Empirical Paper

EUROPEAN JOURNAL OF Personality

European Journal of Personality
2021, Vol. 35(6) 833–860

DOI: 10.1177/08902070211014029
journals.sagepub.com/home/ejop
SAGE

Distinguishing simple and residual consistencies in functionally equivalent and non-equivalent situations: Evidence from experimental and observational longitudinal data

Kai T Horstmann[1], John F Rauthmann[2], Ryne A Sherman[3] and Matthias Ziegler[1]

Abstract

The current work examines consistencies of personality state scores across functionally equivalent and non-equivalent situations. We argue that *simple consistency*, defined as the correlation between state scores without taking people's straits into account, needs to be distinguished from *residual consistency* that does account for traits. The existence of residual consistency reflects systematic interindividual differences in how people respond to situations, above and beyond what is expected from their traits. We examine the level and individual differences in all of these forms of consistency. In four micro-longitudinal studies (total $N = 671$), participants first provided trait self-ratings and then state ratings, either in response to two situation vignettes presented at separate testing occasions (Studies 1 and 2) or during experience sampling in daily life (Studies 3 and 4). In all studies, simple consistency was substantial, and the level of residual consistency varied with the level of functional equivalence of the situations. Further, individual differences in both simple and residual consistencies were only weakly correlated, suggesting no underlying general factor but only trait-specific consistencies. We conclude that there are consistent individual differences in how people respond to equivalent situations, even when their personality trait scores have been taken into account.

Keywords

Traits, states, consistency, situations, experience sampling

Received 12 October 2020; accepted 3 December 2020

You cannot step twice into the same river.

– Attributed to Heraclitus

This quote describes the notion that no person can ever be in the same situation more than once as both the situation as well as the person will have changed. But even if it may not be possible to be in the exact same situation twice, it is still possible to be in situations that are at least *functionally equivalent*, that is, they may be seen and interpreted in similar ways and thus might elicit the same or highly similar thoughts, feelings, desires, or behaviors. Consider, for example, a conference dinner. This situation can be seen as rather formal, partly as a requirement, and, for some, it certainly has the affordance to dance. A functionally equivalent situation would be a conference dinner at another conference, which is perceived in a similar way (see below for a more detailed explanation). Imagine now a person who is only observed during conference dinners (e.g., by colleagues from other universities), someone who is consistently the first on the dance floor and the last at the bar. How much, one may wonder, of their consistent behavior is

[1]Institute of Psychology, Humboldt-Universität zu Berlin, Germany
[2]Institute of Psychology, University Bielefeld, Germany
[3]Hogan Assessments, USA

Corresponding author:
Kai T Horstmann, Humboldt-Universität zu Berlin, Rudower Chaussee 18, Berlin 12489, Germany.
Email: kaitobiashorstmann@gmail.com

due to their personality characteristics, and how much can be attributed to the situation? Importantly, observing a person's behavior only under specific, recurring situations makes it impossible to distinguish behavior that is due to the person, compared to behavior that is due to the consistent occurrence of the functionally equivalent situations. It could even be possible that the relation between functional equivalence of a situation and behavior (e.g., dancing wildly) might, at least in part, be independent of the trait (e.g., general tendency to dance, or the trait extraversion) under focus. If this was the case, and such consistency of behaviors *independent* from a corresponding trait existed, it would alert us to systematic portions of variance in responses to situations that cannot be attributed to trait levels alone. However, previous studies have not quantified such a form of consistency. Nonetheless, it is important to demonstrate such consistency because people may regularly inhabit functionally equivalent—that is, similar—situations in their daily lives that repeat themselves and provide for routine (e.g., playing with one's children, commuting, working with colleagues, meeting friends for a coffee), perhaps sometimes even happening without strong correspondence to their personalities (Ickes et al., 1997). This study examines state consistencies across different yet functionally equivalent, situations (Studies 1, 3, and 4) as well as different and functionally non-equivalent situations (Studies 2, 3, and 4). To this end, we introduce the concept of "residual consistency" alongside the traditional "simple consistency" which has already been examined in the literature.

Background

Traits and states

A personality *trait* is usually conceptualized as a stable characteristic of a person (Funder, 2001), and trait levels may differ between people. For example, some people can be described as *generally* more extraverted, whereas others can be better described as *generally* more introverted. The trait of a person was recently defined as their consistent patterns of behavior, thoughts, and feelings (Fleeson, 2001; Fleeson & Jayawickreme, 2015; Jayawickreme et al., 2019): If a person consistently expresses high levels of extraversion (e.g., dancing), then this person is considered an extravert. Such general descriptions (across many time-points) notwithstanding, people regularly exhibit a range of personality *states*. The latter have been defined as "having the same affective, behavioral, and cognitive content as a corresponding trait (...), but as applying for a shorter duration" (Fleeson & Jayawickreme, 2015, p. 84). Notably, states can form distributions within persons, and the parameters of these distributions (especially the central tendencies) are often substantially associated with self-reports of broad traits (Fleeson & Jayawickreme, 2015; Horstmann & Rauthmann, in preparation; Rauthmann et al., 2019). Such parameters have additionally been shown to be stable, meaningful, and consequential (Fleeson, 2001; Jones et al., 2017). A conceptualization of personality – with stable components (traits) and variable components (states) – reconciles structure- and process-oriented approaches (Baumert et al., 2017; Fleeson, 2001).

Consistency of states

Whereas personality traits are supposed to be stable (i.e., stability is seen as a pre-requisite for establishing that something can be considered a personality trait; Funder, 2001), consistency refers to the manifestation of a trait (Fleeson & Noftle, 2008) under specific circumstances in a specific situation. However, when examining consistency of personality states, there is a fundamental problem: If states are seen as a product of both the situation and the person (Horstmann & Ziegler, 2020), then both the person and the situation need to be considered to examine state consistency. In other words, a state can be consistent (a) because of a consistent influence of the personality trait on the state, (b) because of a consistent effect of the situation on the personality state, or (c) both. It has been shown that states are associated with situations (Horstmann et al., in press; Sherman et al., 2015), which leads to the implicit assumption (and testable hypothesis) that states enacted in similar situations should be more consistent compared to states enacted in dissimilar situations (Sherman et al., 2010), even after controlling for personality trait levels. This does not preclude the influence of a trait. Rather than that, the gist of prior findings suggests that both, trait and situation influence a state. However, the latter influence has not been tested for consistency across situations.

In our introductory example, the behavioral state "dance" is assumed to be to some degree composed of at least a trait (e.g., general tendency to dance or extraversion) and a state (e.g., extreme dancing, more so than is usual per the typical dancing tendency of the person). Consequently, we can define consistency in two ways. If raw state scores (e.g., the average of several items assessing state dancing) are correlated with each other across situations, then *simple consistency* is estimated. If state residual scores are correlated (i.e., state scores that have been controlled for a trait score), then *residual consistency* is estimated (see below for a more detailed description).

The literature on consistency has so far only quantified simple forms (Fleeson & Noftle, 2008), but not residual forms. Here, we propose the idea of residual consistency which allows testing to what extent there are systematic differences in personality states across situations, independent of the trait level. If a person

showed a high degree of residual consistency, that is, consistency in states across situations that cannot be accounted for by the respective personality trait, it means that other quantities than this personality trait must be considered to explain such a consistency. For example, a person could act consistently due to similar social roles in different situations, similar expectations, or, more general, stable, trait-like tendencies to perceive situations in a certain way. Note that the difference between simple and residual consistencies depends on the relation of the trait and the state: If no trait variance is present in the states, then estimates of simple and residual consistencies must be the same.

Nonetheless, the situations in which states occur have often been neglected, and it remains to be empirically shown to what extent *state residuals* that have been ridded from trait variance can be consistent across situations. In other words, if the state residual in one situation differs from the state residual in another yet functionally equivalent situation, then these state residuals would not be consistent. Here, we argue that state residuals should at least in parts be consistent *if* situations are indeed perceived to be functionally equivalent by the person in the situation and furthermore if the situation actually exerts an influence on the manifestation of a trait.

A key component thus far has been situational similarity. The only way to show that residual consistency is related to situations is by showing that estimates of both consistency and residual consistency are higher in situations that are more similar to one another, or, as noted earlier, the more *functionally equivalent* they are.

Functional equivalence of situations. What does functional equivalence of situations mean, and when can two situations be considered functionally equivalent? In the most extreme case, two situations would be functionally equivalent if they were identical. However, as already noted in the introductory quote, it is impossible to enter the exact same situation twice – at the very least, time has passed and the person's knowledge about the situation has changed.

Rauthmann, Sherman, and Funder (2015) suggested that situations can be described at three levels: cues (e.g., physical objects, time, location, etc.), psychological characteristics (i.e., the psychological meaning of a situation), and classes (i.e., situations that are similar either in their situational cues or their characteristics). The authors further suggested that behaviorally manifest personality states are not influenced by cues of a situation per se, but that it is the psychological interpretation that influences behavior. In other words, it is not the dancefloor per se that makes a person dance, but the interpretation of the dancefloor as an invitation to dance.

The level of functional equivalence can therefore be defined via the similarity of situations, either at the level of situation cues or at the level of situation characteristics. With respect to situation cues, one would argue that two Situations A and A′ are higher in their functional equivalence the more similar their situation cues are. With respect to situation characteristics, two situations are high in their functional equivalence if they are perceived and interpreted similarly. Note that two seemingly different situations (as simply judged by their cues) may still be perceived similarly; one may thus attest them a high functional equivalence. For example, a party at a conference dinner may, after all, not be perceived too differently by some participants compared to a party at their favorite hometown bar. Even if the situations are objectively dissimilar, relevance for behavior can be highly similar and, hence, functionally equivalent. At the same time, objectively similar or even identical situations could be perceived rather differently, for example, if the mood of the perceiver has changed substantially from one situation to another (Horstmann et al., 2020; Horstmann & Ziegler, 2019). Functional equivalence therefore resides on a continuum, and situations may be more or less functionally equivalent. We employ an operationalization of functional equivalence via cues in Studies 1 and 2, and via situation characteristics in Studies 3 and 4.

Consistency in functional (non-)equivalent situations. A high degree of residual consistency in functionally equivalent situations suggests that state residuals are systematic and potentially important towards understanding persons and situations better. On the other hand, if situations are functionally non-equivalent, state residuals should show a lower degree of consistency, especially vis-à-vis those from functionally equivalent situations.

Figure 1 illustrates our explanations above, depicting (a) a single trait of the domain "P" (Trait *P*), (b) two regular states of this trait P (hence, State *P*) measured at two time-points in functionally equivalent situations (here: Situations A and A′), and (c) two state residuals or profiles of state residuals (State Residual *P*). First, traits predict states because they manifest or can be expressed in them. This *trait expression* corresponds to a correlation between a trait variable and a state variable across individuals. Trait expression can be estimated for both time-points. Notably, if the situations are indeed perceived to be functionally equivalent, trait expressions should be highly similar across the two measured time-points—as long as the state is influenced by the situation. On the other hand, if situations are functionally non-equivalent, trait expressions should be less similar. Second, *simple consistency* would refer to simply correlating two state scores at different time-points (e.g., *r* across individuals for extraversion). Third, state residual scores—as state scores that are controlled for trait variance—can be correlated across

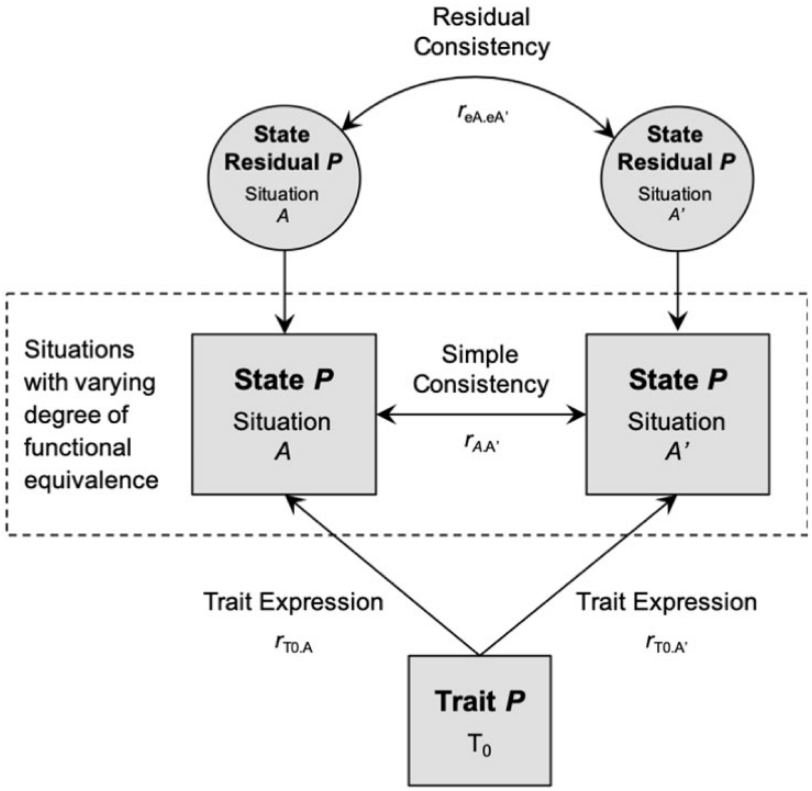

Figure 1. Visualization of residual consistency and simple consistency. *Note*. One domain *P* for traits and states is given as an illustration. Situations A and A′ represent two different situations that may be either functionally equivalent or functionally non-equivalent. A double-headed arrow reflects a correlation.

time-points to obtain estimates of *residual consistency*.

Individual differences in consistency. For the different forms of consistency outlined, we can ask how strong individual differences in them are. We then would want to know to what extent being more consistent in one variable (e.g., extraversion) also entails being more consistent in another one (e.g., neuroticism). Indeed, if people were simultaneously consistent in several variables, then this would point towards a general factor of consistency. Previous research has, however, provided mixed evidence for such general consistency (Bem & Allen, 1974; Chaplin & Goldberg, 1984; Schmitt, 1990a, 1990b), and it is thus an open question if general consistency can be found across different traits. The current study will thus seek to quantify individual differences and their interrelations for simple and residual consistencies in functionally equivalent *and* non-equivalent situations.

