Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes

René Mõttus *

Department of Psychology and Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh

Institute of Psychology, University of Tartu

Timothy C. Bates

Department of Psychology and Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh

David M. Condon

Department of Medical Social Sciences, Northwestern University

Daniel K. Mroczek

Department of Psychology and Weinberg College of Arts & Sciences, Northwestern University

Department of Medical Social Sciences, Northwestern University

William R. Revelle

Department of Psychology, Northwestern University

* Corresponding Author:
7 George Square,
EH8 9JZ, Edinburgh, UK
rene.mottus@ed.ac.uk

Authors' Note

Data collection and preparation of this manuscript were supported by the University of Tartu (SP1GVARENG) and by institutional research funding (IUT2-13) from the Estonian Ministry of Education and Science. The assistance of Anu Realo, Jüri Allik, Andres Metspalu and Tõnu Esko in collecting the data is gratefully acknowledged.

Abstract

Among the main topics of individual differences research is the associations of personality traits with life outcomes. Relying on recent advances of personality conceptualizations and drawing parallels with genetics, we propose that representing these associations with individual questionnaire items (markers of personality "nuances") can provide incremental value for predicting and explaining them—often even without further data collection. For illustration, we show that item-based models trained to predict ten outcomes out-predicted models based on Five-Factor Model (FFM) domains or facets in independent participants, with median proportions of explained variance being 9.7% (item-based models), 4.2% (domain-based models) and 5.9% (facet-based models). This was not due to item-outcome overlap. Instead, personality-outcome associations are often driven by dozens of specific characteristics, nuances. Outlining item-level correlations helps to better understand why personality is linked with particular outcomes and opens entirely new research avenues—at almost no additional cost.

Keywords: personality; items; predictive; validity; polygenic; nuances

Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes

A central question of personality research is how personality characteristics are related to life outcomes, defined as socially or personally important phenomena such as educational (Poropat, 2009) and occupational attainment (Judge, Rodell, Klinger, Simon, & Crawford, 2013), romantic relationships (Malouff, Thorsteinsson, Schutte, Bhullar, & Rooke, 2010), pro- and anti-social behavior (Jones, Miller, & Lynam, 2011), health (Goodwin & Friedman, 2006) and longevity (Graham et al., 2017). Building on recent advances in personality conceptualizations and drawing instructive parallels with developments in genetics, we argue that the links between personality and life outcomes are stronger than commonly estimated and yet more nuanced, often at least partly driven by numerous specific personality characteristics called nuances and represented with individual questionnaire items (McCrae & Mõttus, in press). Many of these associations may well be unexpected. We illustrate these ideas by linking educational level, Body Mass Index (BMI) and various health-related life-style aspects with personality characteristics at different levels of specificity (domains, facets, and items that represent nuances). Because the item-level representation of personality-outcome associations can also be based on already-existing data, any predictive and explanatory leverage it entails often comes at low additional cost.

A lesson from genetics

It is instructive to start with a brief review of how geneticists have attempted to unravel the etiology of complex (multiply determined) phenotypes (observed characteristics). Most early attempts were based on the so-called candidate gene approach, whereby hypotheses regarding specific single nucleotide polymorphisms (SNPs) potentially relevant for a given phenotype were tested; for example, variations in the serotonin transporter (SLC6A4) gene were hypothesized to be associated with depression (Lesch et al., 1996). However, such candidate gene studies have yielded few replicable findings for behavioral phenotypes (e.g., Munafò & Flint, 2011). One response to this has been the pathway approach, wherein SNPs are aggregated according to some biological roles (e.g., membrane transport) or based on broad categories such as being expressed in the central nervous system (Wang, Li, & Bucan, 2007), but this strategy has often met only limited success (Hill et al., 2014).

The second response, genome-wide association studies (GWAS), has been entirely different in being atheoretical, grounded in inductive reasoning. Many (often millions) of genetic markers (SNPs) placed across the genome are analyzed for their association with a given phenotype without any *a priori* hypotheses. This has required large samples to control for an inflated false positive rate but has allowed researchers to identify thousands of genetic variants linked with phenotypes whose genetic etiology had previously been largely intractable, including schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), BMI (Locke et al., 2015) and education (Lee et al., 2018), among others. Individually, the identified genetic variants typically account for only a very small fraction of the heritable variance in these phenotypes, to the extent that the "top" SNPs often do not even replicate across studies (e.g., Luciano et al., 2018). However, the *combined* additive effect of hundreds or thousands of genetic variants in explaining complex traits can be substantial; for example, up to about 10 or 15% for personality constructs (e.g., Lo et al., 2017; Luciano et al., 2018) and even more for intelligence (Davies et al., 2018). Moreover, the aggregate association patterns of large numbers of SNPs are highly consistent across samples (e.g., Lo et al., 2017). The realization that many phenotypes are related to a myriad of genetic variants with each conferring only a tiny effect is now known as the fourth law of behavior genetics (Chabris et al., 2015). Such polygeneity itself has been one of the most informative findings.

In hindsight, the associations often appear obvious (e.g., many education-related SNPs are related to aspects of brain development; Lee et al., 2018), although they would have been unlikely to be identified based on theory alone. However, many associations may not immediately make sense, which is equally telling regarding the complexity of the etiology of the phenotypes. For example, among the (relatively few) SNPs associated with Neuroticism in multiple studies, several had not been previously associated with any other relevant phenotypes and/or have very general biological (e.g., metabolism) or neurodevelopmental (e.g., neurogenesis) functions, and one had been previously associated with morningness chronotype (Luciano et al., 2018)—a characteristic not obvious in relation to Neuroticism. Therefore, atheoretical approaches may often not only help to pinpoint specific associations but also tell us something more fundamental about the phenomena of interest—that they are etiologically highly multifaceted and not necessarily in line with our intuitions.

We argue that similar reasoning may benefit personality-outcome research. Hypotheses-driven research is useful for the development and refinement of specific theoretical ideas, but it may be suboptimal for untangling potentially large numbers of nuanced relationships and, more generally, estimating the sheer scope of their complexity (cf. Yarkoni & Westfall, 2017). Also, hypothesis-driven and hypothesis-consistent findings may often appear trivial—we already suspected them. We may learn as much or even more when we explore beyond our intuitions. But is there something in personality that could serve as its numerous markers ("SNPs")?

Trait hierarchy and outcomes

Composite traits: Domains and facets

Personality-outcome associations have most often been explored by correlating each outcome with one or more composite domain-level personality constructs such as those of the Five-Factor Model (FFM; McCrae & John, 1992). Such investigations are framed by prior ideas regarding how personality characteristics can be reduced into composite traits and motivated by the assumption that these composites reflect latent mental structures that drive personality-outcome links. If so, the basic constituents of the domains (items) are of interest only as indicators of the latent structures; they should be inter-changeable and uninformative in their own right.

