European Journal of Personality, *Eur. J. Pers.* **30**: 304–340 (2016)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2060

OPEN PEER COMMENTARY

Causal Unity of Broader Traits is an Illusion

JENS B. ASENDORPF
Humboldt University, Germany
asendorpf@gmail.com

Abstract: Mõttus alerts us to the widespread predictive heterogeneity of different indicators of the same trait. This heterogeneity violates the assumption that traits have causal unity in their developmental antecedents and effects on outcomes. I would go a step further: broader traits are useful units for description and prediction but not for explaining personality development and personality effects. In most cases, the measured trait indicators are closer to relevant causal mechanisms, and within a network perspective on personality, broader traits as entities with causal potential can be dismissed completely. Copyright © 2016 European Association of Personality Psychology

Many personality psychologists share an unnecessary illusion: broader traits, operationalized by multiple indicators, are causal unities such that their indicators are exchangeable. It is assumed that all indicators of the same trait have similar developmental antecedents ("etiologically unitary"), and that they all show similar developmental trajectories and similar predictive relations to developmental outcomes. Within this view, the predictive heterogeneity of different items or facets of the same trait is a problem for explaining trait effects. If the observed predictive relations are not spurious, they have to be causally attributed at least in part to specifics of the indicators, not only to the underlying trait.

Developmental and predictive homogeneity of trait components may apply to very narrowly defined characteristics, but they do not apply to broad traits such as the Big Five. For example, both cross-sectional and longitudinal studies reported considerable inconsistency of age-related changes among NEO-PI-R facets of extraversion (Bleidorn, Kandler, Riemann, Angleitner, & Spinath, 2009; McCrae et al., 1999), and their facet-specific variance was estimated as significantly and predominantly genetic for five of the six facets, suggesting genetic heterogeneity among the facets (Kandler, Riemann, Spinath, & Angleitner, 2010). If the sources and developmental trajectories of different indicators of the same trait are heterogeneous, it comes as no surprise that the predictive strengths of different trait indicators for the same outcome are also heterogeneous even after controlling for different factor loadings of the indicators.

In my view, traits are useful units of analysis for description because they offer taxonomies of personality differences of reasonable complexity (e.g., the Big Five and their facets). Moreover, they are useful units for predicting short-term outcomes, developmental trajectories, and developmental outcomes. However, they are generally *not* useful for in explaining the observed predictive relations because they are too far away from the causal mechanisms in most cases. In the following, I briefly elaborate this argument.

PREDICTIVE HETEROGENEITY AND THE BANDWIDTH-FIDELITY TRADE-OFF

In prediction models, predictors and outcomes vary in breadth (complexity of information, bandwidth), and predictions of outcomes vary in predictive strength (accuracy, fidelity). Shannon and Weaver (1949) postulated a bandwidth-fidelity trade-off in communication methods: if bandwidth increases, fidelity decreases, and vice versa. This trade-off was introduced to educational psychology by Cronbach (1960) and later discussed at length in personnel psychology (e.g., Hogan & Roberts, 1996). Some consensus has been reached that the trade-off is (a) ubiquitous, and (b) a guideline for an optimal choice of predictors for a given outcome: Breadth of the predictors should match breadth of the outcome (the symmetry principle of prediction; Wittmann, 1988). Thus, broad traits are better for predicting broad outcomes, and narrow traits are better for predicting narrow outcomes. But this does not imply that broad traits are good at *explaining* broad outcomes!

AGGREGATION IN PREDICTION VERSUS EXPLANATION

Aggregation increases the reliability of a predictor by adding parallel items (Spearman–Brown formula). Often aggregation also increases predictive strength by adding items with similar predictive strength. However, these advantages to prediction come at a cost. The more you aggregate, the higher the chance that you bring in disparate causal units involved in explaining emergence of the outcome, and the more difficult it is to identify any of them specifically.

For an illustration, consider Lasky et al.'s (1959) study. They predicted the relapse rate of hospitalized psychiatric patients using their psychiatric diagnoses, judgments of the hospital staff, and volume of the patients' psychiatric records

(measured in inches of paper). Record volume was the best predictor. It is a highly aggregated variable that reflects the complexity and duration of a patient's problems, and is therefore useful for prediction. However, it is completely uninformative about the nature of the problems, and, thus, useless for explaining relapse (and silent about how to intervene to decrease the relapse rate).

Moving to trait effects, consider, for example, that conscientiousness predicts longevity (Friedman & Hudson, 2011). But how? Some components of conscientiousness may be involved in the causal chain from current personality to death, but probably to different degrees; other components may be irrelevant. Predictive heterogeneity at the facet or item level informs us about which components might be relevant for explanation, and therefore can be a useful guide in the search for explanations.

Both examples suggest that the symmetry principle for prediction may be somewhat misleadingly formulated. Critical is number of different mechanisms that cause the outcome, not complexity of the outcome measure. Relapse and longevity can be measured on a simple time scale but the measured outcome is complex and has heterogeneous causal mechanisms. The broad trait of conscientiousness is well suited for predicting longevity apparently because it captures many of these causal mechanisms, but what causes longevity is probably not the broad trait of conscientiousness. What causes longevity is those many different individual mechanisms. Conscientiousness apparently aggregates quite a few of them, thereby obscuring the contributions of each of them.

My conclusion is that broad traits are not useful for explanation. In many cases, the specific trait components (facets, items) are more informative because they are closer to the causal mechanisms underlying personality development and personality effects. Within a network perspective on personality that relies only on observed personality indicators and their causal relations (e.g., Schmittmann et al., 2013), broader traits can even be completely dismissed.

EXPLANATION: THE MAIN TASK OF PERSONALITY PSYCHOLOGY IN THE NEXT DECADES

During the past 50 years, personality psychology has made considerable progress concerning personality description, and prediction of and by personality. In contrast, explanation of personality development and personality effects has lagged far behind. In the coming decades, much more inspiration and transpiration are needed to change this unsatisfactory situation. I believe that focusing on broader traits as causal units leads us nowhere. Instead, we should focus on causal mechanisms that link one or a few components of broader traits to one or a few components of the outcomes, including components of the trait itself later in development.

Beware of Indirect Effects: Rigorous Definitions and Methods for Testing the Causality of Traits

ANNA BAUMERT[1,2], MANFRED SCHMITT[2] and GABRIELA BLUM[2]

[1]*University of Koblenz-Landau, Germany*
[2]*University of Western Australia*
baumert@uni-landau.de

Abstract: We argue that facet-level and item-level analyses cannot be used to estimate causal involvement of traits in outcomes. Employing a fictitious example, we explain that a trait can have several indirect effects with opposite signs. Thus, finding discordant associations of facets or items with an outcome does not necessarily imply that unique aspects of those facets/items rather than an underlying trait are the outcome's causes. We stress the need for definitions of traits that allow rigorous tests (experimental and correlational) that traits cause behaviors and outcomes.

Mõttus calls for facet-level and item-level analyses when scrutinizing associations between traits and outcomes. We support this request but disagree with a central aspect of his article. Mõttus argues that if a trait is a cause of an outcome, all facets and manifest indicators of that trait should be associated with the outcome in the same direction and to the extent that they load on the trait factor. This argument is based on the logic that correlations emerge among variables with a common cause. If variables (facets/indicators and outcome variables) do not have causal effects on each other but are affected by the same trait, they will be correlated. However, Mõttus neglects that in cases where a trait exerts indirect effects on an outcome, the argument no longer holds. Indirect effects that a trait has via (some of) its facets or manifest indicators do not have to be concordant.

To outline our disagreement with Mõttus, we use extraversion from Eysenck's hierarchical factor model of personality as an example and construct a possible yet fictitious scenario involving indirect effects (Figure 1). According to

DOI: 10.1002/per

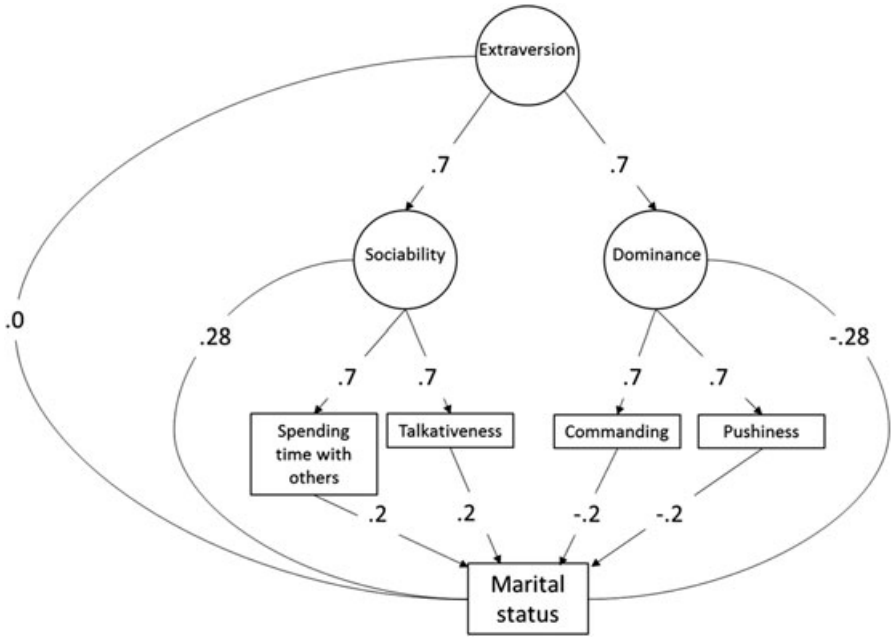

Figure 1. Fictitious example of opposing indirect effects of extraversion on marital status via sociability and dominance. Arrows denote standardized effects, curved lines denote correlations.

Eysenck's model, the type factor extraversion causally affects the trait factors dominance and sociability, among others, both of which exert causal influences on their representative behaviors. For example, dominance shapes commanding and pushy behavior; sociability shapes talkativeness and spending time with others. Suppose that talking a lot and spending time with others is helpful for finding a mate, thus having causal impacts on later marital status. In this fictitious scenario, extraversion has an indirect causal effect on the outcome marital status via the trait factor sociability and the behaviors talkativeness and spending time with others. Further, suppose that being commanding and pushy endanger long-term relationships, thus having negative causal impacts on later marital status. In other words, extraversion has an additional indirect causal effect on marital status via the trait factor dominance and commanding and pushy behaviors. The two indirect effects work in opposite directions, thus reducing the total effect of extraversion on the outcome. We claim that even if the total effect is zero, extraversion is still causally affecting marital status (via two indirect paths).

If one follows Mõttus' argument, finding that sociability is positively correlated and dominance is negatively correlated with likelihood of being married implies that unique aspects of dominance and sociability are causes of marital status but extraversion is not. However, this conclusion might not be correct. To scrutinize the viability of this conclusion, we need to decompose the facets into components that are shaped by extraversion and components that are not shaped by extraversion and therefore unique. Then we can test whether marital status is correlated with the unique components of sociability and dominance or with the extraversion-dependent components.

Going well beyond Mõttus' requests, to achieve a truly explanatory personality psychology, we call for conceptual clarity in what is actually meant by a personality trait. As Mõttus emphasizes, a trait can serve to explain behaviors only if it is not a mere summary term for these same behaviors. Therefore, we need a conceptual definition of a trait that does not rely on the behaviors the trait is assumed to cause.

The literature offers a few examples that approach conceptual clarity. Eysenck (1967) defined extraversion in terms of neural mechanisms by assuming that extraversion *consists of* differential responsiveness of the ascending reticular activating system (ARAS) to stimulation. Sensible definitions of traits also seem possible at levels of conceptualization other than the neural level. One might define extraversion in terms of social-cognitive mechanisms (e.g., Fleeson & Jayawickreme, 2015), for example, by proposing that extraversion *consists of* strength of associations between social cues and response patterns in memory.

Importantly, once committed to a conceptual definition, rigorous tests of causality of the trait for behavior become available. By systematically manipulating personality states, we can experimentally test the causality of a trait for certain behaviors. When conceptualizing trait extraversion as neural responsiveness to stimulation, state extraversion can be manipulated by raising (or lowering) the momentary level of activation in the ARAS. For example, ingesting stimulating substances might raise the momentary level of neural activation and, thus, induce a state of low extraversion. Finding effects of the manipulation on the likelihood of talking, spending time with others, and commanding and pushy behaviors will increase confidence that extraversion is a cause of these behaviors.

If one commits to a different conceptual definition, one will have to choose a different way to manipulate state extraversion. When conceptualizing extraversion

DOI: 10.1002/per

by relying on social-cognitive mechanisms, one could aim to manipulate momentary associations of social cues and response patterns by means of training procedures (e.g., Schnabel & Asendopf, 2015) and test their causal effects on behaviors such as talking or commanding.

After committing to a conceptual definition of a trait that does not rely on the behaviors that the trait is assumed to cause, we will be able to design measures of traits (and corresponding states) that do not overlap in content with measures of behaviors and outcomes. This is important for an experimental approach because the effectiveness of manipulations can be checked prior to testing their effects on behaviors. Moreover, this is important for correlational studies that can complement experiments and serve to test indirect effects of traits on outcomes via repeated behaviors.

In sum, we agree with Mõttus that it is sensible to conduct facet-level and item-level analyses when scrutinizing correlations between trait measures and outcome measures. These analyses are important for distinguishing between correlations due to item content overlap and psychologically informative correlations. However, inspecting the concordance of correlations of items/facets with outcomes cannot determine whether an underlying trait is causal. Mõttus fails to consider that a trait might exert opposing indirect effects. Most importantly, to scrutinize the causality of traits, empirical analyses (correlational or experimental) need to be preceded by theoretical elaboration on precise definitions of personality traits.

Composites Can Be Causal Too

RIET VAN BORK, MIJKE RHEMTULLA and DENNY BORSBOOM

University of Amsterdam
R.vanBork@uva.nl

Abstract: Mõttus gives the impression that composites, as well as other models in which traits are a result rather than a cause of their indicators, require "emergent properties" to have causal power. We argue that this is not necessary; composites can be considered causally relevant by themselves when they mediate the relation between their constituents and the outcome variable. Copyright © 2016 European Association of Personality Psychology

Mõttus describes a number of alternatives to reflective measurement models in personality. Examples are (a) the model proposed by McCrae (2015) in which traits are unions of their semi-autonomous constituents, (b) network models as proposed by Cramer et al. (2012), and (c) Wood, Gardner, and Harms's (2015) model, in which traits are formed by behaviors that covary due to shared functional values. Mõttus suggests that for such models in which semi-autonomous behaviors constitute a trait (rather than reflecting it), causal power is more accurately ascribed to the constituents rather than to the trait:

If [traits are artificial constructions], attributing causality to traits as such seems equally questionable regardless of whether their constituents have similar or different associations with the outcome at hand. Even if the associations generalize across trait constituents, causal interpretations may be more fruitfully based on these constituents rather than the summary-level traits (p. 21).

As an example of an "artificial construction", Mõttus gives socioeconomic status (SES): a composite of education level, income, occupational status and the quality of one's residence. (A composite is a function of its constituents, which completely determine it; an example is a sum score based on questionnaire items, which is completely determined by the item scores). Mõttus concludes that we should not interpret SES as a cause of its associated outcomes "because SES itself is then the *result* rather than the cause of its constituents and, unless it takes on emergent properties, it thereby owes its outcome correlations to these constituents." (p. 22). We disagree. We argue that composites can in fact have causal relevance over and above their indicators, and that this is a realistic possibility in the context of personality.

Introducing a composite as a cause of a particular outcome involves a conjunctive hypothesis; namely, it implies that scores on one constituent can make up for scores on any other. Consider as a small example two constituents: (1) the number of males on a train and (2) the number of females on the train, which together entirely determine the composite 'the number of people on the train'. It is entirely reasonable to conclude that the composite itself (rather than its constituents) causes the outcome variable 'the time it takes to find an empty seat on the train'. In this example, the composite itself is causal because it fully mediates the relation between the constituents and the outcome variable: If one knows the number of people on the train, the number of women on the train does not predict any additional variance in the outcome.

But does that mean that the composite variable 'number of people in the train' has emergent properties with respect to its constituents? That seems implausible. A composite can have causal force without having emergent properties in any interesting sense of the word.

DOI: 10.1002/per

In the example, the constituents show causal unity; they are linked with the outcome in a similar way. However, if the outcome were different (say, 'the number of high heels on the train'), the composite may no longer mediate the relation between 'the number of females on the train' and the outcome variable. Thus, a composite may screen off the relation between indicators and outcomes, but does not necessarily do so. With respect to the variable, 'the number of high heels', a constituent variable (i.e., the number of females on the train) may have a unique causal relation with the outcome.

Could personality traits function like people on a train? We think they could. All that is required is for the constituents to play compensatory roles with respect to the outcomes of interest. Consider the impulsiveness items, "I have trouble resisting my cravings", "When I am having my favourite foods, I tend to eat too much", and "I sometimes eat myself sick": it is not implausible that obesity could be caused by a high sum of these constituents, whether that sum is due entirely to any one, two, or a combination of all three items.

Mõttus argues convincingly that researchers must test whether trait-outcome relations are due to the unique influence of specific items and facets of the trait. We strongly agree and note that it is possible to perform such a test whether the trait is conceptualized as a common-cause latent variable or as a composite. In the former case, a structural equation model can be used to model the item-trait-outcome relations explicitly and examine unique effects. In the latter case, the composite variable must be defined independently of the outcome (e.g., by weighting all constituents equally; Howell, Breivik, & Wilcox, 2007). Figure 1 depicts what these two test models might look like for the Impulsiveness → BMI example from Terracciano et al. (2009).

Rather than dismissing composites as lacking causal power at the level of the trait, we think it is important to take seriously the possibility that personality traits may be the *result* of a set of behaviors. Composites can have causal relevance without being emergent, but constituents can also have unique causal force. As Mõttus argues persuasively, just like we should not *assume* that causality is at the level of the trait rather than the item or facet, but *demonstrate* this, we should do the same when considering the causal power of composites.

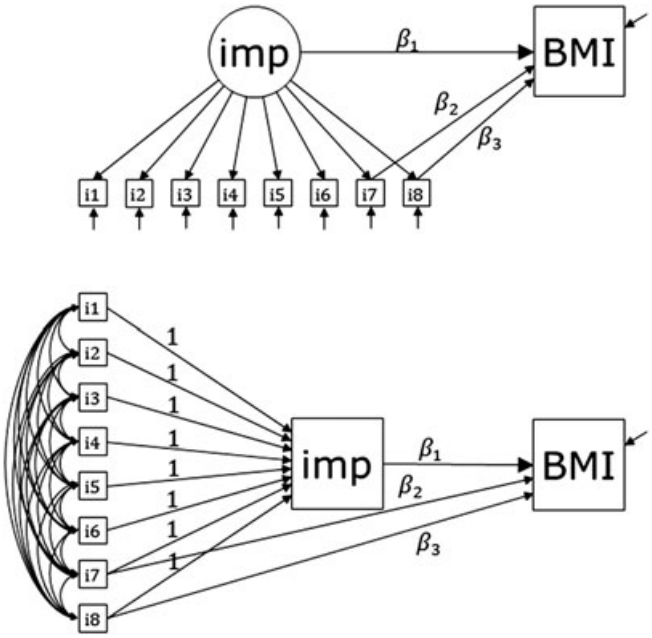

Figure 1. SEM models for testing item-outcome associations for impulsiveness and BMI. Upper panel: reflective latent variable model. Lower panel: composite model. Indicators i7 and i8 refer to eating-related behaviors. A significant β_1 coefficient would mean that the trait as a whole is related to BMI. β_2 and β_3 coefficients reflect unique effects of i7 and i8.

