European Journal of Personality, *Eur. J. Pers.* **34**: 1037–1059 (2020)
Published online 11 May 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2266

Assessing Personality States: What to Consider when Constructing Personality State Measures

KAI T. HORSTMANN* and MATTHIAS ZIEGLER

Institute of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Abstract: Repeated assessments of personality states in daily diary or experience sampling studies have become a more and more common tool in the psychologist's toolbox. However, and contrary to the widely available literature on personality traits, no best practices for the development of personality state measures exist, and personality state measures have been developed in many different ways. To address this, we first define what a personality state is and discuss important components. On the basis of this, we define what a personality state measure is and suggest a general guideline for the development of such measures. Following the ABC of test construction can then guide the strategy for obtaining validity and reliability evidence: (A) What is the construct being measured? (B) What is the intended purpose of the measure? And (C) What is the targeted population of persons and situations? We then conclude with an example by developing an initial item pool for the assessment of conscientiousness personality states. © *2020 The Authors. European Journal of Personality published by John Wiley & Sons Ltd on behalf of European Association of Personality Psychology*

Key words: personality states; assessment; experience sampling; validity; reliability

Most personality theories suggest that personality can be described using a number of entities that all have a unique stable component, personality traits, and also variable aspects, personality states, that fluctuate from moment to moment (Baumert et al., 2017; Funder, 2001; Wrzus & Mehl, 2015). To put these theories to the test and to disentangle the effects of traits and states, researchers frequently rely on experience sampling methods (Horstmann & Rauthmann, n.d.; Wrzus & Mehl, 2015). In many such cases, participants first respond to a one-time assessment of their personality traits and general characteristics, often based on self-report. Subsequently, participants are invited to report their daily behaviour over a longer period of time, for instance, every 3 hours or whenever certain events occurred (Horstmann, 2020). Based on the data collected, theories regarding the interplay between states and traits can be tested. For example, whole trait theory (Fleeson, 2001) postulates that, on the descriptive side, states, repeatedly assessed within one person, should form density distributions of behaviour and that the average personality state should therefore, roughly, correspond to the personality trait of that person. Testing this theory thus requires a repeated assessment of personality states. Yet despite all the theory and studies already existing, state assessments are often constructed in a rather ad hoc manner. In comparison with the abundance of guidelines to construct trait measures (AERA, APA, & NCME, 2014; Borsboom, 2006; Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1980, 1995; Ziegler, 2014), similar literature just begins to emerge for state assessments (Himmelstein, Woods, & Wright, 2019; e.g. Hofmans, De Clercq, Kuppens, Verbeke, & Widiger, 2019; Wright & Zimmermann, 2019; Zimmermann et al., 2019). This does, to our knowledge, apply not only to personality state measures but also to experience sampling items more generally. The current paper aims at providing a first set of such guidelines to further establish quality state assessments and to spur the development of quality standards for state assessments.

Although the construction and psychometric evaluation of global self-reports have been routinely conducted over the past decades, following standard procedures, it is rather unclear how psychometric properties of repeated self-reports of personality states should be constructed and examined and to which benchmarks they should be compared. Although the psychometric properties of scores obtained with the experience sampling method have been discussed at some lengths (e.g. Schönbrodt, Zygar, Nestler, Pusch, & Hagemeyer, n.d.; Furr, 2009; Hektner, Schmidt, & Csikszentmihalyi, 2007; Moskowitz, Russell, Sadikaj, & Sutton, 2009; Nezlek, 2017), some aspects, such as construct validity evidence, were not considered in detail. Additionally, recent technological advancements, such as the ubiquity of smartphones, have increased the usage of experience sampling methods in personality psychology as well as the experience gained from this usage. Although many very interesting and impactful research findings could be obtained using state assessments in daily life, we would argue that there are so far only very limited guidelines regarding the psychometric evaluation as well as theoretical foundation of personality state scores. In the current article, we will first

*Correspondence to: Kai T. Horstmann, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Germany.
E-mail: kai.horstmann@hu-berlin.de

Handling editor: EJP Guest Editor
Received 1 June 2019
Revised 27 March 2020, Accepted 7 April 2020

define what a personality state and a state measure is. We will then review current practices of evaluating evidence regarding state score's reliability and validity, and common reporting standards. Subsequently, we will formulate concrete expectations towards the kind of evidence needed to support state scores' reliability and validity and provide an example for the construction of state items to assess conscientiousness states.

THE DEFINITION OF A PERSONALITY STATE

In our opinion, the most crucial element for the construction of personality state measures is, first and foremost, the definition of (i) the phenomenon personality state and (ii) the specific personality state at hand. By (i), we refer to broader questions, such as 'what is a personality state?', 'how is it related to a personality trait?', and 'under which circumstances does it occur?'. On the other hand, (ii) refers to the definition of a specific personality state, such as 'extraversion states' or 'narcissism states'.

One of the most recent definitions of states (note, not personality states, but states in general) was suggested by Baumert et al. (2017): A state is a 'quantitative dimension describing the degree/extent/level of coherent behaviors, thoughts and feelings at a particular time', and a state level is 'the individual momentary score on a scale measuring a state' (p. 528). Baumert and colleagues further elaborate that state dimensions could be used to describe differences within a person as well as between persons and that states tend to fluctuate from one moment to another (compared with personality traits, which are rather stable over time). Although this definition of states does not explicitly require that states are linked to personality, we would argue that (i) most *personality psychologists* think of states as the manifestation of personality (Baumert et al., 2017; Fleeson, 2001; Fleeson & Jayawickreme, 2015; Wrzus & Roberts, 2017) and that (ii) most recent *personality theories* posit that personality traits are expressed in states (Baumert et al., 2017; DeYoung, 2015; Eaton, South, & Krueger, 2009; Fleeson & Jayawickreme, 2015; Funder, 2001; Horstmann, Rauthmann, Sherman, & Ziegler, in revision; Horstmann, Rauthmann, & Sherman, 2018; Read, Smith, Droutman, & Miller, 2017; Sherman, Rauthmann, Brown, Serfass, & Jones, 2015; Tett & Burnett, 2003; Tett & Guterman, 2000; Wrzus & Mehl, 2015). The expression of a trait is therefore a trait manifestation or personality state (Horstmann, Rauthmann, & Sherman, 2018; Rauthmann, Horstmann, & Sherman, 2019). Note, however, that there are also other conceptualizations of personality traits (i.e. summarized as formative models), where the states form the trait (Buss & Craik, 1983; Horstmann, Rauthmann, Sherman, & Ziegler, accepted). In such cases, the definition of the trait follows from the definition of the state. For the current article, we assume that the state is always a manifestation of a trait and should, therefore, by definition, be linked to the trait.

So what constitutes the difference between a personality state and any other state? Personality states are explicitly linked to personality traits. Some have argued that this means that personality states must serve a specific purpose (Denissen & Penke, 2008) or function (M. Schmitt, 2009a) that is used to fulfil the need that arises from a specific standing on a personality trait. Schmitt (2009a) further elaborated that the quality of behavioural assessments (which, for now, we will equate with personality states, but see below for a discussion) depends on how well the definition of behaviour (i.e. personality states) is grounded in theory, that is, in what way it is linked to a well-defined nomological net. For example, a single personality state score such as choosing to agree five out of seven possible points to the statement 'I am dancing wildly' is most likely connected to the trait extraversion and thus considered an instance of extraverted behaviour. On the other hand, if a professional dancer in a dancing audition for *Step Up* (a bad movie with impressive dance scenes) gave the same response, this could be understood as an instance of conscientious behaviour, which served the need to advance her or his career and make a living. This has, for example, consequences for the examination of consistency: If the seemingly same behaviour does not serve the same purpose, it should also not be considered consistent behaviour (Fleeson & Noftle, 2008). Thus, not knowing why or how a certain behaviour (or personality state) was enacted makes the particular behaviour potentially meaningless with regard to the person's traits (Denissen & Penke, 2008) and unquestionably reduces the certainty with which it can be interpreted as a manifestation of a specific trait.

Consequently, there must be a distinction between instances of behaviour (Furr, 2009), such as watching TV, laughing, scratching ones nose, drinking tea, talking, and then again personality states, which are clearly located in the nomological net of the corresponding personality trait. Similar to emotions, the same instance of behaviour may have many different causes at different times, which is contrary to the current conception of personality traits (M. Schmitt, 2009a). Concerning personality traits, we assume that each trait is reflected in unidimensional indicators or an item in a trait questionnaire should only load on one single dimension. Contrarily, behaviours can be 'factorially complex' (Schmitt, 2009a, 2009b, p. 429). This means that one individual instance of behaviour may be caused by not only one personality trait but by several personality traits at the same time. It could even be argued that the same behaviour can be indicative of different traits in different situations, as exemplified above. Additionally, the expression of a personality trait, that is, the personality state, does not only depend on (multiple) characteristics of the person but may also be influenced by situational factors (Funder, 2006; Horstmann et al., in revision; Mischel & Shoda, 1995; Sherman et al., 2015). This means that a score on a state scale is linked not only to personality traits but also to situational factors. Consequently, state scores comprise several reliable variance sources [Equation 1], which is different to what is usually assumed for trait scores but similar to facet scores that also comprise two reliable variance sources: trait and facet (Ziegler & Bäckström, 2016). Thus, when looking at ways to estimate reliability and validity of behaviour based state scores, there is a need to decompose their variance into its constituents (M. Schmitt, 2009a, 2009b; Schönbrodt et al.,

DOI: 10.1002/per

submitted). Taken together, one can conclude that personality states are multi-determined. Accepting this definition of personality states has far-reaching consequences for the evaluation of their psychometric properties, as we will discuss below. However, not making things any easier, personality states are not only defined by current overt behaviour but also by thoughts and feelings.

Behaviour, thoughts, and feelings?

The previous definition by Baumert et al. (2017) defines states as 'coherent behaviors, thoughts, and feelings'. From there, it follows that one must consider the behaviour, thoughts, and feelings of a person to fully understand the nature of a state. For example, a person may be dancing, thinking about how to dance, and feel stressed. This could mean that this personality state is a function of the trait conscientiousness; on the other hand, if a person danced, thought about how to dance, but felt positive, this might be a function of the trait extraversion. Wilt and Revelle (2015) have examined personality trait scales and concluded that personality traits (at least broad personality traits such as the Big Five) are operationalized with different components, namely, Affect, Behaviour, Cognition, and Desire. They concluded that some personality traits are mostly defined via affect (i.e. neuroticism), overt and observable behaviour (e.g. extraversion or conscientiousness), cognition (i.e. openness), or a mix of affect, behaviour, and cognition (i.e. agreeableness). At the same time, desire was not very present in the content of the examined Big Five items. Yet this means that if a personality state should represent the expression of a personality trait, these components must also be reflected in the state measure. The assessment of a personality state only via the (self-rated) behaviour at a certain point in time will therefore fall short. Instead, and to ensure that the personality state is located in the nomological net of the corresponding trait, the content of the state scale should also correspond to the content of the personality trait. Thus, the inclusion of thoughts and feelings (and potentially desires) imposes conditions on the construction process of personality state items and touches the issue of content validity.

DECOMPOSITION OF STATES

What influences any given personality state? Decomposing a personality state score into different sources of variance allows a better interpretation of a personality state score and the derivation of concrete ideas on how to obtain reliability and validity evidence. For example, if the influences of different personality traits on the same score were consistent, this would mean that this score is not multi-determined. On the other hand, if the state score was highly influenced by momentary situational experiences, but consistently so, as well as by its corresponding personality trait, this would corroborate its interpretation as a valid state score. In Equation 1, we have highlighted different influences on personality states that could potentially be examined. Note that this equation only considers states (i) nested in persons (j).

$$p_{ij} = \gamma_{00} + \underbrace{\gamma_{01}P_{1j}}_{\substack{\text{personality}\\ \text{trait 1}}} + \underbrace{\gamma_{02}P_{2j}}_{\substack{\text{personality}\\ \text{trait 2}}} + \underbrace{\gamma_{0n}P_{kj}}_{\substack{\text{personality}\\ \text{trait } k}} + u_{oj} + \underbrace{\gamma_{10}a_{1ij} + u_{1j}a_{1ij}}_{\text{affect state 1}} + \underbrace{\gamma_{m0}a_{mij} + u_{mj}a_{mij}}_{\text{affect state } m} + \underbrace{\gamma_{30}s_{1ij} + u_{3j}s_{1ij}}_{\substack{\text{situation}\\ \text{characteristic 1}}} + \underbrace{\gamma_{l0}s_{lij} + u_{lj}s_{lij}}_{\substack{\text{situation}\\ \text{characteristic l}}} + e_{ij} \quad (1)$$

Here, p is the personality state of a person j at measurement occasions i. Empirically, this state can be decomposed in a general mean across all persons γ_{00}, which is arbitrary and dependent on the scale used. The state is then decomposed into time/occasion invariant elements (P_j), such as personality traits or response styles, and time-varying elements, such as affective or emotional states (a_{ij}) and psychological characteristics of the situation (s_{ij}). γ_{0k} denotes the effect of the kth trait across all persons j (i.e. person invariant), and γ_{m0}/γ_{l0} denotes the mth effect of affect/emotions a across all time points i or the lth effect of the characteristics of situation s across time points i. Further, u_{mj} denotes the m's deviation of the average effect for person j, of their affect or their current situation on their personality state. Note that this model could be extended, for example, nested in days or families, or with an additional level, accommodating personality traits nested in persons (i.e. the same trait assessed several times within the same person, across a longer period of time). Furthermore, we assume that the effect of any occasion-specific variable will be stable across time (e.g. for a particular person j, affect always has the same effect on the personality state across all occasions i). The equation presented here is thus still overly simplistic (e.g. it does not consider various interaction terms) yet serves the purpose to show that any given state score p_{ij} is influenced by multiple variables.