However, personality characteristics can be grouped into composite traits in different ways. For example, as most already broad FFM traits are at least moderately correlated (van der Linden, te Nijenhuis, & Bakker, 2010), they can be mixed together to form ever broader traits such as stability and plasticity (DeYoung, 2006). At the same time, there are subsets of more strongly correlated items within each of the FFM traits, suggesting that they can also be broken apart into narrower traits such as aspects (DeYoung, Quilty, & Peterson, 2007) or facets (Costa & McCrae, 1992). Representing personality as such a system of increasingly narrow traits is called the personality hierarchy (Eysenck, 1991). We currently have a very limited understanding of the psychological or neurobiological mechanisms that cause specific behavioral, cognitive and affective characteristics to coalesce and vary across individuals in the way they do. As a result, there are no strong theoretical reasons to *a priori* prefer one set of composite personality traits (e.g., one level of trait hierarchy) over others, essentially leaving researchers with a pragmatic choice. When simplicity is desired, FFM or even broader traits may be the most functional operationalizations for personality-outcome associations, whereas narrower traits such as facets tend to provide better value when prediction accuracy is valued (Judge et al., 2013).

Items as traits

But there is yet another layer of personality hierarchy below facets: specific behavioral, cognitive, motivational and affective patterns that can operationalized with individual questionnaire items such as “I enjoy parties” or “I leave my belongings around” (McCrae, 2015). These specific characteristics, called nuances (McCrae, 2015; McCrae & Mõttus, in press) or habitual behavior (Eysenck, 1991), display trait-like properties of stability over time and agreement between different raters, and most of them also show unique etiology in terms of heritable variance and distinct developmental patterns

(Eaves & Eynseck, 1976; Neale, Rushton, & Fulker, 1986; Mõttus, McCrae, Allik, & Realo, 2014; Mõttus et al., 2015; Mõttus, Kandler, Bleidorn, Riemann, & McCrae, 2017). This applies even when variance due to higher-order traits (e.g., FFM domains and facets) is removed from the items, suggesting that they contain *unique* information about how individuals differ from each other in stable, observable and heritable ways; these findings tend to replicate across samples from different countries (Mõttus, Sinick et al., 2018). Also, age differences in personality characteristics may be to a substantial degree driven by the unique variance in items—nuances—rather than the FFM domains and facets (Mõttus & Rozgonjuk, under review). This suggests that nuances have heterogeneous etiology. According to the calculations of McCrae (2015), about two thirds of the error-free variance of a typical FFM questionnaire item is unique to it, reflecting a nuances, and only one third pertains to the FFM domains and facets.

Given that personality is most typically operationalized with questionnaires, items are the lowest operational level of the personality hierarchy. Representing relatively distinct traits of their own (nuances), items then constitute the most fundamental building blocks of personality, which we can call the *persome*. Just as each individual can be characterized by a complete and unique DNA sequence (their genome) or proteins (the proteosome) they can also be characterized by a unique pattern of personality characteristics. Of course, the genome-persome and SNP-item parallel should not be taken literally because the first represents a fixed set of physical structures and the latter does not; yet items are real in representing real behavioural, affective, cognitive and motivational differences between people. But the parallel is instructive in that SNPs tagged in DNA arrays are used as markers of genetic variance similarly to how items can be seen as markers of personality variance.

Not all items of commonly used questionnaires need to capture unique information—a unique nuance. Some of them (e.g., “I am a worrier”) may effectively equate constructs that have been labeled facets. Items of any questionnaire that has been carefully refined to provide efficient measurement of domain- or facet-level constructs may primarily do exactly what they were designed to do. And because test items tend to be selected to maximize the *common variance* purported to a pre-conceived set of composite traits (researchers are rewarded for high internal consistency of their scales), those that contribute unique information are likely underpopulated in existing omnibus personality inventories (Mõttus, Kandler, et al., 2017). That is, if items could be conceived of as markers placed across the persome, the extent to which existing questionnaires cover it ought to be limited. Of course, the variance of many could-be-measured items might be tagged by other items that do not directly map their content but reflect something that is probabilistically related to their content, similarly to how SNPs in DNA arrays also capture the genetic variants in linkage disequilibrium (LD) with the measured ones. There is no physical closeness for items (an important reason for LD in DNA), but there may be other reasons for their linkage such as direct causal connections between them (Cramer et al., 2012) or common situational cues.

Items and outcomes

Most items are thus more than mere indicators of the composites they are designed to operationalize (Mõttus, 2016; Mõttus, Sinick et al., 2018). Could their unique variance also enhance the predictive strength of personality and perhaps even help elucidate explanatory models of outcomes?

As one hypothesis, item-level analyses could only confer incremental value for specific (narrow) outcomes hinging on a particular set of narrow personality characteristics that happen to have corresponding items included in the personality test or are, at least, in linkage with the characteristics covered by the test. If so, item-level analyses could provide little predictive and explanatory value for broad outcomes such as life-satisfaction (Mõttus, Kandler, et al., 2017). Instead, as broad outcomes themselves aggregate accumulating effects of multiple behaviors, cognitions, feelings and motivations, they could be more strongly linked with higher-order composite traits that entail similar aggregation on the personality side (cf. Wittmann, 1988). If so, one would expect that the incremental predictive value resulting from item-level analyses is small or even non-existent for, say, educational attainment (a very broad life-outcome), whereas it might be larger for more specific outcomes such as liking Harry Potter films, preferring a particular type of food, or habitually smoking, provided that relevant psychological characteristics are tagged with the items included in the test.

Alternatively, item-level analyses may confer a predictive advantage regardless of the breadth of the outcome. Paralleling the fourth law of behavior genetics, this would indicate that outcomes tend to be linked with a multitude of specific behaviors, thoughts, feelings and motivations reflected in test items, either directly or by means of their "linkage" with each other, rather than with the broader underlying structures that these items ostensibly measure. If so, aggregating items into composites could mask their individual effects and, with large enough samples, any increases in reliability resulting from aggregation (Wittmann, 1988) would no longer outweigh this loss of substantive information (Goldberg, 1972).

In possibly one of the first studies to focus on items in relation to criteria, Weiss and colleagues (2013) linked items of the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1940) with mortality. The authors called their approach "questionnaire-wide association studies" (QWAS), explicitly re-casting the concept of GWAS. The QWAS is an apt parallel: GWAS atheoretically link individual markers of genetic variance, placed across the genome, with given phenotypes, whereas QWAS link individual markers of personality variance (items), placed across the persome, with outcomes. However, Weiss and colleagues' (2013) focus on items was simply a consequence of how personality self-reports had been historically collected in this particular study – MMPI does not map directly onto currently popular measurement frameworks such as the FFM – than a deliberate choice to represent personality-mortality associations in such a nuanced way. Few studies to date have systematically tested the usefulness of QWAS, or even discussed its advantages and disadvantages in relation to traditional, broad trait-based representations of personality-outcome associations.

In what may be the only relevant study, Seeboth and Mõttus (2018) assessed the predictive power of 50 personality questionnaire items for 40 different criterion variables, systematically comparing it to the predictive power of the FFM domains. For almost all outcomes, item-based models "trained" in one sample partition out-predicted the FFM-based models in an independent sample partition, with an average of 30% of more variance explained. The incremental predictive value did not depend on the breadth/specificity of the criteria as rated by independent judges. What is more, the unique variance in items also tended to drive the predictive power of the FFM domains: systematically dropping the most predictive items from the domain scores notably reduced their predictive power (more than expected by reduced reliability), whereas residualizing items for the domain scores only marginally reduced their predictive power. Seeboth and Mõttus (2018) could not consider the possibility of unmeasured facets rather than items *per se* driving the incremental predictive power of items. As another major limitation, the 50 items contained a substantial amount of redundancy by often being very similar and therefore provided a limited coverage of the persome. Plausibly, more comprehensive item-pools could out-predict the FFM domains to a larger degree. More research is needed.