The Roles of Personality Traits in Health Outcomes

BENJAMIN P. CHAPMAN

Departments of Psychiatry and Public Health Sciences, University of Rochester Medical Center
ben_chapman@urmc.rochester.edu

Abstract: Broad domains of personality traits have organizational utility in grouping specific traits and can sometimes be effective predictors of health outcomes. However, theories about how and why personality affects differing aspects of health often require refinement. This refinement is best achieved by moving beyond broad, multifaceted personality constructs to their constituent subcomponents. Such specificity can also facilitate translational work parleying basic personality research into health intervention and prevention efforts. Multiple levels of analysis in the trait hierarchy are useful in the study of personality and health outcomes. Copyright © 2016 European Association of Personality Psychology

Mõttus provides a valuable contribution to the long history of calls for the decomposition of multifaceted scales to test theories (Carver, 1989) or maximize predictive accuracy (Mershon & Gorsuch, 1988). These

DOI: 10.1002/per

calls seem to have gone largely ignored, or at least underappreciated. One reason may be a drift from views of the Big Five as an ideal *structural taxonomy* for classifying specific traits, to an ideal *level of personality aggregation* for outcomes research. To be sure, compound traits (even broader than the Big Five in some cases) sometimes have strong associations with outcomes, particularly when the outcomes are multifaceted themselves (Ones, Chockalingham, & Dilchert, 2005). Nevertheless, I share Mõttus' concerns over "causal (dis)unity" in compound trait associations with outcomes. In health research, it is often helpful to isolate the "active ingredient(s)" of a compound trait, because some degree of specificity is usually necessary to translate findings into actionable prevention and intervention efforts (Chapman, Hampson, & Clarkson, 2014).

Philosophically, I believe the problems Mõttus describes extend even to scientific instrumentalist views, where compound traits are seen as heuristically useful "metaphors" describing some phenomena (Cacioppo, Semin, & Berntson, 2004). If it is unclear which elements of the phenomenon underlying responses to an "Extraversion" scale are actually linked to the outcome, the explanatory and heuristic utility justifying the instrumentalist invocation of the "Extraversion" concept is also lost. I suspect Mõttus' concerns also extend beyond causal assumptions to situations of causal agnosticism, in which traits may be markers or proxies of some other underlying cause. For instance, risk prediction models are proliferating in medicine and present opportunities to harness the predictive power of personality scales, since any form of data with actuarial power can help improve health outcome forecasts (Chapman et al., 2015). Prevention strategies suggested by such prediction models are to some degree guided by causality of the constituent factors, but often involve general strategies not dependent on causality. For instance, frequency of checkups and screenings might be increased to identify disease onset as soon as possible for early intervention (e.g., catching cancer at Stage 1 rather than Stage 4), or generic preventive measures initiated which are known to be effective regardless of what has caused elevated risk (e.g., initiating daily baby aspirin if one's cardiovascular risk score is high). Facets or pockets of items, rather than broad trait scales, may provide the most efficient predictive gains (Chapman et al., 2015).

Mõttus' considerations have interesting implications for increasingly popular brief Big Five scales (some as few as two items per factor), often utilized in epidemiologic studies. "Causal unity" demonstrations would involve only a few items, but as he points out, interpretation must be restrained to the actual scale content. If a Conscientiousness scale composed of the two trait adjectives "reliable" and "organized" shows an association with some inflammatory marker, for instance, interpretation is most safely centered on these particular trait adjectives. They are anchor items of a broader construct, of the other elements of which might be at play, we are simply not sure.

Item analysis can help sharpen interpretations despite aggrieving classical test theorists, whose domain sampling model demands large numbers of items for effective construct measurement. But in sociology and epidemiology, single items are often intended to be used alone and crafted carefully to capture a particular phenomenon. Sometimes, the object of measurement is narrow, like opinions about a specific political issue in the General Social Survey, but it can be broad as well. Health is a tremendously multi-faceted, complex construct, but the item "In general, how would you rate your health?" [poor, fair, good, excellent] is so predictive of mortality that it is often accorded the status of a health outcome in its own right (DeSalvo, Bloser, Reynolds, He, & Muntner, 2006). Could any personality items function similarly to the general health item, singly measuring some dispositional quality substantial enough to call a "narrow trait"?

Trait-descriptive adjective (TDA) measures would appear candidates, since by definition under the lexical hypothesis, each adjective describes a trait distinct enough to warrant a separate (English) word with distinct meaning. Of course there is covariance between them and a multivariate structure—Mõttus' concern is simply that only some of this structure, rather than all of it, may be relevant to a given outcome. Measurement error in such single items arises from idiosyncratic interpretations and differing reference group effects, among other things. Mõttus' own work on anchoring vignettes seems potentially useful in calibrating these out (Mõttus et al., 2012), and perhaps might aid single-item measures immune by definition to causal disunity.

The strategy he proposes of testing a single scale, reconstituted in several ways by omitting facets (or items), is an interesting expansion to one-at-a-time facet analysis. Scale permutations of this sort may differ in associations with outcomes, but also in their overlap with potential confounders and mediators and in internal consistency. Yet these latter sources of variation can also reflect meaningful impact of the deleted content, rather than statistical artifact. If formal hypothesis testing is conducted with scale permutations, those less sympathetic to Mõttus' concerns may bridle at the multiple comparisons. A False Discovery Rate might be incorporated to address this. One might also compare parameter point and variance estimates for the various scale permutations without formal hypothesis testing, or compare them on scales of evidence using Bayes Factors. There are surely many possibilities, and new approaches often encounter either neophobic dismissal or mechanical imitation. Let us hope that methodological development proceeds as carefully and thoughtfully as Mõttus' arguments. In the meantime, greater nuance in personality and health outcomes research is certainly an imperative and will profit from the considerations he forwards.

DOI: 10.1002/per

Madam, We Are Going to Have to Amputate Your Conscientiousness: Why Personality Traits Are Not (Necessarily) 'Inside' Us

D. ANGUS CLARK[1], C. EMILY DURBIN[1] and BRIAN M. HICKS[2]

[1]*Michigan State University*
[2]*University of Michigan*
cemilydurbin@gmail.com

Abstract: Mõttus' article raises compelling questions about extent to which self-report models of personality can capture with fidelity causal processes that link personality dimensions to life outcomes. However, we take issue with the conceptualization of personality traits put forward. We doubt that different approaches to analysis of self-report data will provide a clear demonstration of causality and argue that trait measures provide a means of describing, not explaining, individual differences in behavior. Copyright © 2016 European Association of Personality Psychology

We appreciate the opportunity to reflect on Mõttus' stimulating article, which emphasized thoughtful considerations of causality and the importance of examining personality at different levels of abstraction. However, we take issue with the conceptualization of personality traits put forward, and find the injunction that all studies include item/facet level analyses both impractical and unlikely to achieve the goal of demonstrating causal impacts of traits on outcomes.

In personality research, there is often a discrepancy between what investigators want traits to be and what traits (as commonly operationalized) actually represent. Personality traits are often granted status of within-person causal drivers of behavior. An individual *is* Conscientious; the high level of Conscientiousness explains *why* the person thoroughly proofreads the commentary before submitting. Mõttus references this notion of traits in which they both exist as real entities and are responsible for individuals' behavior (traits… "initiate the kinds of behaviors or biological processes…", "are 'in our skin'", " can be thought of as unobserved generators that…produce observable behaviors, thoughts, and feelings"). We argue that this conceptualization implies a causal reality of self (and other)-reported trait measures that can neither theoretically nor empirically be supported.

Mõttus focused his discussion on the Big 5, a framework established primarily via exploratory factor analysis (EFA). Big 5 traits are typically operationalized as sum-scores based on these EFAs of items and scales. EFA is a latent variable model that specifies causal relations running from latent factors to their indicators. Thus, the Big 5 traits are conceived as latent variables that cause their indicators. These causal relations can take one of two forms. The first is within-person – Cletus talks a lot at parties *because* he is Extraverted; his Extraversion leads him to talk a lot. The second is between-person – Cletus talks a lot at parties *because* someone at his level of Extraversion is more likely to talk a lot at parties than someone who ranks lower on this dimension. There are several comprehensive treatments showing that only the latter between-person causal interpretation of traits, not the former within-person interpretation, is justified (Borsboom, 2009; Borsboom, Mellenbergh, & van Heerden, 2003; Borsboom & Dolan, 2006; Cervone, 2005; Epstein, 1994; Molenaar, 2004; Pervin, 1994).

This is partly because the EFAs used to establish the Big 5 are between-person models, based on analysis of a correlation matrix that captures covariation between different individuals' responses to items at a single time point. The Big 5 could be treated as within-person "process" variables (i.e., generators) if they also characterized the structure of personality at the individual level (i.e., if factor analyses were run on individuals' patterns of behavior over time). Unfortunately, the Big 5 structure is about as unreliable within people as it is reliable between them (Hamaker, Nesselroade, & Molenaar, 2007; Molenaar, 2004). As it cannot be assumed that the Big 5 structure exists in every individual, the Big 5 cannot legitimately be used to explain individual behavior (Borsboom, 2009).

These arguments are not new, but we trait researchers often find it difficult not to overextend the explanatory power of our constructs. The elegance and replicability of structural models can encourage a confusion between description and explanation; organizing individual differences into dimensional models is *not* the same as studying why individuals do the things they do that give rise to the observed differences among them. Ability to use similar words to describe dimensions that co-vary in self-reports and purported neurobiological individual differences does not mean traits have any greater reality "inside of us" beyond their ability to paint a parsimonious picture of the broad ways in which persons differ. Trait constructs allow us to talk about people in predictable ways (including outcomes more common among some individuals than others), but they do not provide a scientifically defensible means of demonstrating mechanistic causality at the individual level.

Our second concern regards Mõttus' recommendations for exploring causal links between traits and outcomes. Testing for similarity in correlations between outcomes and specific items from trait questionnaires introduces several problems. First, it neglects the key principle of aggregation as a means of increasing reliability and construct fidelity. Psychometrically, single items lack important advantages that broader scales possess. Differences in associations between an outcome and various items could be attributable to many sources, including differential reliability or construct validity of the items. Second, there is not an infinite number

DOI: 10.1002/per

of items that could be employed to assess any trait. Third, following this suggestion would lead researchers to conduct many more analyses in each dataset (every item, every outcome), leading to the well-known problems inherent in multiple comparisons. The results of such comparisons could prove difficult to organize and interpret, thus providing minimal gain with considerable risk.

We are sympathetic to the issues Mõttus raises. Demonstrating causality has a privileged status in science, and the theoretical and practical questions central to personality science that could be advanced by use of causally informative techniques are important. However, we doubt that "better" approaches to analysis of between-subject self-report and informant-report instruments will advance this cause. To the extent that personality constructs reflect real processes, self-report and other-report measures are at best uncertain guideposts to some aspects of those processes rather than markers of precise "things" within us waiting to be "discovered". There is no mono-method route to demonstrating causality, only the painstaking work of scientists piecing together testable hypotheses about the great complexity of individual differences, interpreting data derived from different strategies, designs, and operationalizations, and placing this agenda within the broader framework of understanding what personality can tell us about other important constructs.

ACKNOWLEDGEMENTS

This work was supported by United States Public Health Service grants R01 DA034606 and DA039122 from the National Institute on Drug Abuse.

Time to Move Beyond the Big Five?

DAVID M. CONDON and DANIEL K. MROCZEK

Northwestern University
david-condon@northwestern.edu

Abstract: We agree with Mõttus' recommendations, and suggest it may be time to move beyond the Big Five. The issues Mõttus discusses are the result of a quest for simple structure, especially orthogonality, by personality scientists. This has obscured many important facets and items that do not cleanly fit the Big Five and occupy the interstitial spaces. This in turn has influenced efforts to understand how personality relates to important life outcomes such as health and longevity. Traits of narrower width may be necessary for maximal prediction, as Mõttus points out.

There is a curious technical detail in the history of personality assessment: the persistent theoretical yearning for simple structure. We believe that a considerable portion of Mõttus' article is grounded in observations that are consequences of this yearning. We applaud Mõttus' deft and clear review of the problematic nature of prediction with latent factor models, and we wholeheartedly agree with his recommendations.

Curiosity about simple structure is rooted in early factor analytic work, especially that of Thurstone (1947), who is often cited for laying out the criteria needed to describe a factor analytic solution as psychometrically simple. Bear with us through a summary of the technical details. The most interpretable factor solutions are those which are simplest – the items are close to the axes of each factor. Simplicity can be improved after the factors have been extracted by using matrix algebra to transform the item-level factor loadings. This is theoretically relevant because many test developers have rotated the factor loadings to force uncorrelated factors (orthogonal rotation) rather than to produce the most simple solution overall (oblique rotation). The two types of rotation will produce different rank orderings for the item-level loadings. Historically, orthogonality at the domain/factor level has been argued to improve validity, though most test developers forego orthogonality by subsequently using the highest-loading items for their scales (Saucier, 2002). Choosing the items with highest loadings increases internal consistency, a fact which psychometrically underlies Mõttus' theoretical comments regarding causal unity.

Evidence for the Big Five is not predicated on orthogonality, and several prominent researchers have been proponents of oblique rotations, including Cattell, Norman and occasionally Goldberg (1993). We view orthogonality as theoretically problematic for hierarchical personality inventories where the scores at one level are dependent on other levels. Administration of public-domain correlates of the NEO-PI-R to large online samples (Condon, 2014) produced five correlations between the factors/domains with magnitudes greater than | 0.2 | (95% CI: .16-.24) and 34 meaningful correlations ($r > .30$) between facets and their non-primary domains. Even when the factors were intended to be orthogonal, it seems that the facets – and presumably their constituent items – occupy the interstitial spaces.

How can we reconcile this circumstance with Mõttus' observations and recommendations? For one, it suggests that his facet-level recommendations may need to be extended. Researchers should not only evaluate the facet-level associations

DOI: 10.1002/per

to an outcome within a factor but also across factors, and, whenever possible, on the item level. Anecdotally, we have evaluated the item-level associations for many health, occupation, education, and demographic variables using large samples ($N > 200{,}000$) and consistently find that items from two or more different Big Five factors are among the most highly correlated. This should not be surprising, for the purported orthogonality between the factors (which is usually absent) says nothing about the orientations of the facets to one another.

For another, it suggests need to emphasize that breadth of the outcome being predicted tends to match breadth of measurement. Broad and distal outcomes (e.g., longevity) can be well-predicted by coarse measurement models (e.g., the Big Five); more specific outcomes will be better-predicted by more narrowly operationalized traits. Hierarchically nested facets (like those for the NEO-PI-R) already allow for more narrow trait assessment, though these have complications. For one, they are intrinsically limited in scope to the multidimensional space of the highest level; they are a 30-factor model of the Big-Five space but not a 30-factor model of personality space. Another problem is that these models are not amenable to revision due to dependency in scoring between levels. It is unclear how the model could be improved to accommodate cases where an outcome is best predicted by novel combinations of items (whether items on different facets, different factors, or from outside the extant set). These models of facets are hierarchically rigid and static, though we (Condon, 2014) and Mõttus have suggested heterarchical features of the items and facets depending on the outcome.

Finally, this view of the problems created by simple structure bias can be reconciled with Mõttus' arguments by applauding his call for better means of predictive disambiguation. This is not simply a technical or methodological issue. Mõttus' thesis suggests that *personality scientists may need to move beyond the Big Five*. This idea was heretical in the recent past; the Big Five saved personality from the confused and chaotic mess of personality constructs that preceded its emergence. Now, 25 years later, personality science should be secure enough to consider whether the Big Five structure is too coarse. We think this is the heart of what Mõttus is saying; regardless, it is what we believe.

While Mõttus does also advocate for development of better analytic methods, he consistently implies that the most predictive power comes from the bottom up rather than the top down (e.g., the items rather than the factors). Our view is that new measurement models should be considered. Prediction will be maximized if their levels are independent of one another and if the factors are obliquely oriented within each level. Moving beyond the Big Five also suggests integrating trait assessment with other models of important individual differences, including cognitive abilities, vocational and avocational interests, values, and motivations. These approaches, and others, will help personality to further and fulfill its potential as a predictive science.

A Network Perspective on Causality in Personality Psychology

GIULIO COSTANTINI and MARCO PERUGINI

University of Milan-Bicocca
giulio.costantini@unimib.it

Abstract: The causal roles of personality traits depends on their ontological status. One of the key points made in by Mõttus is that, without assuming existential realism and holism, causal claims involving personality traits do not hold. However, the network perspective on personality offers an alternative account, in which traits are conceptualized as (weakly) emergent properties. From a network perspective, causality can be attributed to traits even without assuming realism and holism. Copyright © 2016 European Association of Personality Psychology

Mõttus (2016) excellent article brings to light some important contradictions of research linking personality and outcomes: While researchers often acknowledge that their results do not warrant causal inferences, they implicitly suggest causality and draw conclusions that would not be warranted if causality did not hold. One of the consequences of this state of affairs is that thorough discussion of the legitimacy of causal inferences in personality has been delayed, although with some important exceptions (Borsboom et al., 2003; Borsboom, Mellenbergh, & van Heerden, 2004; Cramer et al., 2012; Wood et al., 2015).

As Mõttus has shown, the ontological status of personality traits has implications for their causal relevance. One of the key points of his article is that causal claims involving personality traits are not justified without assuming that traits are existentially and holistically real. While we do agree with most points contained in his article, we disagree with this one, especially since existential realism and holism are interpreted as existence of traits as specific and unitary psychobiological attributes. We argue that a network perspective can offer an alternative account of personality traits as causally relevant, while assuming a different ontological status that does not entail realism and holism.

According to the network perspective, the personality system is conceived as a network of elements that interact in complex ways (e.g., Costantini, Epskamp, et al., 2015; Cramer et al., 2012). Personality networks can be defined on relatively microscopic levels, for instance, when the

DOI: 10.1002/per

elements represent momentary experiences of specific emotions (van de Leemput et al., 2014), but also on more macroscopic levels, for instance when elements represent personality facets (Costantini, Epskamp, et al., 2015; Costantini, Richetin, et al., 2015). There are no theoretical boundaries that prevent considering even more microscopic elements (e.g., neurotransmitters, chemical reactions), as well as more macroscopic elements (e.g., broader personality dimensions). Though people could maintain that equating elements to single items in personality questionnaires is the most appropriate choice, it is important to consider that the breadth of items can vary widely (e.g., a single item can assess a whole personality dimension; Woods & Hampson, 2005) and that items often aggregate across more basic phenomena (see Costantini & Perugini, 2012).

Since personality networks can be investigated at very different scales, one could wonder how different levels are related and what this implies for study of causality. We consider higher-levels of the personality network as *emergent* from the microscopic levels (Cramer et al., 2012) and in particular, we see the relations among different network levels as cases of *weak emergence*: Though it would be possible *in principle* to trace patterns of complex causal interactions that are responsible for translating to more microscopic levels (e.g., a complex interplay of chemical reactions and environments) into macroscopic properties (e.g., a recurrent pattern of thoughts, feelings, and behaviors that constitute a personality trait), this enterprise would be too complex *in practice* and the results of this analysis would be of limited use, given their irreducible complexity (Bedau, 2003, 2008). This is not unlike what happens in other fields of science: The microscopic behavior of many agents can result in macroscopic properties that can be causally explained by the microscopic level, but not in simple ways. Examples of emergent properties are traffic jams, that emerge from unsupervised behaviors of many drivers (Bonabeau, 2002), macroscopic organization of the world-wide-web (Albert, Jeong, & Barabási, 1999), and several properties of cellular signaling pathways (Bhalla & Iyengar, 1999) just to name a few.

Although causal relationships are not easy to investigate in emergent phenomena, we envisage ways in which causality can and should be investigated in personality. First, though it is prohibitive to determine the exact causal path that generates a certain macroscopic state, some recurrent patterns often emerge that allow connecting properties of the microscopic structure to macroscopic events (Bedau, 2012). An example in psychology is the work by van de Leemput and colleagues, who showed that some patterns of interaction among momentary experiences of specific moods (microscopic network) are connected to the onset and termination of depressive episodes (macroscopic event; van de Leemput et al., 2014). Other examples are simulation studies that manipulate certain elements of microscopic networks and investigate the effects on macroscopic levels (Read et al., 2010).