To further explain the formula, we come back to the dancing behaviour from above. Dancing at a specific occasion i by person j is clearly influenced by the person's interpretation of her or his situation (e.g. audition vs. party), the affective components (e.g. neutral vs. positive), and her or his personality traits (e.g. conscientiousness vs. extraversion). Of course, the act of dancing does not have to be influenced by only one domain per component; that is, both conscientiousness and extraversion could simultaneously influence dancing. Furthermore, during experience sampling of self-reports, we are not observing states but the self-reports of states. The process that leads from the recollection of 'I was dancing wildly' to 'choosing 4 out of 5' on a rating scale will itself be influenced by person-specific or occasion-specific characteristics. If, in the very best case, the person reported her or his state directly after its occurrence, she or he might be out of breath, or even annoyed by the prospect of having to fill out a survey. Similarly, having seen others dance even wilder could lead to the selection of a lower score, which would be a momentary frame of

DOI: 10.1002/per

reference. At the same time, stable response styles of that person can furthermore make her or his shift towards the middle or the extremes on the rating scale. As this example highlights again, the state score at occasion *i* is influenced by multiple constructs and thus multi-determined.

What is a state measure?

Before going into details about the construction of personality state measures and the estimation of evidence supporting the reliability and validity of personality state scores, it is necessary to define what a (personality) state measure is:

The aim of a personality state measure is to assess the manifestation of a personality trait in a random or pre-defined situation.

As an example, one could think about the assessment of extraverted or conscientious behaviour. Note that there are instances of personality assessment that may not fall under this definition, such as the repeated assessment, say, every 6 months, of developing or changing traits. In other words, the assessment of *current* levels of traits is different from the assessment of states. State measures are furthermore mostly ultra-brief scales that can be used to intensely and repeatedly assess states in everyday life, imposing as little burden on the participant as possible. Approaching state measures from the perspective of ultra-brief scales already has implications for the evaluation of validity or reliability evidence (Kemper, Trapp, Kathmann, Samuel, & Ziegler, 2018; Ziegler, Kemper, & Kruyen, 2014). For example, scores from short scales usually have lower internal consistency. Therefore, validity estimates of short scale scores that are based on correlations can be lower than are longer scales under certain conditions (Heene, Bollmann, & Bühner, 2014; Thalmayer, Saucier, & Eigenhuis, 2011).

CURRENT PRACTICES

Before engaging in a broader discussion about the development of personality state measures, we sought to examine how personality states are currently used and reported in the literature. We therefore reviewed several major personality journals[1] and extracted all studies since 1990 that examined personality states. For each journal, we used the search string '(ESM OR Experience Sampling OR Daily Diary) AND (State* OR Behavior OR Trait Manifestation)' and found $N = 156$ matching articles across all selected journals (the complete list of articles and search strings can be found online on the OSF[2]). These articles were then reviewed by three research assistants who extracted the constructs that were assessed during the experience sampling phase.[3] From these articles, the first author extracted those that explicitly examined trait manifestations. Studies that did not target trait manifestations, but, for example, happiness, positive affect, or current social context, were thus excluded from all further analyses. There were no other exclusion criteria. We also did not evaluate the studies on the appropriateness of the implementation (e.g. a sample size that would usually be too small for the envisioned analyses). This resulted in a total list of 24 studies (Table 1). From these 24 studies, we coded the scale used for all states, the number of items per scale, whether the scale was created ad hoc (i.e. items were newly generated or adapted instead of an existing scale used), and the estimators for reliability and validity evidence presented, as well as the reliability estimate.

As the results from our literature overview show, a large variability exists in the way measures for personality states have been developed, what scores were formed, and how evidence regarding the scores' reliability and validity has been reported. Most strikingly, the most common way of establishing evidence for the validity and reliability of state score interpretations is by averaging state scores within participants and treating the so-resulting scores as person variables. Subsequently, internal consistencies are reported as estimates of reliability (Fleeson, 2001; McCabe & Fleeson, 2016; Moskowitz, 1994). However, this approach does not estimate the reliability of the state score but the reliability of the average-state score, and these two scores can represent entirely different constructs (Fisher, Medaglia, & Jeronimus, 2018; Hektner et al., 2007; Schönbrodt et al., submitted). A similar approach has been taken to showcase evidence supporting the validity of state scores. Some studies reported the correlation of average-state scores with one-time trait self-reports (Fleeson, 2001; Moskowitz, 1994; Sherman et al., 2015). However, as Hektner and colleagues pointed out, an aggregate of a person-level variable, assessed during experience sampling, must not necessarily measure the same as a one-time trait assessment of the 'same' construct (Hektner et al., 2007). In other words, a correlation or internal consistency of an aggregated state score must not necessarily represent an adequate estimate of the individual state score's validity or reliability.

Secondly, huge differences exist in how state measures have been developed. With few exceptions (Bleidorn, 2009; Himmelstein et al., 2019; Moskowitz, 1994; Newman, Sachs, Stone, & Schwarz, 2019; Ostojic-Aitkens, Brooker, & Miller, 2019; Zimmermann et al., 2019, Study 2), the state measures were not tested and validated in an independent sample, before the data collection of the substantive study. The most common way to developing state measures thus far seems to be to take items or adjectives that were used to assess personality traits and transform them into state measures (e.g. Horstmann et al., in revision; Ziegler, Schroeter, Lüdtke, & Roemer, 2018). Fleeson (2001), for example, described how he developed the measures on the basis of existing adjective lists that were used to describe personality traits (Goldberg, 1992). Specifically, he extracted items that (i) loaded on the correct factor (at trait level), (ii) represented the content of the factor, (iii) could be used to describe

[1] *Journal of Personality and Social Psychology* (JPSP), *Journal of Personality* (JoP), *Psychological Science* (PsychScience), Psychological Assessment (PsychAssessment), *European Journal of Psychological Assessment* (EJPA), and *European Journal of Personality* (EJP). The *Journal of Research in Personality* was excluded, as our institution does not have access to it.
[2] https://osf.io/s7tu2/
[3] On the basis of feedback from an anonymous reviewer, we supplemented the initial search by the search term 'EMA' and 'Ecological Momentary Assessment'. We identified 46 additional articles. These were coded by the first author. Six additional articles were included in the review.

DOI: 10.1002/per

Table 1. Overview of personality state measures used in published literature since 1990

Article	Personality state	Ad hoc?	Construction	Rel-estimator	r_{tt}	Validity estimated via	# items
Moskowitz (1994)	Dominance	No	Pilot study	Cronbach α, average across all reports	0.54	Convergent and discriminant validity of aggregated scores	12
	Submissiveness	No			0.45		12
	Agreeableness	No			0.53		12
	Quarrelsome-ness	No			0.84		12
Harlow and Cantor (1995)	Socializing	Unclear	Unclear	-	-	-	1
Fleeson (2001), Study 1	Extraversion	Yes	Selected from Goldberg (1992) and De Raad, Hendriks, and Hofstee (1994). Selection criteria: loading on the correct factor; represent breadth of factor; easily used to describe behaviour; contain no emotion words	Cronbach α, test–retest reliability of means	0.72, 0.90	Convergent and discriminant associations across first and second half of reports	4
	Agreeableness	Yes			0.66, 0.94		4
	Conscientiousness	Yes			0.68, 0.87		4
	Emotional Stability	Yes			0.62, 0.90		4
	Intellect	Yes			0.69, 0.94		4
Fleeson (2001), Study 2	Extraversion	Yes	Selected from Goldberg (1992). Selection criteria: loading on the correct factor; represent breadth of factor; easily used to describe behaviour; a consistent and appropriate dictionary definition can be found	Cronbach α	0.75	Convergent and discriminant associations across first and second half of reports	5
	Agreeableness	Yes			0.78		5
	Conscientiousness	Yes			0.74		5
	Emotional Stability	Yes			0.75		5
	Intellect	Yes			0.51		5
Fleeson (2001), Study 3	Extraversion	Yes	Selected from Goldberg (1992). Selection criteria: loading on the correct factor; represent breadth of factor; easily used to describe behaviour	Cronbach α	0.78	Convergent and discriminant associations across first and second half of reports	6
	Agreeableness	Yes			0.83		6
	Conscientiousness	Yes			0.86		6
	Emotional Stability	Yes			0.76		6
	Intellect	Yes			0.68		6
Weinstein and Ryan (2010)	Prosocial behaviour	Yes	Unclear	-	-	-	1 + 2[a]
Hofmann, Baumeister, Förster, and Vohs (2012)	Self-control	Yes	Item assessing content of desire based on 'self-regulation literature and […] pretesting' (p. 1323) otherwise, unclear	-	-	-	Adaptive, depending on previous responses
Murray, Gomillion, Holmes, Harris, and Lamarche (2013)	Rejecting behaviour within relationship with partner	Yes	Unclear	-	-	-	9
	Selfish behaviour within relationship with partner	Yes		-	-	-	4
	Comforting behaviour within relationship with partner	Yes		-	-	-	7
	Communal behaviour within relationship with partner	Yes	Adapted from Clark and Grote's (1998) index	-	-	-	10
Tomko et al. (2014)	Impulsivity	Yes	Items selected from two other scales that 'reflected both behavioural and cognitive aspects of impulsivity'	Within-person and between-person reliability	$r_{within} = 0.56$ $r_{between} = 0.99$	Difference on average-state scores between groups (BPD and not BPD)	4

(Continues)

DOI: 10.1002/per

Table 1. (Continued)

Article	Personality state	Ad hoc?	Construction	Rel-estimator	r_{tt}	Validity estimated via	# items
Sherman et al. (2015)	Honesty/humility	Yes	'Inspired by Fleeson (2007) as well as Denissen, Geenen, Selfhout, and van Aken (2008)' (p. 877)	-	-	Convergent and discriminant validity of mean state scores with trait scores.	1
	Emotional stability	Yes		-	-		1
	Extraversion	Yes		-	-		2
	Agreeableness	Yes		-	-		1
	Conscientiousness	Yes		-	-		1
	Openness	Yes		-	-		1
Crowe et al. (2018)	Vulnerable narcissism	Yes	Based on expert ratings and factor analysis, selection of final set of 6 items is unclear	Within- and between-person reliability	r_{within} = 0.95 $r_{between}$ = 0.82	Correlations of average states with other narcissism and personality traits (convergent and discriminant)	6
Fleeson and Law (2015)	Emotional stability	Yes	Adjectives selected from Goldberg (1992)	Cronbach α	0.76	Discriminant: predicting second-half mean from all first-half means simultaneously, then comparing stand. Beta for the prediction from same trait's first-half mean with predictions from other traits' first-half means	4
	Extraversion	Yes			0.80		4
	Agreeableness	Yes			0.73		4
	Conscientiousness	Yes			0.73		4
	Openness	Yes			0.84		4
McCabe and Fleeson (2016)	Extraversion	Yes	Adjectives selected from Roberts et al. (2014); Saucier and Ostendorf (1999)	Cronbach α	0.82	-	6
	Conscientiousness	Yes			0.82	-	6
Pihet, De Ridder, and Suter (2017)	Antisocial behaviour	Yes	Items were developed for the current study	Cronbach α for each of the first 15 measurement occasions, then averaged. Within and between reliability	av. α = 0.79 r_{within} = 0.62 $r_{between}$ = 0.96	-	3
	Fear of punishment	Yes			av. α = 0.85 r_{within} = 0.78 $r_{between}$ = 0.98	-	3
	Impulsivity	Yes			av. α = 0.89 r_{within} = 0.53 $r_{between}$ = 0.92	-	3
	Negative Affect	Yes	Items taken from three different prior scales, and two items added additionally		av. α = 0.91 r_{within} = 0.69 $r_{between}$ = 0.95	-	7
Bleidorn (2009)	Emotional Stability	No	Adjectives selected from NEO-PI-R (Costa & McCrae, 1992) based on pre-study	MLM-based within- and between-person reliability estimate	0.66; 0.95	-	6
	Extraversion	No			0.77; 0.82	-	6
	Agreeableness	No			0.63; 0.97	-	6
	Conscientiousness	No			0.65; 0.96	-	6
	Openness	No			0.77; 0.93	-	6
Timmermans, Van Mechelen, and Kuppens (2010)	Interpersonal behaviour (mean communion, mean agency, flux communion, flux agency, interpersonal pulse, and interpersonal spin)	No	Interpersonal grid (Moskowitz & Zuroff, 2005)	Spearman–Brown split-half reliabilities	0.61, 0.65, 0.58, 0.82, 0.57, 0.64	-	1[b]
	Dominance	Yes		-	-	-	

(Continues)

DOI: 10.1002/per

Table 1. (Continued)

Article	Personality state	Ad hoc?	Construction	Rel-estimator	r_{tt}	Validity estimated via	# items
Himmelstein et al. (2019)			Items from the Social Behaviour Inventory (Moskowitz, 1994)				23, 6 items to measure each pole
	Affiliation	Yes		-	-	-	23, 6 items to measure each pole
Zimmermann et al. (2019), Study 1	Sociability	Yes	Selection from item pool, based on ML-EFA	Within-person and between-person reliability using MLM α, based on polychoric correlations	r_{within} = 0.63; $r_{between}$ = 0.77	Discriminant: correlations with other PDD scale means and	4
	Perfectionism	No			r_{within} = 0.63; $r_{between}$ = 0.71	with PID-5 domain scales	2
Zimmermann et al. (2019), Study 2	Sociability	No	Selected in Study 1		r_{within} = 0.68; $r_{between}$ = 0.79		4
	Perfectionism	No			r_{within} = 0.64; $r_{between}$ = 0.81		2
Forgeard et al. (2018)	Openness	Yes	Selection of items from Goldberg (1992) + 1 new item	Cronbach α at day 1	0.85	Longitudinal measurement invariance, convergent/discriminant at day 1	5
Aschwanden, Luchetti, and Allemand (2019)	Openness	No	Daily Behaviour Checklist (Timothy, Church et al., 2008)	-	-	-	10
	Neuroticism	No		-	-	-	10
Giacomin and Jordan (2016)	Narcissism	Yes	Adapted from NPI-16 (Ames, Rose, & Anderson, 2006)	Cronbach α	0.76	Discriminant: do relations between state narcissism and several outcomes hold when controlling for state self-esteem?	16
Wilt, Noftle, Fleeson, and Spain (2012), Study 1	Extraversion	Yes	Selection from Goldberg (1992)	Cronbach α	0.74		4
Wilt et al. (2012), Study 2	Extraversion	Yes[c]	Selection from Goldberg (1992)	Cronbach α	0.58		3
Newman et al. (2019)	Nostalgic experiences	No	Series of studies that led to a four-item measure	Nested α (Nezlek, 2017)	0.90	Correlations of average states with other narcissism and personality trait scores (i.e. convergent and discriminant validity)	4
Ostojic-Aitkens et al. (2019)	Mind-wandering	No	Taken from Killingsworth and Gilbert (2010)	-	-	Correlations of average states with other mind-wandering trait scores	1

Note: ML-FEA, multi-level exploratory factor analysis; MLM, multi-level model; Ad hoc?, Was the measure designed ad hoc, without a pre-test?; Rel-estimator, method used to estimate reliability, if 'Cronbach α', then the within-person mean was used as an estimator for internal consistencies; r_{tt}, estimation of reliability.

