 2001); both obtain indigenous trait measures and compare species or cultures on empirically shared rather than on imported and modified trait dimensions. In this thought tradition, the behavioral repertoire approach provides a solid base for future empirical studies on species differences (Uher, submitted for publication).

Although the small sample sizes preclude comparing the four species in the present data, this study demonstrates a suitable design to study multiple species using the same theoretical and methodical approach. Such designs are important to ensure comparability for species comparisons, especially for contrasts between phylogenetically distant species.

Our findings suggest that the biological standard of knowledge on the species' general behavior, which is unbiased by previous personality research, constitutes a solid basis for systematic and representative trait generation. We could also show that the number of generated trait constructs and their operationalizations are manageable, which shows that the bottom-up approach from behavior is not only theoretically valid, but also empirically viable. This is an important point because more molecular bottom-up approaches like endophenotype approaches that start with neurotransmitters or hormone cascades are equally valid, but more difficult to put into practice for large-scale structural analyses.

Not only the generation of trait constructs but also their operationalizations have to be appropriate for the specifics of animals. Behavior measures have high face validity, but operationalizations in ratings may be prone to bias because they inevitably involve human raters and human language. We compared two different kinds of ratings and analyzed their accuracy in more detail.

### *4.2. Trait operationalization: Adjective ratings versus behavior ratings*

The two item pools varied systematically in their degree of inference from perceivable behavior. Adjectives are highly inferential and require the raters to interpret the animals' overt behavior. In contrast to previous item lists, the adjectives of the *Great Ape Personality Inventory (GAPI-A)* are not complemented with additional definitions that predetermine behavioral or situational contexts. Instead, raters have to rely completely on the adjectives' implicit meanings to form their assessments. The behavior items of the *Great Ape Personality Inventory (GAPI–B),* on the other hand, describe prototypical behavior with verbs, without using any adjectives. These items require frequency assessments of specific behaviors (for example, "walking or brachiating") in specific situations (for example, "when being fed") and are therefore less inferential than adjectives.

In hierarchical models of human personality, behavior-descriptive verbs typically describe lower levels, whereas adjectives typically describe higher levels (Eysenck, 1947; Guilford, 1959). Categories at the higher levels are often broad and refer to more diverse but less specified exemplars than narrow (often lower-level) categories that refer to comparatively small, but highly homogeneous sets of specified and less diverse exemplars.

The higher coherence between behavior ratings and behavior measures (B–M) than between adjective ratings and behavior measures (A–M, see Fig. 1) could thus indicate a fidelity-bandwidth trade-off. That is, manifest behavior measures and behavior ratings both refer to narrow but homogeneous sets of specific trait-relevant behaviors and situations. Adjective ratings, by contrast, refer to a broader bandwidth of less specified and more heterogeneous behaviors and situations which represent more diverse trait aspects. This suggests that the adjectives' coherence to manifest behavior may be lower than that of behavior ratings because the adjectives' bandwidth is broader.

However, a study on personality descriptors in humans shows that bandwidth and grammatical form (adjectives versus verbs) have separate effects on fidelity. When category breadth was held constant, behavior-descriptive verbs were shown to be more trait-prototypical than adjectives (Borkenau & Müller, 1991). Thus, for humans, beyond the effect of category breadth, behavior-descriptive verbs refer to perceivable behavior more precisely than adjectives, and are therefore more appropriate to describe how individuals actually behave than adjectives.

The present data on Great Apes square nicely with these findings (see Fig. 1 and the mediation analyses in Table 4 for the specific traits). This suggests that adjectives were less coherent with behavior measures than behavior ratings not only because adjectives have a broader bandwidth, but also because they may be the less informative behavior descriptors. These findings suggest, in turn, that predictions of manifest behavior are more precisely achieved by behavior ratings than by adjective ratings.

Being more inferential, adjectives are also more susceptive to rater biases such as halo effects or anthropomorphic biases than behavior-descriptive verbs. For example, liking of an animal may cause both biased inferences from the animal's behavior and an increase of all positive adjective ratings. Adjectives may also be prone to anthropomorphic interpretations of behavior, in particular adjectives derived from human personality inventories. Are adjectives thus not suitable for animal ratings?