Second, one can investigate causality within specific levels of the personality network. Though considering only the most microscopic levels might seem the best option, this is not necessarily the case. Other levels can be quite informative, since macroscopic emergent properties are causally dependent but also relatively autonomous from the microscopic networks that generated them (Bedau, 2008). A simple example is the traffic jam: Asserting that being late for work has been *caused* by a traffic jam is a legitimate statement (Bedau, 2003; Mackie, 1965) even though the traffic jam had itself a very complex causal origin that involves the interdependent behavior of many individuals, including oneself. Some personality-outcomes relations can be better understood by considering more microscopic levels, while other relations should be investigated at more macroscopic levels of abstraction. Mõttus reviews several methods that allow identifying which level of abstraction (e.g., items, facets, dimensions) is more suitable for determining at which level a phenomenon relates to personality. We praise this effort and argue that similar analyses can be also performed from a network perspective (Costantini, Richetin, et al., 2015).

The study of causal relationships in personality networks at all levels can be further facilitated by assuming a functional-cognitive perspective that clearly distinguishes the functional level of analysis from a cognitive level, both in defining the elements of the networks and in drawing conclusions from the analyses of such networks (Perugini, Costantini, Hughes, & De Houwer, 2016). Functional-cognitive analyses that translate in broader behavioral principles can be especially important in identifying regularities that generalize across specific domains (Hughes, De Houwer, & Perugini, 2016).

What We Talk about When We Talk about Causes: The Case of Personality Traits

JEREMY FREESE

Stanford University
jfreese@stanford.edu

Abstract: Counterfactual perspectives on causality have become very influential in social science, and often dissolve conceptual problems posed by less specifically grounded discussions of causality. While particularly ardent counterfactual perspectives deny that personality traits are properly considered causes at all, more ecumenical positions are possible, and these better reflect the potential value of personality psychology to social science. Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Social scientists often find personality psychology intriguing for its elaborate questionnaires and intricate applications of factor analysis. Yet this intrigue has not led social scientists to regularly treat personality psychology as potentially valuable for their own projects. Personality psychology offers a vocabulary for describing relatively stable behavioral differences that may generally be useful for understanding outcomes that substantially involve accumulation of many different behaviors and choices over time.

Nevertheless, interested social scientists might see personality psychology's own efforts to articulate the causal significance of personality as confusing. Many social scientists have become much fussier in recent years about how they talk about causes. Counterfactual perspectives on causal inference have been particularly influential (Morgan & Winship, 2014; Imbens & Rubin, 2015). Counterfactual thinking has had far less sway in personality psychology, as evidenced by its absence in Mõttus' article principally about causal description. I will briefly explain the counterfactualist approach and how it might resolve concerns raised there.

COUNTERFACTUALS AS COGNITIVE SOLVENT

Counterfactualism's big idea is to consider causal claims as fundamentally claims about how something would be different had something else been different. Mõttus offers a saucy example involving John's attending fewer parties than Jane as a cause of him having fewer one-night stands. The counterfactualist implication is that if John had attended more parties, he would have had more one-night stands. Supporting this causal claim requires adducing grounds for inferring John's tally had he attended as many parties as Jane.

Pressing a counterfactualist perspective too far yields various philosophical conundrums, sometimes illustrated by exotic examples with simultaneous assassination attempts or time machines. This should not detract from the clarifying value of counterfactualist thought for practically-minded researchers. In this respect, it offers a sort of *cognitive solvent*, pre-empting ways in which scientific reasoning may become convoluted.

TRAITS AS ATTRIBUTES

Mõttus is concerned principally with interpreting personality traits as causes of outcomes. Scientific wariness about explicitly inferring causes often follows from the various complaints summarized in saying "correlation is not causation." Strong statements of the counterfactualist perspective, however, assert a more fundamental problem with presenting personality traits in causal terms. Holland (2003, p. 8) draws a sharp distinction between attributes and causes, in which "causes are experiences that units undergo and not attributes that they possess." Under this interpretation, personality traits are attributes, so dwelling on questions about their causal interpretation may be considered misguided in basic premise.

This strong position identifies causation with intervention; Holland (1986, p. 959) offers the all-caps slogan "NO CAUSATION WITHOUT MANIPULATION." Party attendance works as a possible cause of one-night stands because we can readily conceive at least hypothetical interventions for it: perhaps John's roommate could have been recruited to host more parties, or scientists could organize some systematic 'disinvitation' of Jane.

On other hand, Mõttus' example also takes John's Extraversion as causing him to attend fewer parties. To translate the idea into something that would satisfy the strong position, we would need to think in terms of some intervention that would change John's Extraversion and potentially also his tally of one-night stands. Doing so, notably, dissolves Mõttus' main conceptual complaints. We have redefined causes as events that happen to individuals – interventions – which circumvents the ontological murkiness that attends trying to divvy up hierarchies of human traits. Events are existentially real and whole by virtue of their happening.

TRAITS AS CAUSES

Binding causal statements to interventions grounds discussion and clarifies interpretation in a way amenable to strategies for empirically estimating causal effects. Yet it is easy to see why this interpretation of causes has not caught on with personality psychologists. For many individual attributes of social-science interest (e.g., income), substantial change at least by hypothetical interventions can be readily envisioned. Personality traits, in contrast, are renowned for their stability, making this more difficult.

The strong interpretation of counterfactual thinking I describe has been criticized as leading to "many good ideas [being] stifled or dismissed from causal analysis" (Pearl, 2009, p. 361). John's modest number of one-night stands, per Mõttus' article, is not just due to avoiding parties but also to being less generally socially talkative. Failing to see these behaviors as co-occuring and mutually reinforcing, we would overestimate the causal effect of getting John to attend more parties, if he just ends up standing awkwardly in a corner. Traits like Extraversion offer utility to social science precisely by providing ways of describing how the internal coherence of behavior transcends the particular events.

Nevertheless, it is frankly difficult to discern what personality psychologists imagine themselves to mean with phrases like "causal links" without referencing counterfactual dependence somehow. Useful here may be the metaphor of hypothetical "surgical" intervention by Pearl (2009). Causal talk could describe precise, hypothetical change in a causal variable that is intentionally decoupled from whether or how such changes might be brought about in fact.

DOI: 10.1002/per

Such formulations once again dissolve many issues Mõttus raises, this time as a matter of semantics. For example, he argues that causal interpretation requires "causal unity," meaning equal effects of any subclassification of a trait. But until facets or "nuances" (McCrae, 2015) or sub-nuances are demonstrably indivisible, this argument applies all the way down, implying any causal project of personality psychology should be postponed altogether.

Instead, talk of Extraversion causing some outcome might be better considered ambiguous or agnostic about "causal unity," rather than assumptive. Obviously, if some effects of Extraversion are confined only to some facets, this is informative to know, just as it would be informative to know if effects in some facets are confined only to some "nuances." Evidence of causal effects at one level of description entreats further inquiry at more specific levels. Not pursuing this specificity might be rightly lamented as a scientific failing to be remedied. But asserting anything more fundamentally amiss requires a clearer position on what one means by traits as causes in the first place.

Implementing a Mõttus Index of Causal Plausibility (MICP): Let's Give It a Try!

LEWIS R. GOLDBERG

Oregon Research Institute
lewg@ori.org

Abstract: I applaud Mõttus' arguments and suggest a method for testing causal plausibility in future studies.

In his brilliantly crafted article, Mõttus sets out like a legal brief an argument for testing plausibility of causal explanations linking psychometric measures of personality-trait concepts to those behavioral outcomes they are asked to predict. This jewel of an article should now be required reading in all future assessment courses.

As part of his conclusions, Mõttus writes: "Third, I argued that, for the causal interpretations to be possible, traits have to display evidence of causal unity: constituents of trait-operationalizations (or trait-indicators) have to be linked with outcomes in similar ways, save for variability in factor loadings. In other words, associations should not depend on how traits happen to be operationalized. Fourth, I argued that such causal unity can be and should be tested in each and every study that seeks to link personality traits with outcomes, although formal methodology for doing this robustly requires further development."

I focus solely on his fourth conclusion, and argue that the germ of a formal methodology for testing causal plausibility is embedded in these ideas: One calculates the correlation between the item loadings on each factor and the item correlations with the outcome variable, across all the items that have been administered to the sample of research participants. Let's call that correlation the Mõttus Index of Causal Plausibility (MICP).

Items vary enormously in reliability, but the MICP accounts for this, because item unreliability should affect factor loadings (which are correlation coefficients) roughly the same as it affects outcome correlations. In the extreme case of perfect unidimensionality of indicators, all the items associated with a factor would have the same size factor loadings and the same size correlations with the outcome variable, and all the remaining items would have zero loadings on that factor and correlations of zero with the outcome variable: The resulting MICP correlation is then +1 or −1.

It is important that the MICP be computed across all the items administered to the sample, and not to the facet scales made up of those items. In most cases, there should be a reasonable number of such items, whereas there will be far fewer facets – rarely, if ever, enough for a reliable index. The set of items used to calculate the MICP should include those associated with multiple factors (e.g., five or six), and the MICP should then be calculated for each of those factors. (An exception will occur when only the items associated with one factor have been administered, which renders this a bad practice.)

If journal editors would routinely ask authors to report the MICP whenever authors suggest the possibility of trait-outcome causality, then over the course of a few years we would learn a lot about the properties of the MICP, and thus test its effectiveness as an index of causal plausibility. Let's do this!

Finally, Mõttus' argument concerning the indifference of indicators in the measurement of personality trait concepts can and should be generalized to a much wider audience of investigators and topics. One could argue that the so-called "replication crisis" in psychological research comes about in part because investigators try many methods of operationalizing a psychological construct until they find a method that provides significant results, and then other investigators use that exact same methodology. As Mõttus would argue, these results are method-

DOI: 10.1002/per

specific, and thus, they should not be ascribed to the psychological concept itself. One should always try to provide conceptual replications, not simply exact replications, if one wants to argue for causal links between theoretical concepts. But that is a topic for another article. Perhaps Mõttus will be its author?

Complicated Issues, Practical Suggestions

ROBERT R. MCCRAE

Gloucester, Massachusetts
RRMcCrae@gmail.com

Abstract: Analysis of causality is extremely complicated; our understanding of personality is very crude. For most purposes, it suffices to act as if personality traits are indeed causal agents. The research strategies that Mõttus recommends as a way to make causal attributions have other practical benefits, including internal replication, more precise generalization, and more powerful prediction. They also imply that meta-analyses should be conducted at the lowest feasible level of the hierarchy. I point out that what might be regarded as tautologies can sometimes offer important insights. Copyright © 2016 European Association of Personality Psychology

Causality is a deep and difficult topic in philosophy of science (Bradley, 1966), involving such problems as infinite regression in chains of causal mechanisms, the nature of time's arrow, and justification of scientific induction. Most scientists, including most psychologists, blithely ignore these problems, although they do consider plausible alternate explanations for causal claims (e.g., they address direction of causality.) The only real justification for this cavalier approach is that it has worked pretty well; we really do seem to understand, predict, and control many phenomena.

This pragmatic view seems especially appropriate for the study of personality traits. Within Five-Factor Theory (FFT; McCrae & Costa, 2008), traits are regarded as basic tendencies, which are hypothetical constructs: There is *something* about the individual that is postulated to give rise to observable patterns of thoughts, feelings, and actions. Traits are not features of the brain (although they are biologically based), nor are they collections of the thoughts, feelings, and actions to which they dispose the individual. They are psychological abstractions. To give a truly causal account of how genes produce traits, and traits behaviors, one would need to solve the mind/body problem, and that is not likely to happen soon. It is perhaps wisest to view trait accounts as models that depict how people function *as if* they had causal agents like Neuroticism or Openness operating inside their heads.

Causality is also difficult for personality psychologists because, by and large, traits cannot be manipulated. For example, FFT asserts that only biological interventions (often impractical or unethical) can modify traits. Although this is surely an oversimplification, it points to the fact that it has been extremely difficult to identify life situations or intentional interventions that modify trait levels (Ellis, 1987) – although many interventions successfully modify problematic behaviors and thoughts. If and when psychologists identify a set of manipulations that systematically and reliably alter trait levels, psychology can enter onto a new level of causal analysis of traits. In the meantime, the niceties of the "coherence and existential realism" of traits are, I think, moot.

METHODOLOGICAL AND ANALYTIC CONSIDERATIONS

Thus, I believe the major contribution of Mõttus' article is methodological: Given the fact that traits are hierarchical – composed of traits at a lower level – and given that associations of component traits with a given outcome may vary, what data should we collect, and how should we analyze them? These are fundamental issues for correlational research.

Mõttus' basic idea is that researchers need to examine correlation of the outcome not only with a trait but also with its components: facets (subscales) or nuances (items; see McCrae, 2015). If most of the components have the same relation to the outcome as the trait does, we can assume the trait is the causal source; if the components have very different relations to the outcome, then causal interpretation ought to focus on the individual components.

I applaud this strategy, because it requires researchers to go beyond global correlations to more fine-grained analyses, a strategy I have advocated for many years (e.g., Costa & McCrae, 1995). It encourages internal replication of results. It leads to more differentiated conclusions from data, such as the generalization that (on average) women are higher than men in Openness, *except for* Openness to Ideas (McCrae, Terracciano, & 78 Members, 2005). It reinforces critiques of very brief measures of broad traits (Credé, Harms, Niehorster, &

DOI: 10.1002/per

Gaye-Valentine, 2012), because they offer no possibility of determining whether an association is due to the broad trait itself or to the specific items by which it was operationalized. By identifying specific facets that chiefly account for outcomes, we can make more powerful predictions (e.g., in occupational selection contexts) and can perhaps be guided more rapidly to useful interventions.

Is any other approach defensible at all? Certainly. If researchers have time to administer only brief, broad measures, they may still find associations worth pursuing. They can reasonably conclude that "Domain X, *or some of its facets or nuances*, is associated with this outcome," and if this finding is novel and the outcome is important, other researchers are likely to follow-up and provide the missing details.

Two scales that ostensibly assess the same trait will in fact be different to the extent that they incorporate different components. An Agreeableness scale consisting of Straightforwardness and Modesty will have rather different correlates than one consisting of Trust, Altruism, Compliance, and Tender-Mindedness (as work on the HEXACO model demonstrates; see Ashton & Lee, 2005). This has clear implications for meta-analyses, where results from studies using many different measures are combined. Analyses of broad traits may underestimate magnitudes or replicability of findings if the true associations are confined to subsets of their components. Meta-analyses ought to be conducted at the lowest feasible level of the trait hierarchy, which will usually mean the facet level.

TAUTOLOGIES

One of the side issues that Mõttus mentions is the claim that purported causal associations of traits with outcomes are often mere tautologies: Is it really surprising or informative that the Positive Emotions facet of Extraversion is associated with measures of happiness? Similar words and phrases are used to assess both predictor and criterion, so it appears that all we have learned is that people are reasonably consistent in their responses to the same question.

One might argue that our measures of happiness are merely disguised measures of traits; then happiness is not caused by traits, it *is* a trait. But that itself would be hugely informative. For centuries, it has been assumed that when people are asked how happy they feel, their response reflects the objective quality of their life circumstances. This (mis)attribution leads to causal interventions – "I will be happier when I get rich/publish my dissertation/find my soulmate" – that often disappoint. How happy we feel may be caused by traits.

Robert R. McCrae receives royalties from the NEO Inventories.

Multi-level Analyses, Multiple Methods and Other Considerations to Enhance Research on Connections between Personality Traits and Outcome

CHRISTOPHER S. NAVE[1] and DAVID C. FUNDER[2]

[1]*Rutgers University, Camden*
[2]*University of California, Riverside*
christopher.nave@rutgers.edu

Abstract: We agree with Mõttus that research needs to examine personality from multiple levels of analysis, but we question the utility of continued lengthy discussions of degree to which traits can be considered "real." We offer three additional suggestions for improving personality research: 1) increased emphasis on studies employing multiple methods, 2) direct and conceptual replications of trait-outcome relationships at multiple levels, and 3) deeper exploration of mechanisms and processes that may drive associations between traits and outcomes. Copyright © 2016 European Association of Personality Psychology

After decades spent establishing traits as predictive of behavior and useful for explaining life outcomes, now is a good time to reflect on how personality psychology can advance even further. Mõttus' article challenges the field to take a deeper look at "whether and when causal interpretations are justified" (p.3), and correctly describes why links between traits and outcomes are important: 1) they explain variability in outcomes 2) they establish utility for traits in that they predict something of value, and 3) links between traits and behavior have implications for designing behavioral interventions. His article also includes useful discussion of the different levels of analysis at which personality can be linked with outcomes.

A frequent practice in personality research is to present findings only at the level of general factors or immediately to reduce or factor analyze individual items into something more resembling the Big 5 or HEXACO. We agree with Mõttus that this knee-jerk reaction is a mistake. Indeed, in our own work, we have been asked

DOI: 10.1002/per

countless times to reduce our large correlate tables into something "more interpretable" and less prone to noise. Looking at individual items, facets scores and factor scores for overlap and for unique predictive validity is an important and often underutilized approach to understanding a trait-outcome relationship. We believe that transparency is key in research and that, in many cases, exclusive use of general factors can obscure what is going on underneath the psychological hood. In contrast, examinations of individual items or larger correlate tables may enable deeper understanding of how traits and outcomes are related to one another. Encouraging researchers to design studies that allow them to examine relations at multiple levels of analysis (and to make the data available to all through online depositories like the Open Science Framework) would aid both transparency and conceptual clarity. Advances in randomization analyses (Sherman & Funder, 2009) and in assessing the reliability of the rank ordering of correlates in a table (Sherman & Wood, 2014) help alleviate concerns about capitalizing on chance when looking at large correlate tables, and encourage broad-based exploratory research.

We do have concern with Mõttus' article in one area: its lengthy philosophical discussion of existential and holistic reality as necessary criteria to discuss causality. One of us has participated in seemingly endless discussions of whether traits can be considered "real" and whether "accuracy" of trait judgments is a meaningful concept (summary: *yes*; Kenrick & Funder, 1988; Funder, 1987, 1991, 1995). At the end of his own examination of reality even Mõttus concedes that "we have sufficiently good reason to believe that, in principle, personality traits as such exist and can exert forces outside the personality domain in real and holistic manners" (p. 9). We think it is wise not to let our field yet again get held back by philosophical discussions of this nature, and instead to forge ahead with empirical research and theoretical development concerning the origin, operation, and consequences of assumptively *real* personality traits.

We have three brief, additional suggestions for improving personality research.

PRIORITIZE MULTI-METHOD RESEARCH

In addition to choosing well-validated personality questionnaires, we believe it is important to bring in multiple methodologies whenever possible. Multi-method approaches reduce issues of shared method variance, improve rigor, and enhance conceptual clarity. Do we find the same patterning of links between traits and outcomes when the traits are assessed via self-reports as opposed to peer reports? Self and clinician report? Self-ratings and directly observed behavioral ratings? Consistent patterning of trait-outcome links establish the robust predictive validity of a trait and any differences between methodologies may yield important psychological insights. Merely, to utilize self-reports is a failure of due diligence. For example, for a long time, research in behavioral genetics relied exclusively on self-reports of personality, leading to widespread conclusion that shared family environment has no effect on personality development, a conclusion that was overturned when multiple methods of assessment and behavioral observation were finally employed (Borkenau, Riemann, Angleitner, & Spinath, 2001). What other seemingly established findings will be challenged when more diverse methods are used?