The current studies

To examine simple and residual consistencies in functionally equivalent and non-equivalent situations, we present four different studies. All studies were conducted online. Studies 1 and 2 examined simple and residual consistencies, respectively, of self-reported personality states in an experimental longitudinal design. In Studies 3 and 4, we reanalyzed previously published experience sampling data to test our hypotheses in a naturalistic field/daily life environment. The concept and computations of relevant statistics are explained in Study 1, and then later applied again in Studies 2, 3, and 4. We initially pre-registered the general research question, methods, and parts of the data-analytical strategy of Study 1 on the Open Science Framework (OSF) at osf.io/8u7ka/. Table 1 gives an overview of the research questions, studies, and elements that were pre-registered.

Study 1

The primary purpose of this study is the examination of different types of consistency (Figure 1), specifically illuminating the newly proposed idea of residual consistency. Using hypothetical but functionally equivalent situational vignettes, we examine to what extent (a) participants' state scores as well as (b) their state residual scores are correlated between two functionally equivalent Situations A and A′. The correlation of state scores provides an estimate of simple consistency, whereas the correlation of regression residuals (RRs) provides an estimate of residual consistency.

First, we report mean-level change in state scores between Situations A to A′. As pre-registered, we

Table 1. Overview of study questions, tables, and figures.

Point	Pre-registered?[a]	Table		Figure	
		S12	S34	S12	S34
Study modalities					
General research question	Yes			1	
Methods					
Sample size	Yes[b]				
Procedures	Yes			2	
Materials	Yes			2	
Data pre-processing[c]	Yes				3
Analyses and results					
(No) Mean-level change in states between Situations A and A′	Yes[a]	2	6		
Trait expressions ($r_{T0.A}$ and $r_{T0.A'}$)	No	2	6		
Simple consistency ($r_{A.A'}$)	No	2	7		
Correlation of state scores across time and aspects	No	3	OSM B		
Quantification and intercorrelation of individual differences in $r_{A.A'}$	No	OSM A	8		
Residual consistency ($r_{eA.eA'}$)	No	2	7		
Correlation of residual state scores ($r_{eA.eA'}$) across time and aspects	No	3	OSM B		
Quantification and intercorrelation of individual differences in $r_{eA.eA'}$	No	OSM A	8		

S12: Study 1 and Study 2; S34: Study 3 and Study 4.

[a]If "Yes", this applies only to Study 1. However, Study 2 was designed after Study 1 and may therefore be considered confirmatory as well. Studies 3 and 4 were not pre-registered. Study 4 can be understood as a replication of Study 3.

[b]We pre-registered $N = 250$ for Study 1 (based on recommendations for cross-sectional studies from (Schönbrodt & Perugini, 2013), but were only able to gather full usable data from $N = 158$. Based on our experience with Study 1 and simulations, we planned to collect at least $N = 100$ for Study 2 and sampled there $N = 115$.

[c]Recoding, scale score computations, handling of missing data, etc.

OSM A and OSM B can be found here online at osf.io/pdu23/

expected no significant changes here because the two situations were supposed to be functionally equivalent. For example, if two situations are perceived as equal opportunities to dance, the average level of dancing should also be comparable across both situations.

Second, we report trait expressions which are correlations of a trait score assessed at an initial measurement occasion with each corresponding state score at later Situations A and A′, respectively; in our example, this would be the correlation between general extraversion and dancing at two similar conference dinners. If the situations are functionally equivalent (and if the perception of the situation has an influence on the personality state), then trait expression correlations ($r_{T0.A}$ and $r_{T0.A'}$) should be similar.

Third, we report estimates of simple consistency ($r_{A.A'}$) as the correlation between corresponding state scores at Situations A and A′. For example, state extraversion (e.g., dancing) at Situation A should be substantially correlated with state extraversion at Situation A′; or, in other words, those who dance a lot at one conference dinner should also dance a lot at another conference dinner. The evidence for such simple consistency will be strongest when convergent correlations (e.g., Extraversion at A and A′) exceed the discriminant correlations (e.g., Extraversion at A with Volatility at A and A′). Additionally, we compute correlations of interindividual differences in these simple consistencies, examining if consistency in one domain (e.g., Conscientiousness) is related to consistency in other domains (e.g., Extraversion). Uniformly high positive (and substantial) correlations would point towards a positive manifold and thus to a potential general factor of consistency underlying the covariation of individual simple consistency scores. Low correlations would suggest otherwise.

Fourth, we report findings for residual consistency using the same data-analytical steps as for simple consistency (i.e., quantification of residual consistency, intercorrelations of state residual scores, and intercorrelations of individual differences in residual consistencies). In our example, residual consistency would indicate the correlation of dancing at two conference dinners, while controlling for the person's general tendency to dance. This would then show that, regardless of the general tendency to dance, a person has a specific and consistent tendency to dance at conference dinners. Computing both simple and residual consistencies enables their comparison head to head, with the hypothesis that residual consistencies will be somewhat lower than simple ones yet still substantial. This would demonstrate that state consistency across situations cannot solely be attributed to the respective trait.

Methods

Participants. A total of $N = 158$ participants completed all stages of the first study (see Online Supplemental Materials (OSM) at osf.io/xfhdu/for details).

Of these, 84.18% of participants were female, 15.19% were male, and 0.63% chose not to indicate their gender. Their mean age was $M = 25.49$ ($SD = 7.46$). No other socio-demographic characteristics of the participants were assessed in Study 1. Participants received detailed feedback on their personality traits. If participants were psychology undergraduate students, they could also receive course credit. We initially planned to sample $N = 250$ participants (see below), based on simulations on the precision of correlations by Schönbrodt and Perugini (2013). However, given the high correlations that could be expected Study 1, Schönbrodt and Perugini's simulations as well as our own simulations suggest that $N = 158$ is sufficient to obtain reliable results.

Procedure. First, participants were informed about the study, gave their consent, and were registered with their e-mail address if they wanted to participate. Second at stage T_0, participants indicated their age and sex and responded to items measuring the trait-version of the Big Five Aspect Scales (BFAS; DeYoung et al., 2007). After three weeks, participants were invited via e-mail to participate again and randomly assigned to either of the hypothetical Situation Vignettes A or A′ (see *Materials*).[1] Participants had to read the vignette and rate their hypothetical state in this situation on a state-version of the BFAS. After another three weeks, participants were invited again and rated their hypothetical state in the *other* situation (either A′ or A). At both measurement occasions, participants were explicitly asked not to think about or recall their previous responses.

Materials. All measures were administered online using the platform formr.org (Arslan et al., 2020). Participants were not allowed to skip items (hence, there were no missing data), but could abort at any point. All materials necessary for a direct replication or reproduction of this study (situation vignettes, translated items, questionnaire, study set-up, data, and code) can be found online at osf.io/vwzs5/.

Situation vignettes. The hypothetical situation we chose was a party situation because most participants in our intended student sample could be expected to have attended one (high familiarity, close to real life). Further, a party situation is relevant to the expression of at least two important interpersonal domains (high content specificity)[2]: agreeableness and extraversion (Gurtman, 2009). The participants were instructed to imagine a situation where they had been taken along to a private party. They only know one person there and have to start a conversation with other people. Vignette A is 261 words long, and Vignette A′ 244 words. The vignettes can be found on the OSF page. They were constructed to be similar, though somewhat different in wording (to avoid strong memory effects after three weeks).

Traits and states. Traits at T_0 and states in Situations A and A′ were assessed with 100 items from the BFAS by DeYoung et al. (2007), using a 6-point rating scale (1 = *disagree*, 6 = *agree*). We used the exact same items for trait and state assessment and changed only the instruction, asking participants how well each item would describe them *in* the given situation. Each Big Five domain was measured with two aspects and 10 items each: Openness with Openness and Intellect, Conscientiousness with Industriousness and Orderliness, Extraversion with Assertiveness and Enthusiasm, Agreeableness with Compassion and Politeness, and Neuroticism with Volatility and Withdrawal. As no German version of the BFAS was available, we translated the items with multiple translators (see OSF page and materials at osf.io/3xvc5). We chose the BFAS since it allows a more detailed look at personality than pure Big Five measures, while being reasonably short at the same time. We chose to report all findings at the level of aspects, not domains. Aspects are narrower compared to domains, and it can therefore be expected that they reflect personality states better (Revelle & Condon, 2015).

Descriptive statistics and reliability estimates for the scores of all 10 aspects assessed at the three measurement occasions (trait at T_0, state at Situations A and A′) are displayed in Table 2. The internal consistency estimates (Cronbach's α and McDonald's Omega) for all scale scores were similar to the ones presented by DeYoung et al. (2007, Table 5).

Data analyses. All analyses as well as all anonymized data can be found at osf.io/nueq6. We used the software R (R Core Team, 2020) and the following packages: *dplyr* (Wickham & Francois, 2016), *stringr* (Wickham, 2016b), *purrr* (Wickham, 2016a), *psych* (Revelle, 2016), and *car* (Fox & Weisberg, 2011).

All coefficients in Figure 1 and questions in Table 1 were addressed (see *The Current Studies*). Most data-analytical issues involved are straightforward and will be referenced in *Results*. However, two special issues are addressed here: the derivation of (a) the residual state scores and (b) interindividual differences in consistencies.

Deriving residual state scores. To obtain residual state scores, we regressed a BFAS state score at Situations A or A′, respectively, on the corresponding BFAS trait score at T_0 (see Figure 1). We then extracted the RRs, which are state residuals. All analyses that refer to residual consistency were conducted using the so obtained state residual scores. Conceptually, the correlation between these state residuals can be understood as a partial correlation of two states, while partialing out the trait score.

Deriving interindividual differences in consistencies. Asendorpf (1990) provided an index (*I*) for the

Table 2. Descriptive statistics of trait and state scores in Study 1 and Study 2.

	Trait at T_0				State at Situation A/B				State at Situation A′/C				$\Delta M_{A\text{-}A'}$/ $\Delta M_{B\text{-}C}$			
Domains and aspects	*M*	*SD*	α	ω	*M*	*SD*	α	ω	*M*	*SD*	α	ω	*d*	*t*	*p*	95% CI
Study 1																
Openness																
Intellect	4.47	0.66	.79	.84	4.40	0.79	.86	.90	4.38	0.76	.85	.88	0.03	0.66	.511	[−0.04; 0.08]
Openness	4.50	0.72	.75	.81	4.34	0.78	.79	.84	4.34	0.77	.78	.84	0.01	0.17	.864	[−0.06; 0.07]
Conscientiousness																
Industriousness	4.00	0.81	.83	.87	3.92	0.72	.76	.83	3.90	0.76	.79	.83	0.02	0.49	.626	[−0.05; 0.09]
Orderliness	3.91	0.88	.84	.89	3.97	0.85	.84	.88	3.87	0.83	.82	.87	0.12	3.29	<.001	[0.04; 0.16]
Extraversion																
Assertiveness	4.20	0.68	.81	.85	4.03	0.76	.85	.89	3.98	0.76	.86	.89	0.07	1.49	.138	[−0.02; 0.12]
Enthusiasm	4.29	0.85	.88	.91	4.21	0.84	.87	.91	4.23	0.83	.88	.91	−0.03	−0.72	.472	[−0.08; 0.04]
Agreeableness																
Compassion	4.97	0.67	.87	.90	4.88	0.69	.88	.90	4.86	0.71	.88	.90	0.02	0.47	.641	[−0.04; 0.07]
Politeness	4.52	0.61	.70	.79	4.48	0.61	.72	.79	4.53	0.63	.75	.81	−0.09	−1.69	.092	[−0.12; 0.01]
Neuroticism																
Volatility	3.30	0.91	.88	.92	3.13	0.89	.89	.92	3.12	0.87	.88	.92	0.01	0.19	.848	[−0.07; 0.09]
Withdrawal	3.35	0.86	.85	.89	3.32	0.79	.83	.87	3.27	0.81	.83	.87	0.05	1.26	.209	[−0.02; 0.11]
Study 2																
Openness																
Intellect	4.49	0.76	.87	.90	4.17	0.86	.89	.92	4.42	0.78	.86	.91	−0.31	−5.21	<.001	[−0.36; −0.16]
Openness	4.64	0.70	.78	.84	4.03	0.97	.84	.88	4.36	0.80	.80	.86	−0.36	−4.5	<.001	[−0.48; −0.18]
Conscientiousness																
Industriousness	4.12	0.79	.85	.87	4.15	0.76	.82	.85	4.10	0.70	.76	.81	0.07	1.07	.289	[−0.04; 0.15]
Orderliness	4.07	0.82	.83	.88	4.21	0.76	.80	.86	4.08	0.80	.80	.86	0.17	2.35	.021	[0.02; 0.24]
Extraversion																
Assertiveness	4.14	0.74	.85	.88	3.94	0.73	.82	.87	3.82	0.85	.88	.91	0.15	1.85	.068	[−0.01; 0.24]
Enthusiasm	4.31	0.86	.89	.92	4.06	0.78	.86	.89	4.26	0.83	.89	.92	−0.24	−3.33	<.001	[−0.31; −0.08]
Agreeableness																
Compassion	5.00	0.61	.81	.86	4.49	0.88	.89	.92	4.89	0.66	.81	.86	−0.5	−5.24	<.001	[−0.55; −0.25]
Politeness	4.51	0.60	.69	.74	4.43	0.62	.67	.76	4.61	0.65	.74	.81	−0.28	−3.66	<.001	[−0.28; −0.08]
Neuroticism																
Volatility	3.14	0.94	.91	.94	3.28	1.02	.91	.94	2.94	0.94	.89	.92	0.35	5.29	<.001	[0.21; 0.47]
Withdrawal	3.27	0.88	.85	.89	3.51	0.93	.88	.90	3.23	0.93	.87	.90	0.31	3.95	<.001	[0.14; 0.43]

α: internal consistency (Cronbach's alpha); ω: factor reliability (McDonald's omega). A and A′ refer to the Situations in Study 1, and B and C refer to the Situations in Study 2. $\Delta M_{A-A'}$: mean difference of an aspect scale between Situations A and A′ in Study 1 (upper half); ΔM_{B-C}: mean difference of an aspect scale between Situations and B and C in Study 2 (lower half).
Note. Study 1: $N = 158$; Study 2: $N = 115$.

computation of individual consistency scores across two measurements where their mean is virtually identical to the sample-level observed rank-order consistency. For the general case of two Situations A and A′, $I_{A.A'}$ is computed as follows

$$I_{A.A'} = 1 - \frac{(z_A - z_{A'})^2}{2} \quad (1)$$

where $I_{A.A'}$ is the individual consistency score for one person regarding one BFAS state, z_A and $z_{A'}$ are the *z*-scores on the state at Situations A and A′, respectively.

