Descriptive, predictive and explanatory personality research

Forthcoming in *European Journal of Personality*

Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits

RENÉ MÕTTUS[1,2*], DUSTIN WOOD[3], DAVID M. CONDON[4], MITJA BACK[5], ANNA BAUMERT[6], GIULIO COSTANTINI[7], SACHA EPSKAMP[8], SAMUEL GREIFF[9], WENDY JOHNSON[1], AARON LUKASZEWSKI[10], AJA MURRAY[1], WILLIAM REVELLE[11], AIDAN WRIGHT[12], TAL YARKONI[13], MATTHIAS ZIEGLER[14], and JOHANNES ZIMMERMANN[15]

[1]*University of Edinburgh, UK*

[2]*University of Tartu, Estonia*

[3]*University of Alabama, USA*

[4]*University of Oregon, USA*

[5]*University of Münster, Germany*

[6]*Max Planck Institute for Research on Collective Goods, Bonn, and TUM School of Education, Germany*

[7]*University of Milan-Bicocca, Italy*

[8]*University of Amsterdam, Netherlands*

[9]*University of Luxembourg, Luxembourg*

[10]*California State University, Fullerton, USA*

[11]*Northwestern University, USA*

[12]*University of Pittsburgh, USA*

[13]*University of Texas at Austin, USA*

[14]*Humboldt Universität zu Berlin, Germany*

[15]*University of Kassel, Germany*

Abstract: We argue that it is useful to distinguish between three key goals of personality science – description, prediction and explanation – and that attaining them often requires different priorities and methodological approaches. We put forward specific recommendations such as publishing findings with minimum a priori aggregation and exploring the limits of predictive models without being constrained by parsimony and intuitiveness but instead maximising out-of-sample predictive accuracy. We argue that naturally-occurring variance in many decontextualized and multi-determined constructs that interest personality scientists may not have individual causes, at least as this term is generally understood and in ways that are human-interpretable, never mind intervenable. If so, useful explanations are narratives that summarize many pieces of descriptive findings rather than models that target individual cause-effect associations. By meticulously studying specific and contextualized behaviours, thoughts, feelings and goals, however, individual causes of variance may ultimately be identifiable, although such causal explanations will likely be far more complex, phenomenon-specific and person-specific than anticipated thus far. Progress in all three areas – description, prediction, and explanation – requires higher-dimensional models than the currently-dominant "Big Few" and supplementing subjective trait-ratings with alternative sources of information such as informant-reports and behavioural measurements. Developing a new generation of psychometric tools thus provides many immediate research opportunities.

Keywords: prediction; explanation; cause; hierarchy; personality

*Correspondence to: René Mõttus, 7 George Square EH8 9JZ Edinburgh, Scotland; rene.mottus@ed.ac.uk

This manuscript is based on an Expert Meeting jointly supported by European Association of Personality Psychology and European Association of Psychological Assessment, and held from 6th to 8th September 2018 in Edinburgh, Scotland (https://osf.io/fn5pw). Authors are grateful to Tom Booth, Jaime Derringer, Ryne Sherman and David Stillwell for their contributions to the Expert Meeting, and to Samuel Henry for his comments on the manuscript. Not all authors agree with all arguments put forward in this paper.

Personality psychology has come a long way in describing how people differ in thinking, feeling, behaving, and wanting. This has been facilitated by agreement among researchers on a limited number of broad personality dimensions, organizing research and allowing observations to accumulate. The largely overlapping Big Five (Goldberg, 1990), Five-Factor Model (FFM; McCrae & John, 1992), and HEXACO domains (Ashton & Lee, 2020) have been particularly instrumental broad personality constructs, so much so that they have become the default way of operationalizing personality differences among people; we refer to them as the Big Few.

Yet it is not evident that the Big Few "carve nature at its joints". They are useful for conveniently *summarizing* a variety of ways in which people can differ with a manageable number of dimensions. But there is little evidence that they are particularly good units for *explaining* behaviour or psychological processes underlying it (Baumeister et al., 2007; Wood et al., 2015; Jonas & Markon, 2016) or even that they are the best predictors of real-world outcomes (Mõttus & Seeboth, 2018; Elleman et al., 2020). The Big Few were formed by combining subjective perceptions of traits[1] that statistically co-vary among people rather than based on models of processes that happen in individuals. Currently we do not know of many genetic variants, neurobiological systems, experiences, or developmental processes that specifically contribute to variance in the certain Big Few domains such as Extraversion or Conscientiousness and set them apart from other domains such as Openness and Honesty-Humility or from traits allegedly beyond the Big Few such as motives, beliefs, or abilities. Moreover, the domains partly overlap and can be combined into even broader ones (DeYoung, 2006), but also broken into numerous more specific traits (McCrae & Sutin, 2018).

None of this is necessarily a problem. But it means that the variance found in typical personality measures can be described as a hierarchy of traits, that there are few reasons to automatically prefer any one of its levels over others, and that the mechanisms of the variance can be highly multiply determined. Researchers are also increasingly considering processes and related variance within individuals, besides differences between individuals; it is a crucial question how these variation levels are connected or whether they can be addressed with the same statistical and/or theoretical models at all. Likewise, there may be personality variance both between and within individuals (e.g., behavioural frequencies or relationship dynamics) that is not captured in the subjective perceptions commonly used for personality assessment.

As a result, particular models of personality may work better for some purposes than others. This leads to a central idea of this special issue generally and this article specifically: as our knowledge of personality grows and research questions become increasingly diverse, it may no longer be optimal for researchers to coalesce around a single or even a few ways of operationalizing personality (e.g., the Big Few). We distinguish between three broad aims of personality research – *description*, *prediction*, and *explanation* – and argue that these aims may entail disparate and sometimes even opposing research strategies. We advocate for the explicit articulation of these aims when designing, conducting, and reporting the results of personality research rather than defaulting to research practices that are widely used but may in fact be suboptimal for any given research project. For example, we propose that:

- Descriptive findings should be published in as much detail as possible (e.g., at the individual item level) besides being organized (e.g., according to attributes such as the strength of relations or the psychological modalities of the characteristics involved) or aggregated into broader constructs such as the Big Few. This offers more flexibility than the common practice of *a priori* aggregating findings for simplicity.
- Although traits' predictive validity is often seen as a major reason for doing personality research in the first place, its robustness and ways of maximising it remain under-explored. Availability of large datasets and advanced statistical tools are beginning to improve this. Predictive models should always be independently cross-validated and should not depend on parsimony or consistency with researchers' theoretical intuitions.
- Many phenomena that interest personality scientists such as broad patterns of naturally occurring individual differences (e.g., constructs in the personality trait hierarchy) may not have individually tractable *causes*, at least as this term is typically understood and/or in ways that are meaningfully interpretable and allow for targeted interventions. This is because the phenomena are inherently decontextualized and relative, and their indistinguishable levels can arise through many combinations of processes and may not result from unidirectional cause-effect associations, among other reasons. When this applies, useful explanations may be narratives that integrate many pieces of descriptive findings into broad principles rather than attempts to identify individual and potentially intervenable cause-effect associations. If so, for example, individual regression coefficients provide poor causal explanations. However, by defocusing from broader variability patterns and meticulously studying specific and contextualized behaviours, thoughts, feelings and goals, individual causes of variance may ultimately be identifiable in useful and potentially even controllable ways. Still, such causal explanations may be more complex, phenomenon-specific and person-specific than anticipated thus far.
- Progress in all three areas – description, prediction, and explanation – will likely require availability of far higher-dimensional models based on traits much more

1 Here, we define traits similarly to Baumert and colleagues (2017): trait is a descriptive dimension of any kind of relatively stable psychological and behavioural differences between people, independent of its content and breadth.

specific than the Big Few, as well as supplementing typical subjective trait-ratings with alternative sources of information such as informant-reports and behavioural measurements (Rauthmann, 2020). Therefore, an area with immediate and immense opportunities is developing a new generation of psychometric tools that allow sampling *persome* – the universe of variables capturing personality variability – more broadly than currently available measures do.

Descriptive personality science

Descriptive personality research explores associations between the measurements of personality constructs and/or their links with phenomena allegedly beyond the personality domain (e.g., demographic characteristics, experiences, and behavioural outcomes). The results can and do contribute to explanatory or predictive research, but they are also important in their own right and should not be constrained by theoretical models (purview of explanatory research) or attempts to maximise prediction (aim of predictive research).

For example, there is ample evidence that individual differences in personality characteristics can be clustered into replicable groups such as the Big Few (Schmitt et al., 2007), are relatively stable over several years (Terracciano et al., 2006), are persistently correlated with a variety of life outcomes (Roberts et al., 2007; Soto, 2019), and perceived at least somewhat similarly by different observers (Connelly & Ones, 2010). Genetically related individuals resemble each other in personality characteristics, accounting for most of the similarity of family members (Briley & Tucker-Drob, 2014), although the specific genetic variants correlated with the characteristics have remained elusive (Lo et al., 2017). Changes in personality characteristics barely track with specific life experiences (Bleidorn et al., 2020; Denissen et al., 2019), are similarly distributed across geographically diverse regions (Allik et al., 2017), but vary systematically across genders (Lee et al., 2020). Recently, research has also started to describe systematic patterns of short-term variations in personality as another aspect of individual differences (e.g., Danvers et al., 2020; Horstmann et al., 2020, Sosnowska et al., 2020).[2]

The trait hierarchy

Within the descriptive kind, a lot of research has been carried out on the relations between (subjectively perceived) trait scores with the aim to reduce personality variation among people to as few broad trait dimensions as possible. Summarizing variance with a small number of traits has been a practical approach, both in terms of data collection and reporting. For example, accessing sufficient participant numbers and tabulating data can be burdensome, especially when each trait is measured with numerous items, and analyzing many-dimensional data and communicating findings that involve numerous statistical associations may seem overwhelming.

But these difficulties have recently become less relevant. Technological progress has made accessing participants and collecting data much easier, with sample sizes now routinely in the thousands (Gosling & Mason, 2015). Self-report scales have turned out to be more reliable than previously thought, with their many-dimensionality often mistaken for measurement error (e.g., because internal consistency systematically underestimates reliability; Cronbach & Shavelson, 2004; McCrae, 2015). This allows us to measure a broader selection of narrower traits with the same number of carefully selected items, because fewer conceptually interchangeable items are required for each trait (McCrae & Mõttus, 2019; Wood, Nye, & Saucier, 2010; Yarkoni, 2010). Improved computational power and accessible data analytic tools have eased working with many-dimensional data to efficiently summarize, communicate and compare association patterns (Costantini et al., 2015; Revelle, 2020; Ellemann et al., 2020; Stachl et al., 2020).

Many researchers now agree that population-level personality variation is best represented as a hierarchy of increasingly specific traits, with no level uniquely representing nature carved at its joints (DeYoung, 2015; Markon et al., 2005; McCrae & Sutin, 2018). This hierarchy arises because most Big Few traits inter-correlate, suggesting few very general super-traits such as Stability and Plasticity (DeYoung, 2006), although methodological artifacts may contribute to this (Bäckström, Björklund, & Larsson, 2009; Riemann & Kandler, 2010). The Big Few domains also break down into constituents that develop and correlate with life outcomes in distinct ways (Jang et al., 1998; Paunonen & Ashton, 2001). Some models have therefore delineated "aspects" (DeYoung et al., 2007) or "facets" (Costa & McCrae, 1992) for the Big Few. These are more than just different ways in which the Big Few can be expressed (Jang et al., 1998): moreover, the hypothesis that some traits such as the Big Few are more "core" or "temperamental" than other, ostensibly more "surface" traits such as facets, has found limited empirical support (Kandler, Zimmermann, & McAdams, 2014). However, there has been little systematic research yet to delineate an empirically based and comprehensive model of personality facets for researchers to coalesce around (Saucier & Iurino, in press).

Moreover, most personality questionnaire items contain unique personality variance beyond the Big Few domains, aspects and facets they were designed to measure. Therefore, even the most comprehensive of the current facet models (e.g., Costa & McCrae, 1992) can be broken further down into numerous yet more specific traits, or "nuances". Empirically, nuances are every bit as trait-like as the Big Few domains or their aspects and facets, because even the unique variance in hundreds of items, reflecting the nuances but not facets, aspects and domains, has essential trait properties of stability over many years, transcendence across assessment method such as self- and informant-reports, and

2 These are all findings of descriptive research, even though correlations with life outcomes are sometimes called prediction and the correlation between genetic and phenotypic similarities are sometimes taken as the former explaining the latter.

heritability; item-specific variances also have distinct developmental trends and associations with life outcomes (Mõttus et al., 2017; Mõttus, Sinick, et al., 2019). Of the error-free variance of a typical Big Five item, less than half has been estimated to pertain to the domains and their facets, leaving at least a half for nuances (McCrae, 2015). There are also personality traits that are either in the peripheries of the Big Five domains, as commonly defined, or beyond them (e.g., competitiveness, loyalty, jealousy, humour, sexuality, or others; Bouchard, 2016; Paunonen & Jackson, 2000). These traits are often not well covered in currently popular personality measures. The true ubiquity and utility of nuance-like narrow personality traits is thus yet to be properly estimated, as available evidence is based on questionnaires carefully developed to assess little but the Big Few and their selected facets. This universe of narrow personality traits that forms the basis of the personality hierarchy has also been referred to as the *persome* (Mõttus, Bates et al., 2017; Revelle, Dworak, & Condon, 2017).

In principle, therefore, there are many ways for researchers to describe personality variation such as using different levels of the trait hierarchy. In practice, they often default to the Big Few, likely because these trait models appear intuitive and familiar, are already widely used and can be readily measured with existing instruments. Social pressure from peers, reviewers and editors may also play a role. Although these are legitimate practical reasons, there is no inherent scientific reason why this level of the trait hierarchy should be *a priori* and always preferred over others for each and every research purpose. In fact, this may often be counter-productive, in constraining research choices and inspiring potentially misleading generalizations.

What makes good descriptive research?

To select an appropriate way of representing personality variance for a descriptive research question, it helps to outline criteria for what would be a good descriptive account of whatever is being described in relation to particular personality constructs (e.g., other psychological constructs, different measurements of the same constructs, demographic variables or life outcomes). We illustrate this with how personality varies with age.

Information should be elaborate. Is a good descriptive account simple and parsimonious or comprehensive and detailed? The tension between these priorities can be alleviated by recognizing that *parsimonious accounts can always be extracted from detailed ones containing more numerous and less aggregated variables*. The reverse, however, is not possible (Saucier & Iurino, in press). With remarkable flexibility, many-dimensional findings can be subsequently zoomed into or summarized with fewer dimensions, such as for ease of interpretation and communication.

Being able to zoom in rather than *a priori* aggregating can pay off. For example, age differences in personality are often described using the Big Few traits, showing that older adults tend to be somewhat higher in Emotional Stability, Agreeableness, and Conscientiousness but slightly lower in Extraversion and Openness than younger adults. But some of the findings are specific to questionnaires (Costa et al., 2019), hinting that age differences are at least in part driven by narrower traits that are sampled in different proportions across instruments, and thereby potentially misrepresented by the broad trait domains. There is indeed ample evidence that facets of the same Big Few traits vary in their age differences (Terracciano et al., 2005; Jackson et al., 2009; Lucas & Donnellan, 2009). But even facets may not provide a full understanding, because items of the same facets—reflecting nuances within them—often vary in their age trends, conveying unique developmental information. For example, item-level analysis of the Assertiveness facet of the revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) showed that older people were more likely to take charge of situations but less likely to make others do things, and items of the Achievement Striving facets referring to hard work tended to increase with age while items referring to success-motivation trended downwards (Mõttus & Rozgonjuk, 2019). Such examples abound (Mõttus et al., 2015); for example, Mõttus and Rozgonjuk (2019) reported that items within half of the personality facets varied in the directions of their age differences, leading items to contain over 40% more age-sensitive information than facets and over twice as much as the Big Five domains. More nuanced investigations into how personality is linked with various life outcomes or vary across cultures have led to similar conclusions (Achaa-Amankwaa, Olaru, & Schroeders, 2020; Elleman et al., 2020; Seeboth & Mõttus, 2018; Wessels, Zimmermann, & Leising, 2020).

At which level of a personality hierarchy should descriptive findings stop? The answer will depend on the research questions under consideration, but the goal should be to represent descriptive findings such as age or gender differences or links between personality characteristics and other variables *at the level from which going more detailed would not add further useful information*. Technically, this means the level where the measurable constituents of the traits relate to the other variables alike, because traits' associations should not depend on which indicators are used to operationalize them (Mõttus, 2016; Spearman, 1927; Gonzales, MacKinnon, & Muniz, 2020). Often this may mean levels from which we simply cannot go any more detailed, such as individual test items, given that personality is, and possibly will be for some time at least, most commonly assessed with questionnaires. On other occasions, broader traits such as the Big Few or their facets may turn out to be the most suitable levels of description, because their constituents follow the same association patterns. Following this simple principle makes choosing the appropriate level of the personality hierarchy a defensible empirical question rather than a matter of personal preference, peer pressure or editorial policy.

