European Journal of Personality
Eur. J. Pers. **23**: 403–435 (2009)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/per.725

Discussion on 'Personality Psychology as a Truly Behavioural Science' by R. Michael Furr

OPEN PEER COMMENTARY

Yes We Can! A Plea for Direct Behavioural Observation in Personality Research

MITJA D. BACK and BORIS EGLOFF
Department of Psychology, Johannes Gutenberg University Mainz, Germany
mback@uni-leipzig.de

Abstract

Furr's target paper (this issue) is thought to enhance the standing of personality psychology as a truly behavioural science. We wholeheartedly agree with this goal. In our comment we argue for more specific and ambitious requirements for behavioural personality research. Specifically, we show why behaviour should be observed directly. Moreover, we illustratively describe potentially interesting approaches in behavioural personality research: lens model analyses, the observation of multiple behaviours in diverse experimentally created situations and the observation of behaviour in real life.

Furr (this issue) does an excellent job in illuminating the fundamental importance of behavioural personality research and in describing the methods and prevalence of such studies. It becomes clear that personality psychology should study actual behaviour but that this is still rarely done. In our comment, we argue for a broader definition of behaviour, and a stricter and more ambitious understanding of *behavioural* measurement. Moreover, we describe how direct observations of behaviour can fruitfully be used for better understanding the relations between personality and actual behaviour.

Furr (this issue) defines behaviour as 'verbal utterances or movements that are potentially available to careful observers using normal sensory processes.' This definition has immediate appeal and comes very close to what lay people would subsume under the term 'behaviour.' However, we think that there are some important behavioural phenomena that are not captured by this definition, including behaviours more strongly influenced by physiological processes (e.g. blushing), written verbal content

Received 6 March 2009
Accepted 6 March 2009

(e.g. percentage of negative words in self-descriptions), as well as observable behavioural results (e.g. punctuality; Back, Schmukle, & Egloff, 2006).

Our basic concern with this target paper refers to Furr's taxonomy of types of behavioural data. Furr first describes the strengths and weaknesses of nine methods that produce behavioural data. He then concludes that there are three equally strong measures of actual behaviour: direct behavioural observation, experience sampling self-reports of current or recent behaviour and acquaintance reports of recent behaviour. In contrast, we think that only direct observations of behaviour measure actual behaviour. Of the remaining eight measures, four produce self-concept data and four produce interpersonal perception data (see Figure 1).

We agree that experience sampling (as a self-concept measure) and acquaintance reports of recent behaviours (as a measure of interpersonal perception) are the measures most strongly related to behaviour. However, they should not be equated with behavioural measures themselves (Gosling, John, Craik, & Robins, 1998; Vazire & Mehl, 2008).

The problem of measuring behaviour via self- or acquaintance reports becomes obvious when analysing personality-behaviour relations (e.g. Do extraverts smile more?) or behaviour-interpersonal perception relations (e.g. Are smilers liked more?). Self-reports of behaviour (e.g. I smiled a lot) and acquaintance reports of behaviour (e.g. S/he smiled a lot) contain a substantial amount of self-concept and interpersonal perception variance. When correlating these measures with personality self- or acquaintance reports (e.g. I am a happy person; S/he is a happy person) or interpersonal perceptions (e.g. I like her/him), results can be distorted dramatically. In many cases such an approach will result in overestimations of effects due to shared method variance. In our view, the only solution is to directly observe how people actually behave (e.g. counting the smiles and relating this directly observed behaviour to variables of interest).

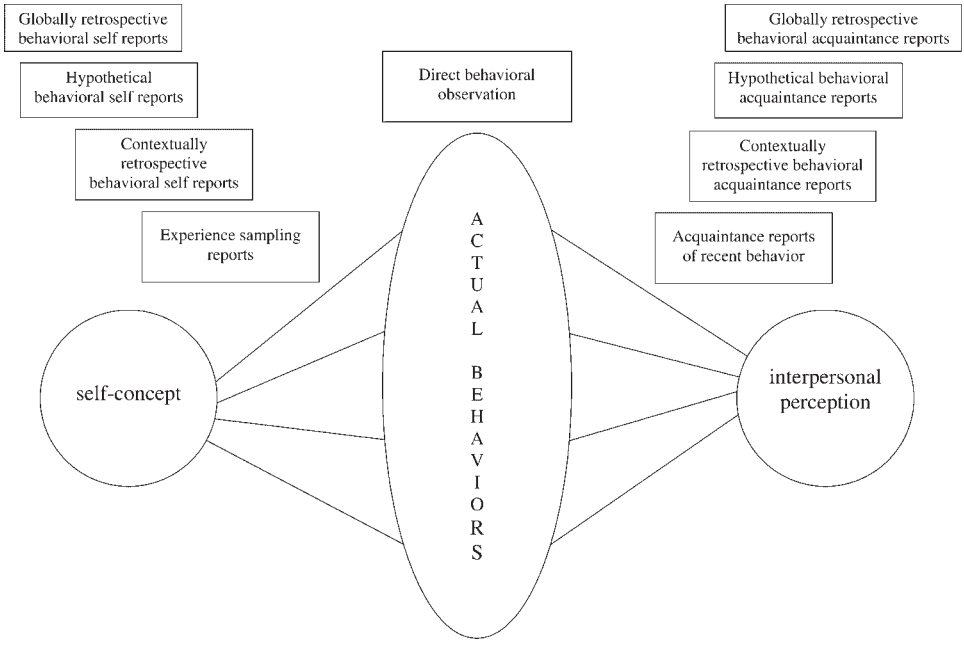

Figure 1. Using the lens model to describe nine methods of studying behaviour in personality research.

DOI: 10.1002/per

But what about the weaknesses of direct behavioural observation? What Furr describes as weaknesses are not inherent features of direct behavioural observation. He simply, and insightfully, describes the procedural burden that it takes to actually observe behaviour systematically: one needs time, effort, money, carefully selected experimental or real-life situations and extensively trained observers (in conjunction with specialized observational technology).

We now want to illustratively describe three approaches that we regard as interesting prospects for behavioural personality research: lens-modelling, multiple behaviours and real-life contexts. The lens-model (Brunswik, 1956) is an ideal framework for studying behaviour (see Figure 1). It incorporates determinants of behaviour (self-concept) as well as consequences of behaviour (interpersonal perception). In doing so, it bridges the traditional gap between personality research and social psychology (Funder, 1999). As an example, we applied a combination of the lens model and the social relations model (Kenny, 1994) to examine the interplay of personality, actual behaviour and attraction at zero acquaintance: a group of psychology freshmen was investigated upon encountering one another for the first time. Personality traits and attraction ratings were assessed using a round-robin design. The direct observation of more than 60 behaviours and physical appearances based on short videotaped self-introductions allowed us to explain the influence of personality on popularity and mutual liking at first sight (Back, Schmukle, & Egloff, 2008a).

Moreover, according to our view, future behavioural personality research would benefit from studying multiple theoretically derived behaviours, including a variety of relevant social situations. For example, in a recent study, we reliably measured more than 50 behavioural indicators that were *a priori* assigned to the Big Five dimensions based on theory and expert ratings. This data set was successfully used to validate explicit and implicit measures of the self-concept of personality within a behavioural process model of personality (Back, Schmukle, & Egloff, in press).

Besides carefully conducting laboratory behavioural studies, researchers could directly observe behaviour in real-life. In groundbreaking recent studies, Matthias Mehl and colleagues creatively used latest technology to unobtrusively observe actual behaviour in natural settings (Mehl, Gosling, & Pennebaker, 2006; Mehl & Pennebaker, 2003). Additionally, people's natural and virtual environments are a rich source for behavioural personality research (Back, Schmukle, & Egloff, 2008b; Back, Stopfer et al., 2008; Gosling, Ko, Mannarelli, & Morris, 2002; Vazire & Gosling, 2004).

Unfortunately, behavioural personality research of this kind is still rare: according to the numbers presented by Furr (this issue), only 5% of the published articles in the leading journals of our field contain behavioural data. The reason for this suboptimal situation is easy to understand: directly observing behaviour is hard work. Moreover, due to the multi-determination of actual behaviour and the lack of shared method variance, personality-behaviour correlations will be much lower when compared to studies in which behaviour is measured by self- or acquaintance reports. We nevertheless encourage researchers to conduct systematic and realistic behavioural studies using a variety of relevant social situations that allow for the measurement of multiple behavioural indicators. Although such studies are painful for researchers, they are rewarding in the long run because they bring us closer to the key mission of psychology: understanding the determinants and consequences of what people actually do. In sum, yes we—as personality researchers—should, and yes we can observe actual behaviour: Just do it!

DOI: 10.1002/per

The Categories of Behaviour Should be Clearly Defined

PETER BORKENAU
Department of Psychology, Martin-Luther University Halle-Wittenberg, Germany
p.borkenau@psych.uni-halle.de

Abstract

The target paper is helpful by clarifying the terminology as well as the strengths and weaknesses of several approaches to collect behavioural data. Insufficiently considered, however, is the clarity of the categories being used for the coding of behaviour. Evidence is reported showing that interjudge agreement for retrospective and even concurrent codings of behaviour does not exeeced interjudge agreement for personality traits if the categories being used for the coding of behaviour are not clearly defined. By contrast, if the behaviour to be registered is unambiguously defined, interjudge agreement may be almost perfect. Copyright © 2009 John Wiley & Sons, Ltd.

Personality psychology has been criticized for relying too strongly on retrospective decontextualized self-reports and for paying insufficient attention to the research participants' actual behaviour (Mischel, 1968). Furr (this issue) gives reasons why the study of behaviour is important for personality research, suggests a definition of behaviour and provides an overview and evaluation of different ways to study it. Although it is not particularly new that personality should be studied by multiple methods, including behaviour observations, the target paper is helpful by clarifying the terminology as well as the strengths and weaknesses of several approaches to collect such data. Particular care is given to distinguishing between types of informants (self, acquaintance, independent observer), the time lag between observation and description (current/recent versus retrospective) and to contextualized versus decontextualized behaviour descriptions.

What is hardly addressed, however, is the level of abstraction of the categories used for the coding of behaviour that is linked to the extent of inferences required by the judges and the interjudge agreement that might be expected. For example, Furr makes a sharp distinction between the Riverside behavioural Q-sort (Funder, Furr, & Colvin, 2000) item 'Exhibits an awkward interpersonal style' being classified as a behaviour description, and the statement 'Is a friendly person' being classified as a dispositional inference. This is reasonable as far as the first statement describes an attribute of a person's behaviour, whereas the second statement describes an attribute of a person. But in many research contexts, the judges can hardly distinguish between behaviour and personality like the friendliness of behaviour and friendliness as a trait. This is because judges have to infer the personality trait from a sample of observed behaviours (Kenny, 2006). Thus behaviour descriptions and trait inferences that rely on the same behaviour observations are likely to be quite similar, as will be the agreement among the judges.