In fact, items out-predicting domains and possibly even facets has also implications beyond prediction *per se*, revealing the *architecture* of personality-outcome links. Specifically, this suggests that the associations are driven by narrow characteristics aggregated into composite traits rather than the latent traits that these composites purport to measure. This very conclusion would be no less valuable than documenting the specific associations between particular personality traits and particular outcomes. It is only possible to test this possibility, however, if we link outcomes with numerous specific traits.

Polytrait scores

The realization that complex phenotypes are associated with thousands of specific genetic variants with each having only a small effect has lead quantitative geneticists to develop the method of polygenic scoring. Polygenic scores estimate individuals' additive genetic propensities for a phenotype by “recycling” the results of a previous GWAS (i.e., training sample) in the target sample. For *each* SNP that meets the inclusion criteria, the effect size of the selected “effect” allele (in relation to the phenotype in the training sample) is multiplied by the count of this allele for every individual in the target sample, with the sums of these products across all SNPs constituting individuals' polygenic scores (e.g., Purcell et al., 2009). Polygenic scores can be based on all available SNPs or only selections of them; scores that include more SNPs, even those with statistically non-significant associations with the phenotype in the training sample, tend to increase predictive accuracy (Dudbridge, 2013), underscoring the vastly polygenic nature of the phenotypes. Such polygenic scores, reflecting the joint predictive power of thousands of genetic variants, can often explain significant proportions of variance in the phenotypes they are created for (e.g., around 10% for general intelligence and education; Lee et al., 2018).

Exactly as GWAS summaries are widely used for creating polygenic scores, personality characteristic profiles can be turned into *polytrait scores*. These are sums of all available personality characteristics weighted by their (independent) contributions to the outcome at hand (obtained from an independent sample), possibly even regardless of whether the individual weights are statistically significant according to conventional criteria. To the extent that causal interpretations ought to be evoked, one could think of the polytrait scores as latent personality propensities for their respective outcomes. If it is the QWAS rather than the composite traits that provide the best out-sample prediction of the distinct aspects of outcomes, item-based polytrait scores (*polyitem scores*) may prove most useful. If so, outcomes could be claimed to be polycausal in relation to personality.

For example, polyitem scores can be used for practical purposes such as predicting job applicants' future performance and making selection/promotion decisions accordingly, or predicting children's later school-performance based on their personality characteristics measured at earlier grades and identifying those at risk of poorer performance for possible help. Similarly, the manner in which an individual responds to a health diagnosis (e.g., good treatment adherence, denial of the diagnosis) may be predicted by diverse item-level traits that are not easily delineated by the FFM domains, making polyitem scores better predictors of these clinically highly relevant outcomes. Using such polyitem scores to achieve incremental predictive value over other personality-based prediction models may thus have material and social benefits. Even when an outcome has not been measured in a given sample but it has a known item profile and corresponding item ratings in other samples, the predicted outcome values can be used as proxies for subsequent analyses; to the extent that polyitem scores improve prediction, they are incrementally helpful over more traditional ways of obtaining predicted values.

Besides polygenic scores, polyitem scores also resemble the idea of empirical personality scales (designed to maximize prediction rather than measure pre-conceived traits) implemented in instruments such the MMPI and the California Personality Inventory (CPI; Gough, 1975). However, unlike these unit-weight sum-scored scales, polyitem scores can weigh constituents (very) differently depending on the particular outcomes at hand so as to maximize the predictive accuracy. Of course, nothing prevents polyitem scores from being unit-weighted but just containing different combinations of predictors for each outcome, if and when this proves most useful (in this case, weights for non-selected predictors are 0 and for selected predictors 1). But even then the polyitem scores are "one-off" scales, being tailored to a specific outcome, whereas empirical scales are multi-purpose.

Methodological considerations

Measurement error. It is a common wisdom that single items are unreliable. But short-term (one to two weeks) retest reliabilities of single personality test items average around .65 and only somewhat tip below this estimate for longer retest intervals (Mõttus, Sinick et al., 2018). This may provide a reasonably good estimate of single item reliability, although some researchers have suggested that it may be higher (> .70; footnote 1 of Wood et al., 2018). This level of reliability is not disastrous, although certainly not ideal. However, a few things can further alleviate the concerns over single item (un)reliability.

First, it is important to realize that even though the prediction models may be trained based on the associations of individual items with outcomes, predictions (polytrait scores) are aggregates in which individual errors tend to cancel out; polyitem scores are multi-item scales, too. Second, reduced reliability can be compensated with large sample sizes; that only small samples used to be available has perhaps been one of the reasons why psychologist have been trained to seek aggregation in the first place (Goldberg, 1972). Now, large samples are regularly available. Third, despite being often rewarded, high internal consistency achieved by combining several similar items into a scale is actually a poor predictor of scales' validity—much worse than retest reliability (McCrae et al., 2011). Fourth, the extent to which items out-predict aggregate scales is in itself evidence that they are not notoriously unreliable.

One powerful solution to not only reduce random but also systematic measurement errors is to combine multiple raters such as the self and an informant; arguably, systematic method effects are even bigger a threat to results than random measurement error (McCrae, 2015; McCrae & Mõttus, in press). Combining raters is probably more useful in the model training phase than in their application for prediction, which reduces the cost of the design. Another way of dealing with measurement error is having the same items rated on several occasions—aggregating multiple measurements could reduce random measurement error as well as state-specific variance in ratings. Although combining raters and/or measurement occasions is preferable to combining multiple items, aggregating items may help reduce the ratio of random measurement error to total variance even for the measurement of nuances—so long as this does not involve aggregating *across* nuances, which would merely re-create what we have already in the form of facets.

Model training. The predictive models must be trained and validated in independent samples to guard against models with more parameters outperforming more parsimonious models due to over-fitting (Yarkoni & Westfall, 2017; Seeboth & Mõttus, 2018). Also, one can use regularization techniques such as least absolute shrinkage and selection operator (LASSO) or elastic net (Zou & Hastie, 2005; Tibshirani, 2011; Waldmann et al., 2017) for training models as these are designed to deal with large numbers of inter-correlated predictors and yield more parsimonious (compared to more traditional approaches) models in which many coefficients are effectively shrunk to zero. The extent to which different methods of obtaining weights for creating polyitem scores are useful is something that merits dedicated empirical scrutiny. Regardless of weighting, such prediction models can be based on just a few or hundreds of inter-correlated predictors (domains, facets or items). Importantly, this approach of using training and validation models in independent samples literally estimates the *predictive* value of personality characteristics rather than their correlations with outcomes, which are sometimes misleadingly called predictions (e.g., Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007) and are subject to over-fitting.