CONDUCT SYSTEMATIC DIRECT AND CONCEPTUAL REPLICATIONS AT MULTIPLE LEVELS OF ANALYSIS

We strongly agree with Mõttus' recommendation that personality should be assessed at all levels of analysis and that results should be compared across levels before findings of personality-outcome relationships are presented. In addition, we would encourage a greater emphasis on replication. Direct replications can be useful in that they shed light on how robust identical items and factors relate to identical outcomes in similar samples. For example, does Conscientiousness, as measured by individual items and a factor score, predict academic performance among college students at multiple colleges with similar demographics? Conceptual replications are helpful as well. What happens if different items or scales are used to measure Conscientiousness? Mõttus correctly reminds us that not all scales measuring Big Five constructs cover the same depth and breadth. Can we obtain the same effect using similar college students? Can we obtain similar predictive validity if we diversify our college student sample? Questions like these are important and are not addressed often enough.

CAREFULLY EXPLORE MECHANISMS AND PROCESSES

Despite the well-known dangers of confusing correlation with causation, longitudinal research and large multimethod samples can allow research to address the processes and mechanisms that underlie robust trait-outcome relationships. In a seminal paper, Hampson (2012) argued that processes and mechanisms can be illuminated both by short-term, event-sampling studies (e.g., finding that Conscientious individuals wear seat belts, drink only in moderation, and avoid risky behaviors on a daily basis) and, in parallel, by lifespan approaches that demonstrate long-term consequences (e.g., Conscientious individuals enjoy better health and mortality in the long run, presumably because of their daily behavior in the short run). Use of multiple methods, including directly observed behavior, may shed additional light on processes, particularly in studies designed to assess short-term behavior as a mediator of long-term trait-outcome associations.

The authors gratefully acknowledge the contributions of research assistants from the Personality, Health, and Behavior Lab at Rutgers University, Camden.

DOI: 10.1002/per

Conceptual and Methodological Complexity of Narrow Trait Measures in Personality-outcome Research: Better Knowledge by Partitioning Variance from Multiple Latent Traits and Measurement Artifacts

DENIZ S. ONES, BRENTON M. WIERNIK, MICHAEL P. WILMOT and JACK W. KOSTAL

University of Minnesota - Twin Cities
onesx001@umn.edu

Abstract: Increased conceptual clarity and methodological rigor is needed in personality-outcome research. We describe the hierarchical nature of personality, with implications for prediction and explanation. Latent traits exist at all levels of the hierarchy. Traits should not be confused with their measures. Attention to criteria to be predicted is essential. Measurement error arising from raters, measures, and assessment occasions should be controlled before bi-factor modeling can separate sources of associated variance. We illustrate these points using meta-analytic data to estimate variance sources in typical other-rated Achievement items and scales. Copyright © 2016 European Association of Personality Psychology

We support increased rigor in personality trait-outcome research. Such rigor requires clear conceptualization of the latent structure of personality and appropriate analytic methods to separate myriad trait and artifactual sources of variance present in personality measures. Below, we present three considerations for conducting rigorous applied personality research.

LATENT TRAITS EXIST AT ALL LEVELS OF THE PERSONALITY HIERARCHY

Causal models of personality trait-outcome relations must be grounded in empirically established trait structures. The hierarchical Big Five model is the most robustly supported structure (John, Naumann, & Soto, 2008; McCrae & Costa, 1997). We conceptualize the Big Five and associated lower-order traits at multiple hierarchical levels, with successively greater specificity and narrower bandwidth. The Big Five factors describe broad parameters of individuals' goal-directed behavior (e.g., Conscientiousness reflects capacity to protect one's goals from disruption; DeYoung, 2015). Below factors, *aspects* describe more specific behavioral patterns that, although distinct from the Big Five, often act in service to those broad tendencies (e.g., the Industriousness aspect of Conscientiousness reflects prioritization of long-term goals, which is one way individuals avoid goal disruption; DeYoung, Quilty, & Peterson, 2007). Below aspects, *facets* capture very narrow behavioral patterns (e.g., Achievement is the tendency to pursue challenging long-term/abstract goals). Within a domain, traits at each level covary because the same behaviors fulfill several traits' psychological functions.

Traits at each hierarchical level should be interpreted as psychological entities in their own rights. Until recently, the Big Five were typically described as the shared variance among narrow trait scales (DeYoung, 2015). An unfortunate side-effect is that researchers often consider the Big Five factors as merely sums of facet traits and view lower-order traits as merely indicators of these factors, rather than as distinct traits unto themselves (cf. Mõttus, p. 294) notes. This characterization is inaccurate; the Big Five factors are not simply formative aggregates of their facets. Instead, they represent broad parameters for how individuals act to achieve their goals. Lower-order traits are similarly distinct psychological tendencies that covary with their associated factors because traits across levels share behavioral manifestations. Factors, aspects, facets, and even specific behavioral manifestations (e.g., items measuring punctuality) can each have unique relations with outcomes; consequently, personality research must identify which latent trait is the source of predictive power for any given criterion.

LATENT TRAITS SHOULD NOT BE CONFUSED WITH THEIR MEASURES

Responses to personality scales are multiply determined. Scale variance reflects not only the intended construct but also myriad other latent traits and measurement artifacts. For example, scores on the NEO PI-R Achievement-striving scale reflect both the Industriousness aspect and Conscientiousness factor, in addition to the Achievement facet. Scores also reflect numerous measurement artifacts, including item-specific variance (from the particular items included), scale-specific variance (from a scale's idiosyncratic conceptualization of the trait), rater-specific variance, transient error, and random-response error.

To demonstrate these various effects, we used meta-analytic data to estimate the relative contributions of each source of variance to a typical *other-rated* Achievement item and scale.[1] Results in Table 1 and Figure 1 are revealing. First, for single items, the largest source of variance is random-response error. Importantly, this error is not "item-specific" reliable variance, but reflects truly random error.

[1]Detailed descriptions of the methods and results of these analyses, as well as additional considerations for estimating measurement error variance components are available in Ones, Wiernik, Wilmot, & Kostal (2016; 10.6084/m9.figshare.3100795).

DOI: 10.1002/per

Table 1. Estimated variance components for a typical other-rated measure of the achievement facet of conscientiousness

Source of variance	How to estimate	Var. comp. for item	Var. comp. for facet scale
Latent traits		**.03**	**.06**
Big Five factor	Bifactor model	.014	.030
Conscientiousness		(.463)	(.463)
Aspect	Bifactor model	.006	.010
Industriousness		(.194)	(.194)
Facet	Bifactor model	.010	.020
Achievement		(.343)	(.343)
Measurement artifacts		**.97**	**.94**
Transient error	CE − CES	.04	.10
Item-specific variance	CS − CES	.04	.10
Scale-specific variance	CE − GCE	.03	.08
Random-response error	1 + CES − CE − CS	.66	.13
Rater-specific variance	CE + CS − CES − IRR	.20	.53

Note: Intercorrelations for latent-trait estimates taken from Judge, Rodell, Klinger, Simon, and Crawford (2013); data for measurement artifacts taken from Connelly (2008) and Gnambs (2015); values in parentheses are the proportion of latent trait variance attributable to each trait; var. comp. = variance component; CE = coefficient of equivalence (e.g., coefficient α, coefficient ω, parallel forms reliability); CES = coefficient of equivalence and stability (Schmidt, Le, & Ilies, 2003); CS = coefficient of stability (test-retest reliability); IRR = interrater reliability; GCE = generalized coefficient of equivalence (Le, Schmidt, & Putka, 2009). See Ones et al. (2016; http://doi.org/10.6084/m9.figshare.3100795) for methods used to estimate these values and additional considerations for estimating measurement error variance components.

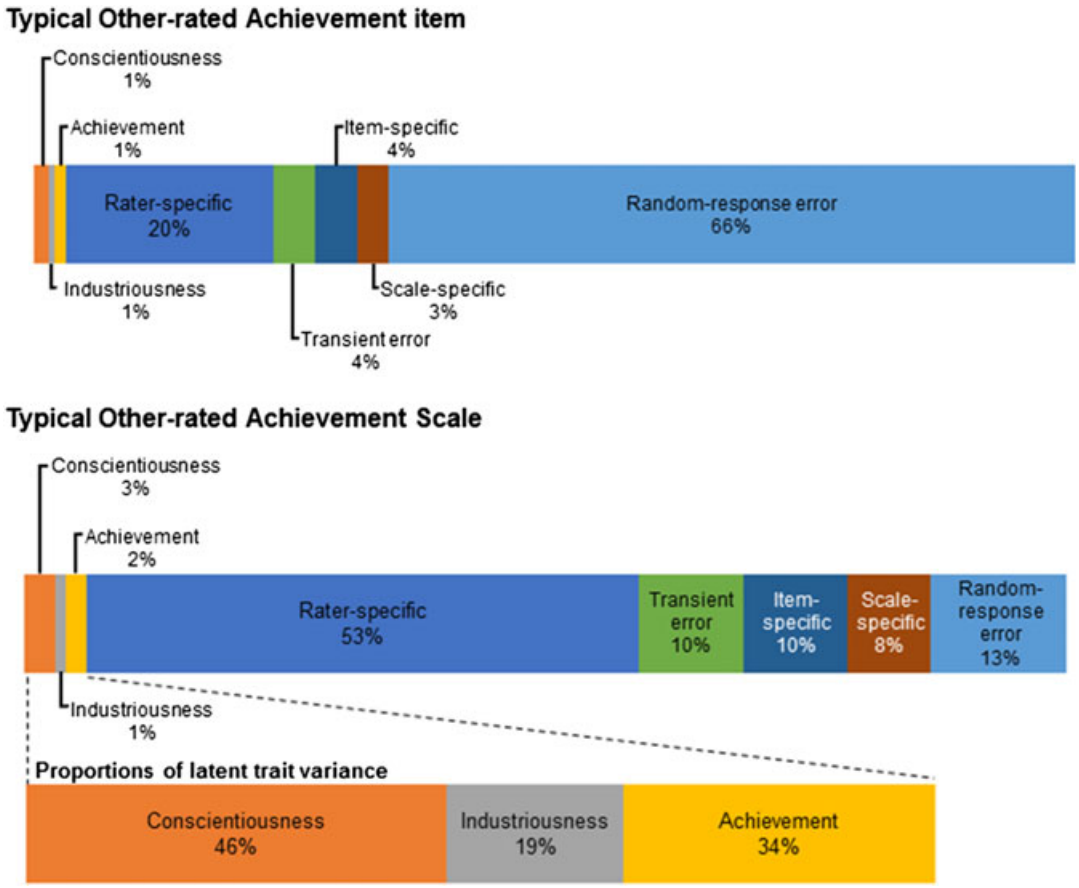

Figure 1. Variance components for a hypothetical other-rated Achievement scale and hypothetical single Achievement item. Values in the cutaway indicate the proportions of latent trait variance attributable to each trait.

This matters because differences in observed item-criterion relations more likely result from measurement error than from systematic features of the items, (cf. Mõttus, p. 298). Because two-thirds of the variance for single items is truly random error, single items are by themselves unacceptable as measures for personality research. Only by aggregating multiple items (and measurement occasions) can systematic variance sources overwhelm artifacts (Epstein, 1983). Second, for a typical Achievement scale, measurement error again dominates—this time from rater-effects. Trait variance accounts

DOI: 10.1002/per

for only 6% of score variance. If, like most personality research, we ignore rater effects, then 41% of the observed variance remains error. Consequently, to accurately capture trait variance, researchers should increase reliability by using multiple raters (Viswesvaran, Ones, Schmidt, Le, & Oh, 2014), occasions (Epstein, 1983), and measures for each trait (DeYoung, 2006). Given the difficulties associated with simultaneously accomplishing these measurement ideals in one primary study, psychometric meta-analysis can address these distorting effects of measurement error (Schmidt & Hunter, 2014). The computational specifications in Table 1 provide helpful guidance.

Retuning to our example of trait-relevant variance in a typical Achievement scale, 46% reflects the Conscientiousness factor, 19% the Industriousness aspect, and only 34% the Achievement facet. This mixture of trait variance makes it impossible to unambiguously interpret criterion correlations based on *observed* scale scores as though it reflected the influence of one trait.

BI-FACTOR MODELING SEPARATES TRAIT VARIANCE

Because personality scales reflect the influences of many latent traits, analytic methods must separate these influences in predicting criteria. Mõttus suggests a 'leave-one-out' procedure to determine whether prediction stems from factor or facet traits. Such analyses can detect whether criterion relations stem from particular items (cf. Mõttus, Realo, Allik, Deary, Esko & Metspalu, 2012), but they cannot disentangle contributions of each latent trait to criteria. Instead, bi-factor models, which analytically separate variance associated with each trait, can be fit to examine criterion relations. Bifactor analyses show that both Big Five factor and facet traits have meaningful relations with work and academic outcomes (Connelly, Wilmot, Hülsheger, Ones, & DeYoung, 2016; McAbee, Oswald, & Connelly, 2014; Salgado et al., 2015); for the investigated criteria, factor traits typically show higher predictive validity than facet traits. Interestingly, some facets predict in directions opposite to their associated factors. Wiernik, Wilmot, and Kostal (2015) provide a primer on using bi-factor models to separate the predictive contributions of multiple latent traits.

CONCLUSION

Personality trait-outcome research must be grounded in clear conceptualizations of both the personality trait hierarchy and criteria and must also attend to statistical and methodological issues associated with partitioning variance.

Our comments should not be mistaken as attacks on empirically oriented approaches to personality-outcome research. Applications of personality traits have a rich predictive history, even without causal explanations, including in clinical (Krueger, Markon, Patrick, & Iacono, 2005), lifespan (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007), and industrial-work-organizational psychology (Ones, Dilchert, Viswesvaran, & Judge, 2007). Specifying the *reasons* for personality-outcome relations is not necessarily essential for applied use. What matters is that the relations exist at all. Prediction and explanation are mutually reinforcing; better prediction informs better explanations, which, in turn, produces better prediction.

All authors contributed equally; order of authorship is arbitrary.

Prediction and Personality-Related Outcomes

DANIEL J. OZER

University of California, Riverside
daniel.ozer@ucr.edu

Abstract: Prediction of important life outcomes from personality attributes is an important endeavor for a variety of reasons, not least of which is developing causal explanations for those outcomes. A 'homogeneity of effects' criterion for attributing causation to a broad superordinate trait is unnecessarily stringent, and current knowledge of personality structure is not yet sufficient to fruitfully implement the proposal. Copyright © 2016 European Association of Personality Psychology

There are many ways in which personality trait-outcome research might be made more rigorous (e.g., better measures, more longitudinal designs, larger and more heterogeneous samples), but Mõttus focuses on meta-theoretical concerns. He suggests that concurrent or predictive relations between traits and outcomes are largely of interest only if they can be understood as causal. He then argues that if the causal source of the associated outcome is to be understood as a unitary trait, then facets (and items) should be equally related to that outcome, with the allowance that this relation may vary to the extent that the facets (or items) are differentially saturated with the putative causal trait—a condition one might describe as 'homogeneity among effects.' I argue that prediction consistent with a causal model is the present reasonable

DOI: 10.1002/per

limit for explaining trait-outcome relations; and however much the homogeneity of effects criterion might serve the development of a causal argument, it is not necessary and is presently beyond routine research practice.

Prediction is not only some substitute for casual explanation. For empirical forecasting, a correlate of the true causal variable may be useful (e.g., a residential postal code may be practically useful when setting car insurance rates). In many instances that would concern personality psychologists, the casual pathway between a set of personality traits and distal outcomes may be so tortuous (Meehl's, 1978 account of 'context-dependent stochastologicals' come to mind) as to undermine any general causal account. In such a circumstance, playing at causal explanation is self-deception: All we can do is predict.

But suppose we do think that some trait-outcome relation is relatively straightforward, and we set out to employ the logic Mõttus describes. Which facets are to be employed? While we have general agreement about the broad factor structure of personality traits (but are there five or six?), there is no consensus on facets; indeed no clear reason to prefer a facet substructure, as opposed to say, a circumplex-like model. Understanding a trait as the cause of an outcome should immediately engage theorizing about mediating processes and the testing of more demanding models. In the development of a causal argument to explain a trait-outcome association, I grant that homogeneity among facet effects provides support for attributing the cause to the broader trait; and the absence of such homogeneity may lead to causal attribution to a facet. While Mõttus' discussion of the implications of homogeneity of effects for causal arguments makes a persuasive case in principle, we know far too little about relations of facets to traits (or items to facets) to believe that considerations other than differential factor saturation can be set aside. There are likely causal relations among facets; and even if two facets have identical relations to a primary factor, they may be differentially related to other factors which are themselves related (differentially) to the outcome of interest. If item-facet and facet-trait relations are truly described by simple structure effect indicator models, then Mõttus' homogeneity of effects condition does strengthen causal claims for the broad trait, but given the absence of empirically grounded agreement about the facet structure of any broad personality trait, and lack of real discussion about what would count as appropriate boundaries for any facet item pool, it seems premature to claim that simple structure effect indicator models can be presumed as the foundational personality structure and that we are ready to undertake the path Mõttus envisioned.

There are some difficult but surely not insurmountable methodological problems that would arise even if a simple structure effects indicator structure of a broad trait was clearly sufficient: How would tests for facet effect homogeneity be undertaken? Simple differences between correlations are notorious lacking in power. If regression models and incremental validity methods are applied, measurement error issues becomes especially problematic; and if testing models where correlated latent facets with coefficients fixed equal are estimated as predictors of an outcome, sufficient power will guarantee that homogeneity of effects will be rejected. If across multiple sufficiently powered studies, one facet carries the entire predictive burden, then attributing causality to that facet rather than the larger trait seems appropriate, but this is a far different standard than the argument for homogenous effects I understand Mõttus to be making.

Despite my misgivings, I agree with Mõttus in recognizing that if facets differentially predict outcomes, it is problematic to act as if they don't by attributing the relationship to the global trait. But I take from this insight a different lesson: It is important to engage with the question of the substructure of personality factors, to understand constituent elements, some of which may be partial causal functions of the unitary existential trait, and also partial functions of other facets within that and other domains. That is, there may be causal relations among facets, as posited in some network models. Without elucidation of this structure, addressing questions about causal relations of traits to distal outcomes seems premature. Presently, prediction must suffice.

Wrong Premise, Right Direction, but Let's Go Further

ARTHUR POROPAT

Griffith University
arthur.poropat@griffith.edu.au

Abstract: 'Realist' interpretations of personality factors require scales with factorial coherence and independent psychobiological confirmation. Like most (if not all) personality models, the Five-Factor Model traits fail these tests, making them poor candidates for analysing causal relations between personality and either behaviour or life outcomes. More finely-focused scales, such as the facet scales advocated by Mõttus, show more promise for causal analyses, provided they can be shown to be unifactorial, clearly verifiable with psychobiological evidence, and aligned with fundamental psychological theories. Variance decomposition will assist causal analyses using such scales.

DOI: 10.1002/per

Relations between personality and behaviour or life outcomes have been central to scientific personality research since the 19th Century, yet the most important of these relations remain merely correlational. This highlights fundamental problems for personality research, and Mõttus has done the field a favour by suggesting some solutions.

Much of the critique Mõttus presents revolves around indeterminacy of measurement associated with one particular model of personality, the Five-Factor Model (FFM). However, measurement problems extend to other personality models (Corr & Poropat, 2016; Poropat & Corr, 2015). Unfortunately, Mõttus bases his subsequent arguments upon the commonly-made premise that personality factors have ontological, explanatory or causal status, which he refers to as the 'realist position' in which a trait is a 'real psychobiological attribute rather than a shorthand summary of various behaviors, thoughts, feelings and whatever else that happen to correlate'.