It is sometimes thought that theories should constrain research questions. For descriptive research (as well as for predictive, below), we argue the opposite: theory should be

used to expand rather than constrain the personality construct space and thereby descriptive findings. For example, theories of how personality may relate to the phenomenon of interest can be used to suggest items to our item pools to make them more comprehensive and sensitive to the topic at hand. If we only operationalize personality with the Big Few, we *a priori* exclude possibilities to uncover additional aspects of personality, and how they develop and co-vary with other phenomena.

But we can use theoretical models to help with *organizing* our findings (e.g., Bem & Funder, 1978). For example, Mõttus and Rozgonjuk (2019) described age differences in personality using 300 items (many reflecting unique personality nuances), but organized the associations according to the Big Five and their facets using a Manhattan plot (Revelle, 2020; Revelle, Dworak, & Condon, 2020). This allowed them to show the general organisation of age differences in personality (they were wide-spread across hundreds of items) and how they were distributed across particular Big Five domains and their facets (i.e., mean difference between domains and facets in age-trajectories), but also how the age differences deviated from the patterns expected under the Big Five model (i.e., items of the same domains/facets often substantially varied in age differences).

Or, item-level findings can be organized according to the degrees to which the items represent affect, behaviour, cognition, or desires/motivation (ABCDs; Wilt & Revelle, 2015). For instance, a mental health variable could be most strongly linked with affective items, regardless of which Big Few domain or facet they belong to; a physical health outcome may be mostly linked with behavioural items; and other outcomes may predominantly track with other types of items. For a few more examples, findings could be organized according to the extents to which items reflect universal traits as opposed to contextual adaptations (Henry & Mõttus, 2020), social desirability (Wessels, Zimmermann, Biesanz, & Leising, 2020; Leising et al., 2020), visibility (Funder & Dobroth, 1987), social maturity (Caspi & Roberts, 2001), pathology (Vachon et al., 2013; Bleidorn et al., 2019) or stability, cross-method agreement, and associations with other variables (Mõttus, Sinick, et al., 2019). This way, we can use theory to expand association maps to hundreds of variables and still extract intelligible information from these, especially when we use suitable (e.g., interactive) visualization tools. Large samples and cross-validations are vital, but this is no longer an insurmountable barrier in the current data-centric age.

Patterns in how personality differences relate to the variables of interest can also be explored atheoretically. For example, item- or facet-level associations can be organized in the descending order of effect size to highlight the strongest associations and find commonalities in them (e.g., Achaa-Amankwaa et al., 2020; Elleman et al., 2020; Bem & Funder, 1978; Block, Block, & Gjerde, 1986; Block, Gjerde, & Block, 1991). In some fields such as genetics, recent progress has almost entirely resulted from atheoretically scanning association patterns rather than imposing theoretical constraints on the findings (e.g., Nagel et al., 2018; Plomin & von Stumm, 2018) and there is no reason why following suit could not help personality scientists.

More detailed findings can be aggregated into any trait construct, either at the time when they are first published or in subsequent research. This flexibility is especially useful, because most items represent several traits at different levels of the trait hierarchy or even at the same level; think of the International Personality Item Pool as an example of how items are "recycled" to measure disparate constructs (Goldberg, 1999). For example, to estimate how a (latent) trait correlates with a criterion from the correlations of k items with this criterion, the item-criterion correlations can be multiplied by the traits' loadings on the items (which can be extracted from correlations among items) and the sum product divided by the sum of the squared factor loadings:[3]

$$r(Trait, Criterion) = \frac{\sum_{i=1}^{k} \left(r(X_i, Criterion) * r(X_i, Trait) \right)}{\sum_{i=i}^{k} \left(r(X_i, Trait)^2 \right)}$$

The same applies to facet-level findings, of course.

As a general rule for basic research, thus, comprehensive and detailed descriptions of personality-related phenomena are preferable to those that *a priori* impose parsimony. But this does not mean that each and every study should necessarily measure hundreds of constructs, nor that each paper reports many hundreds of associations. Instead, personality scientists should *collectively* (across studies) aim towards maximum comprehensiveness. This can be achieved if individual studies a) consider diverse constructs rather than focus all on the same trait model (e.g., a Big Few), thereby distributing the workload and pooling findings either in a directed co-ordination or spontaneously, and b) provide their findings at various levels of specificity and aggregation (including disaggregated, item-level findings). Also needed are accessible tools for integrating the findings of different studies (e.g., for meta-analysing findings for available constructs, collating and publicly depositing them). Individual research reports can then contribute to, and draw from, a central repository of descriptive findings. This is not the default *modus operandi* of current personality research although it is common in some other fields such as genetics and neuroscience.

Findings should not depend on methodologies. When we link something to personality constructs, we typically expect that the associations pertain to psychological characteristics that exist independently of how they happen to be assessed (Hilbig, Moshagen, Zettler, 2016; Mõttus, 2016; Thielmann & Hilbig, 2016). When conclusions reliably differ, say, as a function of which personality questionnaire was used for

3 If the combinations of items ought to represent summary-traits rather than shared variance-based latent traits, principal component loadings can be used instead of factor loadings.

assessing the construct (e.g., the associations of Openness and Extraversion with age or that of Neuroticism with Body Mass Index vary across studies; Costa et al., 2019; Vainik, Dagher et al., 2019), this points to the association being driven by narrower traits that are captured by differing degrees across measurement tools. This implies labelling issues (or "jingle fallacies"; Block, 1995; Larsen & Bong, 2016), whereby investigators mean different things when invoking the same scale or construct name. If so, these narrower traits should be isolated, because generalizing associations beyond them is misleading. Reporting item-level association in particular can help to reveal jingle as well as jangle fallacies.

Unless there are explicit reasons for the contrary, the associations should also generalize across assessment methods such as, most readily, self- and informant-reports (ideally, the aggregate ratings of multiple informants). Findings that self- and informant-reports are measurement invariant are consistent with this (e. g., Mõttus, Allik, et al., 2019). For some traits, self- and informant-ratings may in part measure different aspects of personality (Vazire, 2010; McAbee & Connelly, 2016), in which case discrepant findings may be expected, and even hint at what contributes to the observed associations in the first place. For example, associations between personality traits and age tend to be stronger in self- than informant-reports (Costa et al., 2019), possibly because people have clearer perceptions of their own changes than they do of changes in others, or because age differences in self-reports are inflated due to increasing socially desirable responding with age (Soubelet & Salthouse, 2011).

Researchers should explore the generality of associations across contexts and other potential moderators. We should routinely strive to replicate findings in multiple diverse cultures, clarifying the extents to which the observed associations characterise larger populations than our typical study participants (e.g., Henrich, Heine, & Norenzayan, 2010), or even *humans in general*. Some already have been. For example, age differences in personality are fairly robust across cultures (McCrae et al., 2005), even at the levels of facets and nuances (Mõttus, Sinick, et al., 2019). Other findings may vary *systematically* across context; in these cases, we should establish that the variabilities themselves are replicable and attempt to identify their sources (moderators). For example, the magnitudes (but not profiles across multiple traits) of gender differences vary systematically between cultures and we know how: gender differences are larger in more prosperous societies (Schmitt et al., 2008; Mac Giolla & Kajonius, 2019; Lee & Ashton, 2020). It has been reported that the timing of age trajectories may also systematically vary across cultures (Bleidorn et al., 2013), but these findings have not yet been successfully replicated (McCrae et al., in press).

One possible benefit of routine attempts to replicate findings across cultures is diversifying the range of researchers participating in personality research, including those from currently less represented regions and backgrounds.

Some recommendations for descriptive research

A new trait taxonomy and instruments for it. Besides the Big Few, we need a more encompassing trait taxonomy to be able to comprehensively describe associations of personality traits among themselves and with other phenomena, coupled with instruments for measuring these traits. In other words, we need to sample the persome more broadly than the available taxonomies allow for. This does not mean doing away with the Big Few, but developing a properly hierarchical model in which traits can be investigated at lower (nuance) levels as well as aggregated into increasingly broad traits, including the Big Few. It may also be that the Big Few models eventually require a revision to account for lower level traits that are informative but do not easily fit into the current Big Few models (Saucier & Iurino, in press). Likewise, many lower-level traits may belong to more than one of the Big Few.

Such models are not unrealistic, nor impractical. For example, careful item selection – such as avoiding items with low retest reliability and excessive redundancy (McCrae & Mõttus, 2019) – may allow measuring a usefully comprehensive pool of nuances with one or perhaps two items each. Remember: nuances are narrow, so no broad content sampling is required for them because measurement breadth comes from the pool of nuances collectively, not from items within individual nuances. If so, a say 100- or 200-item test can encompass around 100 nuances that can be aggregated into a few dozens of facets, and still fewer aspects and domains. Common psychometric concerns about the use of short scales can be addressed. For example, the typical retest reliability of single items of existing questionnaires over a one-week or two-week interval is around .65 (e.g., Mõttus, Sinick et al, 2019; Henry & Mõttus, 2020), even though these instruments have rarely been constructed with item-level reliability in mind. Therefore, after careful item selection the majority of them can have reliabilities well above .60, with the average plausibly at about .70.[4] This means that the retest reliability of most two-item scales can be notably higher, often above .80.[5]

Findings obtained with such multi-nuance tests can be interpreted at any one trait hierarchy level or at multiple levels at the same time, as appropriate for the goal at hand. For example, broad-trait associations can be qualified by which specific narrower traits drive them, in the likely case that the associations within the scale have meaningful heterogeneity. Importantly, the measurement of broader traits themselves could also improve as a result of their encompassing more lower-level traits because good measures of broad trait domains sample their content broadly. This is therefore a win-win scenario.

4 Retest-correlations over shorter testing intervals can be higher still (Lowman, Wood, Armstrong, Harms, & Watson, 2018) and may provide even more accurate reliability estimates.

5 For an example of creating a high-dimensional personality trait pool, see Saucier, Iurino, & Thalmeyer (2020).

Of course, several of the Big Few instruments already allow for the measurement of their facets, but few authors have provided comprehensive, empirical evidence-based facet taxonomies (but see MacCann et al., 2009; Roberts et al., 2005; Ziegler et al., 2019) and these facet models are by definition constrained to the Big Few that have been defined *a priori.* Little taxonomic research yet has simultaneously encompassed the Big Few, their aspects and facets as well as traits beyond them (Condon, 2018; McCrae & Costa, 1996), and there has been virtually no taxonomic research for nuances yet (but see Wood et al., 2010).

Being realistic, it may never be possible to devise the ultimate hierarchical model of personality variance that covers all narrow personality traits in the persome, as somehow carved out by nature. There may be too many of them, their boundaries are likely inherently as fuzzy as those of broader traits, and many might apply to only some individuals and thereby have limited variance across individuals. But it is almost certainly plausible to develop models that sample from among the universe of important traits far more comprehensively than the currently popular, Big Few-centric models do.

Additional sources of information. To validate findings based on self-reports and explore patterns that may not be accurately captured with self-reports, researchers should use alternative sources of information about personality variation, while also being mindful of the limitations of these.

For example, technological progress has provided new sources of information (Rauthmann, 2020). Several recent articles describe how personality and its associations with other variables can be assessed through objectively measured behaviour or digital traces of behaviour (e.g., Cooper et al., 2020; Wiernik et al., 2020; Hall & Matz, 2020; Stachl, Au et al., 2020). These approaches offer great potential for non-invasively collecting personality-related information about large numbers of people and possibly over extended periods of time, hence allowing measurement of short- and even longer-term changes in personality. But often these assessment methods have to be given personality-relevant interpretation in relation to subjectively rated personality traits before they become useful. For example, on its own, mobile phone sensor data do not have psychological meaning; they do once we know how they track with self-reported traits (Wiernik et al., 2020; Stachl, Pargent et al., 2020; Hall & Matz, 2020). As a result, these methods often approximate subjectively rated traits rather than provide entirely new information, and any issues with self-reports can spill over to their digital approximations (Tay et al., 2020). Currently, typical correlations between self-reported traits and their digital approximations are in the range from .30 to .40 (Tay et al., 2020; Stachl, Au et al., 2020), so the gap between them remains non-trivial. It may narrow as research progresses, though.

Likewise, many researchers may strive towards objective, laboratory measurements of personality traits such as asking people to persevere with tedious and boring tasks (e.g., Gniewosz, Ortner, & Scherndl, 2020) or keep their hand in cold water (e.g., Schmeichel & Vohs, 2009) to measure their self-control, or asking them to categorize adjectives to measure their implicit self-concept (Greenwald & Farnham, 2000). But despite circumventing the biases of subjective ratings, these methods may not always enable as comprehensive personality measurements as self-reports do. They may also lack inherent psychological meaning (face validity) comparable to typical questionnaire items. Also, the objective measurement approaches may often have poor convergent and discriminant validity (Dreves et al., 2020; Mazza et al., 2020; Schimmack, 2020), possibly in part due to low reliability (e.g., Egloff et al., 2010; Wood & Brumbaugh, 2009).

Measurements with likely greater face validity are direct observations of behaviour and temporal and cross-situational patterns in this. These may include *in situ* self-or informant-reports of behaviour (via experience sampling) and visual and/or audio recordings taken in labs or everyday settings (Breil et al., 2019; Geukes et al., 2019; Schmid, Gatica-Perez, Frauendorfer, Nguyen, & Choudhury, 2015; Wrzus & Mehl, 2015). Indeed, there is a long-standing tradition in personality science to call for greater use of behavioural observations (e.g., Baumeister, Vohs, & Funder, 2007; Back, 2020; Back, in press), and well-cited articles have discussed suitable methods for this (e.g., Furr, 2009). We fully join with these calls and second that personality psychology that exclusively relies on subjective ratings, especially self-ratings, can only provide understanding of subjectively perceived variations and inevitably ignores anything not detectable, or inaccurately detected, by subjective perceptions. However, direct observations of behaviour have remained comparatively rare in personality research, likely because they are harder to obtain for sufficiently large samples and broad domains of behaviour. We hope that recent technological advances, such as those described in a recent special issue of European Journal of Personality (Rauthmann, 2020), will improve the situation.

Combining self- and informant-reports. Objective and / or *in situ* measurements of personality variance are highly desirable and increasingly practical, without any doubt. But it is also likely that subjective and decontextualized ratings will remain among the cost-efficient and ecologically valid methods of measuring stable personality traits, all the more so because the Big Few-centric research strategies have not yet fully exhausted this method's potential (e.g., Wood, Gardner, and Harms, 2015). A well-established but still underused way to improve the reliability and validity of subjective personality ratings is to supplement one rater (e.g., the self) with others (e.g., well-informed other people). With online testing, this is far easier than it was during the paper-and-pencil testing era (e.g., participants can nominate an informant for them, who is sent an automatic invitation to participate in the study).

Combining multiple raters can reduce systematic idiosyncrasies inherent in only one ratings source (McCrae et al., 2019; Vazire, 2006); indeed, such method-specific

effects may make up a large proportion of observed variance (McCrae, 2015). Self-ratings capture self-identity while informant-reports capture reputation; both are likely biased in their own ways, but what is shared between them is more likely to provide valid information. For example, most people have developed an implicit theory of which traits go together and adjust their self-ratings or ratings of someone else accordingly, which can lead to distorted correlations between data-points obtained with one rating source (McCrae et al., 2019). Combining ratings can also reduce random measurement error, especially in single- or few-item nuances where its proportion is higher than in broad aggregate traits.[6] This in turn can result in stronger associations with other variables of interest (e.g., Wright et al., 2019). Of course, informants may have different or less information about their target than the targets themselves do and they may often be biased towards the targets because of being non-randomly selected (Wessels et al., 2020). Likewise, we rarely know how discrepancies between self- and informant ratings arise – from biases in the former, latter, or both – and thereby how to weigh them in the combined results. No single source of information is perfect – but, again, combining them is very likely to improve data quality in most cases.

Multiple sources of ratings can sometimes be "triangulated" to estimate associations a) with reduced single method effects while b) also accounting for imperfect agreement between raters due to different information, rating biases or error (e.g., Biesanz & West, 2004; Eid et al., 2008; Riemann & Kandler, 2010). For example, using cross-trait, cross-twin ratings and cross-trait, cross-time ratings, Mõttus and colleagues (2017) calculated bias-and-error-reduced estimates of heritability and rank-order stability of personality nuances and found that the average estimates were comparable to those of aggregate traits, defying the intuition that broad psychological traits are more "biological" than circumscribed behaviours, feelings, thoughts and motivations.