An empirical illustration

We used a large sample of Estonians (N = 3,561) who had completed the the NEO Personality Inventory-3 (NEO-PI-3; McCrae & Costa, 2010), a 240-item instrument that assesses the five FFM domains and their 30 facets, to compare the predictive powers of FFM domains, facets and items for ten disparate outcomes. The sample is described in Mõttus and colleagues (2017). The NEO-PI-3 is similar to the Revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992), items of which have been shown to contain substantial degrees of unique variance and therefore constitute a useful pool of nuances. For example, on average, more than 60% of the genetic variance in NEO-PI-R items is unique to them, not shared with the variance of any facet or FFM domain: median reliability-corrected heritability estimates of raw item scores and items' unique variances were .42 and .28, respectively (Mõttus, Sinick et al., 2018). Likewise, of the stable variance in NEO-PI-R items, over 75% is their unique variance (Mõttus, Sinick et al., 2018). For most participants ($N = 3{,}241$), an informant (typically partner/spouse, friend or parent/child) had also completed the observer-form of the NEO-PI-3, allowing us to use the averaged self- and informant-ratings in addition to personality self-ratings alone.

Among the ten outcomes that we selected, education was arguably the broadest, reflecting the cumulative effects of numerous choices and life circumstances; it has previously been linked with several FFM facets (Mõttus et al., 2017) and, almost to the same degree, with each of the FFM domains (Damian et al., 2015). Another relatively broad outcome was BMI, an important risk marker for a range of health conditions (Mokdad et al., 2003): it is likely to reflect the accumulation of numerous lifestyle choices over extended periods of time. The bivariate personality trait-BMI associations are generally weak in size and known to be not aligned with how the characteristics map into FFM (Mõttus, Kandler, et al., 2017; Sutin, Ferrucci, Zonderman, & Terracciano, 2011; Vainik, Mõttus, Allik, Esko, & Realo, 2015; Mõttus et al., 2018). Other outcomes represented potentially health-related lifestyle choices: the numbers of hours spent on exercising and walking, the quantity of alcohol (units) consumed within the last year, the frequencies of drinking alcohol, soft drinks and eating sweets and vegetables (or fruits) and whether people had ever been habitual smokers or not. Further details on data are reported in the Supplemental Online Material (URL MASKED).

We trained the models using a regularized (elastic net) regression in 75% of the sample and validated them in the remaining 25% of the sample, repeating the procedure 500 times for each outcome in random splits of the sample. This sample split was chosen to provide the training phase with more statistical power than the validation phase, as the latter only involved calculating correlations between predictions and observed values of the target outcomes. The model training was based on 50-fold cross-validation and the shrinkage parameter allowing for the smallest cross-validation error was selected. As demonstrated by Seeboth and Mõttus (2018) using a simulation, these procedures efficiently guard against over-fitting whereby models with more arguments inevitably out-predict those with fewer parameters: that is, the procedure ensured that if the associations were driven by latent traits underlying items, trait-based models would almost always out-predict item-based models. Age and gender were used as co-variates in model training but not validation (i.e., associations controlled for the demographic background, but predictions were not inflated by its contributions).

Predictive accuracy

Self-reports alone. Table S1 (Supplemental Online Material, URL MASKED) shows the prediction strengths for each model, averaged across the 500 random splits of the sample into training and validation subsamples. With the exception of alcohol quantity consumed within the last year, facet models outperformed domains in their predictive strength (on average, facet-based prediction strength was 6.15 times that of domain-based prediction), whereas items invariably outperformed domains (on average, item-based prediction strength was 9.51 times that of domain-based prediction). Except for the frequency of eating sweet, item models also outperformed facet models (on average, item-based prediction strength was 1.47 times that of facet-based prediction). On average, domain- and facet-based models accounted for 3.77% (*Mdn* = 4.10%) and 6.51% (*Mdn* = 5.60%) of outcome variance, respectively, whereas item-level models accounted for an average of 9.37% (*Mdn* = 8.05%). Therefore, item-based analyses at least doubled the average out-of-sample predictive accuracy compared to domains and yielded, on average, around 40% higher predictive accuracy than facets.

Combined self- and informant-reports. Table 1 shows the prediction strengths for each model based on combined self- and informant-reports (for the majority of participants for whom both types of ratings were available). Compared to self-reports based analyses, the average predictive accuracy increased for item-based models by 16%, but remained comparable for domain- and facet-based models. Facet models out-predicted domains for all outcomes but the alcohol quantity (on average, facet-based prediction strength was 6.00 times that of domain-based prediction, whereas the median of the prediction accuracy increases was 1.63 times) and items always outperformed domains (on average, item-based prediction strength was 10.43 times that of domain-based prediction, *Mdn* = 2.11). Except for the frequency of eating sweets, item models also outperformed facet models (on average, item-based prediction strength was 1.82 times that of facet-based prediction, *Mdn* = 1.53). On average, domain- and facet-based models accounted for 3.67% (*Mdn* = 4.20%) and 6.38% (*Mdn* = 5.90%) of outcome variance, respectively, whereas item-level models accounted for 10.48% (*Mdn* = 9.70%) of variance. The item-level predictions were strongest for educational level (26.58%) and BMI (14.13%) and the lowest for the hours spent walking (2.63%) and frequency of eating sweets (3.15%). Therefore, combining self-reports with informant-ratings allowed item-based analyses to at least double the average out-of-sample predictive accuracy compared to domains and to provide, on average, over 50% higher predictive accuracy than facets. Recall, this came for free as the information was captured in the already collected data![1]

[1] Note that items' unique variance also contributed to facets' and domains' abilities to predict outcomes in the first place. Had the domains and facets been measured with alternative items that reflected them to the same degree but did not have unique associations with the outcomes, the domains and facets would have had weaker links with these outcomes (Seeboth & Mõttus, 2018). Therefore, the degrees to which items out-predicted facets and domains as traits that ostensibly exist and could be measured independently of particular items were likely underestimates.

Table 1 also shows the predictive strength of item-models after residualizing them for the 30 facets (and thereby all FFM domains) using linear regression (the item being residualized was excluded from the facet score at the time; see Mõttus, Kander et al., 2017 for a similar procedure). The extent to which this reduced items' predictive accuracy is a straightforward measure of how much it was driven by the FFM domains and facets. For several outcomes, the unique variance in items allowed for substantial predictive accuracy: the item residuals-based models even out-predicted facet-based models for education, BMI, alcohol units and smoking history (for them, residualizing items for facets and domains reduced their predictive power by 10%, 22%, 62% and 41%, respectively). The item residual-based models did not allow for the prediction of the frequencies of eating sweets and walking and residualizing also substantially (about 70% or more) reduced items' predictive power for the frequencies of consuming soft drinks and vegetables. For some outcomes, thus, the associations are mostly driven by the unique variance that items capture (nuances), for some less so, and for some not at all. Outcomes differ in their personality-related architecture, which is informative in its own right.

Subsequent analyses were based on combined self- and informant-ratings of personality.

The architecture of personality-outcome intersections

The predictions were driven by multiple items: when the elastic net models were run in the whole sample, from 37 (frequency of eating vegetables) to 143 (education) items had non-zero regression weights (i.e., they tended to uniquely contribute to polynuance scores and thereby the prediction of outcomes), with a average of 89 items (*Mdn* = 91). Generally, what could be seen as narrower outcomes were predicted by somewhat but not substantially fewer items (smoking history 116 items, alcohol units 96, alcohol consumption frequency 87) than arguably broader outcomes (education 143 items, BMI 108 items).

There was no evidence for items offering less incremental predictive value than domains or facets for the arguably broadest outcome, education: it was worst predicted by domains, with the prediction more than doubling based on facets and improving by further 59% based on items. Outcomes for which item-models more than doubled the predictive power of facets included smoking history and the quantity of alcohol consumption; they more than doubled domain-level predictions also for education, BMI, and frequencies of walking and soft drink consumption. Importantly, there are no items in the NEO-PI-3 that in any direct way refer to smoking or alcohol use, suggesting that the incremental predictive power of items for these rather specific outcomes was *not* driven by item-outcome content overlap. Direct content overlap was also unlikely for the frequencies of walking, exercising and consuming soft drinks and vegetables as well as for educational level (see also Table 2). Two NEO-PI-3 items referring to over-eating (from the N5: Impulsiveness facet) were among the main drivers of personality-BMI associations, suggesting content overlap; this finding has been documented previously (e.g., Vainik et al., 2015). However, when we dropped the contributions of these two Impulsiveness items from the item-prediction models in the validation sub-sample, the model still explained a substantial proportion (on average 8.13%) of variance in BMI—just as much as facet-based models that did rely on the contributions of these items.

To further illustrate a) why items predicted more variance than domains and facets in some outcomes than in others and b) how the drivers of the predictions were distributed across the domain / facet spectrum of the NEO-PI-3, we created “Manhattan” plots depicting item’s correlations with six of the ten outcomes. These outcomes were smoking history and alcohol units for which items conferred one of the highest incremental value over domains and facets; BMI for which associations were partly driven by selected items; education and frequency of having soft drinks for which items conferred an average level of incremental predictive value over facets (but their overall predictability varied), and frequency of eating sweets for which items conferred no incremental value over facets. The outcomes were residualized for age and sex in using a generalized linear model either with Gaussian (education, BMI, sweets and soft drink consumption frequencies), Poisson (alcohol quantity) or logit (smoking history) link. The associations were grouped according to the FFM domains and facets (Figure 1). Note, however, that these are zero-order correlations, whereas elastic net models had estimated conditional associations (controlling for all other predictors as well as age and gender).

If the associations were driven by domains (see top-left panel of Figure 1 for a hypothetical scenario), correlations of the items of all facets of the domains should have had roughly comparable values, barring variability due to error and the degrees to which the items reflected the domain; this was almost never the case (middle and bottom panels of Figure 1). If associations were driven by facets (see top-right panel of Figure 1 for an hypothetical scenario), correlations of the items within these facets should have had roughly similar values. Items' correlations with educational level, but also with smoking history and frequency of having soft drinks varied both between and within facets of the same FFM domains, resulting in both facets and items conferring incremental predictive value over domains. For alcohol units, facets of the same domains were somewhat more consistent in their average item-associations, but items within facets still varied in their correlations. Despite the overall predictability of these outcomes from personality being quite different, for all of them the *associations were scattered across the spectrum of personality traits* captured by the NEO-PI-3. This suggests that the intersections of personality with these outcomes are both driven by specific characteristics and yet wide-spread, despite educational level being a very broad phenomena but others referring to more specific behaviors. For BMI, although the predictive power was notably driven by two Impulsiveness items, there was further variability between and within facets of the same FFM domains. This is why even without these two items, item-models could make sufficient incremental contributions to the prediction of BMI, as was shown above and has also been shown by Mõttus, Sinick and colleagues (2018). For the frequency of eating sweets, there just were not many sufficiently sizeable associations to begin with.

This demonstrates how item-level analyses do not only confer incremental predictive value, but also illuminate *how* a particular outcome is intersecting with personality. For some outcomes, it is a broad spectrum of narrow characteristics that drives the associations (e.g., education, smoking history or frequency of having soft drinks), even though the overall magnitudes of the associations can vary across the outcomes. In other instances, the main drives can be fewer, efficiently summarized using facets (e.g., frequency of eating sweets). And sometimes such analyses can identify where the associations are inflated by predictor-outcome overlaps (e.g., BMI). This can tell us something about the general architecture of personality-outcome associations, potentially providing insight that may be no less valuable than meticulously documenting which specific constructs intersect with which specific outcomes to which specific degrees. This insight could not be provided by linking outcomes with only the FFM domains or even with their facets.

Examples of specific associations

Table 2 also reports the "top 10" strongest item-level correlates for each outcome, adjusted for age and gender; the items are paraphrased as in Mõttus and colleagues (2018)[2]. Many of the associations make immediate sense. For example, to the extent that NEO-PI-3 items cover relevant nuances, the "top 10" items portray an educated person as somebody who likes mental challenges, is curious and interested in new hobbies, and is expected to take lead; a person with higher BMI as somebody who eats too much, cannot resist carvings, fails with self-improvements and has trouble with keeping things organized; a person who spends time on exercising as someone who is busy, disciplined and organized and carries through with self-improvements (among the top associations form the elastic net models were also items referring to liking excitement and rivalry); and a person who consumes

[2] Across the 240 items, correlations with outcomes were moderately similar to betas from the elastic net models (for the 10 outcomes, Spearman correlations ranging from .25 to .76, *Mdn* = .52, all $p < .001$). The differences is that betas were conditional on the contributions of all other items, but correlations were not. The top 10 betas for each outcome partly overlapped with the top 10 correlations.

greater quantities of alcohol as someone who likes everything social (among the top associations form the elastic net models were also items referring to being reckless and sarcastic and not being concerned about future). But less self-evident associations may prove more useful for generating new insight and hypotheses, especially if replicated in future studies. For example, higher education was linked with not believing that people are after each other (one of the top elastic net associations also referred to trusting others) and higher BMI with being talkative (top elastic net associations also referred to taking lead but also to low self-esteem). A particularly interesting pattern was almost the opposite item-level correlates between the frequency and quantity of alcohol use; those who drink often but in lower quantities are in many social and non-social ways inhibited whereas those who drink less often but in greater quantities may do this due to social reasons. Of course, cross-sample replication is required for such associations; it has already been undertaken for BMI and the effects tend to replicated to a moderate extent across a variety of cultural backgrounds and sample demographics (Mõttus, Sinick et al., 2018).[3]

Final remarks

The bottom-up approach does not preclude aggregation

The described QWAS approach does not deny the existence or utility of taxonomic models of higher-order aggregate traits. Instead, it defies the (often tacit) expectation that *all* information in items pertains to these high-order constructs. Moreover, QWAS does not preclude subsequent aggregation and theorizing around it. For example, numerous genetic variants associated with complex traits may appear to cluster into fewer biological systems (Wood et al., 2014), potentially helping to elucidate the biological pathways of these traits. In the same way, the outcome-specific sets of items may appear to form meaningful clusters, identifiable either psychometrically (Weiss et al., 2013) or conceptually. These items may not be correlated themselves, in which case their clusters could not emerge from a

[3] Elastic net regression coefficients for all items in relation to the ten outcomes are given in the Supplemental Online Material (URL MASKED), as are all item-outcome correlations. Among other things, this allows researchers to calculate polyitem scores for these outcomes in independent NEO-PI-3 datasets.

factor-analytic top-down approach at all. For example, based on the items of the Revised NEO Personality Questionnaire (Costa & McCrae, 1992) it would appear possible that among the strongest correlates of BMI are individual items that refer to eating too much, trying various foods, giving up on self-improvements, being riled by others or having conservative moral principles (Mõttus, Kandler, et al., 2018). Although these items fall into different traits in the FFM on which the questionnaire was based, they may form a conceptually meaningful aggregate in relation to BMI. The strongest associations given in Table 2 provide similar information.

New research avenues

That personality-outcome associations are often largely driven by narrow personality characteristics may seem like a bad news at first glance—there is so much to describe and explain! But we see this as an opportunity for new kinds of research questions. Specifically, this opens the possibility to investigate systematic differences between traits in how they are linked with outcomes, with both practical and theoretical benefits. Practically, this allows identifying the subsets of traits that are most strongly linked with particular outcome domains and use these for prediction (e.g., for hiring purposes of for identifying people at risk for poor academic or health outcomes); essentially, only measuring traits that can be most usefully included in polytrait scores. This could literally allow for more (prediction) with less (items).

Theoretically, a multi-dimensional representation of personality-outcome associations allows for addressing entirely new kinds of research questions. For example, we could examine the extent to which predictive validity is a general property of traits, with some traits more likely to intersect with any kind of life outcome than others. This is plausible as desirable traits generally tend to go with desirable outcomes. Possibly, it is the traits showing the strongest rank-order stability and heritability (e.g., Mõttus et al., 20018) that that also tend to have the strongest links with outcomes as these traits are most likely to constantly pull people in certain life trajectories, allowing outcomes to build. Alternatively, we may examine the properties of traits that have strongest links with particular outcome domains. For example, we might expect that traits with strongest cross-informant agreement are most strongly linked with social outcomes, because these are traits for which accurate person-perception is particularly important and consensus on them may also facilitate achieving the outcomes (e.g., via smoother co-operation). For another example, if health outcomes tend to be predicted by traits representing affect rather behaviour (Wilt & Revelle, 2015), this will illuminate the mechanisms by which personality and health are linked (i.e., affective regulation vs life style). Such research questions are difficult to address with just, say, five traits. We rarely do empirical studies with a sample of just five people.

Sampling of items

An important limitation of using items to represent personality-outcome associations is exactly the limitation of representing the associations using higher-order traits: they depend on which items happened to be included in the particular personality measure (Mõttus, 2016). Had different items representing different nuances been included in the NEO-PI-3, item-outcome and thereby trait-outcome associations could have been different and, as a result, the polytrait scores could also have been different. We do not know yet how different. One way to test the robustness of polyitem scores would be to compare them based on different questionnaires.

DNA arrays used in GWAS studies attempt to place markers throughout the genome in order to capture as much genetic variance as possible (information about additional markers is often imputed based on available information even though the ultimate goal is complete sequencing). In the same way, QWAS studies should strive to base their explorations on the most comprehensive item pools possible. Ideally, these should not be constrained by any pre-existing structural personality model. An example of such an atheoretical set of items is the International Personality Item Pool (IPIP; Goldberg, 1999) that now contains more than 3,500 items.

The most prominent challenge of dealing with item sets this large is data collection. One means for addressing this challenge is to use planned missingness designs like synthetic aperture personality assessment (SAPA; Revelle et al., 2016). Using data collection platforms like the SAPA-Project (Condon, 2017; sapa-project.org), it is possible to derive synthetic correlation matrices by administering random subsets of very large item pools to survey participants. Over time, the empirical associations between many individual items and a host of behavioral outcomes can be identified. While it is unlikely that each of the 3,500 IPIP items conveys unique information for each outcome (in fact, some may not be informative for any outcome), this approach provides a very thorough coverage of the persome. With enough data, it may eventually be possible to identify a *parsimonious* subset of items that predicts a subset of important outcomes better than existing factor and facet models.

Reporting item-level raw data

If, as we argue, item-level analyses confer substantial additional predictive value, reliably detecting this will require large samples—another lesson learned in genetics. It is a common practice in GWAS studies to aggregate samples, often using harmonized or linked measures of the outcomes (e.g., Lee et al., 2018). Also, GWAS studies often make their findings publicly available to facilitate collaborative efforts (e.g., the LD Hub; ldsc.broadinstitute.org). Similarly, personality researchers should begin publishing item-level raw data and outcomes, as this will facilitate the identification of item-level associations and creating polyitem scores across multiple studies. At least, item-level association profiles should be published, so that they could be meta-analyzed into ever more precise estimates. Examples of such datasets include publicly-available data from the Eugene-Springfield Community Sample (Goldberg, 2008) and the SAPA-Project (Condon, Roney, & Revelle, 2017).

Conclusion

Sometimes, broad composite traits such as the FFM domains or narrower composite traits such as aspects or facets may constitute the most instrumental levels of the personality hierarchy (Judge et al., 2013). Oftentimes, however, even narrower characteristics such as nuances, conveniently represented by single items, may provide the best value for both predictions of outcomes and meaningful descriptions of their intersections with personality. We argue that the most appropriate level of personality hierarchy for representing how personality intersects with a given outcome needs to and indeed can be empirically tested rather than assumed to be always the same (e.g., the FFM domains). There is little value in discussing a life outcome in relation to a broad FFM domain when their association is only driven by a narrow subset of traits subsumed under the domain (Mõttus, 2016).

Personality researchers deservedly take great pride in documenting correlations between personality traits and important real-world phenomena. Robust findings such as high Conscientiousness being associated with most socially valued outcomes and, conversely, Neuroticism being correlated with most apparently maladaptive outcomes have been and continue to be among the most important achievements of personality research. But ways of presenting these associations more accurately should be of general interest, especially when no additional data collection is required. A more detailed approach cannot only substantially increase the predictive power of personality characteristics but also provide insights into the general mechanisms by which personality intersects with life. Moreover, when the predictive value of personality characteristics can be more than doubled at no additional cost, their practical utility should increase dramatically. Why would anyone want to ignore these opportunities?

References

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

Condon, D. M., Roney, E., & Revelle, W. (2017). A SAPA Project Update: On the Structure of phrased Self-Report Personality Items. *Journal of Open Psychology Data, 5(1).*

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

Cramer, A. O. J., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., … 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

Damian, R. I., Su, R., Shanahan, M., Trautwein, U., & Roberts, B. W. (2015). Can personality traits and intelligence compensate for background disadvantage? Predicting status attainment in adulthood. *Journal of Personality and Social Psychology*, *109*, 473–489. https://doi.org/10.1037/pspp0000024

Davies, G., Lam, M., Harris, S. E., Trampush, J. W., Luciano, M., Hill, W. D., … Deary, I. J. (2018). Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature Communications*, *9*, 2098.

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., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, *93*(5), 880–896. https://doi.org/10.1037/0022-3514.93.5.880

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLOS Genet*, *9*, e1003348. https://doi.org/10.1371/journal.pgen.1003348

Eaves, L., & Eysenck, H. (1976). Genetic and environmental components of inconsistency and unrepeatability in twins' responses to a neuroticism questionnaire. *Behavior Genetics, 6(2),* 145-160.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?—Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*, 773–790. https://doi.org/10.1016/0191-8869(91)90144-Z

Friedman, J., Hastie, T., Simon, N., & Tibshirani, R. (2016). glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models (Version 2.0-5). Retrieved from https://cran.r-project.org/web/packages/glmnet/index.html

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs*.