In the current study, we presented participants two functionally equivalent situations, hence A and A′. To use the so obtained $I_{A.A'}$ scores (e.g., to correlate them), they needed to be transformed due to their heavy skewness (for the exact procedure, see Asendorpf, 1990, p. 9), which is nearly identical to the Fisher *r*-to-*z* transformation, but is also defined for $r = 1$. The resultant transformed scores $TI_{A.A'}$ are approximately normally distributed and thus can be used for further analyses. $TI_{A.A'}$ scores were derived for simple ($r_{A.A'}$) and residual consistencies ($r_{eA.eA'}$). Intercorrelations of $TI_{A.A'}$ scores allow to examine the structure of individual differences in consistencies across two Situations A and A′. In that sense, strong correlations point towards a possible general factor of consistency.

Deviations from the ore-registration. Study 1 was pre-registered. Although our research question remained unchanged, we deviated from our pre-registration. First, we intended to sample $N = 250$ participants, but only 158 participants completed the final study. We initially planned to compute reliable change scores (RCS) for change from T_0 to Situation A as well as for the change from T_0 to Situation A′. We conducted these analyses in an earlier version of this study (which can be found online at osf.io/nueq6). However, it turned out that computing RRs was much more informative compared to RCS. Two RCS (say, at Situations A and A′) will suffer from autocorrelations if the same score (i.e., the

assessment from T_0) is subtracted from scores at these two Situations A and A′. We furthermore intended to compute latent change score models (McArdle, 2009). However, they are affected by the same limitation as RCS. We therefore concluded that the current analyses based on RRs would be more defensible and informative than those other analyses. Furthermore, we contrasted all three approaches in a simulation study (which can be found online), which substantiates this position.

Results

Mean-level state change. We examined to what extent state means differed between Situations A and A′. The results are displayed in Table 2 under "$\Delta M_{A\text{-}A'}$". As can be seen, there was only one significant yet small mean-level difference for the aspect Orderliness ($d = 0.12$, $t = 3.29$, $p < .001$), which was contrary to our expectations. All other aspects did not differ substantially (dependent ts between -1.69 and 1.49, ds from $-.09$ to .12, with an average absolute d of 0.045; $SD = 0.03$), as expected. We interpret this pattern of findings as support for the functional equivalence of the two situation vignettes.

Trait expressions. Correlations between the trait scores at T_0 and the corresponding state scores at Situations A and A′ ($r_{T0.A}$, $r_{T0.A'}$) are presented in Table 3 under "Trait Expressions". The average trait expressions were $r_{T0.A} = .81$, $p < .001$, and $r_{T0.A'} = .81$, $p < .001$. Additionally, as expected, they were very similar for Situations A and A′ (only the two aspects of Neuroticism were statistically different at $p < .05$, see Table 3, Study 1, "Trait Expressions"), again speaking for the functional equivalence of the vignettes.

Simple consistencies. Correlations between the state scores at Situations A and A′ ($r_{A.A'}$) are presented in Table 3 under the first column of "Consistencies". As can be seen, all simple consistencies were substantial, ranging from .82 (Politeness) to .91 (Orderliness). The average simple consistency was .86 ($SD = .12$). Thus, participants showed strong levels of simple rank-order consistency, which is in line with the functional equivalence of the situation vignettes (i.e., rank-orders among participants are preserved in same situations).

Next, we examined the full intercorrelations of raw state scores for Situations A and A′, as presented in the upper half of Table OSM A[3] (lower triangle) under "For State Scores". The correlation matrix allows drawing conclusions based on convergent and discriminant associations. First, the gray-shaded cells capture convergent associations, which have already been detailed as $r_{A.A'}$ in Table 3. These convergent correlations were, as already described above, on average .86 ($SD = .12$), while the off-diagonal discriminant correlations (i.e., different aspects correlating at the same or different situations) amounted to an average *absolute* correlation of only .25 ($SD = .18$). Thus, we can conclude that state aspect scores for different traits measured twice were associated in a way providing convergent and discriminant validity evidence for state aspect scores.

Lastly, we analyzed how interindividual differences in simple consistencies of each aspect were related to each other. These findings are presented in the upper half of Table 4 under "For State Scores" (lower triangle). As can be seen, correlations were in general rather small (and failed to reach conventional levels for statistical significance in all but one case). The average *absolute* intercorrelation amounted to .10 ($SD = .06$), with the highest correlation being .26 (between Volatility and Assertiveness). This lack of covariation does not speak for a general factor of simple consistency and thus no formal general factor confirmatory factor analysis was performed.

Residual consistencies. Residual consistencies ($r_{eA.eA'}$), as the correlations between state residual scores at Situation A and state residual scores at Situation A′, can be found in Table 3 under the second column of "Consistencies". As can be seen, residual consistencies were generally moderate to high, ranging from .48 (Politeness) to .67 (Compassion), with an average of .59 ($SD = .10$). Thus, participants showed substantial levels of residual rank-order consistency which means that residual state variance was systematic across functionally equivalent situations.

Second, we again inspected the convergent and discriminant correlations of residual state scores, which are presented in Table OSM A in the lower triangle under "For Residual State Scores". Only very few different aspects were substantially correlated with each other, indicating good discriminant validity of the residual state scores. Specifically, for residual state scores, the convergent correlations (as already presented in Table 3 under $r_{eA.eA'}$) amounted on average to .59 ($SD = .10$), while the off-diagonal discriminant ones amounted to an average *absolute* correlation of only .18 ($SD = .13$). These patterns suggest that state residuals are aspect-specific and intercorrelations not likely reducible to method or artifact effects.

Lastly, we also analyzed how individual differences in residual consistencies (across aspects) were related to each other (see Table 4 in the lower triangle under "For State Residuals"). Correlations were rather small (as has already been the case for interindividual differences in simple consistencies) and only one was significant at $p < .05$ (Volatility with Assertiveness). The average *absolute* intercorrelation amounted to .10 ($SD = .06$), with the highest correlation being .26 (between Volatility and Assertiveness). This lack of covariation does not speak for a general factor of

Table 3. Trait expression, simple consistency, residual consistency, and study differences in consistencies.

	Study 1						Study 2						Differences between studies			
	Trait expression				Consistency		Trait expression				Consistency		Simple consistency		Residual consistency	
Domains and aspects	$r_{T0.A}$	$r_{T0.A'}$	t^a	p	$r_{A.A'}$	$r_{eA.eA'}$	$r_{T0.B}$	$r_{T0.C}$	t^a	p	$r_{B.C}$	$r_{eB.eC}$	z^b	p	z^b	p
Openness																
Intellect	.83	.84	−0.48	.315	.87	.57	.81	.85	−1.35	.09	.79	.33	2.11	**.017**	2.47	**.007**
Openness	.76	.78	−0.78	.218	.86	.65	.59	.77	−3.43	**<.001**	.62	.32	4.58	**<.001**	3.52	**<.001**
Conscientiousness																
Industriousness	.79	.80	−0.37	.355	.82	.53	.65	.68	−0.63	.263	.75	.56	1.48	.069	0.32	.375
Orderliness	.86	.84	1.09	.138	.89	.61	.67	.79	−2.72	**.004**	.70	.38	4.47	**<.001**	2.60	**.005**
Extraversion																
Assertiveness	.79	.78	0.39	.348	.85	.60	.64	.75	−2.16	**.016**	.65	.33	3.88	**<.001**	2.87	**.002**
Enthusiasm	.84	.85	−.56	.289	.90	.66	.70	.76	−1.33	.093	.70	.36	4.88	**<.001**	3.46	**<.001**
Agreeableness																
Compassion	.77	.78	−0.41	.342	.87	.67	.52	.62	−1.35	.090	.47	.23	6.64	**<.001**	4.67	**<.001**
Politeness	.79	.76	1.00	.160	.79	.48	.70	.77	−1.50	.069	.66	.27	2.25	**.012**	2.00	**.022**
Neuroticism																
Volatility	.84	.80	1.70	**.046**	.84	.52	.80	.76	1.07	.143	.75	.37	2.00	**.023**	1.55	.061
Withdrawal	.84	.80	1.80	**.037**	.86	.57	.67	.78	−2.30	**.012**	.65	.28	4.18	**<.001**	2.89	**.002**

Note. Study 1: $N = 158$; Study 2: $N = 115$.

[a]Differences between trait expressions were computed using a dependent *t*-test, with the simple consistency estimate as correlation with the third variable (Steiger, 1980).

[b]Differences of simple consistency and residual consistency between studies were computed using an independent z-test.

$r_{T0.A}$, $r_{T0.A'}$, $r_{T0.B}$, and $r_{T0.C}$: bivariate correlation of an aspect at T0 with state at Situations A and A′ (Study 1) and at Situations B and C (Study 2), respectively.

$r_{A.A'}$ and $r_{B.C}$: simple consistency, as the bivariate correlation of a state at Situation A with itself at Situation A′ (Study 1), and at Situation B with itself at Situation C (Study 2).

$r_{eA.eA'}$ and $r_{eB.eC}$: residual consistency, as the correlation of a state residual (here a regression residual score where a state was predicting from the corresponding trait) at Situation A with itself at Situation A′ (Study 1) and at Situation B with itself at Situation C (Study 2), respectively.

Significant *p* values are bolded.

residual consistency and thus no further analyses were conducted.

Discussion

In Study 1, we demonstrated that participants showed high levels of simple *as well as* residual consistency in functionally equivalent situations. We interpret this finding as support for our model presented in Figure 1. Both stable interindividual differences (personality traits, in this case, aspects), as well as specific characteristics of the Situation (A and A′), contribute to participants' response to state items. To go back to our example, if persons were to dance at a conference dinner, it could in part be explained by their general tendency to dance, but also by the specifics of the situation. Here, variance in dancing across situations is consistent, but unrelated to persons' general tendency to dance.

Furthermore, as discriminant validities of RRs indicated (Table OSM A), the correlations of state residuals cannot be attributed to overall method effects or other unspecific artifacts of the situation, as this would have influenced responses of participants across different aspects (and thus RRs across aspects would have been correlated). Furthermore, response styles or other method artifacts related to questionnaires can also not explain these findings, as this variance would be present in the trait scores as well as in both state scores – and thus removed when regressing states at Situations A and A′ on traits at T_0.

However, as seen in the current study, it is still possible that state residuals correlate, even after controlling for personality traits. If this finding can be replicated, it would mean that other stable person characteristics than the personality traits assessed at T_0 are present when assessing states in a specific situation. Besides other unmeasured personality traits, these could be stable trait-like tendencies in situation perception that may be drivers of residual consistency (Ziegler & Horstmann, 2015; Ziegler et al., 2019). Further, simple consistency could be high due to either stable tendencies within the person (i.e., traits) that were not controlled for or stable tendencies within the situation (i.e., functionally equivalent situations harbor situation characteristics that evoke similar mental and behavioral reactions). Indeed, several studies have sought to rule out the second explanation of consistency by having participants engage in *functionally non-equivalent* situations (Back et al., 2009; Borkenau et al., 2004; Fleeson & Law, 2015; Leikas et al., 2012; Morse et al., 2015; Riemann et al., 1998; Weisbuch et al., 2010) so that any observed consistency would be due to the actor or traits within individuals. However, these studies

Table 4. Intercorrelations of individual differences in simple and residual consistency scores in Study 1 and Study 2.

Domains and aspects	1	2	3	4	5	6	7	8	9	10
For state scores										
Openness										
1. Intellect	–	.20	.31*	.20	.16	.30*	.10	.07	.10	.26
2. Openness	.16	–	.06	.23	.18	.24	.27	.05	.16	.15
Conscientiousness										
3. Industriousness	.21	.11	–	.12	.10	.23	.07	.12	.10	.30*
4. Orderliness	.15	−.06	.08	–	.31*	.25	.15	.33*	−.01	.17
Extraversion										
5. Assertiveness	.10	.16	.04	.15	–	.10	.24	.32*	.04	.26
6. Enthusiasm	.14	−.07	.10	−.05	.06	–	.30*	.05	.16	.25
Agreeableness										
7. Compassion	.10	.09	.13	.00	.10	.13	–	.22	.26	.28
8. Politeness	.19	.04	.12	.06	.07	.07	.05	–	.07	.10
Neuroticism										
9. Volatility	.06	−.01	.10	.18	.26*	.12	.01	.08	–	.37**
10. Withdrawal	.00	.12	.01	.08	.18	−.06	.08	.03	.16	–
For state residuals										
Openness										
1. Intellect	–	.21	.36**	.13	.15	.18	.10	.11	.13	.21
2. Openness	.18	–	.11	.24	.21	.26	.15	.09	.11	.03
Conscientiousness										
3. Industriousness	.20	.12	–	.11	.11	.22	.10	.10	.10	.28
4. Orderliness	.14	−.05	.09	–	.25	.14	.05	.19	−.12	−.02
Extraversion										
5. Assertiveness	.11	.19	.07	.15	–	.13	.23	.31*	.02	.19
6. Enthusiasm	.15	−.07	.10	−.05	.08	–	.31*	.10	.09	.19
Agreeableness										
7. Compassion	.10	.10	.13	−.02	.10	.13	–	.21	.28	.30*
8. Politeness	.20	.04	.14	.05	.06	.09	.07	–	.10	.06
Neuroticism										
9. Volatility	.03	.03	.11	.11	.26*	.13	.00	.06	–	.39***
10. Withdrawal	.00	.15	.03	.08	.20	−.06	.12	.02	.17	–

Note. Study 1: $N = 158$; Study 2: $N = 115$.
Upper half: Correlations of individual simple consistency scores (based on states). The lower triangle displays the results from Study 1, and the upper triangle results from Study 2.
Lower half: Correlations of individual residual consistency scores (based on state residuals). The lower triangle displays the results from Study 1, and the upper triangle results from Study 2.
*$p < .05$, **$p < .01$, ***$p < .001$.

Table 5. Descriptive statistics and internal consistencies of trait scores in Studies 3 and 4.