Unfortunately, without independent confirmation of the ontological status of traits, the realist position leads to the type of reasoning presented in the so-called 'Five-Factor Theory' of personality (McCrae et al., 1999). In that 'theory', latent causal factors are established by identification of the FFM, and the FFM is caused by the latent causal factors, which are both unobserved and unobservable. So this theory requires personality researchers to trust in the unknowable, much as medieval thinkers were required to accept unknowable 'spirits' as explanations for human action. Latent factors may be pragmatically convenient for statistical analyses, but this circular reasoning is ontologically and empirically vacuous without confirmation from some source other than personality ratings. To date, psychobiological evidence for the FFM has been questionable (Vul, Harris, Winkielman, & Pashler, 2009), unreliable (Bjornebekk et al., 2013), and modest.

Independent confirmation of current personality models appears unlikely. For example, Mõttus argues that Spearman's (1927) *theorem of indifference of indicator* can form the basis for arguments for the realist position on personality traits. This theorem implies that reliable variance on an indicator will be shared with equivalent indicators, because they will primarily reflect the same latent construct. This implication can be tested using inter-scale correlations and reliability estimates. For example, intelligence measures have typical inter-scale correlations of .8 to .9 and reliabilities of .9 or higher (Urbina, 2011), meaning the proportion of reliable variance shared across intelligence scales is somewhere between 80% and 100%, providing strong support for the latent *g* factor of intelligence.

Pace and Brannick (2010) reported meta-analytic estimates of scale internal consistency and inter-scale correlations that can be used to calculate similar estimates for FFM scales. Scales commonly used to assess the FFM share only a third (34.7%) of their reliable variance, while scales specifically designed to measure the FFM share slightly more than half (56.1%, but still only 38.7% for emotional stability). These figures imply that FFM scales do not measure unitary, common latent factors but instead really *are* shorthand summaries, the outcomes of the complicated processes that produce personality ratings.

This does not mean personality factors are fictions, any more than other shorthand summaries of observations of real events. For example, 'health' is a shorthand summary of multitudinous psycho-socio-biological states and processes, and health ratings are empirically-reliable and pragmatically-valuable for medical practitioners. But if medical practitioners were unable to link health ratings with biological theories and measurements, they would still be in the dark ages. Likewise, our personalities are real and important but as with health, relying on global factors for causal understanding of personality will typically, and perhaps inevitably, be misleading.

Although these comments have focused upon the FFM, the same criticisms apply to any of the currently-dominant personality models because of the manner in which personality is assessed. As outlined recently (Corr & Poropat, 2016; Poropat & Corr, 2015), personality assessments confound multiple types of variance arising from predictable influences. The most important of these confounds is that between personality as expressed (what individuals does) and personality as perceived (how individuals are evaluated). Without decomposing these and other types of variance, causal analyses will produce misleading and confusing results.

Approached from a different direction, Mõttus provides a recommendation that complements this concern about variance sources. Specifically, Mõttus advocates focus on facets, tightly focused personality scales, to understand causal relations with personality. Tightly focused personality scales allows greater variance decomposition and are also more likely to result in genuinely unifactorial scales. In turn, there is greater potential for verification against psychobiological realities, especially if combined with a variance decomposition methodology. And biologically verifiable, unifactorial scales will be far more amenable to causal analyses than the FFM.

This approach should also help to address one of the central problems with the FFM: the atheoretical nature of the FFM traits (Block, 2001). But rather than creating facets to suit the FFM (or any other model), we need facets that specifically reflect independently verified theories. For example, cognitive-learning theory probably warrants being described as the keystone of scientific psychology and has promising associations with personality (Poropat, 2015, 2016). Facets that clearly reflect learning theory, including facets that are positioned within specific contexts or life domains, are likely to provide compelling arguments for causality. The potential is great if we take Mõttus' ideas these few steps further.

DOI: 10.1002/per

Factors are Still Fictions

WILLIAM REVELLE and LORIEN G. ELLEMAN

Northwestern University
revelle@northwestern.edu

Abstract: Mõttus considers the causal relationship between traits as they relate to outcomes. We applaud his efforts and add that all latent traits identified by factor analysis are convenient mathematical fictions. Traits are the formative results of the (perhaps) non-linear sums of basic biological and social mechanisms. We suggest that personality is fractal and has an equally complex structure (is self-similar) at any level of analysis. Traits are useful fictions for relating the myriad of fundamental causes with the seemingly infinite types of behavioral observations we may make.

At least once a decade, it is time to remind personality researchers that factors are indeed fictions (Loevinger, 1957; Revelle, 1983), and that we should not reify the factors known as 'The Big Five' (Block, 1995, 2010). Mõttus does this, and does it well. He focuses on the supposed causal relationship between traits (as exemplified in the 'Big Five') and outcomes. The argument could equally be applied to causal sources of these traits. Just as some theorists explain individual differences in health or longevity in terms of individual differences in conscientiousness or neuroticism (e.g., Weston & Jackson, 2015), many theorists with biological bent like to 'explain' individual differences in traits such as extraversion with individual differences in strength of biological mechanisms such as the Behavioral Activation System (Corr, 2008; Gray & McNaughton, 2000; Smillie, 2008) or in terms of interactions of the six relatively autonomous systems (sensory, motor, cognitive, affective, value and style) of Royce (1983).

In his use of traits, Mõttus refers to the many who take the perspective termed 'realist' by Borsboom et al. (2003)[2] and claim that traits are real psychological attributes. This is reminiscent of the earlier claim by Cattell (1943, 1945) that factors are source traits that can be used to explain the observed correlations between surface traits of items or of behavior clusters. It is also reminiscent of Royce's use of hierarchical factor analysis to identify 'invariant dimensions of individuality' (Royce, 1983, p. 684). To use the terminology of Bollen and Lennox (1991), implicit in this explanation of outcomes is the reflective latent trait model where traits are common causes of items or behaviors. The items are locally independent; that is, when controlling level of the trait, the items do not correlate. This might well be, but these traits are themselves presumably formative results of the (perhaps non-linear) sums of basic biological and social mechanisms.

That reflective source traits can be used as explanations of covariances of observed items, and behaviors is a convenient mathematical fiction. That five (or three, or six, or ten) factors can be extracted from the matrix of intercorrelations of 67 or 171 paragraph descriptors (Cattell, 1945), 100 (Goldberg, 1992) or 540 adjectives (Hofstee, de Raad, & Goldberg, 1992), or 696 short stemmed items (Condon, 2014), is merely a way of representing (modeling) covariances with factors that can, with a suitable choice of items, lead to near local independence of the items. Unfortunately, any particular exploratory factor solution may be subjected to an infinite number of alternative rotations, all of which are mathematically identical in fitting the covariances. The well-known debate between Eysenck (1967) and Gray (1981) as to whether to rotate towards Extraversion-Neuroticism or Impulsivity-Anxiety cannot be resolved on psychometric grounds.

Furthermore, such factor analytic solutions, although near approximations, are just that: approximations. The optimal number of factors to describe any particular covariance matrix is a tradeoff between parsimony (few factors) and goodness of fit (more factors). But most measures of goodness of fit vary as functions of sample size. As sample sizes increase beyond the 100 of Cattell (1945), or the 800–1000 participants of the Eugene-Springfield data set of Goldberg and Saucier (2016), to the sample sizes available through web-based data collection, e.g. $>$ 24 000 in Condon and Revelle (2015),$>$ 65 000 in Revelle, Wilt, and Rosenthal (2010) or $>$ 200 000 in Revelle et al. (2016), the number of interpretable factors increases. Indeed, it appears that the factorial structure of personality is fractal, that is, it is equally complex (self-similar) at all levels of analysis.

Each of three to five high-level factors shows horizontal as well as vertical structure (Goldberg, 1993) and can be subdivided into three to five lower-level factors which in turn yield three to five homogeneous item composites. As is usually the case, Lew Goldberg has made this point before:

'Because one always loses specific variance as one amalgamates measures, the optimal level of prediction is completely a function of statistical power and thus of sample size. In the population (i.e. samples of unlimited size) optimal prediction by regression analysis will always be at the level of individual items; that is, for huge samples it would be silly even to amalgamate the items into scales because one would inevitably lose some specific variance at the item

[2]Although Borsboom et al. (2003) credit Spearman (1904) as originator of the concept of latent variables as theoretical constructs, the concept of unobserved (latent) cause of observations goes back at least 2400 years to the Allegory of the Cave in Plato's *The Republic*.

DOI: 10.1002/per

level that could serve to increase predictive accuracy.' (Goldberg, 1993, pp. 181–182).

Mõttus gave one example of the need to consider item-level data when examining trait-behavior correlations. When predicting variation in Body Mass Index (BMI), rather than the very broad trait of Neuroticism, or the facet of Impulsivity, it is at the item level, e.g., 'When I am having my favorite food, I tend to eat too much' that the best predictors of BMI are found (Terracciano et al., 2009). Support for this also comes from our finding using the SAPA methodology (Revelle et al., 2016) with N > 50 000 where the best measures predicting BMI are 'I ate too much' (r = 0.25) or 'used public transportation' (r = 0.20).

Traits, facets, or items are useful links between fundamental causes (genes, enzymes, transmitters, brain structures, environmental experiences) and behavior. They are useful because they summarize broad patterns of relationships with observed regularities of behavior. They lead to appropriate levels of specificity within the broad framework of symmetry of predictor and criterion (Wittmann, 1988). Although useful, they are fictions created for the purpose of telling a coherent story relating the myriad of fundamental causes with the seemingly infinite types of behavioral observations we may make.

Partially supported by National Science Foundation grant: SMA-1419324 to William Revelle.

Establishing that Estimated Trait-Outcome Associations Aren't Artefactual, Inflated, or Attenuated by Specific Indicators

KENNETH J. SHER

University of Missouri
sherk@missouri.edu

Abstract: Mõttus' suggestions of ways to establish the validity of observed trait-outcome associations are well reasoned. I consider one general issue, the extent to which some observed associations could be artefactual due to criterion contamination of specific trait indicators, and briefly review alternative ways of assessing this problem. Although optimal ways are not clear-cut, approaches Mõttus and I suggest could be viable candidates. Moreover, these same approaches could be used to identify trait indicators that attenuate observed trait-outcome associations artefactually. Copyright © 2016 European Association of Personality Psychology

Mõttus highlights several important issues often overlooked in everyday practice, focusing on importance of insuring validity of estimated trait-outcome associations. Among other points, he highlights renewed recognition of the importance of treating experimental conditions as random rather than fixed effects, a position long-espoused in experimental and clinical research (e.g., Brunswik, 1955; Hammond, 1954), and that the same logic could be extended to conceptualizing personality tests items.

The question of 'whether the associations of traits with outcomes are independent of which indicators are employed rather than being specific to particular indicators' (p. 19) should logically be related to the generalizability coefficient of the scale measuring the trait. In principle, this coefficient represents how well the employed set of indicators generalizes to the universe of suitable indicators. If generalizability is high, the average inter-item correlation is of moderate magnitude, and the number of indicators is large, the extent to which a given indicator is artefactually inducing a trait-outcome association should be, *a priori*, very low. However, in common practice, there is often opportunity for indicator-specific contamination to create appearance of association at the trait level.

The issue of predictor-criterion overlap or 'criterion contamination' has long been recognized in research in personality and substance use (e.g., Darkes, Greenbaum, & Goldman, 1998) and failure to recognize this potential problem can probably be blamed on complacency or lack of due diligence in knowing the item content of personality scales employed in one's research. In my work as a reviewer and editor, I have been surprised at how often this issue has figured prominently in an editorial decision, suggesting its importance has not been sufficientlyly recognized. Moreover, the prototypic exemplars noted by Mõttus and Darkes et al. (1998) and others may represent only the tip of an iceberg in that it seems likely there can be considerable predictor-criterion overlap that is much subtler and more implicit than explicit. For example, if liking 'wild parties' is an indicator of extraversion, its association with drinking alcohol may not be due to blatant criterion contamination but rather implied contamination (i.e., 'wild parties' are often accompanied by alcohol excess). Scrutiny of personality trait indicators and outcome measurements can reveal clear instances of likely contamination, instances of unlikely contamination, but also instances in a 'gray zone' and subject to interpretation and judgment. Due diligence in considering this issue in all work and attempting to address it empirically would appear to be foundational to good research practice.

Although logical and semantic analysis of indicators and outcomes is a reasonable place to start, statistical approaches such as examining the residual correlation between an

DOI: 10.1002/per

indicator of a factor and an outcome (either on *a priori* basis or based on a statistical procedure such as a model modification index like a Lagrange multiplier test) can inform us whether there is unique variance in an indicator that is associated with an outcome beyond the common variance shared with the factor. In general, there is no substitute for thoughtful examination of research materials employed and statistical analyses that attempt to resolve whether there are suspicious anomalies that would lead one to entertain the possibility that some observed associations are artefactual and should not be interpreted substantively. Mõttus cites recent work (Vainik, Mõttus, Allik, Esko, & Realo, 2015) that provides 'formal procedures based on the idea of systematically dropping trait indicators and recalculating the associations' (p. 19) and proceeding iteratively.

Logically, one could take this approach further by undertaking exhaustive permutation analyses in which all possible subsets (of all possible sizes) are drawn from a larger set of trait indicators, and the correlation of each subset with an outcome estimated. The distribution of the results of these permutations can not only be used to generate median and mean associations of the trait measure (assessed by varying number of indicators) with the outcome but also to identify those subsets that produce the largest and smallest estimates. That is, say, for a trait measure with 10 indicators, one would calculate: (1) all 10 permutations of set size = 1 (i.e., the individual correlations), (2) all 45 permutations of set size = 2, and so on through the 10 permutations of set size = 9 and for each set size generate the permutation distribution of the trait-outcome correlations. Then 'extreme' subsets can be used to identify individual items and groups of items that are potentially problematic (while recognizing that such extremes can be generated by chance, and there should be a consistency check such as replicating this in multiple partitions of the data to verify the reproducibility of such effects).

Mõttus' concern and the Vainik *et al.* procedure focus on the question of artefactually large associations, but it seems possible there could be one or more items that conspire to *reduce* the magnitude of the trait-outcome association artefactually in a manner opposite of that obvious criterion contamination. Whether or not this would occur in practice is an open empirical question but by exhaustively examining the entire permutation distribution of trait-outcome associations, one could identify problematic items at both extremes. Although permutation analyses is not common in personality research, in our own work on nosology, we have found it to be a powerful tool in assessing the validity of basic assumptions (Vergés, Steinley, Trull, & Sher, 2010) as well as a tool for discovery (Steinley, Lane, & Sher, 2016), especially in situations where there is interest in interchangeability of items or criteria (Lane & Sher, 2015).

Mõttus touches on a number of other important issues with the common theme of how to improve our theoretical and methodological rigor. His call is well argued, and he provides some preliminary guidance on how to achieve it. The challenge is to motivate researchers to undertake the additional computational work necessary to establish the validity of estimated trait-outcome associations, especially in the context of multivariate research programs where other covariates can condition observed outcome associations at the trait and item levels.

Supported in part by NIH grant K05AA017242

A Hierarchical Defense of Broad Traits

SARA J. WESTON and JOSHUA J. JACKSON

Washington University in St. Louis
sweston@wustl.edu

Abstract: We support Mõttus' call to consider relations at multiple levels of analysis. However, he takes a step too far in insisting that, if only some facets or items are associated with outcomes, the broad trait should be left out of the interpretation. Such insistence ignores (1) the future of the field; (2) the hierarchal nature of personality that hampers the ability to tease apart and measure lower levels of analysis easily; and (3) that most current studies do not adequately separate broad trait variance from specific facet variance. Copyright © 2016 European Association of Personality Psychology

We commend Mõttus for his article, specifically for highlighting need for wide-spread assessment of lower-order facets of personality in the study of real-world outcomes. As he correctly points out, use of other lower-order assessments can help to clarify the relations between personality and outcomes. We and others wholeheartedly agree that a lower-order assessment below the Big Five can be helpful in better ruling out measurement overlap (Suls & Bunde, 2005), clarifying the content of existing measures of personality traits (Mike, Harris, Roberts, & Jackson, 2015), describing development of traits (Jackson et al., 2009), uncovering the mechanisms that link traits with outcomes (Turiano, Chapman, Gruenewald, & Mroczek, 2015), and understanding the trait itself (Jackson et al., 2010).

Despite our general agreement with the article's sentiment, we argue for continued justification of ascribing causal status to broad personality traits, not simply lower-order facets or items. We see at least three main reasons for this.

DOI: 10.1002/per

First, broad traits lose usefulness if they are not causal entities but instead are merely descriptions of co-varying behaviors/thoughts/feelings. In this, we echo realistic interpretations of traits that suggest that the power of personality traits is that they provide meaningful explanatory power at a broad level (e.g., Funder, 1991). Documenting associations at a lower level of analysis contributes to an overall nomological network of hierarchical trait space. It is this nomological network of personality traits that has proven useful for personality psychology in the last two decades. Taken to extreme, the alternative vision set forth by Mõttus is a field that may resemble the pre-Big Five/Three era, where jingle/jangle was the sound of the land.

Second, we believe that the hierarchical organization of trait space justifies continued use of ascribing causal status to related traits at different levels of the trait hierarchy. One consequence of viewing traits hierarchically is that it is difficult to discuss a single level as the 'true' or 'optimal' level – one could always go slightly broader or more narrow in assessment. If it is challenging to distinguish one level from another, and the differences among levels are largely arbitrary, where should causal status start and end? At the item level, the facet level, somewhere in-between? Where would causal status stop when climbing up the hierarchy? And would researchers put in the efforts to measure all of these levels?

Similarly, stating personality is causal implies that causality is somehow embedded as a part of the person, in a realistic sense. If one assumes that trait space is hierarchically organized, then the neurological features responsible for traits must also be ordered hierarchically. If only facets have causal status, it implies that facets are located in unique and unrelated neurophysiological structures, unrelated to broader traits, and solely responsible for manifestation of these facets. However, personality facets from the same trait likely share some underlying brain structure. Similarly, broad traits and thus broad neurological systems responsible would undoubtedly share similar neurophysiology with facets. Thus, in an overly simplified reductionist manner, we might say impulse control influences health and does so as a result of particular neurological features. Later, when we refer to conscientiousness, the overarching trait, we also refer to the same neurological features that are responsible for the link between impulse control and health (plus additional ones). Thus we cannot separate causal status neurologically or physically, because broad traits relate to more specific facets due to the hierarchical nature of personality.

A third reason for pause, due to the way that much work is currently conducted, is that it is unknown whether associations at lower-order levels reflect specific lower-order associations, or if they are driven by broader trait content. Mõttus argues that trait-outcome associations should only be interpreted as pertaining to the unique variance of those facets or items (p. 14). But discussing unique variance of facets or items cannot be done with simple bivariate associations. A simple composite measure of a facet combines both general trait variance and specific facet variance, meaning that bivariate correlations with these facets and some outcome could be driven by either trait variance or facet variance or both. Instead, we must turn to more advanced methods, such as bifactor analysis, to appropriately isolate the unique variance of a facet that is not shared with the broad trait (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; McAbee et al., 2014).

The problem of isolating unique variance is equally difficult when we move from the factor level to the item level. Assessment of a single item cannot benefit from aggregation techniques used to isolate and remove measurement error. Thus, we cannot distinguish null associations between items and outcomes that reflect measurement error from null associations that reflect true independence.

Overall, we agree with Mõttus in calling for analyzing relations at the trait-, facet- and item-levels. However, he takes a step too far in insisting that, if only some facets or items are associated with outcomes, the broad trait should be left out of the causal interpretation. Such discussion ignores (1) the potential future of the field; (2) the hierarchal nature of personality that hampers the ability of researchers to easily tease apart and measure lower levels of analysis and (3) that most current studies do not adequately separate broad trait variance with more specific facet variance.