I am going to illustrate that with some data published quite a time ago. Different from the previous publications using these datasets; however, all correlations reported here refer to the agreement between individual judges (instead of the reliability of the average rating by multiple judges), as the number of judges varied between tasks. Borkenau and Ostendorf (1987) videotaped eight 45-minute six-person group discussions on, for example, speed limits on German highways. The behaviour of the 48 participants was then coded in four different

DOI: 10.1002/per

ways, using concurrent as well as retrospective coding schemes. To obtain concurrent codings, the videotapes were subdivided into 15-second segments and 3696 verbal activities were identified. After having observed a segment, two judges assigned the behaviour of the verbally active participants to one of 16 categories, including *supports*, *directs the discussion*, *criticizes* and *ridicules* (see Borkenau & Ostendorf, 1987). Interjudge agreement for these forced-choice concurrent codings was $\kappa = .30$. Two other judges provided somewhat different concurrent codings: they successively rated the prototypicality of each identified activity for each of the 16 categories of behaviour, watching each video 16 times. For example, they rated whether a particular activity was a good or a poor example of supporting behaviour. Here, the correlations between individual judges varied from $r = .11$ (*ridicules*) to $r = .53$ (*criticizes*) with an average of $r = .42$. Thus for both concurrent coding schemes, interjudge agreement was modest.

In addition, the same videotapes were coded retrospectively. Here, the judges watched each 45-minute group discussion in its entirety and then either: (a) provided rankings of the six participants on who had shown a given behaviour most frequently, second most frequently, etc. or (b) provided unconstrained estimates on how frequently each participant had shown the 16 varieties of behaviour. Interjudge agreement for the retrospective frequency rankings varied from $r = .29$ (*supports*) to $r = .64$ (*directs the discussion*) with an average of $r = .47$, whereas interjudge agreement for the retrospective frequency estimates varied from $r = .18$ (*ridicules*) to $r = .51$ (*directs the discussion*) with an average of $r = .28$. Thus interjudge agreement for retrospective ratings was not systematically lower than interjudge agreement for the on-line codings. Moreover, all interjudge agreement coefficients varied substantially between the 16 categories of behaviour.

In another study (Borkenau & Müller, 1992), the videos of four of the eight discussions were reanalysed, but now the judges provided ratings of the participants' personality traits after having watched a discussion in its entirety. These ratings were provided on eight unipolar (e.g. *dominant*) and seven bipolar (e.g. *dominant-submissive*) scales. Here the agreement between individual judges varied from $r = .02$ to $r = .48$ with an average of $r = .29$, that is, it resembled the agreement for the forced-choice concurrent codings of behaviour and the retrospective act frequency estimates in the earlier study. Thus interjudge agreement was not substantially higher for categories of behaviour than for personality traits. Moreover, it varied substantially between traits but not between unipolar and bipolar rating scales.

Obviously, codings of clearly defined behaviours may be highly objective. An example is the number of spoken words counted in the study by Mehl, Vazire, Ramirez-Esparza, Slatcher, and Pennebaker (2007). According to the supporting online material for that paper, interjudge agreement concerning the number of words spoken was .99. That such a high interjudge agreement was obtained; however, does not imply that high objectivity may be obtained for descriptions of behaviour exclusively. Rather, that study is on the talkativeness of women and men, that is, on gender differences in a personality trait. The authors implicitly assume (and I agree) that many words spoken (a behaviour) and high talkativeness (a personality trait) are undistinguishable, if the coded behaviour was observed in a representative sample of situations. Indeed, representative sampling of situations is the main strength of Mehl et al.'s (2007) study.

Thus the objectivity of behavioural data depends crucially on how clearly the categories used for the coding of behaviour are defined. For instance, what constitutes an awkward interpersonal style is not well defined, in contrast to what constitutes the number of words spoken in a particular situation. Without clear definitions of the behaviour categories being used, interjudge agreement is likely to be modest even if the behaviour is real, contextualized, directly observed and coded concurrently. If the behaviour categories lack precision, interjudge agreement is unlikely to exceed interjudge agreement for personality traits.

DOI: 10.1002/per

This does not at all imply that careful and systematic observations of representative samples of ongoing behaviour are not worth the effort. But they will result in strong data only if the categories used for the coding of behaviour are precise.

Behaviour Functions in Personality Psychology

PHILIP J. CORR

Department of Psychology, Faculty of Social Sciences, University of East Anglia, Norwich, UK

Philip.Corr@btopenworld.com

Abstract

Furr's target paper highlights the importance, yet under-representation, of behaviour in published articles in personality psychology. Whilst agreeing with most of his points, I remain unclear as to how behaviour (as specifically defined by Furr) relates to other forms of psychological data (e.g. cognitive task performance). In addition, it is not clear how the functions of behaviour are to be decided: different behaviours may serve the same function; and identical behaviours may serve different functions. To clarify these points, methodological and theoretical aspects of Furr's proposal would benefit from delineation.

As a journal editor and reviewer, I am heartened to receive manuscripts reporting personality studies that contain some overt form of behaviour, either as target dependent variables or in the form of manipulation of independent variables. There is a certain directness about these studies and appealing face validity (e.g. time delay to knock on a professor's office door as a function of social anxiety). The fact that restrictions are not placed on the subject's behaviour, and they are behaving naturally, often in naturalistic settings adds theoretical creditability. Furr makes a strong case that behaviour should be awarded a more prime position in personality psychology, although researchers may not entirely agree with his definition of behaviour as 'verbal utterances or movements that are potentially available to careful observers using normal sensory processes.' However, this definition is parsimonious, plausible and potentially useful in arriving at a consensus as to what constitutes 'behaviour.'

Furr does an excellent job of providing a viable taxonomy (target article Table 2) of the ways of measuring overt behaviour. His careful delineation of the pros and cons of the different methods of behaviour measurement should prove valuable. I agree that Furr has '. . . offered as a starting point for focussed discussion of these important issues, potentially enhancing the field's standing as a truly behavioural science.'

To some extent, there is still a tension between behavioural approaches and the concerns of the personality psychologist, who frequently is interested in internal states of emotion, motivation, concepts, etc. very often couched in terms of traits conceived as internally organized structures capable of agency. There are (I would guess) few of us who would want to abandon these internal state/trait variables in favour of an exclusive focus on observable behaviour. To many of us, these are vital explanatory concepts—indeed, these internal processes are often demanded by careful analysis of behaviour itself (e.g. the concept of frustration as explaining resistance to extinction on partial reinforcement schedules; or behaviourally silent learning as seen in sensory preconditioning).

DOI: 10.1002/per

One major attraction of considering behaviour as the central unit of analysis is its 'down-stream' nature: it embodies the accumulation of up-stream causal influences that manifest in ways that impact on the environment, from which the organism receives feedback in the form of reinforcement, etc. On evolutionary grounds alone, this is a compelling point: selection does not work directly on thoughts, feelings, etc. but on their behavioural consequences in the real world. It is this level of analysis that really matters. For this reason, geneticists often use overt behaviour in preference to specific cognitive tasks or endophenotypes. One eminent behavioural geneticist told me that he prefers to measure actual behaviour because many experimental tasks in psychology are 'hokey'—if they have reliability and validity, often they have little power of generalization or real-world application.

Furr's paper raises a number of issues that may benefit from clarification. Furr's focus is on overt behaviour, but it is unclear (at least to me) how this level of analysis should relate to others (e.g. computerized cognitive performance). I agree that tapping the keyboard in some computerized cognitive task (e.g. Stroop test) is behaviour in only a trivial sense, but performance on the cognitive task may be far from trivial—the results of which may throw light upon the causes (e.g. weakened inhibitory control) of the carefully observed overt behaviour (e.g. impulsivity) that meets Furr's criteria. Assuming that we have agreed upon a theoretically coherent and consensual model of behaviour, then we would surely be drawn back to the question of the meaning of such behaviour, entailing consideration of underlying cognitive, emotional and motivational antecedents? Furthermore, behaviour often contains the conflation of multiple separate causal influences and it is difficult to decompose these sources of influence: causes are difficult to infer from effects, especially multiply interacting ones. The unscrambling egg problem.

In addition, two quite distinct behaviours (e.g. submissive vs. assertive behavioural styles) may have similar, or identical, meaning and causal antecedents (e.g. a specific motivation to elicit help from others); and similar (even identical) behaviours shown by two people (e.g. anger) may have different meanings and causal antecedents (e.g. defensive vs. predatory aggression). The inclusion of verbal utterances in the definition of behaviour adds a further complication, because this 'behaviour' contains a high potential for deception and manipulation especially in environments where there is high motivation to dissemble. Thus, we cannot take overt behaviour at face-value.

One way around these problems is to undertake rigorous manipulation of independent variables in order to isolate separate causal influences; but this requirement must be absent when observing naturally occurring behaviour where experimental manipulation and control are neither desirable nor possible. In addition, some causal influences are simply not 'available to careful observers using normal sensory processes,' especially those relating to internal processes (e.g. sensitivity to rewarding and punishing stimuli). Determining the meaning of overt behaviour is difficult, chiefly because behaviour itself does not encapsulate function. Interpretation of behaviour requires analysis, and this would (of necessity?) involve the collection of non-behavioural data (e.g. cognitive task performance and questionnaire data).

At the heart of these concerns is the following issue: is Furr's call for more rigorous classification and recoding of behaviour principally methodological (i.e. of improving the quality of data) or theoretical (i.e. improving understanding of personality psychology, over and above that achieved by enhanced methodology)? If the latter, given the problems identified above, how is this to be achieved in terms of uncovering primary and separable causal influences on behaviour?

DOI: 10.1002/per

On the Difference Between Experience-Sampling Self-Reports and Other Self-Reports

WILLIAM FLEESON
Department of Psychology, Wake Forest University, Winston-Salem, NC, USA
fleesonW@wfu.edu

Abstract

Furr's fair but evaluative consideration of the strengths and weaknesses of behavioural assessment methods is a great service to the field. As part of his consideration, Furr makes a subtle and sophisticated distinction between different self-report methods. It is easy to dismiss all self-reports as poor measures, because some are poor. In contrast, Furr points out that the immediacy of the self-reports of behaviour in experience-sampling make experience-sampling one of the three strongest methods for assessing behaviour. This comment supports his conclusion, by arguing that ESM greatly diminishes one the three major problems afflicting self-reports—lack of knowledge—and because direct observations also suffer from the other two major problems afflicting self-reports. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) has done a great service to the field by writing this difficult paper about the need for, definition of and assessment of behaviour in personality psychology. This is a great service, because, as he points out, the ambiguity and disagreement about studying behaviour are matched only by the exhortations and the need to study behaviour. His paper is also a service to the field because he has made himself vulnerable to quite a bit of potential disagreement. No matter how gentle, respectful and humble Furr is as a person, a paper of this sort can't help but be provocative, in the positive, inspirational sense.

Most importantly, Dr Furr performed this service with insight and sophistication. The centrepiece of his paper is its fair consideration of the strengths and weaknesses of each method. It is very important, yet also very difficult, to be both fair and evaluative at the same time. Oftentimes, evaluation degenerates into argumentation and advocacy; conversely, in a vain effort to maintain fairness, many individuals refrain from evaluation and apply simple rules, such as splitting disputes evenly. What is so difficult, and what takes time, is evaluating claims on their merits while applying standards fairly and evenly. Because Furr has done this in his paper, on a topic that cannot help but be controversial, he is to be applauded.