	Study 3				Study 4			
	n	*M*	*SD*	α	*N*	*M*	*SD*	α
Honesty/Humility	196	4.14	0.78	.70	202	3.32	0.55	.63
Emotionality	196	3.82	0.83	.79	202	3.29	0.67	.76
eXtraversion	196	3.86	0.83	.80	202	3.56	0.62	.79
Agreeableness	196	3.75	0.75	.75	202	3.30	0.63	.75
Conscientiousness	196	4.25	0.80	.83	202	3.58	0.57	.75
Openness	196	4.47	0.77	.73	202	3.21	0.66	.74

α: Cronbach's alpha as reported in the original studies by Horstmann et al. (in press, Table 3).

have not investigated the degree of residual consistency in those functionally non-equivalent situations. Before we engage in a more specific discussion of our findings, we sought to rule out this alternative explanation by conducting the exact same study using functionally non-equivalent situations.[4]

Study 2

In Study 2, we used the exact same design as in Study 1, but with functionally *non-equivalent* situations. We expected to obtain outcomes in line with results from Study 1: Both simple and residual consistencies

should be lower in Study 2 as compared to Study 1, and no general factor of interindividual simple or residual consistency should emerge. Study 2 was not pre-registered.

Methods

Participants. $N = 115$ participants completed all three stages of Study 2. Of these, 78.26% were female, 20.87% were male, and the rest preferred not to indicate their gender. Their mean age was $M = 25.57$ ($SD = 7.64$). Most participants (92%) were students. No other characteristics of the participants were assessed.[5] Concerning age and gender, the sample is comparable to that of Study 1. Based on our results from Study 1 and subsequent simulations, we planned to sample $N = 100$ participants for Study 2.

Procedure. The study design was very similar to the design of Study 1 and was conducted online. Participants were invited via websites (e.g., Facebook) and e-mail lists to participate. After clicking on the survey platform (hosted again with formr.org, Arslan et al., 2020), participants were required to give informed consent and register with their e-mail address. Participants were then invited to indicate their age, gender, and occupation. Subsequently, participants rated their personality traits (T_0), using the same instructions as in Study 1. Two days later (T_1), participants were again invited via e-mail to rate their hypothetical personality state in a specific situation (Situation B) on the BFAS. Two days later (T_2), participants were invited again to participate and rate their hypothetical state in another situation (Situation C). The order of Situations B and C was determined at random for each participant. After participants indicated their state in both situations, they received feedback on their personality trait scores on all 10 Big Five aspects. Psychology undergraduate students could also obtain course credit for completion of all three measurement occasions.

Materials. Similar to Study 1, all measures were assessed online on the platform formr.org (Arslan et al., 2020). Participants could again not skip items but could decide to abort participation at any time. All materials relevant to replicate and reproduce the current study can be found at osf.io/nueq6.

Situation vignettes. Contrary to Study 1, we presented participants with two situations that were intended to be functionally different.[6] The first situation described a more or less common situation in the life of an undergraduate student: meeting other students to prepare for an upcoming exam. In this scenario, participants should imagine meeting with other students, while they are not equally well prepared for the upcoming examination. In the second situation, participants should imagine meeting with another attractive person who they have already dated several times. In the current scenario, the new date brings the participant to a party where they are about to meet some friends who are so far unknown to the participant. Vignettes B and C are both 235 words long.

Traits and states. We used the exact same measures and instructions as in Study 1.

Differences between Study 1 and Study 2. The intention of Study 2 was to test our suggested effects in functionally non-equivalent situations. Furthermore, we shortened the time interval between the assessment of personality traits at T_0 and states at T_1 and T_2. Whereas participants had to wait three weeks between assessments in Study 1, participants only had to wait two days between assessments in Study 2. We deliberately chose a smaller interval as this would make it easier to recruit participants and reduce attrition from one measurement occasion to another. Moreover, any carry-over memory effects would make it *more difficult* to observe our hypothesized effect: If a participant remembered their response from a previous assessment and aimed at responding similarly, this would likely *increase* consistency across functionally non-equivalent situations, which would be contrary to our expectations. Choosing a smaller interval in Study 2 therefore is more conservative in our case.

Data analyses. For the computation of the residual state scores, their intercorrelations, and the interindividual differences in consistencies as well as their intercorrelations, we followed the exact same data-analytic strategy as in Study 1. Both residual state scores as well as interindividual differences in consistencies were computed exactly as in Study 1. Descriptive statistics and internal consistency estimates of all aspects are displayed in Table 2. Again, the internal consistency estimates were similar to those reported by DeYoung et al. (2007) and comparable to those from Study 1.

Results

Mean-level state change. As we designed the two Situations B and C to be functionally non-equivalent, we expected significant and sizable mean-level changes in participants' responses from Situations B to C. From 10 possible comparisons, one was significant at $p < .05$, and seven were significant at $p < .001$ (see lower half of Table 2 in column "$\Delta M_{\text{B-C}}$"). Standardized mean-level differences ranged from $d = -0.50$ ($t = -5.24$, $p > .001$, for Compassion) to $d = .35$ ($t = 5.29$, $p < .001$, for Volatility), with an average absolute d for all comparisons of 0.27 ($SD = 0.12$). In accordance with our interpretation in Study 1, we interpret this pattern as evidence for the functional non-equivalence of Situations B and C.

Trait expressions. The correlations of the trait scores at T_0 with the states scores at Situations B and C ($r_{T0.B}$ and $r_{T0.C}$) are displayed in Table 3 under "Trait Expressions". The trait expressions in Study 2 were comparable in size to those from Study 1 (see Table 3), with the average trait expression at $r_{T0.B} = .68$, $p < .001$, and $r_{T0.C} = .76$, $p < .001$. However, $r_{T0.B}$ and $r_{T0.C}$ were not as similar to each other as $r_{T0.A}$ and $r_{T0.A'}$ from Study 1. Although trait expressions were rather high, they differed depending on the Situation (see $r_{T0.B}$ vs. $r_{T0.C}$ in Table 3, and in four cases, this difference was statistically significant). Together, the current findings can be interpreted as further evidence for the functional non-equivalence of situations in Study 2.

Simple consistencies. The correlation between the state scores at Situations B and C ranged from .47 (Compassion) to .79 (Intellect, see Table 3, column $r_{B.C}$ as well as Figure 2). The average simple consistency was $r_{B.C} = .68$ (SD = .16), and is thus lower than the average simple consistency observed in Study 1 ($r_{A.A'} = .86$).[7] Although somewhat lower, the simple consistencies are nevertheless high, which means that participants still showed a remarkable degree of simple consistency across functionally non-equivalent Situations B and C.

Similar to Study 1, we also examined the full intercorrelations of state scores at Situations B and C. These correlations are presented in the upper triangle of the upper half of Table OSM A under "For State Scores". The gray-shaded cells depict the convergent correlations on the diagonal, and the correlations on the off-diagonal depict the discriminant correlations. The average *absolute* discriminant correlation was $r = .27$ ($SD = .23$), whereas the average absolute convergent correlation was .68 ($SD = .16$). This pattern speaks for the convergent and discriminant validity of the state aspect scores.

Finally, we examined interindividual differences in simple consistency of each aspect. These results are displayed in Table 4 in the upper triangle under "For State Scores". Seven correlations were significant (at $p < .05$), but a general pattern of correlated interindividual differences in simple consistencies failed to emerge. The average *absolute* correlation was .19 ($SD = .10$). We interpret this pattern as evidence against the existence of a general factor of consistency and have thus refrained from any further analyses.

Residual consistencies. The residual consistencies ($r_{eB.eC}$) are displayed in Table 3. We expected the residual consistencies in Study 2 to be substantially lower compared to those from Study 1. In line with this, the residual consistencies ranged from .27 (Politeness) to .56 (Industriousness), with an average of .35. Participants therefore showed lower, but still substantial rank-order consistency of RRs compared to Study 1 (on average, .35 vs. .59, respectively).

Second, we investigated the full intercorrelations of residual state scores (see Table OSM A in the upper triangle under "For Residual State Scores"). The gray-shaded convergent correlations (which have also been displayed in Table 2, under $r_{eB.eC}$) were substantially lower than those from Study 1, and only 4 out of 10 were significant at $p < .05$. The average *absolute* discriminant correlation was $r = .20$ ($SD = .17$), whereas the average absolute convergent correlation was .35 ($SD = .11$). In line with our interpretation from Study 1, these patterns suggest that the state residuals are still aspect-specific and cannot be simply explained away as methods effects.

Finally, we again analyzed the correlations of individual differences in residual consistencies across aspects (see Table 4 in the upper triangle under "For State Residuals"). Similar to Study 1, the correlations were comparatively small, and only five were statistically significant (one at $p < .001$, one at $p < .01$, and three at $p < .05$). The average *absolute* intercorrelation amounted to .17 ($SD = .09$). The highest correlation occurred between Volatility and Withdrawal ($r = .39$), two aspects from the same domain of Neuroticism. Similar to Study 1—and even more expected due to functionally different situations—this pattern again did not support a general factor of residual consistency and thus no further analyses were warranted.

Discussion

Summary and interpretation. Study 2 was conducted to investigate the degree of trait expression, simple and residual consistencies in functionally non-equivalent situations. First, trait expression was comparable to that of Study 1. Although the situations were designed to be different, looking simply at the trait expressions would not have alerted us to any differences of the situations.

Second, and although lower than compared to Study 1, the simple consistencies were rather high (i.e., state levels in Situation B corresponded to those in Situation C). This simple consistency in functionally non-equivalent situations provides further evidence for stable interindividual differences in states, even across functionally different situations. However, simple consistency may also be high due to other artifacts, such as method effects (e.g., using the same questionnaire at T_0, Situation B, and Situation C) or simply because participants remembered their responses from the previous assessment. However, simple consistency has been established previously and was not the main focus of the current study.

The main finding of Study 2 was that simple and residual consistencies were indeed lower compared to Study 1 (see Table 3 under "Differences between Studies", where 17 out of 20 comparisons are significant), which is in line with our expectations. Yet,

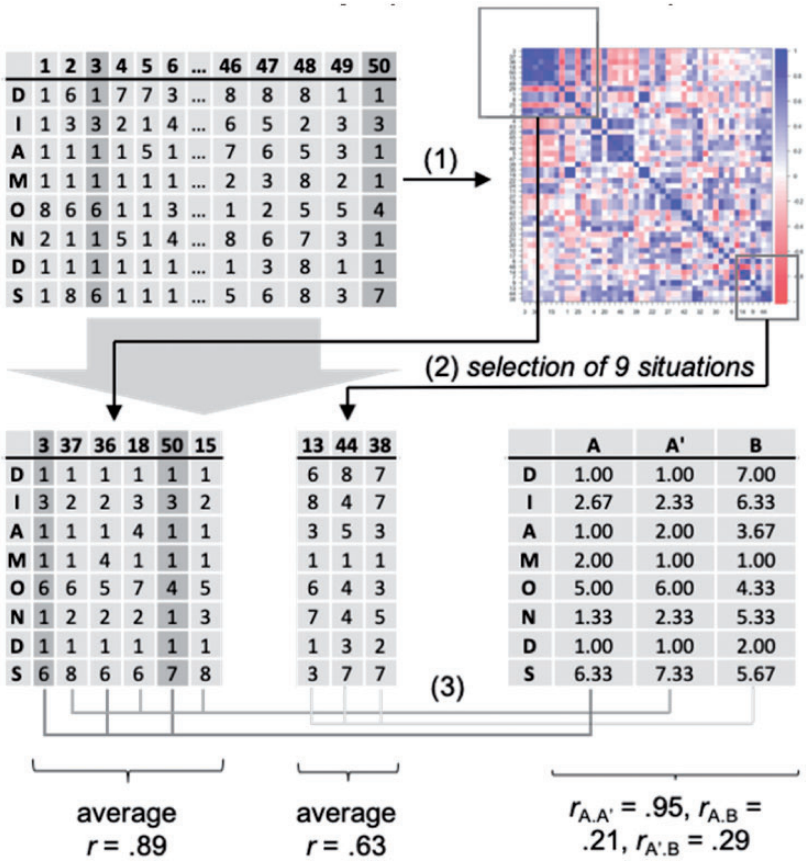

Figure 2. Selection of functionally equivalent and non-equivalent situations. *Note*. Selection of functionally equivalent and non-equivalent situations based on DIAMONDS profiles in Study 3 and Study 4 as exemplified for participant 193 from Study 3. For explanation of coefficients and arrows, see text.

residual consistency in Study 2 was substantially different from zero, although the situations were intended to be functionally non-equivalent (see Table OSM A the gray-shaded cells in the upper triangle of the lower half). Thus, there still seem to be some (unmeasured) person characteristics operating that result in shared systematic variance in residual state scores across functionally non-equivalent situations.

Limitations. Although Studies 1 and 2 provided an initial insight into the possible nature of (a) functionally equivalent and non-equivalent situations and (b) simple and residual consistencies, both had some limitations. First, both studies resorted to hypothetical situation vignettes and people's self-reports of their *imagined* Big Five aspect states in those situations. Thus, these findings hold for people's self-views and imagined states. Second, the items that were used to capture traits may not fully apply to personality states, which could have led to increased trait-state correlations. Further, it is important to note that laboratory or experimental studies will yield opportunities for easily standardizing situations and thus ensuring that people can enter a certain situation twice. Assigning participants randomly to such experimental situations, their traits and the characteristics of the situations will not be substantially correlated as could be the case in real life where there is non-random assortment of phenotypical traits and situations (e.g., Ickes et al., 1997) as people may encounter functionally similar situations that they may have chosen, changed, or evoked (Emmons & Diener, 1986; Rauthmann & Sherman, 2016a; Rauthmann, Sherman, Nave, et al., 2015). If this is the case, associations between trait scores and state scores (trait expressions) as well as between different state scores (simple consistency) or state residual scores (residual consistency) might also change. Thus, our findings from the previous studies, where people were selected into situations (as opposed to selecting situations on their own) need to be replicated in naturalistic settings within people's everyday lives. It was therefore our goal to examine the same research questions in additional data from experience sampling studies.