Only systematic validation can clarify whether adjectives are useful descriptors of animal personality or not; it must be shown empirically that the adjectives have the assumed meaning in the particular target species. Lexically derived descriptors from human personality inventories are inherently anthropocentric. As long as systematic studies are missing that validate each item for each target species, it remains unknown to what these adjectives actually refer. Of course, ratings can only be done on verbal descriptors; given the sedimentation hypothesis it is even very likely that descriptors that are valid for animals are also valid for humans. But the reverse does not need to be true.

We therefore selected adjectives that describe the traits in the most direct and most easiest way in the everyday language of the keepers. The nomological networks around each trait construct permitted systematic comparisons of these adjectives with direct measures and ratings of specific behaviors that we assumed to be trait-prototypical.

Most adjectives had strong empirical associations with behavior ratings; the coherence with manifest behavior measures was moderate to substantial in most traits (see Table 4). Thus, if trait constructs are ecologically valid and if the operationalizing adjectives are selected carefully rather than directly imported from human personality inventories then adjectives can have substantial validity in animals. This shows that adjectives can be quite useful for animal personality ratings.

However, some adjectives had only little or even no empirical relation to behavior. For example, friendly to conspecifics was unrelated to grooming and contact sit both in manifest behavior measures and in behavior ratings. Excitable was associated with arousability in strange situations but not in prefeeding situations as expected (see Table 4). Thus, the adjectives' implicit connotations can differ from their intended meaning. This puts the common practice of adapting adjectives to animals by simply adding definitions in another perspective. For example, the most widely used adjective list in primate personality research defines the adjective gentle with "responds to others in an easy, kind manner" (Stevenson-Hinde & Zunz, 1978). One would expect that gentle, kind, and friendly apes are more involved in grooming than others, and that apes that respond to others in an easy, kind manner are also allowed to spend more time in body contact with them. Surprisingly, we found zero correlations between friendliness ratings and these behaviors.

These puzzling findings highlight that without substantial validation it remains unclear what precisely is at the bottom of adjective ratings in animals. They also suggest that the use of *a priori* definitions as a means to adapt adjectives to animals should be reconsidered carefully. It is an important finding that most adjectives of the *Great Ape Personality Inventory (GAPI–A)* have high validity in Great Apes. Why other adjectives were not valid merits further empirical exploration. Concluding, adjective inventories can be useful for animal personality judgments, but they have to be designed with much greater caution than previously done.

The development of valid adjective lists for animal personality research could profit from a better understanding of how raters come to personality judgments on animals. The present data on interrater agreement, stability, and cross-method coherence reflect the raters' ability to assess these species in general, and to differentiate reliably and accurately between individuals. According to the Realistic Accuracy Model (Funder, 1995,

1999), the present correlations between interrater agreement and the traits' availability, in both adjective and behavior ratings, provide encouraging evidence that the raters based their judgments on and derived their consensus from behavior observation rather than from collaboration and informal exchange. This suggests that the raters successfully managed the complex task of processing much information about the apes' habitual behaviors across a diversity of situations and occasions. These findings also argue against the assumption that animal personality ratings would merely reflect the raters' implicit personality theories rather than the targets' internal dispositions.

Mediation analyses revealed more detailed information on how raters probably develop adjective judgments. They showed that the relations between adjective ratings and manifest behavior measures were partially or even completely mediated by behavior ratings for many traits. This suggests that the raters relied on frequency assessments of specific, trait-indicating behaviors to form their adjective judgments. This also corresponds to previous results on the higher prototypicality of behavior-descriptive verbs and supports the assumption that behavior ratings are more closely bound to manifest behavior than adjectives, and may therefore be the more accurate behavior predictors.

Adjectives may be less accurate, but because they are more intuitive, they may facilitate for raters to form a picture of the individuals and to hold it in mind, which is particularly important for everyday management. The higher effect sizes of adjective rating stabilities compared to those of the behavior ratings seem to support this idea. The substantial effect sizes of both rating methods compared to manifest behavior measures show that rating stabilities in general should be considered carefully when interpretations of personality stabilities are only based on rating data.