A good example of this sophistication is his subtle distinction between different types of self-report measures of behaviour. It is easy enough to dismiss self-report on principle: because self-report is a poor method in some cases, it is easy to conclude that self-report is a poor method in all cases (a generalization further facilitated by the fact that self-report does have limitations in all cases). However, Furr more carefully weighed the pros and cons of self-reports in each case independently. He concluded that one type of self-report, experience-sampling, is a strong method for measuring behaviour.

I would like to elaborate on this subtle distinction by highlighting one advantage of experience-sampling methodology (ESM) that drew me to it years ago. The best place to start is with the problems generated by self-report. For example, trait assessments ask participants to self-report what they are like 'in general,' e.g. how talkative they are in

DOI: 10.1002/per

general. Three main problems have been identified with such self reports (adapted from Paulhus & Vazire, 2007).

(1) *Bias.* There are a variety of potential biases that may lead to inaccurate ratings, including self-enhancement, acquiescence and extreme responding.
(2) *Subjectivity.* When individuals complete rating scales, several judgment processes are required, and judgment process may contaminate accurate assessment.
(3) *Lack of knowledge*. Finally, participants may not know the answer to the self-report questions, because of insufficient information or lack of introspective access.

Lack of knowledge may be especially problematic when making the 'in-general' ratings used to assess traits. There is no single behaviour or experience that determines a given response to an 'in-general' rating, meaning that respondents must somehow combine many behaviours and experiences together to arrive at their one, single response. The processes respondents go through, the information they consider and the rules they use to combine this information are all unknown. Klein et al. (e.g. Klein, Cosmides, Tooby, & Chance, 2001) suggest that typically people do not compute an answer at all, but rather read a ready-made answer off memory. Furthermore, to the extent that actual behaviour and experiences do play a role in such trait ratings, the probability of accurate recall of all relevant behaviour and experience from a period of years seems low. In sum, self-reports of what people are like in general require a process of unknown steps and validity, and require excellent comprehensive memory. Thus, scepticism about the relationship of these self-reports to what people actually do when placed in real situations with real pressures, is reasonable.

Lack of knowledge is also a problem for hypothetical self-reports. Hypothetical self-reports require participants to predict how they will act under various circumstances. People may not know how they will act, and may be surprised at how different situational pressures and forces become when the pressures are real rather than imagined.

The lack of knowledge individuals have about what they are like in general is one reason it is so important for personality psychologists to study what people actually do. What people report they are like in general could very well be a fiction, in part or even totally. In fact, the gist of one standing criticism of personality psychology is that it is only the study of self-concepts, because the relationship of in-general ratings to what people actually are like is so uncertain. Many psychologists have even concluded that the few times the issue has been investigated empirically, the result has been that in-general ratings do not reflect what people actually do when placed in actual situations (Ross & Nisbett, 1991).

Furr concludes that ESM is a stronger measure of behaviour because it is a different kind of self-report. Two problems afflicting self-report are still present in ESM. Biases, including self-enhancement, acquiescence and extreme responding biases, likely affect what people report they are doing in the moment. ESM also requires subjective judgments like all self-reports do. However, ESM greatly diminishes the third problem, lack of knowledge, which is so troubling for in-general ratings. ESM self-reports have very little reliance on memory, because the reported-on behaviour occurs concurrently or immediately prior to the report. ESM has very little need for a generalization process, because participants report on a particular or specific event. Thus, ESM bypasses most of the shaky reliance on self-concepts or on unknown, mysterious generalizing processes: participants report what is happening, now, right in front of them. To be clear, participants still do not have perfect knowledge—my claim is only that participants have much more

DOI: 10.1002/per

knowledge than they have for in-general ratings. This improvement is a huge advantage over in-general ratings.

My argument does not deny the other two problems with ESM self-reports (bias and subjectivity). But, as Furr pointed out, no method is perfect, because behaviour is a latent construct. In fact, direct observations of behaviour have more or less equivalent problems with these two problems of bias and subjectivity. Observers typically do not suffer from self-enhancement bias, but they introduce their own biases less troublesome in self-reports, most notably stereotype biases. More importantly, observers lack the contextualizing information of intention and habit that are critical for disambiguating behaviour (Funder, 1991; Paulhus & Vazire, 2007). For example, whether a gift was a friendly act depends to some degree on whether the gift was meant to be generous, was meant to curry favour, or was meant for some other purpose. This ambiguity forces observers to rely on superficial interpretations of actions, creating a potentially important superficiality bias. Observers also suffer from the subjectivity problem because their reports require the same level of judgments that self-reports require.

Thus, I believe that Furr has good reason to conclude that ESM is one of the three strongest methods for assessing behaviour. The use of ESM (1) greatly diminishes one of the three major problems with self-report, namely, lack of knowledge due to retrospection, generalization or speculation and (2) may not suffer from the remaining two problems of self-report, namely bias and subjectivity, to a greater extent than direct observer reports suffer.

What and Where is 'Behaviour' in Personality Psychology?

LAURA A. KING and JASON TRENT
Department of Psychology, University of Missouri, Columbia, USA
kingla@missouri.edu

Abstract

Furr is to be lauded for presenting a coherent and persuasive case for the lack of behavioural data in personality psychology. While agreeing wholeheartedly that personality psychology could benefit from greater inclusion of behavioural variables, here we question two aspects of Furr's analysis, first his definition of behaviour and second, his evidence that behaviour is under-appreciated in personality psychology.

WHAT IS BEHAVIOUR IN PERSONALITY PSYCHOLOGY?

Furr's definition of behaviour is limited to an almost Skinnerian degree as 'verbal utterances or movements that are potentially available to careful observers using normal sensory processes.' Most introductory psychology texts define (not just personality but) psychology as the science of human behaviour, *broadly defined*, to include such

DOI: 10.1002/per

phenomena as dreams, thoughts, feelings, etc. Should personality psychology embrace a different and decidedly more limited definition?

By requiring that behaviour be observable using only typical sensory processes, Furr excludes a number of behaviours that are routinely dubbed behavioural. For instance, in neuroscience, reaction time data are considered behavioural data. Dismissing reaction times as behaviour seems to impose an arbitrary distinction. Essentially, very slow reaction times would seem to 'count' as behaviour, while exceptionally fast ones would not. The role of physiological responses in behaviour also adds ambiguity. Furr explicitly 'excludes external physiological responses such as blushing or sweating' but includes 'potentially non-intentional responses such as trembling or gaze aversion—psychical movements with potentially important social implications.' Yet, blushing and sweating can have social implications, as habitual 'blushers' and voluminous 'sweaters' would attest. The effort to delineate the phenomena that count and those that do not count as behaviours leads to a great deal of confusion but such confusion should not be taken as detracting from the important contribution of Furr's efforts.

Indeed, we can see why these distinctions were made by Furr. The type of behaviour that is missing from personality research is a particular kind, we might call it 'real' behaviour—things that people actually do and that can be observed. In this same spirit, we would add that many variables that have been typically considered 'life events' in the personality literature (e.g. getting married, divorced, having children, etc.) might instead be viewed as behavioural outcomes. Furthermore, including behaviours that are representative of the kinds of behaviours individual engage in daily life would expand the applicability of personality research to the daily lives of human beings. Such behaviours include activities such as shopping, eating, gossiping, etc. From our perspective, and we suspect Furr's as well, it is not simply that behaviour is missing from personality psychology but *real* behaviour. A call for the increased use of behavioural data in personality research should not be taken as simply encouragement to devise clever laboratory observations of artificial activities. In this regard, we applaud Furr's inclusion of experience sampling methodologies as representing potentially flawed but valuable aspects of behavioural data.

WHERE IS BEHAVIOUR IN PERSONALITY PSYCHOLOGY?

Like previous scholars before him, Furr examined the best journals in personality psychology. This choice places notable limits on what can be found. Personality encompasses many things, with behaviour being but one of these. Personality psychologists have long been interested in varied mental processes. One could argue that understanding these intrapersonal processes may ultimately inform more distal behavioural outcomes. Taken even further, one might note that the results of Furr's analysis would seem to indicate that research that does not readily implicate behaviour may, nevertheless, be valuable to the science of personality. As such, journals that are centered on personality may not be the sole place where the relations of personality to behaviour can be found.

Indeed, there may be behaviours which are themselves of central importance to researchers and such behaviours may be represented in various scholarly outlets, not nominally dedicated to personality. For instance, research on the relations of personality traits to health behaviours might be more likely to be found in journals dedicated to health

DOI: 10.1002/per

psychology than those dedicated to personality. Further, journals dedicated to alcohol use, addictive behaviour, disordered eating, and a host of other behaviours may include numerous papers on the relations of individual differences to these behaviours. Even if such research is not published in leading personality journals, it exists as a testament to the importance of personality characteristics to adaptive and maladaptive behaviours. Clearly, considering a wider pool of outlets might provide surprising results with regard to the appreciation of both behaviour and personality in psychology, more generally.

Even if the conclusion drawn is more circumscribed, that there appears to be a lack of appreciation for behaviour in the mainstream personality journals, the key issue that remains to be addressed is the consequence of the lack of behavioural data in personality psychology. Are there notable places in the literature where personality psychologists are substituting inferior self-report measures where superior behavioural data would be more appropriate? Are there important research questions that are not being asked because of the under-appreciation of behaviour? If these questions are answered in the affirmative, then the questions that demand attention are why this is the case and what might be done?

As Furr acknowledges, behavioural data present a risk for researchers, requiring great amounts of effort, time, and sometimes, money. With scholarly careers in the balance, it might seem imprudent to devote oneself to such a gamble when easier paths to publication exist. In this sense, the scholarly community must create a context in which such data are valued. If we, as personality psychologists, wish to see more behavioural research in our field, we need to represent these values as editors, reviewers and scholars. To be sure, such a commitment might render that 'easier path' a suddenly bumpy one. But to encourage scholars to incorporate actual behaviour into their research programmes, we must foster an intellectual climate that not only appreciates such data but demands it when appropriate.

Naturalistic Observation of Daily Behaviour in Personality Psychology

MATTHIAS R. MEHL

Department of Psychology, University of Arizona, Tucson, AZ, USA

mehl@email.arizona.edu

Abstract

This comment highlights naturalistic observation as a specific method within Furr's (this issue) cluster direct behavioural observation and discusses the Electronically Activated Recorder (EAR) as a naturalistic observation sampling method that can be used in relatively large, nomothetic studies. Naturalistic observation with a method such as the EAR can inform researchers' understanding of personality in its relationship to daily behaviour in two important ways. It can help calibrate personality effects against act-frequencies of real-world behaviour and provide ecological, behavioural personality criteria that are independent of self-report. Copyright © 2009 John Wiley & Sons, Ltd.