Study 3

In Studies 1 and 2, we showed that simple and residual consistencies can be examined using situation vignettes, asking participants about their hypothetical

personality states (i.e., "how would you behave in this situation?"). However, as outlined above, there are several reasons why this should also be tested in a naturalistic setting. To this end, we examined data from an experience sampling study (Horstmann et al., in press) in which participants first reported their personality traits and then repeatedly for several days their momentary behavior (states) and the psychological characteristics of their current situation. Materials for the replication of Study 3 can be found at osf.io/zctv4/; data and analyses code for the reproduction of results are available at osf.io/xfhdu/.

Method

Participants. After removing participants who had less than 12 measurement occasions, data from $N = 196$ (mean age $M = 24.19$, SD = 6.37, 86% female) participants were analyzed in the current study. Participants were mainly psychology students from a German university. They were recruited via e-mail lists and social media. Participants could obtain course credit for the completion of the study, as well as personalized feedback on their personality traits.

Procedure. The whole study was conducted online, using the platform formr.org (Arslan et al., 2020). Participants were first required to answer several demographic questions. Subsequently, participants continued in the experience sampling part of the study. Here, participants received an e-mail every three hours and were asked to complete a brief survey about their current situation, their affect, and their behavior in this situation. Participants could complete up to 50 experience sampling surveys. This study ended if either the participants had completed 50 measurement occasions or chose to no longer participate. The median measurement occasion per participant was 41, with 70 participants completing all possible 50 measurement occasions.

Materials. All materials for the replication of this part of the study can be found at the corresponding OSF website from the original article by Horstmann and colleagues (in press) at osf.io/zctv4/. Not all measures that were initially assessed were analyzed in the current study. However, a complete overview of all scales assessed can be found in Horstmann et al. (in press).

Traits. To assess personality traits, the German translation of the HEXACO-60 (Moshagen et al., 2014) was used. Each dimension is assessed with 10 items. Participant responded to items on a 6-point rating scale (1 = *does not apply at all*, 6 = *does fully apply*). The descriptive statistics and internal consistency estimates of trait scores are presented in Table 5.

Situational ratings. During experience sampling, participants rated the psychological characteristics of their current situation on the S8-I for each of the DIAMONDS (Rauthmann & Sherman, 2016b) in its German translation. The DIAMONDS dimensions contain Duty (*Work has to be done*), Intellect (*Deep thinking is required*), Adversity (*Someone is being threatened, blamed, or criticized*), Mating (*Potential romantic partners are present*), pOsitivity (*Situation is enjoyable*), Negativity (*Situation includes negative feelings (e.g., stress, anxiety, guilt)*), Deception (*Someone is being deceived*), and Sociality (*Social interaction is possible or required*). Participants indicated how characteristic each statement was for each situation on an 8-point rating scale (1 = *extremely uncharacteristic*, 8 = *extremely characteristic*).

States. Personality states were assessed using a bipolar adjective 8-point rating scale. The states assessed correspond to the Big Six personality domains[8]: Honesty/Humility (*humble–arrogant*), Neuroticism (*nervous–calm*), Extraversion (*outgoing–reserved*), Agreeableness (*agreeable–quarrelsome*), Conscientiousness (*organized–disorganized*), and Openness (*creative–uncreative*).

Data analyses. All data relevant for the replication of the results of Study 3 can be found online at osf.io/xfhdu/, alongside the R code for data cleaning and data analysis. Participants were excluded if they completed less than 12 measurement occasions. Fifty-eight participants did not meet this criterion and where removed. We chose 12 as the minimum number of required situations as this gave some degrees of freedom regarding the selection of functionally equivalent and non-equivalent situations.

Selection of functionally equivalent and non-equivalent situations. For each participant, three situations were created, A, A′, and B. A and A′ denote functionally equivalent situations, and B a situation which is non-equivalent to A and A′. Note, however, that we only examined the comparison of A vs. B, not A′ vs. B, as A and A′ are highly similar[9] (see below). Running all additional analyses A′ vs. B would therefore not provide any additional information. The selection of situations for the current analyses is depicted in Figure 2.

As indicated previously, each situation was rated on the Situational Eight DIAMONDS. We first computed profile correlations of the DIAMONDS profiles of all situations *within* one participant (see Figure 2, here exemplified for Participant 193 who completed $m = 50$ measurement occasions). Subsequently, the correlation matrix was sorted, such that situations that had very similar DIAMONDS profiles were ranked close to one another. We then selected those six situations that

were highly similar to each other, and three situations that were similar to another, but dissimilar to the first six situations (top-left corner of correlogram and bottom-right corner of correlogram in Figure 2). For Participant 193, the DIAMONDS profiles of the six similar situations correlated on average at $r = .89$, and the three other similar situations correlated on average at $r = .63$. We then averaged three state scores within each set of situations to form the state score at Situations A, A′, and B, respectively[10]: To exemplify, for Participant 193, measurement occasions 3, 36, and 50 formed Situation A, measurement occasions 37, 18, and 5 formed Situation A′, and measurement occasions 13, 44, and 38 formed Situation B. The profile correlations of the averaged DIAMONDS for Participant 193 were $r_{A.A'} = .95$, $r_{A.B} = .21$, and $r_{A'.B} = .29$. The averaged states within the situations then formed the state scores for each participant that went into further analyses analogous to those from Studies 1 and 2.

Deriving residual state scores. Similar to Studies 1 and 2, state residual scores were obtained by predicting state scores in Situations A, A′, and B with the corresponding trait scores at the initial measurement T_0. We then extracted RRs for each of the Big Six personality domains.[11]

Results

Mean-level state change. We again examined mean-level changes in personality states[12] from Situations A to A′ and B, and Situations A′ to B (see Table 6), as indicated in the columns $\Delta_{A,A'}$, $\Delta_{A,B}$, $\Delta_{A',B}$. For self-reported states, only Conscientiousness changed significantly from A to Á, $p = .033$, $d = -0.12$. Thus, there were no substantial changes in mean level states from Situations A to A′. Contrary, the differences of A to B, and A′ to B were all substantially different (average absolute $d = .44$, $SD = .10$, and $d = .35$, $SD = .15$, respectively), and all but Honesty/Humility were statistically significant at $p < .05$. Thus, we can conclude that the self-reported personality states in Situations A and A′ were more similar to each other than to the states reported in Situation B.

Trait expressions. The trait expressions are displayed in Table 7. The average trait expressions were $r_{T0.A} = .25$, $p < .001$; $r_{T0.A'} = .23$, $p < .01$; and $r_{T0.B} = .18$, $p = .013$. We further examined the differences between trait expressions. With only one exception (extraversion across functionally non-equivalent situations), the domain-specific trait expressions did not differ significantly per functionally equivalent and non-equivalent situations (see column t and p under "Trait Expression").

Simple consistency. Simple consistencies (i.e., $r_{A.A'}$, $r_{A.B}$) are displayed in Table 7 under "Consistency". The average simple consistency for functionally equivalent situations was $r_{A.A'} = .67$ ($SD = .07$), and the average simple consistency for functionally non-equivalent situations was $r_{A.B} = .35$ ($SD = .19$). Simple consistency in functionally equivalent situations was always higher compared to simple consistency in functionally non-equivalent situations (see Table 7, column "Differences between FE and FNE").

Similar to Studies 1 and 2, we also examined intercorrelations of state scores across domains (see in Table OSM B the left half of the upper panel under "For State Scores", osf.io/pdu23). The lower triangle displays the correlations of state scores assessed at functionally equivalent situations, whereas the upper triangle displays the correlations for functionally non-equivalent situations. The average discriminant state score correlation was .29 ($SD = .24$) for functionally equivalent situations, and .22 ($SD = .19$) for functionally non-equivalent situations (off-diagonal), compared to an average of .67 and .35 for convergent correlations (Table OSM B, grey shaded areas), respectively. Similar to Studies 1 and 2, we thus conclude that state scores showed good convergent and discriminant validity.

Finally, we also investigated the correlation of interindividual differences in simple consistency across domains (see in Table 8 the left half of the upper panel, "For State Scores"). The *absolute* average interindividual consistency across functionally equivalent situations was .17 ($SD = .12$). Although four correlations were significant, no overall pattern of individual consistency across domains emerged. For functionally non-equivalent situations, a similar picture emerged, with an average correlation of .15 ($SD = .15$).

Residual consistency. Residual consistencies (i.e., $r_{eA.eA'}$, $r_{eA.eB}$) are displayed in Table 7. For both functionally equivalent and non-equivalent situations, residual consistencies were significantly different from zero. The average residual consistency in functionally equivalent situations was $r_{eA.eA'} = .65$ ($SD = .08$), and $r_{eA.eB} = .33$ ($SD = .18$) in functionally non-equivalent situations. Furthermore, residual consistency of each domain was always lower in functionally non-equivalent compared to functionally equivalent situations. Additionally, we also examined the discriminant associations of residual states scores. The average discriminant correlation of state residuals was .27 ($SD = .23$) for functionally equivalent and .21 ($SD = .17$) for functionally non-equivalent situations.

Similar to previous analyses, we investigated interindividual differences in residual consistency (see in Table 8 the left half on the upper panel, "For State Residuals") in functionally equivalent (lower triangle) and non-equivalent situations (upper triangle). The average correlation amounted to .17 ($SD = .12$) for

Table 6. Descriptive statistics of state scores in Situations A, A′, and B in Studies 3 and 4.

	A		A′		B		$M\Delta_{A,A'}$				$M\Delta_{A,B}$				$M\Delta_{A',B}$			
	M	*SD*	*M*	*SD*	*M*	*SD*	$d_{A,A'}$	$t_{A.A'}$	$p_{A.A'}$	95% $CI_{A.A'}$	$d_{A,B}$	$t_{A.B}$	$p_{A.B}$	95% $CI_{A.B}$	$d_{A',B}$	$t_{A'.B}$	$p_{A'.B}$	95% $CI_{A'.B}$
Study 3																		
Honesty/Humility	6.23	1.04	6.18	1.08	6.05	1.16	0.05	0.97	.332	[−0.06; 0.17]	0.16	**2.47**	**.014**	[0.04; 0.33]	0.11	1.56	.121	[−0.03; 0.29]
Emotionality	3.89	1.38	3.99	1.40	4.38	1.48	−0.07	−1.34	.181	[−0.26; 0.05]	−0.34	**−4.68**	**<.001**	[−0.7; −0.29]	**−0.27**	**−3.43**	**<0.001**	[−0.61; −0.17]
eXtraversion	5.62	1.30	5.53	1.30	4.96	1.37	0.07	1.26	.209	[−0.05; 0.25]	0.49	**5.27**	**<.001**	[0.41; 0.91]	**0.42**	**4.42**	**<0.001**	[0.31; 0.82]
Agreeableness	6.50	1.00	6.38	1.04	5.84	1.17	0.12	1.93	.055	[0.00; 0.25]	0.61	**7.00**	**<.001**	[0.48; 0.85]	**0.49**	**5.54**	**<0.001**	[0.35; 0.73]
Conscientiousness	5.54	1.29	5.37	1.33	4.90	1.31	0.12	2.15	.033	[0.01; 0.31]	0.49	**5.55**	**<.001**	[0.41; 0.87]	**0.36**	**4.10**	**<0.001**	[0.25; 0.71]
Openness	5.65	1.13	5.59	1.25	5.06	1.30	0.05	0.90	.371	[−0.07; 0.20]	0.48	**6.29**	**<.001**	[0.41; 0.78]	**0.42**	**5.20**	**<0.001**	[0.33; 0.73]
Study 4																		
Honesty/Humility	5.78	1.26	5.79	1.21	5.43	1.20	−0.01	−0.23	.817	[−0.11; 0.09]	0.29	**4.93**	**<.001**	[0.21; 0.49]	**0.30**	**4.84**	**<.001**	[0.22; 0.51]
Emotionality	3.31	1.52	3.38	1.58	3.77	1.45	−0.04	−0.78	.435	[−0.24; 0.10]	−0.3	**−3.85**	**<.001**	[−0.69; −0.22]	**−0.26**	**−3.17**	**.002**	[−0.63; −0.15]
eXraversion	5.05	1.56	5.11	1.55	4.22	1.43	−0.04	−0.79	.431	[−0.23; 0.10]	0.56	**6.52**	**<.001**	[0.58; 1.08]	**0.60**	**6.51**	**<.001**	[0.63; 1.17]
Agreeableness	5.64	1.27	5.58	1.28	4.92	1.36	0.05	0.85	.396	[−0.08; 0.20]	0.55	**7.46**	**<.001**	[0.53; 0.92]	**0.50**	**6.44**	**<.001**	[0.46; 0.87]
Conscientiousness	4.97	1.47	4.96	1.43	4.75	1.37	0.01	0.21	.837	[−0.15; 0.18]	0.16	2.11	.036	[0.01; 0.44]	0.15	1.95	0.053	[0.00; 0.42]
Openness	5.30	1.27	5.25	1.29	4.89	1.27	0.04	0.70	.483	[−0.09; 0.18]	0.33	**4.66**	**<.001**	[0.24; 0.59]	**0.29**	**4.20**	**<.001**	[0.19; 0.54]

A, A′, B: scores at Situations A, A′, and B; *d*: standardized mean difference between Situations A and A′, A and B, and A′ and B (Cohen's *d*); *t*: *t* value of difference, *p*: *p* value of difference, 95% CI: 95% confidence interval of mean difference (*d*).

Note. Study 3: $N = 196$; Study 4: $N = 202$.

Significant effects are bolded.

Table 7. Trait expressions, simple consistency, residual consistency, and study differences in consistencies for Study 3 and Study 4.

	Functionally equivalent situations						Functionally non-equivalent situations						Differences between FE and FNE[b]			
	Trait expression[a]				Consistency		Trait expression[a]				Consistency		Simple consistency		Residual consistency	
Domains	$r_{T0.A}$	$r_{T0.A'}$	t	p	$r_{A.A'}$	$r_{eA.eA'}$	$r_{T0.A}$	$r_{T0.B}$	t	p	$r_{A.B}$	$r_{eA.eB}$	z	p	z	p
Study 3																
Honesty/Humility	.27	.19	1.41	.16	.71	.70	.27	.32	−0.81	.42	.57	.53	**2.43**	**.01**	**2.72**	**<.001**
Emotionality	.29	.25	0.86	.39	.70	.68	.29	.27	0.27	.79	.47	.42	**3.47**	**<.001**	**3.61**	**<.001**
eXtraversion	.38	.37	0.14	.89	.66	.60	.38	.10	**3.12**	**<.001**	.14	.11	**6.39**	**<.001**	**5.78**	**<.001**
Agreeableness	.22	.19	0.50	.62	.60	.59	.22	.13	1.02	.31	.26	.23	**4.28**	**<.001**	**4.24**	**<.001**
Conscientiousness	.17	.19	−0.24	.81	.67	.66	.17	.12	0.63	.53	.22	.21	**5.76**	**<.001**	**5.73**	**<.001**
Openness	.18	.17	0.16	.87	.68	.67	.18	.12	0.83	.41	.42	.41	**3.77**	**<.001**	**3.72**	**<.001**
Study 4																
Honesty/Humility	.25	.24	0.34	.74	.84	.83	.25	.22	0.61	.54	.66	.64	**4.11**	**<.001**	**4.10**	**<.001**
Emotionality	.23	.10	**2.47**	**.01**	.69	.69	.23	.12	1.43	.15	.37	.35	**4.69**	**<.001**	**4.86**	**<.001**
eXtraversion	.29	.28	0.25	.80	.72	.69	.29	.27	0.26	.80	.27	.21	**6.22**	**<.001**	**6.39**	**<.001**
Agreeableness	.24	.23	0.27	.78	.70	.68	.24	.24	0.06	.95	.45	.42	**3.75**	**<.001**	**3.83**	**<.001**
Conscientiousness	.32	.21	**2.09**	**.04**	.66	.64	.32	.30	0.38	.70	.43	.37	**3.35**	**<.001**	**3.73**	**<.001**
Openness	.12	.21	−1.81	.07	.73	.72	.12	.26	**−2.08**	**.04**	.51	.50	**3.60**	**<.001**	**3.66**	**<.001**

Note. Study 3: $N = 196$; Study 4: $N = 202$.

[a]Differences between trait expressions were computed using a dependent t-test, with the simple consistency estimate as correlation with the third variable (Steiger, 1980).

[b]Differences of simple consistency and residual consistency between studies were computed using an independent z-test.

$r_{T0.A}$, $r_{T0.A'}$, $r_{T0.A,}$ and $r_{T0.B}$: bivariate correlation of a trait score at T0 with state score at Situations A and A′ and at Situations A and B, respectively.

$r_{A.A'}$ and $r_{A.B}$: simple consistency, as the bivariate correlation of a state at Situation A with itself at Situation A′, and at Situation A with itself at Situation B.

$r_{eA.eA'}$ and $r_{eA.eB}$: residual consistency, as the correlation of a state residual (here a regression residual score where a state was predicting from the corresponding trait) at Situation A with itself at Situation A′ and at Situation A with itself at Situation B.

Significant p values are bolded.

Table 8. Intercorrelations of individual differences in simple and residual consistency scores in Studies 3 and 4.

	Study 3						Study 4					
Domains	1	2	3	4	5	6	1	2	3	4	5	6
For state scores												
1. Honesty/Humility	–	.18	.13	.10	−.02	.05	–	.07	.24**	.09	.09	.08
2. Emotionality	−.07	–	.20	.25**	.21*	.17	.12	–	.11	.04	.22*	.02
3. eXtraversion	.06	.08	–	.25**	.11	.11	.13	.12	–	.20	.25**	.17
4. Agreeableness	.22*	.09	.43***	–	.08	.19	.34***	.22*	.33***	–	.24**	.20
5. Conscientiousness	.07	.05	.09	.14	–	.48***	.05	.04	.14	.12	–	.43***
6. Openness	.07	0	.16	.12	.50***	–	.22*	.04	.01	.15	.18	–
For state residuals												
1. Honesty/Humility	–	−.07	.08	.21*	.06	.07	–	.14	.13	.34***	.05	.20*
2. Emotionality	.19	–	.07	.09	.05	.01	.07	–	.13	.23*	.05	.04
3. eXtraversion	.11	.19	–	.40***	.05	.17	.23*	.10	–	.32***	.14	.01
4. Agreeableness	.10	.25**	.25**	–	.14	.12	.09	.04	.20*	–	.12	.14
5. Conscientiousness	−.01	.21*	.11	.09	–	.51***	.09	.18	.25**	.21*	–	.18
6. Openness	.05	.17	.11	.19	.49***	–	.09	−.01	.18	.23*	.42***	–

Note. Study 3: $N = 196$; Study 4: $N = 202$.

Upper half: Correlations of individual simple consistency scores (based on states). The lower triangle displays the results from functionally equivalent situations, and the upper triangle results functionally non-equivalent situations.

Lower half: Correlations of individual residual consistency scores (based on state residuals). The lower triangle displays the results from functionally equivalent situations, and the upper triangle results functionally non-equivalent situations.

*$p < .05$, **$p < .01$, ***$p < .001$.

functionally equivalent situations, and .14 ($SD = .15$) for functionally non-equivalent situations.

Discussion

The aim of Study 3 was to investigate simple and residual consistencies using a different method. Comparing the results from Study 3 with those of Studies 1 and 2, there are some differences that warrant further attention despite the general replication of the effects we expected from Studies 1 and 2.

First, trait expression was lower compared to Studies 1 and 2. There are several possible explanations for this. The trait measures and state measures were different in this study (i.e., different item formats, more items at trait compared to state level). Further, each participant experienced a situation in daily life that may or may not relate to each trait (e.g., the situation experienced may be relevant only for the expression of conscientiousness, but not for agreeableness). The situations were selected based on their similarity regarding their DIAMONDS profiles, but we do not know how relevant each situation was for the expression of certain traits. If the situation is not relevant for the expression of a trait, then this may lead to lower levels of trait expression (Tett & Burnett, 2003; Tett & Guterman, 2000).

Second, and comparable to Studies 1 and 2 as well as expected, both forms of consistency were lower in functionally non-equivalent situations compared to functionally equivalent situations. This substantiates the idea that state consistency is also a function of the functional equivalence of the situations in which it occurs, and that people behave more consistently in situations that are similar (Sherman et al., 2010).

Third, and most striking, residual consistencies were on average only marginally smaller than simple consistencies ($r_{A.A'} = .67$ vs. $r_{eA.eA'} = .65$ for functionally equivalent situations, on average, and, $r_{A.B} = .35$ vs. $r_{eA.eB'} = .33$ for functionally non-equivalent situations). Technically, simple consistencies can be regarded as the theoretical maximum for residual consistencies: Depending on the amount of trait expression that is controlled for, residual consistencies will be smaller. However, as the trait expression was rather small compared to Study 1 and Study 2, no substantial reduction in consistency was to be expected after controlling for trait scores. An alternative explanation would be that states have been assessed with one item only in Study 3, leading to lower reliability, and therefore to lower trait expression. However, the fact that consistency was nevertheless high suggests that at least some amount of specific, systematic variance was present.

Based on these findings alone, one could surmise that consistency of personality states across situations in daily life could be attributed to momentary, situational factors. However, before interpreting these findings, we aimed at replicating them in an additional, independent sample from a different country (Studies 1-3 used German samples). To this end, we reanalyzed data in Study 4, which parallels Study 3 very closely, to get a better and more robust picture of consistency when examined using experience sampling.

Study 4

Study 4 is a replication-of-analyses from Study 3, using an independent sample from the United States. The data, which were previously published by Sherman et al. (2015), have also been analyzed in other studies (Jones et al., 2017; Rauthmann et al., 2019; Rauthmann et al., 2016; Sherman & Pashler, submitted). The analyses and results we report here are new and have not yet been published previously. The data presented in Study 3 (Horstmann et al., in press) were originally collected to replicate findings from Sherman et al. (2015). Thus, the study design, procedure, and materials are very similar across Studies 3 and 4.

Method

Participants. $N = 210$ participants completed the whole study. Due to a computational error, trait scores of one participant were lost. This participant was removed for the current analyses. Seven further participants completed less than 12 measurement occasions and their data were excluded. Thus, the final sample in Study 4 contained $N = 202$ participants. Their mean age was $M = 18.61$ ($SD = 1.96$), and 66% were female.

Procedure. The data collection procedure was similar to that of Study 3. However, contrary to Study 3, trait measures were assessed in the laboratory, not online. After completing the trait assessments, participants received eight text messages per day for seven days and were required to respond to brief surveys, inquiring about their momentary behavior, happiness, and current situation on a predefined schedule (see Sherman et al., 2015, for details).

Materials. A full description of the study procedure and materials can be found in Sherman et al. (2015). Participants also completed other measures that were not analyzed here.

Traits. Personality traits were assessed using the HEXACO personality inventory (Ashton & Lee, 2009). Participants indicated on a 5-point rating scale how much they agreed with a certain statement (1 = *strongly disagree*, 5 = *strongly agree*). Items were averaged per scale to form the scale composite. Descriptive statistics and internal consistency estimates of trait scores are presented in Table 5.

Situational ratings. Participants rated the characteristics of their current situation on the S8-I (Rauthmann & Sherman, 2016b), indicating how characteristic a certain DIAMONDS item was for the current situation. Participants responded on a 7-point rating scale (1 = *extremely uncharacteristic*, 7 = *extremely characteristic*).

States. Self-reported personality states were assessed using the same bipolar rating scales as in Study 3. Participants reported their momentary state levels on a 7-point rating scale.

Data analyses. The data analyses parallel exactly those of Study 3. The same R code was used to select functionally equivalent and non-equivalent situations as well as for the extraction of residual state scores (see explanation in Study 3).

Results

Mean-level state change. Similar to the previous studies, we first examined mean-level change of self-reported, average personality states from Situations A to A′, and A and A′ to B (see Table 6 the lower half, columns $\Delta_{A,A'}$, $\Delta_{A,B}$, $\Delta_{A',B}$). While there were no substantial mean-level changes in states from Situations A to A′ (average absolute $d = 0.03$, $SD = .02$), nearly all comparisons from Situations A to B, and A′ to B were substantial (average absolute $d = .37$, $SD = .19$, and $d = .36$, $SD = .20$, respectively) and significant (with the exception of Conscientiousness).

Trait expression. Trait expressions (correlation between trait scores and state scores) are displayed in the lower half of Table 7. The average trait expressions amounted to $r_{T0.A} = .24$, $p < .001$; $r_{T0.A'} = .21$, $p < .01$; and $r_{T0.B} = .23$, $p = .001$. We again examined the differences between trait expressions, i.e. compared trait expressions $r_{T0.A}$ with trait expressions $r_{T0.A'}$, or $r_{T0.A}$ with trait expression $r_{T0.B}$. Differences of trait expressions were small and mostly not significant (with three exceptions: Neuroticism and Conscientiousness at functionally equivalent situations, and Openness at functionally non-equivalent situations).

Simple consistency. Simple consistencies from Study 4 are displayed in the lower half of Table 7. The average simple consistency for functionally equivalent situations was $r_{A.A'} = .73$ ($SD = .14$), and $r_{A.B} = .46$ ($SD = .17$) for functionally non-equivalent situations. Across all domains, simple consistency for functionally non-equivalent situations was again lower compared to functionally equivalent situations. We also contrasted the discriminant associations of state scores with the convergent associations of state scores (see in Table OSM B the right half of the upper panel under "For State Scores"). The lower triangle shows the correlations of state scores assessed at functionally equivalent situations, and the upper triangle shows correlations of state scores assessed at functionally non-equivalent situations. The average *absolute* discriminant correlation (off-diagonal) of state scores was .47 ($SD = .28$) for functionally equivalent situations, and .35 ($SD = .24$) for functionally non-equivalent situations, compared to an average

Table 9. Summary of average simple and residual consistencies in Studies 1–4.

	Simple consistency		Residual consistency	
Study	FE	FNE	FE	FNE
1 & 2	.86	.68	.59	.35
3	.67	.35	.65	.33
4	.73	.45	.71	.42

FE: functionally equivalent situation; FNE: functionally non-equivalent situation.

Note. Study 1: $N = 158$; Study 2: $N = 115$; Study 3: $N = 196$; Study 4: $N = 202$.

Displayed are the average simple and residual consistencies across studies.

Note that while we investigated only functionally equivalent situations in Study 1 and functionally non-equivalent situations in Study 2, Studies 3 and 4 allowed examining both functionally equivalent and non-equivalent situations at the same time.

of .73 and .46 for convergent correlations, respectively. Similar to Study 1, we therefore conclude that state scores showed good convergent and sufficient discriminant validity.

Finally, we again examined the full correlation matrix of interindividual differences in simple consistency (see in Table 8 the right half of the upper panel, "For State Scores"). The average *absolute* correlation of individual consistency scores amounted to .17 ($SD = .11$) for functionally equivalent situations (lower triangle) and .15 ($SD = .10$) for functionally non-equivalent situations (upper triangle). Although some correlations were significant in both cases, a general pattern of individual consistency across domains could not be established.

Residual consistency. Residual consistencies of Study 4 are displayed in Table 7, lower half. The average residual consistency amounted to $r_{eA.eA'} = .71$ ($SD = .14$) for functionally equivalent situations and $r_{eA.eB'} = .42$ ($SD = .19$) for functionally non-equivalent situations. We further investigated the discriminant associations of residual state scores (in Table OSM B the right half of the lower panel under "For State Residuals"). The average discriminant correlation was $r_{eA.eA'} = .44$ ($SD = .26$) for functionally equivalent situations, and $r_{eA.eB} = .31$ ($SD = .22$) for functionally non-equivalent situations. Similar to simple consistencies, the residual state scores thus showed reasonable convergent and discriminant associations.

Lastly, we examined again the full intercorrelation matrix of interindividual residual consistency (see in Table 8 the right half of the lower panel under "For State Residuals"). The average *absolute* correlation of individual residual consistency scores was .16 ($SD = .11$) for functionally equivalent situations (lower triangle), and .15 ($SD = .10$) for functionally non-equivalent situations. Again, a consistent pattern of intraindividual consistency could not be detected.

Discussion

We conducted Study 4 to replicate our results from Study 3 in an independent sample. Although across different analyses and indices the effect sizes were somewhat larger, the general pattern of findings replicated well. First, trait expressions in Study 4 were comparable to those from Study 3, and again lower than those from Studies 1 and 2. This allows us to interpret the differences between Studies 1 and 2 vs. Studies 3 and 4 as systematic. Second, replicating previous results, simple and residual consistencies were lower in functionally non-equivalent compared to functionally equivalent situations. Third, residual consistencies were again only somewhat smaller than simple consistencies, regardless of the functional equivalence ($r_{A.A'} = .73$ vs. $r_{eA.eA'} = .71$, and $r_{A.B} = .45$ vs. $r_{eA'.eB} = .42$)—which was what we also found in Study 3. We therefore conclude that, if assessed in daily life using experience sampling, simple and residual consistencies do not differ as much.

Overall discussion

Across four different studies using two different methodological approaches and samples from two different language areas, we showed that state scores consist of systematic trait variance as well as systematic situation variance. The newly introduced concept of residual consistency represents a valuable extension to the concept of simple consistency and also supports the notion of a simultaneous systematic influence of the situation and personality traits on personality states. Further, functional equivalence of situations is related to the level of simple and residual consistencies.

Simple and residual consistencies

Across all studies, trait expressions were comparatively high, though they were lower in Studies 3 and 4 compared to Studies 1 and 2. There are several possible explanations. First, whereas the same measure was used for traits and states in Studies 1 and 2, a different measure for traits and states was used in Studies 3 and 4. Second, given the single item measures that were used in Studies 3 and 4, the content of the assessed states is likely less homomorphous compared to the rather broad state measures used in Studies 1 and 2 (Horstmann & Rauthmann, in preparation; Horstmann & Ziegler, 2020; Rauthmann et al., 2019). Finally, the situations assessed in Studies 1 and 2 were much more specific and may even be referred to as "episodes" (i.e., longer situations), while the instructions during experience sampling asked participants what they were doing "right now". Although we averaged within persons three functionally similar situations in Studies 3 and 4, this may still explain lower trait-state-correlations in Studies 3 and 4 compared to Studies 1 and 2

(Horstmann & Rauthmann, in preparation). Similarly, both consistency estimates were lower in Studies 3 and 4 compared to Studies 1 and 2.

Concerning the difference of residual and simple consistency, residual consistency was always lower compared to simple consistency (which necessarily needs to be the case) across all studies (Table 9). Furthermore, estimates of consistency in functionally non-equivalent situations were even lower compared to functionally equivalent situations. Together, this pattern of findings suggests that state residuals—or "states freed from broad traits"—indeed contain reliable variance, and estimates of residual consistency of personality states can be traced back to the functional equivalence of two situations: The higher the functional equivalence of two situations, the higher the estimates of both simple and residual consistencies.

It is, however, unclear, why the differences of simple and residual consistencies in Studies 3 and 4 were only that small (see Table 9), both among functionally equivalent and functionally non-equivalent situations. As explained earlier, simple and residual consistencies must necessarily be the same if the trait expression is zero. However, in Studies 3 and 4, trait expressions were always positive and different from zero (see Table 7). This means that—at least in the case of experience sampling—estimates of consistency in self-reported states may have more to do with situations than broad personality traits. This is not meant to suggest that traits play no role or that situations are stronger than traits. Personality trait scores, especially when they represent broad traits as in our study, may be more suited to explain between-person variation, whereas momentary situation experiences can explain within-person variation. Further, it is possible that traits we have not measured could explain the consistencies, as we discuss later.

Further, residual consistencies under functionally non-equivalent situations were also not zero. There are a few possible explanations for why this could have occurred. First, the situations may simply not be completely non-equivalent. As outlined above, functional equivalence of situations should be conceptualized as continuous rather than dichotomous. Both Situations B and C contained, for example, social aspects (i.e., other people). Further, we selected situations in Studies 3 and 4 based on the reported situation characteristics (as assessed with the DIAMODNS). Within each participant, it is very likely that functionally non-equivalent situations were not completely different. For example, if all 50 situations of one participant were more or less similar, it would still only be the *least* similar that were selected as Situation B, which was *least* similar to Situations A and A′. However, this does not mean that the situations were completely different with respect to their situation characteristics.

Additionally (and very similar to the point mentioned above concerning trait expressions), it may be possible that the state measures may not be homomorphous, that is, the state measures could assess slightly different constructs than the trait measures (Horstmann & Ziegler, 2020; Rauthmann et al., 2019). Although the same measure was used in Studies 1 and 2 to assess traits and states, some items may not readily apply to states or apply in a different manner. However, if this variance was systematic and only present in the state measures at both measurement occasions, it would lead to an increased residual consistency.

Intercorrelations of individual differences in consistencies: No underlying general factor

Our findings allow new insights into the structure of consistency. For simple and residual forms, interindividual differences in consistency scores from one aspect or domain were in some instances correlated with scores from another aspect or domain (see Tables 4 and 8), especially in Studies 3 and 4. Although some correlations were quite substantial (e.g., up to .51 between conscientiousness and openness; see "For State Residuals" under Study 3 in Table 8), the covariation among individual consistency scores was in all cases too low and too unsystematic to warrant fitting a general factor model to it. Furthermore, although some correlations were quite substantial, they were not consistent across studies. Thus, we failed to find support for broad cross-aspect or cross-domain consistency in our analyses; rather, our findings point towards *aspect-* or *domain-specific* forms of individual differences in consistency.

Functional equivalence of situations

Our work may also be interpreted as a powerful and stringent test of the functional equivalence of situations as we have (a) explicitly manipulated the level of functional equivalence in two studies by manipulating situation cues and (b) selected situations based on their DIAMONDS profiles (i.e., perceived situation characteristics) and assumed that situations with similar DIAMONDS profiles are also functionally equivalent, while those that have dissimilar profiles are also less functionally equivalent. In the first case, however, to test for functional equivalence of situations, computing simple consistency estimates between two supposedly functionally equivalent situations will not be enough because they contain substantive amounts of trait variance (see, e.g., our trait expression estimates). Controlling for trait levels and using resulting state residuals provides a stronger test of situations' functional equivalence. In other words, the higher the functional equivalence of situations is, the higher estimates of residual consistencies need to be – and vice

versa. In Studies 1 and 2, we used situation vignettes that were, judged on face validity, pretty equivalent (Study 1) or non-equivalent (Study 2). Because we were interested in disentangling simple from residual consistency *given that* our situations were supposed to be functionally (non-)equivalent, we did not interpret our findings as evidence for the functional (non-) equivalence of our situations (which would be circular and then only a psychometric exercise of testing situation vignettes). Yet, after having conducted Study 2 and having explicitly manipulated the functional equivalence of the Situations B and C, it has not escaped our attention that the data-analytical procedures presented here can be used in other research as stringent tests for situational functional equivalences. Note, however, that the results of Study 2 show that a small correlation of residual state scores can be expected even if the situations are supposedly non-equivalent. Thus, functional equivalence of situations can either be assessed via their similarities in situation cues or their similarities in situation characteristics. Studies 3 and 4 allow us to conclude that functional equivalence may indeed be assessed directly by computing the similarity in DIAMONDS profiles of two situations. The higher the profiles, the more similar the situations are in (the *rank-orderings* of) their psychological characteristics—and vice versa. Our results from Studies 3 and 4 support this interpretation.

Residual consistency in personality theory

The most important finding from our studies is that the correlations of state residuals were substantial, and that the correlation was higher in functionally equivalent situations compared to functionally non-equivalent situations. What does this mean, and how does it relate to personality theories?

Residual consistency is a correlation of an interindividual difference variable, manifested in two situations. In Studies 1 and 2, the two situations across which consistency was computed were assumed to be fixed (or constants), that is, they were assumed to be the same for all participants, and treated as such in Studies 3 and 4. Given that the situations did not vary between people, they cannot have caused between-person differences; yet, stable interindividual differences were present, even after controlling for self-reported personality traits—clearly because self-reports of the Big Five personality traits do not capture all consistent variance of personality states across situations. Note that it is still correct to say that estimates of residual consistency were influenced by the functional equivalence of the situation as functional equivalence of situations varied between studies. The level of functional equivalence (across studies) is related to the estimates of simple and residual consistencies displayed at the between-person level. However, we do not know—and can only speculate—what has caused a person to report states in one situation that are similar to their self-reported states in another situation, independent of the trait level of that person. Clearly, it needs to be a stable interindividual difference that has led to the more similar self-reported states in functionally equivalent compared to functionally non-equivalent situations, and this stable interindividual difference is not represented in the self-reported personality trait that was controlled for.

Measurement of traits and states. The most parsimonious explanation for the fact that not all consistent variance could be explained by the self-reported personality traits (that we controlled for) may be the measurement approach taken. In all studies, we measured personality traits by using comparatively broad measures of the Big Five or Big Six traits. However, these were not able to sufficiently capture the consistent variance in self-reported states that occurred in the two situations, thus leading to estimates of residual consistency that were substantially different from zero. Alternatively, it is possible that the appropriate personality traits have not been assessed to explain consistent variance in the particular states sampled.

Situation perception. Another possibility that could explain residual consistency would be a stable and reoccurring relation between the person and the situation. Among these may be people's tendencies of situation management (Rauthmann & Sherman, 2016a), that is, how they construe as well as shape (e.g., select, modify) their daily situations. For example, how people tend to perceive situations generally (Rauthmann & Sherman, 2019; Ziegler & Horstmann, 2015; Ziegler et al., 2019) may be associated with state consistencies. In this case, interindividual differences in situation perception could explain the stable interindividual differences that are so far not explained by the self-reports of global personality traits. In this case, a person's general tendency to perceive and interpret situations would constitute a new layer of personality. It has previously been shown that the general tendency of a person to perceive situations is indeed stable across time (Ziegler et al., 2019).

Self-reports of personality traits. A key distinction that has to be made is the difference between personality traits and self-reports thereof. The more unique variance is present both in the self-reported states and the self-reported traits, the higher the difference between simple and residual consistencies. So far, we have only considered systematic variance from the state score that is not shared with the trait scores, thereby focusing on states. The question could also be asked the other way around: Why is it that some consistent variance of the state scores is not presented in the self-reports of the trait scores? One could argue that self-reports of traits may reflect—among other things—an

aggregate of previously experienced states (Fleeson & Jayawickreme, 2015; Judge et al., 2014; Turner, 1978). It is an open question, though, *which* past experiences form the self-report of a person. If the states that form the self-reports are sampled from situations that are functionally equivalent to the situations for which state consistency is computed, state expression must be high, and the difference between simple and residual consistencies must also be high. If the states that form the self-report are functionally non-equivalent to the ones for which state consistency is computed, then trait expression must be low, and the difference between simple and residual consistencies should be smaller.

Note that if states are influenced by a reoccurring set of functionally equivalent situations (i.e., context), but the self-reported trait is not formed based on experiences from this context, then this has several consequences. First, it can help explain why contextualized trait measures, as they are oftentimes used in an occupational context (Klehe et al., 2012; Lievens, 2017; Rockstuhl et al., 2015), have a higher predictive validity than non-contextualized measures, as the consistent state variance is now present in the trait self-report, thereby leading to higher trait expression (i.e., predictive validity).

Second, the fact that self-reported traits cannot explain all consistent variance in self-reported states may also explain discrepancies in self- and observer rated personality scores (Funder, 1995; Rogers & Biesanz, 2019; Vazire, 2010). As the realistic accuracy model (Funder, 1995) suggests, a person may only be judged accurately if a potential judge has the option to observe behaviors of person, and uses these behaviors correctly to form inferences about a person's stable characteristics. However, if a person shows residual consistency—that is, stable patterns of behavior that are by definition not related to their personality trait—and if a judge observes these consistent patterns of behavior, then the judge may form an impression of the person's personality that is consequently unrelated to the personality trait.

Of course, it would now be possible to examine the consistency of state scores that have been ridded from both self-ratings as well as informant ratings, and maybe even informants that are present in the situations the states were assessed in. If estimates from residual consistency were still substantial, this could alert us to variance that is not accessible (or represented) in neither self-reports or informant reports.

The hierarchical structure of personality traits. Personality traits are often assumed to be hierarchically structured (Costa & McCrae, 1995; Mõttus et al., 2017; Soto & John, 2017): Nuances (specific thoughts, feelings, desires, and behaviors) nested in facets, which are nested in aspects, which are again nested in domains (DeYoung et al., 2007; McCrae, 2015; Mõttus et al., 2017). It is important to make explicit that domains, aspects, facets, and nuances are not interchangeable with traits and states. Admittedly, domains are most often conceptualized as broad, heterogenous personality traits (such as the Big Five), and it is difficult to imagine how an entire broad personality trait is supposed to manifest itself in one specific situation. It is much more likely that it is a facet or even simply a nuance of the trait that manifests as a personality state in a given situation. At the same time, however, nuances and facets can both be conceptualized as narrow, yet stable personality traits. For example, the nuance "tendency to dance" is clearly a part of the broader trait extraversion. Dancing in a specific situation would then be the manifestation of that nuance, and thereby a state. Thus, the hierarchical structuring pertains to the breadth of the content covered, and trait or state variance may be present in different proportions in domains, aspects, facets, and nuances depending on how they are measured.

Nuances are theoretically similar to residual consistency as they both consist of systematic variance that is not represented in broad trait scores. Importantly, nuances were conceptualized as traits independent of a situation, whereas residual consistency is an indicator for systematic variance which is related to a situation. This systematic variance could then be understood as a *situation-specific trait*. The existence of such situation-specific traits has previously been suggested and examined by Geiser et al. (2015).[13] Geiser and colleagues extended Latent-State-Trait (LST) theory (Steyer et al., 1999, 2015). In LST, situations are assumed to be sampled at random (i.e., in our terms, it is assumed that they have the same level of functional equivalence). In this case, traits are then assumed to generalize "perfectly across situations" (Geiser et al., 2015, p. 169). Further, effects of the situation cannot be disentangled from effects of person-situation-interactions. The extension proposed by Geiser et al. (2015) considers that some characteristics, such as mood, are characterized by "a sizable level of occasion-specificity (situation-dependence)" (p. 179). The authors then examined if occasion-specific variance is consistent within contexts (i.e., within the same random set of fixed situations, e.g., in "alone situations") and across contexts (i.e., across two fixed situations, e.g., across "alone situations" and "not-alone situations"). Geiser and colleagues could show that, for example, some aspects of mood (e.g., feeling happy or cheerful) show occasion-specific variance, that is, reliable or unique variance that is consistently expressed within the same context. As our research shows, the variance of residualized state scores correlates across two assessment situations, and this correlation is influenced by the functional equivalence of the situation. Situation-specific traits must therefore be understood as consistent patterns of thoughts, feelings, desires, and behaviors that are consistently

expressed in specific situations and similar in similar situations (or, in Geiser and colleagues' words, within the same context).

Limitations and future directions

The limitations to this study point towards future research that would ideally replicate, corroborate, and also extend our findings. First, we aimed at sampling 250 participants in Study 1, but only 158 participants completed it, and 115 completed Study 2. Due to the within-person design and the clear patterns established here and in the simulations (see supplemental materials at osf.io/xfhdu/), findings are unlikely to change using a larger sample size with the data-analytical plan used.

Second, we only used one situation that was encapsulated in two vignettes in Study 1, and two different situations in Study 2. Future research may seek to employ different kinds of situations (i.e., variation in situation content) and examine to what extent our findings can then be replicated and generalized. Notably, it will be interesting to estimate to what extent different combinations of different person populations (e.g., pure community sample) and situation vignette contents would lead to similar or different patterns of findings. Examining such constraints on generality will be an important endeavor (Simons et al., 2017).

Third, future studies should employ actual situations, likely standardized within a laboratory setting (because at least two functionally equivalent situations need to be administered to each person to replicate Study 1), and measure states not just via self- or experience sampling reports, but also via behavioral observation or peer-ratings of actual behavior shown in these situations. Likewise, the measure of the trait (i.e., global explicit self-reports) may be inappropriate to capture all aspects of the trait that could explain consistent variance across situations.

Fourth, we only examined consistency of states across a relatively short period of time. Although states were more or less enacted consistently, depending on the level of functional equivalence, we cannot say anything about their long-term temporal stability. However, personality traits are also considered to be stable across longer time spans, not only consistent across situations. Although the situations in all studies were at least a few hours apart, this does not speak to the stability of states across situations. Future studies will have to examine how longer intervals influences the estimates of residual and simple consistency. For example, it could be possible that residual consistency decreases with longer time intervals, which would alert us to the fact that influences on states were at play that are only of short to medium duration (e.g., prolonged affect, which could have been persistent across all assessment situations in all Studies 1–4).

Lastly, future studies should examine the role of individual differences in situation perceptions in individual differences in consistencies. Note that while we have used situation perceptions in Studies 3 and 4 *as indicators of functional equivalence*, it is trivial to examine how similar the situations in these studies are on their perceived situation characteristics. On the other hand, our results already exemplify that similarity of situations on their psychological situation characteristics has a similar effect on simple and residual consistencies than the vignettes that were designed as functionally more or less equivalent. Thus, similarity of situation perceptions (or, in other words, their consistency) may yield behavioral consistency—a concept that has been explored before (Sherman et al., 2010).

Conclusion

We have examined several research questions for simple and residual consistencies under different circumstances (Table 1). Under functionally equivalent situations, we found that all forms of consistency were relatively high, and we were especially interested in residual forms of consistency: enacting certain states consistently in similarly functionally equivalent situations independent of one's trait levels. Examining functionally non-equivalent situations, we corroborated our interpretation, showing that functionally non-equivalent situations lead to lower simple and residual consistencies. Not all consistent personality states are reflected in the personality trait, and the search for additional predictors for consistent states above and beyond broad personality traits has just begun.

Data accessibility statement

The study materials, data, and analysis scripts used for this article can be accessed online at https://osf.io/xfhdu/.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Authors' note

Study 1 of this article was pre-registered on the Open Science Framework (see https://osf.io/8u7ka/). Additional materials that allow a complete reproduction of the analyses and a replication of the study can be found at https://osf.io/nueq6/. In line with the 21-word solution (Simmons et al., 2012), we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. Materials and data for the study can be found

at https://osf.io/xfhdu. For an overview of tables and supplementary materials, please refer to https://osf.io/a9r74/.

Acknowledgments

We thank Katja Witte and Julia Kim Walther for their support with data collection.

ORCID iD

Kai T Horstmann https://orcid.org/0000-0003-3224-1880

Notes

1. As the order of the vignettes was randomly assigned, we do not refer to the measurement points as T_1 and T_2, respectively. We simply refer to A and A′.
2. It is difficult creating a vignette with a real-world, familiar situation that activates all Big Five domains at once.
3. Table OSM A and Table OSM B can be found at https://osf.io/pdu23/
4. This study was suggested by two anonymous reviewers who had commented on an earlier version of this article that contained only Study 1.
5. This unfortunately means that we cannot rule out the possibility that some participants took part in both studies.
6. Although we could have used one of the two situation vignettes from Study 1 (A or A′), we did not do so as it was possible that the same participants from Study 1 would also participate in Study 2.
7. See Table 3 for statistical differences between simple and residual consistencies between Studies 1 and 2.
8. Note that these are translations of the German adjectives used. Furthermore, two adjectives (instead of one) were used for each pole of the rating scale; all items can be found online at the OSF website of Horstmann et al. (in press), osf.io/zctv4/.
9. Given that A and A′ are selected to be similar and that trait scores across analyses would not change, additional analyses (i.e., A′ vs. B) would yield highly dependent results, making their interpretation very difficult.
10. We initially computed these analyses with only one situation, that is, without averaging across three situations. However, we were concerned that these estimates from single items were not sufficiently reliable and therefore averaged states across three similar situations. The results did not deviate substantially (see osf.io/xfhdu/ under results, analyses with one item).
11. Facet-level analyses were not possible because states in the experience sampling portion of the study were only assessed as expressions of domains.
12. For mean-level changes of situation perception scores, see supplementary material on mean level change.
13. This is similar to the idea of narrow traits suggested by Ones and Viswesvaran (1996).

References

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, *52*, 376–387. https://doi.org/10.3758/s13428-019-01236-y

Asendorpf, J. B. (1990). The measurement of individual consistency. *Methodika*, *4*, 1–22.

Ashton, M. C., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*, 340–345. https://doi.org/10.1080/00223890902935878

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting actual behavior from the explicit and implicit self-concept of personality. *Journal of Personality and Social Psychology*, *97*, 533–548. https://doi.org/10.1037/a0016229

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, William Fleeson, J. J. A., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., & 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

Bem, D., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review*, *81*, 506–520. https://doi.org/10.1037/h0037130

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, *86*, 599–614. https://doi.org/10.1037/0022-3514.86.4.599

Chaplin, W. F., & Goldberg, L. R. (1984). A failure to replicate the Bem and Allen study of individual differences in cross-situational consistency. *Journal of Personality and Social Psychology*, *47*, 1074–1090. https://doi.org/10.1037/0022-3514.47.5.1074

Costa, P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*, 21–50. https://doi.org/10.1207/s15327752jpa6401_2

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

Emmons, R. A., & Diener, E. (1986). Situation selection as a moderator of response consistency and stability. *Journal of Personality and Social Psychology*, *51*, 1013–1019. https://doi.org/10.1037/0022-3514.51.5.1013

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., & 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

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

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

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

Geiser, C., Litson, K., Bishop, J., Keller, B. T., Burns, G. L., Servera, M., & Shiffman, S. (2015). Analyzing person, situation and person × situation interaction effects: Latent state-trait models for the combination of random and fixed situations. *Psychological Methods*, *20*, 165–192. https://doi.org/10.1037/met0000026

Gurtman, M. B. (2009). Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass*, *3*, 601–619. https://doi.org/10.1111/j.1751-9004.2009.00172.x

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

Horstmann, K. T., Rauthmann, J. F., Sherman, R. A., & Ziegler, M. (in press). 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. (2020). Assessment of situational perceptions. In J. F. Rauthmann, R. A. Sherman & D. C. Funder (Eds), *The Oxford handbook of psychological situations* (Vol. *1*, pp. 342–359). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190263348.013.21

Horstmann, K. T., & Ziegler, M. (2019). Situational perception and affect: Barking up the wrong tree? *Personality and Individual Differences*, *136*, 132–139. https://doi.org/10.1016/j.paid.2018.01.020

Horstmann, K. T., & Ziegler, M. (2020). Assessing Personality States: What to Consider when Constructing Personality State Measures. *European Journal of Personality, 34*, 1037–1059. https://doi.org/10.1002/per.2266

Ickes, W., Snyder, M., & Garcia, S. (1997). Personality influences on the choice of situations. In *Handbook of personality psychology* (pp. 165–195). Elsevier. https://doi.org/10.1016/B978-012134645-4/50008-1

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

Judge, T. A., Simon, L. S., Hurst, C., & Kelley, K. (2014). What I experienced yesterday is who I am today: Relationship of work motivations and behaviors to within-individual variation in the five-factor model of personality. *Journal of Applied Psychology*, *99*, 199–221. https://doi.org/10.1037/a0034485

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance*, *25*, 273–302. https://doi.org/10.1080/08959285.2012.703733

Leikas, S., Lönnqvist, J.-E., & Verkasalo, M. (2012). Persons, situations, and behaviors: Consistency and variability of different behaviors in four interpersonal situations. *Journal of Personality and Social Psychology*, *103*, 1007–1022. https://doi.org/10.1037/a0030385

Lievens, F. (2017). Assessing personality-situation interplay in personnel selection: Toward more integration into personality research. *European Journal of Personality*, *31*, 424–440. https://doi.org/10.1002/per.2111

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McCrae, R. R. (2015). A more nuanced view of reliability. *Personality and Social Psychology Review*, *19*, 97–112. https://doi.org/10.1177/1088868314541857

Morse, P. J., Sauerberger, K. S., Todd, E., & Funder, D. (2015). Relationships among personality, situational construal and social outcomes. *European Journal of Personality*, *29*, 97–106. https://doi.org/10.1002/per.1987

Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars. *Diagnostica*, *60*, 86–97. https://doi.org/10.1026/0012-1924/a000112

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*, 474–490. https://doi.org/10.1037/pspp0000100

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior*, *17*, 609–626. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<609::AID-JOB1828>3.0.CO;2-K

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

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

Rauthmann, J. F., Jones, A. B., & Sherman, R. A. (2016). Directionality of person-situation transactions: Are there spillovers among and between situation experiences and personality states? *Personality and Social Psychology Bulletin*, *42*, 893–909. https://doi.org/10.1177/0146167216647360

Rauthmann, J. F., & Sherman, R. (2019). Toward a research agenda for the study of situation perceptions: A variance componential framework. *Personality and*

Social Psychology Review, *23*, 238–266. https://doi.org/10.1177/1088868318765600

Rauthmann, J. F., & Sherman, R. A. (2016a). Situation change: Stability and change of situation variables between and within persons. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01938

Rauthmann, J. F., & Sherman, R. A. (2016b). Ultra-brief measures for the situational eight DIAMONDS domains. *European Journal of Psychological Assessment*, *32*, 165–174. https://doi.org/10.1027/1015-5759/a000245

Rauthmann, J. F., Sherman, R. A., & Funder, D. C. (2015). Principles of situation research: Towards a better understanding of psychological situations. *European Journal of Personality*, *29*, 363–381. https://doi.org/10.1002/per.1994

Rauthmann, J. F., Sherman, R. A., Nave, C. S., & Funder, D. C. (2015). Personality-driven situation experience, contact, and construal: How people's personality traits predict characteristics of their situations in daily life. *Journal of Research in Personality*, *55*, 98–111. https://doi.org/10.1016/j.jrp.2015.02.003

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research. Retrieved from https://cran.r-project.org/package = psych

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

Riemann, R., Angleitner, A., Borkenau, P., & Eid, M. (1998). Genetic and environmental sources of consistency and variability in positive and negative mood. *European Journal of Personality*, *12*, 345–364. https://doi.org/10.1002/(SICI)1099-0984(1998090)12:5<345::AID-PER332>3.0.CO;2-Y

Rockstuhl, T., Ang, S., Ng, K.-Y., Lievens, F., & Van Dyne, L. (2015). Putting judging situations into situational judgment tests: Evidence from intercultural multimedia SJTs. *Journal of Applied Psychology*, *100*, 464–480. https://doi.org/10.1037/a0038098

Rogers, K. H., & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*, *117*, 186–200. https://doi.org/10.1037/pspp0000197

Schmitt, M. (1990a). Further evidence on the invalidity of self-reported consistency. In P. J. D. Drenth, J. A. Sergeant, & R. J. Takens (Eds), *European perspectives in psychology* (Vol. I, pp. 57–68). New York: John Wiley & Sons, Inc.

Schmitt, M. (1990b). *Konsistenz als Persönlichkeitseigenschaft? Moderatorvariablen in der Persönlichkeits- und Einstellungsforschung*. Springer.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Sherman, R. A., Nave, C. S., & Funder, D. C. (2010). Situational similarity and personality predict behavioral consistency. *Journal of Personality and Social Psychology*, *99*, 330–343. https://doi.org/10.1037/a0019796

Sherman, R. A., & Pashler, H. (submitted). Powerful moderator variables in behavioral science? Don't bet on them. Retrieved from 10.31234/osf.io/c65wm

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. https://doi.org/10.1177/1745691617708630

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

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits – Revised. *Annual Review of Clinical Psychology*, *11*, 71–98. https://doi.org/10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, *13*, 389–408. http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-0984(199909/10)13:5%3C389::AID-PER361%3E3.0.CO;2-A/abstract

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

Turner, R. G. (1978). Consistency, self-consciousness, and the predictive validity of typical and maximal personality measures. *Journal of Research in Personality*, *12*, 117–132.

Vazire, S. (2010). Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281–300. https://doi.org/10.1037/a0017908

Weisbuch, M., Slepian, M. L., Clarke, A., Ambady, N., & Veenstra-VanderWeele, J. (2010). Behavioral stability across time and situations: Nonverbal versus verbal consistency. *Journal of Nonverbal Behavior*, *34*, 43–56. https://doi.org/10.1007/s10919-009-0079-9

Wickham, H. (2016a). purrr: Functional programming tools. Retrieved from https://cran.r-project.org/package = purrr

Wickham, H. (2016b). stringr: Simple, consistent wrappers for common string operations. Retrieved from https://cran.r-project.org/package = stringr

Wickham, H., & Francois, R. (2016). dplyr: A grammar of data manipulation. Retrieved from https://cran.r-project.org/package = dplyr

Ziegler, M., & Horstmann, K. T. (2015). Discovering the second side of the coin. *European Journal of Psychological Assessment*, *31*, 69–74. https://doi.org/10.1027/1015-5759/a000258

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