Combining test-retest data. The reliability of personality trait assessments and thereby their associations with other variables can also be substantially improved by measuring presumably enduring traits twice over reasonably short time intervals (e.g., two weeks); besides, the associations can then be corrected for unreliability. Again, with online testing, organizing two or more measurement occasions is not as taxing as it used to be when testing was done on paper and when much of our current assessment practices were set, including the one-assessment-only tradition. It is especially useful if multiple self-ratings can be supplemented with informant-ratings: combining multiple pieces of information allows breaking correlations between variables into several components such as the association net of single-rater and occasion-specific biases, rater-specific effects and occasion-specific effects (e.g., Koch et al., 2017).

6 For example, if 50% of an item's variance is free of measurement error and single source method biases, then combining two raters yields a reliability of .67 for the aggregate, according to Spearman-Brown formula.

Better use of already existing data. Researchers can help to describe the associations of personality constructs among themselves and their relations with other variables in more detail than has been typically done – in fact, with little additional effort and by using data already collected.

For this, we recommend routinely a) using facets of the Big Few and/or b) testing extents to which associations are driven by narrower-still traits such as nuances (e.g., single items). Where the associations are driven by particular facets or nuances, they should not be automatically generalized beyond these, including to broader domains. Faceted and nuanced association patterns can be as informative and hypothesis-generative as the comfortingly predictable association patterns typical to the Big Few – desirable traits all too often going with desirable outcomes and the other way around, with most "significant" correlations somewhere between .10 and .30. We recommend that facet- and/or item-level findings be routinely published in article supplementary materials; this costs very little to authors (calculation and tabulation of findings) or journals, but it adds transparency to findings and facilitates their subsequent re-analysis and (e.g., meta-analytic) integration. This is different from making raw data available, because calculating the correlations of interest from these can often be cumbersome, unless very easy-to-use statistical programming code is made available.

Some may think that item-level findings are notoriously unreliable. But as was discussed before, items often have retest reliabilities of .65 to .70 or higher (Lowman et al., 2019; Mõttus et al., 2019; Wood et al., 2010; Henry & Mõttus, 2020), which may be higher than many intuitively expect. Higher-than-assumed single item reliability is also consistent with findings that items out-predict scales for outcomes and other variables (Mõttus & Rozgonjuk, 2019; Seeboth & Mõttus, 2018; Vainik et al., 2015; Achaa-Amankwaa et al, in press; Ellemann et al., 2020). Therefore, the allegedly low reliability of items should not be a reason for not reporting item-level findings. Where reliability is a concern, however, it can be compensated with large samples, meta-analytic integration of findings, and by aggregating or triangulating self- and informant-reports.

A Personality Research Hub. We recommend developing a central repository of descriptive findings. These findings could involve anything from associations among personality traits or their associations with demographic characteristics, life events and outcomes to their heritability, stability, and cross-method agreement estimates. We think that findings are best deposited disaggregated (e.g., at the item level), allowing for a flexible aggregation into different scales as well as for analysis at the item level. Centrally and publicly available findings can be tested for robustness across studies, as well as for moderators that help to understand why they vary from study to study or from scale to scale. They can also be meta-analytically combined and used for setting up and testing novel hypotheses (e.g., a routine practice in quantitative genetic research; Lee et al., 2018). Some such datasets have already been published (Mõttus,

Sinick, et al., 2019; Condon, Roney, & Revelle, 2017; Goldberg & Saucier, 2016), but there is no central repository of personality research findings yet.

For integrating findings across studies it is not necessary that all or even most studies use similar instruments. In fact, having all researchers assessing the same personality traits may not even be preferable for many research questions, because this would constrain the range of traits for which findings can become available over time. Instead, it is sufficient when studies rely on at least partly overlapping measures so that their associations can be compared for robustness and integrated into larger association networks. This directly parallels the idea of Synthetic Aperture Personality Assessment (Revelle et al., 2016), which allows calculating "synthetic" correlation matrices from only partly overlapping sets of participants. That is, not only can correlation matrices be based on different participant combinations of the same study, they can also be based on combined (synthetic) correlations from different studies. A similar procedure is routinely used in modern genetic research (e.g., Bulik-Sullivan et al., 2015). For working with such data, it is sufficient if (nearly) identical items and traits share annotation (common labels) – something that also helps against jingle-jangle fallacies.

Readily available descriptive findings, especially if they are not *a priori* aggregated into the Big Few, would facilitate a currently underused research strategy: setting up and testing hypotheses that rely on systematic variability between personality traits in their attributes such as demographic differences, stability, heritability, or links with outcomes (e.g., Funder & Dobroth, 1987; Block, Block, & Gjerde, 1986; Funder & Sneed, 1993; Mõttus et al., 2017; Vainik, Misic et al., 2019). That is, much like we study differences between people, we can also study *quantitative differences between traits* such as facets and nuances. This is not possible with only, say, five traits, but becomes increasingly viable as the number of traits increases.

For example, we could numerically test the hypothesis that personality development reflects social maturation (Caspi & Roberts, 2001). If associations between hundreds of items with age are meta-analyzed into reliable estimates, one could select, say, 200 diverse items, quantify their degrees of reflecting social maturity (e.g., using expert ratings or correlations with objective maturity-criteria) and expect these degrees to track with empirical age differences in the items. This would be a powerful and quantitative alternative to eyeball-judging that mean-level change patterns in traits such as the Big Few look like people are generally becoming socially more mature. For other examples, Henry and Mõttus (2020) examined whether items that corresponded to the definition of traits as opposed to characteristic adaptations demonstrated empirical properties often associated with traits such as stability, cross-rater agreement, and heritability; Hang, Soto, Lee and Mõttus (under review) studied whether items representing traits with stronger social expectations had larger age differences in means and variances throughout childhood and adolescence; Kööts-Ausmees and colleagues (in preparation) found that more socially desirable traits showed stronger age-differences in self-reports than in informant-reports, suggesting that age-differences may be inflated in self-reports; and Wood and Wortman (2012) showed that traits which varied least in their desirability across participants were least stable over time.

For a parallel, recent developments in quantitative genetics have been substantially facilitated by a wide-spread practice of sharing genotype-phenotype associations at the most fine-grained level (millions of single nucleotide polymorphisms) in repositories such as the LD Hub (Zheng, et al., 2017). Geneticists routinely (meta-analytically) integrate and re-analyze such data for various research questions, developing novel methodologies in the process. Much of this work is based on examining variabilities between genetic markers in their phenotype-associations or other attributes (e.g., allele frequencies or linkage disequilibrium), exactly as we recommend examining systematic variabilities between personality traits in their quantifiable attributes. The high-dimensional findings are filtered and aggregated in various ways such as by chromosome or gene expression patterns, to test hypotheses and summarize patterns. This is a fundamentally more flexible approach to data than the *a priori* aggregation of data-points that has prevailed in personality research.

New data analytic tools. In conjunction with depositing (disaggregated) findings, we recommend that researchers develop tools for collecting, annotating, archiving, processing, meta-analysing, and processing many-dimensional personality data. For example, we can imagine a software package (e.g., in R, possibly in combination with other platforms) that facilitates:

- administering subsets of item pools, selected according to pre-defined criteria;
- scoring them into various scales (e.g., the Big Five, HEXACO, Dark Triad, or well-being);
- uploading and downloading data from a central repository of findings according to specified criteria;
- automatically meta-analyzing new and/or existing findings for user-selected variables;
- cross-validating findings across different subsets of existing data and identifying candidate moderators;
- leveraging existing information (covariances among items) to impute unmeasured variables and to cross-walk from measured scales to (partly) unmeasured scales;
- summarising findings (e.g., personality-outcome correlations) at different levels of aggregation (personality hierarchy);
- identifying the variables (pre-defined scales, individual items, or computer-identified item collections; Ellemann et al, 2020) that uniquely (over and above other variables) drive particular associations (e.g., Vainik et al., 2015);

- testing the extent to which items' or broader traits' associations with particular variables track with their previously established properties such as reliability, social desirability, degrees of reflecting affect, motivation, and other psychological domains, developmental trajectories, and so forth, so as to better understand the associations and detect possible confounders;
- visualizing association patterns according to user-selected filters (e.g., compare item-outcome correlations in whether they pertain more to affective or behavioural items).

Some of these functions have already been implemented (e.g., Arslan, 2019; Arslan, Walter, & Tata, 2020; Revelle, 2020; Rosenbusch, Wanders, & Pit, 2020), but there is no comprehensive toolbox yet. Possibly, the main reason for why this does not already exist is lack of suitable databases; to date, personality researchers simply have not pooled their (disaggregated) findings, as some other fields have done to a good effect. We hope this will change. For a relevant example in cooperation research see Spadaro and colleagues (2020).

If personality science is moving towards higher-dimensional representations of phenomena, as we hope, this will also have implications for which skills needed to be taught to, and expected from, graduate students pursuing personality research.

Collaborations. Any one researcher or research group can collect only so much data. Individually, even the largest panel studies with often brief measures of personality traits may provide increasingly diminishing returns when the phenomena they explore are many-dimensional. But there is no rule that all research teams have to rely on the same omnibus model of personality and be constrained by the same practical limitations that prevent them from comprehensive measurement. Instead, we may need collaborations where different researchers explicitly set out to examine different aspects of personality (e.g., different traits) and only subsequently integrate their findings.

Within-individual variance

We have focused on variance between individuals in enduring patterns of thinking, feeling, behaving, and motivation, partly because this is what much of personality science is about. But recent years have seen the emergence of a powerful new stream of research that maps variance *within individuals* over very short time-periods and across situational experiences in what is often called personality states (Wendt et al, 2020, Sosnowska et al., 2020, Danvers et al., 2020, Horstmann & Ziegler, 2020), as well as stable individual differences in the distributions of these. This will likely provide more detailed descriptions of how particular individuals and people more generally interact with their environments and differ in this. Here, we do not describe this new and blooming stream of research in detail only because this special issue, as well as another recent special issue (Rauthmann, 2020), already contain papers that do exactly this. Here, we only note two things.

First, much of the research on short-term variance in personality states repurposes the descriptive models developed for summarizing individual differences such as the Big Few. But the extent to which this is appropriate needs to be studied not presumed (e.g., Molenaar & Campbell, 2009; Fisher et al., 2018). There is no reason to assume personality hierarchy operates the same way for individual difference traits and within-individual variance states, although sometimes it may. Many trait models are designed and measured with the specific purpose of glossing over temporal and situational variations, because personality is often conceived of as broad and decontextualised patterns of individual differences (Funder, 1991; McCrae & Sutin, 2018). It is useful to recall that the adjective pools that were used to derive the Big Few systematically excluded terms concerning moods or states (Saucier, 1997). For this and other reasons, employing the Big Few-like broad traits in studies on how personality states fluctuate just because this model is often used in individual differences research may not be a good idea, just as assuming that narrower traits such as facets or nuances are somehow more contextual-situational and thereby more appropriate candidates for personality states may be ill-conceived (Horstmann & Ziegler, 2020). Being artistic may be a useful narrow trait, but uninformative as a personality state. We suspect that some phenomena – for example, being talkative or sad – may constitute reliable variance units both as traits and states (Zimmermann et al., 2019), whereas others may only be appropriate as either.

Second, many of the recommendations that we propose for descriptive research on individual differences should also apply to descriptive research on within-individual variance in personality states. Among them are the need to develop a flexible descriptive framework that allows measuring phenomena with the most appropriate level of granularity for the purpose at hand, validating findings across methods, measures and contexts, combining self- and informant-reports, developing tools for flexibly working with and efficiently summarizing many-dimensional data, and developing efficient tools for data sharing and collaboration (e.g., Kirtley, 2020).

Predictive personality science

Personality researchers often take pride in how personality traits "predict" life outcomes such as academic performance, relationship satisfaction, or health. Strictly speaking, however, many of these findings – correlations or regression coefficients calculated using the same observations being predicted – are actually descriptive. Truly predictive research aims to create models where characteristics such as personality traits are used to model *the best possible predictions* of outcomes *in data that have not yet been accessed* or even collected (out-of-sample prediction). First, this means that the observations that are used to create, or "train", predictive models must not be the same observations

that will eventually be predicted (Yarkoni & Westfall, 2017; Stachl, Pargent et al., 2020). Second, such research should explore the limits of predictive accuracy, whereas descriptive models often have other priorities, as we argue below.

Given that the scientific value of personality traits is often said to hinge on their predictive power for important life outcomes (Ozer & Benet-Martínez, 2006; Roberts et al., 2007; Soto, 2019), it may come as a surprise that this power and ways of maximizing it have rarely been directly assessed in empirical studies. We suspect that this is in part due to a common failure to distinguish predictive research from other kinds of research and a tacit—but often likely mistaken—assumption that priorities and methodologies most suitable for descriptive or explanatory objectives must also be optimal for predictive purposes.

Why do predictive personality research?

Maximizing the out-of-sample predictive utility of personality traits can be an end in itself, sometimes even irrespective of its potential descriptive or explanatory utility. Consider, for example, using personality assessments for candidate selection (Lievens, 2017): what matters most is the accuracy of the estimated probability that the candidates will succeed in the job. Although for transparency it is useful to know which individual traits contribute to these predicted probabilities, the implications of those contributions for our understanding of personality more broadly are less important. Where the most accurate estimates of future job performance are based on the Big Few scores, it makes sense to use them. But where the best predictions are achieved by measuring, say, 100 unrelated personality items and feeding them directly into a predictive model, it may be counterproductive to combine them into broader trait constructs and use these for predictions, however descriptively elegant or comfortingly familiar this may seem. The same applies to using personality traits to decide which products are best advertised to which people (Matz et al., 2016) or for predicting important outcomes in medical and academic contexts, among other possible applications.

Maximising predictive accuracy has theoretical importance, too. Quite simply, to the extent that predictive accuracy is one of the main reasons for pursuing personality research, the case for this pursuit will be even stronger if we manage to increase the predictive accuracy. Likewise, one of the main theoretical implications of the pervasive personality trait-life outcome associations is that the traits may partly shape everyday experiences linked to these outcomes (e.g, differential education, career and relationship success confer different life trajectories and subsequent experiences) and thereby also shape psychological development more broadly (e.g., Scarr, 1983; Roberts & Nickel, 2017). That is, many psychologically consequential experiences are unlikely random but related to pre-existing psychological characteristics: traits' predictive accuracy is the formal measure of how pervasive this tendency is.

Why is predictive research different from descriptive and explanatory research?

It may not be obvious why descriptive models are not necessarily optimal for prediction. For example, doesn't R^2 of a regression model provide a good estimate of its predictive accuracy, even if that model was intended as a descriptive research tool to show how the variables in the model are linked with an outcome?[7] It can, especially when the model comprehensively covers relevant variables at the appropriate level of the personality hierarchy, as we recommended for descriptive research, and was developed on a sufficiently large sample to obtain stable parameter estimates. However, the best descriptive models do not have to be the most predictive ones, because efforts to optimize models for descriptive as well as explanatory appeal often *decrease* their predictive power, for two reasons.

First, a failure to cross-validate performance estimates (e.g., reporting an adjusted R^2 estimate derived from the same data the model was trained on) may result in overfitting (Yarkoni & Westfall, 2017; Stachl et al., 2020) and give overly optimistic impressions of predictive accuracy, while estimating how individual variables in models contribute to their cross-validated prediction reduces the models' descriptive simplicity (for examples, see Stachl, Pargent et al., 2020). To be fair, the issue of overfitting is probably less prevalent in more recent personality research and compared to many other fields of psychology, because often sufficiently large samples are used. But even so, an adjusted R^2 estimates a model's predictive performance in a hypothetical and infinitely large sample that was compositionally exactly identical to the one in which the model was fitted, whereas cross-validation allows one to estimate the robustness of the model across different kinds of samples. Researchers often assume that their findings are robust to variations in sample composition, but R^2 is insensitive to this.[8]

Second, human researchers' and their readers' cognitive constraints introduce a tension between descriptive/explanatory and predictive research objectives, because increased predictive accuracy is often achieved by increasing model complexity, which reduces interpretability and theoretical parsimony. For example, for descriptive and explanatory purposes researchers tend to look for and group correlated variables, whereas sets of variables that capture maximally unique portions of variance likely confer better prediction (Saucier, Iurino, & Thalmeyer, 2020). The increased complexity of predictive models may not only mean including many predictor variables (we do recommend high-dimensional descriptive research!), but also

7 In fact, many studies linking personality traits with outcomes only report correlations and not R^2 estimates.

8 One may expect that increasingly common meta-analyses provide average association estimates across different samples that are more generalizable than estimates from individual studies, and therefore less overfit. However, although likely more accurate due to aggregation, meta-analytic estimates may also be inflated due to overfitting in individual samples.

capitalizing on often uninterpretably small differences between already small weights of individual predictors and sometimes also incorporating non-linear associations and/or interactions between the predictors.

For example, Mõttus and Rozgonjuk (2019) reported that age could be out-of-sample predicted (in statistical, not substantive sense) more strongly from 300 individual test items (r = .65) than from 120 items (r = .54), 30 personality facets (r = .44) or the Big Five domains (r = .28). This shows that hundreds of items contain reliable and age-sensitive information about individual differences that is not fully exhausted by a set of 119, or possibly even 299, other items and that including this information in predictive models makes a material difference in their performance. But from a descriptive/explanatory standpoint, a model with 300 small regression coefficients that are carefully selected to maximize prediction may be suboptimal, because human researchers struggle to reason in so many dimensions and fathom the small differences between the coefficients. The findings have to be filtered or organized somehow to make them useful for descriptive and explanatory purposes. This predictive research just revealed that the Big Five (or any Big Few) may be a particularly suboptimal way of organizing items in their age differences.

For a parallel, the same applies to quantitative genetics, where polygenic models based on contributions from more numerous genetic variations (e.g., 100,000) generally allow for stronger out-of-sample predictions of phenotypic variables than models based on fewer genetic variants (e.g., 50,000), even though the contributions of individual variants are mostly far too small to be meaningfully interpretable (Plomin & von Stumm, 2018). Likewise, in fields like computer vision and natural language processing, opaque and complex statistical learning methods such as deep neural networks (DNNs) vastly outperform simpler, more interpretable statistical models (for review, see LeCun, Bengio, & Hinton, 2015). Many of these models capitalize on so many parameters and small variations in them that they may never be fathomable by humans: not because the models are overly complex *per se*, but because human minds have constraints that models do not have to obey (Hasson, Nastase, & Goldstein, 2020). We don't know yet whether the same will prove true for the prediction of individual differences in behavior (e.g., DNNs often require volumes and quality of training data rarely available in personality research), but this is not an unreasonable hypothesis. As it stands, there have simply been too few attempts to systematically explore the limits of personality traits-based predictions.

But initial evidence does suggest that techniques providing less human-interpretable model parameters such as regularized regressions or random forests may at least sometimes substantially out-perform more intuitive modeling approaches (e.g., Elleman et al., 2020). For example, regularized regression models often shrink many coefficients to a range that descriptively looks close to zero; even ordinary regression models with many predictors tend to do this. And sometimes comparatively more accurate predictions result from even more counter-intuitive modeling. For example, Mõttus & Rozgonjuk (2019) unsurprisingly found that regularized regression models predicted age from items much better than models based on the zero-order correlations of these items with age (i.e., if the predictions were formed by multiplying the standardized score of each item by its correlation with age in another sample and subsequently summing the products). But using zero-order correlations calculated with items' standardized *residuals* (i.e., after removing the variance of Big Five domains and facets from them) to create the prediction models improved their performance to levels comparable to regularized regression models. That is, removing the variance of the Big Five domains and facets from items prior to using them in the models *increased* their ability to out-of-sample predict, despite these items having been selected to measure the domains and facets in the first place.

This surely leaves classical test theorists scratching their heads: how can what is supposed to be error (i.e., left-over variance in items beyond the traits that they were designed to measure) out-predict traits? A plausible explanation is that predictive modeling benefits from uncorrelated predictors and minimizing their redundancy (Saucier, Iurino, & Thalmeyer, 2020). If so, capturing personality variation using sparsely placed markers (items) throughout the persome is more useful for prediction than relying on intuitive variables such as the Big Few or even their facets that capitalize on, and aggregate, correlated traits (i.e., oversample certain areas of the persome). This means a very different measurement philosophy than classical test theory.

It is important to avoid pejoratively calling predictive models with predictors and parameters that are not intuitive or familiar to human researchers "black box" models. They are not black boxes because, having designed them, humans can understand their working principles (Hasson et al., 2020). Besides, researchers know the data on which the models are trained because they designed the measures and collected the data. It is just that the specific parameter values that the models develop to do what modellers designed them to do are often not interpretable to these modellers, possibly due to their own cognitive constraints, but possibly also due to insufficient research and familiarization yet. Personality researchers should be open to the possibility that some, perhaps even many, of their familiar tools may become suboptimal when we start to systematically explore the limits of real-world predictions.

Thus, there may often be an inherent tension between parsimony and predictive power that forces researchers to choose between descriptively/theoretically elegant models that have lower predictive power and better-performing predictive models that benefit from the contributions of numerous variables with sometimes very small coefficients that individually make limited sense. Of course, other things being equal, it is always better to understand how a system operates than not. But sometimes, and maybe even very often, the true data-generating processes underlying

behaviour are too complex for a model to be simultaneously both comprehensible to humans and predictively maximally useful.

Can predictive models help descriptive and explanatory ones, and vice versa?

Predictive modeling can also facilitate progress in other kinds of research, where maximizing out-of-sample prediction is not an end in itself (for review, see Yarkoni and Westfall, 2017).

First, routine cross-validation can provide researchers with more realistic estimates of not only the predictive, but also the descriptive and explanatory capacity of their models. Impressive in-sample performance estimates derived from small-to-medium samples may decrease substantially when evaluated in independent samples, whereas the predictive power can hold up well with larger samples. But regardless of this, where predictive models with tens of well-chosen and well-measured predictors are able to account for only a fraction of the variability in the phenomenon of interest, researchers may want to remain humble about being able to map out the causes of this phenomenon, at least using the kinds of explanatory variables approximated by their predictors. That is, because one could argue that something can only be mechanistically explained to the extent it behaves predictably, predictive accuracy may often signal the limits of the explanatory powers of causal models.

Second, predictive models can help researchers understand the trade-offs inherent in emphasizing certain goals over others and identify important lacunae in descriptive or explanatory models. For example, even if one's goal is to develop a readily interpretable prediction equation using only the Big Few domains, quantifying the performance improvement one *might* obtain by using a more expansive set of predictors can help calibrate expectations about what "good" performance constitutes. It is not uncommon to learn that the Big Few are "powerful" predictors of life outcomes: comparing the predictive power of the Big Few to other trait models would help to either support or at least qualify such claims. The predictive models may also help to identify additional sources of variance for further descriptive/explanatory model development such as facets or nuances that could be included into the Big Few or besides them.

Third, predictively comparing different kinds of models can also shed light on the *general architecture* of personality variation in relation to predicted outcomes. For example, models based on hundreds of predictors out-performing those based on the Big Few or their facets would suggest that the associations of personality with the outcome could be driven by numerous specific processes, rather than a few broad mechanisms – to the extent that causality is involved, of course. Among other possibilities, this can be tested by dropping the strongest predictors from the model and estimating changes in the collective predictive power of the remaining predictors: it may be that this changes the results only minimally (Mõttus & Rozgonjuk, 2019).[9] Likewise, a finding that predictive models allowing for non-linear and/or interactive associations (e.g., recursive partitioning, random forests) do (or do not) out-perform those that only allow for linear additive associations can be equally informative about possible causal mechanisms, at least when the underperformance of complex models is not due to measurement error (Jacobucci & Grimm, 2020). Such findings can also inform intended personality-based interventions, not least about their likely limits in real-life settings.

Fourth, cross-validation as it is routinely done in predictive modeling provides an elegant way of estimating systematic (lack of) generalizability of results across measurable factors. For example, one can train a model on only some samples (e.g., only for men, North Americans, people younger than 50 years) and evaluate its performance on others (e.g., women, Asians, those aged over 50); if the models perform equally well, the factors that differentiate between the samples do not moderate the associations captured by the model.

On the other hand, attending to descriptive and explanatory concerns can also help improve the performance of predictive models. Most importantly, researchers can draw on their domain expertise to facilitate better "feature engineering"; that is, choosing which variables are used in the predictive models and how they are pre-processed (Stachl, Pargent et al., 2020). No amount of machine learning expertise is likely to produce optimal predictions if the available predictors contain mostly noise (Jacobucci & Grimm, 2020) or lack coverage of the critical features of the target phenomenon. An understanding of the sources and structure of human personality and psychometric expertise can be particularly helpful for maximising predictive potential and for anticipating issues with generalizing the models beyond original training settings. For example, it is likely that personality trait inventories that contain items with high reliability but relatively little redundancy are particularly useful for prediction, despite the trait scales having lower internal consistencies and thereby potentially putting off users with less or outdated psychometric expertise (Yarkoni, 2010).

In particular, because accuracy of out-of-sample predictions entirely depends on comprehensive, well-measured and generalizable sets of predictors, theoretical accounts

9 When predicting age from personality test items, Mõttus and Rozgonjuk (2019) tried removing items of several facets that had the strongest correlations with age. Surprisingly, they found that the overall predictive capacity of the models decreased minimally, suggesting that the bulk of the predictive information was not uniquely concentrated to a small selection of items or the traits that they were supposed to index. Not reported in the original paper, but specifically for the current article, we ran additional out-of-sample predictions of age in these data, by dropping 5%, 10%, 25% and 50% of the most predictive items from the total of 300 items: the correlation between predicted and actual ages dropped from .65 to .61, .59, .51 and .41, respectively. These predictions were still far more accurate than those provided by the Big Five domains (.28) and mostly also more accurate than those of the facets (.44), even with these including all their items. This suggests that small amounts of unique age-sensitive information were allocated across many individual items.

elucidating the processes by which personality relates to the outcome and descriptive accounts showing how the outcome is correlated with personality traits can both be useful for *expanding* the range of predictors included in predictive models. This may go against the intuition of some researchers to use prior knowledge to constrain models. For training predictive models, however, it does not matter how many predictors are initially involved or what putative personality hierarchy levels they come from, so long as they help maximize suitably generalizable out-of-sample prediction accuracy. As long as the models are not validated using the observations on which they were trained, any excesses in predictor selection will become apparent in the model validation phase and can be corrected.

Some recommendations for predictive research

Cross-validation. For an accurate evaluation of the predictive value of personality traits, it is most important to use cross-validation procedures that distinguish between the training sample and the validation sample (Yarkoni & Westfall, 2017; Stachl, Pargent et al., 2020). These can be independent partitions of one larger sample (as in *k*-folds or leave-one-out cross validation), but it is even better if they are independently collected datasets, potentially with somewhat varying demographic characteristics. Cross-validation helps to mitigate against model overfitting due to random sampling variance as well as due to systematic biases in sampling (e.g., demographic imbalances), and it can guard against the effects of idiosyncracies in data collection, processing, and statistical modeling. It is especially valuable if the training and validation data were collected by different researchers.

Sufficiently large datasets. Predictive performance tends to improve with increasing model complexity, so long as the training data is sufficiently large to mitigate over-fitting. As a general rule, the more predictors in a model and/or the more complex the functional form relating predictors to the criterion (e.g., allowing for non-linear associations), the larger the training sample that is required. The incremental gains associated with larger sample sizes also depend on the effect sizes in question, as large effects require smaller samples, and the amount of missing data (Elleman et al., 2020). For example, Mõttus & Rozgonjuk, found that prediction models stabilized with a few hundred observations when based on up to 30 variables, but required about 3,000 observations when based on 300 predictors with the smallest individual effect sizes and presumably most measurement error. We therefore do not suggest universally "acceptable" sample sizes; instead, this can be estimated with simulations for individual study designs. For many predictive modeling applications in personality psychology, it is possible that increased sample sizes will have diminishing returns beyond a few thousand observations.

But more variables is often preferable to more participants. Researchers rarely have the luxury of acquiring massive samples with many well-measured variables, and often face a choice between collecting less data from more people or more data from fewer people. In such cases, larger participant numbers are not always desirable. Instead, prioritizing the coverage of the persome by increasing the number of variables at the expense of the number of participants may confer substantial predictive advantages (the same likely applies to descriptive research), provided that the variables used during training are also available in the validation data and any future observations for which predictions are intended. A large number of responses to a short personality questionnaire can be a poor substitute for a rich dataset, even if the latter contains fewer observations. For example, a sample of 3,000 participants measured with 200 items may often enable more predictive (as well as descriptive) models than a sample of 12,000 participants measured with 50 items, and a sample of 60,000 measured with 10 items is likely to fare worse still. Ultimately, the predictive information is in the variables and most outcomes are highly multiply determined, with observations only needed to reliably estimate relevant information in the variables. Besides, many statistical estimation methods such as regularized regressions are designed to help stabilize predictions even in cases where the number of variables exceeds the number of independent observations. A particularly useful solution to balance participant and item numbers is to collect data with massively planned missingness, where each participant provides responses to a different random subset of variables (e.g., Revelle et al., 2017; Elleman et al., 2020).

Flexibility in selecting and transforming predictors. When constructing predictive models from personality data, researchers have flexibility over how, or whether at all, to transform single data points such as item scores into predictive variables; this may involve aggregating, raising to powers or grouping values, for example. In machine learning, this process is termed *feature engineering*. Aggregation tends to filter out potentially useful information, so measuring many traits with one item each can result in more predictive models than measuring few traits with many items. But aggregation may be useful when this demonstrably improves the generalizability of the prediction models across contexts and instruments. For example, it may be that an item-based prediction model vastly out-predicts a model based on fewer aggregate traits in a given sample, but when the trained model is applied in a different demographic group, the gap may close or even reverse. As a general rule, different ways of aggregation could be empirically compared to each other as well as to completely disaggregated models in their ability to predict outcomes in independent data (e.g., Mõttus, Bates et al., 2017).

Comparing statistical models in their performance. Sometimes, well-tuned regularized regression models may provide far more robust and accurate predictions than "standard" (i.e., ordinary least-squares) regression models; sometimes the latter may work just as well. Also, models that allow for non-linear and interactive associations may sometimes provide the most accurate predictions, even if they require larger training samples. In some circumstances

such as high levels of missing data, less sophisticated and less data-hungry models may provide comparably accurate predictions: for example, Elleman and colleagues (2020) introduced the Best Items Scales that are Cross validated, Unit weighted, Informative, and Transparent (BISCUIT) model that allows researchers to create bespoke personality scales for particular outcomes, consisting of as few items as possible and each contributing exactly the same amount towards the prediction for greater interpretability. Our general point is: to date, there has been too little research that has systematically explored the ways of maximising the predictive accuracy of personality variables and therefore we cannot know yet which modeling practices are generally preferable.

Alternative sources of personality information

Predictive personality research may not only use personality traits as predictors, but also as outcomes. A wealth of recent research has explored the possibility to extract personality-relevant information not only from traditional sources like self-reports, but from records that people leave behind such as social media or credit card records, mobile sensor data or diaries (Kosinski et al., 2013; Stachl, Au et al., 2020; Weston et al., 2019; Wiernik et al., 2020). Typically, such data is given psychological meaning by first collating them into scores that approximate self-reported personality traits (e.g., using machine learning techniques; Wiernik et al., 2020) and then using these digital records-based self-report-approximations for descriptive or predictive purposes. The standard approach so far has been to predict the Big Few first and then use these predictions for whatever is their intended purpose, but recent evidence suggests that predicting narrower traits such as nuances first and using these predictions in subsequent analyses may be preferable (Hall et al., 2020). Again, more research is needed before we could recommend generally preferable research practices, and therefore it may be useful to systematically compare different approaches in their performance.

Competitions among research teams

Prediction is different from description and explanation in that there is an objective ground truth for assessing performance: the agreement of predictions with actual observations. This creates an opportunity for researchers to directly compete against one another in developing the best possible prediction models, which could go a long way towards eventually establishing the best practices for the field. For example, teams of researchers could be given similar training data with the only instruction to develop the most accurate prediction models for given outcomes, and the submitted models could be compared in their performance in hold-out data that were not available to model developers (e.g., Salganik et al, 2020).

Explanatory personality science

Many psychologists are not satisfied with describing and predicting personality-relevant phenomena (e.g., traits or their correlates; events, actions, affects, goals, life outcomes) and also aspire to *explain* them (e.g., Baumert et al., 2017). Few would disagree, however, that explaining something is harder than describing and predicting it, not only because of methodological challenges but also because of more fundamental questions about the very nature of useful explanations. In fact, even the authors of this article could not entirely agree on some fundamental questions around causes, explanations and their roles in personality science. Fortunately, there have been other recent contributions regarding how to explain phenomena that personality scientists consider as falling into their jurisdiction (e.g., Baumert et al., 2017; Briley et al., 2018; Grosz et al., 2020), including articles in this issue (e.g., Quirin et al., 2020; Costantini et al,. 2020; Lukaszewski et al., 2020). Here we offer general ideas about how one could think of causes and useful explanations – and why these are not necessarily the same things.

Crucially, there are different approaches to personality that vary in what their advocates may consider useful and realistic goals of explanation. Some conceive of personality as broad regularities in relatively stable individual differences, whereas others think of it as a dynamic and potentially idiosyncratic within-person system, and see the role of personality science as providing an integrative account of how the mind and behaviour come together. The former approach focuses on decontextualised patterns in *naturally occurring, normal and continuous (dimensional) variance among individuals* (e.g., Funder, 1991; McCrae & Sutin, 2018); in this view, personality is a population-level variance phenomenon such as the trait hierarchy. The latter approach is primarily about *specific processes pertaining to individuals* and resultant variability within them, as well as about how individuals may differ in these processes and/or their distant causes (e.g., Quirin et al, 2020). In many cases, it is not evident that variability/processes taking place within individuals and variability among people arise for similar reasons (e.g., DNA structure, anthropometry, parental socioeconomic status or other possible sources of individual differences do not even vary much within individuals), although sometimes they may (see Lukaszewski et al., 2020; Quirin et al, 2020). But even more importantly, while advocates of the latter approach may hope to identify specific causes of specific phenomena (why a particular person reacts to a situation in a particular way) and eventually perhaps even individual differences in these, advocates of the former approach may prefer explanations that propose general principles rather than target specific causes, for reasons that we'll describe shortly.

Causes

Causes can be defined as broad and specific factors (e.g., neurological structures or repeated experiences) or processes (e.g., situation selection or associations among

psychological constructs) that play roles in producing particular responses to environments, or vice-versa, either psychologically or behaviourally. Even if inferred from comparing individuals, causes and effects pertain to processes and variability *within particular individuals in their particular circumstances*. Causal relations have boundary conditions, which can range from the exceptionally narrow (e.g., where affecting X should only affect Y in rare circumstances and needs to be studied idiographically) to very broad (e.g., on Earth, releasing an object almost always causes it to fall toward the Earth). Explanations that target causes thus mean specifying (1) the nature of the cause-effect relation or process (such as $X \rightarrow Y$ or $X \rightarrow M \rightarrow Y$) and (2) the circumstances under which the relation or process is expected to occur.

The gold-standard for identifying causes is the *potential to control* the outcome by experimentally manipulating these conditions and/or processes. For instance, if we have learned that Helen exercises *because* of X_1, or Tom parties *because* of X_2, we should be able to at least in principle influence the levels of X_1 and X_2 to change Helen's rate of exercising and Tom's rate of partying. This involves *counterfactual arguments*: if X and Y occurred and we assert that X was a cause of Y, then we have to be able to show that, without X, Y would not have happened in the way or to the degree that it did, all else being equal. Formalized models of hypothesized processes that enable controlling them at least conceptually (e.g., Directed Acyclic Graphs, DAGs, with *do*-operators; Pearl, 2018) can be particularly useful for probing such specific causal relations.

To be clear, causes do not have to be deterministic; for example, smoking causes lung cancer, but not every smoker gets it. But the probabilistic link between the cause and effect has to be consistent and strong enough such that changing the former makes a non-trivial difference for the latter. Indeed, the risk of smokers developing lung cancer is about 20 times higher than that of non-smokers (Surgeon General's Reports, 2004), so starting to smoke makes a material difference for the probability of developing lung cancer. In contrast, very small individual causal effects have commensurately small explanatory power.

Defined as such, identifying causes may be a useful target for approaches that see the primary role of personality science as identifying potentially controllable processes that underlie within-individual variance and perhaps subsequently also individual differences in these processes (e.g., Quirin et al., 2020). For example, a therapist may be able to identify causes of a patient's problematic behaviours and perhaps even help the patient to control them to facilitate desired personality change (Hopwood, 2018; Magidson, Roberts, Collado-Rodriguez, & Lejuez, 2014). Likewise, functionalist and process approaches may attempt to explain how particular beliefs and skills interact to produce certain behaviours or self-perceptions, which can similarly provide 'levers' for influencing behaviours or trait change (Wood, Spain, Monroe, & Harms, in press; Metcalfe & Mischel, 1999).

However, it is not self-evident that researchers who think of personality as population-level patterns in naturally-occurring individual differences and seek to make sense of these should target their individual causes. This is because these patterns *may not have many tractable causes* to begin with, at least according to our definition of cause, or they may be too numerous and too complex to provide explanations that are interpretable for the human mind and therefore useful. Instead, useful explanations for these patterns could postulate general principles that may or may not apply to potentially controllable processes in particular individuals. We now elaborate on this position, because we feel that it is implicitly adopted by many personality researchers but may cause unrealistic expectations when left unarticulated (Grosz, Rohrer, & Thoemmes, 2020). We will later return to the alternative view according to which personality researchers should hope to reveal the individual causes of personality-relevant phenomena in the strict sense of the term.

Why many causes may be inherently elusive

In one part, causes may often remain elusive because the phenomena that personality scientists seek to explain and/or their plausible explanatory variables are, by definition and intentionally, abstract hypothetical constructs that cut across different circumstances within and across individuals (Funder, 1991), with quantitative levels that are inherently relative.

Think of individual differences in neuroticism, self-esteem, agency, trustfulness, or procrastination as quintessential examples of the kinds of personality constructs many researchers work with. To be personality constructs rather than just specific instances of behaviour, thoughts, feelings, and desires, they represent individual differences in reactions that integrate across many kinds of situations and over time, and are therefore taken out of their specific circumstances. Unless one commits to the view that they represent singular traits (like height) that exist independently of how and where they are expressed and measured (arguably, most personality researchers do not; e.g., Baumert et al., 2017), this inevitably makes them decontextualized aggregates that correspond to different things in different people and circumstances. Also, individuals' "raw" scores on them can only be interpreted in comparison to those of others, because there are few if any concrete "anchors" (e.g., specific behaviours) that invariably correspond to specific trait levels and ground these in individuals.[10] According to our definition, however, causes need to represent concrete "things" (e.g., thoughts, feelings, behaviours, desires, skills, experiences, brain structures,

10 There have been attempts to create personality rating scales that provide raters with concrete behavioural anchors rather than the typical disagreement-agreement dimension such as Likert scale (e.g., Muck, Hell, & Höft, 2008). These may be useful for assessing the manifestation of personality traits in specific circumstances in a non-relativistic way, but the measures tend to be too context-specific to be of general use and to allow for comparing individuals from different circumstances.

even genes) that do correspond to specific circumstances and apply to particular individuals, irrespective of other individuals.

Of course, although many personality constructs are, by nature, decontextualized and relativistic aggregates, their constituents such as behaviours measured with questionnaire items could be concrete enough to also represent situation-specific reactions of particular individuals. If so, we could work backwards from construct levels to what they correspond to in individuals. This may sometimes be the case, especially for narrower constructs that aggregate few constituents; this alone is a good reason to consider lower levels of the trait hierarchy. However, not many personality constructs can boast a well-defined set of concrete constituents: The Act Frequency Approach (e.g., Buss & Craik, 1983) was one prominent attempt to delineate them, but has been largely abandoned for decades. Even for narrow constructs such as the tendency for aggressive behaviour, researchers often ask those who provide information on it to make abstract inferences ("I tend to get into fights", "S/he often hits others") rather than count the frequencies of the specific behaviours involved – because these are too context-bound to be meaningfully comparable across people.

But even if researchers did have, or will manage to reach, a consensus on what are the concrete constituents of specific traits and how to measure them in a non-relative way, they will face another challenge: there are often so many different configurations of these constituents through which any given aggregate value can arise that it is virtually impossible to connect a specific construct score to the values of its constituents in individuals. Any non-extreme level of a construct with even just a handful of facets or nuances can correspond to hundreds of unique facet/nuance configurations, with even the most common of them remaining rare. Intuitively, we may expect that if a person has a medium score on a construct they must also have a medium level on most of its constituents; in fact, generally this is *not* the case.[11] This is a mathematical and empirical fact that may be greatly underappreciated among researchers.

Given this, why do personality scientists not work with variables (e.g., individual genes, brain variables, life experiences, personality nuances, behaviours, or feelings) that are sufficiently concrete and measurable in a non-relative way to serve as causes per our definition? Some already do, and many more may think that they should in order to make progress. For example, we made the case for a greater use of personality nuances in other sections of this article; these are at least somewhat more concrete than broad trait domains. Likewise, we echo those arguing for the importance of moving beyond subjective trait-ratings to objectively measured behaviour (see also Back, 2020). But, again, for many researchers the core of personality science is just something else by definition – broader and decontextualized patterns of individual differences (e.g., McCrae & Sutin, 2018) – so asking them to study only highly specific and contextualized variables instead amounts to asking them to redefine their field of study. The decontextualized nature of personality traits, for example, is often seen as their particular strength (Funder, 1991; McAdams, 1994) and something that makes personality science unique among other fields such as social, developmental, cognitive, or clinical psychology. This is hard to argue with.

But equally importantly, the specificity required of variables that could have causal impacts on personality phenomena such as patterns in naturally-occurring individual differences may often mean that they are too numerous to be individually useful as explanations (Yarkoni, 2020). Besides, many causes can have multiple effects, which further complicates disentangling them. For an extreme example, even if individual DNA base pair variations directly cause individual differences in personality constructs, it will take many thousands of them to account for even a small fraction of the variance, because their individual effects are miniscule (e.g., Lo et al., 2017; Nagel et al., 2018). Most of the individual effects are not even statistically significant in any given sample. This is now so well established that it is called the Fourth Law of Behaviour Genetics (Chabris et al., 2015). Likewise, the very same genetic variants pervasively matter for variations in a whole range of behavioural, social and somatic traits, known as pleiotropy (e.g., Turkheimer, Pettersson, Horn, 2014; Mõttus, Realo et al., 2017; Nagel et al., 2018).

In many cases, the number of potentially relevant causes may be smaller than the very high number of somehow-personality-related genetic variants. But the typical effect sizes in psychology and the pervasive tendency for all things to correlate (a manifestation of the psychological "pleiotropy" that is sometimes called the crud factor; Orben & Lakens, 2020) make it unlikely for many personality phenomena to have distinct causes that are sufficiently strong to explain both behaviour and psychological processes in particular individuals and a non-trivial amount of normal variability between people. Among other things, this is consistent with the lack of robust evidence for the effects of specific life experiences on personality constructs, even in the most powerful studies to date (e.g., Asselmann

11 For illustration, we simulated an unrealistically simple construct (N = 10,000,000) that was defined by only five independent constituents, each having only three levels (-1, 0, 1 with 25%, 50%, and 25% probabilities), and a small amount of uniformly distributed "error" (ranging from -1 to 1 and accounting for about 12% of variance in construct scores). We then extracted about 20,000 scores of this construct with nearly identical values (0 +/- .005) and found that these corresponded to hundreds of configurations of their five constituents. By far the most obvious configuration of the five constituents (all 0s) corresponded to only 7% of the scores and each of the second most prevalent combinations (three 0s, one -1, and one 1) corresponded to less than 2% each. In the real world, of course, few personality-related constructs are almost completely defined by only a handful of well-defined constituents, so our ability to deduce from a construct score to what this may represent in particular individuals is much smaller still. If the constituents are not completely independent (e.g., as semantically non-redundant items of a scale), some configurations become relatively more likely, but this does not change the conclusion. See also Østergaard, Jensen, and Bech (2011).

& Specht, 2020[12]; Chopik et al., 2020; Denissen et al., 2019). Bleidorn and colleagues (2020) recently called for far more detailed examinations of the effects of life experiences on changes in personality constructs than are available to date ("Longitudinal Experience-Wide Association Studies" or LEWAS, p. 285). If Genome-Wide Association Studies are anything to go by, then linking numerous life experience "variants" with changes in personality constructs in large samples will indeed account for a fraction of variance in them, although the findings should always be cross-validated in independent samples to avoid overfitting. This would be an impressive and important empirical feat, but whether this could help us towards potentially controllable and theoretically meaningful *causes* of why particular individuals do what they do, or why they differ in this, is another question.

An equally fundamental reason that identifying specific cause-effect associations is often impractical is that it requires unrealistic assumptions, in particular that causality runs in only one direction (Pearl, 2018). Naturally occurring personality variability represents how free-ranging individuals spontaneously differ when left to their own devices in largely self-created environments. In fact, the very essence of personality is the means by which people choose, adapt to, and modify their real-world situations and experiences to suit them (Buss, 1987). As a result, what may be considered causes of personality characteristics often do not happen to people randomly, but are influenced by something coming from within them – their personalities, potentially including the variables to be explained and other variables linked with these. For example, people's experiences, not just observable traits, are correlated with genetic variance among them (Scarr, 1983). Where this applies, there are no clear cause and effect associations and formal models of causality (e.g., DAGs) and counterfactuals fail: flipping an explanans to its counterfactual state automatically means flipping its explanandum as one of its causes, .suggesting that we cannot eliminate "back-doors" to explanandum (Pearl, 2018). This is also a reason that experimental manipulations and other interventions, even if feasible practically and ethically, could sometimes misrepresent causality in personality science and beyond. In real life, people often choose the "manipulations" and "interventions" that suit them and do all they can to avoid others, in part based on their personalities.

Finally, in some and maybe even many cases, links between phenomena and their plausible causes exist in such narrow circumstances as to be unique to individuals or only small subsets of them (e.g., Beck & Jackson, 2020), which further complicates connecting them with population-level variance in individual differences constructs (cf. Beltz et al., 2016; Dotterer et al., in press; Woods et al., 2020; Wright et al., 2019). The more idiosyncratic the associations are, the less practical and even plausible it is to identify the specific causes of individual differences, at least as long as these are defined as dimensions along which individuals vary. At present, far more research is needed to better estimate the extent to which this applies and whether this generalizes across types of associations (e.g., links between psychological or behavioural phenomena as well as their links with physiological, anatomical, and genetic variables) or levels of the trait hierarchy (Wright & Zimmermann, 2019).

Given all this, it may seem sensible to keep explanations that could apply to what particular individuals do in their particular circumstances and that could potentially be manipulated separate from explanations of population-level variability in situation-general patterns such as traits in the personality hierarchy. These may end up being very different kinds of explanations.

Explanations short of specific and potentially modifiable causes

Where identifying specific causes is not feasible or reasonable, internally coherent and consistent-with-available-observations *narratives* of how normal variation in clearly defined phenomena comes about may serve as the most useful explanations. A useful explanation may state its scope (what kinds of variance patterns are being explained) and premises (what is assumed and not further explained), and specify its observed and unobserved components and general principles of relations among them (how they are organized or tend to inter-relate over time, and in which circumstances they are likely to occur or not occur).[13] For example, abstract narratives about developmental principles of individual differences (e.g., Caspi & Moffitt, 1993; Roberts & Nickel, 2017) may be good candidates to become useful explanations, despite – and maybe exactly because of – not attempting to outline the specific causes of the patterns that they try to explain. Articulating only a few causes would explain just about nothing, whereas attempting to list a sufficient number of them, even if feasible at some point, could make explanations unintelligible.

It is particularly useful if such explanations can be formalized as computational models (Quirin et al., 2020). Although these cannot provide empirical proof and are unlikely to reveal causes in the strict sense of the term, they allow playing through complex hypotheses that involve large numbers of hypothetical variables with potentially many-to-many and bidirectional relationships that can unfold over many iterations. Setting up a computational model that runs and produces results that are even broadly consistent with observations of relevant real-world phenomena often takes a lot of rigorous thinking and is all too likely to identify gaps in verbal-only explanations

12 One may want to adjust the associations reported for personality change in this study for multiple testing. Depending on method of adjustment, this may result in only one significant association between life events and trait change ($\beta = .08$ for decrease in emotional stability after divorce).

13 Besides the 'how' part, there may also be a 'why' part of an explanation, referring to the function (outcome) of the phenomenon in relation to a broader phenomenon (e.g., the function of anger may be to restore equity in social transactions; Lukazsewski et al., 2020); assuming that every explanation involves a function may be problematic, however (e.g., some phenomena are no longer functional or may even appear dysfunctional, but still require an explanation).

(Mõttus, Allerhand, & Johnson, 2020). Examples of the use of computational models in personality science include Revelle and Condon's (2015) dynamics of action model, Read and colleagues' (2010) neural network model, Smaldino and colleagues' model of niche diversity (Smaldino et al., 2019), or Mõttus and colleagues' (2020) model of person-environment transactions and the corresponsive principle.

But are explanations, defined this way, really more than descriptions? We argue that they are if they help to interpret, organize, and integrate descriptive observations. That is, if they fill in knowledge gaps, help researchers to envisage yet-to-be made observations, and suggest possible directions for more detailed explanations. However, we realize that the line between the explanations, defined this way, and descriptive findings is probably far less clear than many would prefer. Indeed, what may seem as identifying causal explanations may often, at a closer look, amount to more detailed and better organized descriptions (Yarkoni, 2020). If so, well-documented and detailed basic descriptive findings are and likely will be central parts of many personality scientific explanations. Descriptive findings are then not just uninspired examples of personality research to be replaced with "proper" causal explanations; they are the ingredients that useful explanations organize into coherent narratives.

For example, theories that seek to explain personality variations through social interactions may benefit from a large-scale project (say $N = 10{,}000$) that documents, in both lab and naturalistic settings, hundreds of objectively measured behaviours, social interaction processes and their subjective perceptions, besides including detailed trait ratings of the participants (see also Back, in press). Using such data, researchers could look for patterns in behaviour, perception and relationship dynamics and link these to measurements of individual differences, possibly being able to account for a non-trivial fraction of variation in personality nuances, facets and domains. Almost certainly, however, a large number of such patterns would uniquely contribute to accounting for trait variance. These findings, such as those from LEWAS (Bleidorn et al., 2020), would be descriptive and unlikely to reveal causes of naturally occurring individual differences in the strict sense of the term. But they could help to identify recurring regularities in behaviour and psychological processes and thereby develop and refine useful explanatory models of personality variation.

Grosz, Rohrer, and Thoemmes (2020) have recently argued that there is a widely-spread taboo against causal inference in non-experimental personality science in that what researchers are allowed to explicitly claim to have achieved is often not what their findings and interpretations actually imply – between the lines. We suspect that this is in part because of a failure to distinguish useful explanations from causes in the sense that we defined them above and many other researchers do as well, at least implicitly. In many cases, researchers can hope to achieve explanations, but not necessarily identify specific causes, because these are either intractable, unintelligible, or both.

It may help to tackle this taboo to realize and accept that many phenomena that personality scientists are focused on may, by their very nature, be distinctly unique in *not* having tractable unidirectional causes (Yarkoni, 2020). The best explanations for these phenomena may often hinge on the most coherent available narratives that combine many pieces of, and patterns in, descriptive findings rather than rely on specific and definitive experiments or statistical models. For example, whether a particular regression coefficient does or does not represent a causal effect in a strict sense may often be a moot question and (suppressing) arguments over this may simply reflect naivety. Regardless of this, regression coefficients *alongside* other findings of descriptive research can be a useful basis for narrative explanations.

Alternative view: Identifying tractable causes may be a tractable problem after all

On the other hand, many researchers – including several authors of this article – disagree with the view that attempts to explain personality may often be best off not targeting its specific causes. Instead, they believe that researchers will eventually identify the specific and potentially even controllable causes of key personality phenomena, including naturally-occuring individual differences in them and broader patterns in these. This will require better methods, measures, and models. But even more importantly, this likely entails (a) defocusing from the broad and situation-general patterns of variation as the starting points of explanations in favour of specific and contextualized within-individual processes and (b) tolerating the complex and potentially phenomenon- and person-specific (idiographic) explanations that result from this shift. In what follows, we discuss what may be particularly important to facilitate moving towards causes-based explanations in personality science.

Some recommendations for explanatory research that seeks to identify causes

Identifying the right level of analysis for explanation. Units at certain levels of analysis may be too far apart to construct meaningful causal accounts, at least without intermediate steps. For instance, reductionists may argue that all psychology can be understood by biology, all biology by chemistry, and all chemistry by physics. But it is unlikely that we will ever identify a tractable explanation of how a leader's personality affects her organization's longevity through particle physics. Instead, explanations using units at more proximal levels to the phenomena we wish to explain may be more useful and appropriate (Borsboom, Cramer, & Kalis, 2019; Dennett, 2013; Hofstadter, 2007; Sperry, 1966). Social cognitive, learning, or functionalist accounts which explain personality trait levels as arising through the interactions of units such as goals, expectancies, affordances, and perceptual processes, may be more appropriate and necessary components of

causal accounts of the phenomena than explanations through specific genes or even specific neurological structures (Back, in press; Baumert et al., 2017).

Once armed with proximal causal explanations, however, researchers can move on to identify the causes of these causes, which ultimately can serve as a strategy for making sense of associations across different levels of analysis. For instance, given the extremely distal relations between genes and psychological traits (e.g., Johnson & Edwards, 2002), identifying genetic variants responsible for between-individual variation in dominance might be aided by first identifying the major proximal causes of the variation, and then working backwards. Trait dominance tends to be elevated among individuals high in physical formidability, which in turn tends to be correlated with the individual's physical height (Lukaszewski, Simmons, Anderson, & Roney, 2016). If so, understanding the genes affecting height can help to understand the genetic variants affecting formidability, which can help to understand some part of the genes affecting dominance. The large number of specific genes affecting height in turn can be organized into smaller sets of specific genes affecting narrower biological processes such as those affecting bone lengths, cartilage production, hormone production, skeleton morphology, and other processes (e.g., A. Wood et al., 2014). Thus, as we improve our accounts of the important proximal causes of a phenomenon of interest, we can in turn identify the most important proximal causes of these variables, at each step identifying more specific targets we can place as intermediators to bridge the gulf across more distal levels of analysis. Those versed in the structural modeling literature may think of this strategy as building a series of multiple indicators, multiple causes (MIMIC) models.

Even if we can eventually identify a tractable number of major proximal causes of our phenomenon of interest, this strategy of iteratively identifying the proximal causes of each proximal cause as outlined in this example will likely result in hundreds or thousands of distal causes with miniscule effects. However, at each end of the long and complex causal chains linking one of these distal causes to the outcome of interest, we could be able to identify stronger causal associations. For instance, on one end of the chain linking specific genes to height or dominance, the NOX4 gene's association with height is likely mediated through stronger effects on the number of osteoclasts cells produced, which aid in bone repair and maintenance (Marouli et al., 2017). On the other, physical formidability and other proximal causes may each have moderate to large main effects (e.g., $r > .30$) on dominant behaviour (Lukaszewski et al., 2016). This strategy may thus help to organize the legions of variables showing small distal effects by showing how they contribute to more proximally related variables and processes, such as the psychological mechanisms or systems that calibrate dominant and aggressive behavior (e.g., Balliet, Tybur, Van Lange, 2017; Lukaszewski et al., 2020). The successful identification of the most proximally related processes in turn offers the greatest potential for intentionally affecting outcomes of interest. For instance, a man might try to facilitate his displays of dominant behaviour by increasing his formidability, perhaps by 'bulking up' by downing protein shakes and spending hours lifting weights at the gym.[14]

It is important, however, not to confuse variation within individuals with individual differences. The former may, and in many cases likely does, contribute to the latter. But the individual differences in within-individual processes that could contribute to other individual differences have to come from somewhere in the first place (Lunansky, Borkulo, & Borsboom, 2020; Quirin et al., 2020) and, as we know from well documented behaviour genetics findings (e.g., Briley & Tucker-Drob, 2014), many sources of individual differences are a) hardly random and b) often not something in which individuals even vary greatly over time (e.g., DNA structure). It may thus be that to a large extent the processes reflected in within-individual variance either amplify (e.g., corresponsive processes between traits and experiences; Nickel & Roberts, 2007) or dampen/reverse (e.g., somebody with maladaptive characteristics seeking to change these) pre-existing individual differences, or translate some other traits (e.g., non-psychological characteristics such as height, metabolic, endocrine or other traits) into psychological traits, rather than create individual differences from scratch.

Working with cleaner units. As noted above, there is a tendency for personality psychologists to combine diverse, causally efficacious sets of variables into single aggregates. However, excessive emphasis on broad all-purpose domains such as the Big Few impedes representing the personality processes or dynamics underlying the phenomena (e.g., Block, 1995; Mischel and Shoda, 1995; Cramer et al., 2012; Wood, Gardner, & Harms, 2015; van der Mass et al., 2006). This is a point that we consistently make throughout this paper: we should be flexible about how, and whether at all, we aggregate variables. For instance, we might imagine that tendencies toward [1] *liking and caring about people* increases a person's likelihood of [2] *doing favours for other people*, which in turn can increase a person's likelihood of [3] *being liked by other people*. Averaging such tendencies into a single scale score complicates understanding the nature of the causal relationships that the conceptually distinguishable attributes have with one another (van der Maas et al., 2006; Wood, Gardner, & Harms, 2015; Epskamp, Waldorp, Mõttus, & Borsboom, 2018).[15] This can also contribute to the view that even moderate (possibly) causal relations among personality variables are hard to find when in fact they are often hiding in plain sight – within our scales. A key recommendation, then, is that researchers a)

14 If the likelihood of increasing formidability turns out to be systematically linked to its plausible downstream causes such as dominance (less dominant people may bother less with having physical means of appearing threatening), the situation becomes more complicated, though, because the cause and effect become entangled, as we discussed above. Scenarios such as this may in fact be uniquely prevalent for personality-related phenomena.

15 It will also often result in putting indicator items of the outcomes we want to predict with personality scales directly into the personality scale, making it difficult to rule out that the correlations may reflect uninteresting tautologies (Mõttus, 2016; Nicholls, Licht, & Pearl, 1982).

aim for constructs and their measures that prioritize *conceptual* distinctions between variables (e.g., items that concern self-perceptions of behaviour *vs* affect or motivation; Wilt & Revelle, 2015; Wood, Gardner, & Harms, 2015) over purely *empirical* ones (e.g., average all items with factor loadings over .40) or b) deliberately create measures for distinct classes of personality-relevant phenomena (e.g., Jackson et al., 2010; Costantini, Saraulli, & Perugini, 2020).

Extending our range of methods and models. Establishing causal relations between variables often requires stronger evidence than cross-sectional correlations. It is ultimately important to provide evidence that manipulating *X* within a potential *X*→*Y* relationship would alter the level of *Y*. But this is often difficult as many of the *X*'s that we examine as potential causes of personality phenomena, such as specific genes, or the size or connectivity of neurological areas, do not lend themselves to manipulation and many *Y*'s also influence their *X*'s, entangling the causes with effects.

Meanwhile, what is almost certain to help is greater use of repeated measures designs, over both long (e.g., multi-wave longitudinal studies such as Denissen et al., 2019) and short measurement windows (e.g., experience sampling studies such as Sosnowska et al., 2020; Danvers et al., 2020). Within such studies, finding that the levels of *X* at one time point *t* are associated with the levels of *Y* concurrently (at time *t*) or even prospectively (e.g., can predict how *Y* will change from *t* to *t+1*; e.g., Epskamp et al., 2018) is useful for bolstering evidence of causal associations, also allowing to separate within- and between-individual variances. Time-series data may be combined with experimental designs, such as by experimentally manipulating the *X* state – for example, instructing people to pursue certain goals or to act extraverted – and see if the *Y* state tends to increase in response (e.g., Margolis & Lyubomirsky, 2019; Steiger et al., 2020). There remain important questions about the extent to which experimentally manipulating psychological states serves as an ecologically valid means of understanding how the states naturally covary, however, due to issues such as self-selection effects (i.e., reverse causality) and issues of finding the ideal time intervals to identify causal effects (e.g., Jacques-Hamilton et al., 2019).

We also encourage within-individual variance designs that focus on estimating idiographic association patterns besides nomothetic ones (Beck & Jackson, 2020; Wright, Gates et al., 2019). It is crucial that we understand how far our typical nomothetic models of variance can go in principle – that is, how broad are the boundary conditions of possible causal effects. The broader the boundary conditions and less idiosyncratic personality processes are, the more useful nomothetic models can be in identifying the causes of personality phenomena, however numerous and multi-leveled these end up being, and vice versa.

Concluding remarks

In this article, we discussed three main kinds of personality research – descriptive, predictive, and explanatory – and argued that they involve different priorities and face different methodological and practical challenges. Descriptive research aims to delineate associations among personality-relevant phenomena and their link with other variables as comprehensively as possible, while also doing this in ways that allow flexibly summarizing and organizing this information; predictive research aims to maximize generalizable out-of-sample predictive power without much regard to the descriptive or explanatory elegance of the statistical models; and approaches aiming to explain personality phenomena need to be clear about their levels of analysis (patterns in naturally occurring individual differences *vs* psychological processes and behaviour of particular people) and set targets that are appropriate and realistic for the type of variability or processes that are being explained.

It does not seem to us that these research kinds should strive towards homogenization between and even within them, at least not any time soon. An approach that aims to achieve all goals may eventually not achieve any of them particularly well. Descriptively most useful models may not be most predictive or provide satisfactory explanations; most predictive models may be too complicated to be useful for description or explanation; and limiting descriptive research or predictive modeling to variables and associations that make conceptual sense may be counterproductive.

That said, it would be equally wrong to suggest that they are in isolation from one another. For example, descriptive findings can be the basis for building predictive and explanatory models, predictive models can help to expand the range of descriptive research, hint at the limits of explanatory models (e.g., how much variability among people in a phenomenon can models hope to account for), and explanations can suggest which further descriptive research is needed or what could be included in prediction models. For these reasons, it is important that descriptive, predictive and explanatory approaches rely on at least partly overlapping sets of constructs wherever possible. However, we argue that the commonly-used Big Few alone is suboptimal for this and we need to develop flexible models of personality variance that fully embrace its hierarchical organization and do not confuse patterns of individual differences with variance and processes within individuals. We also need tools to assess the variance and processes that rely on different sources and types of information, not just self-reports.

References

Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2020). Coffee or Tea? Examining Cross-Cultural Differences in Personality Nuances Across Former Colonies of the British Empire. https://doi.org/10.31234/osf.io/dpqrx

Allik, J., Church, A. T., Ortiz, F. A., Rossier, J., Hřebíčková, M., de Fruyt, F., Realo, A., & McCrae, R. R. (2017). Mean Profiles of the NEO Personality Inventory. Journal of Cross-Cultural Psychology, 48, 402–420. https://doi.org/10.1177/0022022117692100

Arslan, R. C. (2019). How to Automatically Document Data With the codebook Package to Facilitate Data Reuse: Advances in Methods and Practices in Psychological Science. https://doi.org/10.1177/2515245919838783

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

Ashton, M. C., & Lee, K. (2020). Objections to the HEXACO Model of Personality Structure—And Why Those Objections Fail. European Journal of Personality, 34, 492–510. https://doi.org/10.1002/per.2242

Asselmann, E., & Specht, J. (2020). Taking the ups and downs at the rollercoaster of love: Associations between major life events in the domain of romantic relationships and the Big Five personality traits. Developmental Psychology. doi:10.1037/dev0001047

Back, M. D. (2020). Editorial: A Brief Wish List for Personality Research. European Journal of Personality, 34, 3–7. https://doi.org/10.1002/per.2236

Back, M. D. (in press). Social interaction processes and personality. In J. Rauthmann (Ed.), The handbook of personality dynamics and processes. Elsevier.

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

Balliet, D., Tybur, J. M., & Van Lange, P. A. (2017). Functional interdependence theory: An evolutionary account of social situations. Personality and Social Psychology Review, 21, 361-388.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the Science of Self-Reports and Finger Movements: Whatever Happened to Actual Behavior? Perspectives on Psychological Science, 2, 396–403. https://doi.org/10.1111/j.1745-6916.2007.00051.x

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., 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

Beck, E. D., & Jackson, J. J. (2020). Idiographic Traits: A Return to Allportian Approaches to Personality. Current Directions in Psychological Science, 29, 301–308. https://doi.org/10.1177/0963721420915860

Beltz, A.M., Wright, A.G.C., Sprague, B., & Molenaar, P.C.M. (2016). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. Assessment, 23, 447-458.

Bem, D. J., & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. Psychological Review, 85, 485–501. https://doi.org/10.1037/0033-295X.85.6.485

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the big five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. Journal of Personality, 72, 845–876.

Bleidorn, W., Hopwood, C. J., Ackerman, R. A., Witt, E. A., Kandler, C., Riemann, R., Samuel, D. B., & Donnellan, M. B. (2020). The healthy personality from a basic trait perspective. Journal of Personality and Social Psychology , 118, 1207. https://doi.org/10.1037/pspp0000231

Bleidorn, W., Hopwood, C. J., Back, M. D., Denissen, J. J. A., Hennecke, M., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Roberts, B. W., Wagner, J., Wrzus, C., and Zimmermann, J. (2020) Longitudinal Experience-Wide Association Studies—A Framework for Studying Personality Change. European Journal of Personality, 34, 285– 300. https://doi.org/10.1002/per.2247.

Bleidorn, W., Klimstra, T. A., Denissen, J. J. A., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality Maturation Around the World A Cross-Cultural Examination of Social-Investment Theory. Psychological Science, 24, 2530–2540. https://doi.org/10.1177/0956797613498396

Block, J. (1995). A contrarian view of the Five-Factor Approach to personality description. Psychological Bulletin, 117, 187215.

Block, J. H., Block, J., & Gjerde, P. F. (1986). The Personality of Children Prior to Divorce: A Prospective Study. Child Development, 57, 827–840. https://doi.org/10.2307/1130360

Block, J. H., Gjerde, P. F., & Block, J. H. (1991). Personality antecedents of depressive tendencies in 18-year-olds: A prospective study. Journal of Personality and Social Psychology, 60, 726–738. https://doi.org/10.1037/0022-3514.60.5.726

Borsboom, D., Cramer, A., & Kalis, A. (2019). Brain disorders? Not really...: Why network structures

block reductionism in psychopathology research. Behavioral and Brain Sciences, 42, 1–11

Bouchard, T. J. (2016). Experience producing drive theory: Personality "writ large." Personality and Individual Differences, 90, 302–314. https://doi.org/10.1016/j.paid.2015.11.007

Breil, S. M., Geukes, K., Wilson, R. E., Nestler, S., Vazire, S., & Back, M. D. (2019). Zooming into Real-Life Extraversion – how Personality and Situation Shape Sociability in Social Interactions. *Collabra: Psychology, 5,* 7

Briley, D. A., & Tucker-Drob, E. M. (2014). Genetic and environmental continuity in personality development: A meta-analysis. Psychological Bulletin, 140, 1303–1331. https://doi.org/10.1037/a0037091

Briley, D. A., Livengood, J., & Derringer, J. (2018). Behaviour Genetic Frameworks of Causal Reasoning for Personality Psychology. European Journal of Personality, 32, 202–220. https://doi.org/10.1002/per.2153

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L., & Neale, B. M. (2015). An atlas of genetic correlations across human diseases and traits. Nature Genetics, 47, 1236–1241. https://doi.org/10.1038/ng.3406

Buss, D. M. (1987). Selection, evocation, and manipulation. Journal of Personality and Social Psychology, 53, 1214-1221.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. Psychological Review, 90, 105–126. https://doi.org/10.1037/0033-295X.90.2.105

Caspi, A., & Moffitt, T. E. (1993). When Do Individual Differences Matter? A Paradoxical Theory of Personality Coherence. Psychological Inquiry, 4, 247–271. https://doi.org/10.1207/s15327965pli0404_1

Caspi, A., & Roberts, B. W. (2001). Personality Development across the Life Course: The Argument for Change and Continuity. Psychological Inquiry, 12, 49–66. https://doi.org/10.2307/1449487

Chabris, C. F., Lee, J. J., Cesarini, D., Benjamin, D. J., & Laibson, D. I. (2015). The Fourth Law of Behavior Genetics. Current Directions in Psychological Science, 24, 304–312. https://doi.org/10.1177/0963721415580430

Chopik, W. J., Oh, J., Kim, E. S., Schwaba, T., Krämer, M. D., Richter, D., & Smith, J. (2020). Changes in optimism and pessimism in response to life events: Evidence from three large panel studies. Journal of Research in Personality, 88, 103985. https://doi.org/10.1016/j.jrp.2020.103985

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

Condon, D.M., Roney, E. and Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. Journal of Open Psychology Data, 5, p.3. DOI: http://doi.org/10.5334/jopd.32

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological Bulletin, 136, 1092–1122. https://doi.org/10.1037/a0021212

Cooper, A. B., Blake, A. B., Pauletti, R. E., Cooper, P. J., Sherman, R. A., & Lee, D. I. (2020). Personality Assessment Through the Situational and Behavioral Features of Instagram Photos. European Journal of Psychological Assessment. https://doi.org/10.1027/1015-5759/a000596

Costa, P. T., & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual. Psychological Assessment Resources.

Costa, P. T., McCrae, R. R., & Löckenhoff, C. E. (2019). Personality Across the Life Span. Annual Review of Psychology, 70, 423–448. https://doi.org/10.1146/annurev-psych-010418-103244

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. Journal of Research in Personality, 54, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003

Costantini, G., Saraulli, D., & Perugini, M. (2020). Uncovering the Motivational Core of Traits: The Case of Conscientiousness. European Journal of Personality, n/a(n/a). https://doi.org/10.1002/per.2237

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., Kendler, K. S., & Borsboom, D. (2012). Dimensions of Normal Personality as Networks in Search of Equilibrium: You Can't Like Parties if You Don't Like People. European Journal of Personality, 26, 414–431. https://doi.org/10.1002/per.1866

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. Educational and Psychological Measurement, 64, 391418. https://doi.org/10.1177/0013164404266386

Danvers, A. F., Wundrack, R., & Mehl, M. (2020). Equilibria in Personality States: A Conceptual Primer

for Dynamics in Personality States. European Journal of Personality. https://doi.org/10.1002/per.2239

Denissen, J. J. A., Luhmann, M., Chung, J. M., & Bleidorn, W. (2019). Transactions between life events and personality traits across the adult lifespan. Journal of Personality and Social Psychology, 116, 612–633. https://doi.org/10.1037/pspp0000196

Dennett, D. C. (2013). Intuition pumps and other tools for thinking. Oxford, England: Norton.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. Journal of Personality and Social Psychology, 91, 1138–1151. https://doi.org/10.1037/0022-3514.91.6.1138

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

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

Dotterer, H.L., Beltz, A.M., Foster, K.T., Simms, L.J., & Wright, A.G.C. (in press). Personalized models of personality disorders: Using a temporal network method to understand symptomatology and daily functioning in a clinical sample. Psychological Medicine. https://psyarxiv.com/bnxkq/

Dreves, P. A., Blackhart, G. C., & McBee, M. T. (2020). Do behavioral measures of self-control assess construct-level variance? Journal of Research in Personality, 88, 104000. https://doi.org/10.1016/j.jrp.2020.104000

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

Egloff, B., Schwerdtfeger, A., & Schmukle, S. C. (2005). Temporal Stability of the Implicit Association Test-Anxiety. Journal of Personality Assessment, 84, 82–88.

Elleman, L. G., McDougald, S. K., Condon, D. M., & Revelle, W. (2020). That takes the BISCUIT: A comparative study of predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. European Journal of Psychological Assessment.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. Multivariate Behavioral Research, 53:4, 453-480. https://doi.org/10.1080/00273171.2018.1454823

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

Funder, D. C. (1991). Global Traits: A Neo-Allportian Approach to Personality. Psychological Science, 2, 31–39. https://doi.org/10.1111/j.1467-9280.1991.tb00093.x

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. Journal of Personality and Social Psychology, 52, 409–418. https://doi.org/10.1037/0022-3514.52.2.409

Funder, D. C., & Sneed, C. D. (1993). Behavioral manifestations of personality: An ecological approach to judgmental accuracy. Journal of Personality and Social Psychology, 64, 479–490. https://doi.org/10.1037/0022-3514.64.3.479

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

Geukes, K., Breil, S. M., Hutteman, R., Nestler, S., Küfner, A.C.P., Back, M.D. (2019). Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The PILS and the CONNECT study. *PlosOne, 14,* e0210424

Gniewosz, G., Ortner, T. M., & Scherndl, T. (2020). Personality in Action: Assessing Personality to Identify an 'Ideal' Conscientious Response Type with Two Different Behavioural Tasks. European Journal of Personality. https://doi.org/10.1002/per.2296

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology, 59, 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf , Personality Psychology in Europe (Vol. 7, pp. 7–28). Tilburg University Press.

Goldberg, L. R., & Saucier, G. (2016). ORI Technical Report. (Vol. 56 No. 1). Eugene, OR.

Gonzalez, O., MacKinnon, D. P., & Muniz, F. B. (2020). Extrinsic Convergent Validity Evidence to Prevent Jingle and Jangle Fallacies. Multivariate behavioral research. https://doi.org/10.1080/00273171.2019.1707061

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. Annual Review of Psychology, 66, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. Journal of Personality and Social

Psychology, 79, https://doi.org/10.1037/0022-3514.79.6.1022

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. Perspectives on Psychological Science. https://doi.org/10.1177/1745691620921521

Hall, A. N., & Matz, S. C. (2020.). Targeting Item-level Nuances Leads to Small but Robust Improvements in Personality Prediction from Digital Footprints. European Journal of Personality. https://doi.org/10.1002/per.2253

Hang, Soto, Lee and Mõttus (under review). Social expectations and abilities to meet them as possible mechanisms of youth personality development.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron, 105, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. Behavioral and brain sciences, 33, 61-83.

Henry, S., & Mõttus, R. (2020). Traits and Adaptations: A Theoretical Examination and New Empirical Evidence. European Journal of Personality, 34, 265–284. https://doi.org/10.1002/per.2248

Hilbig, B. E., Moshagen, M., Zettler, I. (2016). Prediction consistency: A test of the equivalence assumption across different indicators of the same construct. European Journal of Personality, 30, 637–647. https://doi.org/10.1002/per.2085

Hofstadter, D. R. (2007). I am a strange loop. Basic books.

Hopwood, C. J. (2018). Interpersonal Dynamics in Personality and Personality Disorders. European Journal of Personality, 32, 499–524. https://doi.org/10.1002/per.2155

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

Jacques-Hamilton, R., Sun, J., & Smillie, L. (2019). Costs and benefits of acting extraverted: A randomized controlled trial. Journal of Experimental Psychology: General, 148, 1538–1556.

Jackson, J. J., Walton, K. E., Harms, P. D., Bogg, T., Wood, D., Lodi-Smith, J., Edmonds, G. W., & Roberts, B. W. (2009). Not all Conscientiousness Scales Change Alike: A Multimethod, Multisample Study of Age Differences in the Facets of Conscientiousness. Journal of Personality and Social Psychology, 96, 446–459. https://doi.org/10.1037/a0014156

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

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement: Perspectives on Psychological Science. https://doi.org/10.1177/1745691620902467

Jang, K. L., McCrae, R. R., Angleitner, A., Riemann, R., & Livesley, W. J. (1998). Heritability of facet-level traits in a cross-cultural twin sample: Support for a hierarchical model of personality. Journal of Personality and Social Psychology, 74, 1556–1565.

Johnston, T. D., & Edwards, L. (2002). Genes, interactions, and the development of behavior. Psychological Review, 109, 26-34.

Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. Psychological Review, 123, 90–96. https://doi.org/10.1037/0022-3514.74.6.1556

Kandler, C., Zimmermann, J., & McAdams, D. P. (2014). Core and Surface Characteristics for the Description and Theory of Personality Differences and Development. European Journal of Personality, 28, 231–243. https://doi.org/10.1002/per.1952

Kirtley, O. J., Hiekkaranta, A. P., Kunkels, Y. K., Eisele, G., Verhoeven, D., Van Nierop, M., & Myin-Germeys, I. (2020). The Experience Sampling Method (ESM) Item Repository. https://doi.org/10.17605/OSF.IO/KG376

Koch, T., Schultze, M., Holtmann, J., Geiser, C., & Eid, M. (2017). A Multimethod Latent State-Trait Model for Structurally Different And Interchangeable Methods. Psychometrika, 82, 17–47.

Kööts-Ausmees, L., Kandler, K., McCrae, R. R., Realo, A., Allik, J., Borkenau, P., Hřebíčková, M., & Mõttus, R. (in preparation). Social Desirability and Age Differences in Personality Traits: A Multi-Rater, Multi-Sample Study

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110, 827–840. https://doi.org/10.1073/pnas.1218772110

Larsen, K. R., & Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. MIS Quarterly, 40, 123.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444. https://doi.org/10.1038/nature14539

Lee, K., & Ashton, M. C. (2020). Sex differences in HEXACO personality characteristics across countries

and ethnicities. Journal of Personality. https://doi.org/10.1111/jopy.12551

Lee, J. J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T. A., Bowers, P., Sidorenko, J., Linnér, R. K., Fontana, M. A., Kundu, T., Lee, C., Li, H., Li, R., Royer, R., Timshel, P. N., Walters, R. K., Willoughby, E. A., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature Genetics, 50, 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Leising, D., Vogel, D., Waller, V., & Zimmermann, J. (2020). Correlations between person-descriptive items are predictable from the product of their mid-point-centered social desirability values. European Journal of Personality.

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

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., Smeland, O. B., Schork, A., Holland, D., Kauppi, K., Sanyal, N., Escott-Price, V., Smith, D. J., O'Donovan, M., Stefansson, H., Bjornsdottir, G., Thorgeirsson, T. E., Stefansson, K., McEvoy, L. K., … Chen, C.-H. (2017). Genome-wide analyses for personality traits identify six genomic loci and show correlations with psychiatric disorders. Nature Genetics, 49, 152–156. https://doi.org/10.1038/ng.3736

Lowman, G. H., Wood, D., Armstrong, B. F., Harms, P. D., & Watson, D. (2018). Estimating the reliability of emotion measures over very short intervals: The utility of within-session retest correlations. Emotion, 18, 896–901. https://doi.org/10.1037/emo0000370

Lucas, R. E., & Donnellan, M. B. (2009). Age differences in personality: Evidence from a nationally representative Australian sample. Developmental Psychology, 45, 1353–1363. https://doi.org/10.1037/a0013914

Lukaszewski, A. W., Lewis, D. M. G., Durkee, P. K., Sell, A. N., Sznycer, D., & Buss, D. M. (2020). An Adaptationist Framework for Personality Science. European Journal of Personality. https://doi.org/10.1002/per.2292

Lukaszewski, A. W., Simmons, Z. L., Anderson, C., & Roney, J. R. (2016). The Role of Physical Formidability in Human Social Status Allocation. Journal of Personality and Social Psychology, 110, 385-406. https://doi.org/10.1037/pspi0000042

Lunansky, G., Borkulo, C. van, & Borsboom, D. (2020). Personality, Resilience, and Psychopathology: A Model for the Interaction between Slow and Fast Network Processes in the Context of Mental Health. European Journal of Personality. https://doi.org/10.1002/per.2263

Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. International Journal of Psychology, 54, 705–711. https://doi.org/10.1002/ijop.12529

McAdams, D. P. (1994). A Psychology of the Stranger. Psychological Inquiry, 5, 145–148. https://doi.org/10.1207/s15327965pli0502_12

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

Magidson, J. F., Roberts, B. W., Collado-Rodriguez, A., & Lejuez, C. (2014). Theory-driven intervention for changing personality: Expectancy value theory, behavioral activation, and conscientiousness. Developmental Psychology, 50, 14421450. https://doi.org/10.1037/a0030583

Margolis, S., & Lyubomirsky, S. (2019). Experimental manipulation of extraverted and introverted behavior and its effects on well-being. Journal of Experimental Psychology: General. Advance online publication. https://doi.org/10.1037/xge0000668

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S., Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., … Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. Nature, 542, 186–190. https://doi.org/10.1038/nature21039

Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the Structure of Normal and Abnormal Personality: An Integrative Hierarchical Approach. Journal of Personality and Social Psychology, 88, 139–157. http://dx.doi.org/10.1037/0022-3514.88.1.139

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences, 114, 12714. https://doi.org/10.1073/pnas.1710966114

Mazza, G. L., Smyth, H. L., Bissett, P. G., Canning, J. R., Eisenberg, I. W., Enkavi, A. Z., Gonzalez, O., Kim, S. J., Metcalf, S. A., Muniz, F., III, W. E. P., Scherer, E. A., Valente, M. J., Xie, H., Poldrack, R. A., Marsch, L. A., & MacKinnon, D. P. (2020). Correlation Database of 60 Cross-Disciplinary Surveys and Cognitive Tasks Assessing Self-Regulation. Journal of Personality Assessment. https://doi.org/10.1080/00223891.2020.1732994

McAbee, S. T., & Connelly, B. S. (2016). A multi-rater framework for studying personality: The trait-reputation-identity model. Psychological Review, 123, 569-591

McCrae, R. R. (2015). A More Nuanced View of Reliability: Specificity in the Trait Hierarchy. Personality and Social Psychology Review, 19, 97–112. https://doi.org/10.1177/1088868314541857

McCrae, R. R., De Bolle, M., Löckenhoff , C. E., & Terracciano, A. (in press). Lifespan trait development: Towards an adequate theory of personality. In J. F. Rauthmann (Ed.), Handbook of personality dynamics and processes. Amsterdam: Elsevier.

McCrae, R. R., & Costa Jr., P. T. (1996). Towards a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins , The five-factor model of personality: Theoretical perspectives (Vol. 51, pp. 51–87). Guilford Press.

McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. Journal of Personality, 60, 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McCrae, R. R., & Mõttus, R. (2019). What Personality Scales Measure: A New Psychometrics and Its Implications for Theory and Assessment. Current Directions in Psychological Science, 28, 415–420. https://doi.org/10.1177/0963721419849559

McCrae, R. R., & Sutin, A. R. (2018). A Five-Factor Theory Perspective on Causal Analysis. European Journal of Personality, 32, 151–166. https://doi.org/10.1002/per.2134

McCrae, R. R., Mõttus, R., Hřebíčková, M., Realo, A., & Allik, J. (2019). Source method biases as implicit personality theory at the domain and facet levels. Journal of Personality, 87(4), 813–826. https://doi.org/10.1111/jopy.12435

McCrae, R. R., Terracciano, A., & 78 Members of the Personality Profiles of Cultures Project. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. Journal of Personality and Social Psychology, 88, 547–561. https://doi.org/10.1037/0022-3514.88.3.547

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. Psychological Review, 106, 3–19.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psychological Review, 102, 246–268. http://dx.doi.org/10.1037/0033-295X.102.2.246

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. Current Directions in Psychological Science, 18, 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x

Mõttus, R. (2016). Towards more rigorous personality trait-outcome research. European Journal of Personality, 30, 292–303. https://doi.org/10.1002/per.2041

Mõttus, R., Allerhand, M., & Johnson, W. (2020). Computational Modeling of Person-Situation Transactions: How Accumulation of Situational Experiences Can Shape the Distributions of Trait Scores. In D. C. Funder, R. A. Sherman, & J. F. Rauthmann (Eds.), *Handbook of Psychological Situations*. (pp. xx – xx).

Mõttus, R., & Rozgonjuk, D. (2019). Development is in the details: Age differences in the Big Five domains, facets and nuances. Journal of Personality and Social Psychology. http://dx.doi.org/10.1037/pspp0000276

Mõttus, R., Allik, J., & Realo, A. (2020). Do Self-Reports and Informant-Ratings Measure the Same Personality Constructs? European Journal of Psychological Assessment, 36, 289–295. https://doi.org/10.1027/1015-5759/a000516

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D., & Revelle, W. (2017). Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. https://doi.org/10.31234/osf.io/4q9gv

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. https://doi.org/10.1037/pspp0000100

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait heterogeneity in age group differences in personality domains and facets: Implications for the development and coherence of personality traits. PLoS ONE, 10, e0119667. https://doi.org/10.1371/journal.pone.0119667

Mõttus, R., Realo, A., Vainik, U., Allik, J., & Esko, T. (2017). Educational attainment and personality are genetically intertwined. Psychological Science, 28, 1631–1639. https://doi.org/10.1177/0956797617719083

Mõttus, R., Sinick, J., Terracciano, A., Hrebickova, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability and utility of personality nuances. Journal of Personality and Social Psychology, 117, e35–e50. https://doi.org/10.1037/pspp0000202

Muck, P. M., Hell, B., & Höft, S. (2008). Application of the principles of Behaviorally Anchored Rating Scales to assess the Big Five personality constructs at

work. In J. Deller , Research contributions to personality at work (pp. 77-97). München, Germany: Rainer Hampp

Nagel, M., Watanabe, K., Stringer, S., Posthuma, D., & Sluis, S. (2018). Item-level analyses reveal genetic heterogeneity in neuroticism. Nature Communications, 9, 905. https://doi.org/10.1038/s41467-018-03242-8

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. Psychological Bulletin, 92, 572-580. https://doi.org/10.1037/0033-2909.92.3.572

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. Advances in Methods and Practices in Psychological Science, 3, 238–247. https://doi.org/10.1177/2515245920917961

Østergaard, S.D., Jensen, S.O.W. and Bech, P. , The heterogeneity of the depressive syndrome: when numbers get serious. Acta Psychiatrica Scandinavica, 124: 495-496. doi:10.1111/j.1600-0447.2011.01744.x

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. Annual Review of Psychology, 57, 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Paunonen, S. V., & Ashton, M. C. (2001). Big Five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology, 81, 524–539. psyh. https://doi.org/10.1037/0022-3514.81.3.524

Paunonen, S. V., & Jackson, D. N. (2000). What is beyond the big five? Plenty! Journal of Personality, 68, 821–835. https://doi.org/10.1111/1467-6494.00117

Pearl, J. (2018). The Book of Why: The New Science of Cause and Effect. New York: The Basic Books.

Plomin, R., & von Stumm, S. (2018). The new genetics of intelligence. Nature Reviews Genetics, 19, 148–159. https://doi.org/10.1038/nrg.2017.104

Quirin, M., Robinson, M. D., Rauthmann, J. F., Kuhl, J., Read, S. J., Tops, M., & DeYoung, C. G. (2020). The Dynamics of Personality Approach (DPA): 20 Tenets for Uncovering the Causal Mechanisms of Personality. European Journal of Personality. https://doi.org/10.1002/per.2295

Rauthmann, J. (in press). A (More) Behavioral Science of Personality in the Age of Multi-Modal Sensing, Big Data, Machine Learning, and Artificial Intelligence. *European Journal of Personality.*

Read, S. J., Monroe, B. M., Brownstein, A. L., Yang, Y., Chopra, G., & Miller, L. C. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117*, 61–92.

Revelle, W., & Condon, D. M. (2015). A model for personality at three levels. *Journal of Research in Personality, 56,* 70–81.

Revelle, W. (2020) psych: Procedures for Personality and Psychological Research. (Version 2.0.9). Northwestern University. http://CRAN.R-project.org/package=psych

Revelle, W., Condon, D. M., Wilt, J., French, J. A., Brown, A., & Elleman, L. G. (2016). Web and phone based data collection using planned missing designs. Sage handbook of online research methods (2nd ed., p. 578-595). Sage Publications, Inc.

Revelle, W., Dworak, E. M., & Condon, D. M. (2020). Exploring the persome: The power of the item in understanding personality structure. Personality and Individual Differences, 109905.

Roberts, B. W., & Nickel, L. B. (2017). A critical evaluation of the Neo-Socioanalytic Model of personality. In J. Specht , Personality Development Across the Lifespan (pp. 157–177). Academic Press. https://doi.org/10.1016/B978-0-12-804674-6.00011-9

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The Structure of Conscientiousness: An Empirical Investigation Based on Seven Major Personality Questionnaires. Personnel Psychology, 58, 103–139. https://doi.org/10.1111/j.1744-6570.2005.00301.x

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. Perspectives on Psychological Science, 2, 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. Advances in Methods and Practices in Psychological Science, 1, 27-42. https://doi.org/10.1177/2515245917745629

Rosenbusch, H., Wanders, F., & Pit, I. L. (2020). The Semantic Scale Network: An online tool to detect semantic overlap of psychological scales and prevent scale redundancies. Psychological Methods, 25, 380-392. http://dx.doi.org/10.1037/met0000244

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., … McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences, 117, 8398. https://doi.org/10.1073/pnas.1915006117

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. Journal of

Personality and Social Psychology, 73, 12961312. https://doi.org/10.1037/0022-3514.73.6.1296

Saucier, G., & Iurino, K. (2019). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. Journal of Personality and Social Psychology. https://doi.org/10.1037/pspp0000273

Saucier, G., Iurino, K., & Thalmayer, A. G. (2020). Comparing predictive validity in a community sample: High-dimensionality and traditional domain-and-facet structures of personality variation. European Journal of Personality. https://doi.org/10.1002/per.2235

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→ environment effects. Child Development, 424–435. https://doi.org/10.2307/1129703

Schimmack, U. (2020). The Implicit Association Test: A Method in Search of a Construct: Perspectives on Psychological Science. https://doi.org/10.1177/1745691619863798

Schmeichel, B. J., & Vohs, K. (2009). Self-affirmation and self-control: Affirming core values counteracts ego depletion. Journal of Personality and Social Psychology, 96, 770–782. https://doi.org/10.1037/a0014635

Schmid, M. M., Gatica-Perez, D., Frauendorfer, D., Nguyen, L., & Choudhury, T. (2015). Social sensing for psychology: Automated interpersonal behavior assessment. Current Directions in Psychological Science, 24, 154–160.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martinez, V. (2007). The geographic distribution of big five personality traits—Patterns and profiles of human self-description across 56 nations. Journal of Cross-Cultural Psychology, 38, 173–212. https://doi.org/10.1177/0022022106297299

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. Journal of Personality and Social Psychology, 94, 168–182. http://dx.doi.org/10.1037/0022-3514.94.1.168

Seeboth, A., & Mõttus, R. (2018). Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions. European Journal of Personality, 32, 186–201. https://doi.org/10.1002/per.2147

Smaldino, P. E., Lukaszewski, A., von Rueden, C., & Gurven, M. (2019). Niche diversity can explain cross-cultural differences in personality structure. *Nature Human Behaviour, 3,* 1276–1283. https://doi.org/10.1038/s41562-019-0730-3

Sosnowska, J., Kuppens, P., Fruyt, F. D., & Hofmans, J. (2020). New Directions in the Conceptualization and Assessment of Personality—A Dynamic Systems Approach. European Journal of Personality. https://doi.org/10.1002/per.2233

Soto, C. J. (2019). How Replicable Are Links Between Personality Traits and Consequential Life Outcomes? The Life Outcomes of Personality Replication Project. Psychological Science, 30, 711–727. https://doi.org/10.1177/0956797619831612

Soubelet, A., & Salthouse, T. A. (2011). Influence of Social Desirability on Age Differences in Self-Reports of Mood and Personality. Journal of Personality, 79, 741–762. https://doi.org/10.1111/j.1467-6494.2011.00700.x

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Teije, A. t., & Balliet, D. (2020). The Cooperation Databank. https://doi.org/10.31234/osf.io/rveh3

Spearman, C. (1927). The abilities of man. Macmillan.

Sperry, R. W. (1966). Mind, brain, and humanist values. Bulletin of the Atomic Scientists, 22, 26. https://doi.org/10.1080/00963402.1966.11454956

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National Academy of Sciences, 117, 17680–17687.

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality Research and Assessment in the Era of Machine Learning. European Journal of Personality. https://doi.org/10.1002/per.2257

Surgeon General's Report (2004). *The Health Consequences of Smoking*. Retrieved from https://www.cdc.gov/tobacco/data_statistics/sgr/2004 on 14th October 2020.

Riemann, R., & Kandler, C. (2010). Construct validation using multitrait-multimethod-twin data: The case of a general factor of personality. European Journal of Personality, 24, 258–277.

Stieger, M., Wepfer, S., Regger, D., Kowatsch, T., Roberts, B. W., & Allemand, M. (2020). Becoming More Conscientious or More Open to Experience? Effects of a Two-Week Smartphone-Based Intervention for Personality Change. European Journal of Personality. https://doi.org/10.1002/per.2267

Tay, L., Woo, S. E., Hickman, L., & Saef, R. M. (2020). Psychometric and Validity Issues in Machine Learning Approaches to Personality Assessment: A Focus on Social Media Text Mining. European Journal of Personality. https://doi.org/10.1002/per.2290

Terracciano, A., Costa, P. T., & McCrae, R. R. (2006). Personality Plasticity After Age 30. Personality & Social Psychology Bulletin, 32, 999–1009. https://doi.org/10.1177/0146167206288599

Terracciano, A., McCrae, R. R., Brant, L. J., & Costa, P. T., Jr. (2005). Hierarchical linear modeling analyses of the NEO-PI-R scales in the Baltimore Longitudinal Study of Aging. Psychology and Aging, 20, 493–506. https://doi.org/10.1037/0882-7974.20.3.493

Thielmann, I., & Hilbig, B. E. (2019). Nomological consistency: A comprehensive test of the equivalence of different trait indicators for the same constructs. Journal of Personality, 87, 715–730. https://doi.org/10.1111/jopy.12428

Turkheimer, E., Pettersson, E., & Horn, E. E. (2014). A phenotypic null hypothesis for the genetics of personality. Annual Review of Psychology, 65, 515–540. https://doi.org/10.1146/annurev-psych-113011-143752

Vachon, D. D., Lynam, D. R., Widiger, T. A., Miller, J. D., McCrae, R. R., & Costa, P. T. (2013). Basic Traits Predict the Prevalence of Personality Disorder Across the Life Span: The Example of Psychopathy. Psychological Science, 24, 698–705. https://doi.org/10.1177/0956797612460249

Vainik, U., Misic, B., Zeighami, Y., Michaud, A., Mõttus, R., & Dagher, A. (2019). Obesity has limited behavioural overlap with addiction and psychiatric phenotypes. Nature Human Behaviour. https://doi.org/10.1038/s41562-019-0752-x

Vainik, U., Mõttus, R., Allik, J., Esko, T., & Realo, A. (2015). Are trait-outcome associations caused by scales or particular items? Example analysis of personality facets and BMI. European Journal of Personality, 29, 622–634. https://doi.org/10.1002/per.2009

Vainik, U., Dagher, A., Realo, A., Colodro-Conde, L., Mortensen, E. L., Jang, K., Juko, A., Kandler, C., Sørensen, T. I. A., & Mõttus, R. (2019). Personality-obesity associations are driven by narrow traits: A meta-analysis. Obesity Reviews, 20, 1121–1131. https://doi.org/10.1111/obr.12856

van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. Psychological Review, 113, 842861. https://doi.org/10.1037/0033-295X.113.4.842

Vazire, S. (2006). Informant reports: A cheap, fast, and easy method for personality assessment. Journal of Research in Personality, 40, 472–481. https://doi.org/10.1016/j.jrp.2005.03.003

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

Wendt, L. P., Wright, A. G. C., Pilkonis, P. A., Woods, W. C., Denissen, J. J. A., Kühnel, A., & Zimmermann, J. (2020). Indicators of Affect Dynamics: Structure, Reliability, and Personality Correlates. European Journal of Personality. https://doi.org/10.1002/per.2277

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2020). Differential associations of knowing and liking with accuracy and positivity bias in person perception. Journal of Personality and Social Psychology, 118, 149–171. https://doi.org/10.1037/pspp0000218

Wessels, N. M., Zimmermann, J., & Leising, D. (2020). Who Knows Best What the Next Year Will Hold for You? The Validity of Direct and Personality-based Predictions of Future Life Experiences Across Different Perceivers. European Journal of Personality. https://doi.org/10.1002/per.2293

Weston, S. J., Gladstone, J. J., Graham, E. K., Mroczek, D. K., & Condon, D. M. (2019). Who are the scrooges? Personality predictors of holiday spending. Social Psychological and Personality Science, 10, 775-782.

Wiernik, B. M., Ones, D. S., Marlin, B. M., Giordano, C., Dilchert, S., Mercado, B. K., Stanek, K. C., Birkland, A., Wang, Y., Ellis, B., Yazar, Y., Kostal, J. W., Kumar, S., Hnat, T., Ertin, E., Sano, A., Ganesan, D. K., Choudhoury, T., & al'Absi, M. (2020). Using Mobile Sensors to Study Personality Dynamics. European Journal of Psychological Assessment. https://doi.org/10.1027/1015-5759/a000576

Wilt, J., & Revelle, W. (2015). Affect, Behaviour, Cognition and Desire in the Big Five: An Analysis of Item Content and Structure. European Journal of Personality, 29, 478–497. https://doi.org/10.1002/per.2002

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics, 46, 1173–1186. Scopus. https://doi.org/10.1038/ng.3097

Wood, D., & Brumbaugh, C. C. (2009). Using revealed mate preferences to evaluate market force and differential preference explanations for mate selection. Journal of Personality and Social Psychology, 96, 1226–1244.

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can

explain trait covariation. Psychological Review, 122, 84–11.

Wood, D., Nye, C. D., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive trait markers from the English lexicon. Journal of Research in Personality, 44, 258–272. https://doi.org/10.1016/j.jrp.2010.02.003

Wood, D., Spain, S. M., Monroe, B. M., & Harms, P. D. (in press). Using functional fields to represent accounts of the psychological processes that produce actions. In J. F. Rauthmann , Handbook of Personality Dynamics and Processes. San Diego, CA: Academic Press.

Wood, D., & Wortman, J. (2012). Trait Means and Desirabilities as Artifactual and Real Sources of Differential Stability of Personality Traits. Journal of Personality, 80, 665–701. https://doi.org/10.1111/j.1467-6494.2011.00740.x

Woods, W.C., Arizmendi, C., Gates, K.M., Stepp, S.D., Pilkonis, P.A., & Wright, A.G.C. (2020). Personalized models of psychopathology as contextualized dynamic processes: An example from individuals with borderline personality disorder. Journal of Consulting and Clinical Psychology, 88, 240-254. https://psyarxiv.com/amdu8/

Wright, A.G.C., Gates, K.M., Arizmendi, C., Lane, S.T., Woods, W.C., & Edershile, E.A. (2019). Focusing personality assessment on the person: Modeling general, shared, and person specific processes in personality and psychopathology. Psychological Assessment, 32, 502-515. https://osf.io/nf5me/

Wright, A. G., Creswell, K. G., Flory, J. D., Muldoon, M. F., & Manuck, S. B. (2019). Neurobiological functioning and the personality-trait hierarchy: Central serotonergic responsivity and the stability metatrait. Psychological Science, 30, 1413-1423

Wright, A.G.C. & Zimmermann, J. (2019). Applied ambulatory assessment: Integrating idiographic and nomothetic principles of measurement. Psychological Assessment, 31, 1467-1480. https://psyarxiv.com/6qc5x/

Wrzus, C., & Mehl, M. (2015). Lab and/or field? Measuring personality processes and their social consequences. European Journal of Personality, 29, 250–271.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. Journal of Research in Personality, 44, 180–198. https://doi.org/10.1016/j.jrp.2010.01.002

Yarkoni, T. (2020). Implicit realism impedes progress in psychology: Comment on Fried (2020). https://doi.org/10.31234/osf.io/xj5uq

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12, 1100–1122. https://doi.org/10.1177/1745691617693393

Zheng, et al. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics, 33, 272-279.

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

Zimmermann, J., Woods, W. C., Ritter, S., Happel, M., Masuhr, O., Jaeger, U., Spitzer, C., & Wright, A. G. C. (2019). Integrating structure and dynamics in personality assessment: First steps toward the development and validation of a personality dynamics diary. Psychological Assessment, 31, 516–531. https://doi.org/10.1037/pas0000625