PREPRINT

Original Article

Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments

Jana Uher *[a,b], Elisabetta Visalberghi [c]

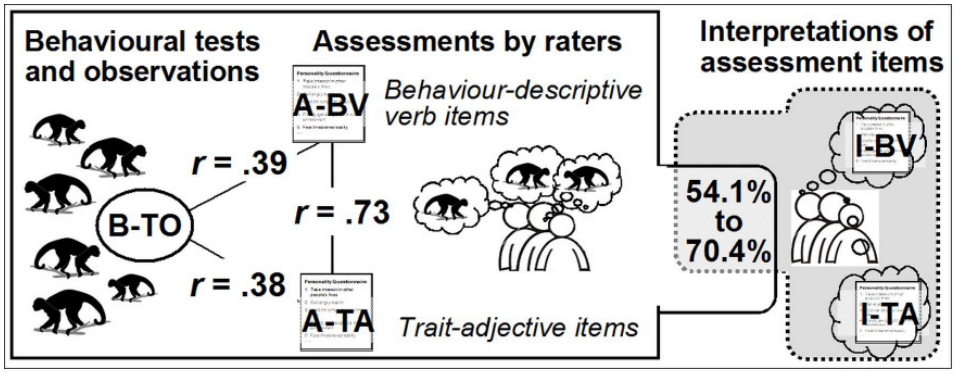
[a] *The London School of Economics and Political Science, United Kingdom*
[b] *Comparative Differential and Personality Psychology, Free University Berlin, Germany*
[c] *Unit of Cognitive Primatology and Primate Centre, Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR), Rome, Italy*

* Correspondence:
London School of Economics and Political Science
Department of Social Psychology
Houghton Street, WC2A 2AE London
United Kingdom
e-mail: uher@primate-personality.net

Abstract
Personality assessments and observations were contrasted by applying a philosophy-of-science paradigm and a study of 49 human raters and 150 capuchin monkeys. Twenty constructs were operationalised with 146 behavioural measurements in 17 situations to study capuchins' individual-specific behaviours and with assessments on trait-adjective and behaviour-descriptive verb items to study raters' pertinent mental representations. Analyses of reliability, cross-method coherence, taxonomic structures and socio-demographic associations highlighted substantial biases in assessments. Deviations from observations are located in human impression formation, stereotypical biases and the findings that raters interpret standardised items differently and that assessments cannot generate scientific quantifications or capture behaviour. These issues have important implications for the interpretation of findings from assessments and provide an explanation for their frequent lack of replicability.

Key words: assessment method; Behavioral Repertoire x Behavioral Situations Approach; capuchin monkey (*Sapajus spp.*); meaning construction; item interpretation; Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm); observation; personality; questionnaire; replicability

Graphical Abstract



Contents

1. Introduction

After almost a century of assessment-based research, psychologists and social scientists are increasingly criticising the shortcomings of assessment methods (Baumeister, Vohs & Funder, 2007; Hammersley, 2013; Rosenbaum & Valsiner, 2011; Uher, 2013, 2015a,b,c) and intensifying their behavioural methods, such as ambulatory monitoring (Fahrenberg et al., 2007; Mehl & Connor, 2012), life-logging (Gurrin et al., 2014), reality-mining (Dong et al., 2011), subjective evidence-based ethnography (SEBE, Lahlou, 2011; Lahlou et al., 2015) and behavioural observations (Furr, 2009; Uher et al., 2008, 2013a). But why are assessments criticised? Here, we explored the methodological differences between observations and assessments for research on "personality". Using a five-method study, we analysed the ways in which assessment-based categorisations of individual-specific behaviours deviate from those obtained with observations and explored the possible sources of these differences.

As humans, we are intimately familiar with the behaviours and individual differences of our own sociocultural community and species. We have developed comprehensive bodies of pertinent social knowledge, beliefs and values that are encoded in everyday language (Ah-King, 2013; Allport, 1937). Our socio-cognitive abilities to recognise individual-specific behaviours are not limited to human individuals, however; they were also fundamental for the domestication of animals (Belyaev, 1969; Trut, 1999; Uher et al., 2013b). Multi-species studies are particularly illuminating for exploring these abilities because other species' behaviours differ from ours, and our pertinent everyday knowledge about them and their individual differences is comparably limited (Uher, 2008a, b). Moreover, nonhuman individuals do not adapt their behaviours to our human beliefs and values as human individuals do (Lloyd & Duveen, 1992) so that attribution biases become particularly apparent.

This research explores a multi-species sample comprising 49 human raters and 150 capuchin monkeys from various research institutions worldwide. Capuchin monkeys, a nonhuman primate species endemic to South America, are interesting for "personality" research because they have large brains, extended periods of maternal dependency and long life spans (Byrne & Suomi, 1998; Fragaszy et al., 2004). Their manipulative skills and flexible tool use are comparable to those of chimpanzees (Visalberghi & Fragaszy, 2012). Capuchins exhibit pronounced individual-specific behaviours (Uher et al., 2013a) and are therefore interesting for exploring how humans mentally represent individual behavioural differences and how they assess individuals' "personality".

We applied the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm), a novel paradigm aimed at exploring and making explicit the most basic assumptions that are being made in a given scientific system and the metatheories and methodologies derived from them for enabling critical analyses and further developments. The TPS-Paradigm comprises metatheoretical, methodological and methodical frameworks for research on both human and nonhuman individuals—and on their "personality". In these frameworks, concepts, approaches and methods from various disciplines are systematically integrated, further developed and complemented by novel ones (Uher, 2011a, 2013, 2015a,b,c,d,e, 2016, in press).

"Personality" is commonly defined as "an individual's consistent patterns", "characteristics" and "uniqueness". But these vague notions do not specify what is supposed to be consistent with what else and what is considered to be characteristic and unique and why. The TPS-Paradigm showed that all concepts of "personality" (see e.g., Allport, 1937) incorporate the idea of *individual-specificity* but differ in the particular phenomena considered. Individual-specificity is studied, amongst others, in behavioural, psychical, physiological and morphological phenomena, in people's sociocultural knowledge, belief and value systems and semiotic encodings, such as everyday language (Uher, 2013). Given this, the term "personality" is presented in quotes in the TPS-Paradigm to indicate its status as a construct that refers to different kinds of phenomena depending on the particular definition used.

The concept of individual-specificity highlights central methodological criteria. To be specific to an individual, patterns in the phenomena under study must differ between individuals, thus be *differential*. But in momentary and fluctuating phenomena, such as behaviours, within-individual variations are often pronounced so that individuals' scores can be only *probabilistic*. To differentiate individual-specific from random between-individual variations, individuals' differential probabilities must be shown to be *temporally extended*. But it is important to note that differential, probabilistic and temporal patterns cannot be directly observed. Individual-specificity—"personality"—is an abstract idea constructed by humans (a *construct*) to denote regularities that occur in many individuals in different ways and over some time. Thus, "personality" cannot be directly perceived at any given moment; this has important implications for research methodology. Moreover, the different kinds of phenomena in which individual-specificity is conceived as "personality" feature different properties that require different methods of investigation. That is, not every method is suitable for exploring every kind of phenomenon.

Behaviours can be directly perceived by multiple observers because behaviours are external to individuals' bodies (for a meta-theoretical definition of behaviour, see Uher, in press). But observations are complicated by the behaviours' momentariness. This requires methods for recording events *while or immediately after they have occurred* (live or from audiovisual records), so-called *nunc-ipsum methods*[1]. Thus, observational data reflect (non-)occurrences of behavioural events that are perceptible by multiple observers and specified in the encoding scheme used. It is only *after their collection (post-hoc)* that observational raw data are explored for differential, probabilistic and temporal patterns and used, in a second step, to create measures reflecting individual-specific patterns (Uher, 2015a, e).

"Personality" assessments, by contrast, require raters to judge *directly* how a target individual typically behaves in comparison with others. But because individual-specific patterns cannot be perceived directly, raters must *retrospectively* construct their assessments on the basis of memory. But memory recall is known to be affected by many biases (Fahrenberg et al., 2007; Schacter, 1999). Moreover, the impressions that people form of individuals are influenced by their everyday knowledge, beliefs and language and by the social values attributed to behaviours (e.g., negative valences of aggressive acts may lead to overestimations of their occurrence). Therefore, assessments cannot reflect occurrences of behaviours as observed in the past. Assessments reflect the ideas and representations that the raters have developed of the target individual on the basis of diverse sources of personal and social knowledge (Uher, 2013; Uher et al., 2013b).

These differences between observations and assessments entail fundamental differences for *quantification*. Assessments are considered subjective quantifications. But how are quantifications generated with other methods, such as observations? To elaborate the essential differences, the TPS-Paradigm scrutinised the concepts of quantification that were established in *metrology,* the science of measurement (JCGM, 2008), highlighting two basic requirements. First, in the phenomena under study, scientists must specify the sets of the elements to be quantified; this is called the *set-theoretic requirement of scientific quantification*. Second, the elements thus defined must be compared with designated standards of measurement to express their ratio as a real number; this is called the *algebraic requirement*. Accordingly, numerical data that fulfil both requirements are called *scientific* quantifications as opposed to (subjective) quantifications in which these requirements are not fulfilled (Uher, 2015a,b,c,e).

Observational methods require researchers to specify in the encoding schemes all elements of the sets *B* of behaviours, *S* of situations, *I* of individuals and *T* of occasions and time spans explored in a study, thus to fulfil the set-theoretic requirement of scientific quantification. This means that observers can be trained to perceive, categorise and encode behaviours in standardised ways such that independent persons encode the occurrences of events in the same target individual and on the same occasions in highly similar ways as

[1] From the Latin *nunc ipsum* for at this very instant.

analysed in terms of inter-*observer* reliability. Comparisons of specified events with fixed standards of measurement, as needed to fulfil the algebraic requirement, however, are often complicated. Behaviours are momentary and often highly fluctuating (e.g., social interactions). Moreover, behavioural events of the same kind often vary in their spatio-temporal extensions (e.g., a gesture can be made quickly or slowly). Therefore, spatial standards of measurement (e.g., metric units of length) often cannot be used.

To meet these peculiarities, the TPS-Paradigm introduced the concept of *time-relative probabilities*. This novel type of probability sets the empirical occurrences of specified events (the elements studied in sets *S* and *B*) in the individuals under study (the elements studied in set *I*) in relation to specified time periods (the elements studied in set *T*). Therefore, unlike other types of probability, absolute time-relative probabilities have a unit (e.g., frequency per observation hour) and are not constrained to values between 0 and 1 (but they cannot become negative). Time provides the fixed standards of measurement that are needed to fulfil the algebraic requirement of scientific quantification. Given that the international time standards provide equal units (e.g., minutes) and that the non-occurrence of the events under study in the time periods under study defines an absolute point of zero, the scientific quantifications thus obtained are *ratio scaled*. This allows all arithmetic operations to be applied and for direct comparisons to be made within and between individuals, groups, species and studies (for details, see Uher, 2013, under review; for empirical applications, see Uher, 2015e, Uher et al., 2013a).

In assessments, quantifications are generated in fundamentally different ways. The item statements and answer categories comprised by an inventory constitute the encoding schemes (i.e., variables and values) that are provided to the persons who generate the data (raters). Item statements often comprise adjectives from everyday language because these abstract and decontextualised terms can be applied to diverse kinds of behaviours and situations without specifying any particular ones. By using everyday language, researchers capitalise on the raters' common-sense knowledge to interpret the meaning of the item statements. Because of this, raters are commonly not trained to use the encoding scheme of a given inventory to generate data in standardised ways as is the case for observers.

But unlike scientific terms and concepts, people's everyday terms and concepts are often fuzzy and context-sensitive (Hammersley, 2013). Therefore—and despite all efforts to improve item readability, clarity and simplicity during instrument development, people construct a broad range of meanings for the same standardised item statement—a *field of meanings* (Arro, 2013; Diriwächter et al., 2005; Rosenbaum & Valsiner, 2011; Uher, 2015b). Thus, people's understanding of standardised item statements often varies between and within individuals—raters and researchers alike. An item's meaning evolves as a product of cognitive information processing. But in standardised assessments, researchers record only the outcome of this mental activity without enabling the raters to provide information about the particular behaviours, situations, individuals and time periods they may have considered and how they may have arrived at their overall judgement (Wagoner & Valsiner, 2005). In behaviour-descriptive items, researchers specify particular behaviours that raters are asked to consider. But the extent to which raters may (unintentionally) also consider other behaviours and other persons' reports about the target individual cannot be determined.

Answer categories as well are commonly encoded with abstract and general descriptors to make them applicable to various kinds of assessments (e.g., "seldom", "often"). But how raters actually interpret and use these quantitative categories for a given assessment remains unspecified. For example, it was shown that instructing raters to consider different sets *I* of individuals results in reference-group effects that influence the quantifications that are obtained (Heine et al., 2002). It is important to note that high inter-*rater* reliability shows only that persons agree about the *encodings* that they have generated with a particular inventory but not whether they have actually encoded and quantified the *same* elements in the *same* standardised ways as is required for observations and analysed as inter-*observer* reliability.

In sum, assessment methods do not allow researchers to specify the particular elements of the sets *B, S, T* and *I* that raters may consider or how raters may quantify their occurrences and mentally compute interrelations within and between these sets to arrive at an overall judgement. The quantitative data that are generated by recoding the verbal answer categories into numeric encodings therefore cannot fulfil the two requirements of scientific quantification. Because *what* has been quantified and *how* this was done remain unspecified, quantitative comparisons between individuals, situations and groups are precluded (Uher, 2015e).

These peculiarities of assessment methods may be a major source of within- and between-person variability in assessments and of the frequent lack of replicability of many findings in psychology and the social sciences, which is currently under intense discussion (Asendorpf et al., 2013; Carpenter, 2012; Yong, 2012). These methodological problems are often overlooked and are therefore the focus of this research.

2. Methods

2.1 Capuchin individuals

We studied *N* = 150 capuchin monkeys (*Sapajus spp.*[2]) housed in 27 groups at 9 research institutions, 6 in the United States, 1 in the United Kingdom, 1 in Japan and 2 in Italy. Across all institutions, the sample comprised 74 males and 76 females, ranging in age from 1 to 41 years; their mean age was 13.1 years with a median of 11 years (*SD* = 9.49; Table 1). The multi-method comparisons focussed on the largest of our sub-samples, housed at the Primate Centre of the ISTC-CNR and hosted by the Bioparco of Rome in Italy. The monkeys were treated in accordance with local regulations and the Guidelines for the Treatment of Animals in Behavioural Research and Teaching (ASAB/ABS, 2012).

Table 1 Capuchin individuals at each research institution: Sample sizes, age and sex

Research Institution, Country	***n***	**Age in years**				**Sex**
		M	*Mdn*	*Range*	*SD*	(M;F)
National Institute of Health, Animal Center, US	24	7.6	6.0	1-30	6.45	15;9
University of Georgia, Primate Cognition and Behavior Laboratory, US	7	21.1	21.0	17-26	3.08	7;0
Yale Canine and Primate Laboratory, US	10	10.1	10.5	316	5.26	4;6
Georgia State University Language Research Center, US	12	9.9	9.0	3-21	5.58	6;6
Nathan Kline Institute for Psychiatric Research, US	20	24.0	30.0	2-34	11.46	3;17
Living Links Research Centre Edinburgh Zoo, UK	24	7.1	5.5	1-41	8.42	15;9
Kyoto University, School of Letters, JP	9	11.7	14.0	2-18	6.36	3;6
Parco dell' Abatino, IT	16	12.3	10.0	2-23	7.05	9;7
Primate Centre of the ISTC-CNR, IT	28	16.3	15.0	1-33	7.91	12;16

2.2 Human individuals providing assessments of capuchin individuals and interpretations of standardised item statements

Across all 9 institutions, 49 persons provided "personality" assessments of the total sample of 150 capuchins studied; each monkey was assessed by 1 to 5 raters (*M* = 2.6, *SD* = 0.7). All 49 raters had been working with capuchins for periods ranging from several months up to 29 years (*M* = 4.4 years; *SD* = 6.74). They had already known the particular capuchins who were being assessed for an average of *M* = 3.7 years (*SD* = 2.0). The raters judged their own familiarity with their target capuchins ("Please assess yourself: How well do you know [Name]?") on a five-point rating scale ranging from (1) *a bit* to (5) *very well*, scoring their average familiarity as *moderately well* (*Mdn* = 3, *M* = 3.2, *SD* = 1.18). The multi-method

[2] The species' former taxonomic name was *Cebus apella*; it has been changed as the result of recent molecular-genetic analyses (Lynch Alfaro et al., 2012).

comparisons focussed on the ratings provided by a sub-sample of 13 raters (9 women, 4 men) at the ISTC-CNR. They knew the target capuchins from different contexts (multiple responses possible): ethological observations (48.7%), lab-based experiments (51.3%), care taking (33%) and non-systematic observations (57.7%). Six of these raters (all women) also provided interpretations of the item statements used in this study.

2.3 Generation of constructs of individual-specific behaviours ("personality")

We applied the Behavioural Repertoire x Behavioural Situations Approach (BR_xBS-Approach; Uher, 2008a, 2011a, 2015b) to generate constructs of individual-specific behaviours ("personality") from the known behavioural repertoire of captive and wild capuchin monkeys. In a review of 68 publications, we compiled all major behavioural categories that were used in these studies along with the categories of situations in which the behaviours were reported to occur. The linking of a behavioural and a situational category is called a *behaviour$_x$situation-unit*. After organising the categories and eliminating redundancies across studies, we generated *working constructs* of "personality" by *hypothetically assuming individual-specific patterns* in the given behaviours and situations described (details are reported in the Supplemental Material and Uher et al., 2013). For captive capuchins, this procedure yielded 21 constructs, excluding constructs that involved behaviours and situations that occurred only in the wild (e.g., territoriality). The construct describing individual-specific behaviours in relation to youngsters was considered only in the analyses of the item interpretations because not all capuchin groups had young individuals at the time of our study.

2.4 Research design: Operationalisation in nomological networks

To operationalise the 21 working constructs, we used five methods that establish a nomological network around each construct. Operationalisations were based on the *behaviour$_x$situation-units* that were used to generate and to define a given construct. Behavioural measures were obtained from tests and observations (B-TO). Assessment measures were obtained from human raters in two standardised formats: behaviour-descriptive verb items (A-BV) and trait-adjective items (A-TA). We also collected raters' open-ended interpretations of the behaviour-descriptive verb items (I-BV) and the trait-adjective items (I-TA; Figure 1).

2.5 Contextualised behavioural measurements: Tests and observations (B–TO)

We comprehensively investigated the behaviours of the capuchins at the ISTC-CNR. To operationalise the 20 working constructs (excluding the youngster-related construct), we developed 15 laboratory-based tests featuring situational properties that are relevant for particular behaviours (e.g., multiple objects). To operationalise constructs describing social behaviours, we observed the monkeys in two different group situations in the outdoor enclosures (Prefeeding and Social observation). For most constructs, we were able to obtain behavioural measurements in several situations. Whenever it was possible, we measured multiple construct-related behaviours in the same situation to increase measurement reliability and to analyse cross-situational consistency and internal consistency. For all N = 20 working constructs, we obtained N = 146 contextualised behavioural measurements.

2.5.1 Scientific quantification of individual and individual-specific behaviours

To reduce the impact of day-to-day fluctuations and to generate data reflecting time-relative probabilities, behavioural measurements were repeated in various test sessions and trials and on different observation days within a 2- to 2.5-week *study block*. All behavioural tests were videotaped and coded for latencies, frequencies and durations of specified behaviours using a detailed encoding scheme and the coding software INTERACT (Rel. 9.2.1, www.behavioural-research.com; Mangold, 2010). In the Prefeeding observations, we used one-zero sampling with 10-sec time intervals to estimate frequencies that included any amount of time spent engaged in the given behaviours. In the Social observations, we

combined three observational methods to estimate time distributions of frequency and duration behaviours using interactive computer software programmed by JU (ObsTool) that logged all data entries with a precise timestamp.

To identify individual-specificity ("personality"), we repeated the entire process of data collection after a break of about a fortnight in a second 2- to 2.5-week block using the same scheme of repetitions and randomisation. In total, each monkey participating in all tests and observations was recorded for 320 min in the tests, 50 min in the Prefeeding observations and 25 h in the Social observations for a total of 31.2 h within a 60-day period. More details and an overview of all behavioural tests and observations are provided in Table S1 in the Supplemental Material and in Uher et al. (2013a).

These observational methods allowed us to fulfil the set-theoretic requirement of scientific quantifications. The elements that were studied from sets *B* and *S* are specified in the behavioural coding schemes and are encoded in the $N = 146$ contextualised behavioural variables generated (listed in the Supplemental Material of Uher et al., 2013a). The elements that were studied from set *I* are specified in Section 2.1. above; those from set *T* are specified in this Section 2.5 and the Supplemental Material. From the behavioural raw data obtained, we generated post-hoc scientific quantifications of individuals' behaviours by accumulating each individual's records over repeated measurement occasions (e.g., test trials, sessions, observation days). Given the fluctuations in behaviour, the measurements thus obtained reflect *probabilistic* patterns. We then obtained time-relative probabilities by relating these probabilistic measurements to the periods of time during which they were recorded in the tests and observations, thus fulfilling the algebraic requirement of scientific quantification.

In a second step, these scientific quantifications of individual behaviours were used to generate measurements reflecting individual-specific behaviours. This involved analyses of differential patterns and their temporal stability and stepwise aggregations of the data into more abstract construct measures as described below (Sections 2.8 and 3).

2.6 Assessments of capuchin individuals by human raters: The Capuchin Personality Inventory (CPI)

To study raters' impressions and mental representations of individual capuchins, we operationalised the working constructs equivalently with two assessment formats phrased in English, one comprising behaviour-descriptive verbs and one trait-adjectives. All items were discussed with several capuchin experts to improve item readability and clarity and to reduce ambiguities. Because raters assessed many (up to 27) individuals, we had to minimise the number of items[3]. Our goal was to develop instruments that are applicable within the constraints of animal research yet still allow assessments of all major domains of behaviour frequently studied in capuchin monkeys to be covered. Smaller sets of items necessarily compromise the ability to maximise a scale's internal reliability as widely discussed, amongst others, in research on very brief measures of the Big Five Model (Donnellan et al., 2006; Gosling et al., 2003).

2.6.1 Assessments using behaviour-descriptive verb items (A–BV)

Each working construct was operationalised with one to three verb-based sentences describing behavioural and situational events that are specified in the *behaviour*$_x$*situation-units* used to generate and define a construct. These statements involve less complex processes of mental construction and inference than abstract and decontextualised trait-adjectives (TA). For example, Gregariousness was operationalised with "[Name] sits close together with other members of the group". We operationalised 11 constructs with two items,

[3] Note that the meanings that can be conceived for BR$_x$BS-Approach-generated working constructs are narrower than those that are conceivable for broad "personality" factors. This also becomes apparent in our findings that the items of several working constructs loaded on the same factor and that the composite measures of the working constructs were substantially more internally consistent than the factor measures that were extracted from all items (Section 3.5).

9 constructs with one item and one with three items. Overall, we constructed 34 behaviour-descriptive verb items of which four could be reversed in meaning without using negation (e.g., "During the day, [Name] spends much time on his/her own" to operationalise the low end of Gregariousness). All items are provided in Table S3 in the Supplemental Material.

2.6.2 Assessments using trait-adjective items (A–TA)

Each construct was operationalised with a trait-adjective and without stating specific behaviours or situations. For example, Social orientation was operationalised with "[Name] is friendly to group members". Trait-adjectives denote abstract and decontextualised ideas that are distant from immediate perceptions of behavioural and situational events. Therefore, trait-adjective items involve more complex processes of mental abstraction and construction than behaviour-descriptive verb items (BV) and may also encode more heterogeneous meanings. In total, we constructed 21 trait-adjective items; none of them were reversed in meaning to avoid negations. All items can be found in Table S4 in the Supplemental Material.

2.6.3 Subjective quantifications of individual-specificity

In their assessments, the raters were required to judge how frequently the target individual typically showed the behaviours described in the item statements in relation to other individuals, thus to *directly quantify* the target's individual-specific patterns. In both formats, assessments were indicated on a 5-point frequency scale with the answer categories *(1) hardly ever, (2) rarely, (3) sometimes, (4) often* and *(5) almost always*. Following established practices for generating quantitative data with assessment methods, we recoded these lexically encoded answer categories into numerical ones (as indicated by the numerals placed in parentheses next to each category).

2.6.4 Assessment procedure and repetitions

Assessments were collected via the Internet portal *www.primate-personality.net*. Because all raters repeatedly used the same items to assess different capuchin individuals, we presented the inventories in an interactive PHP- and SQL-based user interface that provided a personalised online presentation for each rater. The programming also inserted the name of the target monkey into the wording of each single statement to help the raters focus on the particular capuchin individual who was being assessed. In the instructions, raters were explicitly cautioned not to discuss their assessments with the other raters at their institution. All 34 behaviour-descriptive verb statements and all 21 trait-adjective statements were presented together in a fixed randomised order in chunks of five items to avoid cross-checking between responses to items with related content. The programme automatically randomised the order in which each rater assessed his or her particular set of monkey individuals to avoid effects of familiarisation with the inventories on the assessments of single capuchins. At the ISTC-CNR, ratings occurred at the end of the first study block of tests and observations and again after an interval of about 4 weeks. Each time, raters were asked to assess how the monkeys were *currently behaving*. Three of these raters recorded behaviours for this study, and 10 persons did not.

2.7 Raters' interpretations of the questionnaire items: Fields of meanings

To explore the meanings that raters may construct for the assessment items, we asked six of the ISTC-CNR raters to describe their spontaneous interpretations of these items in open-ended statements. To achieve maximum independence from the assessments, item interpretations were collected one year after these persons had assessed capuchins for this study. The interpreters were explicitly cautioned not to discuss their interpretations with the others. All interpretations were collected in writing, first for the behaviour-descriptive verb items and then for the trait-adjective items. To avoid effects of familiarisation with the task on the interpretations of the single items, we presented the items from each inventory in a different order for half of the interpreters. Raters could describe their interpretations in phrases or sentences using about one to three page lines per item

statement. We used different instructions for the two kinds of items to enable illuminative contrasts to be applied.

2.7.1 Interpretations of behaviour-descriptive verb items (I–BV)

To collect interpretations of the behaviour-descriptive verb items, we asked: ”How would you describe a capuchin individual who displays the described behaviours more often than other capuchins? What type of individual is this in your opinion?”. Thus, these instructions asked raters to provide global interpretations that would have a greater likelihood of being adjectives than descriptions of behaviours or contexts.

2.7.2 Interpretations of trait-adjective items (I–TA)

For the trait-adjective items, we asked: “How does a capuchin individual described as being more [e.g., physically active] than other capuchins typically behave? What behaviours and situations come to your mind?”. Thus, raters were required to provide more specific interpretations that would more likely be descriptions of behaviours or contexts than adjectives.

2.8 Data aggregation and data analyses

2.8.1 Technical terminology

The TPS-Paradigm adopts a more technical terminology than commonly used in “personality” research (Uher, 2013). A precise terminology is needed to refer unambiguously to different concepts and kinds of phenomena and to the different levels of aggregation where they are being described and analysed (e.g. behavioural measurements, behavioural composite construct measures, mean rating scores). Terms that are insufficiently defined or used inconsistently were shown to be a major source of conceptual misunderstandings between biological and psychological researchers of “personality” (Uher, 2011a). A glossary of terms relevant for this study is provided in Table 2 the Supplemental Material.

2.8.2 Levels of aggregation

Behavioural measures. Within each study block, we accumulated the *behavioural raw measurements* to derive $N = 146$ *contextualised behavioural measurements* that reflected individuals’ absolute time-relative probabilities for showing specific behaviours in specific situations. Through *z*-standardisation, we obtained individuals’ differential scores of time-relative probabilities that were then aggregated across all construct-relevant behaviours and situations into decontextualised *behavioural composite construct measures* (for contextualised analyses, see Uher et al., 2013a). To identify individual-specific patterns, we analysed the test-retest reliability between study blocks for the $N = 146$ contextualised behavioural measurements and the $N = 20$ decontextualised composite construct measures. Finally, the behavioural construct measures were aggregated across the two blocks.

Assessment measures. The *assessment raw scores* were comprised of each single rater’s assessments of each capuchin on each item within each study block (two blocks at the ISTC-CNR, one block at all other institutions). We aggregated these raw scores on different levels. First, we computed each capuchin’s *assessment mean score* per item across raters within each block. For some analyses, the scores on the behaviour-descriptive verb assessments were further aggregated on the construct level, thereby reversing the scores for some items so that they could share the same meaning. These aggregates are referred to as *assessment-based composite construct measures*. Because trait-adjective assessments comprised just one item per construct, the corresponding scores on the trait-adjectival construct measures were identical to their assessment mean scores. For the ISTC-CNR capuchins, we also computed mean scores on the item and construct levels across the two study blocks. Finally, all capuchins’ assessment mean scores on all single items were statistically summarised into *assessment factor scores*.

Interpretation measures. The raters’ interpretations of the assessment items represent *textual raw materials* that were reduced in three steps. Each open-ended statement provided for a given item was broken down into *lexical elements* comprising either

one adjective (e.g., “despotic”) or one behaviour-descriptive verb (e.g., “jumps”) and, if provided, a contextual description (e.g., “at the mesh”). For each item, the lexical elements were pooled across all interpreters and then further reduced by identifying all *unique* lexical elements and by coding them into the 21 working constructs. Quantitative data were generated, first, by counting the occurrences of all lexical elements consisting of identical words (i.e., the occurrences of all unique lexical elements) and then, after coding, by counting the occurrences of all lexical elements that were encoded into the same constructs.

2.8.3 Analyses of behavioural data, assessment data and interpretation data

On these levels of aggregation, we explored the matrices of *i* individuals by *j* variables for both the behavioural and the assessment data using variable-oriented analyses.

First, we analysed the *measurement reliability* for the behavioural raw measures in terms of agreement between two persons who coded 20% of all test sessions and who recorded 15% of all observations independently from one another; inter-coder and inter-observer reliability, respectively, were explored using intra-class correlation coefficients. The measurement reliability of the assessment raw scores was explored in terms of inter-rater agreement within each study block by computing intraclass correlation coefficients; the reliability of single assessments is indicated by *ICC*(3,*1*), and the mean reliability of the *k* assessments per capuchin monkey is indicated by *ICC*(3,*k*) (Shrout & Fleiss, 1979).

We analysed inter-rater agreement in both the pattern of the items’ mean scores across all capuchins (average patterns) and the relative ordering of the capuchins on the single items (differential patterns). Because raters could not have compared their target capuchins with all other capuchins in our sample, we analysed inter-rater agreement, first, separately for each institution and, second, across all institutions. This also allowed us to disentangle possible response biases (e.g., at some institutions, raters may have scored their capuchins generally higher than raters at other institutions) that could artificially inflate inter-rater agreement scores in the complete sample.

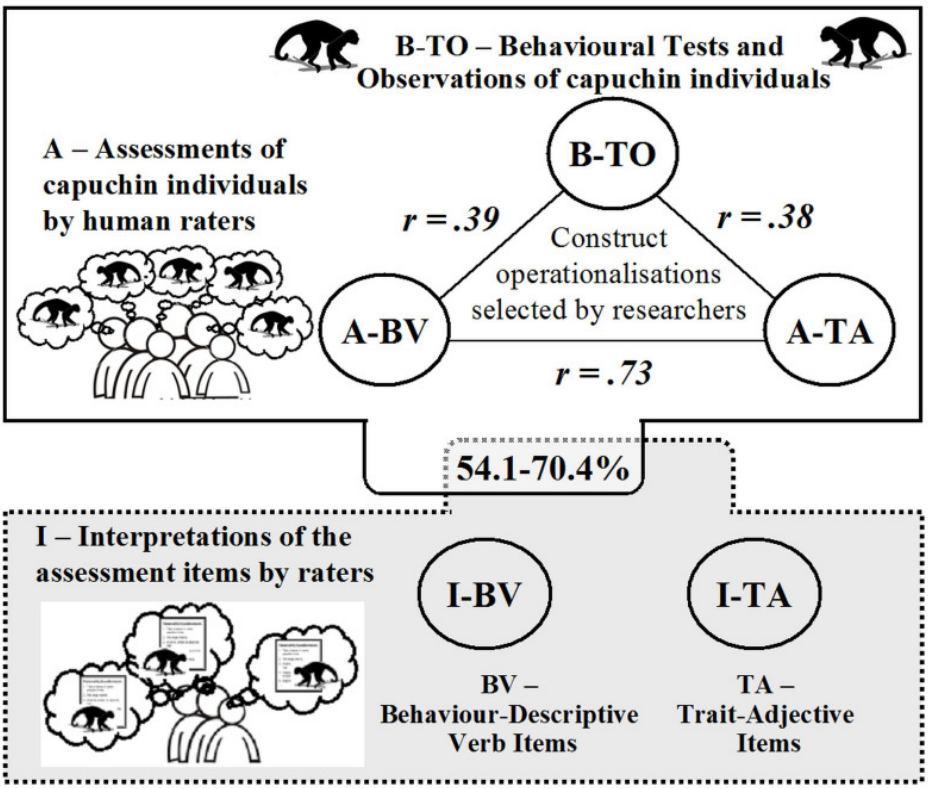
To analyse individual-specificity in the ISTC-CNR capuchins, we studied the *temporal reliability* between study blocks for the behavioural measures and the two assessment measures on different levels of aggregation using Pearson test-retest correlations and compared their magnitudes of temporal reliability with one another. Using the three composite construct measures aggregated across the two study blocks, we analysed the *cross-method coherence* and explored how raters may have arrived at their adjectival assessments by applying a mediation analysis. Then, we investigated the *taxonomic structure* of the differential patterns in the assessment data of all 150 capuchin monkeys.

We compared the internal reliability (internal consistency) of the factor measures thus constructed with that of the behavioural and the assessment-based measures on the level of working constructs. To unravel possible assessment biases, we explored associations of the behavioural composite measures and the two different assessment measures with the capuchins’ socio-demographic factors (age, sex, rearing history) on the level of working constructs. We also explored the fields of meanings that were reflected in the raters’ item interpretations by conducting a content analysis, a method for analysing the content of textual materials by encoding key words (lexical elements) and quantifying their occurrences in textual data (Bauer & Gaskell, 2000; Weber, 1990).

To compute mean correlations, to test correlation scores for differences between methods and rater groups and to explore associations with socio-demographic factors, we always used Fisher’s *r*-to-*Z* transformation. We calculated the magnitude of differences between groups of individuals or groups of variables with Cohen’s effect size *d* with pooled standard deviations (Cohen, 1992). These effect sizes can also be interpreted in terms of the percentage of non-overlapping score distributions between the two contrasted groups (Cohen, 1988). Given the small sample sizes available for some analyses, we computed post-hoc power analyses using the G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007) to explore which of the results were likely to be replicable in larger samples.

All data are available upon request from the first author (JU).

Figure 1 Research design and coherence between all five methods



Note. Upper part: Cross-method coherence between Behavioural Tests and Observations (B-TO), Assessments using Behaviour-descriptive Verb items (A-BV) and Trait-Adjective items (A-TA) across the nomological networks of 20 working constructs as defined by the researchers. Lower part: The researchers' operationalisations of all 21 BR_xBS-Approach-generated working constructs (including the youngster-related construct) and six raters' open-ended interpretations of the behaviour-descriptive verb items (I-BV) and the trait-adjective items (I-TA) showed an overlap of 54.1% when the interpretations were encoded into the 21 working constructs. When working constructs of highly similar (sometimes inversed) meanings (e.g., Distractibility and Persistency) were considered together, the overlap was 60.5%. When raters' item interpretations were set in relation to the five factor-analysed assessment constructs, which each summarise several working constructs, the overlap was 70.4%.

3. Results

3.1 Measurement reliability

Inter-coder and inter-observer agreement in the recording of individual behaviours in the tests and observations were high; the median was ICC(3,*2*) = .89 (range .71 – .98).

Inter-rater agreement was substantial. On the *level of institutions*, the k = 1 to 6 raters (M_k = 2.6, SD_k = 0.74) agreed substantially about both the average pattern of the item means across all capuchins and the differential patterns for each item. On average, across all institutions, the interrater agreement on the *item means* was *ICC(3,1)* = .621 (range .354 to .827) and *ICC*(3,*k*) = .837 (range .522 to .914) for the behaviour-descriptive verb items and *was ICC(3,1)* = .694 (range .464 to .879) and *ICC*(3,*k*) = .877 (range .634 to .935) for the trait-adjective items. Interrater agreement about the *differential patterns* for each item was on average *ICC(3,1)* = .435 (range .235 to .530) and *ICC*(3,*k*) = .636 (range .537 to .766) for the behaviour-descriptive verb items, and it was *ICC(3,1)* = .443 (range .291 to .592) and *ICC*(3,*k*) = .625 for the trait-adjective items (range .501 to .786; scores on all single items are provided in Tables S3 and S4 in the Supplemental Material). At the ISTC-CNR, inter-rater reliability 4 weeks later was virtually identical: *ICC(3,1)* = .327 and *ICC*(3,*k*) = .653 for the behaviour-descriptive verb items and *ICC(3,1)* = .354 and *ICC*(3,*k*) = .664 for the trait-adjective items.

For some items, the inter-rater reliability for differential patterns was low, *ICC*(3,*k*) < .50; specifically, it was low for the behaviour-descriptive verb items operationalising Arousability *ICC*(3,*k*) = .29 to .48, Curiousness *ICC*(3,*k*) = .35 to .48, Cleanliness *ICC*(3,*k*) = .26, and Persistency *ICC*(3,*k*) = .26, and for one item from each of the operationalisations of the constructs Food orientation[4] *ICC*(3,*k*) = .19, Physical activity *ICC*(3,*k*) = .40, and Social orientation *ICC*(3,*k*) = .25. But none of these items generally lacked inter-rater agreement at all institutions; rather, inter-rater agreement for these items varied considerably between institutions. At the ISTC-CNR, the items operationalising Arousability, Cleanliness and Physical activity showed low inter-rater reliabilities in both study blocks. Interrater agreement was also low for the five trait-adjective items operationalising Anxiousness *ICC*(3,*k*) = .420, Distractibility *ICC*(3,*k*) = .320, Gregariousness *ICC*(3,*k*) = .420, Cleanliness *ICC*(3,*k*) = .001, Social orientation *ICC*(3,*k*) = .440 and Vigilance *ICC*(3,*k*) = .340. But because none of these items lacked inter-rater reliability at all institutions and because we wanted to study a sample of assessments that were representative of the raters' ideas about capuchins' individual-specific behaviours (even if they might be inconsistent) rather than a sample of scores that were artificially selected on the basis of test-theoretical assumptions, we retained all items for the subsequent analyses.

On the level of *the sample of all of the 150 capuchins at all institutions,* inter-rater agreement on the *item means* was *ICC(3,1)* = .630 and *ICC*(3,*k*) = .872 for the behaviour-descriptive verb items, and it was *ICC(3,1)* = .640 and *ICC*(3,*k*) = .877 for the trait-adjective items. Cross-institutional inter-rater agreement in the *differential patterns* of all 150 capuchins on the single items was *ICC(3,1)* = .470 and *ICC*(3,*k*) = .635 for the behaviour-descriptive verb items and *ICC(3,1)* = .454 and *ICC*(3,*k*) = .617 for the trait-adjective items[5]. Although virtually identical to the averages for the institution-specific reliability scores, these results have to be considered with caution because they ignore the fact that no rater could have known all of the individuals in the entire sample. Thus, in their assessments, raters could not have considered all capuchins in this multi-institutional sample in order to make relative comparisons between individuals. This is seldom considered in animal studies, most of which report only cross-institutional reliabilities (e.g., Morton et al., 2013). In fact, substantial inter-rater agreement across institutions could also be simply based on systematic mean-level differences in the assessments between institutions (as explored in Section 3.5.1).

3.2 Temporal reliability: Identifying individual-specificity

3.2.1 Behavioural measures (B–TO)

Temporal reliability was high at the different levels of aggregation. The test-retest correlations for the *N* = 146 variables of contextualised behavioural measurements was on average r_m = .60 (range -.09 to .99). Of these, 86 variables met the significance criterion (p < .05) with an average test-retest correlation of r_m = .74 (range .43 to .99). Of the behavioural construct measures that were composed *exclusively of temporally reliable* measurements, 19 measures showed significant test-retest reliability (p < .05) with an average of r_m = .76 (range .47 to .91). Of the behavioural construct measures composed of both temporally reliable *and* non-reliable contextualised behavioural measurements, 18 constructs showed

[4] Given their origins in the behaviour-scientific knowledge base, BR$_x$BS-Approach-generated constructs are labelled with terms that are much less colloquial than those derived from humans' everyday languages. This meets efforts to reduce the impact of implicit meanings and anthropomorphic biases (see the Supplemental Material; Uher, 2015b; Uher et al., 2013a).

[5] The present level of inter-rater agreement, for direct comparisons between studies relying on different numbers *k* of raters considered in terms of the *ICC(3,1)*, was substantially higher than those obtained in other capuchin "personality" studies for assessments using trait-adjective items adapted from a human Five-Factor Model inventory (*ICC*(3,*1*) = .36; Morton et al. 2013) and trait-adjective items taken from studies on various other primate species (*ICC*(3,*1*) = .09; Manson & Perry, 2013), which showed almost zero agreement between raters. However, the levels of inter-rater agreement found in the current study were comparable to those obtained for assessments of BR$_x$BS-Approach-generated constructs in crab-eating macaques (Uher et al., 2013b) but lower than those obtained for assessments of great apes (Uher, 2011b; Uher & Asendorpf, 2008).

significant test-retest reliability, with an average of r_m = .66 (range .22 to .91). As expected, the differences between these two composite measures were significant, *t*-test for dependent samples, $t(18) = 3.172$; $p = .005$. Nevertheless, we retained all variables and constructs for our present analyses because we wanted to obtain representative measures that reflected the true-to-life patterns of individuals' behaviours more accurately than those obtained from variables that were selected for significant test-retest correlations and that artificially inflated the true consistency of observable behaviours[6].

3.2.2 Assessment measures (A-BV, A-TA)

The temporal reliability of the "personality" assessments between the two study blocks in the ISTC-CNR sub-sample was high. The average test-retest correlation of the differential patterns on each single item was r_{tt} = .74 (*SD* = .20, range .07 to .93) for the behaviour-descriptive verb items and r_{tt} = .77 (*SD* = .18, range .16 to .93) for the trait-adjective items. Four items lacked test-retest reliability; these were comprised of one behaviour-descriptive verb item from each of the operationalisations of the constructs Anxiousness (r_{tt} = .32, p = .109, N = 27), Cleanliness (r_{tt} = .16, p = .419, N = 27) and Physical activity (r_{tt} = .31, p = .119, N = 27) and the trait-adjective item for Cleanliness (r_{tt} = .07, p = .730, N = 27). Given that the inter-rater reliabilities varied considerably between institutions, it is likely that the test-retest reliabilities would also have varied between institutions. Thus, the lack of reliability in these items may be due more to the small sample sizes available at each institution than to problems in the items themselves. This assumption builds on findings from a previous study of N = 104 crab-eating macaques with an analogous research design and six waves of data collected over three years (Uher et al., 2013b). In that study, single items showed only low inter-rater or test-retest reliability, but no item generally lacked reliability in any of the years of the study. Given this and because it was our aim to obtain a representative picture of the raters' impressions of capuchin individuals even if they might be inconsistent at times, we retained all items for the subsequent analyses including a few items that showed only low temporal reliability.

3.2.3 Comparison of temporal reliability between capuchins' individual-specific behaviours and human raters' assessments on the two formats

The parallel collection of behavioural measures and assessment measures on the same constructs for the same individuals permitted direct comparisons of the temporal reliability of these measures. We compared the temporal reliability scores between the 146 behavioural measurements, the 20 behavioural construct measures composed of all measurements, the 32 behaviour-descriptive verb assessments and the 20 trait-adjective assessments on the single item level. Note that the behavioural measures constitute scientific quantifications in terms of differentially standardised time-relative probabilities, whereas the assessment measures constitute subjective quantifications (see Sections 1., 2.5.1 and 2.6.3).

The temporal reliabilities of these four kinds of measures differed significantly as indicated by a one-way ANOVA, $F(3,220) = 5.567$, $p = .001$. Bonferroni tests showed that the 146 behavioural measurements were significantly less reliable than the two assessment measures (p = .013); additional differences were substantial, as indicated by Cohen's effect size *d*, though not significant. Specifically, the behavioural measurements were much less temporally reliable than both the trait-adjective assessments (d = -0.77; 45.0% non-overlapping score distributions and 89% achieved power for detecting such a difference) and the behaviour-descriptive verb assessments (d = -0.61, 38.2% non-overlapping, 89% achieved power) and slightly less temporally reliable than the behavioural composite construct measures (d = -0.24; 17.3% non-overlapping, 17% achieved power). In turn, the

[6] The temporal reliability scores of all single contextualised behavioural measurements, all single composite construct measures (both contextualised and decontextualised) composed of all or of only temporally reliable measurements, temporal reliability scores of individual behavioural (response) profiles, individual situation-behaviour profiles within constructs and "personality" profiles across all working constructs along with findings on cross-situational consistency are reported in Uher et al. (2013a).

behavioural composite construct measures were less reliable than both the trait-adjective assessments (d = -0.65; 41.7% non-overlapping, 52% achieved power) and the behaviour-descriptive verb assessments (d = -0.45; 30.4% non-overlapping, 35% achieved power). The temporal reliabilities of the two assessment measures were hardly different (d = 0.16; 11.2% non-overlapping, 8% achieved power).

3.3 Validity of assessments: Cross-method coherence on the level of working constructs

On the construct level, aggregated across study blocks, we explored the relationships between capuchins' individual-specific behaviours and raters' pertinent mental representations in terms of the correlations between the behavioural and the two assessment-based composite construct measures.

Across all of the working constructs that we studied, coherence between the three methods studied with Pearson correlations r was substantial and differed significantly from zero when one-sample t-tests were applied: $t_{A\text{-}BV–A\text{-}TA}(19) = 8.752$, $p = .000$; $t_{A\text{-}BV–B\text{-}TO}(19) = 4.618$, $p = .001$; $t_{A\text{-}TA–B\text{-}TO}(19) = 4.835$, $p = .001$ (see Table 2). The strength of coherence between methods differed significantly. Across all 20 constructs, the two assessment methods were significantly more strongly correlated with one another (mean $r_{A\text{-}BV–A\text{-}TA} = .73$) than the behaviour-descriptive-verb assessments were with the behavioural construct measures (mean $r_{A\text{-}TA–B\text{-}TO} = .39$; $t_{A\text{-}BV\text{-}A\text{-}TA–B\text{-}TO\text{-}A\text{-}BV}(19) = 5.530$; $p = .000$) and the trait-adjective assessments were with the behavioural construct measures (mean $r_{A\text{-}BV–TO} = .38$; $t_{A\text{-}BV\text{-}A\text{-}TA–B\text{-}TO\text{-}A\text{-}TA}(19) = 4.753$, $p = .000$). The behaviour-descriptive verb assessments and the behavioural construct measures did not show significantly higher correlations with each other than the trait-adjective assessments and the behavioural composite measures ($t_{A\text{-}BV\text{-}B\text{-}TO–A\text{-}TA\text{-}B\text{-}TO}(19) = -0.211$, $p = .835$). The cross-method coherence across all 20 "personality" constructs is depicted in the upper part of Figure 1.

For some constructs, such as Aggressiveness to conspecifics, Curiousness, Playfulness and Sexual activity, all three kinds of operationalisations were substantially correlated within their nomological networks. For some other constructs, such as Anxiousness, Competitiveness, Impulsiveness and Vigilance, the two assessment measures had high correlations with one another but not with the behavioural measure, indicating incongruencies in the meanings of the assessments and the particular behaviours and situations under study. For other constructs, such as Gregariousness, the behavioural measures were correlated with just one of the two assessment measures. In Arousability and Physical activity, coherence was generally low between all three kinds of operationalisations (Table 2). These findings point to important differences between behavioural and assessment measures that are explored in more detail below (Section 3.6).

3.4 Mediation analyses: How raters may have developed impressions of the capuchins' individual-specificity ("personality")

We analysed potential pathways for how the raters may have formed their impressions and mental representations of the capuchins' individual-specificity ("personality"). Specifically, we explored whether the raters may have developed abstract mental representations of the capuchin individuals (as expressed in the trait-adjective assessments) rather directly from observations of a broad range of behavioural events (as reflected in the behavioural composite construct measures), or whether more specific mental representations that referred to only a few indicative behaviours (as expressed in the behaviour-descriptive verb assessments) may have served as mediators. Partial mediation would be evidenced when, controlling for behaviour-descriptive verb assessments, the behavioural composite construct measures still directly affected the trait-adjective assessments and complete mediation when they no longer directly affected the trait-adjective assessments.

We estimated and tested this model by computing multiple regression analyses for 20 working constructs according to Baron and Kenny's (1986) guidelines. The behavioural

composite construct measures as predictors were significantly correlated ($p < .05$) with both the trait-adjective assessments as criteria and the behaviour-descriptive verb assessments as potential mediators for six constructs and for four additional constructs when (because of the small sample size) we relaxed the significance levels to $p < .10$. For eight of these constructs, multiple regressions of trait-adjective assessments on behavioural composite construct measures and on behaviour-descriptive verb assessments *(a)* showed a significant impact of the mediator (behaviour-descriptive verb assessments) on the criterion (trait-adjective assessments) and *(b)* rendered the effect of the behavioural construct measures on the trait-adjective assessments non-significant. In these cases, the effects of the behavioural construct measures on the trait-adjective assessments were fully mediated by the behaviour-descriptive verb assessments. For one additional construct (Aggressiveness to conspecifics), the behavioural construct measures still directly affected the trait-adjective assessments when controlling for behaviour-descriptive verb assessments, thus fulfilling the criteria for partial mediation (Table 2).

Table 2 Cross-method coherence and mediation analyses on the level of the BR_xBS-Approach-generated working constructs

Working constructs	**Coherence between methods**[a]						**Mediation analyses between methods**[b]			
	B-TO – A-BV		A-BV – A-TA		B-TO – A-TA		B-TO–*(A-BV)*–A-TA		A-BV–*(B-TO)*–A-TA	
Aggressiveness to conspecifics	**.77**	(.000)	**.97**	(.000)	**.81**	(.000)	**.16**	(.042)	**.85**	(.000)
Aggressiveness to humans	**.48**	(.027)	**.66**	(.000)	.42	(.056)	.13	(.507)	**.60**	(.008)
Arousability	.00	(.993)	.24	(.226)	.22	(.279)	.22	(.260)	.32	(.104)
Anxiousness	.12	(.553)	**.75**	(.000)	.23	(.253)	.15	(.317)	**.69**	(.000)
Competitiveness	-.03	(.923)	**.90**	(.000)	.01	(.972)	.04	(.728)	**.94**	(.000)
Creativeness/ inventiveness	**.77**	(.001)	**.87**	(.000)	**.61**	(.016)	-.09	(.709)	**.92**	(.002)
Curiousness	**.57**	(.027)	**.79**	(.000)	**.68**	(.005)	.31	(.100)	**.66**	(.003)
Distractibility	-.11	(.709)	**.42**	(.031)	-.35	(.207)	-.28	(.178)	**.67**	(.005)
Dominance[c]	.43	(.051)	**.40**	(.039)	**.55**	(.009)	**.51**	(.030)	.10	(.636)
Food orientation	**.57**	(.002)	**.68**	(.000)	.33	(.097)	-.06	(.766)	**.69**	(.002)
Gregariousness	**.25**	(.254)	.31	(.112)	**.72**	(.000)	**.68**	(.001)	.17	(.324)
Impulsiveness	.10	(.726)	**.70**	(.000)	.06	(.830)	-.03	(.864)	**.87**	(.000)
Physical activity	-.03	(.894)	.34	(.087)	.01	(.995)	.01	(.980)	.15	(.528)
Persistency	.24	(.393)	**.88**	(.000)	.40	(.141)	.22	(.211)	**.74**	(.001)
Playfulness	**.65**	(.002)	**.80**	(.000)	**.49**	(.025)	-.03	(.865)	**.81**	(.001)
Self-cleanliness	.38	(.088)	**.42**	(.029)	**.51**	(.018)	.27	(.109)	**.62**	(.001)
Social orientation to conspecifics	.25	(.226)	**.78**	(.000)	.13	(.537)	-.07	(.620)	**.79**	(.000)
Social orientation to humans	**.73**	(.000)	**.75**	(.000)	**.62**	(.003)	.09	(.680)	**.74**	(.002)
Sexual activity	**.66**	(.001)	**.89**	(.000)	**.57**	(.008)	-.03	(.865)	**.89**	(.000)
Vigilance	-.12	(.661)	**.72**	(.000)	-.06	(.819)	-.03	(.919)	.29	(.314)

Note. Methods: B-TO behavioural composite construct measures obtained in test and observations, A-BV behaviour-descriptive verb assessments, A-TA trait-adjective assessments. Mediation analyses: B-TO – A-BV Correlations of predictors (B-TO) with potential mediators (A-BV); B-TO – A-TA Correlations of predictors (TO) with potential criteria (A-TA); B-TO – *(A-BV)* – A-TA Regression coefficients of predictors (B-TO) on criteria (A-TA) controlled for mediators (A-BV); A-BV – *(B-TO)* – A-TA regression coefficients of mediators (A-BV) on criteria (A-TA) controlled for predictors (B-TO). Correlations of trait-adjective assessments and behaviour-descriptive verb assessments based on $N = 26$, correlations with behavioural composite construct measures based on $N =$ 15 to 26 capuchin individuals. Significant coefficients are bold; p values in parentheses (correlations one-sided, regression coefficients two-sided). [a] Pearson correlations r; [b] Standardised regression coefficients β in multiple regression equations. Bold coefficients are significant at least at the $p < .05$ level. [c] We studied Dominance as "personality" construct, that is, as individual-specific patterns in dominant-submissive behaviours in which all individuals can be quantified and compared with one another, rather than as social status, which refers to only a few individuals per group (e.g., alpha male status).

3.5 Taxonomic structures

In animal research, the opportunities to fulfil the case-to-variable ratio that is required for structural analyses of between-individual differences are generally compromised. In many animal studies, the number of variables even exceeds the number of cases, thus rendering the results prone to sample biases. Our sample sizes were large enough for us to explore taxonomic structures in the raters' assessments. We did not explore taxonomic structures in the behavioural measures given the much smaller sample that was available for the tests and observations. But our samples allowed us to explore and compare the internal consistencies of the behavioural construct measures and the various assessment-based construct measures (Section 3.5.2).

3.5.1 Exploratory factor analysis of the raters' assessments

We explored the taxonomic structures of the assessments of all 150 capuchins from study block one. The nine sub-samples differed in size and composition; some institutions had more males than females or only males (see Table 1), which may have affected the raters' opportunities to compare individuals. Therefore, we first explored the assessment data for possible between-institution differences. One-way ANOVAs revealed that raters' assessments differed significantly between institutions for 37 of the 52 items, $F(8,139) = 2.08$ to 7.63, $p = .000$ to .041. Given this, we *z*-standardised the assessment data within each institution and then pooled the *z*-standardised scores from all nine institutions into one sample. This was done separately for each item.

Because no item generally lacked inter-rater reliability in all sub-samples (see Section 3.1), all items were included in an exploratory R-factor analysis. To further explore the cross-method relations, we analysed all 52 items jointly. R-factor analysis seeks to construct the smallest number of latent composite variables (factors) that can statistically explain the common variance of individuals' scores on the variables studied; hence, it is a variable-oriented method of analysis. We applied principal axis factoring with oblique promax rotation, which aims to identify simple structures. Unlike orthogonal rotation methods (e.g., varimax), oblique rotation methods allow for possible intercorrelations at the latent factor level. This allows for the consideration that people's mental representations are generally associated with one another in highly complex ways rather than being neatly partitioned into distinct units and this is even more so the case for individuals' behaviours (Allport & Vernon, 1933; Blurton Jones, 1967; Costello & Osborne, 2005; Uher, 2015b; Uher et al., 2013b). Oblique rotation should therefore render a more accurate factor solution for describing the complex structures underlying people's mental representations as reflected in the assessments.

First, we explored the factorability of the assessment data. The Kaiser-Meyer-Olkin measure of sampling adequacy was KMO = .815, well above the commonly recommended value of .60. Bartlett's test of sphericity was significant, $\chi^2 (1326) = 6260.458$, $p = .000$, indicating that the correlations in the data set of 52 items were appropriate for factor analysis. The mean item communality was $h^2 = .75$ (range .46 to .91, all but one communality exceeded $h^2 > .50$; see Table 3), indicating that each item shared some common variance with the other items. Based on principal axis factoring with squared multiple correlations as communality estimates, a parallel analysis and the initial eigenvalues suggested the extraction of 11 factors that explained 73.54% of the variance. The initial eigenvalues from the first five factors (each comprising 5 to 11 items) were 10.83, 8.43, 5.28, 3.51 and 2.39, respectively, corresponding to 20.8%, 16.2%, 10.2%, 6.8% and 4.6% of the explained item variance. The sixth factor (comprising 3 items) had an eigenvalue of 1.7 and explained 3.3% of the variance. Factors 7 to 11 (comprising only 3, 4, 3, 2 and 2 items, respectively) had eigenvalues between 1.0 and 1.4, and each explained from 2% to 2.7% of the variance.

Given that there were many small factors that each explained only a small amount of variance, substantial cross-loadings of many items resulting in several substantial correlations between different factors and the "levelling off" of eigenvalues in the scree plot

after six factors, we examined five- and six-factor solutions that explained 57% and 54% of the variance, respectively. These two solutions yielded highly similar structures. Specifically, the sixth factor comprised only two trait-adjective items operationalising Social orientation to humans and Aggressiveness to humans. But it showed negative internal consistencies, *ICC(*3*,1)* = -.560 and *ICC*(3,*k*) = -2.55, and therefore had to be discarded. The five-factor solution agreed with the graphical elbow in the scree plot; the primary loadings of all items were at least ≥ .30 except for one item (operationalising Cleanliness) with an acceptable primary loading of .29. All five factors together explained 54% of the item variance (Table 3). The five factor measures were moderately correlated with one another, $r_{F1\text{-}F2}$ = .15; $r_{F1\text{-}F3}$ = -.02; $r_{F1\text{-}F4}$ = .21; $r_{F1\text{-}F5}$ = .26; $r_{F2\text{-}F3}$ = .38; $r_{F2\text{-}F4}$ = .29; $r_{F2\text{-}F5}$ = -.13; $r_{F3\text{-}F4}$ = .00; $r_{F3\text{-}F5}$ = -.15; $r_{F4\text{-}F5}$ = -.34; corresponding to a maximum of 14.4% common variance.

The meanings of the items with dominant loadings allowed for clear interpretations of the five factors (see Table 3 below). The first factor, labelled *Dominant-competitive-aggressive*, mainly explained variance in the items operationalising the working constructs Dominance[7], Competitiveness, Aggressiveness to conspecifics, (inverse) Anxiousness, Food orientation, Sexual activity and Impulsiveness. The second factor, labelled *Curious-inventive-persistent*, primarily explained variance in the items operationalising Curiousness, Creativeness/Inventiveness, (inverse) Distractibility, Persistency, Vigilance, (inverse) Anxiousness and Social orientation to humans. The third factor, labelled *Playful-active-impulsive,* mainly explained variance in the items operationalising Playfulness, Physical activity, Impulsiveness, Arousability, Distractibility, Aggressiveness to humans and (inverse) Cleanliness. The fourth factor, labelled *Gregarious-prosocial*, primarily explained variance in the items operationalising Gregariousness and Social orientation to group members. The fifth factor, labelled *Excitable-vigilant*, mainly explained variance in the items operationalising Arousability, Vigilance, Anxiousness, Food Orientation and Aggressiveness to humans.

For 14 out of 20 working constructs, behaviour-descriptive verb and trait-adjective assessments operationalising the same construct loaded highest on the same factor, supporting the finding that assessments in both formats converged notably for most constructs (Section 3.3). For the first four factors, there were two to five working constructs for which all of the items had their highest loadings on the same factor. For Impulsiveness, Vigilance, Distractibility, Arousability and Aggressiveness to humans, items that were supposed to operationalise the same construct loaded highest on two different factors. Anxiousness was the only construct operationalised by items that loaded on three different factors (see Table 3). But considering the item content, many of these split loadings are meaningful. For example, the behaviour-descriptive verb item “In social conflict situations, [Name] screams quickly and flees from others”, meant to operationalise Anxiousness, had a high negative loading (-.76) on the *Dominant-competitive-aggressive* factor. Thus, in the raters’ view, capuchin individuals scoring high on this factor seldom screamed quickly and fled from others; this is often the case for dominant and aggressive individuals. The second behaviour-descriptive verb item operationalising Anxiousness, “[Name] keeps a distance from unknown objects, persons, and/or avoids uncertain situations”, showed a moderate negative loading (-.56) on the *Curious-inventive-persistent* factor. Thus, raters mentally represented capuchins scoring high on this factor as individuals who also tended to approach unknown objects and persons and did not avoid uncertain situations. The trait-adjective item (“anxious”) loaded highest (.58) on the fifth factor *Excitable-vigilant*.

The behaviour-descriptive verb item “When [Name] does not get his/her food or reward immediately, he/she quickly bangs against the mesh or tries to get it forcefully”, used to operationalise Impulsiveness, had a moderately high loading (.48) on the *Dominant-competitive-aggressive* factor. The food-related behaviours described in this item are well-matched with the behaviours described in the three Food orientation items that also had their highest loadings on this factor. The behaviour-descriptive verb item “[Name] can focus for a

[7] We studied Dominance as a “personality” construct, that is, as individual-specific patterns in dominant-submissive behaviours in which all individuals can be quantified and compared with one another, rather than as social status, which refers to only a few individuals per group (e.g., alpha male status).

long time on activities that take effort and time", meant to operationalise inverse Distractibility, had a high loading (.69) on the *Curious-inventive-persistent* factor. This item shares the meaning that can be conceived for the working construct Persistency, which also had its highest loadings for both of its items on this factor. The behaviour-descriptive verb item "[Name] quickly spots small food items, potential prey or changes in the environment", meant to operationalise Vigilance, also had a high loading on the *Curious-inventive-persistent* factor (.69), which matches well with the meanings of the three Curiousness items and the two Creativeness/Inventiveness items, which also had their highest loadings on this factor. The three trait-adjective items operationalising Impulsiveness, Arousability and Distractibility had moderate to high loadings (.58 to .70) on the *Playful-active-impulsive* factor. They reflect a field of meanings also captured by the behaviour-descriptive verb item "When somebody stands in front of the cage, [Name] jumps at the grate and may also try to grab that person", which was meant to operationalise Aggressiveness to humans and also had a moderately high loading on this factor (.57).

The behaviour-descriptive verb items of Arousability "When awaiting the feeding, [Name] paces restlessly and/or scratches him/herself" and "When there are unusual noises outside the cage, [Name] starts pacing and/or scratching" both had moderately high loadings on the *Excitable-vigilant* factor (.52 to .62). This finding supports the meanings that can be conceived for the behaviour-descriptive verb item "[Name] watches everything around him/her closely", meant to operationalise Vigilance, and the pertinent trait-adjective "vigilant", which loaded on this factor (.55 to .59). This factor reflects a field of meanings that may also include the meaning of the trait-adjective "anxious", which also loaded on this factor (.58). The trait-adjective item of Aggressiveness to humans had moderate loadings on the *Dominant-competitive-aggressive* factor (.44) and the *Excitable-vigilant* factor (.47). Given that the pertinent behaviour-descriptive verb item had moderate loadings on both the *Playful-active-impulsive* factor (.57) and the *Curious-inventive-persistent* factor (.51), these split loadings may indicate that these items refer to different kinds of Aggressiveness to humans that the raters conceived for these monkeys.

3.5.2 Internal reliability of behavioural versus assessment-based composite measures

For structural comparisons between behavioural and assessment-based measures, we analysed and compared their internal reliabilities. The *behavioural composite construct measures* of the first study block had low to moderate internal consistencies; on average across 17 working constructs, they were $ICC(3,k)$ = .632 (range -.131 to .859) and $ICC(3,1)$ = .268 (range -.040 to .750). Behavioural construct measures that were composed of a greater number of *k* measurements were not more internally consistent than those composed of fewer measurements (r = .17, p = .515).

Given that each working construct was operationalised with just one trait-adjective item and that assessments in the two formats were strongly interrelated (see Sections 3.3 to 3.5), we analysed the internal consistencies of *assessment-based construct measures composed of both behaviour-descriptive verb items and trait-adjective items*. Their average internal consistency was $ICC(3,k)$ = .731 (range .483 to .907) and $ICC(3,1)$ = .526 (range .276 to .796) across 20 working constructs. For comparison, we also explored the internal consistencies of the *factor-analysed assessment measures* that comprised items operationalising different working constructs (see Section 3.5). Across the five factors, the average internal consistency was $ICC(3,k)$ = .746 (range .593 to .837) and $ICC(3,1)$ = .237 (range .139 to .322).

Between these three kinds of composite measures, the internal consistencies of the average measurements $ICC(3,k)$ did not differ, $F(2,39) = 1.689$, p = .198, indicating that differences in the number *k* of measurement variables of which these different construct measures were composed did not affect their internal consistencies. But the internal consistencies of the single measurements $ICC(3,1)$ differed significantly, $F(2,39) = 8.730$, p = .001. Bonferroni tests showed that the assessment-based composite measures of the working constructs were significantly more internally reliable than those of both the behavioural construct measures (d = 1.62) and the factor-analysed assessment-based

construct measures (*d* = 1.77), indicating greater heterogeneity in the measurements of which these latter two were composed.

3.6 Associations of the capuchins' socio-demographic factors with their individual-specific behaviours and how these were mentally represented by the human raters

On the level of the working constructs, we explored associations of the capuchins' age, sex and early rearing history with their individual scores on the behavioural construct measures, the behaviour-descriptive assessments and the trait-adjective assessments (each aggregated across study blocks). For each of these three methods, we calculated the magnitude of the difference in the scores between males and females (11;16), between two age groups created by median split (14;13) and between mother-reared and hand-reared individuals (15;12).

For many constructs, we found substantial cross-method coherence in these socio-demographic associations. For example, the males' higher Aggressiveness to conspecifics found in the behavioural construct measures (*d* = 1.69) was also reflected in the pertinent behaviour-descriptive verb assessments (*d* = 1.06) and trait-adjective assessments (*d* = 1.29). In the behavioural measures, age-group differences were absent for 19 working constructs, and this was also reflected in the pertinent assessment-based measures of 18 of these constructs (see Table 4 below).

In some cases, between-group differences found in the behavioural measures were also reflected in the raters' assessments but differed in magnitude. For example, males' behaviours were substantially more dominant than those of the females (*d* = 1.68); raters' assessments reflected this difference but in much less pronounced ways (*d* = 0.42 to 0.90). Compared with mother-reared capuchins, hand-reared capuchins behaved much less aggressively towards humans (*d* = -1.79) and spent less time in close proximity to conspecifics (*d* = -1.17). These differences were also reflected in raters' behaviour-descriptive and trait-adjectival assessments but in less pronounced ways (*d* = -0.73 to -0.54 and *d* = -0.85 to -0.26, respectively) and usually with only low power for detecting such differences (see Table 4).

But there were also profound divergences between the three methods. Some group differences found in the capuchins' behavioural measures were not reflected in the raters' pertinent assessments. For example, younger capuchins tended to behave much more impulsively than older ones (*d* = 2.00), but the assessment-based measures did not reflect this substantial age difference (*d* = -.09 to 0.02). Similarly, compared with mother-reared individuals, hand-reared capuchins were more easily distracted (*d* = 1.37) by humans who produced noises near their cages, but these differences were not reflected in the raters' pertinent assessments (*d* = -0.29 in both formats). Conversely, we also found associations reflected in the raters' assessments that did not emerge in the behavioural measures. For example, males were assessed to be more competitive, more curious, more food-oriented and more persistent than females but this was not found in the behavioural measures. Hand-reared capuchins were assessed to be cleaner and less sexually active than mother-reared capuchins, but such differences were not reflected in the behavioural measures (see Table 4).

Some assessments reflected even *inverse* associations with socio-demographic factors. For example, males were assessed to be considerably less anxious than females (*d* = -1.29 to -1.33) but this was not found in the behavioural measures that rather showed a tendency for males to be more anxious than females (*d* = 0.48; but with only 31% power for detecting such a difference). Such attributions of sex differences that went in the direction opposite the ones found in the behavioural measures also occurred for the constructs Arousability, Impulsiveness, Physical activity and Vigilance. Hand-reared individuals were assessed to be much less competitive than mother-reared individuals, but the behavioural measures tended to show the opposite pattern (see Table 4 for all effect sizes and power estimations).

3.7 Content analyses of raters' item interpretations

To further explore these deviances between the behavioural and assessment-based construct measures, we analysed raters' open-ended interpretations of the item statements using content analysis. These textual materials were broken down into lexical elements that were each composed of either one behaviour-descriptive verb or one trait-adjective (e.g., "moves quickly"; "despotic"); these are called *unique lexical elements*. Lexical elements were reworded in the third-person singular form when necessary. Enumerations of different behaviours or different trait-adjectives in the interpretation of the same item statement were split into separate lexical elements even if they had similar meanings (e.g., "dominant", "assertive"). Only pieces of contextual information of similar content that were provided for the same verb or the same adjective were summarised by enumeration (e.g., "for resting, to stay").

Altogether, we identified 490 unique lexical elements. Per item statement, raters' interpretations contained an average of $M = 9.65$ different lexical elements ($SD = 2.98$), ranging from 4 to 18 elements. Thus, the interpretations of some items contained only a few different elements, each of which was mentioned in identical ways by multiple interpreters. But the interpretations of other items contained many different lexical elements, both within and between interpreters. For example, the behaviour-descriptive verb item "When there is food, [Name] is quickly on the spot", meant to operationalise Food orientation, was interpreted by one and the same rater as describing capuchins who are "alert, vigilant, fast, dominant". This interpretation reflects a much broader field of meanings covering more diverse behaviours than just feeding-related behaviours. Interpretations also varied between raters; rater 2 interpreted monkeys showing the behaviours described by this item more often than others as "reactive", rater 3 as "dominant, hungry, voracious", rater 4 as "gluttonous", rater 5 as "dominant, no fear" and rater 5 as "active, possibly more dominant". All lexical elements and the frequencies of their occurrence per item are provided in Tables S3 and S4 in the Supplemental Material.

To further explore the fields of meanings that are reflected in the raters' item interpretations, we coded the meaning of each lexical element using the 21 BR_xBS-Approach-generated constructs (including the youngster-related construct) and their definitions and operationalisations as used in the observations and assessments. Intra-coder reliability, analysed by applying independent codings obtained after a 6-month break, was excellent ($\kappa = .94$). Of the 490 unique lexical elements, 396 elements could be clearly assigned to one construct, 32 were assigned to two constructs (e.g., "despotic" was assigned to both Aggressiveness to conspecifics and Dominance) and five lexical elements were assigned to three constructs (e.g., "starts conflicts over food or foraging places with others" was assigned to Food orientation, Competitiveness and Aggression). Fifteen lexical elements specified information that was not directly reflected by the BR_xBS-Approach-generated working constructs (e.g., "precise", "accurate", "obsessive", "probably is a relative"), and these elements were therefore excluded. In the coding, we also considered the idea that some lexical elements reflected a construct's inverted meaning (e.g., "subordinate" was encoded as inverse Dominance). Finally, for each given item, we summarised the number of lexical elements that were encoded into each of the 21 working constructs, considering the elements that reflected inverted meanings separately.

In the interpretations of all items, the 490 unique lexical elements occurred a total of 821 times. In 444 cases, the lexical elements were encoded into the working construct that the interpreted item was supposed to operationalise. In 497 cases, lexical elements were assigned to constructs of highly similar meaning, as can be conceived for the constructs Curiousness and Creativeness/Inventiveness, the constructs Gregariousness and Social orientation, and the constructs Distractibility and (inversed) Persistency. Thus, overall, 54.1 to 60.5% of all lexical elements contained in the raters' item interpretations corresponded to the meanings that we, as the researchers, constructed for these items.

But 39.5 to 45.9% of the lexical elements contained in the raters' item interpretations referred to other constructs that we did not intend to be measured by the given items. This

proportion is substantial. Even when compared with the constructs derived by factor analysis describing the latent structure of 49 persons' assessments of 150 capuchins, still 29.6% of the lexical elements contained in the raters' item interpretations referred to factors other than those on which the given items showed their highest loading (in 578 cases, lexical elements were coded into the working construct that had its highest loading on the same factor as the item under interpretation).

For example, the above-mentioned interpretations of the item statement "When there is food, [Name] is quickly on the spot" covered lexical elements that could be encoded into the constructs Arousability, Vigilance, Physical activity, Dominance, Food orientation or (inverse) Anxiousness. There were even cases in which different raters interpreted the same item with regard to the same working construct but in the opposite direction of meaning. For example, "When somebody stays in front of the cage, [Name] jumps at the grate and may also try to grab that person" was interpreted by some raters as describing capuchins who are "not confident with humans" and "maybe feel threatened by the person" but by other raters as describing capuchins who are "bolder" and "not fearful towards humans". Table 5 shows for each given item the frequencies for the lexical elements in the raters' interpretations that were assigned to each of the 21 BR_xBS-Approach-generated working constructs (below).

The interpretations of trait-adjective items (I-TA) contained significantly more unique lexical elements (M = 17.29, SD = 6.43) than those of the behaviour-descriptive verb items (I-BV; M = 14.35, SD = 2.65), $F(1,53) = 5.593$, $p = .022$. But the number of different constructs into which the lexical elements could be encoded did not differ between interpretations of trait-adjective items (M = 4.00, SD = 2.07) and behaviour-descriptive verb items (M = 4.29, SD = 1.42), $F(1,53) = 0.389$, $p = .536$. That is, the fields of meanings reflected in the interpretations of trait-adjective items were not broader or more diverse than those of the behaviour-descriptive verb items; they were only more differentiated.

It is interesting that the fields of meanings that raters constructed for trait-adjective items corresponded significantly more to the meanings of the constructs that these items were meant to operationalise (M = 72.0%, SD = 23.0) than was the case for the behaviour-descriptive verb items (M = 53.9%, SD = 24.3), $F(1,53) = 7.535$, $p = .008$. The fields of meanings reflected in the interpretations of trait-adjective items also corresponded more to the fields of meanings reflected by the assessment factors (M = 81.1%, SD = 17.1) than was the case for the behaviour-descriptive verb items (M = 66.7%, SD = 23.9), $F(1,53) = 5.559$, $p = .022$. Recall that interpretations of the trait-adjective items primarily comprised behavioural and situational descriptions, whereas interpretations of the behaviour-descriptive verb items primarily comprised trait-adjectives. That is, raters' adjectival interpretations of observable behaviours, as described in the behaviour-descriptive verb items, reflected more heterogeneous fields of meanings than their behavioural and contextual interpretations of trait-adjective items. This result may indicate an effect of the semiotic meanings contained in trait-adjectives (see Section 4.5).

4. Discussion

We applied the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) to highlight essential methodological differences between observations and assessments for research on "personality" and to specify the central points of the increasing criticism of assessment methods. To enable illuminating contrasts, we used a five-method multi-species study comprising human raters and capuchin monkeys to analyse the ways in which two assessment-based categorisations of individual-specific behaviours deviated from those obtained with observations. Possible sources of these differences were identified in the different kinds of phenomena that can be captured with observations versus assessments, in the processes of human impression formation and in the ways in which data are generated with these methods, highlighting various biases and serious methodological limitations in the use of questionnaires, none of which have been previously well considered. Our findings offer a novel approach to explaining the frequent

lack of replicability of findings in psychology and the social sciences, a topic that is currently discussed intensely.

4.1 Capuchins' individual-specific behaviours versus raters' pertinent mental representations—two different kinds of phenomena

The TPS-Paradigm emphasises the idea that individuals' behaviours are different from the ideas, beliefs and mental representations that humans develop of them. These different kinds of phenomena require different methods of exploration; therefore, such methods are not interchangeable (Uher, 2015a, b, c). The behaviours' momentariness requires real-time recordings, thus observations. But assessments are inherently retrospective and memory-based and therefore cannot be used to explore behaviours. This explains why assessments using two different formats yielded similar results, whereas their relations to the observational measures were much weaker. This also explains why assessments overestimated the temporal reliability of observable individual behaviours (d = 0.61 to 0.77), a finding that is in accordance with previous ones obtained with the same research design for assessments of great apes (d = 0.73 to 0.91; Uher & Asendorpf, 2008) and crab-eating macaques (d = 0.92 to 1.33; Uher et al., 2013b).

Given these findings, taxonomic models derived from assessments summarise structural patterns that may underlie raters' mental representations, but such models cannot reflect structural patterns in the behaviours of the individuals who were assessed. This point is frequently overlooked in taxonomic "personality" research. Previous studies have shown that the latent structures of assessment data are more coherent and much less complex than the latent structures of behavioural data (Allport & Vernon, 1933; Blurton Jones, 1967, 1972; Smith, 1973; Smith & Connolly, 1972, 1980; Uher et al., 2013b). This is because the mental images that people develop of observable behaviours are simplified representations that are consistent with the logic of the human mind (and its many fallacies) and with the implicit structures contained in language—but not necessarily with the structures that can be identified in behaviours.

The frequent interpretation of the more coherent patterns in assessment data as indicating their superior reliability and utility is not warranted. Instead, these patterns result from restricting the empirical values that can be generated with assessments to just a few unspecified categories and from selecting only those items that allow data that best match statistical criteria (e.g., reliability, simple factor structure) to be generated. By doing so, psychometricians align the development of inventories and models to statistical theories rather than to the actual phenomena under study (e.g., mental representations, individual behaviours). In other fields, by contrast, scientists (e.g., physicists) do not discard their measurement tools (e.g., thermometers) just because the data that they produce (e.g., about temperature) do not fit particular statistical models (Chang, 2004; Uher, 2015a,b,c,e, under review). This practice of discarding, which is widespread in psychology and the social sciences, further contributes to the deviation of assessment results from those obtained with observations.

4.2 Formation of "personality" impressions

We explored how raters may have developed abstract mental representations of capuchin individuals in terms of trait-adjectival assessments and analysed how these are related to observable behaviours. Our results on mediation effects suggest (at least for some constructs) that more specific mental representations of particular individual-specific behaviours may have served as intermediate steps in the formation of more abstract representations as encoded with trait-adjectives. For example, observations of behaviours involving social contact with humans (e.g., Scalp lift, Lip-smack, Approach) may have been *abstracted in a bottom-up* fashion first into behaviour-specific representations that, in turn, may have facilitated the development of more abstract representations of monkeys as differing in the degree to which they are "friendly to humans". These findings mirror previous

ones on keepers' assessments of great apes (Uher & Asendorpf, 2008) and on expert and novice observers' assessments of crab-eating macaques (Uher et al., 2013b). These findings are also in accordance with models of impression formation about human individuals according to which, at low levels of experience with a target person, people mentally represent impressions as behavioural exemplars and, with increasing experience, extract abstract impressions that can then be retrieved independently, such as for assessments (Park, 1986; Sherman & Klein, 1994). These models provide additional explanations for the divergence between results from assessments versus observations.

4.3 Assessments contain stereotypical biases

Our analyses of associations with socio-demographic factors revealed that capuchin monkeys showed hardly any age, sex or rearing-related differences in their individual-specific behaviours. This is in contrast with the numerous group differences reflected in the assessments; sex differences, for example, occurred in 12 out of 20 "personality" constructs. Males were judged to be more excitable, less anxious, more competitive, more curious, more food oriented, more impulsive, more persistent and more vigilant than females, amongst other differences. But none of these differences occurred in the behavioural measures. We found behavioural sex differences only in intra-specific Aggressiveness and Dominance; these differences were also reflected in the assessments but underestimated in terms of their magnitude. Behavioural differences between age groups occurred only in Impulsiveness but were not reflected in the assessments. In the behavioural measures, we found three differences related to rearing history, but only two of them occurred in the assessments and in much less pronounced ways. Conversely, the assessments showed four further associations with early life experiences that occurred in the behavioural measures in much less pronounced ways, in the opposite direction or not at all. Such complex patterns of divergence between observations and assessments have also been demonstrated in a study involving 99 human raters and 104 crab-eating macaques with the same research design (Uher et al., 2013b).

The many group differences found in the assessment data likely reflect stereotypical ideas about human individuals that are widespread in the raters' sociocultural communities. These attribution biases became particularly apparent because, in the behaviours of our nonhuman study species, group differences were largely absent. It is important to note that these methodical differences emerged despite the fact that stereotypical biases also influence behavioural observations of both human (Pellegrini, 2011) and nonhuman individuals (Uher, 2011b). This tendency argues for a profound impact of the ways in which quantitative data are generated in observational versus assessment methods as explored in this research (i.e., the real-time recording of occurrences of specified events versus the retrospective and memory-based construction of overall judgements that are based on unspecified events and algorithms).

Given this, our findings provide an additional explanation for the differences between assessment-based and behavioural categorisations of individuals and highlight serious limitations of standardised assessments for analyses of group differences. The research frameworks of the TPS-Paradigm and the study design of the present study allow for exploring these important methodological issues more systematically also in assessments of human individuals.

4.4 Assessment methods do not allow for the generation of scientific quantifications

The quantifications obtained with assessment methods are based on the mental processes through which raters generate their judgements. But despite their importance for many fields of research, these processes have hardly been explored so far (Diriwächter et al., 2005; Rosenbaum & Valsiner, 2011; Uher, 2013, 2015e; Wagoner & Valsiner, 2005).

Inventories and rating scales rely on abstract and decontextualised descriptors from everyday language that are intuitively understandable by laypeople and are applicable to

diverse assessments. But this ease of use has its downside because it prevents researchers from specifying the particular elements that raters may have considered and how raters may have quantified what they consider to be specific to an individual in comparison with others. In fact, for behaviours that generally occur more frequently in a sample, the category "often" should refer to higher numbers of observed occurrences than for behaviours that generally occur much less frequently. That is, raters have to construct for the *same* answer categories *different* meanings. But researchers commonly recode the answer categories into numerals for all items always in exactly the same ways. Observations, by contrast, rely on fixed biunique relations between observed occurrences of specified events and their numerical encoding in the data (e.g., two occurrences are always encoded as "2").

In sum, assessment methods do not allow researchers to fulfil the two requirements of scientific quantification and can produce only subjective quantifications. The fact that such quantifications at best reflect ordinal-scaled data is well-known but seldom considered in "personality" research; instead, the data thus generated are commonly treated as metric data, such as when factor analysis is applied (Michell, 1999). In this research, we followed these established psychometric practices of quantitative psychology to demonstrate that by their very application, essential divergences from behavioural data occur. We aimed to highlight important limitations of assessment methods that are not well considered and that provide an explanation for the frequent lack of replicability of assessment-based findings.

4.5 Standardised assessment items do not represent standardised meanings but reflect entire fields of meanings that vary within and between persons

Standardised assessments are based on the idea that all raters interpret the item statements in the exact same way as the researchers. But this assumption may not hold true as previous studies on assessments of humans have shown (Diriwächter et al., 2005; Rosenbaum & Valsiner, 2011; Wagoner & Valsiner, 2005). The current research is the first to scrutinise this assumption with regard to assessments of nonhuman individuals. Our analyses revealed tremendous variations in the ways in which six independent persons understood the same item statements—both within and between raters. The meanings that they constructed for the item statements were much broader and more heterogeneous than the meanings that we, as the researchers, had aimed to operationalise with them.

Raters' behavioural interpretations of the trait-adjective items conformed better to our own interpretations than did raters' adjectival interpretations of the behaviour-descriptive verb items. Along with our findings on possible pathways in the formation of abstract representations of individual behaviours, this may indicate an effect of the implicit semiotic meanings that trait-adjectives contain. Interpretations of specific behaviours may be more diverse because the meanings of behaviours depend on the contexts in which they occur (Uher, 2015a, e). Trait-adjectives, by contrast, reflect decontextualised mental representations that were already abstracted from specific behaviours and situations. Thus, the meanings of adjectives rely on the implicit meanings that are attributed to them, which may make their interpretation more coherent between different persons.

The interpretations of standardised assessment items of only six persons varied tremendously. The possible variability that may occur in larger samples of raters may be even more pronounced, especially if people from different sociocultural communities are involved. Had we asked all of the raters from the nine participating institutions originating from four different nations with sociocultural and linguistic communities as diverse as North-American, Asian, West- and South-European, very likely we would have revealed far more diverse fields of meanings. These issues profoundly effect the interpretation of results obtained with assessments but have hardly been considered so far in research on nonhuman individuals and in research on human individuals.

4.6 Conclusions

This research highlighted fundamental methodological differences between assessments and observations in research on "personality". Using a five-method study and a multi-species sample, we showed *(1)* that capuchin monkeys exhibit pronounced individual-specific behaviours and that these are reflected in assessments provided by human raters using two standardised inventories. But we also demonstrated *(2)* that assessment-based categorisations of individuals contained several kinds of biases derived from raters' mental abstractions and stereotypical beliefs about individuals. Our study also revealed (3) that standardised items do not reflect narrow standardised meanings as commonly assumed but broad fields of meanings, and this profoundly effects the interpretation of results.

Our findings argue for much more critical applications of assessment methods in "personality" research and highlight that assessments are not equivalent to observations. To reduce the pronounced biases in memory-based methods, "personality" researchers should explore the relations of assessments of individual-specific behaviours to pertinent observations much more systematically and comprehensively than is commonly the case in both human and animal research. In particular, researchers should intensify applications of modern technologies enabling sophisticated techniques for recording individual behaviour in everyday life settings, such as ambulatory monitoring, digital ethnography and reality mining. With these efforts, psychologists and social scientists can draw on the long-standing pertinent expertise of animal researchers. Vice versa, animal researchers can learn from the extensive experience that psychologists and social scientists have gained in assessment-based research over the last century. In particular, animal researchers should resist the temptations of easy-to-use assessment tools to create large data sets quickly and at low cost and should value and capitalise on their behaviour-scientific core competencies.

Assessment methods cannot be used to explore behaviours and to generate scientific quantifications of individual-specificity ("personality"). Assessments cannot replace observations.

Acknowledgements

This research was funded by a grant to Jana Uher from the Deutsche Forschungsgemeinschaft, DFG (German Research Foundation, Grant Nr. UH249/1-1), and the PNR-CNR Aging Program 2012-2014. We thank the editor, Elsa Addessi and three reviewers for their helpful comments on an earlier draft. We are very grateful to all our raters in the collaborating research institutions. Special thanks go to Nicolas Claidière and Gary Lynn. We thank the staff members and interns of the Comparative Differential and Personality Research Group: Federica Dal Pesco, Matilde Sauvaget and Christina Meier for behavioural coding from video and help with the data organisation, and Lisa Rodova for her help with the literature. We also thank researchers and staff of the ISTC-CNR: Antonella Giordano for carrying out most of the behavioural tests and observations, Valentina Truppa, Francesco Natale, Massimiliano Bianchi and Simone Catarinacci for their support and the Fondazione Bioparco for hosting the Primate Centre.

References

Ah-King, M. (2013). *Challenging popular myths of sex, gender and biology*. Springer.
Allport, G. W. (1937). *Personality: A psychological interpretation.* New York, NY: Macmillan.
Allport, G. W., & Vernon, P. E. (1933). *Studies in expressive movement*. New York: Macmillan.
ASAB/ABS (2012). Guidelines for the treatment of animals in behavioural research and teaching. *Animal Behaviour, 83,* 301-309
Arro, G. (2013). Peeking into personality test answers: Inter- and intraindividual variety in item interpretations. *Integrative Psychological and Behavioral Science, 47,* 56-76.
Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Wicherts, J. M., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G. & Weber, H. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality, 27,* 108-119.
Bauer, M. & Gaskell, G. (Eds). (2000). *Qualitative researching with text, image and sound*. London: Sage.
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.
Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science, 2,* 396-403.
Belyaev, D. K. (1969). Domestication of animals. *Science Journal (U.K.), 5*, 47–52.
Blurton Jones, N. G. (1967). An ethological study of same aspects of social behaviour of children in nursery school. In D. Morris (Ed.), *Primate ethology* (pp. 347–368). London: Weidenfeld & Nicolson.
Blurton Jones, N. G. (1972). Categories of child-child interaction. In N. G. Blurton Jones (Ed.), *Ethological studies of child behavior* (pp. 97-127). London: Cambridge University Press.
Byrne, G. & Suomi, S. J. (1998). Relationship of early infant state measures to behavior over the first year of life in the tufted capuchin monkey (*Cebus apella*). *American Journal of Primatology, 44,* 43-56.
Carpenter, S. (2012). Psychology's bold initiative. *Science, 335,* 1558-1561
Chang, H. (2004). *Inventing temperature.* New York: Oxford University Press.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.
Costello, A. B., & Osborne, J. W. (2005). Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10,* 1-9.
Diriwächter, R., Valsiner, J., & Sauck, C. (2005). Microgenesis in making sense of oneself: Constructive recycling of personality inventory items. *Forum Qualitative Sozialforschung/ Forum: Qualitative Social Research, 6*, Art. 11.
Dong, W., Lepri, A., & Pentland, S. (2011). Modeling the so-evolution of behaviors and social relationships using mobile phone data, *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, 134-143.
Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192-203.
Fahrenberg, J., Myrtek, M., Pawlik, K. & Perrez, M. (2007). Ambulatory assessment – monitoring behavior in daily life settings. A behavioral-scientific challenge for psychology. *European Journal of Personality Assessment, 23,* 206-213.
Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.
Fragaszy, D. M., Visalberghi, E., & Fedigan, L. (2004). *The Complete Capuchin*. Cambridge University Press.
Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality, 23,* 369-401.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in Personality, 37,* 504-528.

Gurrin, C., Smeaton, A. F., & Doherty, A. R. (2014). Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval, 8,* 1-125.

Hammersley, M. (2013). *The myth of research-based policy and practice*. London: Sage.

Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (*2002*). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology, 82,* 903-918.

JCGM, Joint Committee for Guides in Metrology. (2008). *International vocabulary of metrology – Basic and general concepts and associated terms (VIM) (3rd ed.),* Working Group 2 (Eds.), Joint Committee for Guides in Metrology.

Lahlou, S. (2011). How can we capture the subject's perspective?: An evidence-based approach for the social scientist. *Social Science Information, 50,* 607-655.

Lahlou, S., Le Bellu, S. & Boesen-Mariani, S. (2015). Subjective evidence based ethnography: method and applications. *Integrative Psychological and Behavioral Science, 49,* 216-238.

Lloyd, B., & Duveen, G. (1992). *Gender identities and education: The impact of starting schoo*l. Hemel Hempstead: Harvester Wheatsheaf.

Lynch Alfaro, J.W., de Sousa e Silva Jr, J., & Rylands, A.B. (2012) How different are robust and gracile capuchin monkeys? An argument for the use of Sapajus and Cebus. *American Journal of Primatology, 74,* 273-286.

Mangold, P. (2010). *Interact User Guide*, V. 9.0. ff. Arnstorf: Mangold International.

Manson, J. H. *&* Perry, S. (2013). Personality structure, sex differences, *and* temporal change *and* stability in wild white-faced capuchins, *Cebus capucinus. Journal of Comparative Psychology, 127,* 299-311.

Mehl, M. R., & Conner, T. S. (Eds.). (2012*). Handbook of research methods for studying daily life*. New York: Guilford Press.

Michell, J. (1999). Measurement in psychology. Cambridge, UK: Cambridge University Press.

Morton, F. B., Lee, P. C., Buchanan-Smith, H. M., Brosnan, S. F., Thierry, B., Paukner, A., Widness, J. & Weiss, A. (2013). Personality structure in brown capuchin monkeys (*Sapajus apella*): Comparisons with chimpanzees (*Pan troglodytes*), orangutans (*Pongo spp.*), and rhesus macaques (*Macaca mulatta*). *Journal of Comparative Psychology, 127,* 282-298.

Park, B. (1986). A method for studying the development of impressions of real people. *Journal of Personality and Social Psychology, 51,* 907-917.

Pellegrini, A. D. (2011). ".. in the eye of the beholder": Sex bias in observations and ratings of students' aggression. *Educational Researcher, 40*, 281 - 286.

Rosenbaum, P. J., & Valsiner, J. (2011). The un-making of a method: From rating scales to the study of psychological processes. *Theory & Psychology, 21,* 47-65.

Schacter, D. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist 54,* 182-203.

Sherman, J. W. & Klein, S. B. (1994). The development and representation of personality impressions. *Journal of Personality and Social Psychology, 67,* 972-983.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Smith, P. K. & Connolly, K. J. (1972). Patterns of play and social interaction in preschool children. In N. G. Blurton Jones (Ed.). *Ethological studies of child behavior* (pp. 65-95). London, U.K.: Cambridge University Press.

Smith, P. K. & Connolly, K. J. (1980). *The ecology of preschool behaviour.* London, U.K.: Cambridge University Press.

Smith, P. K. (1973). Temporal clusters and individual differences in the behaviour of preschool children. In R. P. Michael & J. H. Crook (Eds.). *Comparative ecology and behaviour of primates,* (pp. 751-798). London, U.K.: Academic Press.

Trut, L. N. (1999). Early canid domestication: The farm-fox experiment. *American Scientist, 87,* 160–169.

Uher, J. (2008a). Comparative personality research: Methodological approaches. *European Journal of Personality, 22,* 427-455.

Uher, J. (2008b). Three methodological core issues of comparative personality research. *European Journal of Personality, 22,* 475-496.

Uher, J. (2011a). Individual behavioral phenotypes: An integrative meta-theoretical framework. Why 'behavioral syndromes' are not analogues of 'personality'. *Developmental Psychobiology, 53*, 521–548.

Uher, J. (2011b). Personality in nonhuman primates: What can we learn from human personality psychology? In A. Weiss, J. King, & L. Murray (Eds.), *Personality and Temperament in Nonhuman Primates* (pp. 41-76). New York, NY: Springer.

Uher, J. (2013). Personality psychology: Lexical approaches, assessment methods, and trait concepts reveal only half of the story—Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science, 47,* 1-55.

Uher, J. (2015a). Conceiving "personality": Psychologists' challenges and basic fundamentals of the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals. *Integrative Psychological and Behavioral Science, 49*, 398-458.

Uher, J. (2015b). Developing "personality" taxonomies: Metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integrative Psychological and Behavioral Science, 49*, 531-589.

Uher, J. (2015c). Interpreting "personality" taxonomies: Why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead. *Integrative Psychological and Behavioral Science, 49*, 600-655.

Uher, J. (2015d). Agency enabled by the Psyche: Explorations using the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals. *Annals of Theoretical Psychology, 12,* 177-228.

Uher, J. (2015e). Comparing individuals within and across situations, groups and species: Metatheoretical and methodological foundations demonstrated in primate behaviour. In D. Emmans & A. Laihinen (Eds.). *Comparative Neuropsychology and Brain Imaging (Vol. 2), Series Neuropsychology: An Interdisciplinary Approach.* (pp. 223-284). Berlin: Lit Verlag.

Uher, J. (2016). Exploring the workings of the psyche: Metatheoretical and methodological foundations. *Annals of Theoretical Psychology, 13*, 299-324.

Uher. J. (in press). What is behaviour? And (when) is language behaviour? *Journal for the Theory of Social Behaviour.* DOI: 10.1111/jtsb.12104

Uher, J. (under review). Philosophy-of-science principles of measurement and quantification in the physical sciences, life sciences, social sciences and in psychology.

Uher, J., & Asendorpf, J. B. (2008). Personality assessment in the Great Apes: Comparing ecologically valid behavior measures, behavior ratings, and adjective ratings. *Journal of Research in Personality, 42,* 821-838.

Uher, J., Addessi, E., & Visalberghi, E. (2013a). Contextualised behavioural measurements of personality differences obtained in behavioural tests and social observations in adult capuchin monkeys (*Cebus apella*). *Journal of Research in Personality, 47*, 427-444.

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.

Uher, J., Werner, C. S., & Gosselt, K. (2013b). From observations of individual behaviour to social representations of personality: Developmental pathways, attribution biases, and limitations of questionnaire methods. *Journal of Research in Personality, 47,* 647–667.

Visalberghi, E. & Fragaszy, D. (2012). What is challenging about tool use? The Capuchin's Perspective. In T. R. Zentall & E. A. Wasserman (Eds). *The Oxford Handbook of Comparative Cognition*, Chapter 40. Oxford University Press.

Wagoner, B., & Valsiner, J. (2005). Rating tasks in psychology: From static ontology to dialogical synthesis of meaning. In A. Gülerce, I. Hofmeister, G. Saunders, & J. Kaye (Eds.), *Contemporary theorizing in psychology: Global perspectives* (pp. 197–213). Toronto, Canada: Captus.

Weber, R. P. (1990). *Basic content analysis* (2nd ed). Newbury Park, CA: Sage Publications.

Yong, E. (2012). Replication studies: Bad copy. *Nature, 485,* 298–300.

***

Table 3 Exploratory factor analysis of raters' assessments on behaviour-descriptive verb items (BV) and trait-adjective items (TA) using the Capuchin Personality Inventory (CPI): BR×BS-Approach-generated working constructs operationalised by each item, item content, factor loadings and item communalities.

Working construct	**Item format** [a]	**Item content (abbreviated** [b]**)**	**Assessment factors**					**Item commu-nality h^2**
			Dominant-competitive-aggressive	Curious-inventive-persistent	Playful-active-impulsive	Gregarious-prosocial	Excitable-vigilant	
Dominance	TA	Dominant	**.90**	.05	-.16	.19	.22	.91
	BV	Can occupy the best places	**.83**	.11	-.04	**.40**	-.02	.91
	BV	Makes way for others	**-.75**	-.16	-.07	**-.42**	.16	.86
Competitiveness	BV	Starts conflicts	**.86**	.17	-.01	.10	.29	.90
	TA	Competitive	**.83**	.21	.13	.14	.23	.87
	BV	Displaces others who are with partners	**.79**	.03	-.14	.26	.19	.82
Aggressiveness to conspecifics	BV	Starts agonistic interactions	**.81**	.07	-.01	.11	.20	.90
	TA	Aggressive to conspecifics	**.81**	.02	.00	-.02	.24	.87
Anxiousness	BV	Screams quickly and flees	**-.76**	-.17	.03	**-.40**	.09	.83
Food orientation	BV	Is quickly on the spot when there is food	**.72**	**.40**	.14	**.45**	-.05	.52
	TA	Gluttonous	**.59**	.28	.14	.09	.18	.73
	BV	Spends much time searching for food	.34	.20	-.12	.08	.30	.82
Sexual activity	TA	Sexually active	**.56**	.02	**-.43**	.16	**.43**	.85
	BV	Tries to contact others sexually	**.55**	.00	**-.40**	.25	**.40**	.83
Impulsiveness	BV	Quickly bangs against the mesh when he/she does not get his/her food immediately	**.48**	.27	.26	-.16	**.41**	.70
Curiousness	BV	Readily explores changes in the environment	.21	**.77**	**.42**	.23	-.07	.75
	TA	Curious	-.04	**.75**	**.53**	.28	-.31	.83
	BV	Explores new, potentially edible materials	.15	**.73**	**.46**	.20	-.14	.72
Creativness/ Inventiveness	BV	Involves several objects in his/her activities	-.07	**.77**	**.58**	.14	-.16	.84
	TA	Inventive	-.05	**.69**	.37	.04	-.08	.77

Working construct	**Item format** [a]	**Item content (abbreviated** [b]**)**	**Assessment factors**					**Item commu-nality h^2**
			Dominant-competitive-aggressive	Curious-inventive-persistent	Playful-active-impulsive	Gregarious-prosocial	Excitable-vigilant	
Distractibility	BV	Can focus long on activities	.09	**.69**	-.03	.12	-.09	.71
Persistency	TA	Persistent	.31	**.64**	.28	.09	.12	.74
	BV	Can spend much time without interrupting activity	.24	**.59**	-.09	.02	.05	.69
Vigilance	BV	Quickly spots small food items or changes	.36	**.69**	.23	.18	.24	.75
Anxiousness	BV	Keeps a distance to unknown objects and/or persons	-.36	**-.56**	-.31	-.35	.33	.65
Social orientation to humans	BV	Approaches, lip smacks and/or scalp lifts to persons	.03	**.51**	.23	.17	.03	.66
	TA	Friendly to humans	-.21	**.49**	.07	**.40**	**-.47**	.79
Playfulness	TA	Playful	-.09	**.52**	**.73**	.33	**-.44**	.86
	BV	Plays on his/her own	-.17	**.58**	**.67**	.11	-.25	.78
	BV	Engages in rough-and-tumble play or play chases	-.03	**.42**	**.65**	.31	-.35	.77
Physical activity	TA	Physically active	.10	**.45**	**.69**	.19	-.06	.77
	BV	Takes rests during daytime	.21	-.25	**-.55**	.04	.08	.62
	BV	Constantly moves about	-.06	.23	**.50**	.02	.09	.59
Impulsiveness	TA	Impulsive	.32	.31	**.70**	.00	.21	.77
Arousability	TA	Excitable	.23	.22	**.68**	-.11	.14	.73
Distractibility	TA	Distractible	-.08	-.04	**.58**	-.07	-.02	.62
Aggressivenss to humans	BV	Jumps at the grate when persons are in front	.35	**.51**	**.57**	.16	.04	.71
Cleanliness	TA	Cleanly with him/herself	.04	-.02	-.38	.10	.25	.56
	BV	Cleans him/her-self intensely	.01	-.14	-.29	-.06	.20	.46
Gregariousness	BV	Sits close together with others	**.40**	.10	-.09	**.84**	-.17	.88
	TA	Gregarious	.09	.25	.23	**.67**	-.33	.70
	BV	Spends much time on his/her own	-.37	-.22	-.25	**-.66**	.19	.74
Social orientation to conspecifics	BV	Co-feeds with others	.37	.17	.00	**.73**	-.21	.73
	BV	Touches and grooms others	.35	.09	-.39	**.62**	.17	.83
	TA	Friendly to others	-.32	.23	.03	**.61**	**-.47**	.78
	BV	Approaches and lip smacks to others	-.01	.09	.03	**.49**	-.18	.65

Working construct	**Item format** [a]	**Item content (abbreviated** [b]**)**	**Assessment factors**					**Item commu-nality h^2**
			Dominant-competitive-aggressive	Curious-inventive-persistent	Playful-active-impulsive	Gregarious-prosocial	Excitable-vigilant	
Arousability	BV	Prior feeding, he/she paces restlessly	.13	-.11	-.03	-.24	**.62**	.65
	BV	Paces or scratches when there are unusual noises	.02	-.22	.00	-.22	**.51**	.66
Vigilance	TA	Vigilant	.17	.07	-.04	-.23	**.59**	.66
	BV	Watches everything around him/her closely	.06	.15	-.05	-.08	**.55**	.66
Anxiousness	TA	Anxious	-.26	-.20	.06	-.38	**.58**	.74
Aggressiveness to humans	TA	Aggressive to humans	**.44**	-.04	.32	-.30	**.47**	.76
		Eigenvalues	1.60	8.05	4.70	3.05	1.92	
		Percentage of total variance	2.39	15.48	9.04	5.86	3.70	
		Number of items	15	12	12	7	6	

Note. Based on assessments for $N = 150$ capuchin individuals, principal axis factoring, and promax rotation with Kaiser normalisation. [a] Item format: TA Trait-adjective item, BV Behaviour-descriptive verb item. Factor loadings ≥ .40 in absolute value are bold. Grey cells indicate the primary loadings on the factors constructed. [b] Item content abbreviated; the complete item statements are provided in Tables S3 and S4 in the Supplemental Material.

Table 4 Associations of socio-demographic factors with the capuchins' individual-specific behaviours studied with tests and observations (B-TO) and their reflection in the observers' mental representations studied with behaviour-descriptive verb assessments (A-BV) and trait-adjective assessments (A-TA) on the level of BR_xBS-Approach-generated working constructs

Working construct	Male vs. female[a]						Older vs. younger[b]						Hand-reared vs. mother-reared[c]					
	B-TO		A-BV		A-TA		B-TO		A-BV		A-TA		B-TO		A-BV		A-TA	
	d	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*	*d*	*1-β*
Aggressiveness to conspecifics	**1.69**	(98)	**1.06**	(82)	**1.29**	(93)	0.28	(15)	-0.21	(13)	-0.23	(14)	-0.67	(40)	**-0.95**	(76)	**-0.88**	(70)
Aggressiveness to humans	0.77	(51)	**0.91**	(71)	**1.20**	(89)	-0.62	(38)	-0.47	(32)	0.13	(9)	**-1.79**	(98)	-0.73	(55)	-0.54	(37)
Arousability	-0.39	(24)	**0.93**	(72)	0.82	(63)	-0.32	(20)	-0.40	(26)	0.70	(54)	-0.47	(31)	-0.50	(33)	0.47	(31)
Anxiousness	0.48	(31)	**-1.29**	(93)	**-1.33**	(94)	-0.01	(5)	-0.12	(9)	-0.05	(6)	0.09	(8)	0.47	(31)	0.16	(11)
Competitiveness	-0.02	(5)	0.82	(63)	**1.63**	(99)	0.48	(21)	-0.19	(12)	-0.13	(9)	0.84	(38)	**-1.30**	(94)	**-0.89**	(71)
Creativeness/inventiveness	0.81	(43)	**1.03**	(80)	0.69	(51)	-0.48	(22)	-0.12	(9)	0.01	(5)	0.48	(21)	0.26	(16)	0.13	(9)
Curiousness	0.35	(16)	**1.19**	(89)	0.89	(69)	0.18	(10)	0.23	(14)	-0.02	(5)	1.05	(56)	0.16	(11)	0.12	(9)
Distractibility	0.17	(9)	0.25	(16)	0.32	(20)	0.56	(31)	-0.09	(8)	-0.47	(32)	**1.37**	(76)	-0.29	(18)	-0.29	(18)
Dominance[d]	**1.68**	(97)	0.42	(28)	**0.90**	(72)	0.10	(8)	-0.35	(22)	0.10	(8)	-0.68	(41)	-0.23	(14)	-0.89	(72)
Food orientation	0.39	(24)	**1.20**	(89)	**1.26**	(92)	0.34	(21)	0.19	(12)	-0.09	(8)	0.23	(14)	-0.38	(23)	-0.51	(35)
Gregariousness	0.02	(5)	-0.08	(7)	0.10	(8)	-0.48	(27)	-0.48	(34)	-0.89	(72)	**-1.17**	(79)	-0.85	(68)	-0.26	(16)
Impulsiveness	-0.55	(26)	1.51	(38)	**0.92**	(72)	**-2.00**	(98)	-0.09	(8)	0.02	(6)	-0.73	(35)	-0.19	(12)	-0.29	(17)
Physical activity	-0.23	(12)	-0.66	(50)	0.69	(53)	-0.04	(6)	-0.08	(8)	-0.27	(17)	-0.10	(7)	-0.08	(7)	-0.31	(20)
Persistency	-0.02	(5)	**0.90**	(70)	0.76	(58)	-0.75	(39)	0.34	(21)	0.45	(30)	0.55	(25)	0.01	(5)	0.08	(7)
Playfulness	0.62	(38)	**0.95**	(74)	0.45	(29)	-0.27	(14)	0.19	(12)	-0.43	(28)	-0.50	(27)	0.00	(5)	0.21	(13)
Self-cleanliness	0.59	(35)	0.33	(20)	-0.09	(8)	0.11	(8)	0.09	(8)	-0.32	(19)	0.11	(8)	**0.91**	(72)	0.44	(29)
Social orientation to conspecifics	0.24	(14)	-0.28	(17)	-0.32	(19)	-0.65	(48)	-0.48	(32)	-0.37	(24)	-0.82	(64)	-0.74	(57)	-0.22	(13)
Social orientation to humans	-0.17	(10)	0.31	(18)	-0.04	(6)	0.22	(12)	-0.46	(31)	-0.18	(12)	0.95	(63)	0.21	(13)	0.68	(51)
Sexual activity	-0.34	(18)	-0.08	(7)	0.32	(19)	-0.37	(20)	-0.74	(57)	-0.35	(22)	-0.22	(12)	**-1.01**	(79)	-0.79	(61)
Vigilance	-0.52	(25)	**1.41**	(96)	0.77	(58)	0.18	(10)	-0.07	(7)	0.26	(16)	-0.02	(5)	-0.11	(9)	-0.20	(12)

Note. B-TO – Behavioural composite construct measures obtained in tests and observations; A-BV – Behaviour-descriptive verb assessments aggregated; A-TA – Trait-adjective assessments. Cohen's effect size *d* on pooled standard deviations; estimated power to detect the given difference in parentheses. Bold effect sizes detected with a power > 70%. [a] A positive *d* indicates higher scores for males compared to females; [b] a positive *d* indicates higher scores for individuals older than 15 years (median) as compared to younger individuals; [c] a positive *d* indicates higher scores for hand-reared individuals compared to mother-reared ones. [d] We studied Dominance as "personality" construct, that is, as individual-specific patterns in dominant-submissive behaviours in which all individuals can be quantified and compared with one another, rather than as social status, which refers to only a few individuals per group (e.g., alpha male status).

Table 5 Raters' interpretations of the behaviour-descriptive verb items (CPI-BV) and trait-adjective items (CPI-TA) and the occurrences of lexical elements encoded in the 21 BR_xBS-Approach-generated working constructs based on content analysis

Item code[a]	Item-wise frequencies of lexical elements encoded in the 21 working constructs organised by their associations in the factors																				
	Dominant-competitive-aggressive					Curious-inventive-persistent					Playful-active-impulsive				Gregarious-prosocial		Excitable-vigilant				
	DO	CO	AG	FO	SX	CU	CR	DI	PE	SH	PL	PA	IM	SC	GR	SO	AR	VI	AX	AH	YO
DOCPAD	**18**		**3**	**7**	**1**					1		1			1						
DOCPB1	**8**		**1**																		
DOCPB2	**-6**																	1	6		
COCPB1	**5**	**3**	**8**													-1			-2		
COCPAD		**12**	**4**	**3**	**1**														1; -1		
COCPB2	**4**	**6**	**3**																		
AGCPB1	**6**		**11**													1		1	1		
AGCPAD			**12**																		
AXCPB1	**-4**															2	2		8		
FOCPB2	**4**			**3**								2					1	3	-1		
FOCPAD	**1**	**3**		**14**						1							1				
FOCPB1				**4**		1			2			2			-2	-1					
SXCPAD					**10**																
SXCPB1	**1**				**4**											2	1				
IMCPB1				**1**			1						6				4				
CUCPB2	1					**7**	**1**					3						2	-2		
CUCPAD				2		**14**						1						1	-3		
CUCPB1				1		**12**	**2**									1			-3		
CRCPB1						**5**	**9**					3									
CRCPAD						**2**	**7**														
DICPB1								**-1**	**8**			1	-2								
PECPAD								**-6**	**11**												
PECPB1				2				**-1**	**11**									-1			
VICPB2						**3**						2; -1					1	10			
AXCPB2	-1																1		11		
SHCPB1				1						**12**											
SHCPAD										**16**	2								-1		

Item code	DO	CO	AG	FO	SX	CU	CR	DI	PE	SH	PL	PA	IM	SC	GR	SO	AR	VI	AX	AH	YO
PLCPAD		-1								1	**13**	**2**				2					1
PLCPB1						-5	2				**2**	**2**			-1	-1					
PLCPB2											**5**	**4**			1	5					
PACPAD				1		2	1				**2**	**11**									
PACPB2												**4**					-6				
PACPB1	-1					1						**2**			-1		7	1	1		
IMCPAD	1	1	1						-1				**11**				2				
ARCPAD				1	1					1	**1**	**1**					1	1			
DICPAD								7										2	1		
AHCPB1	1									3	**1**						2		2; -2	5	
SCCPAD														**4**			2				
SCCPB1														**1**	-2		5		2; -1		
GRCPB1	-1															**11**			-1		
GRCPAD	1										3				**6**	**7**					1
GRCPB2	-2														**-8**	**-2**	1		1		
SOCPB3		-1														**10**			-1		
SOCPB2	-1	-1														**12**			-1		
SOCPAD		-1	-2								3				**4**	**11**					
SOCPB1		-1			1					1					**1**	**13**					
ARCPB2	-3												2				**6**		**4**		
ARCPB1																	**4**	**1**	**8**		
VICPAD								2									**1**	**9**	**2**		
VICPB1						7			-1								**1**	**7**	**1**		
AXCPAD												1		2			**8**	**8**	**7**		
AHCPAD										1										**13**	
YOCPAD																					13
YOCPB1											3					3					8
YOCPB2		1														6					5

Note. Absolute frequencies of the 490 unique lexical elements composed of either one contextualised behaviour-descriptive verb or one trait-adjective that occurred in six raters' item interpretations and their encoding in the BR_xBS-Approach-generated working constructs. Negative scores indicate frequencies of lexical elements encoding a construct's inverted meaning. [a] The item code is composed of two digits abbreviating the working construct, two digits indicating the species (CP = capuchin) and two digits indicating the item format (B1 to B3 = behaviour-descriptive verb items, AD = trait-adjective items). Construct abbreviations: AG Aggressiveness to conspecifics, AH Aggressiveness to humans, AR Arousability, AX Anxiousness, CO Competitiveness, CR Creativeness/ Inventiveness, CU Curiousness, DI Distractibility, DO Dominance, FO Food orientation, GR Gregariousness, IM Impulsiveness, PA Physical activity, PE Persistency, PL Playfulness, SC (Self-)Cleanliness, SH Social orientation to humans, SO Social orientation to conspecifics, SX Sexual activity, VI Vigilance, YO Social orientation to youngsters. Grey cells indicate lexical elements that were encoded in the particular working constructs that the items were meant to operationalise. Working constructs are sorted by their interrelations in the assessment factor scores; bold frequencies indicate lexical elements that were assigned to the same factors on which the given item showed its highest loading in the assessment data (see Table 3).