The broader bandwidth of adjectives could also convey other interesting information about personality characteristics that account for perceivable behavior; for example, they could provide information about the individuals' behavior in other situations or in the future (Semin & Fiedler, 1988). Therefore, it is quite possible that the adjectives have a wider predictive range than the behaviors measured in this study. That way, adjectives can constitute an important source for construct validation.

We have argued that the preconditions for valid personality assessments in animals are trait generations and trait operationalizations that appropriately meet the specifics of animals. We have shown that ecologically valid trait constructs and operationalizations successfully meet the standard criteria of human personality psychology. Within these constraints, bias in animal personality ratings is probably not larger than in human personality ratings. Therefore, for realistic and accurate personality judgments in animals, a fifth precondition should be added to the Realistic Accuracy Model (Funder, 1995, 1999). The framework that raters are given to indicate their personality judgments should be appropriate to the specifics of the rated species; that is trait constructs and operationalizations should be ecologically valid.

### *4.3. Limitations and future directions*

The intense research design with multiple and repeated assessments is inevitably at the expense of a larger sample. We decided for a smaller mixed species sample of Great Apes, because already 12 studies in chimpanzees, but only three in gorillas, one in bonobos and no study in orangutans were published when planning the study. In doing so, we could demonstrate the realization of a multimethod design in a multi-species study, which is particularly important for species comparisons. But the present results are clearly limited by the small sample size which also prevented use of factor analysis for an empirical evaluation of personality difference structure.

Psychometric studies of primate personality are still rare, in Great Apes in particular, but they provide interesting opportunities for many important research questions. For example, it was argued that coherence between personality judgments and manifest behavior measures occurs more easily in Great Apes than in humans because human behavior can only be sampled in a limited number of domains, whereas self or peer ratings are based on larger and far more varied domains (King & Figueredo, 1997). This predestines research in animals for basic methodological studies that are relevant to both non-human and human research. Shorter life times and the possibility to control environments permit more detailed studies on genetic and environmental influences on personality development. The complex social systems of non-human primates in particular could be illuminative to understand more about the role of personality in social behavior and group dynamics,

which could also have practical relevance. Because transfers to potential new mates are often very stressful for captive animals and success of new group formations is often not well predictable, information on the personality compatibility between individuals could complement genetic criteria for decisions on animal husbandry management. Last but not least, as human's closest living relatives, non-human primates can be illuminating for studies on the evolutionary basis of personality (Dall, Houston, & McNamara, 2004; Gosling, 2001; Nettle, 2006; Sih, Bell, Johnson, & Ziemba, 2004; Uher, submitted for publication; Wilson, 1994). Which personality traits are universal across a wide range of species, which are only associated with the complex lives of primates, and which traits are uniquely human?

Personality research in animals constitutes a yet rather unexplored research area holding considerable potentials for fruitful and serious research on a wide range of topics. The present study has opened a new avenue for accomplishing this task with behavior ratings grounded in the species' behavioral repertoires. Although we have critically discussed previous approaches to animal personality, we are convinced that, ultimately, trait constructs that emerge alike from different starting points are the most convincing.

## Acknowledgments

We thank David Funder and three anonymous reviewers for valuable comments on the manuscript. We are greatly indebted to the zoo-keepers at the Wolfgang Köhler Primate Research Center in Leipzig, Germany, for their cooperation and for rating the apes. We also thank Josep Call from the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, for supporting this study, and Josefine Kalbitz for recording behavior for reliability analyses.

## References

Allport, G. W., & Odbert, H. G. (1936). Trait names: A psycholexical study. *Psychological Monographs, 47*, 1.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bolig, R., Price, C. S., O'Neill, P. L., & Suomi, S. J. (1992). Subjective assessment of reactivity level and personality traits of rhesus monkeys. *International Journal of Primatology, 13*, 287–306.

Borkenau, P., & Müller, B. (1991). Breadth, bandwidth, and fidelity of personality-descriptive categories. *European Journal of Personality, 5*, 309–322.

Buirski, P., Plutchik, R., & Kellerman, H. (1978). Sex differences, dominance, and personality in the chimpanzee. *Animal Behaviour, 26*, 123–129.