Furr's target paper (this issue) provides an excellent and sorely needed analysis of the nature, potentials and state-of-the-science of behavioural assessment in personality

DOI: 10.1002/per

psychology. This comment highlights *naturalistic observation* as a specific method within Furr's cluster *direct behavioural observation* and discusses ways in which it can help the field become 'a truly *behavioural* science.' Naturalistic observation is the observation of subjects in their natural habitat. Whereas the method is fairly common in neighbouring disciplines (e.g. anthropology, primatology) and areas (e.g. developmental psychology), it has a thin history in personality psychology (Barker & Wright, 1951; Craik, 2000).

Over the last 10 years, I have co-developed and validated the Electronically Activated Recorder or EAR (Mehl et al., 2001) as a naturalistic (acoustic) observation sampling method that can be used in relatively large, nomothetic studies. The EAR is a pocket-sized audio-recorder that periodically samples snippets of ambient sounds from people's momentary environments (e.g. 30 second every 12.5 minute). Participants carry the device around while going about their normal lives. That way, the EAR produces acoustic logs of their daily behaviours as they naturally occur over the course of a day.

In a series of studies, my colleagues and I have used the method to show that a broad spectrum of acoustically detectable behaviours can be assessed reliably and with low levels of reactivity from the sampled ambient sounds (Mehl & Holleran, 2007), show very large between-persons variability and good temporal stability (Mehl & Pennebaker, 2003) and have good convergent validity with theoretically related trait measures such as the Big Five (Mehl et al., 2006) and subclinical depression (Mehl, 2006).

Naturalistic observation with a method such as the EAR can inform researchers' understanding of personality in its relationship to daily behaviour in at least two ways.

NATURALISTIC OBSERVATION CAN HELP CALIBRATE PERSONALITY EFFECTS AGAINST ACT-FREQUENCIES OF REAL-WORLD BEHAVIOUR

The field has long defended itself against criticisms of the limited magnitude of personality effects. Even though two recent landmark studies went a long way to silence them (Ozer & Benet-Martinez, 2006; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007), the field still needs more effective ways to communicate the implications of its findings. In psychology, the vast majority of measures use arbitrary metrics. Calibrating effects based on arbitrary metrics against inherently meaningful, real-world referents is one way to come to a 'better understanding of the measures by which the phenomena with which we concern ourselves are gauged' (Sechrest, McKnight, & McKnight, 1996, p. 1068).

One advantage of the EAR is that its sound-file based behavioural codings can be readily converted into a metric that is non-arbitrary, intuitively meaningful and inherently real-world relevant. If the EAR captures a person talking in 40 out of 120 recordings, one can estimate that the person spent about a third of her time awake (or about 5 hours) talking. By linking EAR-derived act-frequencies of daily behaviour to individual differences, a better understanding of effect sizes can be obtained. For example, in one study conscientiousness correlated $r = -.29$ with EAR-assessed swearing and $r = .42$ with EAR-coded time in class (Mehl et al., 2006). Converted into a more meaningful metric, this suggests that individuals who marked a '4' on the 5-point conscientiousness scale, compared to those who marked a '2,' used profanity at less than half the rate (0.5 vs. 1.2%) and spent about three times as much time in class (11.9 vs. 4.1%).

DOI: 10.1002/per

Similarly, testing the myth that women are by a factor more verbose than men, we estimated based on six EAR studies that both men and women use about 16 000 words per day (Mehl et al., 2007). A sex difference of 546 words compared to a range of over 46 000 words between the least and most talkative individual (695 vs. 47 016) rendered significance testing close to meaningless—and speaks powerfully to the magnitude of individual differences. Thus, in facilitating an absolute metric in the measurement of daily behaviour, naturalistic observation can help 'benchmark' personality effects.

NATURALISTIC OBSERVATION CAN PROVIDE ECOLOGICAL, BEHAVIOURAL PERSONALITY CRITERIA THAT ARE INDEPENDENT OF SELF-REPORT

The 'criterion problem' is a vexing issue in the field. Generally, behavioural criteria are deemed preferable to those based on self- or informant reports. For assessing personality in the field, experience sampling has emerged as the best available proxy to behavioural observation (Spain, Eaton, & Funder, 2000). In cases where it is necessary to measure real-world, behavioural personality criteria independent of self-report, the EAR can help accomplish this.

For example, we tested the accuracy of self- and other-reports by comparing the predictive validity of participants' self-ratings of how much they engage in different daily behaviours to similar ratings obtained from people who knew the participants well. The frequency with which the EAR captured participants actually engaging in these behaviours served as 'impartial' accuracy criterion. Self- and other-ratings showed identical validity but also uniquely predicted certain behaviours (Vazire & Mehl, 2008). Importantly, to avoid giving one perspective an undue advantage, it was critical to minimize shared method variance with the two. The EAR-derived behaviour counts maximally accomplished this while at the same time preserving the study's ecological focus.

Similarly, responding to Terracciano et al.'s (2005) influential finding that national stereotypes have zero validity, Heine, Buchtel, and Norenzayan (2008) argued that 'comparing means on subjective Likert self-report scales is the most commonly used method for investigating cross-cultural differences, yet there are many methodological challenges associated with this approach' (p. 309). Following their solution to concentrate on behavioural trait markers, we compared Americans' and Mexicans' sociability in a binational EAR study (Ramírez-Esparza, Mehl, Álvarez Bermúdez, & Pennebaker, 2009). We found that although American participants reported being more sociable than their Mexican counterparts, they spent less time with others and had fewer social (i.e. non-instrumental) conversations. Intriguingly, whereas Americans rated themselves significantly *higher* than Mexicans on the item 'I see myself as a person who is talkative,' they spent in fact almost 10% *less* time talking (34.3 vs. 43.2%). Again, in providing ecological, behavioural personality criteria, naturalistic observation can help resolve important debates within the field.

To summarize, naturalistic observation clearly occupies a methodological niche; it is not for everyone and everything. It is highly labour-intensive and thus requires careful deliberation as to when it should be used instead of more economic methods. However, in providing arguably the strongest form of behavioural data, it can yield valuable findings that cannot be obtained otherwise and that way contribute to the field becoming a leading behavioural science.

DOI: 10.1002/per

Measuring Behaviour

D. S. MOSKOWITZ and JENNIFER J. RUSSELL
Department of Psychology, McGill University, Montreal, Canada
dsm@psych.mcgill.ca

Abstract

Furr (this issue) provides an illuminating comparison of the strengths and weaknesses of various methods for assessing behaviour. In the selection of a method for assessing behaviour, there should be a careful analysis of the definition of the behaviour and the purpose of assessment. This commentary clarifies and expands upon some points concerning the suitability of experience sampling measures, referred to as Intensive Repeated Measurements in Naturalistic Settings (IRM-NS). IRM-NS measures are particularly useful for constructing measures of differing levels of specificity or generality, for providing individual difference measures which can be associated with multiple layers of contextual variables, and for providing measures capable of reflecting variability and distributional features of behaviour. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) provides a detailed and illuminating comparison of various behavioural assessment methodologies. His careful analysis clearly delineates the strengths and weaknesses of several techniques and concludes that direct behavioural observation, experience sampling reports of current or recent behaviour, and acquaintances' reports of recent behaviours are 'strong' measures of behaviour. The subsequent comments are intended to clarify and expand upon some points relevant to the identification of strong measures and the characteristics of experience sampling measures.

Psychological testing involves the implicit or explicit measurement of some construct (Cronbach & Meehl, 1955), an abstract concept presumed to be indicated by specific, measurable behaviours. To determine whether a construct has been adequately assessed, the construct and the domain of behaviours corresponding to that construct must be carefully defined. Measurement tools are then evaluated with respect to this definition. For example, laboratory-based measures of behaviour (e.g. behaviour tests) usually provide information about a small number of behaviours and may therefore be inadequate to measure constructs referencing a wide range of behaviours.

It is also important to specify the purpose of measurement, including the conditions over which the measure should generalize. When a measure is intended to be specific to a narrowly circumscribed type of event, behaviour tests are suitable due to their specificity. For measures that are intended to generalize across occasions and situations, questionnaires might be a strong method because they permit the broad sampling of situations, occasions and various behaviours that refer to a construct. Some evidence suggests that while recent or current reports provide more accurate assessments of objective behaviour, global retrospective measures might be more predictive of future behavioural choices (Wirtz, Druger, Scollon, & Diener, 2003). Thus, different methods may be 'strong' for different constructs, for different purposes or to address different research questions.

The method referred to as experience sampling by Furr, alternatively identified as ecological momentary assessment or diary research, is more accurately described by the

DOI: 10.1002/per

more inclusive term 'intensive repeated measures in naturalistic settings' (IRM-NS, Moskowitz, Russell, Sadikaj, & Sutton, in press). One of the great strengths of this methodology is the possibility to control the extent of specificity or generalizability in the measurement. A measurement made at a particular moment is specific to that occasion and its concurrent situational cues; different aggregation strategies can produce scores that reduce specificity and that are generalizable across a specifiable number of occasions or set of situations (*cf*., Moskowitz, 1994; Brown & Moskowitz, 1998).

As IRM-NS measures can be constructed to be specific to particular types of situations, these measures can be used to delineate behaviour-situation profiles (e.g. Fournier, Moskowitz, & Zuroff, 2008). We can also use these measures to move beyond the categorization of situations. Simultaneously measuring several situational features of an event can provide a multi-layered characterization of the situation (e.g. the person's role relations to another person in the event, and the person's perceptions of the other person's behaviour in the event), similar to the characterization of the person along multiple dimensions (Moskowitz, 2009). IRM-NS data is particularly suitable for a multi-layer approach to the characterization of events, because the method is usually designed to provide measures of a large sample of events for each research participant.

One advantage of IRM-NS, not mentioned by Furr, is the opportunity to create measures of new kinds of personality constructs. While personality research is frequently focussed on mean level behaviour, IRM-NS measures also permit the examination of distributional features of behaviour (Fleeson, 2001). The most common approach to assessing a distributional feature of behaviour across time is to calculate the within-person standard deviation. Moskowitz and Zuroff (2004) referred to this variability as flux and demonstrated that flux in interpersonal behaviours (e.g. dominant and agreeable behaviours) is a stable feature of the person. IRM-NS measures permit researchers to examine additional features of intraindividual variability, such as the tendency to switch among types of interpersonal behaviour (spin, Moskowitz & Zuroff, 2004). Evidence suggests that individuals with high interpersonal spin are perceived as less effective contributors to work groups and are perceived as peripheral to social networks at work (Côté, Moskowitz, & Zuroff, 2009). Interpersonal spin also has potential for characterizing behavioural differences among individuals with psychopathology (Russell, Moskowitz, Paris, Sookman, & Zuroff, 2007). These and other measures of intraindividual variability show promise for building a new wave of personality characteristics.

There are multiple types of IRM-NS designs, including time-contingent (e.g. record completion once a day), signal-contingent (i.e. record completion upon receipt of a signal from the experimenter) and event-contingent recording (i.e. record completion after an event designated by the investigator). Event-contingent recording designs are particularly well suited to providing measures of particular kinds of events or situations; these events might be missed by the more common signal-contingent methods.