On the TRAPs that Make it Dangerous to Study Personality with Personality Questionnaires

DUSTIN WOOD and P.D. HARMS

University of Alabama
dustinwood79@gmail.com

Abstract: The concerns discussed by Mõttus have been discussed for decades (e.g., Nicholls, Licht, & Pearl, 1982). Here, we emphasize that one of the major problems discussed in his review, which we term tautological relationships between attributes and predictors (or TRAPs), can be attributed more specifically to semantic redundancies between items. To avoid correlations between variables being driven by hidden TRAPs, we suggest that researchers increasingly consider use of item-level analyses, and identify more formalized procedures for indexing semantic redundancy. Copyright © 2016 European Association of Personality Psychology

DOI: 10.1002/per

Mõttus' article brings to mind Nicholls *et al.*'s (1982) earlier warnings of the 'dangers of using personality questionnaires to study personality' (p. 572). Specifically, they detailed that two measures of conceptually distinct constructs (e.g., *masculinity*, *self-esteem*) often correlate for the trivial reason of having semantically redundant items or items used to estimate trait levels across measures (e.g., '[I have a] strong personality', '[I] describe myself as a pretty strong personality'). Mõttus' article shows that correlations driven by such semantic redundancies continue to reach publication at a pace largely unabated 34 years later. (Nor is this issue limited to personality questionnaires; see Larsen & Bong, in press.)

Other labels for similar phenomena include the *jangle fallacy* (Kelley, 1927) and *construct identity fallacy* (Larsen & Bong, in press). We will describe these here as *tautological relationships between attributes and predictors* (TRAPs), due to our understanding that correlations driven by semantically redundant tests or items can exist between measures of distinct constructs for the more valid (i.e., less 'fallacious') reason of the two constructs having overlapping item domains. For instance, *general conscientiousness* and *work-contextualized conscientiousness* (i.e., how conscientious I am *in general* versus *just as an employee*) are conceptually distinguishable, but 'arriving at work on time' or 'completing job assignments' helps to identify the individual as higher on both constructs simultaneously (Wood, 2007).

We continue by providing two of the more important forces which increase the prevalence of TRAPs, which may help researchers to avoid them in the future.

BIAS AGAINST ITEM-LEVEL ANALYSES

Like Mõttus, Nicholls et al. (1982) noted that 'more careful attention to scale content should lead to more consistent detection of such problems' (p. 578). We agree with this suggestion; however ,it is too often the case that researchers are not asked to examine or report their results at the item level. This is unfortunate, because as Mõttus noted, simply eyeballing the items is one of the most straightforward and effective way of identifying whether correlations of interest are influenced by TRAPs.

Recent work by Mõttus and others illustrates that important findings concerning trait-outcome associations can be revealed by simply decomposing broad scales to their narrower 'aspects' or 'nuances'. We would go further and argue: why not prioritize the item-level? Item-level analyses have several virtues. First, because items are generally directly reported in item-level analyses, it is much easier to spot TRAPs. Second, traditional psychometric concerns about item-level analyses are increasingly shown to be unfounded. For instance, recent papers have shown that test-retest reliabilities (which can be estimated on single items) are both higher than might be expected and more *appropriate* reliability estimates than estimates based on inter-item correlations (which cannot; McCrae, Kurtz, Yamagata, & Terracciano, 2011; Wood & Wortman, 2012). Third, due to recently increased ability to provide supplementary materials with published articles, the old barrier of space limitations to the publication of item-level results is now becoming increasingly irrelevant.

LACK OF PROCEDURES FOR INDEXING SEMANTIC REDUNDANCY

Another reason for prevalence of TRAPs in the literature is that they are difficult to formally operationalize. As Nicholls et al. (1982) noted, 'the test of equivalence of item content we urge you to make is necessarily subjective. Item inter-correlations cannot serve in this content because items can be correlated even if there is absolutely no perceived or logical overlap of item content' (p. 573). We agree that determinations of semantic redundancy should be based *solely* on an examination of the content of the tests (e.g., the item stems, response scales). For instance, we should be able to determine that the items 'I like being with people' and 'I enjoy being around other people' are essentially semantically redundant independent of any consideration of their actual correlation. However, we believe it is also important that determinations of semantic redundancy between two items are not *ultimately* of an entirely 'we know it when we see it' character. For instance, Arnulf, Larsen, Martinsen, and Bong (2014) and Larsen and Bong (in press) showed that a range of procedures can be used to index the level of semantic similarity between items via automated analyses of item content formally.

CONCLUSION

There are few less interesting reasons for correlations between measures than the presence of semantically redundant items. What is clear from Mõttus contribution, which we have attempted to amplify here, is that correlations driven by semantic redundancies between measures continue to pervade the literature decades after earlier warnings (e.g., Nicholls et al., 1982). What is less clear is how these TRAPs can be avoided. Given space limitations, we are not able to elaborate all of the sources of TRAPs, nor necessarily even the most important. However, we hope to show that better identifying the major sources of TRAPs may help toward the development of more formalized procedures for detecting and avoiding them in the future.

ACKNOWLEDGEMENT

We thank Kai Larsen for feedback on an earlier draft of this manuscript.

DOI: 10.1002/per

Domain: Facet = A: a?

MATTHIAS ZIEGLER and JOHANNA ZIEGLER
Humboldt Universität zu, Berlin
zieglema@hu-berlin.de

Abstract: Mõttus puts forward some thought-provoking ideas. One of the main issues raised concerns the relations between personality trait domains and facets and between traits and their measurement criteria. While we agree with many of the points raised, we suggest some slightly different conclusions. Based on arguments regarding unidimensionality, makeup of criteria, and situation perceptions, we recommend extending the argument space to scrutinize a complete nomological net and not just abstract definitions. This means substantiating psychological processes behind traits. Copyright © 2016 European Association of Personality Psychology

Mõttus' article is a strong theoretical work questioning trait theories. The main assumption underlying his criticism is that test-criterion-correlations for facets should be similar to the test-criterion-correlation of the domain, only varying with factor loadings. The same argument is brought forward regarding the relations between items and facets. Otherwise, to Mõttus, a causal influence of the trait on a specific criterion is not feasible. The theoretical considerations fit with work by Borsboom and colleagues (e.g., Borsboom et al., 2003; Fried et al., 2016) and read like a swan song for the idea of latent traits.

In general, we agree with most if not all of the arguments presented. However, in our opinion three additional aspects need to be considered: unidimensionality, criterion makeup, and influences of situation perceptions.

UNIDIMENSIONALITY

Generally, it is presumed that unidimensionality means items reflecting one specific trait also represent one specific psychological process. However, depending on the breadth of a construct, this is not necessarily true. Bejar (1983) established that unidimensionality holds if all items are influenced by the same set of psychological processes in the same way (also see Carpenter, Just, & Shell, 1990). Transferring this to the hierarchical conception of traits, an observed domain score would represent a multitude of different psychological processes, all of which influence the underlying facets in similar ways. Taking this definition seriously, there would be no need for facets. Thus, each facet needs to be allowed to have specific and systematic variance reflecting the trait in question but not shared with other facets. This specific makeup of psychological processes within a facet should influence underlying items in a similar way to guarantee unidimensionality. Accepting these ideas explains differences in loadings of domains on facets. Moreover, these ideas have consequences for test-criterion-correlations for domains, facets, and items.

A more direct consequence of these ideas is demand for faceted personality measures with tested factorial validity including tests for unidimensionality for each measured facet and domain (Ziegler, 2014; Ziegler & Hagemann, 2015).

THE CRITERION

Following up on the idea of facets containing specific trait variance not reflected by other facets, it is important to reconsider the criterion itself. Mõttus assumed that facets should be similarly correlated with the criterion and the domain, otherwise causal inferences regarding the domain should be considered illegitimate. This per se is true. However, before throwing out the baby with the bath water, the criterion itself should be considered. Judging the facet-criterion-correlation is based on the assumption that the facet in question predicts the trait-relevant variance within the criterion. Brogden and Taylor (1950) explicated several types of criterion bias undermining this assumption. Of importance here is what they call criterion-contamination, i.e. criterion variance extraneous to the domain score. As explained, facet variance represents psychological processes reflected by all facets but also some unique process reflecting the trait. If we assume that the criterion variance represents the same variance shared by all facets but also the specific psychological process unique to the facet in question, the facet-criterion-correlation for this facet will be higher than for the other facets or for the domain. However, this is not due to trait-irrelevant variance. Instead the specific facet variance, which makes the facet important in the first place, is responsible for this result.

It could also be argued that facet and criterion are more symmetrical in level of abstraction (Brunswik, 1955), rendering a larger correlation than observed on domain level. This would not explain why one facet correlates higher than the other facets of a domain, though. The idea of unique but trait-relevant facet variance, however, does explain such a finding. Of course, this still leaves the question of whether it is correct to attribute this to the domain under research or the facet. Mõttus called the first interpretation into question. Considering Cronbach and Meehl's (1955) notion of a nomological net, it might indeed be more advisable to refer to a specific location within this net as being causal for the observed influence on a criterion. In any case, thorough investigation of the criterion variance and the psychological processes reflecting an assumed causal relationship is necessary. The result does not need to be abandonment of traits but could be a more fine-grained combination of a trait's nomological net and psychological processes and criterion makeup.

DOI: 10.1002/per

SITUATION PERCEPTION

Tett and Burnett's (2003) notion of trait activation theory assumes that individual trait differences only manifest if a situation is trait-relevant and external rewards do not preclude differing behaviors. It has been shown that different job demands lead to different personality facets predicting job training outcomes (Ziegler et al., 2014). Again, this has implications for the ideas presented in the target article. A criterion often is a set of specific behaviors, shown across a specific time period within a specific setting (e.g., academic performance). Thus, the classes of situations (Rauthmann, Sherman, & Funder, 2015) constituting this criterion need to be trait-relevant. Coming back to the idea of a trait being a combination of different psychological processes, the situation classifications need to be relevant for these psychological processes. Again it could be argued that the unique processes inherent in facets are activated differently by those situations leading to different correlations. Thus, considering the specific psychological processes relevant in the situation classes constituting a criterion could be more informative regarding different facet level correlations than doubting the whole idea of domains and facets. Of course, as Mõttus stated, it is correct to question a causal trait influence.

Summing up, we support most of the arguments Mõttus brought forward. However, we recommend considering questions of unidimensionality, criterion makeup, and situational influences to complement evidence produced to claim a causal influence. Moreover, we plea for interpreting nomological nets representing an assortment of specific psychological processes instead of abstract trait definitions.

AUTHOR'S RESPONSE

Item-Level Analyses Should Become Standard – For More than One Reason

RENÉ MÕTTUS

University of Edinburgh
rene.mottus@ed.ac.uk

Abstract: Among the topics discussed in the comments, one idea appeared to be supported by most commenters: when personality trait scores are related to possible outcome variables (or possible causal factors), scale-level analyses should be supplemented by item-level analyses. This could help to corroborate causal inferences, refine interpretations, rule out measurement/construct overlaps and/or lead to new discoveries. This suggestion is consistent with recent evidence regarding single items often reflecting unique personality characteristics ('nuances') with trait-like properties. Future work could focus on improving item properties and delineating a useful set of nuances. Copyright © 2016 European Association of Personality Psychology

I am grateful to all the 35 commenters, authoring 19 comments in total, for their thoughtful contributions. Their comments spanned a number of specific topics, from intricacies of factor analyses (Revelle & Elleman; Condon & Mroczek) and measurement models (Ones et al.) to general principles of understanding causality (Freese; van Bork et al.; McCrae) and no less than the very nature of personality traits (e.g., Asendorpf; Baumert et al.; Constatini & Perugini). The comments sometimes offered very different views on the same questions. For example, while some commenters urged researchers to focus on explaining how personality can be causal to outcomes even to the extent of carrying out experimental manipulations (Asendorpf; Baumert et al.), others argued that we should currently be content with merely documenting associations and postpone causal explanations (Ozer; McCrae). As another example, some suggested that personality traits are real entities (Nave & Funder), while others argued that they are useful fictions for telling coherent stories (Revelle & Ellemann), or that traits could be dismissed altogether as explanatory units (Asendorpf). Several comments put forward specific methodological suggestions for improving personality trait-outcome research (Baumert et al.; van Bork et al.; Condon & Mroczek; Goldberg; McCrae; Sher; Wood & Harms). Some commenters appeared worried that questioning the nature of traits or trait-outcome associations could damage the progress of personality research (Nave & Funder; Weston & Jackson), whereas others seemed to suggest doing exactly this (Condon & Mroczek; Baumert et al.; Poropat; Ziegler & Ziegler). I will not attempt to address all these diverse topics and views in this rejoinder. This is not because I mean to dismiss them. Instead, this is because I want to focus on what seems to be the most important practical conclusion that can be taken from the discussion generated by my article: ***it's time to work with items***.

FROM NOW ON, LET'S CARRY OUT AND REPORT ITEM-LEVEL ANALYSES

There are number of reasons that speak for carrying out item-level analyses. I therefore suggest that reviewers and editors

DOI: 10.1002/per

start encouraging authors to report them, if they do not already do so.

Most of us think that item-level analysis is a good idea

To the extent that the 35 commenters and I constitute a quorum of personality researchers, there is an emerging consensus that item-level analyses are worth doing. This does not mean doing away with scale-level analyses necessarily, though item-level associations might suffice in some cases; rather, it means supplementing scale-level analyses with item-level ones. Most of the commenters appeared to encourage (Baumert et al.; Costantini & Perugini; Freese; Revelle & Elleman; Weston & Jackson) or even strongly advocate item-level analyses (Asendorpf; van Bork et al.; Chapman; Condon & Mroczek; Goldberg; McCrae; Nave & Funder; Sher; Wood & Harms) – and so do I. A few commenters did not explicitly address item-level analyses, but advocated facet-level analyses (Ones et al.; Ozer; Poropat; Ziegler & Zielger). I suspect that their arguments may also justify extending the analyses to single items. However, one comment appeared less sympathetic to the idea: Clark et al. suggested that item-level analyses might be impractical due to items' low reliabilities and increased type 1 error rates; I will discuss these points below.

Item-level analyses can be useful for several reasons, even if one does not agree with all

I suggest that item-level analyses are required when causal inferences are sought. As I stressed in the target article, if an association of a composite trait with an outcome appears to be driven by only a subset of the indicators used for identifying the trait (sometimes only one or a few items), then, as a general rule, causal interpretations should focus on these indicators rather than their ostensible parent trait. That is, although item-level analyses cannot be used to support causal claims per se, as causal unity is not a sufficient condition for causality, they may help to rule out implausible claims. Some commenters appeared to agree with this reasoning (e.g., Asendorpf; Goldberg; Chapman, perhaps also Ziegler & Zielger). However, several commenters (e.g., Baumert et al.; Freese; McCrae; Ozer; Weston & Jackson) did not think that discordant outcome-correlations of items of the same scale – beyond what is expected due to different factor loadings – are necessarily problematic for causal interpretations.

For example, it was suggested that there might be discordant indirect effects from traits to outcomes via items (or facets) and this alone can cause heterogeneity in outcome correlations (Baumert et al.). This may be true, but there is a counter-argument, which follows directly from the basic idea of the target article: if the associations are to be ascribed to existentially real and holistic traits, then they should exist independently of their indicators and purported mediators– and this should be tested. Thus, in the example of Baumert et al., Extraversion should be defined independently of the Sociability and Dominance items and the association that appears then could be taken as an indication of how Extraversion as such may be related to marital status.

It was also suggested that our current understanding of personality traits and their structure is too limited to separate different sources of causal influence (e.g. McCrae; Ozer; Weston & Jackson). I definitely agree with this. But this is exactly why I prefer to think that the best strategy is to base interpretations of observed associations on the level of analysis that yields the most consistent results and not to invoke hypothetical higher-order constructs when there is no consistent evidence pointing to their relevance. Regardless of whether we think that we have evidence for veracity of the higher-order traits, they may not be needed for interpreting 'personality's share' of the variance in an outcome, unless evidence consistently shows that there is something about these particular collections of behaviors, thoughts or feelings that coheres in a way that relates to the outcome more strongly than the pieces do. In other words, I see traits as composites, either reflecting something real or being convenient summaries, whose relevance is not earned by virtue of being part of a model (questionnaire) but depends on the particular purpose at hand. For example, just because we have the FFM traits (or the trait labels) does not mean that we have to or objectively can rely on them to be the causal agents.

Item-level analyses can have value even for those who disagree with their implications for causal interpretability

Several commenters (e.g., Freese; McCrae; Nave & Funder; Ziegler & Ziegler) suggested that item-level analyses can *refine* our understanding of how traits are related to outcomes, perhaps pointing to possible pathways between them. For example, if scores on an Agreeableness scale are correlated with partners' marital satisfaction and this correlation is particularly strong for Agreeableness items referring to being quarrelsome and uncooperative, then this may point to mechanisms by which Agreeableness, whatever it is, relates to marital quality. This is especially plausible if the other items of the scale display correlations in the same direction but smaller in magnitude, so that dropping the quarrelsomeness and uncooperativeness items from the scale would not nullify the trait-level correlation, but only weaken it.

Item-level analyses can reveal associations that would not emerge from trait-level analyses

For example, if only one or two items of a scale are correlated with an outcome, this may not be sufficient to make the scale-level association strong enough to catch researchers' attention (e.g., other items may have near-zero associations – maybe often even in the other direction). As a result, item-level analyses may lead to new discoveries. For example, the Achievement Striving facet of the NEO Personality Inventory (NEO-PI; revised or third version; McCrae & Costa, 2010) is not correlated with Body Mass Index (BMI; Sutin, Ferrucci, Zonderman, & Terracciano, 2011; Vainik et al., 2015), but there is evidence that its item that refers to giving up on self-improvement programs is linked with BMI (Mõttus, Kandler, Bleidorn, Riemann, & McCrae, in press).

DOI: 10.1002/per

Investigating item-level correlations can help to identify possible measurement/construct overlaps

As Wood and Harms put it in their comment, 'There are few less interesting reasons for correlations between measures than the presence of semantically redundant items'. In some cases, overlaps in the content of personality scales and outcomes is obvious (e.g., two items referring to feeling depressed or happy), whereas in some cases it is more subtle. For example, in his comment, Sher discussed that the association between Extraversion and alcohol use might be driven by items such as those referring to 'wild parties'. Of course, such instances of 'indicator-specific contamination' (Sher) might reflect genuinely interesting associations, but generalizing them to broader trait constructs is probably not appropriate.

ITEMS ARE NOT ALWAYS NOISY INDICATORS – MIND THE 'NUANCES'!

Collections of individual items are used to define traits such as those of the Five-Factor Model (FFM) domains or facets, but individual items often reflect specific personality characteristics over and above these broader traits. McCrae (2015) has called these specific personality characteristics 'nuances'. We (Mõttus et al., in press; Mõttus, McCrae, Allik, and Realo 2014) have previously reported that the unique variances of most of the 240 NEO-PI items (after being residualized for the variances of facets and domains) have significant rank-order stability and cross-rater agreement, and a large proportion of them also display significant genetic variance components. Furthermore, a number of item residuals predicted BMI and interests in various life domains consistent with specific hypotheses that had been set up for them (Mõttus et al., in press). Of course, not all items of commonly-used personality questionnaires provide incremental value for the description of individual differences and prediction of outcomes. Perhaps many of them do mostly what they are designed to do: measure broader trait constructs as well as possible, which means contributing to some predefined co-variance structure. Yet, despite having been designed for other purposes, many items *do* contain useful unique signal and thereby constitute nuances. It seems wise to harness this information usefully rather than ignore it or, even worse, risk letting it distort findings at the level of broader constructs.

SINGLE ITEMS ARE NOT HOPELESSLY UNRELIABLE: A SPECIFIC EXAMPLE

Traditional psychometric training (e.g., based on classical test theory) mislabels items' unique variance as measurement error. This may have led to the widespread perception that single items are notoriously unreliable, which, I think, is an exaggeration. Surely, aggregation has the benefit of tending to reduce random error that is undoubtedly present, but individual items seem to be doing quite well in capturing potentially valid signal (Ones et al.; Wood & Harms). In addition to the above-cited evidence regarding the properties of item residuals, I want to nail this point with one more example, pertaining to an already familiar outcome, BMI.