Buss, D. M. (1999). *Evolutionary psychology: The new science of the mind*. Needham Heights, Massachusetts: Allyn and Bacon.

Caine, N. G., Earle, H., & Reite, M. (1983). Personality traits of adolescent pig-tailed monkeys (*Macaca nemestrina*): An analysis of social rank and early separation experience. *American Journal of Primatology, 4*, 253–260.

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

Capitanio, J. P. (1999). Personality dimensions in adult male rhesus macaques: Prediction of behaviors across time and situation. *American Journal of Primatology, 47*, 299–320.

Capitanio, J. P., & Widaman, K. F. (2005). Confirmatory factor analysis of personality structure in adult male rhesus monkeys (*Macaca mulatta*). *American Journal of Primatology, 65*, 289–294.

Carver, C. S., & Scheier, M. F. (2000). *Perspectives on personality* (4th ed.). Boston: Allyn and Bacon.

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of Personality, 69*, 979–1006.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Erlbaum.

Crawford, M. P. (1938). A behavior rating scale for young chimpanzees. *Journal of Comparative Psychology, 26*, 79–91.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Dall, S. R. X., Houston, A. I., & McNamara, J. M. (2004). The behavioural ecology of personality: Consistent individual differences from an adaptive perspective. *Ecology Letters, 7*, 734–739.

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.

Eysenck, H. J. (1947). *Dimensions of personality*. London: Routledge & Kegan Paul.

Figueredo, A. J., Cox, R. L., & Rhine, R. J. (1995). A generalizability analysis of subjective personality assessments in the stumptail macaque and the Zebra finch. *Multivariate Behavioral Research, 30*, 167–197.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102*, 652–670.

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception*. San Diego: Academic Press.