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 (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press.

Goldberg, L. R. (2008). The Eugene-Springfield community sample: Information available from the research participants. *Oregon Research Institute Technical Report, 48(1)*.

Gough, H. G. (1975). *Manual for the California Psychological Inventory*. Consulting Psychologists Press, Palo Alto, CA.

Graham, E.K., Rutsohn, J.P., Turiano, N.A., Bendayan, R., Batterham, P., Gerstorf, D., Katz, M., Reynolds, C., Schoenhofen, E., Yoneda, T., Bastarache, E., Elleman, L., Zelinski, E.M., Johansson, B., Kuh, D., Barnes, L.L., Bennett, D., Deeg, D., Lipton, R., Pedersen, N., Piccinin, A., Spiro, A., Muniz-Terrera, G., Willis, S., Schaie, K.W., Roan, C., Herd, P., Hofer, S.M., & Mroczek, D.K. (2017). Personality predicts mortality risk: An integrative analysis of 15 international longitudinal studies. *Journal of Research in Personality, 70,* 174-186.

Hathaway, S. R. and McKinley, J. C. (1940). A Multiphasic Personality Schedule (Minnesota) : I. Construction of the Schedule. *The Journal of Psychology, 10*(2):249–254.

Hill, W. D., Leeuw, C. de, Davies, G., Liewald, D. C. M., Payton, A., Craig, L. C. A., … Deary, I. J. (2014). Functional gene group analysis indicates no role for heterotrimeric g proteins in cognitive ability. *PLOS ONE*, *9*, e91690. https://doi.org/10.1371/journal.pone.0091690

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, *39*, 329–337. https://doi.org/10.1016/j.jcrimjus.2011.03.004

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *The Journal of Applied Psychology*, *98*, 875–925. https://doi.org/10.1037/a0033901

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., … Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science, 274*, 1527–1531.

Lo, M.-T., Hinds, D. A., Tung, J. Y., Franz, C., Fan, C.-C., Wang, Y., … 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

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*, 197–206. https://doi.org/10.1038/nature14177.

Malouff, J. M., Thorsteinsson, E. B., Schutte, N. S., Bhullar, N., & Rooke, S. E. (2010). The Five-Factor Model of personality and relationship satisfaction of intimate partners: A meta-analysis. *Journal of Research in Personality*, *44*, 124–127. https://doi.org/10.1016/j.jrp.2009.09.004

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., & Costa, P. T. (2010). *NEO Inventories professional manual*. Odessa, FL: Psychological Assessment Resources.

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., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and social psychology review, 15(1),* 28-50.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., & Marks, J. S. (2003). Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, *289*, 76–79.

Mõttus, R. (2016). Towards more rigorous personality trait–outcome research. *European Journal of Personality*, *30*(4), 292–303. https://doi.org/10.1002/per.2041

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

Mõttus, R., Marioni, R., & Deary, I. J. (2017). Markers of psychological differences and social and health inequalities: Possible genetic and phenotypic overlaps. *Journal of Personality*, *85*, 104–117. https://doi.org/10.1111/jopy.12220

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality*, *52*, 47–54. https://doi.org/10.1016/j.jrp.2014.07.005

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., & Rozgonjuk, D. (under review). Devel(opment) is in the details: Investigating age differences in the Big Five domains, facets and nuances with machine learning. https://doi.org/10.13140/RG.2.2.29038.87360

Mõttus, R., Sinick, J., Terracciano, A., Hrebícková, M., Kandler, C., Ando, J., … Colodro-Conde, L. (2018). 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*.

Munafò, M. R., & Flint, J. (2011). Dissecting the genetic architecture of human personality. *Trends in Cognitive Sciences*, *15*, 395–400. https://doi.org/10.1016/j.tics.2011.07.007

Neale, M. C., Rushton, J. P., & Fulker, D. W. (1986). Heritability of item responses on the Eysenck personality questionnaire. *Personality and Individual Differences, 7(6),* 771-779.

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135(2)*, 322–338. https://doi.org/10.1037/a0014996

Purcell, S., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., … Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, *460*, 748–752. https://doi.org/10.1038/nature08185

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research (Version 1.8.12). Retrieved from 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. In Fielding, N. G., Lee, R. M., & Blank, G., (Eds.), *Handbook of Online Research Methods*. Thousand Oaks, CA: Sage Publications.

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

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.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, *511*, 421–427. https://doi.org/10.1038/nature13595

Sutin, A. R., Ferrucci, L., Zonderman, A. B., & Terracciano, A. (2011). Personality and obesity across the adult life span. *Journal of Personality and Social Psychology*, *101*(3), 579–592. https://doi.org/10.1037/a0024286

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society*, *73*, 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.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

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, *44*, 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

Waldmann, P., Mészáros, G., Gredler, B., Fürst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00270

Wang, K., Li, M., & Bucan, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, *81*, 1278–1283. https://doi.org/10.1086/522374

Weiss, A., Gale, C. R., Batty, G. D., & Deary, I. J. (2013). A questionnaire-wide association study of personality and mortality: the Vietnam Experience Study. *Journal of Psychosomatic Research*, *74*, 523–529. https://doi.org/10.1016/j.jpsychores.2013.02.010

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.

Wittmann, W. W. (1988). Multivariate reliability theory. Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 506–560). New York: Plenum Press.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., … 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. https://doi.org/10.1038/ng.3097

Wood, D., Qiu, L., Lu, J., Lin, H., & Tov, W. (2018). Adjusting Bilingual Ratings by Retest Reliability Improves Estimation of Translation Quality. *Journal of Cross-Cultural Psychology, 49(9*), 1325-1339.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12, 1100-1122.*

Figure Captions

Figure 1. *"Manhattan" plots of the associations of two hypothetical and six observed outcomes with 240 NEO-PI-R items. The dots represent correlations, controlling for age and gender. Associations are grouped according to the FFM domains and their facets. Dashed lines represent significance levels (with Bonferroni correction for multiple testing). The plots were drawn with the manhattan function from the psych R package (Revelle, 2018).*

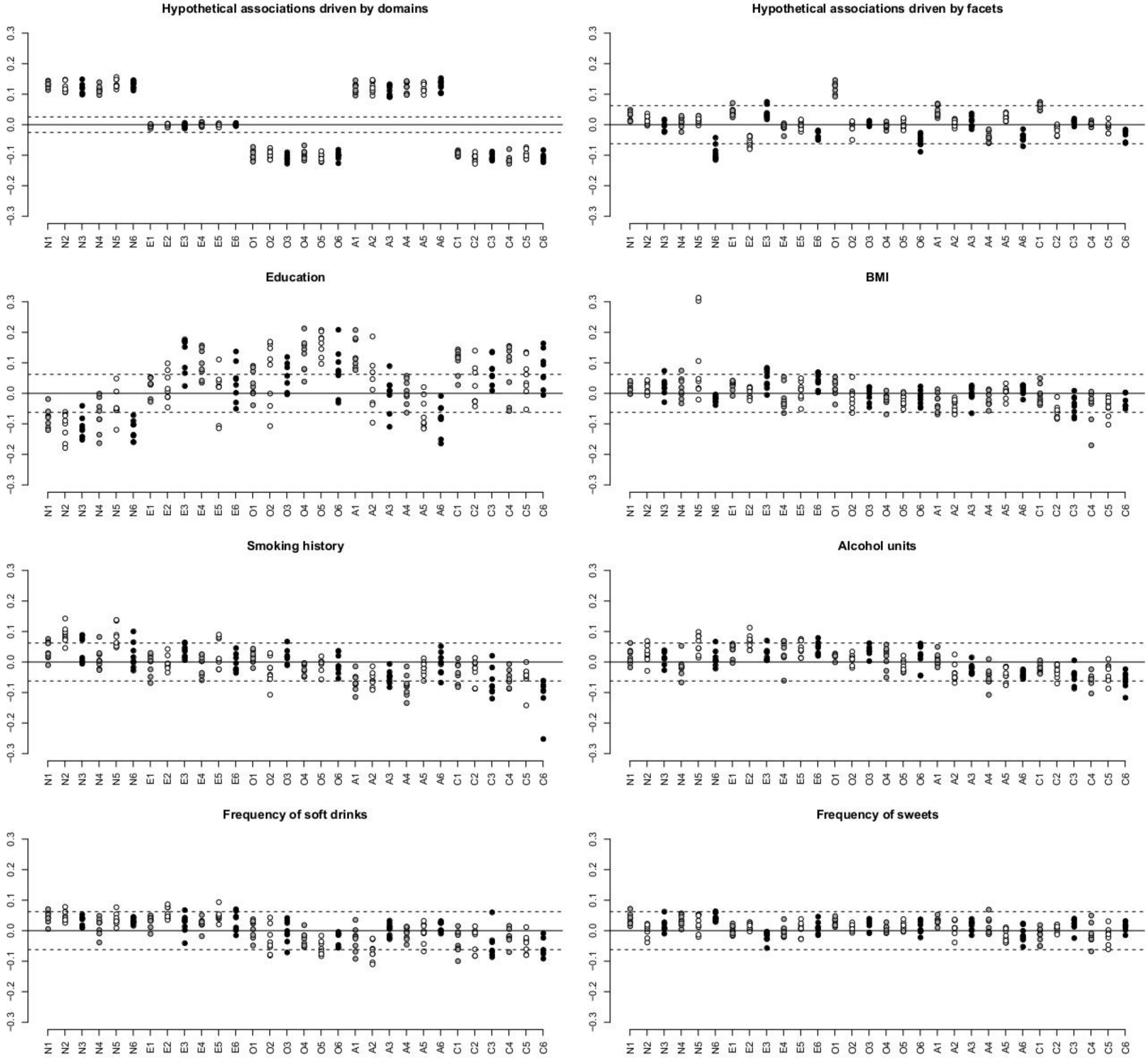

Table 1. *Predictive strengths of models using personality domains, facets, and items (combined self- and informant-reports).*

	Domain-models		Facet-models		Item- models		Item- models (residualized)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Education	.067	.015	.167	.020	.266	.022	.239	.022
BMI	.002	.002	.079	.017	.141	.020	.110	.018
Exercise	.046	.011	.061	.012	.089	.014	.030	.009
Walking	.002	.002	.016	.007	.026	.009	.008	.005
Alcohol frequency	.061	.017	.069	.019	.101	.021	.033	.012
Alcohol units	.056	.019	.026	.013	.112	.028	.042	.018
Soft drinks	.038	.012	.057	.015	.084	.018	.011	.006
Vegetables	.055	.014	.085	.017	.104	.019	.021	.009
Sweets	.024	.009	.041	.012	.032	.011	.002	.002
No smoking history	.016	.006	.037	.010	.093	.015	.055	.012

NOTE: The predictive strength is the squared correlation between the respective outcome’s predicted and observed values in the validation sample. Mean: average estimate across 500 permutations; SD: standard deviation of the estimates across 500 permutations (dividing it by √500 = 22.4 entails standard error of the mean) . Residualized = items were residualized for the scores of all FFM domains and facets.

Table 2. *Ten items with the strongest unique association in elastic net models.*

Education	Body Mass Index	Exercising hours	Walking hours	Alcohol frequency	Alcohol units	Soft drinks frequency	Vegetable frequency	Sweets frequency	Smoking history
Likes mental challenges	Eats excessively	Carries through with self-improvements	Has bustling imagination	Likes crowded parties (-)	Likes crowded parties	Avoids tricking people into things (-)	Tries different foods	Is ambitious in everything (-)	Has avoided being reckless (-)
Enjoys puzzles	Overeats favorite foods	Is always on the go	Doesn't think that people are after (-) each other	Tries different foods (-)	Enjoys social events	Likes crowded parties	Is interested in new hobbies	Finds it hard to fight back	Avoids sarcasm (-)
Has curiosity about many things	Carries through with self-improvements (-)	Likes attending games	Enjoys letting fantasies develop	Wants action (-)	Is merry	Finds philosophy interesting (-)	Has curiosity about many things	Is unable to self-manage in a crisis	Finds it hard to resist temptations
Doesn't think that people are after each other	Cannot resist cravings	Likes expressive dance	Is emotionally attached to friends	Enjoys social events (-)	Prefers company	Acts impromptu (-)	Is joyful	Feels anxiety	Has sometimes felt unbearably ashamed
Is interested in new hobbies	Is very disciplined (-)	Has curiosity about many things	Trusts others’ intentions (-)	Prefers company (-)	Likes jobs that require working with others	Likes to be surrounded by people	Is interested in patterns	Gets easily disheartened and gives up	Is agitated
Tolerates controversial ideas	Keeps possessions tidy (-)	Is very disciplined	Doesn't worry that kinds acts have (-) ulterior meanings	Likes vacations with crowds (-)	Monitors his or her feelings	Likes showy styles	Likes expressive dance	Has poor judgement in difficult situations	Monitors his or her feelings
Is expected to take lead	Is well organized (-)	Does the talking in meetings	Loves talking to people	Likes jobs that require working with others (-)	Works quickly	Comes prepared (-)	Values aesthetics	Is commanding (-)	Works excessively (-)
Feels breaking down under stress (-)	Is meticulous (-)	Is meticulous	Is sometimes overwhelmed by joy	Works quickly (-)	Is no less energetic than others	Doesn't manipulate others (-)	Is usually in a rush	Fears embarrassing himself herself	Is seen as a precarious person
Does the talking in meetings	Handles tasks diligently (-)	Uses extremely positive words to describe things	Thinks others are better than him her	Feels comfortable in crowds (-)	Likes vacations with crowds	Doesn't think that people are after each other (-)	Lives a fast-moving life	Is easily frightened	Is merry
Feels comfortable being a leader	Is the most talkative person in conversations	Is well organized	Doesn't consider him herself better than others	Likes garish destinations (-)	Likes garish destinations	Is emotionally sensitive to environments (-)	Is happy to change environment	Feels very embarrassed when teased or made fun of	Doesn't avoid daydreaming