Furr suggests that electronic data collection devices such as palm pilots provide the best IRM-NS data. While this is true with respect to the ease of processing data for the investigator, electronic devices are not always preferable from the perspective of the research participant. In a study of couples, almost 20% of the participants reported dislike for the electronic recording devices (Green, Bolger, Shrout, Rafaeli, & Reis, 2006). Individuals from vulnerable populations (e.g. with psychopathology) or older participants may have difficulty reading the computer screen and manipulating the device. Participants may become frustrated if the interface is confusing or the device does not operate as expected. This may increase the perceived burden of recording and negatively impact

DOI: 10.1002/per

participants' responsiveness and reactivity. Finally, the 'time stamp' provided by electronic devices is not relevant for assessing compliance when using event-contingent recording designs (Takarangi, Garry, & Loftus, 2006).

In summary, it is difficult to identify strong or weak methods for measuring behaviour without specifying the construct or question under investigation. Given the surge of interest in contextualized behaviours, cross-situational generality, behaviour-situation profiles and within person variability, the best fit to the problem may often be designs that use intensive repeated measurements in naturalistic settings.

Behaviours, Non-Behaviours and Self-Reports

SAMPO V. PAUNONEN

Department of Psychology, University of Western Ontario, London, Canada

paunonen@uwo.ca

Abstract

Furr's (this issue) thoughtful analysis of the contemporary body of research in personality psychology has led him to two conclusions: our science does not do enough to study real, observable behaviours; and, when it does, too often it relies on 'weak' methods based on retrospective self-reports of behaviour. In reply, I note that many researchers are interested in going beyond the study of individual behaviours to the behaviour trends embodied in personality traits; and the self-report of behaviour, using well-validated personality questionnaires, is often the best measurement option. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) has expressed dissatisfaction with the body of contemporary research in personality psychology. Whereas our discipline is supposed to represent the scientific study of human behaviour, most published personality studies are only remotely related to actual behaviours that can be seen and measured directly. To support this assertion, Furr has provided a cursory survey of the personality literature, which revealed that perhaps only one in five published studies report results based on behaviour data. His conclusion from this tally was that 'behaviour is not studied to the degree it merits.' He went on to state that even those relatively few studies that have provided actual behaviour data often rely on 'weak' methods based on retrospective self-reports of behaviour.

I have two questions regarding Furr's conclusions about the personality literature and the type of data typically reported. The first is, if only one in five personality studies deals with behaviour data, how non-behavioural are the other four studies? The second question is, just how weak and problematic are those retrospective self-reports that people are asked to provide about their past behaviours?

BEHAVIOURS AND NON-BEHAVIOURS

Furr was not clear about the type of data represented by the four out of five studies in the personality literature that do not report behaviour data. What he did say was that he did not count in his tally any study that 'was either clearly non-behavioural (e.g. self-concept) or

DOI: 10.1002/per

could be interpreted in many non-behavioural ways (e.g. decision-making, creativity).' He also referred to studies of 'cognition, affect, motivation, self-concept, social perceptions and trait structure.'

Personality psychologists are, of course, wholly interested in human behaviour, even if few of their empirical studies entail coding observable behavioural acts. Much of that research is focussed on hypothetical constructs we call personality traits. These traits, rooted in psychological theory, help to explain temporal and situational consistencies in the types of observable behaviours of interest to all of us. Furthermore, personality traits are generally measured with self-report paper-and-pencil questionnaires. Note that, although the only observable behaviour one might see in a typical personality trait study is a respondent circling a number on a seven-point rating scale, that too is behaviour (see next section).

Because of the attitudinal nature of, say, self-esteem, Furr would probably exclude a self-report study of that trait from his tally of behavioural studies. But should he? Consider the myriad types of responses that can be required of test respondents on a self-esteem scale. Indeed, the questionnaire might have self-esteem items only distantly related to actual instances of behaviour (e.g. 'I think I am attractive'). Yet it could have items that are extremely specific and not fundamentally different from the experience sampling reports recommended by Furr (e.g. 'I look at myself whenever I pass a mirror').

It is not evident which studies using such self-report personality trait questionnaires should make it into a tally of behavioural studies. Should we count measures of behaviour frequency but not measures of behaviour trends? What about behaviour preference ratings versus global trait attributions? Making these distinctions, besides being somewhat arbitrary, would generally require careful reading of the items in the questionnaires. Personally, I believe that all of these approaches to assessment represent the study of human behaviour, have a legitimate place in the literature of personality, and, arguably, could be included in a tally of behavioural studies.

SELF-REPORTS *ARE* BEHAVIOURS

Personality assessors long ago supplemented the sign view of the personality item response with a sample view (see Jackson, 1971; Loevinger, 1957; Meehl, 1945). The sample view of personality item responding is that (a) the act of circling a number on an item's rating scale is a behaviour, (b) it is driven by some underlying personality trait(s) and (c) people with higher/lower levels of the measured trait are pressed to circle higher/lower numbers on the rating scale. Of course, the extent to which the item response predicts (i.e. is a sign of) significant non-test behaviours is an important empirical question, the answer to which can, at once, both instantiate the underlying hypothetical construct as well as validate the measure of that construct.

Under the sample view of item responding, personality questionnaires require a person to describe his or her typical behaviours, and those self-reports are taken largely at face value. Certainly there are problems with people reporting on their own behaviours, and Furr has adequately documented these. They include memory errors, motivated distortions and certain response biases. However, these problems have solutions, equally well documented, that begin with a proper programme of test construction having a strong focus on construct validity.

Measuring personality traits using laboratory-based observations of trait exemplary behaviours would not be feasible as a general assessment procedure. Nor, in my opinion,

DOI: 10.1002/per

would be replacing self-reports on a personality inventory with informant ratings. Even if one was interested in only five dimensions of personality (e.g. the Big Five), the time and resources involved in a proper experimental behaviour observation study would be prohibitive. Furthermore, most psychologists see the imperative in understanding preferences, sentiments, feelings, wishes, values and so on, recognizing the importance of internal motivation in the determination of (observable) behaviours. Assessing such thoughts would be nigh impossible through direct behaviour observation, and would be of questionable validity through informant reports.

CONCLUSIONS

I am personally not bothered by Furr's estimate that only one in five studies in the personality literature is 'truly' behavioural, according to his inclusion criteria. Certainly, as he has argued, there are many good reasons to study behaviour directly. But there are equally good reasons for our science to proceed beyond individual acts of behaviour and their attendant situational constraints. As Furr himself has recognized, many personality psychologists are interested in long-term behaviour trends, with generalizability across time and situations—trends that direct observation and even daily experience sampling procedures are not well equipped to reveal.

I do not accept the notion that self-reports, by their very nature, are the foundations of weak methods of behaviour assessment. Well designed, standardized, personality questionnaires that have undergone a rigorous programme of construct validation can provide substantiated respondent information about a multitude of real behaviours, and their internal causes, that cannot be obtained easily by other means, including informant reports. Such an assessment paradigm is an indispensable expedient for the modern study of human behaviour and personality.

ACKNOWLEDGMENT

This research was supported by the Social Sciences and Humanities Research Council of Canada Research Grant 410·2006·1795.

An Ethological Perspective on How to Define and Study Behaviour

LARS PENKE

Department of Psychology, The University of Edinburgh, Edinburgh, UK

lars.penke@ed.ac.uk

Abstract

While Furr (this issue) makes many important contributions to the study of behaviour, his definition of behaviour is somewhat questionable and also lacks a broader theoretical

DOI: 10.1002/per

frame. I provide some historical and theoretical background on the study of behaviour in psychology and biology, from which I conclude that a general definition of behaviour might be out of reach. However, psychological research can gain from adding a functional perspective on behaviour in the tradition of Tinbergens's four questions, which takes long-term outcomes and fitness consequences of behaviours into account. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) puts his finger on an important topic: the neglect of behaviour in psychology. The most valuable part of his target paper is certainly the systematic evaluation of different ways to assess behaviour. Such a practical guideline is sorely needed, since way too often strong inferences are drawn from studies that report what Furr fittingly refers to as 'weak behavioural data.' Indeed, Furr's differentiation between 'weak' and 'strong' behavioural data should become part of the standard vocabulary in personality psychology, as should his differentiation of 'behavioural data' on the manifest level and 'actual behaviour' on the latent level. Further, the distinction between observable behaviours, behavioural tendencies (personality traits and motivations that predispose individuals to show a certain behaviour) and behavioural intentions (attitudes on or preferences for showing a certain behaviour) is a worthy pursuit. With regard to the latter, my colleagues and I provided evidence for analogue distinctions in the area of mate choice and sexual behaviour (Penke & Asendorpf, 2008; Todd, Penke, Fasolo, & Lenton, 2007).

A part of Furr's paper that I found less convincing is his definition of behaviour: *'verbal utterances or movements that are potentially available to careful observers using normal sensory processes.'* First of all, why do utterances necessarily have to be verbal? What about affirmative or deprecatory mumbles and grumbles, laughing, teeth grinding, humming a melody or all the other sounds people can produce (often without much observable movements, I should add)? If we exclude them from a definition of behaviour, don't we miss something important? And what else should they be, if not behaviours? Similarly, do behaviours that are not verbal utterances really need to include movements? As an example, Furr explicitly excludes blushing from his definition, even though it is an integral component of displaying emotional states like shame. People will also certainly notice (and most often get irritated) if somebody suddenly ceases to move, maybe because he or she freezes in panic, or if an interaction partner shows no facial movements whatsoever. In the end, there is something to Watzlawick's (Watzlawick, Beavin, & Jackson, 1967) famous quote 'One cannot not behave.' Overall, I find Furr's definition of behaviour a bit ad hoc and operational (perhaps reflecting the strong methodological focus of the paper). In the following, I aim to give some broader theoretical, historical and interdisciplinary background that could be helpful to understand behaviour and how it should be defined and studied.

While I agree that behaviour does not receive the attention it deserves in psychology nowadays, this had not always been the case. In the first half of the last century, behaviourism (e.g. Watson, 1925) focussed almost exclusively on behaviour as the subject of psychology, banning everything else into a securely shut 'black box.' The behaviouristic definition of behaviour resembles Furr's in being rather operational and atheoretical: behaviour is what an organism observably does or says (Watson, 1925, p. 6). At roughly the same time, a new wake of interest in behaviour occurred in a different discipline, biology, and led to the formation of the new branch of ethology, which was also exclusively dedicated to the study of behaviour. Since then, ethology has developed into

DOI: 10.1002/per

sociobiology and later behavioural ecology, making biology a discipline where the study of behaviour is at least as established as in psychology. Curiously, however, even landmark publications in ethology (Eibl-Eibesfeld, 1989; Tinbergen, 1963), sociobiology (Wilson, 1975) and behavioural ecology (Krebs & Davies, 1997) failed to provide an explicit definition of behaviour, and modern standard biology textbooks also give only vague definitions like 'what an animal does and how it does it (Campbell, 2008).' Part of the reason for the general lack of a clear definition of behaviour might be that it is more a lay concept from everyday language than a scientific construct, helpful mainly to draw rather artificial lines between movements outside and inside the body (e.g. gut motility), or between body movements and physiological reactions (like blushing or sweating), or movements in animals and reactions to environmental stimuli in plants (including locomotion in slime mould).

Instead of a mere definition; however, ethology provided something else to the study of behaviour that turned out to be even more important—and has since become an integral part of the biology of behaviour: Tinbergen's (1963) four questions that he suggested should be asked about any behavioural phenomenon. Two of them concern the underlying proximate mechanism that causes behaviour (the content of the behaviourists' black box) and its ontogenetic development, and both are all too familiar to psychologists. The other two; however, are not generally acknowledged by all psychologists. They concern the effects that a behaviour ultimately has on evolutionary fitness and the behaviour's phylogenetic history. The advantage of asking all four questions about any behaviour is that proximate, mechanistic and ultimate, functional answers to any 'why'? and 'how?' questions are explicitly separated. It reminds researchers that insight into the *function* of behaviour can only be gained from an evolutionary perspective (Krebs & Davies, 1997).

From such an evolutionary, functional perspective, behaviour can be understood as a way how individuals adjust more or less instantaneously to their current environment (Penke, 2009). Behavioural adjustments can be fitness-increasing if they are conditional to adaptively relevant environmental aspects and if they are guided by evolved adaptations as well as stable behavioural tendencies and intentions (as reflected in personality traits) that have been under balancing selection or recent selective sweeps (Penke, in press; Penke, Denissen, & Miller, 2007). Put differently, behaviour is not merely a function of the situation (i.e. the environment) and the person (Funder, 2006), but a way how persons can fit themselves to situations. Depending on how well such fits are achieved in different stages of the lifespan, behaviours contribute to more or less favourable life outcomes and ultimately fitness differences (Penke, in press). Thus, behaviours are important mediators between person-environment interactions and fitness consequences, and they should be studied as such.

To sum up, Furr's definition does not seem to capture the whole phenomenon that is behaviour, but a more general and conclusive one may be out of reach. Thus, a slightly modified version that allows for non-verbal utterances and certain non-movements might be sufficient for practical purposes and measurement discussions. However, research on behaviour is well advised to take a broader perspective and to attempt to answer all four of Tinbergen's (1963) questions for any behaviour under study.

DOI: 10.1002/per

What is a Behaviour?

MARCO PERUGINI
Faculty of Psychology, University of Milan—Bicocca, Milan, Italy
marco.perugini@unimib.it

Abstract

The target paper proposes an interesting framework to classify behaviour as well as a convincing plea to use it more often in personality research. However, besides some potential issues in the definition of what is a behaviour, the application of the proposed definition to specific cases is at times inconsistent. I argue that this is because Furr attempts to provide a theory-free definition yet he implicitly uses theoretical considerations when applying the definition to specific cases. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) convincingly argues that the personality field should pay much more attention to the study of behaviour so that it can define itself as a truly behavioural science. The first step in such a worthwhile enterprise is to define what is a behaviour. I shall focus my comment on this point as it provides the foundation for all other subsequent issues.

I wish first to clarify that many of the arguments put forward by Furr are convincing and overall this is an important contribution to the personality field. However, although there is much to admire in the effort of Furr as he takes on a task of Herculean proportion, I am not convinced by some key distinctions that he adopts at the onset and, especially, by some key exclusions that he develops in the remaining of his paper.

Furr defines behaviour as '. . . verbal utterances or movements that are potentially available to careful observers using normal sensory processes.' If one accepts this definition and takes it literally, then it is unfortunate that some classes of actions are excluded. Take for instance blushing that is excluded because it is not a movement. However, it can be observed by a careful observer and it is an important indicator relevant for personality research (Leary & Meadows, 1991). Moreover, blushing tends to co-occur with gaze aversion in social anxiety inducing situations (Leary, Britt, Cutlip, & Templeton, 1992). From a theoretical point of view it would seem awkward that the first cannot be considered as a behaviour whereas the second can.

Furr excludes also some other classes of actions although I am not sure on what grounds. Take first reaction times or computerized cognitive assessment. Although Furr does not exclude them altogether ('. . . most indices . . . '), he does not provide any clue on what is meant by most and conversely what can still be included. I would argue that it all depends on what is specifically measured by the task. A finger pressing a key, with or without a time response window, could mean for instance choosing between two different products or judging some personality characteristics on the basis of a rapidly presented picture of a face. On what definitional grounds these actions should be excluded most of the time?

Furr returns on this issue by attempting to further define what are behavioural data, that is, '. . . data intended to represent what a person does rather than what a person thinks, feels or otherwise experiences.' The problem with this distinction is that, unlike what Furr argues, it cannot be used to exclude some behaviours as not relevant. As I have argued before, using a finger can be a behaviour in the sense of Furr's initial definition. It is a

DOI: 10.1002/per

behaviour no matter whether it is a movement on the keyboard to press a button corresponding to some choice or whether it is the middle finger directed towards someone with whom one is angry at. Would he argue that this latter is not a behaviour? If it is not a behaviour, then so many other actions (e.g. gaze aversion) are not a behaviour as well. But if it is considered as a behaviour, then why showing the middle finger is what one *does* whereas pressing a finger on the keyboard is what one thinks or feels?

Moreover, this further qualification could be used to exclude some actions that instead Furr considers as behaviours. Take for instance utterances. It could be easily argued that utterances reflect what one person feels or thinks rather than does. This would imply the paradoxical situation that despite being explicit part of the main definition, utterances should not be considered as behaviour if we were to apply this further qualification.

Perhaps the most arbitrary decision of Furr is to exclude specific behaviours such as to donate money to a charity in a laboratory experiment. Here I cannot see on what grounds this decision can be justified. To argue that such class of actions should be excluded because they do not fall clearly within the types of behavioural data described in his manuscript points to a limitation of the classification table (or of its use, as I shall argue below) rather than anything else. Equally, to exclude them because they are too specific again is not a good reason. For example, in what sense can gaze aversion be considered as a more general behaviour than donating money to a charity?

But I would also argue that in fact this class of specific actions does fall within the different types of behaviours proposed by Furr. Donating money to a charity in a laboratory experiment is one neat example of direct behavioural observation. The main difference with other examples of this class of behaviour provided by Furr is that careful observers will have a 100% agreement on coding this specific behaviour of donating money. Perhaps what can create confusion is that there is no compelling need of careful observers given that these behavioural data can be directly and objectively collected via other means (e.g. stored in a computer file). Yet, the definition refers to behaviour that *in principle* is available to careful observers and not that it *must* be coded by careful observers. Therefore, a specific action such as donating money to a charity is just one example of direct behavioural observation and I do not see in what sense it is too specific or not a general method—exactly the same arguments could be used for most other observed behaviours.

Taken altogether, these examples help me to highlight what in my opinion are the weakest points of Furr's contribution, that is the limitations of the definition of what is a behaviour and especially the inconsistency in the use of the definition. The first issue means that the proposed definition leaves out phenomena that on theoretical grounds would be deemed to have equal status of other phenomena that are instead included. It is fair to note that it is not easy to find a truly better alternative and that every general definition carries this risk. However, the second issue means that, even if we were to accept the definition, then some subsequent arguments would seem inconsistent with the definition itself.

I suspect that the main problem lies in the attempt to define what is a behaviour irrespective of theoretical considerations yet implicitly using them when applying the definition to specific cases. Some economists would argue convincingly that donating real money is a behaviour, whereas gaze aversion is definitely not. Some developmental psychologists would argue that gaze aversion is an indicator of shyness, whereas others that is an indicator of learning (Phelps, Doherty-Sneddon, & Warnock, 2006). Theoretical considerations are essential to identify, measure and justify the use of some classes of

DOI: 10.1002/per

actions that can be considered as behaviour. The theoretical background and purposes of a given study fundamentally influences what can be considered as a relevant behaviour in a specific study. If one recognizes that this is a fundamental fact of scientific research, then it follows that whatever general definition of a behaviour can at most provide a first useful step that needs to be supplemented by a specific explicit evaluation of the *relevance* of an action to the underlying theoretical construct in a specific research with a specific aim. Moreover, the general definition needs to be applied consistently. If we do so in this case, it would seem to me that there are no convincing reasons why some classes of actions should be excluded in general, whereas it is possible that in some specific cases some specific actions may not be deemed to be relevant behavioural indicators. But I fail to see a better alternative than leaving the burden to the interested researchers of justifying theoretically why a certain action is a relevant behavioural indicator in a specific case.

Is Personality Really the Study of Behaviour?

MICHAEL D. ROBINSON
Department of Psychology, North Dakota State University, Fargo, ND, USA
Michael.D.Robinson@ndsu.edu

Abstract

Furr (this issue) contends that behavioural studies of personality are particularly important, have been under-appreciated, and should be privileged in the future. The present commentary instead suggests that personality psychology has more value as an integrative science rather than one that narrowly pursues a behavioural agenda. Cognition, emotion, motivation, the self-concept and the structure of personality are important topics regardless of their possible links to behaviour. Indeed, the ultimate goal of personality psychology is to understanding individual difference functioning broadly considered rather than behaviour narrowly considered. Copyright © 2009 John Wiley & Sons, Ltd.

Early views of personality were integrative in nature (e.g. McClelland, 1951), but this has changed over time. For example, there is very little work today integrating ability and psychosocial perspectives of the individual (Batey & Furnham, 2006). Also, important questions related to the self-concept are typically investigated by social rather than personality psychologists, despite the clear relevance of this interface (Robinson & Sedikides, in press). In my view, personality psychology should retain the integrative potential favoured by early theorists (for a related perspective, see Mayer, 2005). The beauty of personality psychology is that a simple correlation between different sorts of measures or realms of functioning can establish a connection that enriches our understanding of both personality and other areas of psychology.

For example, Rogers and Monsell (1995) showed that there are robust deficits in reaction time performance when individuals are asked to switch between simple cognitive tasks. Subsequently, higher task-switching costs were linked to advanced age, schizophrenic

DOI: 10.1002/per

symptoms and negative affect (for a review, see Ode, Robinson, & Hanson, 2009). Such links inform our understanding of what task-switching costs assess and measure. Moreover, paradigms of this type can now be viewed as an important probe of individual differences in executive function, a probe likely to have considerable value in understanding other personality processes as well. In sum, I view personality psychology as an integrative and inter-disciplinary science rather than one that should pursue a specific agenda or set of topics.

In multiple respects, the target paper should be viewed as a successful and thought-provoking one. Nonetheless, my integrative vision for personality science clearly differs from Furr's (this issue) recommendations. To highlight such points of disagreement, I consider a series of questions motivated by Furr's analysis.

Should Personality be reduced to Its Behavioural Manifestations? In my view, personality psychology encompasses all domains associated with significant and consequential individual differences. Behaviours, especially of an interpersonal kind, fit within this definition. Yet, so do quite a few non-behavioural variables like emotional experiences, motivational states, self-views, biological states and cognitive processing tendencies. The idea that the latter variables are only important to the extent that they manifest themselves in behaviour is questionable. For example, the anxiety-behaviour relationship is often a complicated one, but a subjective life marked by high levels of anxiety is problematic nonetheless (Eysenck, 1997).

Is Behaviour the Ultimate Outcome to Be Explained? Furr (this issue) admirably encourages a focus on what may be viewed as consequential outcomes. However, behaviour is not the only or even the most important personality outcome to be predicted. Furr appears to neglect more consequential outcomes such as death, divorce, criminality, work success and so forth. None of these outcomes are strictly behavioural according to Furr's definition. Nonetheless, it is arguably the case that such outcomes are more important than the behaviours exhibited by the individual. For such reasons, I propose that functioning (broadly defined) is the ultimate outcome to be explained in personality research.

Why Have Behavioural Studies of Personality Decreased in Frequency? Baumeister, Vohs, and Funder (2007) documented a decline in the use of strictly behavioural measures, a theme if not conclusion reached by Furr (this issue). In appreciating any such trend, it must be recognized that the psychology literature of the 1960s–1980s continued to be influenced by Skinnerian psychology, which was strictly behavioural in nature (Dixon, 1981). Major advances in psychology have been made since then, including in the domains of cognition, emotion and neuroscience. It is justifiable that these more recent advances be given greater attention in the personality literature than had been given previously. From the present perspective, then, behavioural studies of personality are well represented in modern personality research.

Are Behavioural Studies of Personality Disadvantaged? Furr (this issue) reports that the average number of studies per paper is lower for investigations reporting high-quality behavioural data. Thus, there appears to be no publication bias to correct in the future: editors and reviewers already recognize the value of high-quality, intensive behavioural studies. Indeed, in the present author's opinion, further insistence on behavioural sources of data is likely to harm rather than facilitate personality psychology's diversity and integrative potential.

Is 30% Really Such a Poor Figure? Furr (this issue) documents the fact that approximately 30% of the papers published in *Journal of Personality* and the personality

DOI: 10.1002/per

section of *Journal of Personality and Social Psychology* include behavioural measures in at least one study. This figure is deemed perhaps too low, but I do not view this to be the case. As Furr notes, personality psychology is concerned with a wide range of topics including cognition, affect, motivation, the self-concept, social perceptions and trait structural considerations. Given this broad integrative scope, it is not surprising or problematic that 30% of papers assess actual behaviour as Furr defines it. This is an adequate figure in my opinion.

Should There Be an Agenda for Personality Research? Furr (this issue) presents what amounts to an agenda for how personality research should be conducted and which specific outcomes are of most importance. In the present author's view, it is dangerous to link agendas to the manner in which science is conducted. When this is done, the science becomes too narrow and the funding opportunities too political.

Conclusions. There is more to personality than behaviour. Topics related to cognition, motivation, emotion, the self-concept and trait structure will continue to be of major interest to personality psychologists, as they should be. Behavioural studies of personality are important as well. However, my view is that behavioural studies of personality should not be valued above other sorts of investigations, nor should an agenda be set for what constitutes appropriate research questions in the future. Furr (this issue) makes a strong set of points, as target papers should, but personality psychology has more value as an integrative science rather than one devoted to a particular set of outcomes or dependent measures.

Linking Personality and Behaviour Based on Theory

MANFRED SCHMITT

Department of Psychology, University of Koblenz-Landau, Landau, Germany

schmittm@uni-landau.de

Abstract

My comments on Furr's (this issue) target paper 'Personality as a Truly Behavioural Science' are meant to complement his behavioural taxonomy and sharpen some of the presumptions and conclusions of his analysis. First, I argue that the relevance of behaviour for our field depends on how we define personality. Second, I propose that every taxonomy of behaviour should be grounded in theory. The quality of behavioural data does not only depend on the validity of the measures we use. It also depends on how well behavioural data reflect theoretical assumptions on the causal factors and mechanisms that shape behaviour. Third, I suggest that the quality of personality theories, personality research and behavioural data will profit from ideas about the psychological processes and mechanisms that link personality and behaviour. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) defines behaviour but not personality. Yet the relevance of behaviour for our field depends on how we define personality. Consider two alternatives. Definition A views personality as a set of latent factors that account for the cross-situational consistency

DOI: 10.1002/per

and stability of behaviour. Definition B views personality as a set of goals plus knowledge about the instrumentality of behaviours (means-end-knowledge). Definition A forms the basis of descriptive trait models (Cattel, Eysenck, Guilford, Five Factor Model). Definition B combines need theories (Jackson, Murray) and cognitive personality theories (Ajzen, Rotter). According to Definition A, behaviour is an integral part of personality, and thus it makes no sense to ask whether personality causes behaviour. Definition B views personality and behaviour as conceptually and empirically independent psychological elements. Asking whether personality causes behaviour is meaningful. Answering this question requires substantive theory and empirical research. I want to make the point that the role of behaviour in personality psychology is not self-evident, but depends greatly on how we define personality. My next comment follows from Definition B.

TAXONOMIES OF BEHAVIOUR SHOULD BE GROUNDED IN THEORY

Furr (this issue) proposed a taxonomy of behavioural data that combines measurement methods and types of data. Furr provides an analysis of the strengths and weaknesses of each category and concludes that direct behavioural observations, experience sampling reports of current behaviour, and acquaintance reports of recent behaviour provide the strongest data. I fully agree that the quality of behavioural data depends greatly on the validity of the measurement methods. I would like to add that the quality of data also depends on how well grounded in theory they are. Behavioural data are strong to the extent that they reflect theoretical assumptions on the causal factors and mechanisms that shape behaviour. I illustrate my point with three examples.

Controllability. Behaviours differ in how well they can be controlled. Some kinds of behaviour, such as speech illustrators and body adaptors, occur automatically and without conscious awareness. Other behaviours, such as starting a conversation, require intentional control. Several studies have shown that automatic behaviour can be better predicted from implicit traits, whereas controlled behaviour can be better predicted from explicit traits (Asendorpf, Banse, & Mücke, 2002). The distinction between automatic and controlled behaviours is important because they have different causal antecedents. Moreover, the distinction is theoretically meaningful because the pattern of correlations and effects can be explained by dual process theories (Friese, Hofmann, & Schmitt, 2009). Behavioural data are strong to the extent that the controllability of behaviour is known and corresponds to the causal factors that are considered in a study.

Factorial complexity. Behaviour differs in its factorial complexity (Borkenau, 1986). Behavioural items of personality scales have one strong factor and little uniqueness (sum of all weak factors). As a matter of fact, factorial simplicity is the very reason for employing such items. They are pure measures of their factor. However, many kinds of behaviour are factorially complex. They are not shaped by a single personality factor, but by several factors. Many real-life behaviours, such as helping a stranger find his way or making choices between consumer products, are multi-determined. Predicting multi-determined behaviours requires complex causal models. Specifying such models requires careful theoretical analysis. Again, behavioural data are strong to the extent that they are consistent with a theoretically sound causal model.

Situational sensitivity. Behaviour depends not only on personality factors, but also on situation factors and person x situation interactions. All meta-analyses of the relative impact of personality factors, situation factors and person x situation interactions have

DOI: 10.1002/per

shown that the proportion of behavioural variance that can be explained by these factors varies across types of behaviour (Richard, Bond, & Stokes-Zoota, 2003). The same observation has been made by research based on latent state-trait theory (Steyer, Schmitt, & Eid, 1999). Interaction and moderator theories offer substantive explanations for inter-behaviour differences in situational sensitivity (Friese et al., 2009; Schmitt, in press). Behavioural data are strong to the extent that they are consistent with these theories and useful for testing hypotheses derived from them.

MECHANISMS LINKING PERSONALITY AND BEHAVIOUR

Personality has been defined as coherent patterns of cognition, emotion, and action (Cervone & Shoda, 1999). Theoretical explanations of behaviour will profit from ideas about the processes and mechanisms that link these elements. Let me use our research on justice sensitivity as an example (Schmitt, 1996). We assume that due to the frequent perception of and rumination about justice issues, persons high in justice sensitivity develop highly accessible injustice concepts as well as a highly differentiated and integrated knowledge structure in the domain of justice. This knowledge structure shapes the processing of justice-related information (Baumert & Schmitt, in press). It guides attention, the interpretation of ambiguous information, and how well just and unjust information is stored in memory. These information processes affect emotional reactions to experienced injustice (anger), emotional reactions to observed injustice (moral outrage) and emotional reactions to committed injustice (guilt). Anger, moral outrage and guilt in turn are motives for specific actions. These ideas have been tested successfully using various kinds of behaviour such as protesting against inequitable treatment (Mohiyeddini & Schmitt, 1997), helping innocent victims (Gollwitzer, Schmitt, Schalke, Maes, & Baer, 2005), behaviour in experimental games (Fetchenhauer & Huang, 2004), civil courage (defending a victim against a perpetrator at the risk of being harmed), and vengeful reactions after having been exploited (Gollwitzer & Rothmund, in press). These studies and research in other domains of personality such as anxiety (Williams, Watts, MacLeod, & Mathews, 1997) suggests that considering mechanisms that mediate the effects of personality on behaviour (Definition B) or explain links among personality components (Definition A) contributes to the quality of personality theories, the quality of personality research and the quality of behavioural data.

The Apparent Objectivity of Behaviour is Illusory

RYNE A. SHERMAN, CHRISTOPHER S. NAVE and DAVID C. FUNDER
Department of Psychology, University of California, Riverside, CA, USA
funder@ucr.edu

Abstract

It is often presumed that objective measures of behaviour (e.g. counts of the number of smiles) are more scientific than more subjective measures of behaviour (e.g. ratings of the degree to which a person behaved in a cheerful manner). We contend that the apparent

Eur. J. Pers. **23**: 403–435 (2009)
DOI: 10.1002/per

objectivity of any behavioural measure is illusory. First, the reliability of more subjective measures of behaviour is often strikingly similar to the reliabilities of so-called objective measures. Further, a growing body of literature suggests that subjective measures of behaviour provide more valid measures of psychological constructs of interest. Copyright © 2009 John Wiley & Sons, Ltd.

Understanding human behaviour is a fundamental goal in psychology. In many instances, behaviour serves as the ultimate outcome of interest: what people *do*. To study behaviour effectively, our field needs to develop a common language for describing behaviour, a clear understanding of what behaviour is, and a useful taxonomy of the ways in which behaviour can be studied. Despite over 100 years of psychological research, none of these aims has been accomplished. The target paper provides a step towards remedying this situation by providing a clear, coherent definition of behaviour as well as a systematic taxonomy for methods of behavioural measurement.

We wholeheartedly agree with Furr that behaviour should not continue to be neglected in personality research. The amount of behavioural data reported in the personality literature has been surprisingly minimal (Furr, this issue) and most studies that attempt to gather such data do not observe behaviour directly, but rely on relatively weak and problematic methods (e.g. retrospective or hypothetical self-reports). This comment applies to social psychology as well (Baumeister et al., 2007). We applaud the target paper's distinction among strong and weak behavioural methodologies, and its encouragement of researchers to use more direct behavioural observation, experience sampling or acquaintance reports of recent behaviour.

While the target paper does a laudable job of distinguishing much of the terminology used in behavioural research, one often raised distinction is not discussed—the difference between objective and subjective behaviours. The purpose of this paper is to illuminate the distinction and argue that ultimately, the objectivity of any behavioural measure is illusory.

OBJECTIVE VERSUS SUBJECTIVE BEHAVIOUR

Most behavioural data and all interpretations of behavioural data are far from objective. Yet there tends to be a misperception that objectively measured behavioural data is somehow better, or more scientific.

What is the difference between objective and subjective behaviours that might lead to this misperception? So-called objective behaviours tend to be fairly micro-level in nature and are often coded using counts or frequencies. For example, counting the number of smiles a person displays might be considered an objective measure. Behaviours regarded as more subjective tend to be mid-to-macro level in nature and are often coded by general overall impressions. For example, rating how much a person, 'Behaved in a cheerful manner' on a Likert-type rating scale would typically be considered a subjective measure.

Counting the number of smiles might be said to be objective because each coder can physically point to each smile and make a note of it, while the basis of a rating of 'cheerful behaviour' cannot be so directly specified. It is perhaps for this reason alone that objective behaviours are sometimes seen as more scientific than subjective behaviours and in some instances even the gold standard. In fact, Kenny (1994) states, '. . . researchers should

DOI: 10.1002/per

attempt to establish high levels of interrater reliability . . . to show that the behaviour ratings are relatively objective.' And elsewhere, '. . . instead of rating friendliness, observers should count or measure the duration of smiles (pp. 135–136).' Thus, the apparent rationale for a preference for objective behaviours is presumed increased agreement, otherwise known as reliability or precision of measurement.

However, this presumption of increased agreement and objectivity is often unfounded. For example, we could take a behavioural frequency approach and count the number of times that a person smiles. But even this seemingly innocent task has complications. What exactly is a smile? Typically operational definitions must be created by the researcher (e.g. lips move into upward shape; cheekbones raised; shows teeth), but ultimately some mouth movements may look like smiles to some coders, but not to others. As a result, the coding of this seemingly objective behaviour might be less reliable than one would expect. Further, the interpretation of smiling is also not objective. Much like the subjectivity that inevitably enters into the interpretation of a factor analyses, these results must be interpreted by people and these interpretations may vary.

The literature and our own research increasingly indicate that coding psychologically meaningful behaviours should receive priority, regardless of perceived objectivity (Martin & Bateson, 1993; Vazire, Gosling, Dickey, & Shapiro, 2007). In fact, the reliabilities of objective and subjective behavioural measurements are often surprisingly similar. High reliability has been frequently obtained using more subjective, macro level accounts of behaviour (e.g. Fast & Funder, 2008; Nave, Sherman, & Funder, 2008; Vazire & Funder, 2006). More importantly, subjective accounts of behaviour tend to provide greater external validity. Objective behaviours tend to be more contextualized, and therefore less generalizeable (Martin & Bateson, 1993; Vazire et al., 2007). For example, how many times someone smiles ought to be related to how happy someone is feeling (and it often is), but sometimes it is not. Sometimes people smile because they are nervous or anxious. Sometimes people smile because they are feeling sneaky. And sometimes people smile to mask their true feelings about another person (e.g. annoyance). So in fact, smiling is not always the best indicator of how a person might be feeling. On the other hand, subjective measures of behaviour can demonstrate impressive relationships with psychological constructs of interest. For example, someone who is perceived to be 'behaving in a cheerful manner' is very often feeling happy because the behavioural description takes into account many things that a single objective behaviour like smiling ignores, such as other non-verbal behaviour as well as what the person says. Further, such subjective coding allows a qualitative interpretation of the physical process of a smile. That is, when a target person smiles the objective coder must make a tally on a page, while the subjective coder may use all available information to determine if the smile was an anxious smile or legitimately a 'cheerful' smile.

CONCLUSION

The bottom line is that any apparent objectivity of narrowly specified behaviours is illusory and ultimately hinges upon agreement among raters. To find out what a behaviour means, it is necessary to look to how behaviour relates to other data. Multi-method approaches greatly aid our understanding. We are grateful that the target paper has helped clarify the common language that psychologists should use when studying behaviour and for its examination of the various behavioural methods that we can employ in our future

DOI: 10.1002/per

research, but we would remind readers that 'behaviour' is best thought of as a broad construct, not a narrow one.

ACKNOWLEDGEMENTS

Preparation of this note was aided by National Science Foundation Grant No. 06422243 to David Funder.

Personality and Behaviour: A Neglected Opportunity?

LIAD UZIEL and ROY F. BAUMEISTER

Department of Psychology, Florida State University, Tallahassee, FL, USA

Baumeister@psy.fsu.edu

Abstract

Personality psychology has neglected the study of behaviour. Furr's efforts to provide a stricter definition of behaviour will not solve the problem, although they may be helpful in other ways. His articulation of various research strategies for studying behaviour will be more helpful for enabling personality psychology to contribute important insights and principles about behaviour. The neglect of behaviour may have roots in how personality psychologists define the mission of their field, but expanding that mission to encompass behaviour would be a positive step. Copyright © 2009 John Wiley & Sons, Ltd.

Furr (this issue) has presented the problem of the neglect of behaviour in personality psychology. We share Furr's view about the importance of studying behaviour and we appreciate the effort that the author has made in promoting progress in this direction. In our commentary we would like to address two issues that relate to reasons for this neglect in social and personality psychology.

The first point concerns the importance that Furr attributes to having a proper definition of behaviour. Although efforts to improve precise definitions are nearly always useful for science and we appreciate Furr's contribution, we do not share Furr's diagnosis that this is *the* fundamental issue at hand. In fact, we suspect that at this point a search for a definition of what constitutes a behaviour runs the risk of turning this pressing practical problem into a philosophical discussion. In the early days of modern social psychology, researchers were not concerned with whether their measures fit strict definitions of behaviour, but with whether their research had relevance to everyday issues and problems, most of which clearly had something to do with behaviour. Many of the classic studies in social psychology were attempts to provide scientific explanations for everyday behaviours. The link between a psychologist's lab and the experiences of the person in the street was intuitive.

The neglect of behaviour by present day social/personality psychology is not a result of lack of a proper definition of behaviour—we suspect most researchers can agree broadly

DOI: 10.1002/per

about what behaviour is—but, fundamentally, a question of prioritization and communication. In essence, the study of behaviour is the central tool for social/personality psychology to communicate with its broader audiences in a common language. By not studying behaviour, psychologists have made an implicit decision to distance themselves from everyday reality. Unfortunately, deciding that a specific act fits (or misfits) a definition of behaviour will not restore the relevance of our field; not every bit as much as a conscious and deliberate decision to study once again everyday acts that people do.

The second point that we wish to make addresses the reasons for neglecting behavioural data by social and personality psychology. Furr's survey of personality research revealed that only about 7% of the studies included direct behavioural observation, experience sampling or acquaintaince reports, and another 8% included behavioural tests. This total of 15% is close to the 20% that was found in a survey of a combined pool of social and personality studies (Baumeister et al., 2007). Behaviour is thus no more popular (and perhaps somewhat less popular) among personality than among social psychologists. There may be particular reasons that personality psychologists are especially unmoved to study behaviour, however.

Behaviour is inextricably central to the project of social psychology. Social psychologists aim to discover, describe and explain social reality. A large proportion of this reality takes place in the open and is expressed in the acts that people do. In neglecting to address behaviour, social psychologists seriously limit their database, and in many respects, neglect part of their field's mission. Baumeister et al. (2007) suggested several reasons that explain why this happens. Among the more disturbing reasons are bureaucratic considerations and logistical concerns (e.g. that obtaining an IRB approval for a study that includes actual behaviour is difficult, or that the collection of behavioural data takes more time). Overcoming many of these barriers is mostly a matter of researchers' decision to change the priorities in the field.

Things are different in personality psychology. Many personality psychologists see their field's primary mission as assessment: finding ways to classify and categorize people according to their stable traits. Psychology has a long tradition of focussing on assessment, and this includes both intelligence testing and clinical diagnosis. No one seemingly complains that intelligence researchers often neglect to identify behavioural results of intelligence or that clinical assessment is done with interviews and questionnaires rather than behavioural observation. Learning about people can be seen as an end in itself, without needing to extend to prediction and observation of behaviour.

A few decades ago strong arguments were made about the inability of personality traits to predict behaviour (Mischel, 1968). In the years that followed this criticism personality psychologists have struggled to make the argument that their constructs are useful in predicting behaviour. Only recently this struggle was declared successful. The person–situation debate has recently been described '98% over' (Funder, 2001) and more recently to be completely over (Fleeson & Noftle, 2008). Still, the personality–behaviour relationship is, arguably, more than a 'neglect of ease.' It is a reflection of fundamental questions in personality psychology, and their recent resolution is a positive sign for the field.

Part of the solution of the person-situation debate was a formation of a general agreement about the basic elements of personality (i.e. the Big Five; John & Srivastava, 1999) and the accumulation of evidence about the ability of traits to predict behaviour and life outcomes with reasonable success (e.g. Roberts et al., 2007). A great deal of thought and work went into identify the Big Five dimensions as a way of making sense of the mountains of questionnaire data and, more profoundly, of the interrelationships among all

DOI: 10.1002/per

the trait constructs that researchers have identified. This was a seen as a major goal of the field, and it did not require validation with behavioural data.

In a sense, personality researchers approach social phenomena with an explanation already at hand (i.e. the personality trait). And this answer often reflects aggregated behavioural tendencies, because traits are frequently defined in terms of behaviours (such as talkativeness). In focussing on behaviour, personality psychologists sometimes risk making true but trivial conclusions (e.g. in finding that extraverts are more talkative than introverts). They also risk making circular arguments (e.g. that extraverts' talkativeness is caused by their extraverted and therefore talkative nature). For a reason, then, personality psychologist are more interested with *why* and *when* some individuals behave (or have the predisposition to behave) in certain ways. These questions call for a focus on motivation, emotion and cognition and these processes take the centre stage in personality research. To be sure, that is not to say that personality researchers should not study behaviour. The argument that we wish to make is that the discovery of new behavioural phenomena is of less crucial importance for personality psychology, and the study of behaviour in personality research is more often in the service of exploring inner processes (and as such is highly needed and valued).

In conclusion, the study of behaviour plays a different role in social and personality psychology, and the reasons for a relative neglect of behaviour may reflect different underlying processes in the two fields. Still, in both fields behavioural data is and should continue to be an essential part of theoretical progress, one which also has the ability to re-establish the applicability of the field to the non-scientific world. Furr's paper points the way for personality psychologists to embrace or reaffirm their field's potential for making a contribution to understanding human behaviour. A greater inclusion of actual behaviour in personality psychology's research designs would be a very positive step.

DOI: 10.1002/per