Funder, D. C. (2004). *The personality puzzle* (3rd ed.). New York: W.W. Norton and Co.
Gold, K. C. (1992). Nonsocial behavior of captive infant gorillas. *American Journal of Primatology, 26*, 65–72.
Gold, K. C., & Maple, T. L. (1994). Personality assessment in the gorilla and its utility as a management tool. *Zoo Biology, 13*, 509–522.
Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229.
Gosling, S. D. (1998). Personality dimensions in spotted hyenas (Crocuta crocuta). *Journal of Comparative Psychology, 112*, 107–118.
Gosling, S. D. (2001). From mice to men: What can we learn about personality from animal research? *Psychological Bulletin, 127*, 45–86.
Gosling, S. D., & John, O. P. (1999). Personality dimensions in non-human animals: A cross-species review. *Current Directions in Psychological Science, 8*, 69–75.
Guilford, J. P. (1959). *Personality*. New York: McGraw-Hill.
Hebb, D. O. (1949). Temperament in chimpanzees: I. Method of analysis. *Journal of Comparative and Physiological Psychology, 42*, 192–206.
John, O. P., & Gosling, S. D. (2000). Personality traits. In A. E. Kazdin (Ed.). *Encyclopedia of psychology* (Vol. 6, pp. 140–144). Washington, DC: American Psychological Association.
Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person situation debate. *American Psychologist, 43*, 23–34.
King, J. E., & Figueredo, A. J. (1997). The five-factor model plus dominance in chimpanzee personality. *Journal of Research in Personality, 31*, 257–271.
King, J. E., & Landau, V. I. (2003). Can chimpanzee (*Pan troglodytes*) happiness be estimated by human raters? *Journal of Research in Personality, 37*, 1–15.
King, J. E., Weiss, A., & Farmer, K. H. (2005). A chimpanzee (*Pan troglodytes*) analogue of cross-national generalization of personality structure: Zoological parks and an African sanctuary. *Journal of Personality, 73*, 389–410.
Lukas, K. E., Hoff, M. P., & Maple, T. L. (2003). Gorilla behavior in response to systematic alternation between zoo enclosures. *Applied Animal Behaviour Science, 81*, 367–386.
Mangold, P. (2006). Interact User Guide, V 7.0 ff. Arnstorf: Mangold International.
Maple, T. L. (1980). *Orang-utan behavior*. New York: Van Nostrand Reinhold.
Martau, P., Caine, N., & Candland, D. (1985). Reliability of the emotions profile index, primate form, with *Papio hamadryas*, *Macaca fuscata*, and two *Saimiri* species. *Primates, 26*, 501–505.
McGuire, M. T., Raleigh, M. J., & Pollack, D. B. (1994). Personality features in vervet monkeys: The effects of sex, age, social status, and group composition. *American Journal of Primatology, 33*, 1–13.
Meder, A. (1993). *Gorillas Ökologie und Verhalten*. Berlin: Springer Verlag.
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268.
Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science, 11*, 50–54.
Murray, L. E. (1998). The effects of group structure and rearing strategy on personality in Chimpanzees (*Pan troglodytes*) at Chester, London and Twycross Zoos. *International Zoo Yearbook, 36*, 97–108.
Nettle, D. (2006). The evolution of personality variation in humans and other animals. *American Psychologist, 61*, 622–631.
Pavlov, I. P. (1906). The scientific investigation of the psychical faculties or processes in the higher animals. *Science, 24*, 613–619.
Pederson, A. K., King, J. E., & Landau, V. I. (2005). Chimpanzee (*Pan troglodytes*) personality predicts behavior. *Journal of Research in Personality, 39*, 534–549.
Rijksen, H. D. (1978). *A field study on sumatran Orangutans (Pongo pygmaeus abelli* Lesson 1827): Ecology, behaviour and conservation. Wageningen, The Netherlands: H. Veenman and Zonen.
Schaller, G. B. (1963). *The mountain gorilla. Ecology and behavior*. Chicago, IL: University of Chicago Press.
Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54*, 558–568.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2*, 420–428.
Sih, A., Bell, A. M., Johnson, J. C., & Ziemba, R. E. (2004). Behavioral syndromes: An integrative overview. *The Quarterly Review of Biology, 79*, 241–277.
Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M. (1980a). Subjective assessment of rhesus monkeys over four successive years. *Primates, 21*, 66–82.
Stevenson-Hinde, J., Stillwell-Barnes, R., & Zunz, M. (1980b). Individual differences in young rhesus monkeys: Consistency and change. *Primates, 21*, 498–509.
Stevenson-Hinde, J., & Zunz, M. (1978). Subjective assessment of individual rhesus monkeys. *Primates, 19*, 473–482.
Suomi, S. J., Novak, M. A., & Well, A. (1996). Aging in rhesus monkeys: Different windows on behavioral continuity and change. *Developmental Psychology, 32*, 1116–1128.
Susman, R. L. (Ed.). (1984). *The pygmy chimpanzee evolutionary biology and behaviour*. New York: Plenum Press.
Uher, J. (submitted for publication). Comparative personality research: Methodological approaches. *European Journal of Personality*.
Uher, J. (2005). Personality in the Great Apes—Methods and approaches. Unpublished Master's thesis. Humboldt-University Berlin & Max Planck Institute for Evolutionary Anthropology. Berlin & Leipzig, Germany.
Uher, J., Asendorpf, J. B., & Call, J. (2008). Personality in the behaviour of Great Apes: Temporal stability, cross-situational consistency and coherence in response. *Animal Behaviour, 75*, 99–112.

van Lawick-Goodall, J. (1968). The behaviour of free-living chimpanzees in the Gombe Stream Reserve. *Animal Behaviour Monographs, 1*, 165–311.

van Schaik, C. P. (2004). *Among orangutans: Red Apes and the rise of human culture*. Belknap/Harvard University Press.

Weiss, A., King, J. E., & Figueredo, A. J. (2000). The heritability of personality factors in chimpanzees (*Pan troglodytes*). *Behavior Genetics, 30*, 213–221.

Weiss, A., King, J. E., & Perkins, L. (2006). Personality and subjective well-being in orangutans (*Pongo pygmaeus* and *Pongo abelii*). *Journal of Personality and Social Psychology, 90*, 501–511.

Wilson, D. S. (1994). Adaptive genetic variation and human evolutionary psychology. *Ethology and Sociobiology, 6*, 219–236.

Yerkes, R. M. (1939). The life history and personality of the chimpanzee. *American Naturalist, 73*, 97–112.