Specifically, I predicted BMI (log-transformed and residualized for age and sex) from a selection of individual items and 30 NEO-PI facets in the Estonian Genome Bank dataset (Leitsalu et al., 2015; Vainik et al., 2015; 3,548 individuals with BMI and personality self-reports, 60% women, mean age 46.8 years with a range of 18 to 91 and standard deviation of 17.0). To reduce the chances of capitalizing on chance, I split the sample into a 'learning sample' (N = 2,500), where I created a prediction model for BMI, and a 'testing sample' (N = 1,548), where I applied the model; I repeated this procedure 1,000 times. Specifically, in the learning sample, I calculated the correlations of individual items with BMI and then fit a linear regression model, whereby BMI was predicted by the scores of those items that had significant correlations with BMI (applying Bonferroni correction, so the threshold p-value was .05/240 = .000208). In most cases, the prediction model included only four to six items, although the number of items varied from three to nine (three items were omnipresent: two came from the Impulsiveness facet and referred to eating too much and one came from the Achievement Striving facet and referred to giving up on self-improvement programs). I then applied the models estimated in the learning sample to the independent testing sample and correlated the predicted BMI with the actual BMI values. I also fit a linear regression predicting BMI from the scores of the 30 facets in the learning sample, applied this in the testing sample and, again, correlated the predicted BMI values with observed values. The average correlations between predicted and observed BMI values were .27 when items were used for the prediction and .25 when the 30 facets were used. I also ran the predictions with only the three omnipresent items in the model, which also resulted in an average predicted-observed BMI correlation of .27. Naturally, the facet scores included the 'top' items: removing them reduced the average correlation between facet-predicted BMI and observed BMI to .18. Increasing the number of items in the prediction formula only slightly increased the predicted-observed BMI correlation. For example, when prediction in the learning sample included top 50 items, the average correlation between predicted and observed BMIs in the testing sample was .30.

This suggests that for predicting this particular outcome, a few selected items can outperform 30 facets (especially when the facets do not include these best-predicting items), and that throwing more items in might slightly increase the predictive accuracy. This is consistent with the findings reported in the comment by Revelle and Elleman: the strongest predictors of BMI were a few single items. If so, there may be no need to implicate broader traits at all: as far as the presented evidence is concerned, individuals with higher BMI eat too much and give up easily on self-improvement. Period.

DOI: 10.1002/per

ITEM-LEVEL ANALYSES ARE 'FREE'

It is obvious, but worth pointing out all the same: carrying out item-level analyses does not require collecting any additional data, nor any new modelling procedures. It just means repeating the scale-level analyses at the item level. In software packages like R, such analyses can easily be automated.

Historically, journal space constraints may have prevented researchers from reporting detailed analyses. However, this is no longer the case, as Wood and Harms pointed out: "… due to recently increased ability to provide supplementary materials with published articles, the old barrier of space limitations to the publication of item-level results is now becoming increasingly irrelevant".

THERE IS A DEDICATED R PACKAGE FOR ITEM-LEVEL ANALYSES

An R package (ionr; Vainik & Mõttus, 2016) can help with testing the degrees to which trait-outcome associations are independent of particular items. The package applies the procedure described by Vainik et al. (2015), whereby items are systematically dropped from scales and significance of the resulting changes in outcome-correlations is estimated. Appropriate levels of significance in the changes of outcome-correlations, given the sample size and other relevant parameters, can also be estimated using the package. Items that significantly increase or decrease the scale-outcome correlations (if any) are dropped from the scales and finally two outcome-correlations are compared: one based on the scale with all items included and the other based on a reduced scale from which any 'bad apples' are removed. The package also plots the associations of single items with outcomes (see Figure 1 for an example). Admittedly, this procedure does not take into account the variability of factor loadings across items (dropping an item with a low factor loading has different implications for the aggregate of the remaining items than dropping an item with a high factor loading). Therefore, attempts to improve on this procedure or devise alternatives should be encouraged. As one example of how the procedure can be extended, Sher proposed a permutation approach whereby all item combinations are related to the outcome and those producing the strongest associations are taken for further consideration.

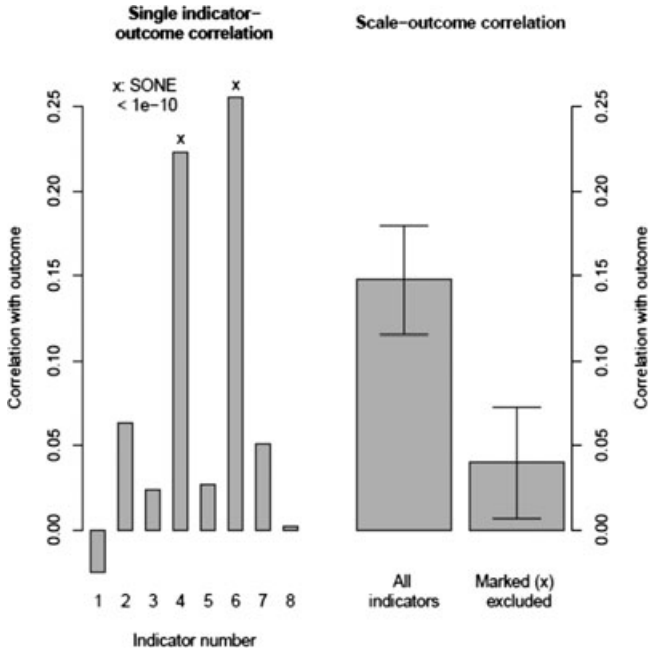

Figure 1. Example plot from R-package ionr.

RISKS OF MULTIPLE COMPARISONS CAN BE MANAGED

As Clark et al. correctly pointed out, running more analyses entails an increased risk of running into problems related to multiple comparisons such as increased type 1 error rates. Chapman, however, suggested that this can be managed by employing procedures such as False Discovery Rate or questionnaire permutations, or relying on Bayesian inference methods. If one wants to be especially stringent, Bonferroni correction might be applied, as I did above, and/or associations can be estimated in multiple partitions of a sample to test their robustness. Also, for parsimonious models that control for inter-correlations among predictors, researchers could be encouraged to employ shrinkage procedures such as LASSO or related methods (Tibshirani, 2011). Of course, large samples and – most of all – independent replications are required, both direct and conceptual (comparing results based on similar but not identical items), as was pointed out by Goldberg and Nave & Funder.

REPLICATIONS AND META-ANALYSES SHOULD CONSIDER FACETS – AND ITEMS

McCrae suggested that possible facet- or item-specificity of associations may lead to underestimations of their replicability and result in attenuated meta-analytic associations, especially when the meta-analyzed studies have used different combinations of specific items to measure the same constructs: 'Analyses of broad traits may underestimate magnitudes or replicability of findings if the true associations are confined to subsets of their components. Meta-analyses ought to be conducted at the lowest feasible level of the trait hierarchy, which will usually mean the facet level.' I agree, but would go further: when the same instruments are used in multiple studies, or even different instruments with similar items, why not carry out meta-analyses at item level? Most of all, this makes sense when there is evidence for item-specificity in the associations from individual studies. In particular, there may be replicable item-specific associations that do not emerge at the level of composite scales, leading to new discoveries.

DOI: 10.1002/per

STUDIES OF DEVELOPMENT AND POSSIBLE CAUSES OF PERSONALITY VARIANCE SHOULD ALSO REPORT ITEM-LEVEL ASSOCIATIONS

Echoing Asendorpf's comment, I think the argument for studying item-level associations should extend beyond personality-outcome research. For example, items of the same NEO-PI facets may have hugely variable developmental trajectories which cannot be explained by their differential factor loadings. Specifically, we (Mõttus et al., 2015) found that none of the 30 NEO-PI facets met the criterion for strong measurement invariance across age groups. Furthermore, even when items had been residualized for their respective facets, nearly half of them had significant ($p < 0.0002$) correlations with age, suggesting that a substantial reason that personality characteristics relate to age may involve items' unique rather than shared variance.

Examination at the item-level associations revealed some interesting patterns (Mõttus et al., 2015). For example, although the Depression facet scores showed a slight downwards trend with age, two items ('I tend to blame myself when anything goes wrong', 'I have a low opinion of myself') had significant positive associations with age. Therefore, based on items' common variance, older people tended to experience slightly less NEO-defined Depression than younger people, but, within that, they were inclined to be more self-critical. As another example, although there were no clear age trends in the Achievement Striving facet scores, some items showed large age-related differences. Scores of the item 'I'm something of a "workaholic"' increased until age 70 and then showed slight decline (possibly partly reflecting physical limitations or retirement), whereas scores of items referring to *not* being 'easy-going' and 'lackadaisical' but being 'driven to get ahead' showed decreases. In fact, mean differences among some of the items was up to 2.1 standard deviations around age 70, though the mean scores were identical in the young-adulthood age group (as this is how the measurement models were set up). Therefore, although older people/cohorts may describe themselves as more hard-working than younger people/cohorts do, they may have less ambition and feel less driven to get ahead. Whatever they show, such item-level patterns can be informative.

Usefulness of item-level analyses also extends to studies that seek to identify potential causes of personality variance such as genetic variants, brain parameters (structural, functional or chemical) or social-cognitive processes. If something is to be a causal factor for a trait as such, then its impact ought to be observable across its manifestations. If this is not the case, then the ostensible causal factor may only be relevant for a specific component of the trait definition and not to the trait as such. Importantly, this argument also applies to experimental approaches that attempt to manipulate personality traits (as suggested by Baumert et al.) by, for example, pharmacological (e.g., Tang et al., 2009) or behavioral interventions (e.g., Jackson, Hill, Payne, Roberts, & Stine-Morrow, 2012). Of course, undertaking such efforts requires some commitment to the idea of traits as real entities in the first place.

MORE ATTENTION TO THE PROPERTIES OF SINGLE ITEMS

Although items seem able to perform well in terms of being associated with outcomes, they might perform even better if we could improve their psychometric properties. I have the impression (based on my own attempts to develop questionnaires, among other things) that one of the principle item properties that test constructors consider when selecting or refining their content or wording, is items' ability to contribute to the variance *shared* among the items of their intended scale (i.e., all else being equal, items should increase internal consistency/factor loadings) and, at the same time, not to contribute to the shared variance of other scales (i.e., items should have as few cross-loadings as possible). These considerations are related to the goal of obtaining simple structure, discussed by Condon and Mroczek. (This is less true when personality test construction is based on Item Response Theory, but this approach has not been taken often.) But if we accept item-level variance as potentially informative, this practice has a major negative consequence, which I will discuss in the next section. Moreover, focus on it may also distract test constructors from other important item properties.

First, unambiguous readability of items is paramount. Second, an important formal property of items is variance. The amount of variance available in the scores of single items is artificially limited by the response options presented to participants, and this may be hard to alter. But even within the limited ranges of response options typically used, distributions of items scores are often extremely skewed (Mõttus et al., 2015). For example, it is not uncommon that more than 80% of the responses given on a, say, 5-point Likert scale fall into just two categories towards one end of the scale. When this is so, it may substantially limit the predictive value of the characteristic that the item reflects. Such problems could be fixed by writing items with reasonable (medium) levels of 'difficulty' and by reducing social desirability from their content. Taking item-level analyses more seriously should thus motivate us more than ever to consider and improve the psychometric properties of single items, and by extension, the scales to which they contribute.

THERE ARE PROBABLY MORE NUANCES THAN WE CAN CURRENTLY (AND MAYBE EVER) IDENTIFY

Current evidence regarding nuances as potentially uniquely informative personality characteristics (Mõttus et al., in press; 2014) may underestimate their pervasiveness, because odds of identifying them have been against us. This is because the evidence is based on questionnaires containing items that have been *designed* to measure particular broader personality traits (facets, domains) as purely as possible, increase scales' internal consistency, and yield 'simple structure', as Condon and Mroczek put it. In other words, items that have appeared to measure something other than the preconceived broader traits (and thus might have predominantly captured nuances

DOI: 10.1002/per

beyond those thought to be most representative or otherwise relevant), have been omitted from them in the first place (for an example of such a procedure see Soto, in press). Furthermore, from the very beginning, structural models such as the FFM have been identified based on the shared variance of personality characteristics. This means that the specific personality characteristics that did not contribute much to the shared trait variance were omitted from the models outright. This practice makes perfect sense when one is looking for broader trait dimensions or wants to reduce the dimensionality of data – but it dismisses nuances.

Future work that seeks to identify and use item-level characteristics, nuances, should be based on (large) item pools that have not been tailored to specific personality models or assessment instruments. Likewise, the goal should not necessarily be designing scales with simple structure and high internal consistency, as this entails loss of specific information. Specific characteristics that nevertheless contain 'signal' should be included in personality taxonomies. This echoes a similar suggestion by Condon and Mroczek.

GENERALLY, BRIEF SCALES ARE NO-GOS IF TRAIT-LEVEL EXPLANATIONS ARE SOUGHT

McCrae and Chapman, correctly in my view, pointed out that quests for causal unity do not go well with the use of brief scales to measure broad traits. This is because brief scales 'offer no possibility of determining whether an association is due to the broad trait itself or to the specific items by which it was operationalized' (McCrae). If brief scales are used, then 'interpretation must be restrained to the actual scale content. If a Conscientiousness scale composed of the two trait adjectives 'reliable' and 'organized' shows an association with some inflammatory marker, for instance, interpretation is most safely centered on these particular trait adjectives. They are anchor items of a broader construct, of the other elements of which might be at play, we are simply not sure' (Chapman).

INCORPORATING WITHIN-INDIVIDUAL VARIABILITY AND SITUATIONAL INFLUENCES WILL BE IMPORTANT FUTURE DEVELOPMENTS

Traits summarize (or reflect, depending on what direction of causality seems more plausible) regularities in behaviour thoughts and feelings, but not only in individual differences. People also vary within themselves in trait expression over time and across situations (Sherman, Rauthmann, Brown, Serfass, & Jones, 2015). What tend to be stable within individuals and differ among them are the distributions and patterns (e.g., contingencies with contextual variables) of personality variation (Fleeson, 2001). Technological and other methodological advances are beginning to make it possible to capture the two levels of variability – within and among people – at the same time. I do not have the faintest doubt that simultaneously considering these two variance levels is how the future personality psychology will look.

This has at least two implications for personality-outcome association research. First, it is possible to consider situational influences on these associations, which is facilitated by emerging taxonomies for situation assessment (Rauthmann et al., 2015). As Ziegler & Ziegler pointed out, outcomes are often context-specific and so may be at least some personality manifestations. Modelling within- and between-individual variability at the same time enables us to see how both levels relate to outcomes and whether, among a multitude of other possibilities, context-specific outcomes (e.g., doing sports with friends) are more relevant for context specific personality manifestations (within-individual variance: being more gregarious than usual), whereas general outcomes (e.g., being physically active by taking a high number of steps most days) are more strongly linked with between-individual differences (e.g., being generally lively and energetic).

Second, considering within-individual variance may, in principle, enhance the plausibility of at least some causal inferences because it allows us to temporally separate ostensible causes and effects. For example, those higher in Extraversion tend to be physically active (Rhodes & Smith, 2006). To the extent that there is any causality at all in this association, this may be because either activity or Extraversion is causal, or because they reinforce each other over time. In a study of within-individual variability, Wichers et al. (2012) provided evidence for increases in physical activity being associated with increased positive affect (a component or manifestation of Extraversion in many models) at a later time-point, but not the other way around (not all studies have confirmed this, however; e.g., see Kühnhausen, Leonhardt, Dirk, & Schmiedek, 2013; Dunton et al., 2014). Given that associations at the two levels of variance are interpretable in the same way and this finding will eventually prove reliable, this would corroborate the hypothesis that physically active lifestyle may contribute to higher Extraversion.

I say 'may', because associations at the two levels of explanation do not have to have similar strength or even direction. In principle, it could be that extraverted individuals are generally more active, but mostly when they have been less extraverted than is usual to them (to regain their normal level of extraversion or overcome boredom), yielding a *negative* temporal association at the level of within-individual variance. By default, however, it may make sense to hypothesize that the associations at the two levels are interpretable in the same terms: individuals with low self-discipline find it difficult to avoid unhealthy food and especially so when they are even less self-disciplined than usual. For physical activity, we have found evidence for associations at within-individual variance level reflecting those at between-individuals level (Mõttus, Epskamp, & Francis, in press). However, other researchers may report different findings.

Freese, who explicitly dismissed within-individual variability in traits, discussed the idea that attributes (trait levels in the sense of individual differences, I take it) may not be suitable causal candidates at all, because there is reason to argue that causes have to vary within individuals (cf. Borsboom et al., 2003). We may or may not agree with this as a general principle (I am not sure Freese did), but I suspect

DOI: 10.1002/per

that recognizing that individuals do vary within themselves at least in personality trait expression removes this possible obstacle to causal interpretations.

Naturally, requirements for causal and aetiological unity in trait components also apply to within-individual variability. For example, if particular situational experiences are relevant for only a subset of a trait's manifestations (items or facets) then this subset is the appropriate unit of causal interpretations and it may be erroneous to generalize associations to broader traits to which they do not pertain.

CONCLUSION

In this rejoinder, I deliberately eschewed the question of the inherent nature of (broad) traits. Are traits real psychobiological entities, emergent from causal networks (Costantini & Perugini) or something else? Can composites be causal even when they do not reflect underlying traits or emergent properties? Perhaps in some specific cases they can (van Bork et al.). These are all important questions, but we may not be able to answer or reach consensus on them at present. This is why I wanted to focus on making the case for item-level analyses. On this question, achieving consensus seems more likely and this will have important and immediate practical implications.

At this point, it may indeed be useful to consider personality traits as fictions that may enable us to tell coherent stories, as Revell and Elleman suggested. But some stories are more plausible than others and our job is to find combinations of personality characteristics that tell the most plausible ones – even if they are more complicated than we initially hoped. In the process, we might even start to identify the inherent nature of traits.

ACKNOWLEDGEMENTS

I am grateful for Wendy Johnson and Tom Booth for their suggestions regarding this rejoinder.

REFERENCES

Albert, R., Jeong, H., & Barabási, A.-L. (1999). Diameter of the world-wide web. *Nature*, *401*, 130–131. doi:10.1038/43601.

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., & Bong, C. H. (2014). *Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour*.

Ashton, M. C., & Lee, K. (2005). Honesty-humility, the Big Five, and the Five-Factor Model. *Journal of Personality*, *73*, 1321–1353.

Bedau, M. A. (2003). Downward causation and autonomy in weak emergence. *Principia*, *6*(1), 5–50.

Bedau, M. A. (2008). Is weak emergence just in the mind? *Minds and Machines*, *18*(4), 443–459. doi:10.1007/s11023-008-9122-6.

Bedau, M. A. (2012). Weak emergence and computer simulation. In P. Humphreys, & C. Imbert (Eds.), *Models, simulations, and representations* (pp. 91–114). New York: Routledge.

Bejar, I. I. (1983). *Achievement testing: Recent advances*. Beverly Hills, CA, USA: Sage Publications.

Bhalla, U. S., & Iyengar, R. (1999). Emergent properties of networks of biological signaling pathways. *Science*, *283*(5400), 381–387. doi:10.1126/science.283.5400.381.

Bjornebekk, A., Fjell, A. M., Walhovd, K. B., Grydeland, H., Torgersen, S., & Westlye, L. T. (2013). Neuronal correlates of the five factor model (FFM) of human personality: Multimodal imaging in a large healthy sample. *NeuroImage*, *65*, 194–208. doi:10.1016/j.neuroimage.2012.10.009.

Bleidorn, W., Kandler, C., Riemann, R., Angleitner, A., & Spinath, F. M. (2009). Patterns and sources of adult personality development: Growth curve analyses of the NEO PI-R scales in a longitudinal twin study. *Journal of Personality and Social Psychology*, *97*, 142–155.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, *117*(2), 187–215. doi:10.1037/0033-2909.117.2.187.

Block, J. (2001). Millennial contrarianism: The Five-Factor approach to personality description 5 years later. *Journal of Research in Personality*, *35*, 98–107.

Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, *21*(1), 2–25. doi:10.1080/10478401003596626.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. doi:10.1037/0033-2909.110.2.305.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *PNAS*, *99*(suppl. 3), 7280–7287. doi:10.1073/pnas.082080899.

Borkenau, P., Riemann, R., Angleitner, A., & Spinath, F. M. (2001). Genetic and environmental influences on observed personality: Evidence from the German Observational Study of Adult Twins. *Journal of Personality and Social Psychology*, *80*, 655–668.

Borsboom, D. (2009). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Borsboom, D., & Dolan, C. V. (2006). Why *g* is not an adaptation: A comment on Kanazawa (2004). *Psychological Review*, *113*(2), 433–437. doi:10.1037/0037/0033-295X.113.2.433.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. doi:10.1037/0033-295X.110.2.203.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Bradley, F. H. (1966). *Appearance and reality: A metaphysical essay* (2nd ed.). London: Oxford University Press (Original work published 1897).

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement*, *10*, 159–183.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*(3), 193–217.

Cacioppo, J. T., Semin, G. R., & Berntson, G. G. (2004). Realism, instrumentalism, and scientific symbiosis: Psychological theory as a search for truth and the discovery of solutions. *American Psychologist*, *59*(4), 214.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*, 404–431.

Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology*, *56*(4), 577–585.

Cattell, R. B. (1943). The description of personality. I. Foundations of trait measurement. *Psychological Review*, *50*(6), 559–594. doi:10.1037/h0057276.

Cattell, R. B. (1945). The description of personality: Principles and findings in a factor analysis. *The American Journal of Psychology*, *58*(1), 69–90. doi:10.2307/1417576.

Cervone, D. (2005). Personality architecture: Within-person structures and processes. *Annual Review of Psychology*, *56*, 423–452.

Chapman, B. P., Hampson, S., & Clarkson, J. (2014). Personality intervention for healthy aging: Conclusions from a National Institute on Aging Workgroup. *Developmental Psychology*, *50*(5), 1411–1426.

Chapman, B. P., Weiss, A., Fiscella, K., Muennig, P., Kawachi, I., & Duberstein, P. (2015). Mortality risk prediction: Can comorbidity indices be improved with psychosocial data? *Medical Care*, *53*(11), 909–915.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219–251. doi:10.1111/j.1467-6494.2011.00739.x.

Condon, D. M. (2014). *An organizational framework for the psychological individual differences: Integrating the affective, cognitive, and conative domains (Unpublished doctoral dissertation)*. Northwestern University.

Condon, D. M., & Revelle, W. (2015). Selected personality data from the SAPA-Project: On the structure of phrased self-report items. *Journal of Open Psychology Data*, *3*(1). doi:10.5334/jopd.al.

Connelly, B. S. (2008). *The reliability, convergence, and predictive validity of personality ratings: An other perspective (Doctoral dissertation)*. University of Minnesota, Minneapolis, MN. Retrieved from University of Minnesota Digital Conservancy. (60223)

Connelly, B. S., Wilmot, M. P., Hülsheger, U. R., Ones, D. S., & DeYoung, C. G. (2016). *Predicting work performance from hierarchical personality traits: A multi-informant study*. Manuscript submitted for publication.

Corr, P. J. (2008). Reinforcement Sensitivity Theory (RST). In P. J. Corr (Ed.), *The reinforcement sensitivity theory of personality* (pp. 1–43). Cambridge University Press: Cambridge.

Corr, P. J., & Poropat, A. E. (2016). Personality assessment and theory. In U. Kumar (Ed.), *The Wiley Handbook of Personality Assessment* (pp. 19–30). Oxford U.K.: John Wiley & Sons.

Costa, P. T. Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, *64*, 21–50.

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

Costantini, G., & Perugini, M. (2012). The definition of components and the use of formal indexes are key steps for a successful application of network analysis in personality psychology. *European Journal of Personality*, *26*(4), 434–435. doi:10.1002/per.1869.

Costantini, G., Richetin, J., Borsboom, D., Fried, E. I., Rhemtulla, M., & Perugini, M. (2015b). Development of indirect measures of conscientiousness: combining a facets approach and network analysis. *European Journal of Personality*, *29*(5), 548–567. doi:10.1002/per.2014.

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*(4), 414–431. doi:10.1002/per.1866.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*, 874–888.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Darkes, J., Greenbaum, P. E., & Goldman, M. S. (1998). Sensation seeking–disinhibition and alcohol use: Exploring issues of criterion contamination. *Psychological Assessment*, *10*(1), 71–76.

DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., & Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of General Internal Medicine*, *21*(3), 267–275.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social Psychology*, *91*(6), 1138–1151. doi:10.1037/0022-3514.91.6.1138.

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, *56*, 33–58 http://doi.org/10/33h.

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 http://doi.org/10/b4xsvz.

Dunton, G. F., Huh, J., Leventhal, A. M., Riggs, N., Hedeker, D., Spruijt-Metz, D., & Pentz, M. A. (2014). Momentary assessment of affect, physical feeling states, and physical activity in children. *Health Psychology*, *33*, 255–263. doi:10.1037/a0032640.

Ellis, A. (1987). The impossibility of achieving consistently good mental health. *American Psychologist*, *42*, 364–375.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360–392. doi:10.1111/j.1467-6494.1983.tb00338.x.

Epstein, S. (1994). Trait theory as personality theory: Can a part be as great as the whole? *Psychological Inquiry*, *5*(2), 120–122.

Eysenck, H. J. (1967). *The biological basis of personality*. Springfield: Thomas.

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

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92.

Fried, E. I., van Borkulo, C. D., Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). *Measuring Depression Over Time . . . or not? Lack of Unidimensionality and Longitudinal Measurement Invariance in Four Common Rating Scales of Depression*. *Psychological Assessment*, *No Pagination Specified*. doi:10.1037/pas0000275

Friedman, H. S., & Hudson, L. R. (2011). *The longevity project*. New York, NY: Hudson Street Press.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, *101*, 75–90.

Funder, D. C. (1991). Global traits: A neo-Allportian approach to personality. *Psychological Science*, *2*, 31–39.

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

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences*, *84*, 84–89. doi:10.1016/j.paid.2014.08.019.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. doi:10.1037/1040-3590.4.1.26.

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development* (pp. 169–188). Washington, DC: American Psychological Association.

Goldberg, L. R., & Saucier, G. (2016). *The Eugene-Springfield community sample: Information available from the research participants (Tech. Rep. No. 56-1)*. Eugene, Oregon: Oregon Research Institute.

Gray, J. A. (1981). A critique of Eysenck's theory of personality. In H. J. Eysenck (Ed.), *A model for personality* (pp. 246–277). Berlin: Springer.

Gray, J. A., & McNaughton, N. (2000). *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system*. Oxford: Oxford University Press.

Hamaker, E. L., Nesselroade, J. R., & Molenaar, P. C. M. (2007). The integrated trait-state model. *Journal of Research in Personality*, *41*, 295–315. doi:10.1016/j.jrp.2006.04.003.

Hammond, K. R. (1954). Representative vs. systematic design in clinical psychology. *Psychological Bulletin*, *51*(2), 150–159.

Hampson, S. E. (2012). Personality processes: Mechanisms by which personality traits "get outside the skin.". *Annual Review of Psychology*, *63*(1), 315–339. doi:10.1146/annurev-psych-120710-100419.

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big-Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163. doi:10.1037/0022-3514.63.1.146.

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior*, *7*, 627–637.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*, 945–960.

Holland, Paul W. (2003). "*Causation and race. Educational Testing Service Research Report RR-03-03*.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*, 205–218.

Hughes, S., De Houwer, J., & Perugini, M. (2016). The functional-cognitive framework for psychological research: Controversies and resolutions. *International Journal of Psychology*, *51*(1), 4–14. doi:10.1002/ijop.12239.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jackson, J. J., Bogg, T., Walton, K. E., Wood, D., Harms, P. D., Lodi-Smith, J., et al. (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*(2), 446–459. doi:10.1037/a0014156.

Jackson, J. J., Hill, P. L., Payne, B. R., Roberts, B. W., & Stine-Morrow, E. A. L. (2012). Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults. *Psychology and Aging*, *27*, 286–292. doi:10.1037/a0025918.

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*(4), 501–511. doi:10.1016/j.jrp.2010.06.005.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: theory and research* (3rd ed., pp. 114–158). New York: Guilford Press.

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. *Journal of Applied Psychology*, *98*(6), 875–925. doi:10.1037/a0033901.

Kandler, C., Riemann, R., Spinath, F. M., & Angleitner, A. (2010). Sources of variance in personality facets: A multiple-rater twin study of self-peer, peer-peer, and self-self (dis)agreement. *Journal of Personality*, *78*, 1565–1594.

Kelley, T. L. (1927). *Interpretation of educational measurements*. Oxford: World Book Company.

Kenrick, D. T., & Funder, D. C. (1988). Profiting from controversy: Lessons from the person-situation debate. *American Psychologist*, *43*, 23–34.

Krueger, R. F., Markon, K. E., Patrick, C. J., & Iacono, W. G. (2005). Externalizing psychopathology in adulthood: A dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology*, *114*(4), 537–550. doi:10.1037/0021-843X.114.4.537.

Kühnhausen, J., Leonhardt, A., Dirk, J., & Schmiedek, F. (2013). Physical activity and affect in elementary school children's daily lives. *Frontiers in Psychology*, *4*, 456. doi:10.3389/fpsyg.2013.00456.

Lane, S. P., & Sher, K. J. (2015). Limits of current approaches to diagnosis severity based on criterion counts: An example with DSM-5 Alcohol Use Disorder. *Clinical Psychological Science*, *3*, 819–835.

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

Lasky, J. J., Hover, G. L., Smith, P. A., Bostian, D. W., Duffendack, S. C., & Nord, C. L. (1959). Post-hospital adjustment as predicted by psychiatric patients and by their staff. *Journal of Consulting Psychology*, *23*, 213–218.

Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, *12*(1), 165–200 http://doi.org/10/c9qbtd.

Leitsalu, L., Haller, T., Esko, T., Tammesoo, M.-L., Alavere, H., Snieder, H., … Metspalu, A. (2015). Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, *44*, 1137–1147. doi:10.1093/ije/dyt268.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports Monograph Supplement*, *9*(3), 635–694. doi:10.2466/pr0.1957.3.3.635.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

McAbee, S. T., Oswald, F. L., & Connelly, B. S. (2014). Bifactor models of personality and college student performance: A broad versus narrow view. *European Journal of Personality*, *28*(6), 604–619. doi:10.1002/per.1975.

McCrae, R. R. (2015). A more nuanced view of reliability: Specificity in the trait hierarchy. *Personality and Social Psychology Review*, *19*, 97–112. doi:10.1177/10888683 14541857.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*(5), 509–516. doi:10.1037/0003-066X.52.5.509.

McCrae, R. R., & Costa, P. T. Jr. (2008). The Five-Factor Theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159–181). New York: Guilford.

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

McCrae, R. R., Costa, P. T., de Lima, M. P., Simões, A., Ostendorf, F., Angleitner, A., … & Chae, J. H. (1999). Age differences in personality across the adult life span: Parallels in five cultures. *Developmental Psychology*, *35*, 466–477.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scaleM validity. *Personality and Social Psychology Review*, *15*(1), 28–50.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, *55*(4), 675–680.

Mike, A., Harris, K., Roberts, B. W., & Jackson, J. J. (2015). Conscientiousness. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 658–665). Elsevier http://doi.org/10.1016/B978-0-08-097086-8.25047-2.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference*. Cambridge University Press.

Mõttus, R. (2016). Towards more rigorous personality trait-outcome research. *European Journal of Personality*.

Mõttus, R., Allik, J., Realo, A., Pullmann, H., Rossier, J., Zecca, G., … et. al. (2012). Comparability of self-reported conscientiousness across 21 countries. *European Journal of Personality*, *26*(3), 303–317.

Mõttus, R., Epskamp, S., & Francis, A. (in press). Within- and between individual variability of personality characteristics and physical exercise. *Journal of Research in Personality*. doi:10.1016/j.jrp.2016.06.017

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (in press). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*.

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. doi: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. doi:10.1371/journal.pone.0119667.

Nicholls, J. G., Licht, B. G., & Pearl, R. A. (1982). Some dangers of using personality questionnaires to study personality. *Psychological Bulletin*, *92*(3), 572–580.

Ones, D. S., Chockalingham, V., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance*, *18*(4), 389–404.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995–1027. doi:10.1111/j.1744-6570.2007.00099.x.

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*(6), 669–676. doi:10.1016/j.paid.2010.06.014.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (Second ed.). Cambridge: Cambridge University Press.

Perugini, M., Costantini, G., Hughes, S., & De Houwer, J. (2016). A functional perspective on personality. *International Journal of Psychology*, *51*(1), 33–39. doi:10.1002/ijop.12175.

Pervin, L. A. (1994). A critical analysis of current trait theory. *Psychological Inquiry*, *5*(2), 103–113.

Poropat, A. E. (2015). Personality and educational outcomes. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., pp. 787–791), *17*. Oxford: Elsevier.

Poropat, A. E. (2016). Beyond the shadow: The role of personality and temperament in learning. In L. Corno, & E. Anderman (Eds.), *Handbook of educational psychology* (pp. 172–185). Washington D.C: American Psychological Association, Division 15 - Educational Psychology.

Poropat, A. E., & Corr, P. J. (2015). Thinking bigger: The Cronbachian paradigm & personality theory integration. *Journal of Research in Personality*, *56*(1), 59–69. doi:10.1016/j.jrp.2014.10.006.

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

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*(1), 61–92. doi:10.1037/a0018131.

Revelle, W. (1983). Factors are fictions, and other comments on individuality theory. *Journal of Personality*, *51*(4), 707–714. doi:10.1111/1467-6494.ep7380795.

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 N. G. Fielding, R. M. Lee, & G. Blank (Eds.), *Sage handbook of online research methods (chap. 37: Mobile Methods)*. Sage Publications, Inc. (in press).

Revelle, W., Wilt, J., & Rosenthal, A. (2010). Individual differences in cognition: New methods for examining the personality-cognition link. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition: Attention, memory and executive control* (pp. 27–49). New York, N. Y.: Springer.

Rhodes, R. E., & Smith, N. E. I. (2006). Personality correlates of physical activity: A review and meta-analysis. *British Journal of Sports Medicine*, *40*, 958–965. doi:10.1136/bjsm.2006.028860.

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*(4), 313–345. doi:10.1111/j.1745-6916.2007.00047.x.

Royce, J. R. (1983). Personality integration: A synthesis of the parts and wholes of individuality theory. *Journal of Personality*, *51*(4), 683–706. doi:10.1111/j.1467-6494.1983.tb00874.x.

Salgado, J. F., Moscoso, S., Sanchez, J. I., Alonso, P., Choragwicka, B., & Berges, A. (2015). Validity of the five-factor model and their facets: The impact of performance measure and facet residualization on the bandwidth-fidelity dilemma. *European Journal of Work and Organizational Psychology*, *24*(3), 325–349. doi:10.1080/1359432X.2014.903241.

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality*, *36*, 1–31.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*(2), 206–224. doi:10.1037/1082-989X.8.2.206.

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*, 43–53.

Schnabel, K., & Asendopf, J. B. (2015). Cognitive trainings reduce implicit social rejection associations. *Journal of Social and Clinical Psychology*, *34*, 365–391.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press ADDIN ZOTERO_BIBL {"custom":[]} CSL_BIBLIOGRAPHYX.

Sherman, R. A., & Funder, D. C. (2009). Evaluating correlations in studies of personality and behavior: Beyond the number of significant findings to be expected by chance. *Journal of Research in Personality*, *43*(6), 1053–1063. doi:10.1016/j.jrp.2009.05.010.

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

Sherman, R. A., & Wood, D. (2014). Simple and intuitive statistics for calculating the expected replicability of a pattern of correlations. *Multivariate Behav. Res*, *49*, 17–40.

Smillie, L. D. (2008). What is reinforcement sensitivity? Neuroscience paradigms for approach-avoidance process theories of personality. *European Journal of Personality*, *22*(5), 359–384. doi:10.1002/per.674.

Soto, C. J. (in press). The Little Six personality dimensions from early childhood to early adulthood: Mean-level age and gender differences in parents' reports. *Journal of Personality*. doi:10.1111/jopy.12168.

Spearman, C. (1904). \General Intelligence," objectively determined and measured. *American Journal of Psychology*, *15*(2), 201–292. doi:10.2307/1412107.

Spearman, C. (1927). *The abilities of man*. London: MacMillan.

Steinley, D., Lane, S. P., Sher, K. J. (2016). Determining optimal diagnostic criteria through chronicity and comorbidity. In *Silico Pharmacology*, *4*(1), 1. doi:10.1186/s40203-016-0015-8.

Suls, J. M., & Bunde, J. (2005). Anger, anxiety, and depression as risk factors for cardiovascular disease: The problems and implications of overlapping affective dispositions. *Psychological Bulletin*, *131*(2), 260–300. doi:10.1037/0033-2909.131.2.260.

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*, 579–592. doi:10.1037/a0024286.

Tang, T. Z., DeRubeis, R. J., Hollon, S. D., Amsterdam, J., Shelton, R., & Schalet, B. (2009). Personality change during depression treatment: A placebo-controlled trial. *Archives of General Psychiatry*, *66*, 1322–1330. doi:10.1001/archgenpsychiatry.2009.166.

Terracciano, A., Sutin, A. R., McCrae, R. R., Deiana, B., Ferrucci, L., Schlessinger, D., … & Costa Jr, P. T. (2009). Facets of personality linked to underweight and overweight. *Psychosomatic Medicine*, *71*, 682–689.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology*, *88*, 500–517. doi:10.1037/0021-9010.88.3.500.

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of vectors of the mind*. Chicago, IL: University of Chicago Press.

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

Turiano, N. A., Chapman, B. P., Gruenewald, T. L., & Mroczek, D. K. (2015). Personality and the leading behavioral contributors of mortality. *Health Psychology*, *34*(1), 51–60. doi:10.1037/hea0000038.

Urbina, S. (2011). Tests of intelligence. In R. J. Sternberg, & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 20–38). Cambridge, U.K.: Cambridge University Press.

Vainik, U., & Mõttus, R. (2016). *ionr: Test for indifference of indicator (version 0.3.0)*. Retrieved from https://cran.r-project.org/web/packages/ionr/index.html

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

van de Leemput, I. A., Wichers, M., Cramer, A. O. J., Borsboom, D., Tuerlinckx, F., Kuppens, P., … Scheffer, M. (2014). Critical slowing down as early warning for the onset and termination of depression. *PNAS*, *111*(1), 87–92. doi: 10.1073/pnas.1312114110 ADDINZOTERO_BIBL{"custom":[]}CSL_BIBLIOGRAPHYX

Vergés, A., Steinley, D., Trull, T. J., & Sher, K. J. (2010). It's the algorithm! Why differential rates of chronicity and comorbidity are not evidence for the validity of the abuse-dependence distinction. *Journal of Abnormal Psychology*, *119*, 650–661.

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I.-S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial and Organizational Psychology*, *7*(4), 507–518 http://doi.org/10/5k5.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x.

Weston, S. J., & Jackson, J. J. (2015). Identification of the healthy neurotic: Personality traits predict smoking after disease onset. *Journal of Research in Personality*, *54*(Special Issue), 61–69. doi:10.1016/j.jrp.2014.04.008.

Wichers, M., Peeters, F., Rutten, B. P. F., Jacobs, N., Derom, C., