Comparing individuals within and across situations, groups and species: Metatheoretical and methodological foundations demonstrated in primate behaviour

Jana Uher [1,2]

Correspondence*:
[1, *] The London School of Economics and Political Science, St Clemens Building, Houghton Street, London WC2A 2AE, United Kingdom
[2] Comparative Differential and Personality Psychology, Free University Berlin, Germany

Individuals are explored in various kinds of phenomena and contexts. But how can scientists compare individual variations across phenomena with heterogeneous properties that require different methods for their exploration? How can measurements of individual variations be made directly comparable between different studies, groups of individuals or even species? This research applies the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) to elaborate metatheoretical concepts and analytical methodologies for quantitative comparisons of individual variations within and across situations, groups and species using behavioural phenomena as examples. Established concepts from personality psychology, differential psychology and cross-cultural and cross-species research are systematically integrated into coherent frameworks and extended by adding concepts for comparing individual-specific variations (i.e., "personality") between species. Basic principles for establishing the functional comparability of behavioural and situational categories are elaborated while considering that individuals from different groups and species often show different behaviours and encounter different situations and therefore cannot be studied with identical variables as is done in assessment-based research. Building on these principles, the chapter explores methodologies for the statistical analyses of the configurational comparability of constructs and of mean-level differences between groups and species. It highlights that situational properties are crucial for quantitative comparisons of individual variations. Fundamental differences between observational methods and assessment methods are explored, revealing serious limitations and fallacies inherent to comparisons of individuals on the basis of assessments. Implementations of the methodological principles and concepts presented are illustrated with behavioural data from four primate species (weeper capuchins, mandrills, toque macaques and rhesus macaques).

Keywords:
- Cross-species and cross-cultural comparative methodology;
- Scientific measurement and quantification;
- Configurational comparability/ structural or construct equivalence;
- Patterning and positioning effects;
- Limitations of assessment methods;
- Weeper capuchin (*Cebus olivaceus*), mandrill (*Mandrillus sphinx*), toque macaque (*Macaca sinica*), Rhesus macaque (*Macaca mulatta*);
- Sex/ gender differences and species differences.

***

14 Comparing Individuals within and across Situations, Groups and Species: Metatheoretical and Methodological Foundations Demonstrated in Primate Behaviour

Jana Uher

Key words:

Cross-species and cross-cultural comparative methodology, Scientific measurement and quantification, Configurational comparability/ structural or construct equivalence, Patterning and positioning effects, Limitations of assessment methods, Weeper capuchin (*Cebus olivaceus*), mandrill (*Mandrillus sphinx*), toque macaque (*Macaca sinica*), Rhesus macaque (*Macaca mulatta*), Sex/ gender differences and species differences

Abstract

Individuals are explored in various kinds of phenomena and contexts. But how can scientists compare individual variations across phenomena with heterogeneous properties that require different methods for their exploration? How can measurements of individual variations be made directly comparable between different studies, groups of individuals or even species? This research applies the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) to elaborate metatheoretical concepts and analytical methodologies for quantitative comparisons of individual variations within and across situations, groups and species using behavioural phenomena as examples. Established concepts from personality psychology, differential psychology and cross-cultural and cross-species research are systematically integrated into coherent frameworks and extended by adding concepts for comparing individual-specific variations (i.e., "personality") between species. Basic principles for establishing the functional comparability of behavioural and situational categories are elaborated while considering that individuals from different groups and species often show different behaviours and encounter different situations and therefore cannot be studied with identical variables as is done in assessment-based research. Building on these principles, the chapter explores methodologies for the statistical analyses of the configurational comparability of constructs and of mean-level differences between groups and species. It highlights that situational properties are crucial for quantitative comparisons of individual variations. Fundamental differences between observational methods and assessment methods are explored, revealing serious limitations and fallacies inherent to comparisons of individuals on the basis of assessments. Implementations of the methodological principles

and concepts presented are illustrated with behavioural data from four primate species (weeper capuchins, mandrills, toque macaques and rhesus macaques).

Introduction

Comparisons between individuals are central to many fields of study, such as to neuropsychological, behavioural, social and "personality" research. In these fields, diverse kinds of phenomena (e.g., morphology, physiology, behaviour, the psyche) are explored in a multitude of contexts in which they are embedded. Individuals are compared with one another within and across different biosocial and sociocultural groups, such as groups of individuals of particular age, sex/gender, socio-economic status, nationality and language (Mischel, Shoda, & Ayduk, 2007).

Researching individuals of nonhuman species provides unique opportunities for comparing between-individual variations also across species that differ in their behavioural and social systems, ecological adaptations and phylogenies. Such comparisons are important for exploring the mechanisms and processes of the evolution of population differences within and between species (Capitanio, 2004; Hammock & Young, 2005; Uher, 2008a,b). Cross-species comparisons are also essential to unravel the ways in which humans are, in fact, unique (Tomasello, 1999; Uher, 2013).

The complexity of each of the various kinds of phenomena studied in individuals and the diversity of comparisons that can be made between individuals within and across different situations, cultural groups and species (e.g., cross-situational, cross-cultural and cross-species) entail particular challenges to research methodology. In many fields of psychological and social science research, assessments have become a primary source of information for the scientific exploration and comparison of individuals (Mischel et al., 2007). Judgements on individuals play important roles not only in research but also in everyday life (e.g., at work, schools).

But assessments are prone to various fallacies of the human mind (Uher, 2015a,b,c). Specifically, because individuals are members of the same species, they share basic organismal properties and are therefore substantially similar to one another (Uher 2011a,b). To judge individuals, people therefore abstract from properties that are common to all individuals and focus on the *minor variations* that are informative for differentiating highly similar individuals from one another, such as fine-grained differences in individuals' faces and behaviours. The particular kinds of minor variations that people use most frequently and in so-cially shared ways for differentiating individuals in their everyday lives are men-

tally represented in more pronounced ways. The corresponding individual variations are then considered to be *salient*, and they appear in people's minds to be much larger than can actually be observed (Lahlou, 1998; Uher, 2015b).

By contrast, to differentiate individuals of different biosocial and sociocultural groups, people tend to homogenise in their minds within-group differences and to exaggerate between-group differences, and they attribute salience to the particular properties on which these discriminations are based (Connolly, 1998; Gillespie, Howarth, & Cornish, 2012). This also influences people's perceptions of individuals because people may pay more attention to observable properties that match their implicit ideas and social beliefs but less attention to other properties that do not match, thus reinforcing stereotypical representations of individuals from different groups. Stereotypical beliefs about differences between individuals of different gender/sex, age group, ethnicity, nationality, language, amongst others, are found all around the globe (Schneider, 2005).

Salient individual and group differences form important parts of the everyday knowledge and social belief systems that human communities develop and maintain (Jovchelovitch, 2007; Uher, 2013). These socio-cognitive category systems enable humans to quickly form impressions of other individuals that are sufficiently viable for the purposes of first encounters. These category systems are lexically encoded in complex everyday vocabularies that enable people to efficiently communicate their impressions of individuals to others and to develop, maintain and propagate pertinent knowledge, attitudes and social beliefs (Dunbar, 1996; Goldberg, 1982; Lahlou, 1996, 2001; Uher, 2015d).

In fact, it was argued that these socio-cognitive and linguistic abilities first enabled our human ancestors to deal with unknown individuals of other communities in peaceful ways, thus enabling peaceful exchange, traffic and trade between different communities—abilities of fundamental importance for human evolution (the so-called *personality-constructs-promote-peaceful-anonymous-contacts-hypothesis*; Uher, 2013). These human abilities are also important for present-day humans' ability to live in large-scale societies in which they have to deal with many anonymous others in their everyday lives (Blaffer-Hrdy, 2009; Uher, Werner & Gosselt, 2013b).

These social beliefs, ideas and vocabularies inevitably influence the assessments that people make about individuals. For standardised assessments, questionnaire scales are widely applied. They require respondents to judge the assumed degree to which particular characteristics described in statements are applicable to a target individual and to indicate their graded judgements using a set of predetermined category units. The data that can be generated by this procedure constitute much of today's empirical research in many fields of psychology and the social sciences (Baumeister et al., 2007; Schwarz, 1999). Assessment data are also used widely in economic, medical, educational and governmental research (Hammersley, 2013; Menon & Yorkston, 2000) and increasingly also in

basic and applied animal behaviour research (Tetley & O'Hara, 2012; Uher & Asendorpf, 2008).

Assessment tools are popular because they are rather easy to implement, time and cost efficient and because they produce highly standardised sets of data. But the ways in which respondents use these tools and by which they generate their answers to standardised scales are still not well understood (Dirnwächter, Valsiner & Sauck, 2004; Rosenbaum & Valsiner, 2011). In assessments, people are required to judge a target individual on the basis of their previous experiences with that individual—thus, in retrospect. Therefore, assessments are inherently based on the raters' memory and mental reconstruction of events that they have perceived in the past and of impressions that they have previously formed (Uher, 2015a,b,c).

But human memories are fallible; the accuracy of reconstructions of past experiences is well known to be constraint and biased by errors in memory and reasoning (Fahrenberg, Myrtek, Pawlik, & Perrez, 2007; Gigerenzer, Todd & the ABC Group, 1999; Schacter, 1999; Shweder & D'Andrade, 1980; Uher, 2013). Moreover, memories are never just recalled in identical ways but are always re-constructed anew in the context of the given situation (Bartlett, 1932; Schacter & Addis, 2007; Uher, 2015a,b,c,d). Memory recall and thus, also the construction of assessments therefore depend on the states of mind that individuals may have in particular situations (Omi, 2012; Uher, 2013).

The item statements and answer categories used in assessment tools are commonly worded in everyday terms so that the raters can understand them without detailed instruction or even training of how to use these tools to generate data (as is the case in observational methods; see below). Thus, scientists capitalise on the raters' commonsense knowledge for interpreting the meaning that the statements and category units provided in a particular assessment tool may have for these raters in view of the specific assessment task at hand. Given this, it may be inevitable that assessment methods activate all kinds of sociocultural and stereotypical beliefs, lay theories and implicit norms and values, such as those about gender/sex (Lloyd & Duveen, 1992), ethnicity (Howarth, 2009), age and social status (Schneider, 2005; Uher et al., 2013b) that are widespread in all human communities.

From the particular requirements of assessment methods it follows that assessment data reflect properties of the raters' personal and socially shared ideas, beliefs, knowledge and psychical representations about the phenomena to be assessed but they cannot reflect properties of these phenomena in and of themselves (Uher, 2015a,b,c). Knowledge and belief systems are always context-dependent (Jovchelovitch, 2007). Therefore, they can be used flexibly and in context-sensitive ways in everyday life—and thus, also for assessments (Hammersley, 2013). As a consequence, assessment-based findings depend on the particular contexts of the raters, such as their particular stages of development (Toomela, 2010), educational level (Vygotsky, 1934) and their abilities to reflect on themselves, their

world (Omi, 2012) and their life histories (McAdams, 1985). Assessments also depend on the raters' particular situation (Bandura, 1986; Lazarus, 1981; Mischel, 1968; Rotter, 1954); therefore, numerous studies could demonstrate that (experimentally induced) motivations and goals significantly influence the outcomes of assessments (Biesanz & Human, 2010).

Information obtained with assessments also depend on the context of particular groups of individuals that raters may implicitly consider (Heine, Lehman, Peng, & Greenholtz, 2002). This accounts for numerous, apparently puzzling findings from comparative research on individuals such as repeated findings showing that the average scores of "personality" assessments of people from different nations did not correspond to national character stereotypes (e.g., McCrae, Terracciano & 79 Members of the Personality Profiles of Cultures Project, 2005; McCrae, Terracciano, Realo, & Allik, 2007a,b; Perugini & Richetin, 2007; Terracciano, Abdel-Khalak, Adam, Adamovová et al., 2005).

Given the context-sensitivity of social beliefs, values and knowledge and the inherently memory-based generation of assessments, standardised assessment methods are not appropriate for valid comparisons between individuals and for exploring and explaining real world issues in the increasingly pluralistic societies we live in today (e.g., in occupational, clinical, educational and economic settings). Moreover, the standardised data formats produced with assessment methods obscure the fundamental metatheoretical and methodological challenges that are inherent in any comparative study. For how can scientists establish comparability between phenomena, events or properties that are heterogeneous and that require for their exploration different kinds of methods and different units of measurement? How can quantitative measurements be made directly comparable between specialised disciplines, as this is possible in the physical sciences? Establishing such comparability is important to enable the interdisciplinary collaboration that is necessary to comprehensively explore the complex interplay between genes, brain and behaviour in individuals, such as the complex interplay between genes, brain and behaviour.

This research applies the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm, see below; Uher, 2015a,b,c) to elaborate the metatheoretical concepts and analytical methodologies that are required for systematic quantitative comparisons between individuals within and across different situations, groups and species. These elaborations are made and illustrated using the phenomena of behaviour as example. Behaviours are chosen because they can be directly perceived in all kinds of individuals including those who are not able to self-report (e.g., young children, nonhuman animals). But behaviours also feature some challenging properties, such as the momentariness and dynamic fluctuations of behavioural events, which hinder direct comparisons between individuals and considerably complicate scientific measurements (Uher,

2015a). Explorations of other physical phenomena of individuals (e.g., morphology, physiology) can directly build on these principles. They also apply to explorations of the phenomena of the psyche; but because psychical[2] phenomena cannot be directly perceived in others, additional methodological peculiarities must be considered (Uher, 2015b,c,d).

First, the chapter outlines the relevant foundations that the TPS-Paradigm provides for research on behavioural phenomena and for scientific measurement and quantification. Using these fundamentals, established concepts for exploring variations within and between individuals and groups are systematically integrated and extended to the level of species, highlighting three central viewpoints on which all kinds of comparative explorations are based. Building on these concepts, the chapter elaborates basic principles for establishing the comparability of behavioural categories between groups and species that show different behaviours and that therefore cannot be studied using exactly the same behavioural variables. These principles are used to scrutinise the widespread practice of using "identical" variables to establish comparability in assessment-based research. Then the chapter explores methodological approaches needed to statistically analyse the configurational comparability of constructs and mean-level differences between groups and species, showing that situational properties are essential for quantitative comparisons. Implementations of the principles and concepts presented are illustrated by behavioural data from four nonhuman primate species.

The Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm)

The TPS-Paradigm comprises interrelated philosophical, metatheoretical and methodological frameworks for exploring individuals. In these frameworks, concepts, approaches and methods from various scientific disciplines were systematically integrated, further developed and complemented by novel ones; therefore it is a *transdisciplinary* paradigm (Uher, 2008a,b, 2011a,b, 2013, Desideratum 1a). It is called a *philosophy-of-science* paradigm because it emphasises the importance of making explicit the philosophical presuppositions that are made in a given scientific system and the metatheories and methodologies that are derived from them, and it aims to make these explicit in the fields that it explores (Uher, 2015a). The TPS-Paradigm therefore provides the necessary foundations from which to scrutinise established theories and methods from various fields of research for their commonalities and differences and to refine and develop metatheories and methodologies that are applicable across disciplines.

[2] The TPS-Paradigm uses the term *psychical* rather than *psychological* because "events, processes and structures that are properly called psychical do not become psychological until they have been operated upon in some way by the science of psychology" (Adams & Zener in Lewin, 1935, p. vii).

The potentials of the TPS-Paradigm for enabling such explorations were already demonstrated, amongst others, in critical analyses of taxonomic "personality" research. In this field, the TPS-Paradigm revealed fundamental mismatches between the researchers' (implicit and explicit) concepts of "personality" and the methodological rationales underlying the selection approaches, methods of data generation and reduction principles that are used for developing comprehensive "personality" taxonomies (Uher, 2015b). Further mismatches were revealed between widely established interpretations of the taxonomic "personality" models, constructs and data thus-derived and their actual abilities to reflect individual-specificity in the targeted phenomena (Uher, 2015c). Basic concepts of the TPS-Paradigm were also used to explore the phenomena of the psyche, their microgenetic, ontogenetic and (potential) phylogenetic development and their roles in the emergence of language, social institutions and societies (Uher, 2015d).

At the heart of the TPS-Paradigm is the metatheoretical definition and differentiation of various kinds of phenomena that are explored in research on individuals (i.e., the phenomena of morphology, physiology, behaviour, the psyche, semiotic representations, artificially modified outer appearances and contexts). These differentiations are made on the basis of *three metatheoretical properties* that can be conceived in different forms for each given phenomenon: *1)* spatial location in relation to the studied individual's body, *2)* temporal extension and *3)* spatial extension, i.e., physicality versus "non-physicality".[3] These three spatio-temporal properties are considered because they determine a phenomenon's perceptibility by individuals in everyday life and—thus, also its accessibility by researchers and the methodologies required for explorations (Uher, 2015a,b,c).

What is Behaviour? Metatheoretical Definition and Methodological Implications

Behaviours are defined in the TPS-Paradigm as the "external changes or activities of living organisms that are functionally mediated[4] by other external phenomena (Millikan, 1993) in the present moment" (Uher, 2013, 2015a,b,c). Thus, behaviours are not just movements[5] (e.g., freezing in prey animals). Vice

[3] In the TPS-Paradigm, the term "non-physical" is put in quotation marks as it denotes properties that are not simply contrasted against the physical but are complementary instead (see Uher, 2015a).

[4] In the TPS-Paradigm, the meaning of the term *mediation* refers to the Latin *mediare*, to be in the middle, not to the meaning established in statistics (where it is differentiated from moderation).

[5] Importantly, the TPS-Paradigm conceives neither of physiological responses nor of mental activities as behaviours, in contrast to behaviourist and cognitivist paradigms (Skinner, 1957; Koffka, 1935). Moreover, it makes no a priori assumptions about the potential causes of behaviours in other kinds of phenomena as is implied by concepts of behavioural "responses" or "goal-directed" actions, amongst others. Instead, the TPS-Paradigm considers that different metatheoretical properties can be conceived for other kinds of phenomena that therefore require different methodologies for their exploration and thus also for explorations of their interrelations with behaviours (Uher, 2015a,b,c).

versa, not all movements, external changes or activities are behaviours (e.g., mere chemical by-products such as heat); they are behaviours only if their functions have *reference to* other external phenomena or to connections with them (Millikan, 1993). Hence, the meaning and function of behaviours can be explored only by considering the particular contexts in which they occur (cf. the behavioural situation, below; Uher, Addessi, & Visalberghi, 2013a; Uher et al., 2013b).

The TPS-Paradigm conceives of *functions* as temporal interrelations that regularly occur between particular kinds of phenomena, events or properties—thus, as established effects (derived from the Latin *effectus* meaning "worked out, brought about, accomplished"). Importantly, functions thus-defined neither imply purpose nor intention because teleological properties presuppose that possible prospective outcomes are simulated and evaluated on the basis of retrospective analyses of experiences made in the past, which is possible only for psychical phenomena (Uher, 2015d). Moreover, functions denote not only causal connections of various kind (Kausal-Zusammenhänge) but also compositional connections (Gefüge-Zusammenhänge) in which the interacting elements co-occur in coordinated ways and match and cooperate with one another such that the entirety of their joint interactions results in complexes and functions of higher organisation (Rothschuh, 1963; Uher, 2015a,c).

Behavioural phenomena are located external to the body of the individual under study and can therefore be directly perceived by observers. But observations are complicated because behaviours are bound to the present moment. Moreover, behavioural events often have only low temporal extensions but unfold continuously and dynamically in time and are therefore often highly fluctuating. Hence, behaviours can be observed and thus be recorded only in the particular moments in which they occur. Methods allowing for the real-time recording of momentary events are referred to as *nunc ipsum* methods (from the Latin *nunc ipsum* for at this very instant) in the TPS-Paradigm (Uher, 2015a).

Behaviours are immaterial physical phenomena in and of themselves (e.g., sound waves, movements); therefore, rather constant spatial and temporal units suggesting clear demarcations of single events are largely absent. But spatio-temporal units can be identified on the basis of the physical matter of individuals' bodies (e.g., arms, torso) and of their external physical surroundings (e.g., trees) to which behaviours are bound. This boundedness to matter allows for identifying behavioural events that are repeatable at least to some extent although the units that can be identified in this way vary in their spatio-temporal extensions (e.g., arm movements to pick apples from a tree). These peculiarities have important implications for scientific measurement and quantification as explored below.

The phenomena of contexts—Situations and behavioural situations

Contextual phenomena (derived from the Latin *contexere* to interweave, link together) can be identified and conceived only with regard to the particular individual under study. Specifically, individuals interact with only those external physical events that are relevant for them given their particular physical and psychical properties; hence, the same physical surroundings are not the same for all individuals (Bandura, 1986; Lewin, 1936; Rotter, 1954; von Uexküll, 1909). Dualistic conceptions, such as of "individual versus environment" or "nature versus nurture" are therefore misleading and inherently circular (Uher, 2015a,b,c).

Considering this, the TPS-Paradigm defines a *situation* as the particular constellation of events of all kinds of phenomena both external and internal to the body of the studied individual that are present in a given moment. Thus from the universe of all external physical events, an individual's *situation* comprises all those particular events that are present in a given moment and that are therefore directly perceptible by the individual, consciously or not. Because individuals' bodies are always present, a situation also comprises all psychical events (e.g., physical condition, food deprivation). It also comprises internal physical events that are ongoing, thus also all those memories and psychical representations that the individual retrieves and reconstructs in the given moment, such as representations of events that are absent in and of themselves (Uher, 2015a).

Behavioural situations are situations of special kind that denote the particular constellation of those external physical events that functionally mediate an individual's external changes or activities in a given moment—his or her behaviours. Thus, the events that constitute behavioural situations are external to the individual in and of themselves. But the criterion for demarcating these particular events from the universe of all external physical events present in a given moment is bound to the properties of that individual. This criterion is the events' effectiveness to make functional the individual's external changes or activities that thereby become behaviours. Importantly, this demarcation is based on individuals' *external* bodily events that can be directly observed (i.e., behaviours) rather than on internal ones (i.e., psychical events). This facilitates the identification of those particular external properties that may functionally mediate the behaviours of a particular individual in a particular situation, which may be neither directly apparent nor known a priori. To identify the functions that a given behavioural event may have for particular individuals in particular situations, behavioural scientists therefore explore proximate spatio-temporal interrelations in the empirical occurrences of different behavioural events with one another and with other events in the individuals' external surrounding (see below; e.g., Preuschoft, 1992; Preuschoft & van Hooff, 1995).

Exploring individuals' behaviours in different kinds of situations, as intersubjectively categorised on the basis of particular constellations of observable

properties (see below), constitutes an important approach for researchers to explore the possible psychical relevance and meaning that particular situational properties may have for particular individuals even if these individuals cannot be interviewed about their personal views, such as young children (Shoda, Mischel, & Wright, 1994; Wright & Zakriski, 2003) and nonhuman animals (Hinde, 1974, 1982). Such analyses presuppose metatheoretical and analytical concepts that enable scientists to quantify their study phenomena appropriately.

Philosophy-of-science principles of scientific measurement and quantification

Using concepts of the science of measurement—metrology (JCGM, 2008), the TPS-Paradigm elaborated the philosophy-of-science foundations of scientific quantification, highlighting two central requirements. First, researchers must specify the sets of the elements to be quantified in the phenomena under study; this is called the *set-theoretic requirement*. Second, the elements thus-defined must be compared with designated fixed standards to express their ratio as a real number; this is called the *algebraic requirement* of scientific quantification (Uher, 2015a,b,c). The TPS-Paradigm refers to numerical data that fulfil the set-theoretic and algebraic requirements as *scientific* quantifications—as opposed to (subjective) quantifications in which these requirements are not fulfilled. It highlights that the opportunities for meeting these two requirements are unequivocally determined by the three spatio-temporal properties that it considers. Importantly, these requirements must be fulfilled *with regard to the phenomena under study* and not only with regard to the phenomena that are used as data (Uher, 2015a,b, submitted).

What are *data*? In many fields of research, scientists seldom directly explore the phenomena under study in and of themselves (e.g., behaviours) as is the case when, for example, in medical investigations body tissue is dissected and dyed, thus changed and destroyed. Research on individuals mostly relies on persons (e.g., researchers, observers) who convert information from their observations of single occurrences of the phenomena under study into information encoded in signs (Uher, 2015c). The TPS-Paradigm highlights that *signs* are composite kinds of phenomena that consist of physical phenomena that are external to individuals' bodies (e.g., vocalisations, ink on paper) and of psychical phenomena (e.g., meanings) that particular individuals attribute to these external physical phenomena and that first make them signs. Behavioural and material signs (e.g., spoken and written language) are developed and used to physically represent meanings and all kinds of phenomena that individuals can perceive and/or conceive of (e.g., behaviours, buildings, deities). This enables the transmission of meanings also in absence of the particular objects to which they refer

and across time and space, making signs important means for communication and the development and transfer of knowledge (Uher, 2015d).

The signs used in science (e.g., numerals, lexically encoded variables) and to which scientists attribute particular meanings (e.g., mathematical properties, behavioural events observed) are commonly called data. Data can be stored and analysed, thus manipulated, decomposed and recomposed *in lieu of* the actual phenomena of interest and in ways that may not be enabled by these phenomena in and of themselves (Uher, submitted).

The essence of data generation

The pivotal point for generating data are the decisions that observers and researchers make on the ways in which particular information from the actual phenomena under study can be appropriately represented in signs (i.e, data). When information from one kind of phenomenon are represented in another kind of phenomenon, this is called *conversion* in the TPS-Paradigm. Human individuals explicitly and implicitly convert information between different kinds of phenomena all the time they are communicating (Uher, 2015d). To enable systematic conversions and standardised encodings of information that are considered to be relevant for a given investigation, researchers must therefore specify the particular operations and practices that are used to implement these decisions—hence, the *methods of data generation* (Uher, 2015a).

The choice of the particular method of data generation depends on the theoretical definition of the object of research and on its operational definition in a given study. Therefore, researchers must specify the particular intersubjective agreements that are made about which particular pieces of information about the properties and events of the phenomena under study as perceived and conceived by human individuals (e.g., observers) can be demarcated and categorised in what particular ways. These specifications are essential for meeting the two requirements of scientific quantification.

Fulfilling the set-theoretic requirement for scientific quantification

The set-theoretic criterion of scientific quantification requires researchers to specify all the elements to be quantified. Thus, to scientifically quantify individual behaviours, researchers must specify all studied elements of the set B of behavioural events, the set S of behavioural situations that (may) functionally mediate these behaviours, the set I of individuals in which they are explored and the set T of occasions and periods of time in which they are recorded in a given study (Uher et al. 2013a).

The fact that quantifications of individual differences always depend on the particular set I of individuals to which a given target individual is compared

is known as *reference group effect* (Heine et al., 2002). It is also well known that differences in the studied elements of the set *T* can influence empirical results, such as the effects that aggregations have on the reliability of measurements (Epstein, 1979, 1980). But it is often not well considered that the same principle equally applies to the elements of the sets *B* of behaviours and *S* of behavioural situations because the functions and meanings that behaviours may have for individuals depend on the particular situation in which they occur. Moreover, in the same sample of individuals, different behavioural events generally occur with different average frequencies in both the same and different kinds of situations (Uher, 2011a,b). Thus, comparing behavioural occurrences between individuals without considering such behaviour-specific and situation-specific variations can bias quantitative measurements.

The methodological necessity to specify the studied elements in *all* sets considered tends to be overlooked likely because the different kinds of phenomena impose different demands on researchers and observers for meeting this requirement. Specifically, the elements of the sets *I* and *T* can be specified easily by conceiving of individuals (on the basis of their bodies) as discrete physical material entities and by using the standard instruments for measuring time that are widely used also in everyday life. But to specify which particular elements of the sets *B* and *S* are being explored in a given study requires researchers and observers to decide on how to demarcate distinct kinds of events in the heterogeneous properties that can be perceived in these complex phenomena and on how to categorise these events. This entails particular challenges.

Challenges and principles for developing category systems in behavioural research

Developing category systems that specify all the elements of the sets *B* and *S* that are explored in a given study is complicated because, in behavioural phenomena, clear-cut and mutually exclusive categories often cannot be easily defined as this may be possible, for example, for categorisations of chemical elements. Demarcating and categorising different kinds of events is complicated because the momentary and fluctuating occurrences of behaviours hinder direct comparisons of their observable properties across different occasions and across different situations and individuals.

Demarcating behavioural events in the flow of their occurrence is also complicated because the spatio-temporal units that can be identified in behavioural phenomena often vary considerably in their extensions and are therefore identically repeatable only to some extent. For example, human individuals can smile in very different ways. They can turn the corners of their mouths up in

different ways, with open or closed mouth. In addition, individuals may also produce laughter lines around their eyes with different levels of intensity. Given this considerable variability that is perceptible in the spatial properties of these behaviours, what entity can be demarcated and categorised as *one* ($n = 1$) smile? Which particular entities can be considered to be of the same kind, thus as identically repeatable at least to some extent? Moreover, smiles also differ in their temporal extension. Can a quick smile and a long-lasting smile be conceived as events of the same kind?

Whereas demarcations of behavioural events are made on the basis of the similarity in their *perceptible spatio-temporal forms*, these properties cannot be used for categorising the events thus-specified. The TPS-Paradigm emphasises that in research and in everyday life categorisations of behaviours are based on their *known or theoretically assumed functions* (e.g., slapping someone on the shoulder to kill a midge versus to bully that other individual). In fact, functionally similar behaviours often differ from one another in their spatio-temporal forms. In school children, for example, slapping and verbally insulting others can be considered functionally similar in particular kinds of situations (e.g., to win over others), but their particular spatio-temporal forms differ considerably (cf. Murray-Close & Ostrov, 2009).

Likewise, empirical co-occurrences of different behavioural events across the individuals studied (as studied, e.g., with factor analysis) cannot be used to derive summary categories of behavioural events—contrary to widespread practices in "personality" research. Rather, empirical co-occurrences of functionally related behaviours can also be low (e.g., of contact and non-contact aggression; Uher, 2011b; Uher et al., 2013a). But no matter such empirical inconsistency, in their everyday lives, people commonly categorise behaviours of *similar function* as behaviours of *similar kind* (e.g., as "aggression") and they also do so in standardised assessments.

In behavioural research, hierarchical *category systems* in which more specific categories of behavioural events are summarised by more abstract categories are established on the basis of the functions and meanings that the behaviours may have for the individuals studied. Hence in category systems, research-ers specify not only events to be studied but also the *theoretical* interrelations of the specified elements both within and between the sets *B*, *S*, *I* and *T*. This is prerequisite for exploring the empirical interrelations between the elements studied—thus, for their scientific quantification as explored below.

As functions denote temporal interrelations that regularly occur between particular kinds of phenomena, events or properties, functions can be identified by means of contextual pre-post analyses of behavioural sequences studied in observations and by experimentation (Hinde, 1974, 1982; Preuschoft, & van Hooff, 1995; Visalberghi, Valenzano, & Preuschoft, 2006). But the particular external situational properties that are, in fact, functionally mediating particular be-

haviours in particular individuals are not always known a priori or directly apparent and they may not even be consciously perceptible by human individuals (e.g., subliminal odours; Li, Moallem, Paller, & Gottfried, 2007). Therefore, it may be possible that observers and researchers specify properties that actually do not fulfil mediating functions for the particular behaviours of the particular individuals studied. This is ultimately a matter of research. A first step towards identifying functional interrelations is to specify the situational properties (i.e., the elements of the set S) that can be directly perceived and intersubjectively conceived by multiple human observers. Given this, initial specifications may refer only to properties that are salient to individuals of particular sociocultural communities or only to humans. But with accumulating evidence, these specifications allow for exploring which particular constellation of properties can legitimately be assumed to fulfil mediating functions for particular kinds of behaviours in particular kinds of individuals.

Behavioural events are sometimes also categorised with regard to their *degrees of functional intensity*. This presupposes that their known or theoretically assumed functions and meanings that they may have for particular individuals, groups or species are considered to be qualitatively comparable. In school children, for example, slapping and verbally insulting others may be considered functionally comparable because both behavioural events represent offences in particular situations. But because of their different perceptible properties, differences in their degrees of intensity—thus quantitative differences—cannot be readily specified. For example, one may assume that slaps constitute a more offensive behaviour than verbal insults. But how could this be quantified? Can five verbal insults be considered to be as equally harmful as one slap?

Moreover, the functional intensity of behaviours also depends on the particular behavioural situations in which they occur. Slapping another child for taking away a toy may be considered less aggressive, and thus less intense, than slapping others who offer to play together (cf. Wright & Zakriski, 2003), even though the perceptible properties of the slap, in and of themselves, may be exactly the same in both kinds of situations. Often there are no natural entities on which decisions can be made about intensity differences between functionally related behaviours. Such decisions are thus arbitrary to some extent; they depend on the current state of knowledge and, in particular, on the particular social values that are socioculturally attributed to behaviours.

Further challenges arise because the spatial, temporal and functional properties that can be identified in behaviours are sometimes interrelated. For example, play behaviours necessarily involve physical activity; an individual that does not move at all does not play. But a motionless position may still have a social function, such as when individuals rest next to others rather than alone. Therefore, functional properties of the same behavioural event can be encoded in different behavioural categories (e.g., locomotive, play and social behaviours)

even in the same study (for an example, see Uher et al., 2013a). This multiple categorisation of functional properties of one and the same behavioural event corresponds to the intuitive categorisations of behaviours that people make in their everyday lives.

In sum, in category systems researchers specify the known or assumed *theoretical* interrelations of the studied elements both within and between the sets B, S, I and T. This specification is prerequisite for exploring the empirical interrelations between the elements studied—thus, for their scientific quantification as explored below.

Principles for developing encoding schemes

Once the category system is developed for an investigation, researchers must specify encoding schemes that enable observers to quickly recognise relevant properties in the phenomena under study and to categorise and encode occurrences of defined events systematically and in standardised ways into raw data. During live observations, these requirements must be fulfilled real-time—i.e., nunc-ipsum, while the behaviours occur. Video-based observations decelerate this task and facilitate perceptions of the relevant properties, such as by enabling repeated and slow-motion playback. To establish behavioural encoding schemes, researchers can capitalise on the broad range of methods developed in ethology (e.g., Martin & Bateson, 2007; Lehner, 1998).

Importantly, researchers must consider that behavioural observations are never free of implicit decisions made by the observers (Pelligrini, 2011). This is because categorising individual behaviour is an important ability that humans use frequently in their everyday social lives and for which they have developed comprehensive sets of socio-cognitive categories. To ensure that observers interpret and use the encoding schemes for generating data as specified for the purposes of a given study, observers are instructed and trained. Therefore, researchers must specify the observable properties that are used to demarcate and categorise the particular kinds of events studied. This is especially important for abstract behavioural categories (e.g., "aggression") that are sometimes used to encode a heterogeneous set of behavioural events that could also be encoded in several more specific coding categories (e.g., "verbal insult", "slap"). The defining properties must be explicitly and sufficiently described in intersubjective ways such that different observers demarcate and categorise behavioural events and encode their occurrences in very similar ways as statistically tested by inter-observer reliability.

The TPS-Paradigm emphasises that encoding schemes can rely *only* on those properties of the events and phenomena under study that are or can be (technically) made *directly perceptible* and intersubjectively conceivable by multiple human individuals. In their encoding schemes, researchers must therefore address the particular forms that can be conceived for the given study phenomenon with

regard to each of the three spatio-temporal properties because these properties determine a phenomenon's general accessibility by human individuals. Therefore, these properties also determine the accessibility of further, more specific properties on which demarcations and categorisations of events may be based in a given study (e.g., sound) and that are therefore also specified in the encoding schemes (Uher, 2015a,b,c, submitted).

For enabling systematic encodings, behavioural categories are defined on the basis of properties that occur in comparable constellations in all kinds of events that are assigned to other categories. Such definitions can be based, for example, on the different communication modes such as by specifying the particular constellations of verbal, vocal, facial, gestural or postural properties that occur in all behavioural assigned to a particular category.

As behaviours are functionally mediated by other external events and are thus always contextualised, behavioural encoding schemes must also specify properties of the particular behavioural situations studied (Uher, 2013, Desideratum 2c). In behavioural studies, these specifications are always contained at least implicitly because they are carried out and their category systems are set up for investigations in particular settings, such as school classes or experimental laboratories. As with behavioural categories, defining encoding schemes for behavioural situations can rely only on those properties that are or can be made consciously perceptible by human observers. Situational categories can be defined in various degrees of abstraction. For example, descriptions of the situations in which school children are observed could refer to social settings in general. More specific sub-categories could differentiate contacts to adults or peers, to familiar or unfamiliar peers, same or opposite sex peers in conflict, play or teaching situations, involving one other individual, a group or an anonymous mass, amongst others. The degree of specification of situational properties depends on the purpose of the given study.

In sum, specifying the perceptible properties on which the demarcations and categorisations of behavioural events are based makes painfully apparent that identifying similarity—thus, establishing comparability—between events other than those that can be conceived to be identically repeatable at least to some extent is always a matter of decision, theoretical interpretation, degree of abstraction and the particular aim and purpose of the given study. These decisions are seldom made explicit but they are always involved in every observation of individual behaviour in both research and everyday life.

Such decisions are also involved in assessments but the standardisations that are implemented in these methods obscure the complexity of these decision. Assessment instruments comprise standardised statements and answer categories that are often global and abstract and not further specified in terms of perceptible

properties of events to which they may refer. In contrast to observers in a behavioural study, raters are neither instructed nor trained to use the fixed encoding schemes provided in systematic and standardised ways for generating data. As a consequence, the particular elements of the set B, S, I and T that raters may implicitly consider when generating assessments and the ways in which they encode the elements considered in the predefined scales remain unspecified. Therefore, assessment methods fail to fulfil the set-theoretical requirement of scientific quantification, in contrast to observational methods (Uher, submitted).

Fulfilling the algebraic requirement for scientific quantification

The algebraic requirement of scientific quantifications states that researchers must directly compare the elements specified with designated standards of measurement to express their ratio as a real number. Most widely used in science are spatial standard units of measurement, such as of distance (e.g., meter). But such comparisons are complicated in immaterial phenomena that are momentary and fluctuating and in which identifiable events vary in their spatio-temporal extension (e.g., behaviours, neuronal action potentials).

Specifically, in behavioural research, direct comparisons with standards of measurement are possible only for some kinds of properties and often only under experimental conditions (e.g., measuring physical force with a dynamometer). But in most observational settings in either research or everyday life, properties of behavioural events can seldom be directly compared with designated standards of measurement in the moments in which the events occur (e.g., for measuring spatial distances between individuals that are socially and physically interacting in a group).

Specifying observable quantitative properties of the events under study

The possibilities to *directly* quantify occurrences of behavioural events are rather limited because, in nunc-ipsum methods, observers must decide quickly on how to demarcate, categorise and encode events in the continuous and dynamic flow of behaviours. Mostly, observers record only the occurrences and non-occurrences of specified events, thus using binary scales. But in some behavioural events, quantitative properties can be directly identified, such as by defining spatial distances between individuals on the basis of their bodily properties (e.g., body contact, arm-length distance) or by counting the occurrences of distinct events. To enable observers to systematically identify quantitative properties and to encode them in standardised ways in the data, researchers must specify the particular properties that are considered to be quantitative in the particular phenomena under study.

Quantitative properties are divisible properties of events that are categorised as being of the same kind, thus of events that are considered to be qualitatively similar (Hartmann, 1964; Uher, submitted). The TPS-Paradigm emphasises that these properties must be directly perceptible and intersubjectively conceivable by multiple human observers. Researchers must also specify rules for the determinative assignment of these properties to numerals so that observers can encode their occurrences always in exactly the same way. For example in the great ape species, scratching usually occurs at a speed that enables human observers to perceive and count the single flexions of their fingers while they occur (Uher, Asendorpf & Call, 2008). In capuchin monkeys, by contrast, scratching occurs so quickly that human observers often cannot perceive each single finger flexion during live observations; this is only possible from video records using slow motion playback (Uher et al., 2013a). Therefore, capuchin researchers commonly record occurrences of *episodes* of scratching that are defined by several consecutive finger flexions and that are demarcated from one other by a short break (Fragaszy, Visalberghi, & Fedigan, 2004).

The observers' possibilities to directly quantify behavioural events are generally limited to behaviours of very short temporal extension (called point or frequency behaviours) that occur in temporal and spatial proximity. Occurrences of behaviours with longer temporal extension (called duration behaviours) are often recorded in terms of their onsets and terminations that are encoded on binary scales in the raw data. This requires a continuous recording of events while observers focus on only one individual and his or her behaviours and interactions. Given the enormous effort involved such focal individual sampling continuous recording methods, duration behaviours are also often estimated from scan sampling instantaneous recording techniques, in which observers record whether at specified points of time defined kinds of duration behaviours do or do not occur, respectively. These time point samples allow for generating only crude quantitative estimates of behavioural durations, but they enable the collection of data on many individuals during the same observation period (as done in the primate studies reported below; Altmann, 1974; Lehner, 1998; Martin & Bateson, 2007).

From the data log files thus-produced, researchers can generate *post-hoc* (i.e., after the collection of raw data is completed) quantitative measures by re-coding, transforming and further abstracting the raw data (e.g., by computing durations and latencies from the onsets and terminations recorded using behavioural coding software). From log files of observational data, scientific quantifications can be generated with the concept of time-relative probabilities.

Scientific quantification using the concept of time-relative probabilities

The TPS-Paradigm introduces the concept of *time-relative probabilities* to meet the algebraic requirement of scientific quantification for the momentary

and fluctuating phenomena explored in individuals that often feature spatio-temporal units of variable extension, thus the phenomena of physiology and behaviour. It can also be applied to explore psychical phenomena (Uher, 2016). This concept relies not on space but on *time* as the physical standard units of measurement to which the occurrences of the defined events can be compared for generating scientific quantifications (Uher, 2013, 2015a).

Time-relative probabilities fundamentally differ from other concepts of probability because researchers of individuals, unlike physicists, cannot deal with random, equiprobable and mutually disjunctive events that are independent from one another and that could be experimentally repeated as in Laplacian theory. Rather, the units that can be identified in the phenomena explored in individuals often vary in their spatio-temporal extension so that demarcations and categorisations of events can be made in various ways. Moreover, different kinds of events may co-occur and even depend upon one another, in particular if they are functionally related (e.g., slapping others requires physical proximity, but both behavioural events do not always co-occur). Because of these peculiarities of living organisms, previous concepts of probability are not applicable (Uher, 2013, submitted).

To generate time-relative probabilities, nunc-ipsum records of the occurrences of defined events are accumulated over repeated occasions to derive post-hoc quantitative measurements. As momentary phenomena often fluctuate, these measurements can be only *probabilistic*. Time-relative probabilities are then obtained by relating these probabilistic measurements to the periods of time during which they were recorded and that are also precisely specified in terms of elements of the set *T* studied. Therefore, unlike other types of probability, time-relative probabilities always have a unit (of time) and their values are not constraint to values between 0 and 1 (but they cannot become negative; for details, see Uher, submitted, 2013, pp. 6-11).

The scientific quantifications thus-obtained are *ratio-scaled* because the international time standards provide equal units (e.g., minutes, hours) and because the non-occurrence of the events studied in the particular set *T* of occasions and periods of time considered in a study defines an absolute point of zero. For example, the aggressive behaviours of a particular child in his or her peer group during free play as measured in 10 hours of observation during 5 consecutive days could be scientifically quantified with an average of 40 seconds of expressing verbal insults to peers, no verbal insults to adults and two slaps to peers per hour. The measurement units thus-created (e.g., overall duration in seconds per hour, frequency per hour) can be understood in the same way all around the globe. Because data reflecting time-relative probabilities are ratio-scaled, they permit applications of all arithmetic operations and they can be directly compared across studies and disciplines (Uher, 2013, submitted). The concept of time-relative probabilities is therefore essential for comparing variations within and across individuals, groups and species.

Comparing variations within and between individuals, groups and species

The concept of time-relative probabilities enables scientifically quantified comparisons not only of *averages* but also of *variabilities* and *ranges* in the occurrences of events across time (e.g., day-to-day variabilities, longitudinal changes) within and between individuals, groups and species (Uher, 2015a,b,c). Exploring such patterns enables systematic contrasts with assessments that people provide on individuals and on differences between individuals, groups and species. Such contrast can shed a new light on the human tendencies to overestimate the magnitude of minor variations between individuals for differentiating highly similar individuals and to homogenise within-group differences and to exaggerate between-group differences for differentiating individuals from different groups and communities.

The TPS-Paradigm provides comprehensive methodological frameworks for enabling such explorations. In these frameworks, concepts from "personality" psychology, differential psychology, cross-cultural psychology and cross-species comparative psychology are systematically integrated and extended by novel ones (Uher, 2008a,b, 2011a,b).

Important insights about patterns of behavioural variations can be gained by comparing different kinds of variations with one another, such as by relating variations *within* particular samples of individuals to differences *between* particular samples. Different kinds of *samples* can be defined that are nested within each other, ranging from single individuals over various kinds of groups of (conspecific) individuals (e.g., of the same age, sex/gender, socio-economic status, level of education, nation, religion or language) up to species and groups of species. The following outlines basic concepts, focussing on individuals, groups and species as samples (see Table 1).

Data reflecting time-relative probabilities can be explored for patterns of within-individual variability around within-individual averages[6], for example, an individual's activity levels can be explored for day-to-day variability (de Weerth, van Geert, & Hoitink, 1999; van Geert & van Dijk, 2002). This allows for quantifying an individual's range of behavioural possibilities in terms of his or her average, minimum and maximum probability to show particular kinds of behaviours in particular kinds of situations (cf. the related concepts of individual density distributions of act frequencies or daily activity states; Buss & Craik, 1985, Fleeson, 2001). Studying within-individual variability is of particular importance for exploring the dynamics and processes of individual change and development (Thelen & Smith, 1993).

6 Such approaches are often called ipsative-normative (Lazarus, 2000), sometimes also ideographic-nomothetic (Conner, Barrett, Tugade, & Tennen, 2007; Shoda, Mischel, & Wright, 1994); but because the connotations of these terms vary they are not used in the TPS-Paradigm.

Patterns of within-individual variability can be explored for commonalities and differences between individuals, such as between-individual differences in the day-to-day variability of individuals' activity (Uher, 2013). Patterns of within-individual variability can also be explored between individuals of different age groups, such as variability in daily activity patterns between individuals of different groups (Table 1; for an overview of methods of analyses of variability, see van Geert & van Dijk, 2002).

Furthermore, individuals can be compared with one another in their averages. Differences between individuals' averages can be compared between different groups, such as between-individual differences in males with those in females. Averages of particular groups of individuals, in turn, can be explored for differences between groups, such as sex/ gender differences (e.g., Mehl, Vazire, Ramírez-Esparza, et al., 2007; Table 1). Between-group differences can be set in relation to differences between higher-order groups, such as by comparing sex/gender differences between different nations, amongst others, as studied in cross-cultural differential psychology (Heine et al., 2002; Leung & Bond, 1989; van de Vijver & Poortinga, 2002).

Research on nonhuman species opens up many further opportunities for comparing behavioural variations. For example, patterns of within-individual variability can be compared between species, such as within-individual variability in daily activity patterns between cats and dogs (e.g., Piccione, Marafioti, Giannetto, et al., 2014). Species can also be compared with regard to their between-individual differences or the between-individual differences of particular groups (e.g., sex differences in between-individual variations between three macaque species; Sussman, Ha, Bentson & Crockett, 2013, see also below) as is done in cross-species comparative differential and "personality" psychology (Uher, 2008a,b). Biologists frequently compare the averages of particular groups between species, such as average age, sex/gender and social status differences between chimpanzees and bonobos (e.g., Boesch, Hohmann, & Marchant, 2000; Table 1).

Table 1 *Basic concepts for comparing variations within and between individuals, groups and species*

Differences between	Variations within		
	Individuals	**Groups**	**Species**
Individuals	Between-individual differences in within-individual variations (e.g., day-to-day variations around each individual's average activity level; Figure 1)	/	
Groups	Between-group differences in within-individual variations (e.g., individual differences in males compared with those in females)	Between-group differences in within-group variations (e.g., patterns of day-to-day variability compared between individuals of different age groups)	/
Species	Between-species differences in within-individual variations (e.g., individual day-to-day variabilities in weeper capuchins compared to those in rhesus macaques; Table 2)	Between-species differences in within-group variations (e.g., individual differences in female bonobos compared with those in female chimpanzees)	Between-species differences in within-species variations (e.g., individual differences in gender differences or average sex/gender differences in mandrills compared with those in toque macaques; Table 6)

Note. Patterns of behavioural variations can be explored by comparing different kinds of variations within and across different sample levels that are nested within each other, such as by relating *within-sample variations* to *between-sample differences*. The table specifies such comparisons, focussing on single individuals, groups of (conspecific) individuals and species as samples.

Applications of such comparative analyses can be demonstrated exemplarily with data from four nonhuman primate species. Individual behaviours were observed in 9 weeper capuchin monkeys (also called wedge-capped capuchins, *Cebus olivaceus*), 8 mandrills (*Mandrillus sphinx*) and 12 toque macaques (*Macaca sinica*) housed at the Berlin Zoological Garden and in 7 rhesus macaques (*Macaca mulatta*) housed at the Berlin animal shelter (Germany). The individuals[7] of each species lived together in one social group. Capuchins are

endemic to South-America, mandrills to Africa and macaques to Asia. All four species live in multi-male multi-female groups but their behavioural and social systems differ (e.g., the dominance behaviours of mandrills and rhesus macaques are described as "despotic", whereas those of capuchins and toque macaques are described as "egalitarian" (Fragaszy et al., 2004; Thierry, 2007; Otovic, Partan, Bryant, & Hutchinson, 2014).

Analyses of between-individual differences in within-individual variations and averages can be illustrated by observational data on the behavioural activities of the 7 Rhesus macaques. Three behaviours representing different degrees of physical activity were differentiated on the basis of observable properties and were encoded in three mutually exclusive categories, thus on a nominal scale. Resting and not moving was encoded as "sit", moving on the spot but not changing the location was encoded as "move" and moving about, thus changing location as "run". Using nunc-ipsum methods and scan sampling instantaneous recording, the observers recorded (amongst others) the activity state of each individual in the group in 20-minute intervals starting from the end of the last scan. On average, data were collected on 14.8 scan sample occasions per day (*Mdn* = 14; range 12-16) from 10 a.m. to 5 p.m.; observations occurred on 12 days within a 4-week period (3 days per week).
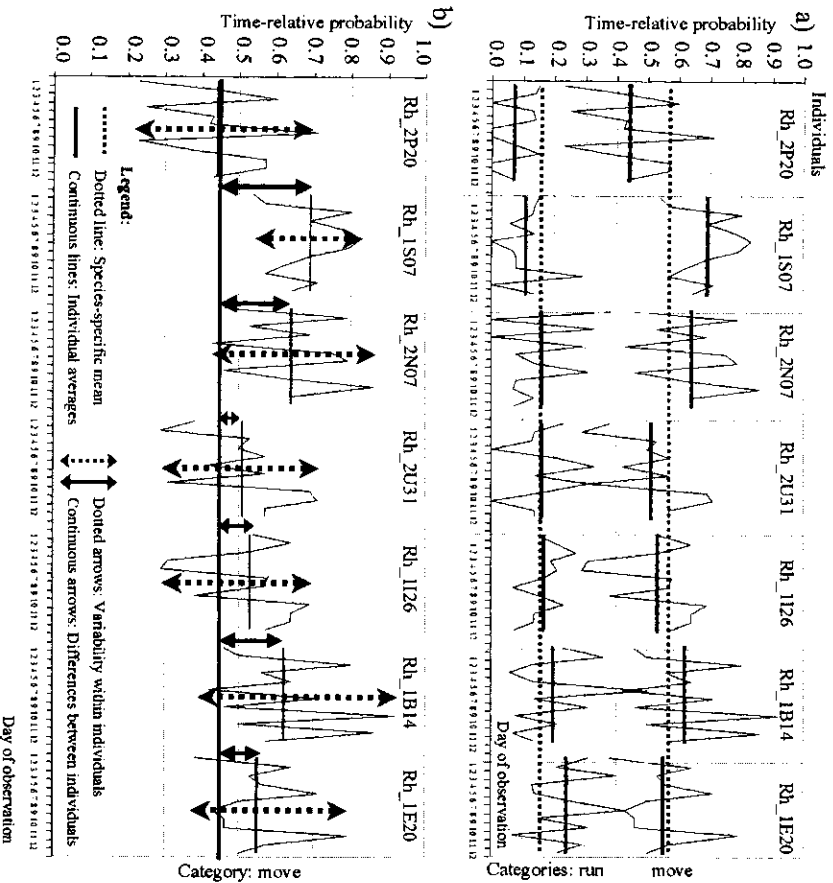
These binary encoded raw data were post-hoc accumulated over all scans within each day and then divided by the number of scans collected on each given day. The ratio-scaled data thus-generated for each day reflect each individual's time-relative probabilities for showing the three activity behaviours at a randomly selected moment during the times of the day studied. Hence, these data allow for estimating each individual's daily time-budgets with regard to his or her physical activity. Figure 1a depicts the individuals' daily time-relative probabilities for showing activity states categorised as "move" (upper graphs) and as "run" (lower graphs); thus, the difference between each individual's daily sum of these scores to 1.0 represents his or her daily time-relative probability for resting ("sit"; not depicted).

The activity levels of the 7 Rhesus macaques varied considerably within each individual from day-to-day over the 12 days of observation (Figure 1a). The patterns of within-individual variability and individuals' averages over these 12 days (shown as continuous horizontal lines across each individual's graph) differed between individuals. The time-relative probabilities for moving about of the first individual, the male Rh_2P20, were similar to those of the second individual, the female Rh_1S07. But his average time-relative probabilities for moving on the spot were much lower than those of that female. In fact, his average was lower than the average of all Rhesus individuals studied (shown as dotted line across the graphs of all individuals), whereas her average was higher than the Rhesus average. But his within-individual variability in moving on the spot, thus

---

[7] The sample consisted 8 mandrills (4 males, 4 females), 12 toque macaques (5 males, 7 females), 9 weeper capuchins (3 males, 6 females) and 7 rhesus macaques (3 males, 4 females). It comprised individuals of all age groups: in the weeper capuchins, mean age was 6 years (*Mdn* = 3 years; range 1-21 years); in the mandrills, mean age was 9.75 years (*Mdn* = 8 years; range 1-26 years); in the toque macaques, mean age was 4.69 years (*Mdn* = 4 years; range 0.25-13 years) and in the rhesus macaques, mean age was 17.68 years (*Mdn* = 20 years; range 7-31 years).

his range of behavioural probabilities for showing moderate activity levels was much higher than hers.

everyday life are much less pronounced than those that can be extracted statistically from data or abstracted mentally from past observations.

The within-individual day-to-day variability was generally much higher than the average between-individual variability as highlighted by the dotted and continuous vertical arrows in Figure 1b for the individual variations in moving on the spot. These findings are in accordance with robust findings from cross-cultural psychology indicating that "within-unit variance overwhelms between-unit variance whatever the unit—individual, family, or culture" (Shweder & Sullivan, 1990, p. 412). Statistically, these differences in magnitude between within-sample variability and between-sample variability can be explained by the higher reliability of the aggregated measurements generated at higher sample levels. But importantly, this broad range of individuals' behavioural possibilities[8] illustrates that individual differences in behavioural patterns—as they occur in everyday life—are much less pronounced than those that can be extracted statistically from data or that can be abstracted mentally from observations made in the past.

Comparative analyses using nunc-ipsum recorded behavioural data therefore enable insightful contrasts to be applied with findings obtained from assessment studies, which are inherently based on strong aggregation (see below). Such contrast allow for exploring the human abilities to quickly form impressions of individuals and to develop for this purpose socially shared socio-cognitive categories in which observable variations are represented in either exaggerated or oversimplified ways. Given the enormous variability that can occur within and between individuals, such analyses presuppose metatheoretical and analytical concepts that enable scientists to disentangle patterns that are specific to particular individuals from those that are common to all of them (Uher, 2011a,b, 2015a).

Disentangling the specific from the common

Comparing individual variations requires the identification of what is specific to a particular individual, group or species and what is common to all individuals of the particular sample considered. Given that in research on individuals the basic ontological entity is the individual, the concept of individual-specificity is of central importance.

Individual-specificity: Metatheoretical concepts and methodological implications

In momentary and fluctuating phenomena like behaviours in which within-individual variations are often pronounced, patterns in the occurrences of

---

Figure 1 *Between-individual differences in within-individual variabilities and averages in the physical activity of seven rhesus macaques*

a)
Individuals: Rh_2P20 Rh_1S07 Rh_2N07 Rh_2U31 Rh_1I26 Rh_1B14 Rh_1E20
Time-relative probability: 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
Categories: run move

b)
Rh_2P20 Rh_1S07 Rh_2N07 Rh_2U31 Rh_1I26 Rh_1B14 Rh_1E20
Time-relative probability: 1.0 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0.0
Category: move
Day of observation

Legend:
.... Dotted line: Species-specific mean
——— Continuous lines: Individual averages
<····> Dotted arrows: Variability within individuals
<———> Continuous arrows: Differences between individuals
Day of observation

Note. The estimated time-relative probabilities of 7 rhesus macaque individuals for showing activity states categorised as "move" (i.e., moving on the spot; upper graphs) and "run" (i.e., moving about and changing location; lower graphs); the difference between each individual's daily sum of these scores to 1.0 thus indicates his or her time-relative probability for resting ("sit"; not depicted). The graphs show each individual's daily averages of his or her time-relative probabilities across 12 days of observation carried out within a 4-week period. a) Individuals' averages over these 12 days are shown as continuous horizontal lines across each individual's graph; the averages of all 7 rhesus individuals over the 12 days are shown as dotted lines across the graphs of all individuals. b) The day-to-day variability within individuals was much higher than the average between-individual variability as highlighted by the dotted and continuous vertical arrows for the individual variations in moving. Statistically seen, this is due to the higher reliability of the aggregated measurements generated at higher sample levels. But this broad range of individuals' behavioural differences shows that individual differences in behavioural patterns as they occur in

---

[8] The daily time-relative probability scores for physical activity shown in Figure 1 were also derived from aggregation over all on average 14.8 data points collected on each given day, which in turn are only time samples from the whole range of behavioural variability that occurred in these individuals.

events that can be empirically obtained for individuals can only be *probabilistic*. To be specific to an individual, these probabilistic patterns must differ between individuals, thus be *differential*. To make sure that these probabilistic patterns between differences reflect not just random variations derived from the dynamic fluctuations of behaviours, these patterns must also be stable for at least some amount of time, thus be *temporally stable* (Uher, 2011a).

Hence, individual-specificity is an abstract theoretical construct denoting the coincidence of differential and temporal patterns and, in momentary and fluctuating phenomena, also of probabilistic patterns in the occurrences of events. It was shown that the idea of individual-specificity is elementary to all definitions of "personality" and that definitions differ in the particular phenomena considered, which depends on the particular philosophical presuppositions made by the authors (Uher, 2013). For enabling transdisciplinary research, the TPS-Paradigm therefore conceives of "personality" as individual-specificity in *all* of the various kinds of phenomena explored in individuals (Uher, 2015a,b,c).

Importantly, neither differential, nor temporal nor probabilistic patterns can be directly observed in any given moment. Instead, to identify individual-specificity, (non-)occurrences of defined behavioural events must be recorded using nunc-ipsum methods and repeatedly across many occasions. The nominal-scaled behavioural raw data must be accumulated over time within each individual to generate post-hoc ratio-scaled scientific quantifications of *individual patterns* using the concept of time-relative probabilities. These individual patterns must then be compared between individuals to identify *differential patterns*.

But direct comparisons of scores within and between individuals are hindered because, in any given sample, different kinds of behaviours generally occur with different frequencies (cf. the differences in the average occurrences of behavioural activities coded as "moving" and "running" in Rhesus macaques; see the dotted horizontal lines in Figure 1). Even the same kind of behaviour commonly occurs with different frequencies across different kinds of situations. Differential analyses therefore require that the individuals' time-relative probabilities for showing a particular behaviour are statistically standardised across all individuals within each kind of situation studied and within the time periods for which the scores are determined (Uher, 2011a, 2013). Time-relative probability scores can be differentially standardised, for example, by using z-standardisation in which the deviation of individual scores from the mean score of all studied individuals is divided by their standard deviation (same time units provided).

Differential standardisation transforms scientific quantifications of *absolute* individual scores into scores reflecting *relative* between-individual differences. Therefore, differentially standardised time-relative probabilities have no time units. But as they were generated in specified steps of data processing rather than implicitly as is the case when people generate "personality" assessments, the absolute time-relative probabilities from which the differential time-relative probabilities were derived can always be traced when needed for interpretation,

analyses and comparisons with other sets of individuals (as demonstrated below). Differential standardisation also enables direct comparison of scores between variables with different time units (e.g., frequencies, durations) both within and between individuals. It also enables statistical aggregations of different variables into composite or summary variables as required for computing construct measures (see below).

To identify individual-specificity, the differential patterns thus-identified must be shown to be stable across time periods longer than those in which the measurements were first ascertained and in ways that are considered to be meaningful (e.g., defined by the strength of statistical correlations over specified time periods). Thus, temporal patterns are to be explored in patterns that are in and of themselves defined by a certain temporal stability. To consider the effects of aggregation on the reliability of measurements, comparisons between individuals presuppose that the number of occasions and the lengths of the time periods considered for estimating individuals' time-relative probabilities are comparable and at best identical between the individuals being compared (Uher, 2013; Uher et al., 2013a).

Such analyses of individual-specificity can be illustrated by observational data on the rhesus macaques studied at the Berlin animal shelter and the weeper capuchins, mandrills and toque macaques studied at the Berlin Zoological Garden. Three constructs of individual-specificity describing social behaviours were studied. *Social orientation* was operationalised with grooming others and three proximity measures indicating different distances to the target individuals (i.e., body contact, arm length distance and 1m distance). *Aggressiveness* was operationalised with behaviours of contact aggression and non-contact aggression. *Dominance* was operationalised with dominant behaviours (i.e., displacing others) and behaviours of subordination (i.e., be displaced by others, subordinate to others) as indicators of low dominance.

The specific perceptible forms of these behaviours slightly varied between these four species also because they differ morphologically from one another to some extent. For example, mandrills are much larger than capuchins and the macaques. Mandrills have one of the greatest sexual dimorphisms of all primates, whereas in the other three species studied, the differences between males' and females' body sizes are still pronounced but much less than in the mandrills; Fragaszy et al., 2004; Thierry, 2007; Otovic et al., 2014). Therefore, it was the observers' task to abstract from this perceptible variability of these species' behaviours and to demarcate and categorise events that can be considered function-ally comparable across the species studied. The inter-observer reliability in encoding the occurrences of these 9 different behaviours in the different species was high[9], demonstrating intersubjective agreement between two persons in the encoding of behavioural occurrences in the raw data.

[9] Inter-observer reliability in the binary scaled encoding of the (non-)occurrences of all 9 behaviours at specified times was studied across all species with Cohen's κ, which was also used as a very

In each of the three species observed at Berlin Zoo, the duration behaviours grooming and proximity were recorded using scan sampling instantaneous recording every 15 minutes starting from the last scan. The frequency behaviours of aggression and dominance generally occur less frequently and were therefore recorded using event sampling event recording, in which the observers recorded these events whenever they occurred in the group during a 5-minute time sample directly after completing the scan sampling records in the given species. On average, data were collected on 21 scan sample occasions and 21 event sample occasions per day and species ($Mdn = 23$; range 11-26) between 9 a.m. and 5 p.m.

In the Rhesus macaques observed at Berlin animal shelter, the same behaviours were recorded using the same behavioural categories and encoding scheme. But a slightly different time sample scheme was used. Scan sampling occurred only every 20 min, enabling the recording of on average 13.8 scan sample occasions per day (see above). Event recording of aggressive and dominant behaviours occurred during 20 minute time samples on 11 occasions on average per day.

For enabling direct quantitative comparisons between all four species, individuals' time-relative probabilities were standardised in two different ways so that the data represent exactly the same time units and the same situational conditions. First, the data obtained from the different time sample schemes were made directly comparable with one another by converting them into the same basic time samples. Thus in both studies, the binary raw data of the scan sample records were accumulated over all scan occasions within each day and then divided by the number of scans collected on each given day. The same was done for the binary raw data of the event samples in the zoo study. In the animal shelter study, the event sample data were accumulated over all 20-minute event samples collected within each day and then standardised to a time sample of 5 minutes, which corresponds to the basic time sample used in the zoo study. Otherwise, the frequencies of the event behaviours (i.e., aggression and dominance) in the Rhesus macaques would be significantly overestimated.

The ratio-scaled data thus-generated for each day reflect each individual's absolute time-relative probabilities for grooming when observed at a randomly selected moment and for showing each of the aggression and dominance behaviours during a randomly selected time sample of 5 minutes (during the times of the day studied). The proximity data reflect the average number of conspecifics in different distances to the target individual when observed at a randomly selected moment. Hence, the data thus-obtained are standardised with regard to the time units that they reflect. For example, an individual may have an absolute time-

strict measure for testing agreement between two observers in encoding the number of conspecifics located at different distances to the target individual. Inter-observer reliability computed over 621 single observations was high ($\kappa_m = .87$, ranging from $\kappa = .74$ to .99). For comparison, mean Pearson-correlations were $r_m = .95$** (computed using r-to-Z transformation), ranging from $r = .81$ to .99.

relative probability of 0.18 (i.e., an 18% probability) for showing contact aggression within 5 minutes of observation and it may be in arm-length distance to on average 0.23 conspecifics when observed at a randomly selected moment.

Second, for enabling direct quantitative comparisons of social interaction behaviours between individuals from social groups of different sizes, individuals' absolute time-relative probabilities were also standardised to a hypothetical number of 10 potential interaction partners within each group. The data thus-standardised reflect the average number of conspecifics in proximity and the average time-relative probabilities for showing the particular behaviours studied if 10 potential partners were available for each individual studied.

Observations in the zoo were carried out on 15 days, 5 days per week each in the first two weeks (t_1 and t_2) and in the last week (t_3) of a 5-week period. The 12 days of observation in the animal shelter were carried out within a 4-week period, thus in a time-span comparable to the time period covered by t_2 and t_3 of the zoo study. For temporal analyses, the data from the animal shelter study were therefore divided into two 2-week periods of 6 days each and assigned to the time periods t_2 and t_3 of the zoo study. This schedule enables temporal analyses over a 2-week period (t_1-t_2; in the zoo study), a 4-week period (t_2-t_3) and a 5-week period (t_1-t_3; in the zoo study). Overall, each individual of the three species studied at the zoo was observed for 80 hours, 26.5 hours on average per study period; each Rhesus individual was observed for 60 hours in total and 30 hours on average per study period. Hence, despite the different time sample schemes used in the zoo study period and in the animal shelter, the time periods used for estimating individuals' time-relative probabilities are comparable, thus enabling direct comparisons of the data between all four species.

To explore whether the individuals' absolute time-relative probabilities obtained from this observation scheme reflect individual-specific patterns, the between-individual variations of each behaviour category were analysed for temporal stability between the three study periods. This was done for each species separately by computing (Pearson) test-retest correlations of the rank-order of individuals' scores for each behavioural variable (i.e., using variable-oriented analyses; Bergmann & Trost, 2006; Uher, 2011a). Rank-order stability was explored in both individuals' averages (using individuals' means within each study period) and their within-individual variations (using the corresponding standard deviations). For an exemplary demonstration, some categories describing functionally related behaviours were summarised into more abstract categories; therefore, the individuals' scores on some more specific behavioural variables were added into scores on summary variables (i.e., the three different proximity measures, the two aggression measures and the two subordination measures; see Table 2). These *summary scores* still reflect *absolute* time relative probabilities (e.g., for showing any kind of aggressive behaviours in a randomly selected 5-minute time sample), which helps to demonstrate basic principles on which the subsequent steps of analyses are based.

In all behavioural (summary) scores, between-individual variations in individuals' *averages* were test-retest reliable over most study periods and in most species (except for the Subordination score SU in the toque macaques over 4-week and 5-week periods and in the weeper capuchins; see Table 2). These findings provide evidence for individual-specific patterns (considering the limited power of these analyses due to the small sample sizes). By contrast, between-individual variations in within-individual *variabilities* tended to be less test-retest reliable. In some behavioural scores, the test-retest reliability was low over some study periods and in some species; in the Proximity summary scores (PX) it was low over all study periods and in all species. This indicates that within-individual variability was pronounced and sometimes not differentially stable, in and of itself, despite the temporally stable and thus individual-specific patterns shown in within-individual averages (Table 2).

Table 2 *Temporal stability of between-individual differences in within-individual averages and variabilities over 2, 4 and 5 weeks: Evidence for individual-specific patterns within each species*

Behaviours xx Species	Within-individual averages (m)			Within-individual variabilities (sd)		
	2 weeks t_1-t_2	4 weeks t_2-t_3	5 weeks t_1-t_3	2 weeks t_1-t_2	4 weeks t_2-t_3	5 weeks t_1-t_3
	r (p)	r (p)	r (p)	r (p)	r (p)	r (p)
GR Grooming score						
Weeper capuchins	.72* (.030)	.96** (.000)	.71* (.034)	.60 (.090)	.87** (.002)	.74* (.023)
Mandrills	.70 (.056)	.91** (.002)	.72* (.045)	.63 (.096)	.74* (.036)	.90** (.002)
Toque macaques	.92** (.000)	.95** (.000)	.94** (.000)	.28 (.374)	.54 (.070)	.43 (.162)
Rhesus macaques	-	.81** (.001)	-	-	.66 (.089)	-
PX Proximity summary score						
Weeper capuchins	.81** (.009)	.94** (.000)	.78* (.012)	-.26 (.492)	.17 (.659)	.46 (.212)
Mandrills	.89** (.003)	.67 (.068)	.52 (.184)	-.19 (.647)	-.27 (.512)	-.25 (.543)
Toque macaques	.92** (.000)	.95** (.000)	.94** (.000)	.28 (.374)	.38 (.217)	.08 (.815)
Rhesus macaques	-	.64 (.125)	-	-	-.17 (.719)	-
AG Aggression summary score						
Weeper capuchins	.93** (.000)	.92** (.000)	.88** (.002)	.81** (.008)	.71* (.034)	.65 (.058)
Mandrills	.68 (.064)	.73* (.039)	.97** (.000)	.71* (.048)	.69 (.056)	.96** (.000)
Toque macaques	.72** (.008)	.78** (.003)	.54 (.071)	.27 (.387)	.59* (.044)	.30 (.351)
Rhesus macaques	-	.71 (.077)	-	-	-.08 (.863)	-
DI Displacement score						
Weeper capuchins	.98** (.000)	.96** (.000)	.99** (.000)	.94** (.000)	.92** (.000)	.99** (.000)
Mandrills	.92** (.001)	.85** (.008)	.93** (.001)	.50 (.204)	.11 (.787)	.86** (.007)
Toque macaques	.92** (.000)	.85** (.001)	.95** (.000)	.59* (.045)	.35 (.267)	.76** (.004)
Rhesus macaques	-	.78 (.038)	-	-	.38 (.399)	-
SU Subordination summary score						
Weeper capuchins	.31 (.421)	.51 (.160)	-.23 (.562)	.45 (.219)	.63 (.068)	-.14 (.718)
Mandrills	.89** (.003)	.82* (.012)	.82* (.013)	.89** (.002)	.86** (.006)	.83* (.010)
Toque macaques	.80** (.002)	.44 (.149)	.08 (.811)	.07 (.830)	.65* (.022)	-.01 (.968)
Rhesus macaques	-	.89** (.007)	-	-	.73 (.065)	-

Note. * $p < .05$; ** $p < .01$. Variable-oriented analyses of the test-retest reliability of between-individual differences within each species (weeper capuchins, $N = 9$; mandrills, $N = 8$; toque macaques, $N = 12$; rhesus macaques, $N = 7$) using Pearson correlation r. The temporal stabilities of individuals' averages were calculated using the mean scores of individuals' absolute time-relative probabilities for showing a given behaviour and the mean scores of the number of conspecifics in proximity, respectively. The temporal stabilities of individuals' corresponding patterns of day-to-day variability were computed using the standard deviations. The summary scores represent more abstract categories of encoding and were derived by summarising the scores of more specific behavioural variables: The Proximity summary score summarises the three proximity measures, the Aggression summary score summarises the categories contact and non-contact aggression, and the Subordination summary score summarises the categories being displaced by others and subordinating to others. The Grooming summary score summarises the categories being displaced by others and subordinating to others. The Grooming

Exploring individual variations from three interrelated viewpoints: Specificity/ uniqueness, comparability and universality/ typicality

The concept of individual-specificity highlights that exploring individual variations involves three inter-related viewpoints—*specificity (i.e., uniqueness)*, *comparability* and *universality (i.e., typicality)*. It shows that *specificity* can be determined and quantified only in relation to other individuals of the same kind. This implies that individuals can be compared with one another, thus *comparability*. The determination of both specificity and comparability depends on the particular set *I* of individuals considered (i.e., the reference population), which, in turn, determines the specifications of *universality* (Uher, 2008a,b).

In research on individual-specificity ("personality"), universality always refers to *populations* of individuals (e.g., groups or species) but not to single individuals because patterns that all individuals of particular a group or species show in the same way cannot be individual-specific. Individual patterns that are universal for all individuals of particular groups or species are explored in other fields of research (e.g., general psychology, biology, anthropology).

Importantly, the highest possible level for determining individual-specificity is the species level because species are considered the smallest biological populations that differ distinctively and not just continuously from another due to intrinsic barriers that prevent interbreeding (Campbell & Reece, 2005; Uher, 2008a,b). Thus, same-species individuals share essential organismal properties and are therefore similar to considerable extent. It is the variations occurring within this essential similarity of same-species individuals that are explored in research on individual-specificity (e.g., differential psychology, "personality" research). In the analyses presented above for illustrating these principles, the temporal patterns of individual differences were therefore analysed separately for each of the four primate species studied (Table 2).

The concepts of specificity, comparability and universality can be applied for explorations of individual variations across all sample levels. For example, if particular kinds of between-individual or between-group variations in humans are found only in a particular cultural community, then they are culture-specific. If all cultural communities exhibit these variations in similar ways, then these variations are universal to the human species (Kluckhohn, Murray, & Schneider, 1953). If particular kinds of variations are shown only by humans but not by any other species, then these variations are human-specific (or uniquely human). If

several species show comparable kinds of between-individual or of between-group variations, then they can be considered universal for this particular group of species, for example, variations may be universal primate, universally mammalian or perhaps even universally vertebrate (Uher, 2008b). If the focus lies on comparisons between groups of species that do not exhibit a particular kind of variation, this variation can be, for example, primate-specific, mammal-specific and so on.

The pivotal point of such explorations is the concept of comparability being applied and the ways in which it is methodologically implemented in a given study.

Principles for establishing comparability

In a strict sense, no individual of whatever group or species is exactly like another, even if the differences are only small. Between-individual variability is pervasive—and it is essential for evolution (Darwin, 1859). Individuals from different groups and, to a much greater extent, individuals from different species often differ from one another in significant ways, such as in their morphology, behaviours and the behavioural situations they encounter in their lives as well as in the functions and meanings that these phenomena may have for these individuals (e.g., head-shaking and nodding signalling agreement or non-agreement; Uher, 2008a,b, 2013). Specifying the perceptible properties on which the demarcations of events can be based makes painfully clear that establishing comparability is *always* a matter of decision and abstraction on the parts of human individuals, whether explicitly or implicitly.

The more diverse the perceptible properties of the phenomena under study are, the more abstract the levels of abstraction need to be in order to establish comparability. Phrased differently, the more abstract the level of consideration, the more comparable will the properties considered appear to be. The particular kinds of abstractions on which observations of individual behaviour are based differ from those required in assessments.

Essential differences between observations and assessments

Observers are confronted with the heterogeneous properties of the fluctuating and dynamic phenomena of behaviour that can be directly perceived in the particular situations and time periods of observation. Observers must accomplish the difficult task to demarcate and abstract from this diversity particular kinds of events that can be considered to be identically repeatable at least to some

extent. Thus, observers must establish *comparability between particular percep-tible properties* to identify commonalities and differences between them in order to demarcate, abstract and categorise behavioural events. Observers must do this in systematic and standardised ways and in the given moments in which the events occur. To achieve this, observers are instructed and trained to focus their perception on particular properties specified in the encoding schemes, which are specifically developed for this purpose and tested for their viability using scientific principles.

Assessments also rely on people's perceptions of the complex properties of behaviours—yet on perceptions that they have made in the past without having been instructed or trained to focus on particular properties and to demarcate and categorise events in particular ways. In their everyday lives, people rely on intuitive abstractions often using the socio-cognitive categories that are encoded in everyday language. Assessment methods capitalise on this everyday knowledge by wording the statements to be assessed in everyday language and by using abstract and global everyday terms that are applicable to diverse kinds of properties and phenomena yet without specifying any particular ones.

But unlike scientific categories, everyday categories are not explicitly defined, well structured and disjunctive; rather, they are often fuzzy and context-sensitive, thus allowing for flexible demarcations (Hammersley, 2013). The words denoting everyday categories are used for making perceivably different properties subjectively comparable. Moreover, words are often erroneously assumed to directly reflect the phenomena that they denote (called the *fallacy of misplaced concreteness*; Bentham in Ogden, 1932; Whitehead, 1929). But this is possible only for words denoting concrete phenomena that can be directly perceived but not for words denoting abstract ideas that can be derived only by mental *construction*—such as words that denote individual-specific behaviours or "personality" (Uher, 2015b,c).

As a result, the particular elements of the sets *B* of behaviours, *S* of situations, *T* of time and *I* of individuals to which particular everyday words and statements may refer for particular individuals in particular situations (e.g., when generating assessments) will necessarily vary. In fact, it was already shown that the meanings that raters construct for the standardised item statements of established assessment instruments often vary considerably both within and between raters (Arro, 2013; Rosenbaum & Valsiner, 2011). This variability in people's interpretations of descriptors and statements about individual behaviours may also derive from the fact that everyday categories reflect not only abstractions from perceptible properties but also abstractions and categorisations that are made on the basis of implicit assumptions about the functions, meanings and values that particular behavioural events may have for particular individuals, thus on assumptions about their *functional comparability*.

In contrast to the forms of particular events and phenomena, the particular functions and meanings that they may have for particular individuals depend on

256

complex constellations and can seldom be directly perceived or straightforwardly inferred. Assumptions about functions and meanings of behaviours therefore vary between groups of individuals, such as between different sociocultural and language communities (e.g., the particular behaviours considered "insulting" and "aggressive"). These different assumptions are implicitly contained in the socio-cognitive categories and pertinent vocabularies that are established and used by human communities in their everyday lives. Therefore, the everyday words of different sociocultural communities may reflect different abstractions and categorisations of observable properties and phenomena.

Consequently, the apparently greater and more direct comparability of assessments of individual behaviours as compared to pertinent observational data solely derives from the much higher levels of abstraction on which everyday categories are based and should therefore not be mistaken as an indicator of higher data quality as this is the case in many fields of psychological and "personality" research (Uher, 2011b, 2015a,b,c). On the contrary, assessment methods fail to fulfil the requirements of scientific quantification because they specify neither the particular elements to be quantified nor the principles used for establishing comparability between particular elements nor the algorithms used for generating quantitative data on the empirical occurrences and interrelations of the elements considered (Uher, submitted). Comparability cannot be established by simply using identical words and statements in assessments because words are only signs that are used for encoding properties of the actual phenomena under study.

Establishing functional comparability of behavioural and situational categories

The TPS-Paradigm emphasises that comparability between kinds of events that differ in their perceptible properties can be established only on the basis of the known or assumed functions or species that they may have for the individuals of the different groups or species that are compared with one another (Uher, 2013; 2015a). Functional commonalities and dissimilarities between different kinds of events specify the *theoretical interrelations* of the studied elements within and between the sets *B*, *S*, *I* and *T*, which is prerequisite for empirical analyses (see below; Uher, submitted).

A systematic theoretical framework for establishing comparability in a behavioural study can be developed on the basis of a category system describing the known or assumed functions and meanings of the particular kinds of behaviours and behavioural situations studied on different levels of abstraction. In such frameworks, abstract functional concepts (e.g., survival) are broken down stepwise into more specific ones (e.g., finding food, finding shelter, defence against attacks from others; cf. Uher, 2008a). On lower levels of abstraction, the functional categories can be systematically linked with abstract categories of the per-

257

ceptible properties of the behaviours studied (e.g., different modes of communication) and of the different kinds of phenomena that are comprised by a given behavioural situation (e.g., physical properties of objects or behavioural properties of other individuals present in a situation).

For example, a theoretical framework of functional categories of social contact behaviours may comprise specific categories that may each describe functions for establishing, maintaining and ending proximity to others, for keeping and increasing distance to others, etc. Each of these functional categories, in turn, may comprise more specific categories comprising the particular vocal, facial, gestural or postural behaviours shown by the individuals of the particular groups or species under study.

Statistical analyses of configurational comparability: Construct equivalence and patterning effects

Theoretical frameworks specifying the known or assumed functional interrelations between the events studied are prerequisite for analysing the *empirical interrelations* of their occurrences, such as to explore the statistical comparability of construct measures. For systematic explorations, scientists develop an empirical framework comprising the particular measurement variables that are used in a given study to operationalise the constructs that are specified in the theoretical framework. Systematic linkages within and between the theoretical framework and the empirical frameworks are specified using the concept of *nomological networks* (cf. Cronbach & Meehl, 1955).

For the primate studies used for illustrating the basic concepts, a theoretical framework was established specifying the three behavioural constructs of individual-specificity studied (Social orientation, Aggressiveness and Dominance) and the theoretical interrelations of the behavioural categories that each construct comprises. An empirical framework was defined specifying for each of the behavioural categories specific behavioural variables (e.g., different proximity measures for physical proximity) that were used in the encoding schemes for generating the data. This nomological network comprising the theoretical and the empirical framework guided the computation of composite measures of both the constructs and the behavioural (summary) scores as well as the analyses of their configurational comparability between species as shown below.

Statistical analyses of the configurational comparability of construct measures between different groups or species can draw on analytical concepts from cross-cultural "personality" psychology (Uher, 2008b), such as on the statistical concepts and methods for exploring *structural or construct equivalence* (van de Vijver & Leung, 1997a,b; van de Vijver & Poortinga, 1997). Configurational comparability in the empirical interrelations of composite measures across

different constructs in terms of their latent structures (as described, e.g., in taxonomic "personality" models) can capitalise on the concepts and methods used for exploring *patterning effects* (Leung & Bond, 1989; Uher, 2008a,b).

Limitations and fallacies of assessment-based comparative analyses

The concepts and methods needed for such statistical analyses were primarily developed in assessment-based research where they are used to explore the empirical comparability between sets of data obtained for different groups of individuals (e.g., national, cultural or linguistic communities). These analyses are based on the widespread idea that comparability of measurements could be established by using the same assessment instruments or translated versions of it (van de Vijver & Tanzer, 2004)—an idea that cannot withstand critical analysis from a philosophy-of-science perspective. Assessment data are generated (mostly by lay people) on the basis of unspecified rules for demarcating, abstracting, converting and encoding information from the properties and phenomena to be assessed. Hence, it remains unspecified what particular properties and events the data thus-generated actually reflect and on what particular considerations the raters may have actually based their assumptions of comparability between observable properties (Uher, 2015b,c). Functionally comparable behaviours need not—and often do not—have spatio-temporal forms that are identical across the different samples studied.

Independent of this, construct measures are commonly derived by aggregating or adding together the raw scores across several measurement variables, thus across different scales. Although it is well known that ordinal-scaled units do not permit such arithmetic operations, these practices are widespread in psychology and the social sciences. Their application is commonly justified by assuming that inferences can be made from the statistical properties and structures of the empirical data (e.g., distribution patterns, latent factors) to the properties of the phenomena that these data are meant to represent. Therefore, metric properties that can be found in assessment data are widely assumed to constitute the ultimate evidence that the phenomena toward which they are targeted could have metric properties as well (cf. Michell, 2011, 2012). But importantly, such inferences are possible only if the ways in which the *data were generated* in the first place fulfil the requirements of scientific measurement (Uher, submitted).

The frequent finding that assessment data meet the assumptions of sophisticated statistical theories merely results from the rigorous selection of only those particular measurement variables that allow for generating sets of data that meet these particular measurement assumptions and from discarding all those variables that do not meet them. This procedure radically matches the data generation to statistical theories rather than to the properties of the actual phenomena under study that therefore cannot be explored in data thus-produced (Uher, 2015b,c). In addition

to these limitations derived from the ways in which assessment data are generated, further limitations and fallacies must be considered in the statistical analyses that are commonly used in assessment-based comparative research.

Limitations and fallacies of redundancy-based statistical analyses

The statistical analyses that are most frequently used for exploring configurational structures within and between construct measures (e.g., factor analysis, cluster analysis) rely on similarities and dissimilarities in the co-variations of scores of different measurement variables. Given that co-variation reflects redundancies (Uher, 2015b,c), this allows for the parsimonious description of the empirical variation contained in the data in aggregate scores on more abstract composite variables (Kaplan & Saccuzzo, 2008). It is widespread practice in many fields of research to reduce sets of variables solely on the basis of their empirical patterns of co-variation and to interpret the composite variables thus-created (e.g., factors, clusters) as reflecting meaningful phenomena. These practices are increasingly applied also to observational data, in particular in animal "personality" research.

But actually what does it mean if particular kinds of behaviours happen to occur in a given sample of individuals in similar ways over a specified period of time even if these behaviours are functionally unrelated or do not occur in spatio-temporal proximity? Researchers of individuals commonly interpret such kinds of overall empirical co-variation as reflecting "common underlying mechanisms" (cf. Allport, 1937; Cattell, 1946). But analyses of between-individual variations cannot yield results that are generalisable to the understanding and explanation of within-individual variations—much less of internal mechanisms or even their dynamics (Block, 2010; Cervone et al., 2001; Omi, 2012; Molenaar, 2004; Toomela, 2011; Uher, 2014, 2015b,c; van Geert & van Dijk, 2002). This type of fallacy, called ecological fallacy (Robinson, 1950), occurs in different forms in all fields of research (Pollet, Tybur, Frankenhuis, & Rickard, 2014; Subramanian, Jones, Kaddour, & Krieger, 2009; Uher, 2015c). It occurs, for example, when researchers erroneously project onto each other patterns of within-individual and between-individual variations (Moolenaar, 2004; Toomela, 2011), or patterns of between-individual and between-group variations (Thorndike, 1939) or patterns of between-individual and between-culture variations (Hofsteede, 2001). Analogous fallacies also occur in cross-species comparisons.

When behavioural events functionally depend on one another (e.g., in primates, grooming involves body contact but not every body contact is due to grooming), then their empirical occurrences necessarily co-vary on the sample level. For example, the occurrences of body contact and grooming were moderately correlated in the weeper capuchins ($r = .28, p < .05$), the mandrills ($r = .20; p < .10$), the toque macaques ($r = .47; p < .01$) and the rhesus macaques ($r = .21; p < .05$). But such overall co-variations may not be found if functionally similar

behaviours are assumed to differ in their functional intensity (e.g., if, in a particular situation, one act of contact aggression may fulfil the same function like three acts of non-contact aggression). In fact, empirical occurrences of functionally similar behaviours often show only low to moderate, sometimes even zero consistency across individuals (Uher, 2011a,b, 2015b; Uher et al., 2008, 2013a). Behavioural scientists studying humans have also reported repeatedly that the factors emerging from factor-analyses accounted for only low to moderate percentages of the variance in behavioural data (Blurton Jones, 1967, 1972; Smith, 1973; Smith & Connolly, 1972, 1980). The analyses revealed "intrinsic statistical inadequacies in the statistical treatment of consistency" in behavioural data (Allport & Vernon, 1933, p. 47).

The TPS-Paradigm therefore questions the utility of redundancy-based reduction principles for studies that are targeted towards exploring real-life issues (Uher, 2015b,c). Far more illuminative than explorations of redundancies between behavioural events are, for example, analyses of their spatio-temporal interrelations. This requires statistical context-based reduction principles that enable explorations of temporally and situationally contextualised behavioural data (amongst others; for details, see Uher, 2015b). Such data cannot be obtained with assessment methods. For exploring configurational comparability, function-based reduction principles are therefore needed as explored now.

Statistical explorations of configurational comparability

Constructs are abstract concepts denoting complex entities that are only constructed by human minds and that therefore cannot be directly perceived (e.g., "personality"; intelligence; Cronbach & Meehl, 1955). As a consequence, the analysis of configurational comparability is primarily a theoretical task and only secondarily an empirical one.

Exploring configurational comparability using function-based reduction principles and ratio-scaled time-relative probabilities

The TPS-Paradigm provides a two-step function-based reduction principle in which the behavioural raw data are first statistically reduced according to the known or assumed functions of the behaviours studied regardless of potentially moderate or even low consistencies in the empirical occurrences of functionally related behaviours. In a second step, the functionally defined composite measures thus-derived, rather than the behavioural raw data, are statistically analysed using redundancy-based or context-based reduction principles (as demonstrated below). This function-based reduction principle corresponds to people's intuitive categorisations of behaviours in everyday life that are primarily based

on the meanings attributed to behaviours rather than on their actual empirical occurrences (Uher, 2015b; Uher et al. 2013a,b).

To statistically explore the configurational comparability of functionally defined constructs, scientists also have to decide what kinds of empirical patterns they consider to be comparable between the particular groups or species studied. Specifically, researchers must decide if they weight functionally similar behaviours equally in their meaning as indicators of a particular construct or if they weight them differently in a particular study. These decisions cannot be made on the basis of mere statistical patterns; instead, they must be made on the basis of theoretical assumptions and with regard to the particular research question posed. In fact, these decisions precede statistical analyses because they determine the analytical algorithms that are used for computing composite measures. These algorithms influence the empirical results on the configurational patterns that can be identified in the behavioural variations of the different groups or species studied—and thus, the results on their configurational comparability.

The necessity of making such decisions does not become apparent in assessment-based research because people implicitly consider the different functions and meanings that they attribute to particular behaviours and may therefore implicitly weight the occurrences of particular behaviours differently when generating the raw data. In the primate studies used for illustrating these principles, all behavioural categories assigned to the same construct were attributed equal meanings. The corresponding behavioural variables were therefore weighted equally in the algorithms used for generating the composite measures of the three constructs studied. To consider that different behaviours generally occur with different frequencies in the same sample, all behavioural variables were first differentially standardised[10] (using z-standardisation). The z-standardised scores were then averaged across variables into a mean score per construct, thus weighted equally. Alternatively, some behaviours could also be considered to be more meaningful than others with regard to a given construct so that, for example, in the computation of the composite measures of Social Orientation, grooming and body contact could be weighted higher than arm-length and 1m proximity to others.

At the bottom of statistical analyses of the configurational comparability of construct measures and of patterning effects between groups and between species, respectively, are data reflecting individuals' *absolute* time-relative probabilities. This is because differential standardisation removes the information of absolute differences between individuals, groups and species, which are in the focus of these analyses. Therefore, configurational comparability across different sample levels is explored using three kinds of analyses that were developed for between-group comparisons in cross-cultural assessment-based research (Leung &

---

10 The behavioural (summary) scores (cf. Table 2), by contrast, were derived by adding scores together across variables, thus without considering differences in the behaviours' general occurrences.

Bond, 1989). Analogous analytical methodologies can be derived for comparisons between species on the basis of scientific quantifications using time-relative probabilities as shown now.

Exploring configurational comparability across different sample levels demonstrated by the example of species comparisons: Species-specific, universal and species-comparative analyses

In *species-specific analyses*, individuals' absolute time-relative probabilities are differentially standardised within each species. The species-specific construct measures derived from averaging these scores can then be explored for individual-specific variations and their configurational patterns. The results of these analyses are collected in matrices consisting of the separate species' co-variation scores of measurements of behaviours that are known or assumed to be functionally comparable between the species studied (e.g., internal consistencies, inter-correlations, test-retest reliabilities, see Tables 3 and 4).

In *universal analyses*, individuals' absolute time-relative probabilities are differentially standardised across the entire sample of all individuals ignoring their species membership. Averaging the differential time-relative probabilities thus-standardised therefore leads to construct measures in which individual variations are confounded by species-typical variations. Such universally standardised measures allow for exploring species-specific variations but not for exploring individual-specific variations ("personality") because species is the highest sample level for identifying individual-specificity. However, universal analyses can be applied to study individual-specificity when there is no evidence for systematic mean level differences between the species studied (as shown, e.g., in methodological studies on small samples of individuals of four great ape species; Uher, 2011b; Uher et al., 2008). Moreover, when species-specific analyses provide evidence for individual-specificity within each species studied and for configurational comparability of the given constructs between these species, then universal analyses can be used to explore the individual-specific variations that are shared by these species for mean-level differences between species (as explored below). The results of universal analyses are collected in matrices comprising the common co-variation scores of measurements of behaviours that are considered to be functionally comparable between the species studied.

In *species-comparative analyses*, individuals' absolute time-relative probabilities are averaged within each species to derive species-specific means of absolute time-relative probabilities for each given behavioural variable. Species-specific means of construct measures are derived by averaging individuals' time-relative probabilities that are first differentially standardised in universal analyses across all individuals of all species studied. By contrast, differential time-relative probabilities derived from the species-specific analyses (i.e., that are differentially standardised within each species) cannot be used because z-

standardisation sets the species' means to zero. The results of species-comparative analyses are collected in the same kind of matrices as those collected for the results from the species-specific and the universal analyses.

The species-specific, universal and species-comparative matrices containing various kinds of co-variation scores form the basis of the comparative analyses of the configurational and temporal patterns identified. The degree of between-species comparability of the empirical configurations of (behavioural) scores within each construct measure (i.e., construct equivalence), the empirical configurations of composite measures across multiple constructs (i.e., patterning effects) and the test-retest reliability of measurement variables and construct measures can be analysed statistically. Suitable methods are, for example, factor congruence coefficients (van de Vijver & Leung, 1997a,b; van de Vijver & Poortinga, 2002) and multi-group confirmatory factor analysis and structural equations modelling, which enable simultaneous tests of the same correlational or latent structure across species or group while keeping each sample distinct (Byrne, 1994; Byrne, Shavelson, & Muthén, 1989; Vandenberg & Lance, 2000).

Applications of these three kinds of analyses for exploring configurational comparability across different sample levels can be exemplarily demonstrated by data on the social behaviour of the four primate species. The left part of Table 3 presents the results on the internal consistency of the construct measures generated for the three functionally-defined constructs; the right part presents the results on the intercorrelations between these construct measures. From top to bottom, the table contains the three different kinds of matrices used for cross-level comparative analyses. Part 1) contains the species-specific matrix of the within-species findings on the internal consistency and the intercorrelations between the three construct measures. Part 2) contains the results from the corresponding universal analyses across the joint sample of all individuals of all four species. Part 3) shows the results from the corresponding species-comparative analyses performed on the species-specific mean scores (Table 3).

Table 3 *Configurational comparability: Species-specific, universal and species-comparative analyses of construct equivalence and of patterning effects*

Kind of analysis	n	Construct equivalence (Internal consistency)			Patterning effects (Intercorrelations)		
xx Species	f_m	SO-Index	AG-index	DO-index	SOi-AGi	SOi-DOi	AGi-DOi
			a r (p)	r (p)	r (p)	r (p)	r (p)
1) Species-specific analyses of individual differences							
Weeper capuchins	9	.51	a .95**(.000)	-.51 (.165)	-.15 (.709)	-.13 (.774)	.98**(.000)
Mandrills	8	-.13	a .96**(.000)	-.29 (.487)	.21 (.618)	.47 (.244)	.60 (.119)
Toque macaques	12	.18	a .95**(.000)	-.49* (.103)	.08 (.794)	-.07 (.833)	.70**(.012)
Rhesus macaques	7	.03	a .56**(.188)	-.28 (.624)	.03 (.942)	.12 (.804)	-.01 (.977)
2) Universal analyses of individual differences							
All species	36	.05	a .69**(.000)	.14 (.404)	.04 (.838)	.08 (.668)	.55**(.001)
3) Species-comparative analyses of species-specific means							
All species	4	-.05	a .27 (.726)	.99**(.000)	.33 (.672)	.17 (.833)	.51 (.495)

Note. * $p < .05$; ** $p < .01$. n The number of individuals or species, respectively, studied in each given kind of analysis. a Mean correlations were computed using r-to-z transformation; therefore, p-values cannot be reported. Exemplary demonstrations of analyses of configurational comparability of the three functionally-defined constructs in the four primate species studied: The *Social Orientation Index (SOi)* is based on grooming and the number of conspecifics in body contact, arm-length and 1m distance; the *Aggressiveness Index (AGi)* is based on non-contact aggression and contact aggression, and the *Dominance Index (DOi)* is based on displacing other, being displaced and subordinate to others (the latter two were a priori summarised into one variable because subordination occurred only rarely). Computations are based on individual means and species means of absolute time-relative probabilities, averaged across study periods t_2 and t_3 in which data from all four species are available. Construct measures (indices) were computed as mean scores of differentially standardised behavioural variables (using z-standardisation). In the *species-specific analyses*, z-standardisation was performed for each species separately before construct measures were computed. In the *universal and species-comparative analyses*, z-standardisation was performed on the individuals of the joint sample of all four species before individual construct measures and species means were computed. Universal and species-comparative analyses are used to explore whether the between-individual and between-species differences are comparable across the different sample levels. Construct equivalence was analysed in terms of the internal consistency of the composite measures using Pearson correlation r. Intra-class correlation could not be used because the internal consistency of the Social Orientation Index was low and tended to be negative and because the two measures summarised by the Dominance Index have opposite meanings (displace others and subordinate), thus resulting necessarily in negative co-variations. Patterning effects were analysed in terms of Pearson correlations r between the species-specific, universal and species-comparative measures of all three constructs.

The findings, although limited by the small sample sizes, illustrate basic principles. The species-specific internal consistencies of the Social Orientation Index (SOi) were generally low, except for the weeper capuchins (Table 3; left columns). The internal consistencies of this construct were also low in the universal analyses and in the species-comparative analyses. In psychometrics, it is common practice to discard measurement variables that yield only low internal

consistency (Block, 2010). Such redundancy-based reduction principles were developed in assessment-based research that inherently relies on language, in which redundancies can be created at low cost and are therefore widespread (cf. Lahlou, 1996). But in behaviour, redundancies are rare and may be even constraint because they are too costly in ecological and evolutionary regards (Uher, 2015b; Uher et al., 2013a). Given the shared function and meaning of grooming and proximity to others, there is no reason to discard any of these behavioural measurements or even the construct just because the empirical occurrences of these behaviours do not co-vary on the sample level.

Species-specific internal consistencies of the Aggressiveness Index (AGi), by contrast, were high in three species and moderate in the rhesus macaques (Table 3). This is interesting because in some previous studies, occurrences of contact and non-contact aggression were not associated (e.g., in 21 tufted capuchins, *Cebus apella*; Uher et al., 2013a). High internal consistency of the Aggressiveness Index was also found in the universal analysis performed on the joint sample of all individuals from all four species. But in the species-comparative analysis performed on the species-specific mean scores, consistency was only moderate. These findings indicate absence of mean-level differences between the species as explored below.

The internal consistencies of the Dominance Index (DOi) showed inverse patterns (Table 3). The species-specific internal consistencies were only moderate in all four species, the universal internal consistency was low and the species-comparative internal consistency was high. The Dominance index comprises behaviours with opposite meanings (i.e., displace others and subordinate to others). Therefore, species-specific internal consistencies were negative as expected. But despite the opposite valences of these behaviours with regard to the construct Dominance, their occurrences in individuals are not mutually exclusive (e.g., because individuals may show both kinds of behaviours but to different social partners). This may explain the only moderate internal consistencies found within each species. The low but positive internal consistency found in the universal analysis and the high internal consistency found in the species-comparative analysis therefore indicate the presence of mean-level differences in the occurrences of these behaviours between at least some of the species studied as explored below.

The intercorrelations between these construct measures showed much more coherent patterning effects across the different levels of analysis (Table 3, right columns). The Social Orientation and Aggressiveness indices were not associated within and across the species. This is interesting because in a previous study on crab-eating macaques the behaviours described by these construct were associated (e.g., $r = .75, p < .001, n = 61$; Uher et al., 2013b). Likewise, the Social Orientation and Dominance indices were not associated on any level of analysis, except for the mandrills in which low to moderate species-specific associations

occurred. The Aggressiveness and Dominance indices, by contrast, were substantially associated within three of the four species (except for the rhesus macaques) as well as in the universal analyses. The species-comparative analysis indicated similar associations.

In larger samples, the results collected in the three kinds of matrices could now be explored statistically for the degrees of construct equivalence and the degrees of similarity in the patterning effects between the different species studied and within and across the different sample levels explored. For the purposes of the present study and considering the small sample sizes, the similarities between the three matrices shall be sufficient to assume that the three constructs are configurationally comparable between the four species studied.

Additional evidence for this assumption is provided by results on these constructs' temporal reliability derived from species-specific, universal and species-comparative analyses. The three matrices containing the corresponding test-retest correlations between all three study periods are shown in Table 4, which is structured identically to Table 3. The high temporal reliabilities shown in all species-specific analyses over 2 and 4 weeks for all constructs and over 5 week periods for most constructs provide evidence that the construct measures reflect individual-specific patterns in each of the four species. Temporal reliabilities were also high in all universal analyses. But universal analyses can yield high temporal stabilities even in absence of individual-specific variations because they are confounded by species-typical variations. First indications for systematic mean-level differences between the species therefore derive from high variable-oriented temporal stability of the species-comparative analyses as shown in Table 4.

Variable-oriented analyses explore the rank ordering of scores on the variables studied; thus without mean-level differences between the species, high temporal reliability cannot be found in species-comparative analyses. Therefore, the high test-retest reliabilities of the Social Orientation index across all study periods indicate the possible presence of mean-level differences between species, whereas the varying and low temporal reliabilities of the Aggressiveness and Dominance indices indicate the possible absence of temporally stable between-species differences. The TPS-Paradigm elaborates basic principles that allow for scientifically quantifying mean-level differences between groups and species as described in the following.

4.

Table 4 *Species-specific, universal and species-comparative analyses of the temporal stability of individual behavioural variations over 2, 4 and 5 weeks: Evidence for individual-specific variations within and across species*

Kind of analysis Construct	Species	n	Test-retest reliability of construct measures					
			2 weeks t_1-t_2 r	p	4 weeks t_2-t_3 r	p	5 weeks t_1-t_3 r	p
1) Species-specific analyses of individual differences								
SO-Index	Weeper capuchins	9	.73*	(.026)	.95**	(.000)	.80**	(.010)
	Mandrills	8	.73*	(.041)	.61	(.108)	.20	(.642)
	Toque macaques	12	.92**	(.000)	.90**	(.000)	.93**	(.000)
	Rhesus macaques	7	-	-	.60	(.153)	-	-
AG-Index	Weeper capuchins	9	.80**	(.009)	.89**	(.001)	.66	(.051)
	Mandrills	8	.68	(.065)	.72*	(.043)	.97**	(.000)
	Toque macaques	12	.67*	(.017)	.76**	(.004)	.55	(.062)
	Rhesus macaques	7	-	-	.73	(.065)	-	-
DO-Index	Weeper capuchins	9	.86**	(.003)	.86**	(.003)	.78*	(.010)
	Mandrills	8	.93**	(.001)	.86**	(.006)	.87**	(.005)
	Toque macaques	12	.85**	(.000)	.74**	(.006)	.70*	(.012)
	Rhesus macaques	7	-	-	.79*	(.036)	-	-
2) Universal analyses of individual differences								
SO-Index	All species	36	.91**	(.000)	.86**	(.000)	.84**	(.000)
AG-Index	All species	36	.65**	(.000)	.59**	(.000)	.87**	(.002)
DO-Index	All species	36	.92**	(.000)	.80**	(.000)	.82**	(.000)
3) Species-comparative analyses of species-specific means								
SO-Index	All species	4	.97*	(.030)	.89	(.106)	.92	(.077)
AG-Index	All species	4	.72	(.277)	.37	(.630)	.74	(.262)
DO-Index	All species	4	.74	(.261)	-.67	(.330)	-.15	(.846)

Note. * $p < .05$; ** $p < .01$. SO Social Orientation, AG Aggressiveness, DO Dominance. Variable-oriented analyses of the test-retest reliability (using Pearson correlation r) of between-individual differences across different sample levels. *n* The number of individuals or species, respectively, that were studied in each given kind of analysis. Computations are based on individual means and species means of differential time-relative probabilities, averaged across study periods t_2 and t_3 in which data from all four species are available. In the species-specific analyses, z-standardisation of the behavioural variables comprised by the construct measures was

performed for each species separately. In the *universal and species-comparative analyses*, z-standardisation was performed on the individuals of the joint sample of all four species before Individual construct measures and species means were computed. But importantly, only species-specific analyses can be used to identify individual-specificity ("personality") because species is the highest population level for determining the individual-specific variations identified are comparable across the different sample levels as well as possible mean-level differences in these variations between species.

Statistical analyses of mean-level differences between groups and species: Positioning effects

Constructs describing individual-specific variations that are considered to be functionally and configurationally comparable between particular groups or species also allow for exploring quantitative differences between the groups or species studied. Such mean-level differences along shared dimensions of between-individual differences are also referred to as *positioning effects* (Leung & Bond, 1989; Uher, 2008a,b).

Scientific quantifications of positioning effects on the level of variables must be based on individuals' absolute time-relative probabilities; pertinent explorations on the construct level must be based on individuals' time-relative probabilities that are differentially standardised across the joint sample of individuals from all groups or species studied (i.e., in universal analyses). The TPS-Paradigm emphasises that scientific quantifications of behavioural mean-level differences between individuals, groups and species depend on the particular behavioural situations studied.

The role of situations for scientific quantifications of behavioural variations within and between individuals, groups and species

Given that behaviours are functionally mediated by other external phenomena that are present in the moments of their occurrence, analyses of mean-level differences require that relevant properties of the studied behavioural situations are known or assumed to be functionally and quantitatively comparable between the particular groups or species studied. Therefore, these situations need not—and often cannot—be nominally identical. Rather, comparability of behavioural situations must be established on higher levels of abstraction.

Higher levels of abstraction are also required because the functional relevance of situational properties varies also between individuals of the same group (Bandura, 1986; Mischel & Shoda, 1995; Rotter, 1954, 1981). In fact, it is well known that situational properties are directly related to individual variations. The concept of *weak situations* describes constellations of situational properties in which individual differences emerge in a given sample; such constellations have high power to discriminate between individuals. But there are also constellations of situational properties in which individual differences do not occur, thus resulting

ing in floor or ceiling effects in the distribution of empirical scores. Such constellations do not discriminate well between individuals and are therefore referred to as *strong situations* (in the sense that they restrict the emergence of individual variations; Mischel, 1977; Tett & Guterman, 2000; Uher, 2011a,b).

The particular quantitative properties of behavioural situations that differentiate well between individuals and that result in comparable distribution patterns may vary between groups or species. For example, in some nonhuman species, most individuals may flee from unfamiliar objects sized 10% of their body size, whereas in other species most individuals may do so only if unfamiliar objects are sized 25% their body size. This in itself reflects quantitative differences between the particular groups or species studied. The concept of situational strength clarifies that statistical comparisons of behavioural scores without considering the particular properties of the behavioural situations in which they were obtained may bias findings on mean-level differences between individuals, groups or species.

In observational studies, situational properties that are considered to be relevant can be quantified and controlled for in the empirical analyses; for example, activity patterns may be weighted against the space available to each individual of the social groups studied. Experimental studies additionally allow for making adjustments and targeted manipulations of particular situational properties; for example, quantitative situational properties in novel object tests can be made comparable by using equal object-size-to-body-size ratios (e.g., what may be a small object for chimpanzees can be life-sized for mice). In the primate studies used here for illustrating the comparative methodologies, situational properties relevant for social behaviours were made quantitatively comparable between the species (and thus between social groups of different size) by weighting individuals' time-relative probabilities for showing particular behaviours against the total number of individuals in each social group to account for differences in the number of potential interaction partners available (see above).

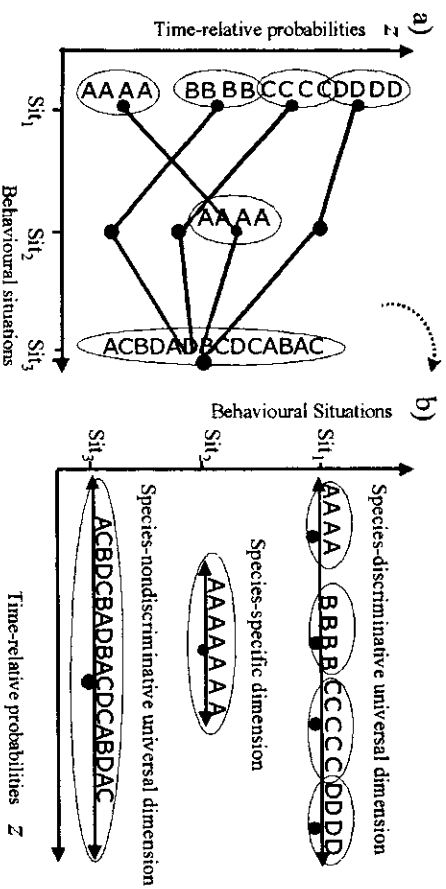*Individual-specific and species-specific situation-behaviour profiles*

Individuals' time-relative probabilities to display particular behaviours often vary across different behavioural situations. In addition to such general shifts in the occurrences of behaviours across situations in a given sample, there are often pronounced between-individual differences in these cross-situational variations that are relatively stable over time. If such individual-specific patterns emerge in a given sample, then the *cross-situational consistency* of the between-individual variations in the occurrences of this particular behaviour can be only moderate (Mischel, 1968; Uher, 2011a,b; Uher et al., 2008, 2013a; Wright & Mischel, 1987).

The cross-situational variations that are specific to particular individuals can be illustrated with *individual-specific situation-behaviour profiles* (sometimes called if...then...profiles; Mischel et al., 2002). In these profiles, an individual's time-relative probabilities to show a particular kind of behaviour are plotted against the particular kinds of situations in which they were measured. For direct comparisons of individual-specific scores between different situations, individuals' scores are differentially standardised within each situation. This allows for exploring, for example, in which kinds of situation a given individual has the highest time-relative probabilities to show a given behaviour and in which ways this may differ between the individuals studied (Uher, 2011a,b; Uher et al., 2008, 2013a).

The TPS-Paradigm extends the concept of individual-specific situation-behaviour profiles to the levels of groups and species for enabling scientific quantifications of mean-level differences in individual-specific behavioural variations between groups and between species. Figure 2a illustrates this concept using the hypothetical example of four species labelled A, B, C and D; the single letters represent single individuals of these species. The individuals' scores of time-relative probabilities to show a particular behaviour are plotted against three kinds of behavioural situations that are assumed to be functionally and quantitatively comparable between these species. Importantly, the scores are differentially standardised across the joint sample of all four species within each situation, thus using universal analyses.

The average scores of the individuals of species A vary across the three situations; the average is the lowest in situation 1 and much higher in situations 2 and 3. As the scores are differentially standardised across the joint sample of individuals from all species, these shifts indicate situation-specific variations in the averages of species A relative to the situation-specific variations in the averages of the other species studied. The averages of the species B, C and D also vary across the three study situations yet in different ways. These differential shifts in species-specific averages can be illustrated as *species-specific situation-behaviour profiles* as shown in Figure 2a.

Figure 2 *The concept of species-specific situation-behaviour profiles and its relations to the concepts of species-discriminative and species-nondiscriminative universal dimensions and species-specific dimensions of individual-specific behaviours*

a)

Time-relative probabilities z

Sit₁ Sit₂ Sit₃
Behavioural situations

b)

Behavioural Situations

Sit₁ Sit₂ Sit₃

Species-discriminative universal dimension
Species-specific dimension
Species-nondiscriminative universal dimension

Time-relative probabilities z

Note. *a)* The concept of species-specific situation-behaviour profiles for enabling scientific quantifications of mean-level differences in individual-specific behavioural variations between groups and between species, illustrated by a hypothetical example of four species labelled A, B, C and D; the single letters represent single individuals of these species. The individuals' scores of time-relative probabilities to show a particular behaviour are plotted against three kinds of behavioural situations and are differentially standardised across the joint sample of all four species within each situation, thus using universal analyses. The average scores of each given species vary across the three situations yet in different ways for each species as indicated by their different profile shapes across the different situations. Situation 1 allows for exploring only between-individual differences and between-species differences. Situation 2 allows for exploring both between-individual differences and between-species differences; between-individual differences can be explored only in species A but not in the three other species. Situation 3, in turn, allows for exploring only in species A but not in the three other species; between-individual differences can be explored only in species A but not in the three other species. *b)* Seen from an orthogonal view on Figure 2a, continuous dimensions of individual behavioural variations within each of the three kinds of situations studied become apparent that constitute prototypes. Species-specific constructs describe kinds of individual-specific variations that occur only in a particular species but not in other species under study. Universal constructs describe kinds of individual-specific variations that occur in various species in comparable ways. Species-discriminative universal constructs describe kinds of individual-specific variation that differentiate also between species. Species-nondiscriminative universal constructs describe kinds of individual-specific variation that do not differentiate between species.

In each kind of situation, the scores of the individuals of species A vary around the species' average. This illustrates the fact that an individual's score always reflects both individual variations and species-typical variations—unless individuals' scores are differentially standardised within each species as done in species-specific analyses (Uher, 2011a,b). In the species B, C and D, between-individual variations occur only in situations 1 and 3 but not in situation 2. Hence for these three species, situation 2 constitutes a strong situation in which all individuals behave in highly similar ways so that between-individual variations do

not occur. Situations 1 and 2, by contrast, constitute weak situations for the species B, C and D because there are pronounced between-individual variations within each species (Figure 2a).

But situations 1 and 3 also differ from one another. Between-species differences occur only in situation 1, whereas the species' averages fall together in situation 3. Analogous to the concept of weak situations established for comparisons of between-individuals within a given sample (Mischel, 1977), the situations also differ in their power for discriminating between species. Situation 1 allows for exploring both between-individual differences and between-species differences. Situation 2 allows for exploring only between-species differences; between-individual differences can be explored only in species A but not in the three other species. Situation 3, in turn, allows for exploring between-individual differences within each species but not differences between species (Figure 2a).

Species-specific, species-discriminative universal and species-nondiscriminative universal variations of individual-specific patterns

An orthogonal view on Figure 2a reveals continuous dimensions of individual behavioural variations within each of the three kinds of situations studied (Figure 2b). Three kinds of between-species variations of individual-specific patterns can be distinguished that constitute prototypes.

Three prototypes of variations of individual-specific patterns within and across species

Situation 2 allows for discriminating between the four species only in terms of their averages but not in terms of between-individual differences. Such kinds of species differences are irrelevant for explorations of individual-specific variations within each of the four species studied but they may be relevant for other fields of research (e.g., biology, anthropology). In cross-cultural psychology, variations in the averages of sociocultural groups in absence of between-individual variations are referred to as cultural dimensions (Leung & Bond, 1989). Between-individual differences and thus, individual-specificity can be identified in Situation 2 only in species A. Such kinds of variations are described in cross-species comparative research by constructs of *species-specific dimensions* (Uher, 2008a,b), derived from analogous constructs of *emic dimensions* in cross-cultural research (Leung & Bond, 1989).

Situation 3 allows for discriminating between individuals within each of the four species studied but not for discriminating between species. In cross-species comparative research, such kinds of variation are described by constructs of

---

[1] This vital point is mistaken in some animal "personality" or temperament studies (e.g. Clarke & Boinski, 1995; Herrmann, Hare, Cissewski, & Tomasello, 2011).

weak universal dimensions (Uher, 2008a,b), derived from analogous constructs of weak etic dimensions in cross-cultural research (Leung & Bond, 1989).

Situation 1 allows for discriminating both between-individual variations within all four species and mean-level differences between species. In cross-species comparative research, such kinds of variation are described by constructs of strong universal dimensions (Uher, 2008a,b; cf. also Capitanio, 2004), derived from the analogous constructs of strong etic dimensions in cross-cultural psychology (Leung & Bond, 1989).

The various meanings of "weak" and "strong" in comparative concepts may cause confusion. In the concept of situational strength from "personality" psychology, the terms refer to constraints that situations may impose on individuals' behaviours; thus, "strong" indicates absence of variations. In the cross-cultural and cross-species comparative concepts, by contrast, the terms refer to discriminative power; thus, "strong" denotes presence of variations.

The TPS-Paradigm therefore applies a more descriptive terminology. Kinds of individual-specific variation that occur only in a particular species but not in other species under study are called species-specific. Those kinds of individual-specific variations that occur in various species are called universal for the particular group of species considered. Universal kinds of individual-specific variation that also differentiate between species are called species-discriminative and those that do not differentiate between species are called species-nondiscriminative. Analogous concepts for comparisons within and across groups of conspecific individuals describe group-specific, group-discriminative and group-nondiscriminative universal variations of individual-specificity.

Empirical demonstration of the comparative methodologies for exploring differences between species

These concepts are directly related to the three different kinds of analyses used for exploring variations within and between groups and species. This can be illustrated by the empirical data on the four primate species. The various species-specific analyses of the three construct measures (cf. Tables 3 and 4) provided evidence for configurationally comparable individual-specific variations in all four species; thus, none of these variations are species-specific. The universal analyses suggested that the three constructs describe individual-specific variations that are universal for these species. Species-comparative analyses indicated some of the four species, i.e., species-discriminative universal variations, and absence of stable between-species differences in the Aggressiveness and Dominance indices, i.e., species-nondiscriminative universal variations of individual-specificity.

These indications were statistically tested using One-way ANOVAs on the individuals' averages across study periods t_2 and t_3 for the three construct

measures, which reflect differentially (z-)standardised scores, and for the more specific behavioural (summary) scores, which reflect absolute-time relative probabilities. The magnitude of the between-species differences was estimated using Cohen's effect size d on pooled standard deviations. This effect size can also be interpreted in terms of the percent of non-overlap of the score distributions between the two groups contrasted (Cohen, 1988). Given the small sample sizes, post-hoc power analyses were conducted using G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007) to explore which of the results are likely to replicate.

Significant species-differences occurred in the Social Orientation index (SOi), $F_{soi}(3, 32) = 8.274$, $p = .000$. Bonferroni tests indicated that the weeper capuchins ($M = -0.41$; $SD = 0.27$) scored significantly lower than the mandrills ($M = -0.04$; $SD = 0.16$), toque macaques ($M = 0.19$; $SD = 0.38$) and rhesus macaques ($M = 0.13$; $SD = 0.21$) with effect sizes ranging from $d = -1.6$ to 2.2 (73-85% non-overlap in the score distributions and .88-.98 achieved power of detecting such a difference; see Table 5, upper part). Hence, the Social Orientation index has a high power to discriminate weeper capuchins from the three other species but not for discriminating between these three species.

These species differences were further explored by analysing the proximity summary score (PX) and the grooming score (GR) that the Social Orientation index comprises. Weeper capuchins ($M = 0.75$; $SD = 0.25$) stayed in proximity to significantly fewer conspecifics than toque macaques ($M = 1.36$, $SD = 0.47$) but not significantly fewer than mandrills ($M = 1.01$; $SD = 0.17$) and rhesus macaques ($M = 0.99$; $SD = 0.15$), $F_{PX}(3, 32) = 6.430$, $p = .002$ (using Tamhane's T2 test; see Table 5, lower part). Hence, the proximity summary score (PX) had high power only for discriminating the capuchins from toque macaques but not from the two other species. The time-relative probabilities for grooming did not differ between weeper capuchins ($M = 0.03$; $SD = 0.03$) and Mandrills ($M = 0.03$; $SD = 0.29$), but both species groomed significantly less often than toque macaques ($M = 0.09$; $SD = 0.05$) and rhesus macaques ($M = 0.10$; $SD = 0.04$); the two macaque species did not differ from one another; $F_{GR}(3, 32) = 8.408$, $p = .000$ (using Bonferroni tests). Thus, the grooming scores had high power for discriminating between these two groups of species ($1-\beta > .90$; see Table 5).

Table 5 *Mean-level differences between species: Evidence for positioning effects in the construct measures and the behavioural (summary) scores*

Species contrasted		Construct measures (universally standardised)												
		SO-Index			AG-Index			DO-Index			DI+SU-Index		
Sp₁	Sp₂	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β
Cw	Md	**-1.6**	73	**.88**	-0.4	27	.12	0.0	0	.05	**-2.2**	85	**.99**
Cw	Tq	**-1.8**	77	**.97**	-0.2	15	.06	0.1	7	.05	**-1.7**	75	**.98**
Cw	Rh	**-2.2**	85	**.98**	0.1	8	.06	0.1	7	.05	-0.3	21	.15
Md	Tq	-0.8	47	.38	0.3	21	.10	0.0	0	.05	**1.5**	70	**.95**
Md	Rh	-0.9	52	.36	0.5	33	.17	0.0	0	.05	**2.1**	83	**.99**
Tq	Rh	0.2	15	.07	0.5	33	.17	0.0	0	.05	**1.6**	73	**.93**

Species contrasted		Behavioural (summary) scores (absolute time-relative probabilities)												
		PX			GR			DI			SU		
Sp₁	Sp₂	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β	d_{sp}	%ₙₒ	1-β
Cw	Md	-1.2	62	.64	0.1	7	.05	-1.4	68	.79	-1.2	62	.64
Cw	Tq	**-1.6**	73	**.93**	**-1.5**	71	**.95**	-0.8	48	.38	-1.2	62	.71
Cw	Rh	-1.2	62	.57	**-2.0**	81	**.98**	-0.3	21	.08	0.1	7	.05
Md	Tq	-1.0	55	.54	**-1.6**	73	**.90**	1.0	55	.53	**1.5**	70	**.95**
Md	Rh	0.1	8	.06	**-2.1**	83	**.96**	1.4	68	.70	**2.1**	83	**.99**
Tq	Rh	1.1	59	.55	-0.2	15	.07	0.7	43	.27	1.1	58	.56

Note. d_{sp} Effect sizes Cohen's *d* for differences between species; bold effect sizes reflect species differences for which the given construct measures have a high test power (1-β > .88) in the samples studied and that are thus more likely to replicate in larger samples. %ₙₒ Percentage of non-overlap of the score distributions of the two species compared. 1-β test power achieved in the given sample sizes. Sp₁ and Sp₂ indicate the species contrasted; their order in the table is relevant for interpreting the direction of the mean-level differences between them such that a negative effect size indicates that the scores of Sp₁ are lower than those of Sp₂. Cw weeper capuchins, Md mandrills, Tq toque macaques, Rh rhesus macaques. Construct measures (upper part of the table) reflect individuals' differential time-relative probabilities and were computed in universal analyses as the means of standardised behavioural scores that were z-standardised across the individuals of all species. Behavioural (summary) scores (lower part) reflect individuals' absolute time-relative probabilities to show particular kinds of behaviours (summary scores and were computed by adding together the scores from the more specific related behavioural variables). Computations are based on individuals' averages across study periods t₂ and t₃ in which data from all four species are available. Results on the aggression summary scores (AG) are not shown because were virtually identical to those of the Aggressiveness index (AGi), which were derived from the same variables yet using a slightly different algorithm. The Dominance index (DOi) was computed from individuals' displacement scores and their inverted (z-standardised) subordination scores to reflect individuals' dominance relationships within their social group and thus within each species; therefore, the Dominance index cannot reflect species-differences. For enabling species comparisons, the DI-index was computed from non-inversed (z-standardised) subordination scores; thus the DI+SU-index reflects the differential occurrences of *any kind* of dominance-related behaviours, i.e., dominant and subordinate.

In the Aggressiveness index (AGi), the mandrills scored the highest ($M = 0.24, SD = 0.99$) but not significantly higher than the weeper capuchins ($M = -0.08; SD = 0.64$), toque macaques ($M = 0.00, SD = 0.38$) or rhesus macaques ($M = -0.15; SD = 0.20$); $F_{AG}(3, 32) = .612, p = .612$. The aggression summary scores (AG) were virtually identical because they were derived from the same variables yet with a slightly different algorithm (Table 5, upper part). Both kinds of measures had no power for discriminating between any of the four species.

The Dominance Index (DOi) did not reflect species-differences, $F_{DO}(3, 32) = .002, p = 1.000$, but the strikingly similar mean-scores of the four species, weeper capuchins ($M = -0.01; SD = 0.11$), mandrills ($M = 0.00, SD = 1.11$) toque macaques ($M = 0.00; SD = 0.42$) and rhesus macaques ($M = 0.00; SD = 0.10$; Table 5, upper part), indicated the possibility of a methodological artefact derived from the algorithm used for its computation. The Dominance index comprises two behavioural variables with opposite meanings (i.e., displace and subordinate) and was therefore computed from the inverted (z-standardised) scores of one of these variables (i.e., subordinate) so as to reflect individuals' dominance rank within each social group. Dominant individuals seldom display subordinate behaviours and subordinate individuals seldom display dominant behaviours. Therefore, possible species differences in the overall occurrences of both behaviours cannot be reflected in the index.
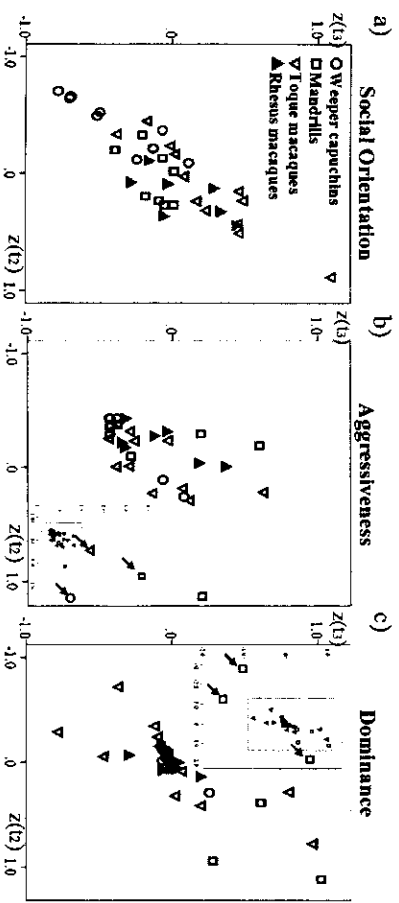
This issue was further explored by analysing the two behavioural summary scores that the Dominance index comprises and that reflect individuals' absolute-time relative probability scores. Displacement (DI) occurred in the man-drills ($M = 0.25; SD = 0.23$) more often than in the weeper capuchins ($M = 0.02; SD = 0.03$) and in the rhesus macaques ($M = 0.02; SD = 0.02$), but the displacement score had only moderate power for detecting such differences in the given samples (Table 5, lower part). Mandrills also displaced conspecifics more often than did the toque macaques ($M = 0.08; SD = 0.11$); but the power of these scores for discriminating between these species was only low. Given the unequal variances, none of these differences were significant as Tamhane's T2 tests showed, $F_{DI}(3, 32) = 6.116, p = .002$. Subordination (SU) also occurred in the mandrills ($M = 0.27, SD = 0.30$) more often than in the weeper capuchins ($M = 0.02; SD = 0.01$), toque macaques ($M = 0.08; SD = 0.07$) and rhesus macaques ($M = 0.02; SD = 0.03$), $F_{SU}(3, 32) = 5.357, p = .004$, but also not in significant ways as Tamhane's T2 tests showed. The power of the subordination score for discriminating between the species was only moderate (Table 5, lower part).

To account for the opposite meanings of displacement and subordination behaviours and for the resulting different distribution patterns of these behaviours within each species, an alternative construct measure was computed using non-inversed (z-standardised) subordination scores. This DI+SU-index thus reflects the differential occurrences of *any kind* of dominance-related behaviours, i.e., displace and subordinate. In this index, the mandrills ($M = 0.87; SD = 0.83$)

scored significantly higher than the weeper capuchins ($M = -0.42$; $SD = 0.08$) and the rhesus macaques ($M = -0.39$, $SD = 0.09$) but not than the toque macaques ($M = -0.08$; $SD = 0.26$), and the toque macaques scored significantly higher than the weeper capuchins and the rhesus macaques; $F_{Px}(3, 32) = 16.595$, $p = .000$ (using Tamhane's T2 tests); $F_{DI+SU}(3, 32) = 16.596$, $p = .000$. Despite the small sample sizes, the DI+SU-index had high power to discriminate between the species except for discriminating between weeper capuchins and rhesus macaques (see Table 5, upper part).

These positioning effects are illustrated in Figures 3a-c in which individuals' differential time-relative probabilities (standardised across all individuals from all species in universal analyses) from study period t_2 are plotted against those from study period t_3. Hence, the graphs depict the temporal stability of the between-individual variations described by the three construct measures that were explored within and across species. As these variations were also stable within each species (see Table 4), they reflect both individual-specific variations (i.e., "personality" differences) and—in some cases—mean-level differences in the species' mean scores around which these individual-specific variations occur. The individuals of all four species vary along the dimensions of individual-specific variations described by the three constructs and that are thus universal for these species.

Figure 3 *Empirical evidence for species-discriminative and species-nondiscriminative universal variations of individual-specific behaviours*

a) Social Orientation | b) Aggressiveness | c) Dominance

Legend: O Weeper capuchins, □ Mandrills, ▽ Toque macaques, ▲ Rhesus macaques

Axes: z(t₃), z(t₂); 1.0, 0, -1.0

Note. Empirical demonstrations of the concepts of species-discriminative and species-non-discriminative universal dimensions of individual-specific variations and the species' positioning effects. Individuals' differential time-relative probabilities (standardised across all individuals from all species in universal analyses) from study period t_2 are plotted against those from study period t_3. The individuals of all four species vary along the dimensions of individual-specific variations described by the three constructs, thus they are universal for these species. For more direct comparisons between the constructs, the three graphs are all scaled such that each axis spans about two standard deviations ($z = -1$ to $+1$). The small graphs inserted in the Figures 3b and 3c are scaled differently. In Figure 3b, the x-axis is scaled from $z = -1.0$ to 2.5 and the y-axis inserted in the Figures 3b and 3c are

In the Social Orientation index (Figure 3a), the scores of the weeper capuchins (depicted as circles) were distributed at the low end of the dimension and those of the two macaque species (depicted as white and black triangles) were distributed toward the high end. Mandrills (depicted as squares) were distributed more in the middle of the dimension. Hence with regard to the weeper capuchins as compared to the three other species, this construct describes a species-discriminative universal dimension of individual-specific variants with only 15-27% overlap in the score distributions between the weeper capuchins and the three other species. If only mandrills and the two macaque species are considered, the construct describes a species-nondiscriminative universal dimension of individual-specific variants. Thus in this construct, the two prototypes of species-discriminative and species-nondiscriminative universal variations merge together (cf. Figure 2).

In the Aggressiveness index (Figure 3b), the scores of the individuals from all four species were distributed along the same dimension. The more scattered distribution patterns reflect the slightly lower temporal stability of these between-individual variations (cf. Table 4). The mixed distribution of the scores from the different species across the entire dimension reflects the absence of species differences. Hence, the Aggressiveness index describes a species-nondiscriminative universal dimension of individual-specific variants.

For more direct comparisons between the constructs, the three graphs are scaled in exactly same way so that the ranges of individuals' score distribution span about two standard deviations ($z = -1$ to $+1$). The small graph inserted in the bottom right corner of Figure 3b, by contrast, is scaled differently and shows the distribution of all Aggressiveness scores including three outliers (a weeper capuchin, a mandrill and a toque macaque) with scores up to 4 standard deviations higher than the individuals at the low end of the Aggressiveness dimension.

In the Dominance-index, individuals' scores also varied along the same dimension. The graph shows that the ranges of the within-species variabilities of the mandrills and toque macaques were much higher than those of the weeper

three outliers are the Capuchin alpha male $z(2.6;0.4)$, a mandrill female $z(1.4; 3.6)$ and the alpha male of the toque macaques $z(0.5;1.4)$. In Figure 3c, the x-axis is scaled from $z = -3.0$ to 2.0 and the y-axis from $z = -.2$ to 1.0); outliers are three mandrills, two low-scoring females $z(-1.2; -1.5)$ and $z(-2.7;-1.1)$ and the alpha male scoring high in study period t_2 $z(1.8; 0.5)$. a) The Social Orientation index describes a species-discriminative universal dimension of individual-specific variants with regard to weeper capuchins as compared to the three other species but a species-nondiscriminative universal dimension if only mandrills, toque macaques and the rhesus macaques are considered. Thus, the two prototypes of species-discriminative and species-nondiscriminative universal variations merge together in this construct (cf. Figure 2a,b). b) The Aggressiveness index describes a species-nondiscriminative universal dimension of individual-specific variants; individuals of all for species vary along this dimension but there are no mean-level differences between species. c) The Dominance-index describes a species-nondiscriminative universal dimension of individual-specific variations in which species differences did not emerge in the species' averages but in their between-individual variabilities in which the within-species variabilities of the mandrills ($M_{SD} = .83$) were much higher than those of toque macaques ($M_{SD} = .26$) that in turn were higher than those of the weeper capuchins ($M_{SD} = .08$) and rhesus macaques ($M_{SD} =.09$), which did not differ from one another.

capuchins and rhesus macaques. Three mandrills scored up to three standard deviations lower and two standard deviation higher, respectively, than the average individual from all four species (see the small differently scaled graph inserted in the left upper corner of Figure 3c). Interestingly, these pronounced within-species deviations from the joint average of all species occurred in both directions. This explains the absence of mean-level differences between species. That is, in the mandrills (and some extent also in the toque macaques) there were some individuals who showed displacement significantly more often and subordination significantly less often than the average of all four species, thus who behaved extremely dominant, and other individuals who showed the reverse pattern, thus who behaved very submissively. This suggests that the mandrills' dominance hierarchy was much steeper than those of the other species. This is commonly referred to as a "despotic" dominance style as opposed the more "egalitarian" styles of other species (Otovic et al., 2014).

Hence, the Dominance-Index describes individual-specific variations in which species differences did not emerge in the species' averages but in their between-individual *variabilities*. Species-comparative analyses of the species' variabilities (using standard deviations of their between-individual differences) yielded high temporal stabilities across all study periods of $r_{t1-2} = .99$ ($p < .032$, $n = 3$), $r_{t2-3} = .98$ ($p < .022$, $n = 4$), $r_{t1-3} = .96$ ($p < .181$, $n = 3$) that contrast with the varying and partly absent temporal stability of the species' averages (see last row in Table 4). Across the study periods t_2 and t_3 depicted in Figure 3c, the average within-species variabilities of the mandrills ($M_{SD} = .83$) were much higher than those of toque macaques ($M_{SD} = .26$) that in turn were higher than those of the weeper capuchins ($M_{SD} = .08$) and rhesus macaques ($M_{SD} = .09$), which did not differ from one another. Consequently, the Dominance-Index describes universal individual-specific variations that are species-discriminative with regard to the species' variabilities but species-nondiscriminative with regard to the species' averages.

Empirical demonstration of the comparative methodologies for exploring between-group differences within and across species

Analogous analytical methodologies can be applied for comparisons of groups of individuals within and across species (see Table 1). An example of such analyses is provided in Table 6 containing tentative findings from species-specific analyses of sex differences in the four primate species using the behavioural (summary) scores, which reflect absolute time-relative probabilities. Given the small samples sizes, sex differences were explored only using Cohen's effect size d yet without testing for significance. Post-hoc power analyses were conducted to explore which of the results are likely to replicate. Sex differences were largely absent in all species with two exceptions. In the toque macaques, females tended to stay in proximity to more conspecifics than males ($M_F = 1.61$, $SD_F = 0.43$; M_M $= 1.02$, $SD_M = 0.29$) and in the rhesus macaques, females groomed others twice as much as males ($M_F = 0.12$, $SD_F = 0.01$; $M_M = 0.06$, $SD_M = 0.02$).

Table 6 *Sex differences within each species based on absolute time-relative probabilities*

Behavioural (summary) score	Weeper capuchins (n = 6/3)			Mandrills (n = 4/4)			Toque macaques (n = 7/5)			Rhesus macaques (n = 4/3)		
	$d_{F\,vs.\,M}$	%NO	1-β	$d_{F\,vs.\,M}$	%NO	1-β	$d_{F\,vs.\,M}$	%NO	1-β	$d_{F\,vs.\,M}$	%NO	1-β
GR	0.9	52	.19	1.8	77	.59	1.1	59	.38	**2**		**.98**
PX	0.3	21	.07	-1.0	55	.21	**1.6**	**73**	**.70**	0.2	15	.05
AG	-0.5	33	.10	0.6	38	.11	-0.6	38	.16	-0.4	27	.08
DI	-0.6	38	.12	0.0	0	.05	-1.3	65	.51	0.9	52	.16
SU	1.2	62	.29	1.7	75	.51	0.8	47	.22	-0.7	43	.12

Note. d_{F} Effect sizes Cohen's d for differences between females (F) and males (M); positive scores indicate that females score higher than males. Bold effect sizes reflect sex differences for which the given construct measures have a moderate or high test power, respectively ($1-\beta > .70$) in the samples studied and that are thus more likely to replicate in larger samples. %NO Percentage of non-overlap of the score distributions of the two sexes compared. $1-\beta$ test power achieved in the given sample sizes. Behavioural (summary) scores reflect individuals' absolute time-relative probabilities to show particular kinds of behaviours (summary scores were computed by adding together the scores from the more specific related behavioural variables): GR Grooming score, PX Proximity summary score, AG Aggression summary score, DI Displacement score, SU Subordination score. Computations are based on individuals' averages across study periods t_2 and t_3 in which data from all four species are available.

The present findings are in accordance with previous findings from a study on 26 tufted capuchins living in four social groups, in which sex differences in individual-specific behaviours were absent in 18 out of 20 different constructs studied (Uher et al., 2013a). These studies, although limited in sample sizes, showed that in five group-living species of nonhuman primates with pronounced sexual dimorphisms in body size, individuals did not exhibit pronounced sex differences in individual-specific behaviours as is often assumed for our human ancestors by evolutionary psychologists (e.g., Buss, 1999).

These findings contrast with findings from assessment-based studies that frequently report on sex and gender differences in "personality" in both nonhuman primates and humans (e.g., King, Weiss, & Sisco, 2008; Soto, John, Gosling, & Potter, 2011). But in lack of systematic contrasts to differences that can actually be observed in individuals' behaviours, it is still largely unknown to what extent assessments reflect socio-culturally shared beliefs, ideas and values. A previous study on 104 crab-eating macaques (*Macaca fascicularis*) and their 99 human observers with Dutch language-background has already demonstrated systematic attribution biases in the observers' assessments of these monkeys' individual-specific behaviours. These attribution biases were likely derived from the observers' socioculturally shared stereotypical beliefs about of sex/gender, age

and status differences in *human* individuals—thus, anthropomorphic biases (Uher et al., 2013b).

Summary

This research applied the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) to elaborate comprehensive metatheoretical and methodological frameworks for quantitative comparisons of individual variations within and across situations, groups and species. These frameworks, basic principles and concepts were elaborated using behavioural phenomena as examples. Their applications were demonstrated empirically using behavioural data from four nonhuman primate species. These principles and concepts were also used to critically analyse the quantifications of individual variations that can be obtained from standardised assessments and to highlight serious methodological deficiencies of assessment methods that are not well considered.

First, the chapter introduced a metatheoretical definition of behaviour that highlights the functionally mediating role of the behavioural situations in which behaviours occur. Particular spatio-temporal properties that can be conceived for behaviours (i.e., external to individuals' bodies, momentary and immaterial physical) were specified and used to derive various methodological implications (e.g., nunc-ipsum methods). Then the chapter outlined some basic concepts of the philosophy-of-science foundations that the TPS-Paradigm provides for scientific measurement and quantification (e.g., the set-theoretic and algebraic requirements, definitions of data, methods of data generation, the concept of time-relative probabilities). Their applications for generating scientific quantifications of individual variations in behaviour were demonstrated empirically by the example of behavioural studies in various primates. The philosophy-of-science foundations were also used to highlight fundamental differences to standardised assessment methods, which, unlike observational methods, do not allow researchers to fulfil the requirements of scientific quantification.

In the second part, the chapter applied these foundations to elaborate systematic metatheoretical and methodological frameworks for comparing individual variations within and across different sample levels using individuals, groups and species as examples. These frameworks were developed on the basis of concepts from "personality" psychology, differential psychology and cross-cultural and cross-species research that were systematically integrated into coherent frameworks and complemented by concepts for comparing individual-specific variations (i.e., "personality") between species.

Then concepts for disentangling specific from common variations were elaborated highlighting three interrelated viewpoints that are involved in any comparative exploration of individuals (i.e., specificity/ uniqueness, comparability, universality/ typicality). The chapter discussed basic principles for establish-

and status differences in *human* individuals—thus, anthropomorphic biases (Uher et al., 2013b).

Summary

This research applied the Transdisciplinary Philosophy-of-Science Paradigm for Research on Individuals (TPS-Paradigm) to elaborate comprehensive metatheoretical and methodological frameworks for quantitative comparisons of individual variations within and across situations, groups and species. These frameworks, basic principles and concepts were elaborated using behavioural phenomena as examples. Their applications were demonstrated empirically using behavioural data from four nonhuman primate species. These principles and concepts were also used to critically analyse the quantifications of individual variations that can be obtained from standardised assessments and to highlight serious methodological deficiencies of assessment methods that are not well considered.

First, the chapter introduced a metatheoretical definition of behaviour that highlights the functionally mediating role of the behavioural situations in which behaviours occur. Particular spatio-temporal properties that can be conceived for behaviours (i.e., external to individuals' bodies, momentary and immaterial physical) were specified and used to derive various methodological implications (e.g., nunc-ipsum methods). Then the chapter outlined some basic concepts of the philosophy-of-science foundations that the TPS-Paradigm provides for scientific measurement and quantification (e.g., the set-theoretic and algebraic requirements, definitions of data, methods of data generation, the concept of time-relative probabilities). Their applications for generating scientific quantifications of individual variations in behaviour were demonstrated empirically by the example of behavioural studies in various primates. The philosophy-of-science foundations were also used to highlight fundamental differences to standardised assessment methods, which, unlike observational methods, do not allow researchers to fulfil the requirements of scientific quantification.

In the second part, the chapter applied these foundations to elaborate systematic metatheoretical and methodological frameworks for comparing individual variations within and across different sample levels using individuals, groups and species as examples. These frameworks were developed on the basis of concepts from "personality" psychology, differential psychology and cross-cultural and cross-species research that were systematically integrated into coherent frameworks and complemented by concepts for comparing individual-specific variations (i.e., "personality") between species.

Then concepts for disentangling specific from common variations were elaborated highlighting three interrelated viewpoints that are involved in any comparative exploration of individuals (i.e., specificity/ uniqueness, comparability, universality/ typicality). The chapter discussed basic principles for establish-

ing the functional comparability of behavioural and situational categories that allow researchers to consider that individuals from different groups and species often show different behaviours and encounter different situations and thus cannot be studied with identical variables as is done in assessment-based research.

Using these concepts and principles, methodologies for the statistical analyses of the configurational comparability (i.e., equivalence and patterning effects) of constructs within and across different sample levels, originally developed in cross-cultural psychology, were discussed and generalised to concepts for cross-species comparisons (e.g., species-specific, universal and species-comparative analyses). The chapter introduced basic principles for enabling such explorations on the basis of function-based approaches and time-relative probabilities that can help researchers to overcome the limitations of assessment-based methods (e.g., subjective quantifications generated on the basis of unknown encoding rules, redundancy-based principles, ecological fallacies). Applications of these principles and analytical methodologies were demonstrated empirically in explorations of three constructs of individual-specific behaviours (Social Orientation, Aggressiveness and Dominance) that were shown to be configurationally comparable between the four primate species studied.

Finally, analytical methodologies needed to scientifically quantify mean-level differences between groups and species were explored (i.e., positioning effects). The chapter introduced the concept of species-specific situation-behaviour profiles that demonstrates the ways in which situational properties are related to quantitative differences between individuals, groups and species. Using these methodologies, the three studied constructs were shown to describe species-discriminative and species-nondiscriminative universal variations of individual-specific behaviours. Applications of the analytical methodologies for comparing groups of individuals within and across species were demonstrated by exploring sex-differences in the individual-specific behaviours described by the three constructs and that were shown to be largely absent in all four species studied.

Implications

Given the complex issues involved in comparative analyses of individual variations as explored in this research, how realistic is it to assume that people can accurately assess individual-specific variations (i.e., "personality")? And how realistic is it to assume that standardised assessments methods enable people to provide quantifications on differences between individuals, groups and species that correspond to patterns that can be identified in observable events? The metatheoretical and methodological frameworks elaborated in this research enable systematic contrasts to be applied with assessments that people provide on individuals and on differences between individuals, groups and species. Such contrast allow for exploring the human abilities to quickly form impressions of individuals

and to develop and use socially shared socio-cognitive categories and pertinent everyday vocabularies.

Such contrast can shed a new light on the human tendencies to overestimate the magnitude of minor variations between individuals for differentiating highly similar individuals and to homogenise within-group differences and to exaggerate between-group differences for differentiating individuals from different groups and communities—and species. A better understanding of these tendencies can help to explore the mechanisms and processes underlying stereotypical beliefs and biases, to investigate their impact on assessments of individuals and to develop methodologies that enable researchers to minimise such biases in empirical comparisons of individuals.

Scientific quantifications of observable individual behaviours could also provide important evidence supporting the findings from cross-cultural psychology that "it is almost as if the full range of human diversity reproduces itself at every level of organization" (Shweder & Sullivan, 1990, p. 412). Exploring this diversity will be important for exploring real world issues in the increasingly pluralistic societies of our times.

Acknowledgements

284

References

Adams, D. K. & Zener, K. E. (1935). Translators' preface. In K. Lewin (1935). A dynamic theory of personality. Selected papers. New York and London: McGraw Hill.

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.

Altmann, J. (1974). Observational study of behavior: Sampling methods. Behaviour, 49, 227-267.

Arro, G. (2013). Peeking into personality test answers: Inter- and intraindividual variety in item interpretations. Integrative Psychological and Behavioral Science, 47, 56-76.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge, UK: Cambridge University Press.

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.

Bergman, L. R. & Trost, K. (2006). The person-oriented versus the variable-oriented approach: Are they complementary, opposites, or exploring different worlds? Merrill-Palmer Quarterly, 52, 377-389.

Biesanz, J. C., & Human, L. J. (2010). The cost of forming more accurate impressions: Accuracy motivated perceivers see the personality of others more distinctively but less normatively. Psychological Science, 24, 589-594.

Blaffer-Hrdy, S. (2009). Mothers and others: the evolutionary origins of mutual understanding. Cambridge MA: Belknap Press.

Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. Psychological Inquiry, 21, 2-25.

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 & Nicholson.

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.

Boesch, C., Hohmann, G. & Marchant, L. (2002). Behavioural diversity in chimpanzees and bonobos. Cambridge: Cambridge University Press.

Buss, D. M. (1999). Evolutionary psychology: The new science of the mind. Boston: Allyn & Bacon.

Buss, D. M. & Craik, K. H. (1983). The act frequency approach to personality. Psychological Review, 90, 105-126.

Byrne, B. M. (1994). Structural equation modelling with EQS and EQS/Windows: Basic concepts, applications, and programming. Thousand Oaks, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychological Bulletin, 105, 456-466.

Campbell, N. A., & Reece, J. B. (2005). Biology (7ᵗʰ Ed.). New York, NY: Benjamin Cummings.

Capitanio, J. P. (2004). Personality factors between and within species. In B. Terry, M. Singh & W. Kaumanns (Eds.) Macaque societies (pp. 13–33). Cambridge, UK: Cambridge University Press.

Cattell, R. B. (1946). The description and measurement of personality. New York: World Book.

Cervone, D., Shadel, W. G., & Jencius, S. (2001). Social-cognitive theory of personality assessment. Personality and Social Psychology Review, 5, 33-51.

Clarke, A. S., & Boinski, S. (1995). Temperament in nonhuman primates. American Journal of Primatology, 37, 103-125.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2ⁿᵈ ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conner, T. S., Barrett, L. F., Tugade, M. M. & Tennen, H. (2007). Idiographic personality: The theory and practice of experience sampling. In R. W. Robins, R. C. Fraley, & R. Krueger (Eds.), Handbook of Research Methods in Personality Psychology (pp. 79 – 98). New York, NY: Guilford Press.

Connolly, P. (1998). Racism, gender identities and young children: Social relations in a multi-ethnic, inner-city primary school. London, UK: Routledge.

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

Darwin, C. R. (1859). On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London, UK: John Murray.

de Weerth, C., van Geert, P., & Hoitink, H. (1999). Intra-individual variability in infant behavior. Developmental Psychology, 35, 1102–1112. cite

Diriwächter, R., Valsiner, J., & Sauck, C. (2004). Microgenesis in making sense of oneself: Constructive recycling of personality inventory items. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 6, Art. 11.

Dunbar, R. (1996). Grooming, gossip, and the evolution of language. Cambridge, MA: Harvard University Press.

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

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

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.

Fleeson, W. (2001). Towards a structure- and process-integrated view of personality: Traits as density distributions of states. Journal of Personality and Social Psychology, 80, 1011-1027.

Fragaszy, D. M., Visalberghi, E., & Fedigan, L. (2004). The complete capuchin. Cambridge: Cambridge University Press.

Gigerenzer, G., Todd, P., & the ABC Group. (1999). Simple heuristics that make us smart. New York, NY: Oxford University Press.

Gillespie, A., Howarth, C., & Cornish, F. (2012). Four problems for researchers using social categories. *Culture and Psychology, 18*, 391-402.

Goldberg, L. R. (1982). 'From Ace to Zombie: Some explorations in the language of personality'. In C. D. Spielberger & J. N. Butcher (Eds), *Advances in Personality Assessment*, Vol. 1, pp. 203-234, Erlbaum, Hillsdale.

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

Hammock, E. A., & Young, L. Y. (2005). Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science, 308*, 1630-1634.

Hartmann, N. (1964). *Der Aufbau der realen Welt. Grundriss der allgemeinen Kategorienlehre* (3. Aufl.). Berlin: Walter de Gruyter.

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.

Herrmann, E., Hare, B., Cissewski, J., & Tomasello, M. (2011). A comparison of temperament in nonhuman apes and human infants. *Developmental Science, 14*, 1393–1405.

Hinde, R. A. (1974). *Biological bases of human social behavior*. New York: McGraw-Hill.

Hinde, R. A. (1982). *Ethology: Its nature and relations with other sciences*. Oxford: Oxford University Press.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. California, CA.: Sage Publications.

Howarth, C. (2009). "I hope we won't have to understand racism one day": Researching or reproducing 'race' in Social research? *British Journal of Social Psychology, 48*, 407-426.

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.

Jovchelovitch, S. (2007). *Knowledge in context: Representations, community and culture*. London, UK: Routledge.

Kaplan, R. M., & Saccuzzo, D. P. (2008). *Psychological testing: Principles, applications, and issues*, (7th ed.). Belmont, CA: Thompson. Wadsworth.

King, J. E., Weiss, A. & Sisco, M. M. (2008). Aping humans: Age and sex effects in chimpanzee (Pan troglodytes) and human (Homo sapiens) personality. *Journal of Comparative Psychology, 122*, 418-27.

Kluckhohn, C., Murray, H. A., & Schneider, D. M. (1953). Personality in nature, society, and culture. New York, NY: Knopf.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York, NY: Harcourt, Brace, & World.

Lahlou, S. (1996). Propagation of social representations. *Journal for the Theory of Social Behaviour, 26*, 157–175.

Lahlou, S. (1998). *Penser-manger*. Paris, France: Les Presses Universitaires de France.

Lahlou, S. (2001). Functional aspects of social representations. In K. Deaux & G. Philogene (Eds.), *Representations of the social: Bridging theoretical traditions* (pp.131–146). Oxford: Blackwell.

Lazarus, A. A. (1981). *The practice of multimodal therapy : systematic, comprehensive, effective psychotherapy*. New York: McGraw-Hill.

Lazarus, R. S. (2000). Towards better research on stress and coping. *American Psychologist, 55*, 665-673.

Lehner, P. N. (1998). *Handbook of ethological methods*. Cambridge, UK: Cambridge University Press.

Leung, K., & Bond, M. H. (1989). On the empirical identification of dimensions for cross-cultural comparisons. *Journal of Cross Cultural Psychology, 20*, 133-151.

Lewin, K. (1936). *Principles of topological psychology*. New York, NY: McGraw-Hill.

Li, W., Moallem, I., Paller, K. A. & Gottfried, J. A. (2007). Subliminal smells can guide social preferences. *Psychological Science, 18*, 1044-1049.

Lloyd, B. & Duveen, G. (1992). *Gender identities and education: The impact of starting school*. London: Harvester-Wheatsheaf.

Martin, P. & Bateson. P. (2007) *Measuring behaviour: An introductory guide* (3rd ed). Cambridge: Cambridge University Press.

McAdams, D. P. (1985). *Power, intimacy, and the life story: Personological inquiries into identity*. New York, NY: Guilford.

McCrae, R. R., Terracciano, A., & 79 Members of the Personality Profiles of Cultures Project (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*, 407–425.

McCrae. R. R., Terracciano, A., Realo, A., & Allik, J. (2007a). Climatic warmth and national wealth: Some culture-level determinants of national character stereotypes. *European Journal of Personality, 21*, 953–976.

McCrae, R. R., Terracciano, A., Realo, A., & Allik, J. (2007b). On the validity of culture-level personality and stereotypes scores. *European Journal of Personality, 21*, 987–991.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science, 317*, 82.

Menon, G., & Yorkston, E. A. (2000). The use of memory and contextual cues in the formation of behavioral frequency judgments. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman & V. S. Cain (Eds.) *The science of self-report*, (pp. 63–79). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Michell, J. (2011). Qualitative research meets the ghost of Pythagoras. *Theory & Psychology, 21*, 241-259.

Michell, J., (2012). The constantly recurring argument: Inferring quantity from order. *Theory & Psychology, 22*, 255-271.

Millikan, R. (1993). *White queen psychology and other essays for Alice*. Bradford: MIT Press.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Mischel, W. (1977). The interaction of person and situation. In: D. Magnusson, & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. (pp. 333-352). Hillsdale, NJ: Erlbaum.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246-268.

Mischel, W., Shoda, Y., & Ayduk, O. (2007). Introduction to personality: Toward an integrative science of the person (8th ed.). Hoboken, NJ: J. Wiley & Sons.

Mischel, W., Shoda, Y., & Mendoza-Denton, R. (2002). Situation-behavior profiles as a locus of consistency in personality. Current Directions in Psychological Science, 11, 50–54.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. Measurement: Interdisciplinary Research and Perspectives, 2, 201–218.

Murray-Close, D., & Ostrov, J. M. (2009). A longitudinal study of forms and functions of aggressive behavior in early childhood. Child Development, 80, 828–842.

Ogden, C. K. (1932). Bentham's theory of fictions. New York: Harcourt Brace.

Omi, Y. (2012). Tension between the theoretical thinking and the empirical method: Is it an inevitable fate for psychology? Integrative Psychological and Behavioral Science, 46, 118–127.

Otovic, P., Partan, S. R., Bryant, J. B., & Hutchinson, E. (2014). Let's call a truce … for now: The silent bared-teeth face expression in mandrills (Mandrillus sphinx) during baseline and post-conflict conditions. Ethology, 120, 1118–1127.

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

Perugini, M. & Richetin, J. (2007). In the land of the blind, the one-eyed man is king. European Journal of Personality, 21, 977–981.

Piccione, G., Marafioti, S., Giannetto,C., Di Pietro, S., Quartuccio, M., & Fazio, F. (2014). Comparison of daily distribution of rest/activity in companion cats and dogs. Biological Rhythm Research, 45, 615–623.

Pollet, T. V., Tybur, J. M. Frankenhuis, W. E. F., & Rickard, I. J. (2014). What can cross-cultural correlations teach us about human nature? Human Nature, 25, 410–429.

Preuschoft, S. & van Hooff, J. A. R. A. M. (1995). Homologizing primate facial displays: A critical review of methods. Folia Primatologica, 65, 121–137.

Preuschoft, S. (1992). "Laughter" and "smile" in Barbary macaques (Macaca sylvanus). Ethology, 91, 200–236.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. American Sociological Review, 15, 351–357.

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

Rothschuh, K. E. (1963). Theorie des Organsimus. Bios – Psyche – Pathos (2. erw. Aufl.). München: Urban & Schwarzenberg.

Rotter, J. B. (1954). Social learning and clinical psychology. New York: Prentice-Hall.

Rotter, J. B. (1981). The psychological situation in social learning theory. In D. Magnusson (Ed.), Toward a psychology of situations: An interactional perspective. Hillsdale, NJ: Lawrence Erlbaum.

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

Schacter, D. L. & Addis, D. R. (2007). Constructive memory: Ghosts of past and future. Nature, 445, 27.

Schneider, D. J. (2005). The psychology of stereotyping. New York: Guilford Press.

Schwarz, N. (1999). Self-reports: How the questions shape the answer. American Psychologist, 54, 93–105.

Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intra-individual stability in the organization and patterning of behavior: Incorporating psychological situations into the idiographic analysis of personality. Journal of Personality and Social Psychology, 67, 674–687.

Shweder, R. A. & D'Andrade, R. G. (1980). The systematic distortion hypothesis. In R. A. Shweder (Ed.). Fallible judgment in behavioral research: New directions for methodology of social and behavioral science (Vol. 4, pp. 37–58). San Francisco: Jossey-Bass.

Shweder, R. A. & Sullivan, M. A. (1990). The semiotic subject of cultural psychology. In L. A. Pervin (Ed.), Handbook of personality, (pp. 399 - 416). New York, NY: Guilford.

Skinner, B. F. (1957). Verbal behavior. Englewood Cliffs, NJ: Prentice-Hall.

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.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big-Five domains and facets in a large cross-sectional sample. Journal of Personality and Social Psychology, 100, 330-348.

Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N. (2009). Revisiting Robinson: The perils of individualistic and ecologic fallacy. International Journal of Epidemiology, 38, 342–360.

Sussman, A.F., Ha, J.C., Bentson, K.L., & Crockett, C. M. (2013) Temperament in rhesus, long-tailed, and pigtailed macaques varies by species and sex. American Journal of Primatology, 75, 303–313.

Terracciano, A., Abdel-Khalek, A. M., Ádám, N., Adamovová, L., et al. (2005). National character does not reflect mean personality trait levels in 49 cultures. Science, 310, 96–100.

Tetley, C. L. & O'Hara, S. J. (2012) Ratings of animal personality as a tool for improving the breeding, management and welfare of zoo mammals. Animal Welfare, 21, 463-476.

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. Journal of Research in Personality, 34, 397-423.

Thelen, E., & Smith, L. B. (1993). A dynamic systems approach to the development of cognition and action. Cambridge, MA: MIT Press.

Thierry, B. (2000). Covariation of conflict management patterns across macaque species. In F. Aureli & F. B. M. de Waal (Eds). Natural conflict resolution, (pp. 106-128). Berkeley: University of California Press.

Thierry, B. (2007). Unity in diversity: Lessons from macaque societies. Evolutionary Anthropology, 16, 224-238.

Thorndike, E. L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. American Journal of Psychology, 52, 122-124.

Tomasello, M. (1999). The Cultural Origins of Human Cognition. Harvard University Press.

Toomela, A. (2010). Modern mainstream psychology is the best? – Noncumulative, historically blind, fragmented, atheoretical. In A. Toomela & J. Valsiner, J. (Eds.). Methodological thinking in psychology: 60 years gone astray? (1 - 26). Charlotte, NC: Information Age Publishers.

Toomela, A. (2011). Travel into a fairy land: A critique of modern qualitative and mixed methods psychologies. Integrative Psychological and Behavioral Science, 45, 21-47.

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

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

Uher, J. (2011a). Individual behavioral phenotypes: An integrative metatheoretical 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. (2014). Fundamental challenges of contemporary "personality" research. Physics of Life Reviews, 11, 695-696.

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(3). DOI: 10.1007/s12124-014-9283-1.

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(3). DOI: 10.1007/s12124-014-9280-4.

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(3). DOI: 10.1007/s12124-014-9281-3

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. (2016). Exploring the workings of the psyche: Metatheoretical and methodological foundations. Annals of Theoretical Psychology, 13, xx-xx.

Uher, J. (submitted). 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 behavior 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.

van de Vijver, F. J. R. & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13, 29-37.

van de Vijver, F. J. R., & Leung, K. (1997a). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), Handbook of cross-cultural psychology (2nd ed., Vol. 1, pp. 257-300). Boston: Allyn & Bacon.

van de Vijver, F. J. R., & Leung, K. (1997b). Methods and data analysis for cross-cultural research. Newbury Park, CA: Sage.

van de Vijver, F. J. R., & Poortinga, Y. H. (2002). Structural Equivalence in Multilevel Research. Journal of Cross-Cultural Psychology, 141, 141-156.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 54, 119-135.

van Geert, P. & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. Infant Behavior & Development, 25, 340-374.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3, 4–70.

Visalberghi E., Valenzano, D.R., & Preuschoft S. (2006). Facial displays in tufted capuchins. International Journal of Primatology, 27, 1689-1707.

von Uexküll, J. (1909). Umwelt und Innenwelt der Tiere. Berlin: Springer.

Vygotsky, L. S. (1934/1962). Thought and language. Cambridge, MA: MIT Press.

Whitehead, A. N. (1929). Process and reality. New York: Harper.

Wright, J. C. & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. Journal of Personality and Social Psychology, 55, 454-469.

Wright, J. C., & Zakriski, A. L. (2003). When syndromal similarity obscures functional dissimilarity: Distinctive evoked environments of externalizing and mixed syndrome children. Journal of Consulting and Clinical Psychology, 71, 516-527.