[a]Adaptive, if first response was positive, two additional items.

[b]The interpersonal grid measure is a 9 × 9 inch large piece of paper with two dimensions, assured–dominant to unassertive–submissive and cold–quarrelsome and warm–agreeable.

[c]Study 1 and study 2 contained different adjectives.

DOI: 10.1002/per

behaviour, and (iv) did not contain 'emotion words'. This approach has indeed led to high cross-temporal stability of average-state scores ($r > 0.44$) as well as high internal consistency estimates (>0.66) of average-state scores. On the other hand, these estimates of evidence supporting the scores' reliability and validity do not indicate whether the individual state scores at each measurement occasion are indeed valid or reliable, which we will come back to below. Note that this approach, although common for the development or adaptation of state measures, would not satisfy current best practices for the development of trait measures (e.g. AERA et al., 2014).

Validity evidence for state scores and validity evidence for average-state scores

It has been noted several times throughout this article that validity of the average-state scores (i.e. evidence, that the interpretation of average-state scores is valid) does not imply validity of the individual, underlying state scores. Why is that? This problem has previously been described using different terminology, namely, ergodicity, ecological fallacy, or the Simpson paradox (Fisher, Medaglia, & Jeronimus, 2018). These terms simply relate to the problem that statistics obtained at group level (such as the distribution of trait scores or the variance of aggregate state scores) must not generalize to the level of the individual. For example, it may be possible that the assumed structure of constructs at trait level (between person) is different compared with the structure at the individual level (within person). Dejonckheere et al. (2018), for example, compared the structure of positive and negative affect. Whereas positive affect and negative affect are independent at the trait level, they are not independent at the state level (Bleidorn & Peters, 2011; Dejonckheere et al., 2018). Although people high on positive affect may similarly experience high levels of negative affect on average, positive affect and negative affect are negatively correlated at the individual level. This may of course also be true for personality traits and states, and thus, the between person structure must not correspond to the within-person level.

Alternative scores estimated on experience sampling data

In the current article, we focus on the validity of individual state scores and average-state scores. Whereas an individual state score is the score p of person j at occasion i, average-state scores are computed as the average of all state scores p of person j across all measurement occasions i to n $\left(\overline{p_j} = \frac{1}{n}\sum p_i\right)$. Of course, it is possible to estimate any other person-specific parameter, such as weighted means, median, person-standard deviations (Jones, Brown, Serfass, & Sherman, 2017), or a number of other parameters (Dejonckheere et al., 2019; Wright & Zimmermann, 2019; Zimmermann et al., 2019), each of which technically needs to be validated in their own right and based on theoretical assumptions. The focus on average-state scores in the current study reflects its popularity as an estimate of a person characteristic (Table 1).

PURPOSES, VALIDITY, AND RELIABILITY OF STATE MEASURES

Before constructing or using any measure to assess personality states, it is important to ask which purpose the measure will serve (Moskowitz et al., 2009; Moskowitz & Russell, 2009; Ziegler, 2014). With respect to the assessment of personality states, at least two purposes come to mind: (i) the use of average-state scores as alternative measures of a person's trait level or (ii) the use of individual states as an assessment of a person's daily experiences. An overview of the different methods and possibilities for the examination of reliability and validity of personality state scores is presented in Table 2.

Average-state scores

Purpose

First, personality states could be assessed to obtain a proxy for the corresponding personality trait. Based on whole trait theory, the average of personality states should correlate with the trait of this person (Fleeson, 2001; Fleeson & Jayawickreme, 2015; Jayawickreme, Zachry, & Fleeson, 2019). Examples of such research include the examination of incremental validity of average-state scores across self-reported personality trait scores to predict informant reports (Finnigan & Vazire, 2018; Vazire & Mehl, 2008) or affect (Augustine & Larsen, 2012).

In case the purpose of the measure is the assessment of stable personality characteristics, the validity of the obtained score depends mostly on (i) the breadth of the construct reflected in the state items and (b) the sampling procedure during the experience sampling phase. If, for example, the personality trait contains strong components of affect and behaviour (e.g. extraversion), the state items should reflect this content. This could mean that several items are assessed at each measurement occasion, or that items are sampled at random at each measurement occasion in a planned missingness design. Planned missingness designs can perform well in experience sampling studies, given a reasonable number of participants and measurement occasions (Silvia, Kwapil, Walsh, & Myin-Germeys, 2014). It is furthermore important to sample states throughout the day, across the whole week. If, for example, participants were only assessed in the morning during work hours, it could result in a bias of assessments at the individual level and, therefore, in a biased person-level estimate of the average personality state (Horstmann & Rauthmann, in preparation).

Validity evidence

Evidence for the validity of aggregated state measures can be obtained similarly to the evidence that is obtained for self-reported or informant-reported personality traits (Table 2). First, structural validity can be obtained by averaging items across measurement occasions and fitting confirmatory factor models with items averaged across assessments, within persons, as indicators.[4] This model then indicates if

[4]Note that it is still an empirical question if ignoring the multi-level structure in experience sampling structure is tolerable. See Sengewald and Vetterlein (2015) for an empirical examination in the context of student evaluation.

DOI: 10.1002/per

Table 2. Recommendations for assessing validity and reliability of state measures

	Question answered	Trait and aggregated states	Individual state
Validity			
Structural validity	Do items that are intended to assess the same construct load on the intended factor?	Exploratory and confirmatory factor analysis	Multi-level confirmatory factor analysis (Muthén, 1994), measurement invariance across time (Vogelsmeier et al., 2019)
Convergent and discriminant validity	Is the score obtained sufficiently distinct from scores that represent different constructs? Is the score obtained sufficiently similar to scores that represent the same construct?	Manifest or latent correlations with scores intended to measure the same construct (convergent) or a different construct (discriminant) (Campbell & Fiske, 1959; Dumenci, 2000; Nussbeck et al., 2009)	Multi-level multi-trait-multi-method analysis (Eid et al., 2008; Maas et al., 2009)
Predictive/criterion validity	Does the score obtained predict relevant outcomes? Do groups that are known or treated to differ on the construct differ on the scores?	Correlation with theoretically relevant outcomes (Borsboom et al., 2004; Horstmann et al., 2019) Differences between groups that are known to differ (Tomko et al., 2014) or treated to differ (van Roekel et al., 2019)	Multi-level models (Snijders & Bosker, 1999) or continuous time modelling (Driver et al., 2017) Do scores obtained in the same situation differ between participants from different groups? (Horstmann et al., accepted)
Nomological homomorphy	Is the score obtained related to other constructs in the same way as a score that reflects the same construct?	Correlations with a set of correlates, regression on a set of correlates (Rauthmann et al., 2019; Ziegler et al., 2014)	-
Variability	Does the score vary sufficiently from one measurement occasion to another?	-	Amount of variance attributed to the person vs. variance attributed to the measurement occasion (intra-class correlation).
Reliability			
Internal consistency	How high are the average correlations among items belonging to one scale?	Cronbach's alpha (Cronbach, 1951), McDonald's omega (McDonald, 1999)	Hierarchical alpha (Nezlek, 2017; Schönbrodt et al., submitted), hierarchical omega (Bolger & Laurenceau, 2013).
Test–retest	Are the scores obtained at one measurement occasion correlated with scores at another measurement occasion?	Correlation between the same score obtained at two different time points (trait) or in two different measurement bursts (aggregated state)	Correlation of two states assessed under the same circumstances (Horstmann et al., accepted)
Split-half[a]	Are scores obtained in the first half of the questionnaire/ assessment period correlated with those from the second half/assessment period?	Correlation between the first set of items and the second set of items	Correlation between the first set of items and the second set of items, within each measurement occasion

[a]Similar to split-half reliabilities, different ways of splitting the items are possible (e.g. odd–even, random).

the average response of the participants per item loads on the same latent factor. Second, further validity evidence can be obtained from multi-trait-multi-method analyses (Campbell & Fiske, 1959). The underlying idea is that correlations of scores obtained to represent the same construct (convergent correlations) should be higher compared with correlations of scores obtained to represent different constructs (discriminant validity). At the same time, correlations of scores obtained with similar methods can give an estimate of the method (co-)variance, if the underlying constructs of these scores are theoretically independent. Furthermore, it is possible to represent the multi-trait-multi-method matrix in a latent model, which allows comparing latent correlations (Campbell & Fiske, 1959; Dumenci, 2000; Nussbeck, Eid, Geiser, Courvoisier, & Lischetzke, 2009). This provides the direct advantage that all scores are estimated without measurement error, which directly corrects for attenuation of correlations (i.e. correction for reliability of obtained scores). Concerning personality states, one would expect, for example, that the average of personality state scores of one domain correlates at least with the corresponding personality trait as a convergent measure. Average extraverted behaviour should, for example, correlate the highest with self-reported or informant-reported trait extraversion. At the same time, one could expect that the correlations with scores obtained to represent other domains should be substantially lower. Currently, this approach to validating personality state scores is probably used most often, and the findings are generally as expected, that is, high convergent correlations and lower discriminant correlations (Horstmann & Rauthmann, in preparation).

Third, evidence for the validity of the score's interpretation can be obtained by examining the extent to which the score predicts theoretically meaningful and relevant outcomes (Borsboom, Mellenbergh, & van Heerden, 2004; Horstmann, Knaut, & Ziegler, 2019). For example, if the average-state score of extraversion correlates with the number of parties one has visited during the last months, this may be seen as evidence for the score's interpretation as an estimate of the person's extraversion. Note that the criterion must not necessarily be obtained at the same time as the average-state score. The average-state score is usually interpreted as a time-invariant characteristic of the person

DOI: 10.1002/per

and therefore as stable (Fleeson, 2001; Jones et al., 2017). The average-state score should therefore be related to other person characteristics, regardless of the time of their assessment. However, this is of course only true if the period during which states are assessed is representative for the criterion focused.

Fourth, one can gather evidence for the average-state score's validity by examining its nomological network, specifically its nomological homomorphy with trait scores (Rauthmann et al., 2019). The idea is that trait scores are related to other correlates in their nomological net, and that if the average-state score and the trait score indeed reflect the same construct (Fleeson & Jayawickreme, 2015), the average-state score should be related to the nomological correlates in a similar way. The congruence of this relation then describes the nomological homomorphy of trait and average-state scores. Note that this is an extension to the examination of construct and criterion validity, as this approach employs regression models to examine the relation of average personality state scores to multiple correlates simultaneously. This approach is also similar to the suggestions for constructing personality short scales (Ziegler et al., 2014).

Finally, average-state measures should also capture differences in average daily experiences between saliently different groups. First, it is possible to manipulate states between groups. Using experience sampling in an experimental design where one group receives treatment to change behaviour and the other does not should be reflected in the average level of behaviour difference between groups, all else being equal (Hudson, Briley, Chopik, & Derringer, 2018; van Roekel, Heininga, Vrijen, Snippe, & Oldehinkel, 2019). Average-state scores should thus be sensitive to manipulations that target an individual's average experience. Second, it is also possible to examine groups that differ in their known levels of personality states and examine if the measure reflects upon these known differences (Tomko et al., 2014).

Reliability

There are many ways to examine an average-state score's reliability each with advantages and disadvantages. Once the items at state level have been averaged within persons, across measurement occasions, they could technically be treated similarly to items from trait questionnaires. First, it is possible to examine the internal consistency of these scales. Note that averaging items across measurement occasions will lead to indicators that are much more unidimensional, as the unique, occasion-specific variance will be minimized, and the application of Cronbach's alpha as an estimator of the internal consistency of the average-state scale is much more appropriate (Cronbach, 1951; Nezlek, 2017). However, a more suitable measure of internal consistency would be McDonald's Omega (McDonald, 1999). In both cases, the underlying assumption is, however, that a latent variable is assessed. If the average-state score is assumed to reflect the personality trait of the person, then this may be appropriate. However, if the purpose of the reliability estimation were to gauge the amount of reliable variance traceable to the specific trait, this procedure would be inappropriate because the state score could be multidimensional as highlighted above. Additionally, if one-item indicators are used, as it is regularly the case (Horstmann et al., in revision; Sherman et al., 2015), an estimation of internal consistency of the aggregate state scores is not possible.

Alternatively, and especially suited for one-item state measures, one can estimate the stability of the average-state score across several measurement occasions (Moskowitz et al., 2009). Correlating the average measurement of the first half of the experience sampling phase with that from the second half of the experience sampling phase yields an estimator of the average scores test–retest reliability.[5] Note, however, that this procedure confounds the stability of the aggregate state score with the reliability of the aggregate state score (which is a common problem when estimating test–retest correlations). The estimation of the reliability via the score's stability will therefore result in lower and more conservative estimates.

Individual state scores

Purpose

If, however, state scores are obtained to examine within-person processes, state measures should be able to validly capture the daily experiences of a person. In other words, it is important to examine 'whether explanations other than the participants' natural experience could account for their [participants'] […] responses' during experience sampling (Hektner et al., 2007). Individual state scores can, for example, be used to get a picture of a patient's everyday life or about within-person effects of situations on behaviour or vice versa.

Validity evidence

To test these assumptions, one can first examine the structural validity of the state items. For state items, one can do this using multi-level confirmatory factor analysis (ML-CFA). ML-CFA simultaneously considers the within-person and between-person structure and allows testing a model that considers between-person and within-person variance at the same time (Muthén, 1994). Second, if more than one construct is assessed during the experience sampling phase, preferably with multiple methods, multi-level multi-method analysis (Maas, Lensvelt-Mulders, & Hox, 2009) allows examining discriminant and convergent validity of the state scores. For example, Bleidorn and Peters (2011) showed that positive affect and negative affect are unrelated at the between level but correlated at the within-person level. Concerning state measures, one could, for example, investigate if the structure of personality (e.g. uncorrelated domain scores) also holds within person or if, as suggested, different tendencies to behave and thus actions (i.e. personality states) inhibit each other (Revelle & Condon, 2015), which must necessarily mean that personality states at the within-person level could not be independent from each other.

[5]Instead of splitting first with second half of t the states, one can also split (i) at random, (ii) by every other measurement occasion (e.g. 'odd–even'), (iii) by time of assessment (e.g. divide split by every third hour) or theoretically guided (e.g. split by weekdays vs. weekends).

DOI: 10.1002/per

Recently, Vogelsmeier, Vermunt, van Roekel, and De Roover (2019) suggested latent Markov (exploratory) factor analysis (LMEFA) to examine the within-person structure of psychological constructs. Specifically, they suggested that the structure of measurement models may change across time and between individuals. For example, it could be assumed that the Big Five personality states are usually represented by five independent domains. Yet it could be possible that, under certain circumstances, for example, stress, the structure of the measurement models for the corresponding Big Five states changed. Under those circumstances, it would no longer be possible to compare means or covariances of individual states across time. LMEFA allows exploring the structure of measurement models across time, within models, and identifying occasions at which measurement models are invariant and state scores are thus comparable. This approach therefore addresses the challenges of state assessment in a very sophisticated manner. At the same time, software for the implementation of such models is not yet openly available, and an implementation of this approach in open source software is needed.

Third, personality state scores should be related to specific, theoretically plausible outcomes at state level, even after controlling for potentially overlapping, other constructs. State scores and correlates can be assessed prior to the state, at the same time as the state, or after the state. Depending on the research question at hand and the structure of the data, these associations can, for example, be examined using multi-level regression models (Snijders & Bosker, 1999), or in case of outcomes that are assessed at a different time, continuous time models (Driver, Oud, & Voelkle, 2017; Voelkle, Oud, Davidov, & Schmidt, 2012). These analyses should reveal that (i) state scores can be used to predict theoretically plausible outcomes, (ii) that the state scores are linked to only these outcomes and not to all outcomes that were assessed at state level, and (iii) that these links remain substantial, even after controlling for potential covariates.

Recently, Sun and Vazire (2019) presented evidence for the validity of momentary state scores by means of structural equation modelling. Coming back to the claim that a state score is valid if it covers the momentary experiences of a participant, Sun and Vazire examined the convergence of self-reported personality states and informant-reported personality states. The informant reports were obtained by having coders rate audio snippets recorded during participants' daily lives, which were matched to the momentary self-reports. Thus, it was possible to examine the extent to which self-reported personality states were congruent with participants' momentary experiences. Although this is probably the most sophisticated way to examine the validity of personality state scores, Sun and Vazire's study shows how taxing such an undertaking can be. Additionally, it has to be noted that the validation of a measure and its use for answering a substantial research question should, technically, be separated. However, given the complexity of required study designs in ESM studies, it is understandable that this is very rarely the case.

Finally, state scores should show reasonable fluctuation from one measurement occasion to another to justify the examination of within-person processes (Ilies et al., 2007; Sherman et al., 2015). Variance of state measures should be attributable not only to the person but also to the measurement occasion or situation. This ratio of variance is usually expressed as an intra-class correlation (ICC; Bliese, 1998). If the ICC of state scores was near one, this would mean that all variances could be attributed to the person and that the measure was not sensitive to situational changes (or that no situational changes occurred, of course). On the other hand, if the ICC were near zero, this would mean that close to all variances had to be attributed to the situation or measurement occasion. However, if personality states are understood as the manifestation of stable personality traits, this would be counterintuitive and in opposition to the theoretical assumptions. Note, however, that the proportion of within-person variance, and thus the ICC, does depend not only on the construct that is examined but also on methodological considerations and study design (Podsakoff, Spoelma, Chawla, & Gabriel, 2019). ICCs for state measures can be expected to be between 0.20 and 0.50 (Horstmann et al., in revision; Podsakoff et al., 2019; Sherman et al., 2015; Sun & Vazire, 2019).

Reliability

Reliability refers to the precision with which a certain score is obtained and can be estimated using either internal consistency or test–retest correlations. For test–retest correlations, this means that if a score was to be assessed with perfect reliability, it would follow that the next time the score is assessed under the exact same circumstances, the same result would have to be obtained, *if* the true value that is reflected in the score had not changed. However, concerning personality states, we require that personality states *can* change and that personality state measures are sensitive to this change. In other words, using the same items across measurement occasions, we would only expect to obtain the exact same score during a second, third, or *n*th assessment, if none of the variables depicted in Equation 1 had changed. Specifically, the psychological situation would have to remain unchanged, the person's affect would be required to be the same, and the person's personality would also be required to be stable. However, it is highly unlikely that a person does ever experience the same situation twice (Horstmann et al., accepted). In other words, any estimate of a state score's reliability must necessarily deal with variability as an essential property of the state score. There are several ways to address this problem.

First, one could aim to estimate a test–retest reliability of state scores. The estimation of test–retest reliability is essentially a question of consistency. Fleeson and Noftle (2008) alert us to the fact that consistency of behaviour can be estimated across many different components, though. Although Fleeson and Noftle suggested a total of 36 different forms of consistency (different enactments: single, aggregate, contingent, and patterned, which were crossed with different definitions of similarity: absolute, relative, and ipsative, which were crossed with different competing determinants: time, situations, or behavioural content), test–retest reliability of state scores would be the examination of *relative stability* of *single enactments* across *time as competing determinant* (and not the situation, as this would have to be kept as similar as possible; see Horstmann et al., accepted). The most

DOI: 10.1002/per

challenging aspect to examine the test–retest reliability of state scores is keeping the personality state as stable as possible. Equation 1 showcases the elements that could be addressed to achieve this goal. These are (i) time point (e.g. of the day/week/month) within person, (ii) situational content, (iii) affective states, and (iv) personality characteristics. First, one could hold the time of assessment as constant as possible. For example, it could be possible to examine the test–retest reliability from one assessment at Monday, 9 a.m., to the next assessment at Monday, 9 a.m., the week after. Here, it is assumed that personality states at similar weekdays and times will also be more similar. Second, one could aim to extract situations that are similar in their psychological characteristics. For example, Horstmann and colleagues (under review) extracted situations from an experience sampling phase that had highly similar situational characteristics profiles, which would then allow the examination of test–retest reliability of personality state scores. Third, one could assess personality states of participants who are in similar affective states. Using mood induction techniques (Larsen & Ketelaar, 1991; Westermann, Spies, Stahl, & Hesse, 1996), participants could be brought into a similar mood and their states could then be assessed. Finally, one should make sure that the personality traits that are related to the personality state that is assessed have not changed. Of course, such changes are rare and slow (e.g. Lüdtke, Roberts, Trautwein, & Nagy, 2011; Roberts & Mroczek, 2008; Roberts, Walton, & Viechtbauer, 2006), but if they occurred, one would have to expect state changes—and thus decreased reliability estimates.

A second way to estimate reliability would be to treat each measurement occasion (or other unit in which several measurements occurred) as a separate study (Nezlek, 2017). In case of one-item measures, one could, for example, treat days or hours as separate studies (i.e. all assessments from Monday are assumed to be from one study, assessing the same construct). For each 'study', one can then estimate the reliability and finally aggregate all reliability estimates (Nezlek, 2017; O'Brien, 1990).

A third possibility to estimate the reliability of state scores is to assess the internal consistency of scales, while simultaneously modelling the nested structure of the data (Nezlek, 2017). To obtain an estimate of item-level reliability, a three-level model must be specified, with items nested in measurement occasions (e.g. days and time intervals), nested in persons. The item-level reliability is then defined as the occasion level variance divided by the occasion level variance plus the item-level variance (Nezlek, 2017, p. 152).[6] If all persons responded differently across occasions, but in the same way to all items *within occasions*, the internal consistency estimate would be close to 1. Note that, similar to the use of Cronbach's alpha at trait level, the estimation of nested alpha assumes that all items are parallel (i.e. that they are interchangeable indicators of the same latent construct and unidimensional and have equal error variances). If, however, items do not homogenously load on their latent variable (i.e. loadings are different), CFAs allow estimating a more precise estimate of reliability, Omega (McDonald, 1999), as it does not require unidimensionality of the items. Similar to alpha, it is also possible to estimate omega for nested data structures using ML-CFA (Geldhof, Preacher, & Zyphur, 2014; Raykov & du Toit, 2005). Geldhof et al. (2014) showed that omega is, under almost all circumstances, preferable to alpha in all cases for the estimation of within-level reliability.

To summarize, it is possible to estimate within-person-level estimates of reliability, similar to test–retest reliability or internal consistency. However, as our literature overview has indicated (Table 1), this is very rare in the current published literature. We therefore recommend publishing these estimates along with descriptive statistics of the scales. At the same time, it is currently (to our knowledge) unknown which effects the reliability of state scores has on the estimation of effects at the between and within levels. For person-level scores, for example, it is long known how the increase of the score's reliability would lead to an increase in its correlation with another score (Spearman, 1904, 1910). Although the same logic clearly applies to within-person estimates, the power to detect effects that are typically examined in an experience sampling study (e.g. fixed effects, random effects, and cross-level interactions) depends on much more than just the reliability of the level 1 score (e.g. the sample size at different levels, the reliability of the level 2 predictors, and the ICC; Bliese, 1998; Mathieu, Aguinis, Culpepper, & Chen, 2012). It is thus still an open question which level of reliability suffices to examine which effects in experience sampling studies, contingent on other study details.

THE ABC OF TEST CONSTRUCTION FOR PERSONALITY STATES

Before engaging in the construction of any psychometric measure, there are at least three different questions that should be answered (Ziegler, 2014): First, what is the construct being measured? Second, what is the intended purpose of the measure? And third, what is the targeted population? Answering these questions, both during the construction of any measure and also during their application, can help with the interpretation of the results.

What is the construct being measured

If constructing state measures, the first question that should be answered is 'what is the construct being measured?' Although this can at times seem straightforward and possible answers may be something like 'personality state of conscientiousness', one should explicitly look at the specific definition of the personality trait (that is expressed). Here, well fleshed-out definitions of personality traits can provide the first and most useful source of information. These definitions should provide descriptions of the personality trait manifestations (i.e. personality states).

[6]Code for running these analyses can be found in the OSM of Nezlek (2017) for the program HLM or at https://github.com/kthorstmann/horst/blob/master/R/nestedAlpha.R for the program R.

DOI: 10.1002/per

Second, the items that are then selected for the state measure should be (i) linked as well as possible to only one facet or trait and (ii) applicable to most situations that participants could encounter during their lives. This can, for example, be answered by asking participants how they did understand or interpret individual questions or by employing think-aloud-techniques (Ziegler, Kemper, & Lenzner, 2015). Note that items that are developed *for* an experience sampling study do not necessarily have to be validated *during* a full-experience sampling study. Alternatively, participants can be asked during a one-time assessment in their daily lives what they are currently doing, what they are feeling, and so on. This procedure may be sufficient to reduce a large number of items to an initial, manageable item pool that can then be further refined using more elaborate designs. As explained above, items that (i) are applicable to everyday contexts and (ii) assess the manifestation of stable interindividual differences should therefore yield medium ICCs. ICCs can therefore be a first indicator whether the items do indeed vary over time but are also related to stable person constructs. Personality state measures that yield extreme ICCs should therefore be avoided.

What is the intended purpose of the measure?

Similar to the previous point, the purpose of the measure should clearly be considered. Many different purposes come to mind. First, the measure could be put to use in a purely scientific context, that is, for the examination and testing of personality theories, such as hypotheses derived from whole trait theory (Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019) or the TESSERA framework (Wrzus & Roberts, 2017). In these cases, personality state measures must be *able* to provide meaningful information to evaluate these hypotheses. Consider, for example, whole trait theory. One of its central claims is that 'traits can be conceptualized as density distributions of states' (Fleeson & Jayawickreme, 2015, p. 82). In other words, state scores should form density distributions within persons. A state measure that would be able to test this hypothesis must therefore, in principle, be able to form density distributions, which is only the case if it can reliably capture within-person variance. With regard to Equation 1, this would mean that a reliability estimate needs to focus on the trait component within the equation. Statistical models employed should therefore be able to extract this variance.

The TESSERA framework, for example, suggests that certain expectations in a situation will lead to trait manifestations. If, for example, this hypothesis was to be tested in an experience sampling study, both the expectations and the states had to be assessed. A test of this hypothesis would only be possible if those two entities (expectations and states) were not correlated 'by default', for example, due to common method variance. As Bäckström, Björklund, and Larsson (2009) reported, formulating items in trait measures more neutrally (e.g. 'I swim regularly' vs. 'I love swimming') reduces their interrelatedness, or their common method variance (Bäckström & Björklund, 2013). Similarly, items that are developed for testing specific hypotheses about state–state relations must have sufficient discriminant validity. Otherwise, findings that, upon first sight, corroborate such theories as whole trait theory could turn out to be of lesser value than initially hoped (Horstmann et al., in revision). As a consequence, for providing reliability and validity evidence, the decomposition of variance is again vital [Equation 1]. Moreover, it is also important to gauge discriminant validity evidence while paying attention to such potential overlaps. Thus, statistical models need not only be able to decompose the variance sources within states for one trait but also to relate those variance sources to each other.

Second, state measures could be used to test the effectiveness of interventions in clinical studies (Magidson, Roberts, Collado-Rodriguez, & Lejuez, 2014; van Roekel et al., 2019). State measures should be sensitive to changes in the targeted variable. For example, Horstmann and colleagues (under review) assessed how consistent participants would behave in hypothetical, dissimilar situations. However, the authors used an altered version of the Big Five Aspect Scale (DeYoung, Quilty, & Peterson, 2007) to test mean changes in personality states. It could be argued that the Big Five Aspect Scale items are not sensitive to momentary changes in behaviour and are therefore unsuitable to capture changes owing to interventions at the state level. This means that state items intended to capture change must reach strong internal consistencies within each situation. Consequently, the statistical modelling approach must be able to decompose variance sources for each state measure occasion and relate those estimates across occasions within persons.

To summarize, when paying attention to the rather abstract purpose question while constructing a personality state measure, two specific questions have to be answered: 'What is the goal of this study?' And, 'Is the measure able to provide sufficient evidence to achieve this goal?' Future experience sampling studies can rely on previous measures that have been successful in providing answers to substantive questions (Table 1). For example, if the scores obtained with a specific measure clearly fit into a specific theory (e.g. the correlations were as expected), then it could be a valid strategy to use the existing tool instead of constructing a new one —especially if resources are scarce. However, the question whether a measure has indeed been successful has to be evaluated using many different criteria, which we aim to develop in the current paper. Alternatively, if no previous measure exists, one needs to construct a new one.

What is the targeted population?

Different from the use of most trait measures, state measures always target two populations: (i) the population of the participants and (ii) the population of situations in which participants are observed (Horstmann, Rauthmann, & Sherman, 2018; Horstmann, Ziegler, & Ziegler, 2018; Ziegler, Horstmann, & Ziegler, 2019).

Considering the population of participants is extremely relevant for the formulation of state items and their applicability to the participants' daily experiences. For example, experience sampling studies that target very young or older participants (e.g. Quintus, Egloff, & Wrzus, n.d.) must consider that their daily experiences are different from the average

DOI: 10.1002/per

daily experience of college students. So far, only very few studies have targeted personality states in older participants, and evidence for the applicability of personality state measures in these populations is therefore scarce. Although elderly participants have been assessed during experience sampling (e.g. Carstensen et al., 2011; Drewelies et al., 2018; Hülür et al., 2016), none of the published studies have examined personality states (defined as the manifestation of personality traits) in particular. Similarly, item difficulty needs to be adjusted for different populations. One can easily see that an item that has been used for the assessment of neuroticism states 'During the last hour, how worried were you?' (Finnigan & Vazire, 2018; Sun & Vazire, 2019) will result in very low scores in a healthy, young population. If the same item was used in a clinical sample, this might be very different. Items with extreme item difficulties can potentially lead to attenuated correlations owing to restricted variance (Sim & Rasiah, 2006) and can make it therefore harder to test hypotheses at the within-person level.

Second, the item content also depends on the targeted population of situations. Consider a regular college student, who is asked 'if, in the last 30 minutes, they talked to strangers'. If this student's life is assessed under normal circumstances, this item will most likely show some variance. If, on the other hand, the experience sampling phase falls into the exam period, the student may not be speaking to strangers as much. Of course, the very reason for examining how much someone spent time talking to strangers is to find out if they do it at all; in other words, it might not always be possible to know in advance which period or population of situations will be targeted. However, in some cases, it is clear that circumstances will be somewhat special (e.g. around Christmas and during holidays or exam periods). As a remedy, one can offer participants the option to indicate when they would like to participate (Roemer, Horstmann, & Ziegler, n.d.).

PRACTICAL RECOMMENDATIONS AND EXAMPLE: AN ITEM POOL FOR CONSCIENTIOUSNESS

Throughout the article, we have highlighted some major theoretical differences between the construction of trait versus state measures. However, these can sometimes be abstract and hard to grasp when they are put to practice. Specifically, it can be difficult to generate an initial set of items to assess states in a systematic way. In this section, we will therefore aim to give practical recommendations for the development of state measures, focusing explicitly on those aspects that are substantially different from the construction of trait measures. We will exemplify this process by developing an initial item pool for conscientiousness states.

For the development of a state measure, one needs to first generate the item content that should be captured with the measure. Moving on, one has to develop potential items. Then the appropriate instructions need to be selected, items need to be finalized, and an appropriate response format has to be selected. Here, we will exemplify how one can go through these stages, especially focusing on the generation of items and the item content. Initially, the generation of item content and the items should only be guided by theoretical principles and not hindered by practical considerations. This means that, at first, an extensive list of potential items should be created. These are then later pruned to a practically manageable size. However, the practical considerations should be ignored at first to come up with the most appropriate measure of the state at hand.

ABC: State measure of conscientiousness

What is the construct being measured?

Conscientiousness at trait level is a very well-researched construct at the between-person level (Bogg & Roberts, 2004; Roberts, Chernyshenko, Stark, & Goldberg, 2005; Roberts, Jackson, Edmonds, & Meints, 2009; Roberts, Lejuez, Krueger, Richards, & Hill, 2014; Soto, 2019; Watson, 2001) and also at the within-person level (Chapman & Goldberg, 2017; Church, Katigbak, Miramontes, del Prado, & Cabrera, 2007; Hudson et al., 2018; Hudson & Fraley, 2015; Jackson et al., 2010; Magidson et al., 2014). It is therefore not surprising to find many slightly varying definitions of the construct in the published literature. Roberts et al. (2014) defined conscientiousness as the 'propensity to be self-controlled, responsible to others, hardworking, orderly, and rule abiding' (p. 1315). This shows, on the one hand, the many different states that may be seen as manifestations of conscientiousness (i.e. self-controlled, responsible, hardworking, and rule abiding). On the other hand, this alerts us to the breadth of the construct, and the fact that it may therefore not be possible to assess conscientiousness at state level with only one or two items.

What is the intended use of the measure?

The state measure for conscientiousness that we envision here is meant to be used for the assessment of conscientiousness states (as opposed to a conscientiousness trait that is based on averaged conscientiousness traits). The measure should furthermore be used by participants several times per day to indicate their momentary levels of conscientiousness states. This means that the state measure must be applicable to a variety of states in different situations as well as a range of conscientiousness states, depending on the targeted population of persons who enact conscientiousness states, and the anticipated situations in which conscientiousness states are enacted.

What is the targeted population?

As argued before, the targeted population of (i) persons and (ii) situations informs the item format, the item content, and the item difficulties. For this conscientiousness state measure, meant for research, we would primarily target students. Fortunately (and, yes, we deliberately chose a comparatively simple case), a lot is known about students' everyday lives. First, most scholars have once been students themselves and have an intuitive understanding of students' lives. Second, there is some research that has examined personality states or manifest behaviours of either the general population (Chapman & Goldberg, 2017) or student populations (Harari et al., 2017; Stachl et al., 2017). This means that it is possible

DOI: 10.1002/per

to build on the available literature to get an understanding of participants' lives, which can inform the relevant content of the items.

Defining the item content and designing the items

After the construct, the intended use of the obtained scores, and the targeted populations are defined, it is first necessary to generate item content that reflects the intended personality states. Generally speaking, this means that—ideally—a very long list of trait manifestations of conscientiousness should be assembled. In a second step, this content will then be transformed into items.

Item content

The search for item content generally benefits from an ever-growing nomological net. In some cases, as it is the case with conscientiousness, the nomological net is already very well established (e.g. Roberts et al., 2005, 2009). When organizing the literature and thinking about the enactments of conscientiousness, it is helpful to think in terms of *antecedents* (How does one change in conscientiousness?), *correlates* (Which acts are associated with conscientiousness?), and *consequences* (What are consequences of conscientiousness?). With respect to conscientiousness, these questions are comparatively easy to answer, given the available literature. The principle on which the development of item content then rests is the assumption that any antecedent, correlate, or consequence of conscientiousness traits *should be tied to the trait via manifestations of the trait, that is, states*. In other words, a correlate or consequence of conscientiousness does not just fall from the sky, and neither does a person change their conscientiousness trait on the fly. For example, conscientiousness has been linked to physical health (e.g. Bogg & Roberts, 2004; Roberts et al., 2014). Now, if one assumes that conscientiousness *leads* to physical health, which might be reasonable to do, one can generate state items that mediate this effect. Any effect of or on a trait must be mediated by a number of related states: Only enacted states have consequences, not traits per se. One should therefore generate state items that could potentially mediate the relation between antecedents, correlates, and consequences of traits.

On the basis of existing literature, we collected a number of possible antecedents, correlates, and consequences of the targeted trait conscientiousness (Table 3). For example, Hudson et al. (2018) showed that completing certain weekly challenges can lead to higher conscientiousness in participants. These challenges are then an antecedent of conscientiousness, and the enactment of pursuing the challenge can be considered a conscientiousness state. Enacting this state over and over again then leads to higher conscientiousness. Similarly, Church, Katigbak, Miramontes, del Prado, and Cabrera (2007) listed correlates of conscientiousness, such as getting a good grade. Here, getting the grade itself (e.g. the moment one sees the grade) clearly has nothing to do with conscientiousness, but some intermediate process, such as learning for the exam, could have led to the good grade. Finally, there are consequences of conscientiousness. Bogg and Roberts (2004), for example, show that persons who score high on conscientiousness drink less alcohol than do low-scoring persons. Note that Bogg and Roberts explicitly assume that drinking less is a consequence of conscientiousness, although the effect might also occur in the opposite direction (i.e. drinking over an extended period of time leads to lower conscientiousness after a while).

ABCDs

With respect to conscientiousness, most states have a behavioural content. This is very similar to what Wilt and Revelle (2015) observed in their analysis of Big Five trait items. Nevertheless, Chapman and Goldberg (2017) also reported that participants who scored high on conscientiousness spent less time daydreaming. Not daydreaming can therefore be seen as one of the rare cognitive components of conscientiousness. Similarly, the item 'I complete tasks well because I want to' can be seen as a desire rather than a behaviour. These items, if correctly identified, can broaden the representation of the construct.

Additional ways to generate item content

It may not always be possible to rely on such a well-developed nomological net as we did here. In such cases, it may be more difficult to come up with a list of potential item contents. To overcome this problem, one may of course first develop such a nomological net and conduct studies that are similar to those listed in Table 3 (especially studies that resemble Hudson et al., 2018; Jackson et al., 2010; or Magidson et al., 2014, as these have provided much detailed information about what conscientious people do). Alternatively, one may be inspired by items from already published trait questionnaires.

Item content from trait questionnaires. Another source of inspiration for item content for state measures is the trait measures of the corresponding trait. For example, the BFI-2 (Soto & John, 2017) contains items such as 'Tends to be disorganized' (reverse coded), 'Tends to be lazy' (reverse coded), or 'Is dependable, steady' (Soto & John, 2017, p. 142). Being organized, not lazy, or dependable is therefore seen as qualities of the conscientious person. This approach may of course be supplemented with items from other sources, especially if the items are freely available (i.e. non-proprietary), as in the International Personality Item Pool (IPIP, 2015), or the Synthetic Aperture Personality Assessment project (Condon, 2018; Condon & Revelle, 2015; Condon, Roney, & Revelle, 2017).

Critical incident technique. The critical incident technique is described as a 'procedure for gathering certain important facts concerning behavior in defined situations' (Flanagan, 1954, p. 9). Although initially designed to define behaviours that were critical for either the failure or the success of a person in a specific situation, this technique can be used to generate descriptions of personality states. For example, to obtain descriptions of the personality state conscientiousness, an interviewer can ask the interviewee (who should, at best, be a member of the population that is later assessed, in this case, a student) to think about typical situations in their everyday life. It is important that the situation is appropriate for the personality state. Appropriateness means that the trait can be manifested in

DOI: 10.1002/per

Table 3. Item content and items for a potential scale to assess state conscientiousness

Source	Method/background	Item content	Actual item
Antecedents			
Magidson et al. (2014)	Behavioural activation to increase conscientiousness in a single clinical case	- Pick up daughter from school - Hand money over to wife after receiving paycheck - Go regularly to meetings of Narcotics Anonymous	- I fulfilled a duty. - I fulfilled a regular obligation. - I attended a regular meeting.
Hudson et al. (2018)	Weekly challenges that needed to be completed to change behaviour	- Show up 5 minutes early for every class, appointment, or activity on your daily schedule - Plan out a full day, hour by hour, putting all classes, appointments, and social activities on a calendar	- I arrived before time for my last meeting. - When I finished my last task, I had already planned exactly what I had to do next.
Correlates			
Church et al. (2007)	Act frequency approach	- Got a good grade on an assignment or exam - Finished a task on time - Did an important task well	- I learned for an upcoming exam. - My last activity took as long as planned. - I devoted attention to important tasks.
MacCann, Duckworth, and Roberts (2009)	Attributes that differentiate between persons with high vs. low conscientiousness scores	- Class absence (r) - SPORT absence (r) - Disciplinary infraction (r) - High honours	- I missed a class. (r) - I skipped gym. (r) - I had trouble with authorities. (r) - I did exceptionally well in a task.
Raynor and Levine (2009)	Self-reported correlates of conscientiousness	- Seatbelt use - Exercise - Restful sleep - Fruit or vegetable servings	- I took care of my safety. - I exercised. - I am well rested. - I ate something healthy.
Jackson et al. (2010)	Behavioural indicators of conscientiousness	- Tell a child a rule for proper etiquette - Leave unfinished food sitting out (r) - Take wrong materials to class or work (r)	- I told someone to behave adequately. - I left something lying around. (r) - I brought along all the utensils I need right now.
Chapman and Goldberg (2017)	Act frequency approach	- Swore around other people (r) - Spend an hour at a time daydreaming (r)	- I've got myself under control. - I daydreamed. (r)
Consequences			
Watson (2001)	Correlation with procrastination domains	- Exams (r) - Reading assignments (r) - Attending meetings (r) - Academic tasks	- I stuck to my schedule. - I read assigned coursework. - I attended a scheduled meeting. - I completed an academic task.
Bogg and Roberts (2004)	Meta-analysis of relation of conscientiousness and health related outcomes	- Excessive alcohol use (r) - Drug use (r) - Unhealthy eating (r) - Risky driving (r) - Risky sex (r)	- I drank more than I planned to. (r) - I could not refrain from visiting places that trigger my drug use. (r) - I thought about getting fast food. (r) - I drove too fast. (r) - I resisted a temptation.
Soto (2019)	Prediction of outcomes	- Antisocial behaviour - Intrinsic success - Religious behaviour	- I littered. (r) - I complete tasks well because I want to. - I did something to fulfil my religious duties.
Wilmot and Ones (2019)	Meta-analysis of relation of conscientiousness and outcomes at work	- Procrastination (r) - Need for competence - Academic dishonesty (r) - Antisocial behaviour (r) - Overall job performance	- I left a task unfinished. (r) - I double-checked something. - I copied a coursework from someone else. (r) - I took materials from work. (r) - I made sure that I complete all my assigned tasks well.
Trait questionnaires			
Soto and John (2017)	Trait scale (BFI-2) to assess the Big Five personality traits	- Tends to be disorganized (r) - Tends to be lazy (r) - Is dependable, steady	- I do not know what to do next (r) - I have fulfilled all my duties. - I kept a promise.
Critical incident technique			
Critical incident techniques with students		- Learning every day for 3 hours before doing something else	- I did what I planned to do before doing something else

Note. (r) indicates a reverse coded item, that is, in the direction of low conscientiousness. Item content is taken from the studies listed in the column *Source*. Items were designed assuming that students are the targeted population and that the items are therefore applicable to students' everyday lives.

DOI: 10.1002/per

this situation; that is, the situation must be relevant to the expression of the trait (Tett & Burnett, 2003; Tett & Guterman, 2000). The appropriateness must be assessed by the interviewer, who therefore needs to be an expert on the construct. Again, the nomological net, especially the assumed consequences of the corresponding personality trait, can be helpful in guiding this decision. If, for example, highly conscientious persons are more successful in their job than are less conscientious persons, what exactly was a critical situation at work that might have led to success or failure? For example, a student could suggest a situation such as 'studying for an exam'. One can now proceed and ask the student about behaviours, thoughts, feelings, and so forth that they experienced that have led to success in this situation. Note that at this stage, the interviewee may name a number of different states that may or may not be classified as manifestations of conscientiousness, such as 'was open to suggestions on how to improve learning', 'learning every day for 3 hours before doing something else', 'partying less during exam period', 'saying no to invitations from others', or 'did not worry about missing out on positive experiences'. Although all of these are manifestations of the Big Five personality traits (i.e. openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism), only the second one can be classified as a manifestation of conscientiousness.

Item wording

When choosing a wording for the items, one should always keep in mind the responses to the questions 'what is the targeted population?' and 'what is the intended use of the measure?' The actual items are developed based on the extracted item content. Sometimes, this is very straightforward, as the item wording is very similar to its content. For example, the correlate 'exercise' can be transformed into an item such as 'I exercised'. Other content, which is very specific, has to be made broader in order to apply to several different occasions. For example, the content 'hand over money to wife after receiving paycheck' is too specific to apply to many occasions, as a paycheck is only obtained once a week or once a month. This content may be transformed into a more general item such as 'I fulfilled a regular obligation'. Finally, some content may be so abstract that it does not reflect a state and must therefore be made more specific. For example, the content 'antisocial behaviour' could be transformed into an item such as 'I littered' or 'I've done something inappropriate'. These items would then be more applicable to a broad range of participants and daily situations.

When developing items based on content from previously published trait scales, one has to keep in mind that the content is typically rather broad (e.g. being disorganized, lazy or dependable, see Table 3). These comparatively general qualities will then also have to be 'translated' into more specific behaviours, thoughts, or feelings that could have occurred in several occasions. Being disorganized, for example, would mean, among other things, 'not knowing what to do next'. This would then be applicable to nearly all potential measurement occasions in daily life.

The item content obtained during the critical incident technique can be transformed to state items in a very similar way. Similar to other content, 'I learned for about three hours before I did something else' might be much too specific. Integrating this statement with statements from the same person as well as other interviewees can however result in general principles such as 'I did what I planned to do before doing something else', or 'I have stuck to my plans', or even more specific with respect to the reference period 'Right now, I am sticking to my plans'.

Items, scales, and instructions

Not only is it necessary to define the wording of the items, but it is also necessary to adhere to general principles of item design, choosing the most adequate rating scale, and to choose the correct instructions.

Item design. Throughout the history of psychology, much has been said about the way items in questionnaires should be designed (e.g. Krosnick & Presser, 2010). For example, items should be written in simple syntax, using familiar words, and the wording should be specific and concrete as opposed to general and abstract. Here it is assumed, however, that the respondent will most likely only take the survey once or, if more than once, with some longer time interval in between assessments, and that a general characteristic of the person is being assessed, such as their personality traits. Both do not apply to the assessment of personality states. The first consequence is that the time to which the item refers (e.g. 'within the last hour', 'just now', 'today', and 'recently') should not refer to a period that is longer than the time between two adjacent measurement occasions. Empirical evidence that can inform this decision more concretely is, to our knowledge, not available, though.

Secondly, a state item does not asses general and time-invariant characteristics of the person, but a momentary state. This is reflected in the level of hierarchy of the construct that it refers to. Therefore, the content to which it relates should be reasonably concrete but, at the same time, applicable to as many assessment situations as possible. On the one hand, items can be worded such that they are broadly applicable in everyday life. At the broadest level, such an item could be 'I behaved conscientiously' or a little less broad 'I did what I planned to do'. Note that these items, although very broad in their description, already focus solely on the behavioural aspect of conscientiousness (i.e. 'behaved' and 'did'), and not on any other aspect, such as thoughts (e.g. the planning itself) or possibly related feelings (e.g. the satisfaction when having followed through with ones plans). On the other hand, items can be worded very narrowly, such as 'I checked my manuscript for spelling errors'. This very specific behaviour is, however, so concrete that it may not readily apply to all possible situations, and certainly only to a very specific population (i.e. those who write their own manuscripts).

Items can now be crafted in line with these two dimensions: reference period and breadth of construct. The reference period can be very short (i.e. 'just now') or indefinitely long (e.g. 'in the last year', 'in general'). For state items, one should aim for a very short interval. How

DOI: 10.1002/per

low one can go, is, however also determined by the item content. Some states may simply not occur that often and asking the participant whether this state was recently manifested may not be applicable. Concerning the breadths of the construct, one should aim to formulate items such that they can be as specific as possible. Again, the content may limit the specificity of the item, depending on the targeted population of persons and situations.

Note that if an event-contingent plan is used, the reference period as well as the content of the items can of course be much shorter and more specific. If, for example, a signal to participate is triggered whenever a person writes on their manuscript, the reference period can be very short, as the item refers to exactly this particular moment (e.g. the item 'I am checking my spelling' will always be applicable if the event 'writing a manuscript' has triggered the signal—we just know that the person was writing right now). Similarly, the content can be much more specific (e.g. 'I was revising what I wrote yesterday' vs. 'I was writing'), as it is already known that the person was writing a manuscript.

Generally speaking, items can be written to combine any reference period with any breadth of item content. The decision should be informed by the anticipated precision with which items *can* be answered. On the one hand, the item could be so specific that it would not be possible to respond to it in most situations, as the specific item content is not relevant. On the other hand, an item could be too broad; in both cases, one is more likely to assess the person's trait-like tendencies compared with their state-like tendencies (Robinson & Clore, 2002a, 2002b).

Number of response options. A number of research articles have examined the effect of the response format of an item on basic psychometric properties of the item and the resulting scale score (e.g. Lee & Paek, 2014; Simms, Zelazny, Williams, & Bernstein, 2019). However, with respect to state measures, such research is practically non-existent (Wright & Zimmermann, 2019). On the one hand, the offered response options should allow a differentiation between a range of state levels. Similar to trait measures, offering between five to eight response options seems reasonable. On the other hand, providing too many response options may confuse participants and lead to longer response rates or additional measurement error. Another factor that may only be relevant for state items is the available screen width of the device that is used during data collection (usually a smartphone). If the screen is narrow, having too many response options may result in each response option being also displayed narrow, which again could result in wrongly selected responses. Similarly, a narrow screen may not allow naming each response option, and a label may only be given to the endpoints and potentially the mid-point of the scale. Alternatively, one can choose sliders for collecting responses from participants. However, these sliders have a particular drawback, namely, that in most cases, a certain response is pre-selected (i.e. the slider has to start somewhere). Whether or not any of the decisions made with respect to answer format affect the psychometric properties of items remains, ultimately, an empirical question yet to be examined.

Instructions. In an experience sampling study, there are two instructions that can, but do not have to be, the same. First, at the initial assessment, participants may be informed on how to respond to the questionnaire and receive detailed instructions. Second, during the experience sampling, participants should receive a short instruction on how to respond to each item. Here, it should be pointed out which aspects of the states the items refer to (e.g. thoughts, feelings, and behaviour) and to which reference period (unless these elements are present in the items themselves). The instructions, especially those during the experience sampling, should be kept reasonably brief to avoid unnecessary participant burden.

Further steps

After the initial item pool has been constructed, the items will have to be presented to members of the targeted population. Note that for an initial test of this item pool, it may not be required to do this in a full-experience sampling study but can also be achieved by asking participants to rate only their current situation once, by means of experimental manipulation of a situation, and so on (Table 2). After the first sample of respondents has been collected, the item statistics, intercorrelations, and estimates of validity and reliability have to be computed. Although the technique may be different compared with trait measures, the principles are comparatively similar, as laid out before.

CONCLUSION AND OUTLOOK

Taking a look back at the rapid development of the experience sampling method, from booklets being handed out to participants, answers being recorded with SMS, to palm-held computers and finally to smartphones that are now readily available, it is clear that data obtained from experience sampling will become more and more important for (personality) psychologists. Additionally, newer methods such as life logging (e.g. Brown, Blake, & Sherman, 2017) or the electronically activated recorder (Mehl, Pennebaker, Crow, Dabbs, & Price, 2001; Sun & Vazire, 2019) further extend the psychologists' toolbox and provide us with a very powerful repertoire to examine personality dynamics. Although research on experience sampling methods and psychometric evaluation of data obtained in experience sampling has come a long way, many challenges remain. We argue that is important to accept that personality state scores are multidimensional. If this idea is accepted, it will help in guiding the interpretation of scores and results obtained using experience sampling.

Next, it is important, as indicated above, to develop quality guidelines for the examination of personality state measures. Not all psychometric evidence reported for a specific measure is useful, and not all evidence is required, providing exactly the evidence that the purpose of the measure requires should be the goal of each test construction (Ziegler, 2014).

DOI: 10.1002/per

We assume that our first overview of current practices, options, and methods will be somewhat outdated in a few years, but we look forward to the broad application of methods that only very few are currently thinking (or, via Twitter, heavily arguing) about.

Psychologists have learned a great deal during the development of theories such as the Big Five and the herein embedded development of trait measures. As a consequence, agreed-upon guidelines exist, and methods for the construction and evaluation of trait questionnaires are (hopefully) included in every undergraduate curriculum. This also means that we have the chance to avoid making the same mistakes that were made during the construction of trait measures. As one of the most outstanding practices, alpha maximization comes to mind (N. Schmitt, 1996). This describes the poor practice of selecting items for a questionnaire such that its internal consistency is maximized often at the cost of heterogeneity and content validity.

Finally, many open questions, mostly of methodological nature, remain unanswered. Given the importance of data from experience sampling for the examination of recent personality theories, we hope that these questions, some of which are listed in this paper, will spur new research. Data on personality states are already widely available and only waits to be analysed. Similar to the beginning of trait research when factor analytical methods were developed to explore personality structure, we may now hope for a new era of method development fostering the examination of dynamic personality.

ACKNOWLEDGEMENTS

We thank our three student assistants, Lilly Buck, Maximilian Ernst, and Aaron Peikert, for their help with the literature review. We also thank Clemens Stachl for helpful suggestions.

REFERENCES

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*, 440–450. https://doi.org/10.1016/j.jrp.2005.03.002.

Aschwanden, D., Luchetti, M., & Allemand, M. (2019). Are open and neurotic behaviors related to cognitive behaviors in daily life of older adults? *Journal of Personality*, *87*, 472–484. https://doi.org/10.1111/jopy.12409.

Augustine, A. A., & Larsen, R. J. (2012). Is a trait really the mean of states? *Journal of Individual Differences*, *33*, 131–137. https://doi.org/10.1027/1614-0001/a000083.

Bäckström, M., & Björklund, F. (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. *Scandinavian Journal of Psychology*, *54*, 152–159. https://doi.org/10.1111/sjop.12015.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-Factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*, 335–344. https://doi.org/10.1016/j.jrp.2008.12.013.

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., … Wrzus, C. (2017). Integrating personality structure, personality process, and personality development. *European Journal of Personality*, *31*, 503–528. https://doi.org/10.1002/per.2115.

Bleidorn, W. (2009). Linking personality states, current social roles and major life goals. *European Journal of Personality*, *23*, 509–530. https://doi.org/10.1002/per.731.

Bleidorn, W., & Peters, A.-L. (2011). A multilevel multitrait-multimethod analysis of self- and peer-reported daily affective experiences. *European Journal of Personality*, *25*, 398–408. https://doi.org/10.1002/per.804.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, *1*, 355–373. https://doi.org/10.1177/109442819814001.

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. *Psychological Bulletin*, *130*, 887–919. https://doi.org/10.1037/0033-2909.130.6.887.

Bolger, N., & Laurenceau, J. P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440. https://doi.org/10.1007/s11336-006-1447-6.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061.

Brown, N. A., Blake, A. B., & Sherman, R. A. (2017). A snapshot of the life as lived. *Social Psychological and Personality Science*, *8*, 592–600. https://doi.org/10.1177/1948550617703170.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review*, *90*, 105–126. Retrieved from. https://psycnet.apa.org/record/1983-23438-001.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105. https://doi.org/10.1037/h0046016.

Carstensen, L. L., Turan, B., Scheibe, S., Ram, N., Ersner-Hershfield, H., Samanez-Larkin, G. R., … Nesselroade, J. R. (2011). Emotional experience improves with age: Evidence based on over 10 years of experience sampling. *Psychology and Aging*, *26*, 21–33. https://doi.org/10.1037/a0021285.

Chapman, B. P., & Goldberg, L. R. (2017). Act-frequency signatures of the Big Five. *Personality and Individual Differences*, *116*, 201–205. https://doi.org/10.1016/j.paid.2017.04.049.

Church, A. T., Katigbak, M. S., Miramontes, L. G., del Prado, A. M., & Cabrera, H. F. (2007). Culture and the behavioural manifestations of traits: An application of the act frequency approach. *European Journal of Personality*, *21*, 389–417. https://doi.org/10.1002/per.631.

Church, A. T., Katigbak, M. S., Reyes, J. A. S., Salanga, M. G. C., Miramontes, L. A., & Adams, N. B. (2008). Prediction and cross-situational consistency of daily behavior across cultures: Testing trait and cultural psychology perspectives. *Journal of Research in Personality*, *42*, 1199–1215. https://doi.org/10.1016/j.jrp.2008.03.007.

Clark, M. S., & Grote, N. K. (1998). Why aren't indices of relationship costs always negatively related to indices of relationship quality? *Personality and Social Psychology Review*, *2*, 2–17. https://doi.org/10.1207/s15327957pspr0201_1.

Condon, D. M. (2018). *The SAPA personality inventory: An empirically-derived, hierarchically-organized self-report personality assessment model*. https://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, 3. https://doi.org/10.5334/jopd.al.

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA project update: On the structure of phrased self-report personality items.

Journal of Open Psychology Data, 5. https://doi.org/10.5334/jopd.32.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. https://doi.org/10.1037/h0040957.

Crowe, M. L., Edershile, E. A., Wright, A. G. C., Campbell, W. K., Lynam, D. R., & Miller, J. D. (2018). Development and validation of the narcissistic vulnerability scale: An adjective rating scale. *Psychological Assessment*, *30*, 978–983. https://doi.org/10.1037/pas0000578.

De Raad, B., Hendriks, A. A. J., & Hofstee, W. K. B. (1994). The Big Five: A tip of the iceberg of individual differences. In C. F. Halverson Jr., G. A. Kohnstamm, & R. P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 91–109). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., … Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*, 323–341. https://doi.org/10.1037/pspp0000186.

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*, 478–491. https://doi.org/10.1038/s41562-019-0555-0.

Denissen, J. J. A., Geenen, R., Selfhout, M., & van Aken, M. A. G. (2008). Single-item big five ratings in a social network design. *European Journal of Personality*, *22*, 37–54. https://doi.org/10.1002/per.662.

Denissen, J. J. A., & Penke, L. (2008). Motivational individual reaction norms underlying the Five-Factor model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality*, *42*, 1285–1302. https://doi.org/10.1016/j.jrp.2008.04.002.

DeYoung, C. G. (2015). Cybernetic Big Five theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*, 880–896. https://doi.org/10.1037/0022-3514.93.5.880.

Drewelies, J., Schade, H., Hülür, G., Hoppmann, C. A., Ram, N., & Gerstorf, D. (2018). The more we are in control, the merrier? Partner perceived control and negative affect in the daily lives of older couples. *The Journals of Gerontology: Series B.*. https://doi.org/10.1093/geronb/gby009.

Driver, C. C., Oud, J. H. L., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software*, *77*. https://doi.org/10.18637/jss.v077.i05.

Dumenci, L. (2000). Multitrait-multimethod analysis. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 583–611). San Diego, CA, US: Academic Press.

Eaton, N. R., South, S. C., & Krueger, R. F. (2009). The cognitive–affective processing system (CAPS) approach to personality and the concept of personality disorder: Integrating clinical and social-cognitive research. *Journal of Research in Personality*, *43*, 208–217. https://doi.org/10.1016/j.jrp.2009.01.016.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230–253. https://doi.org/10.1037/a0013219.

Finnigan, K. M., & Vazire, S. (2018). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology*, *115*, 321–337. https://doi.org/10.1037/pspp0000136.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences*, *115*, E6106–E6115. https://doi.org/10.1073/pnas.1711978115.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin*, *51*, 327–358. https://doi.org/10.1037/h0061470.

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*, 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011.

Fleeson, W. (2007). Situation-based contingencies underlying trait-content manifestation in behavior. *Journal of Personality*, *75*, 825–862. https://doi.org/10.1111/j.1467-6494.2007.00458.x.

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009.

Fleeson, W., & Law, M. K. (2015). Trait enactments as density distributions: The role of actors, situations, and observers in explaining stability and variability. *Journal of Personality and Social Psychology*, *109*, 1090–1104. https://doi.org/10.1037/a0039517.

Fleeson, W., & Noftle, E. E. (2008). Where does personality have its influence? A supermatrix of consistency concepts. *Journal of Personality*, *76*, 1355–1386. https://doi.org/10.1111/j.1467-6494.2008.00525.x.

Forgeard, M., Herzhoff, K., Jayawickreme, E., Tsukayama, E., Beard, C., & Björgvinsson, T. (2018). Changes in daily manifestations of openness to experience during intensive cognitive-behavioral treatment. *Journal of Personality.* https://doi.org/10.1111/jopy.12438.

Funder, D. C. (2001). Personality. *Annual Review of Psychology*, *52*, 197–221. https://doi.org/10.1146/annurev.psych.52.1.197.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, *40*, 21–34. https://doi.org/10.1016/j.jrp.2005.08.003.

Furr, R. M. (2009). Personality psychology as a truly behavioural science. *European Journal of Personality*, *23*, 369–401. https://doi.org/10.1002/per.724.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91. https://doi.org/10.1037/a0032138.

Giacomin, M., & Jordan, C. H. (2016). The wax and wane of narcissism: Grandiose narcissism as a process or state. *Journal of Personality*, *84*, 154–164. https://doi.org/10.1111/jopy.12148.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*, 26–42.

Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z., & Campbell, A. T. (2017). Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior*, *67*, 129–138. https://doi.org/10.1016/j.chb.2016.10.027.

Harlow, R. E., & Cantor, N. (1995). To whom do people turn when things go poorly? Task orientation and functional social contacts. *Journal of Personality and Social Psychology*, *69*, 329–340. https://doi.org/10.1037/0022-3514.69.2.329.

Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? *Journal of Individual Differences*, *35*, 245–249. https://doi.org/10.1027/1614-0001/a000146.

Hektner, J., Schmidt, J., & Csikszentmihalyi, M. (2007). *Experience sampling method*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. https://doi.org/10.4135/9781412984201

Himmelstein, P. H., Woods, W. C., & Wright, A. G. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychological Assessment*, *31*, 952–960. https://doi.org/10.1037/pas0000718.

Hofmann, W., Baumeister, R. F., Förster, G., & Vohs, K. D. (2012). Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, *102*, 1318–1335. https://doi.org/10.1037/a0026545.

Hofmans, J., De Clercq, B., Kuppens, P., Verbeke, L., & Widiger, T. A. (2019). Testing the structure and process of personality using ambulatory assessment data: An overview of within-person and person-specific techniques. *Psychological Assessment*, *31*, 432–443. https://doi.org/10.1037/pas0000562.

Horstmann, K. T. (2020). Experience sampling and daily diary studies: Basic concepts, designs, and challenges. In J. F. Rauthmann (Ed.), *The handbook of personality dynamics and processes*. Academic Press.

Horstmann, K. T., Knaut, M., & Ziegler, M. (2019). Criterion validity. In *Encyclopedia of personality and individual differences* (pp. 1–3). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1293-1

Horstmann, K. T., & Rauthmann, J. F. (n.d.). *How many states make a trait? A comprehensive meta-analysis of experience sampling studies*.

Horstmann, K. T., Rauthmann, J. F., & Sherman, R. A. (2018). Measurement of situational influences. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *The SAGE handbook of personality and individual differences* (pp. 465–484). SAGE Publications.

Horstmann, K. T., Rauthmann, J. F., Sherman, R. A., & Ziegler, M. (in revision). *Distinguishing simple and residual consistency in functionally equivalent and non-equivalent situations: Evidence from experimental and observational longitudinal data*.

Horstmann, K. T., Rauthmann, J. F., Sherman, R. A., & Ziegler, M. (accepted). Unveiling an exclusive link: Predicting behavior with personality, situation perception, and affect in a pre-registered experience sampling study. *Journal of Personality and Social Psychology.*

Horstmann, K. T., Ziegler, J., & Ziegler, M. (2018). Assessment of situational perceptions. (J. F. Rauthmann, R. Sherman, & D. C. Funder, Eds.), *The Oxford handbook of psychological situations* (Vol. *1*). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.21

Hudson, N. W., Briley, D. A., Chopik, W. J., & Derringer, J. (2018). You have to follow through: Attaining behavioral change goals predicts volitional personality change. *Journal of Personality and Social Psychology.*. https://doi.org/10.1037/pspp0000221.

Hudson, N. W., & Fraley, R. C. (2015). Volitional personality trait change: Can people choose to change their personality traits? *Journal of Personality and Social Psychology*, *109*, 490–507. https://doi.org/10.1037/pspp0000021.

Hülür, G., Hoppmann, C. A., Rauers, A., Schade, H., Ram, N., & Gerstorf, D. (2016). Empathic accuracy for happiness in the daily lives of older couples: Fluid cognitive performance predicts pattern accuracy among men. *Psychology and Aging*, *31*, 545–552. https://doi.org/10.1037/pag0000109.

Ilies, R., Schwind, K. M., Wagner, D. T., Johnson, M. D., DeRue, D. S., & Ilgen, D. R. (2007). When can employees have a family life? The effects of daily workload and affect on work–family conflict and social behaviors at home. *Journal of Applied Psychology*, *92*, 1368–1379. https://doi.org/10.1037/0021-9010.92.5.1368.

IPIP. (2015). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*. Retrieved from http://ipip.ori.org/

Jackson, J. J., Wood, D., Bogg, T., Walton, K. E., Harms, P. D., & Roberts, B. W. (2010). What do conscientious people do? Development and validation of the behavioral indicators of conscientiousness (BIC). *Journal of Research in Personality*, *44*, 501–511. https://doi.org/10.1016/j.jrp.2010.06.005.

Jayawickreme, E., Zachry, C. E., & Fleeson, W. (2019). Whole trait theory: An integrative approach to examining personality structure and process. *Personality and Individual Differences*, *136*, 2–11. https://doi.org/10.1016/j.paid.2018.06.045.

Jones, A. B., Brown, N. A., Serfass, D. G., & Sherman, R. A. (2017). Personality and density distributions of behavior, emotions, and situations. *Journal of Research in Personality*, *69*, 225–236. https://doi.org/10.1016/j.jrp.2016.10.006.

Kemper, C. J., Trapp, S., Kathmann, N., Samuel, D. B., & Ziegler, M. (2018). Short versus long scales in clinical assessment: Exploring the trade-off between resources saved and psychometric quality lost using two measures of obsessive–compulsive symptoms. *Assessment*, 107319111881005. https://doi.org/10.1177/1073191118810057.

Killingsworth, M. A., & Gilbert, D. T. (2010). A wandering mind is an unhappy mind. *Science*, *330*, 932–932. https://doi.org/10.1126/science.1192439.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In 313 (Ed.), *Handbook of survey research* (p. 263). Emerald Group publishing limited.

Larsen, R. J., & Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *Journal of Personality and Social Psychology*, *61*, 132–140. https://doi.org/10.1037//0022-3514.61.1.132.

Lee, J., & Paek, I. (2014). In search of the optimal number of response categories in a rating scale. *Journal of Psychoeducational Assessment*, *32*, 663–673. https://doi.org/10.1177/0734282914522200.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

Lüdtke, O., Roberts, B. W., Trautwein, U., & Nagy, G. (2011). A random walk down university avenue: Life paths, life events, and personality trait change at the transition to university life. *Journal of Personality and Social Psychology*, *101*, 620–637. https://doi.org/10.1037/a0023743.

Maas, C. J. M., Lensvelt-Mulders, G. J. L. M., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology*, *5*, 72–77. https://doi.org/10.1027/1614-2241.5.3.72.

MacCann, C., Duckworth, A. L., & Roberts, R. D. (2009). Empirical identification of the major facets of conscientiousness. *Learning and Individual Differences*, *19*, 451–458. https://doi.org/10.1016/j.lindif.2009.03.007.

Magidson, J. F., Roberts, B. W., Collado-Rodriguez, A., & Lejuez, C. W. (2014). Theory-driven intervention for changing personality: Expectancy value theory, behavioral activation, and conscientiousness. *Developmental Psychology*, *50*, 1442–1450. https://doi.org/10.1037/a0030583.

Mathieu, J. E., Aguinis, H., Culpepper, S. a., & Chen, G. (2012). "Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling": Correction to Mathieu, Aguinis, Culpepper, and Chen (2012). *Journal of Applied Psychology*, *97*, 981–981. https://doi.org/10.1037/a0029358.

McCabe, K. O., & Fleeson, W. (2016). Are traits useful? Explaining trait manifestations as tools in the pursuit of goals. *Journal of Personality and Social Psychology*, *110*, 287–301. https://doi.org/10.1037/a0039490.

McDonald, R. P. (1999). *Test homogeneity, reliability, and generalizability.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*, 517–523. https://doi.org/10.3758/BF03195410.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

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

Psychological Review, *102*, 246–268. https://doi.org/10.1037/0033-295x.102.2.246.

Moskowitz, D. S. (1994). Cross-situational generality and the interpersonal circumplex. *Journal of Personality and Social Psychology*, *66*, 921–933. https://doi.org/10.1037/0022-3514.66.5.921.

Moskowitz, D. S., & Russell, J. J. (2009). Measuring behaviour. *European Journal of Personality*, *23*, 417–419.

Moskowitz, D. S., Russell, J. J., Sadikaj, G., & Sutton, R. (2009). Measuring people intensively. *Canadian Psychology/Psychologie Canadienne*, *50*, 131–140. https://doi.org/10.1037/a0016625.

Moskowitz, D. S., & Zuroff, D. C. (2005). Assessing interpersonal perceptions using the interpersonal grid. *Psychological Assessment*, *17*, 218–230. https://doi.org/10.1037/1040-3590.17.2.218.

Murray, S. L., Gomillion, S., Holmes, J. G., Harris, B., & Lamarche, V. (2013). The dynamics of relationship promotion: Controlling the automatic inclination to trust. *Journal of Personality and Social Psychology*, *104*, 305–334. https://doi.org/10.1037/a0030513.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398. https://doi.org/10.1177/0049124194022003006.

Newman, D. B., Sachs, M. E., Stone, A. A., & Schwarz, N. (2019). Nostalgia and well-being in daily life: An ecological validity perspective. *Journal of Personality and Social Psychology.*. https://doi.org/10.1037/pspp0000236.

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149–155. https://doi.org/10.1016/j.jrp.2016.06.020.

Nussbeck, F. W., Eid, M., Geiser, C., Courvoisier, D. S., & Lischetzke, T. (2009). A CTC(M-1) model for different types of raters. *Methodology*, *5*, 88–98. https://doi.org/10.1027/1614-2241.5.3.88.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods & Research*, *18*, 473–504. https://doi.org/10.1177/0049124190018004004.

Ostojic-Aitkens, D., Brooker, B., & Miller, C. J. (2019). Using ecological momentary assessments to evaluate extant measures of mind wandering. *Psychological Assessment*, *31*, 817–827. https://doi.org/10.1037/pas0000701.

Pihet, S., De Ridder, J., & Suter, M. (2017). Ecological momentary assessment (EMA) goes to jail. *European Journal of Psychological Assessment*, *33*, 87–96. https://doi.org/10.1027/1015-5759/a000275.

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within-person variance in applied psychology constructs? An empirical examination. *Journal of Applied Psychology*, *104*, 727–754. https://doi.org/10.1037/apl0000374.

Quintus, M., Egloff, B., & Wrzus, C. under review. *Momentary processes predict long-term development in explicit and implicit representations of Big Five traits: An empirical test of the TESSERA framework.*.

Rauthmann, J. F., Horstmann, K. T., & Sherman, R. A. (2019). Do self-reported traits and aggregated states capture the same thing? A nomological perspective on trait-state homomorphy. *Social Psychological and Personality Science*, *10*, 596–611. https://doi.org/10.1177/1948550618774772.

Raykov, T., & du Toit, S. H. C. (2005). Estimation of reliability for multiple-component measuring instruments in hierarchical designs. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*, 536–550. https://doi.org/10.1207/s15328007sem1204_2.

Raynor, D. A., & Levine, H. (2009). Associations between the Five-Factor model of personality and health behaviors among college students. *Journal of American College Health*, *58*, 73–82. https://doi.org/10.3200/JACH.58.1.73-82.

Read, S. J., Smith, B. J., Droutman, V., & Miller, L. C. (2017). Virtual personalities: Using computational modeling to understand within-person variability. *Journal of Research in Personality*, *69*, 237–249. https://doi.org/10.1016/j.jrp.2016.10.005.

Revelle, W., & Condon, D. M. (2015). A model for personality at three levels. *Journal of Research in Personality*, *56*, 70–81. https://doi.org/10.1016/j.jrp.2014.12.006.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, *58*, 103–139. https://doi.org/10.1111/j.1744-6570.2005.00301.x.

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness. In *Handbook of individual differences in social behavior* (pp. 369–381). New York, NY, US: The Guilford Press.

Roberts, B. W., Lejuez, C., Krueger, R. F., Richards, J. M., & Hill, P. L. (2014). What is conscientiousness and how can it be assessed? *Developmental Psychology*, *50*, 1315–1330. https://doi.org/10.1037/a0031109.

Roberts, B. W., & Mroczek, D. (2008). Personality trait change in adulthood. *Current Directions in Psychological Science*, *17*, 31–35. https://doi.org/10.1111/j.1467-8721.2008.00543.x.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25. https://doi.org/10.1037/0033-2909.132.1.1.

Robinson, M. D., & Clore, G. L. (2002a). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, *128*, 934–960. https://doi.org/10.1037/0033-2909.128.6.934.

Robinson, M. D., & Clore, G. L. (2002b). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, *83*, 198–215. https://doi.org/10.1037/0022-3514.83.1.198.

Roemer, L., Horstmann, K. T., & Ziegler, M. n.d. (submitted). *Sometimes hot, sometimes not: The relations between situational vocational interests and situation perception*.

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, *76*, 613–627. https://doi.org/10.1037//0022-3514.76.4.613.

Schmitt, M. (2009a). Linking personality and behaviour based on theory. *European Journal of Personality*, *23*, 428–429.

Schmitt, M. (2009b). Person × situation-interactions as moderators. *Journal of Research in Personality*, *43*, 267. https://doi.org/10.1016/j.jrp.2008.12.032.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353. https://doi.org/10.1037/1040-3590.8.4.350.

Schönbrodt, F. D., Zygar, C., Nestler, S., Pusch, S., & Hagemeyer, B. (n.d.). *Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples*. Retrieved from https://doi.org/10.31234/osf.io/6mq7t

Sengewald, E., & Vetterlein, A. (2015). Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation. *Diagnostica*, *61*, 116–123. https://doi.org/10.1026/0012-1924/a000140.

Sherman, R. A., Rauthmann, J. F., Brown, N. A., Serfass, D. G., & Jones, A. B. (2015). The independent effects of personality and situations on real-time expressions of behavior and emotion. *Journal of Personality and Social Psychology*, *109*, 872–888. https://doi.org/10.1037/pspp0000036.

Silvia, P. J., Kwapil, T. R., Walsh, M. A., & Myin-Germeys, I. (2014). Planned missing-data designs in experience-sampling research: Monte Carlo simulations of efficient designs for assessing within-person constructs. *Behavior Research Methods*, *46*, 41–54. https://doi.org/10.3758/s13428-013-0353-y.

Sim, S.-M., & Rasiah, R. I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple

choice questions of a para-clinical multidisciplinary paper. *Annals-Academy of Medicine Singapore*, *35*, 67–71.
Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*, 557–566. https://doi.org/10.1037/pas0000648.
Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The life outcomes of personality replication project. *Psychological Science*, *30*, 711–727. https://doi.org/10.1177/0956797619831612.
Soto, C. J., & John, O. P. (2017). The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, *113*, 117–143. https://doi.org/10.1037/pspp0000096.
Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904–1920*, *3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x.
Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., … Bühner, M. (2017). Personality traits predict smartphone usage. *European Journal of Personality*, *31*, 701–722. https://doi.org/10.1002/per.2113.
Sun, J., & Vazire, S. (2019). Do people know what They're like in the moment? *Psychological Science*, *30*, 405–414. https://doi.org/10.1177/0956797618818476.
Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500–517. https://doi.org/10.1037/0021-9010.88.3.500.
Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality*, *34*, 397–423. https://doi.org/10.1006/jrpe.2000.2292.
Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment*, *23*, 995–1009. https://doi.org/10.1037/a0024165.
Timmermans, T., Van Mechelen, I., & Kuppens, P. (2010). The relationship between individual differences in intraindividual variability in core affect and interpersonal behaviour. *European Journal of Personality*, *24*, 623–638. https://doi.org/10.1002/per.756.
Tomko, R. L., Solhan, M. B., Carpenter, R. W., Brown, W. C., Jahng, S., Wood, P. K., & Trull, T. J. (2014). Measuring impulsivity in daily life: The momentary impulsivity scale. *Psychological Assessment*, *26*, 339–349. https://doi.org/10.1037/a0035083.
van Roekel, E., Heininga, V. E., Vrijen, C., Snippe, E., & Oldehinkel, A. J. (2019). Reciprocal associations between positive emotions and motivation in daily life: Network analyses in anhedonic individuals and healthy controls. *Emotion*, *19*, 292–300. https://doi.org/10.1037/emo0000424.
Vazire, S., & Mehl, M. R. (2008). Knowing me, knowing you: The accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *Journal of Personality and Social Psychology*, *95*, 1202–1216. https://doi.org/10.1037/a0013314.
Voelkle, M. C., Oud, J. H. L., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: Relating authoritarianism and anomia. *Psychological Methods*, *17*, 176–192. https://doi.org/10.1037/a0027543.
Vogelsmeier, L. V. D. E., Vermunt, J. K., van Roekel, E., & De Roover, K. (2019). Latent Markov factor analysis for exploring measurement model changes in time-intensive longitudinal studies. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 557–575. https://doi.org/10.1080/10705511.2018.1554445.
Watson, D. C. (2001). Procrastination and the Five-Factor model: A facet level analysis. *Personality and Individual Differences*, *30* (1), 149–158. https://doi.org/10.1016/S0191-8869(00)00019-2
Weinstein, N., & Ryan, R. M. (2010). When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology*, *98*, 222–244. https://doi.org/10.1037/a0016984.
Westermann, R., Spies, K., Stahl, G., & Hesse, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology*, *26*(4), 557–580. https://doi.org/10.1002/(SICI)1099-0992(199607)26:4557::AID-EJSP7693.0.CO;2-4
Wilmot, M. P., & Ones, D. S. (2019). *A century of research on conscientiousness at work*. Proceedings of the National Academy of Sciences, 201908430. https://doi.org/10.1073/pnas.1908430116
Wilt, J., Noftle, E. E., Fleeson, W., & Spain, J. S. (2012). The dynamic role of personality states in mediating the relationship between extraversion and positive affect. *Journal of Personality*, *80*, 1205–1236. https://doi.org/10.1111/j.1467-6494.2011.00756.x.
Wilt, J. A., & Revelle, W. (2015). Affect, behaviour, cognition and desire in the Big Five: An analysis of item content and structure. *European Journal of Personality*, *29*, 478–497. https://doi.org/10.1002/per.2002.
Wright, A. G. C., & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. *Psychological Assessment.*. https://doi.org/10.1037/pas0000685.
Wrzus, C., & Mehl, M. R. (2015). Lab and/or field? Measuring personality processes and their social consequences. *European Journal of Personality*, *29*, 250–271. https://doi.org/10.1002/per.1986.
Wrzus, C., & Roberts, B. W. (2017). Processes of personality development in adulthood: The TESSERA framework. *Personality and Social Psychology Review*, *21*, 253–277. https://doi.org/10.1177/1088868316652279.
Ziegler, M. (2014). Stop and state your intentions! *European Journal of Psychological Assessment*, *30*, 239–242. https://doi.org/10.1027/1015-5759/a000228.
Ziegler, M., & Bäckström, M. (2016). 50 facets of a trait—50 ways to mess up? *European Journal of Psychological Assessment*, *32*, 105–110. https://doi.org/10.1027/1015-5759/a000372.
Ziegler, M., Horstmann, K. T., & Ziegler, J. (2019). Personality in situations: Going beyond the OCEAN and introducing the Situation Five. *Psychological Assessment*, *31*, 567–580. https://doi.org/10.1037/pas0000654.
Ziegler, M., Kemper, C. J., & Kruyen, P. (2014). Short scales—Five misunderstandings and ways to overcome them. *Journal of Individual Differences*, *35*, 185–189. https://doi.org/10.1027/1614-0001/a000148.
Ziegler, M., Kemper, C. J., & Lenzner, T. (2015). The issue of fuzzy concepts in test construction and possible remedies. *European Journal of Psychological Assessment*, *31*, 1–4. https://doi.org/10.1027/1015-5759/a000255.
Ziegler, M., Schroeter, T., Lüdtke, O., & Roemer, L. (2018). The enriching interplay between openness and interest: A theoretical elaboration of the OFCI model and a first empirical test. *Journal of Intelligence*, *6*, 35. https://doi.org/10.3390/jintelligence6030035.
Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., … Wright, A. G. C. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. *Psychological Assessment*, *31*, 516–531. https://doi.org/10.1037/pas0000625.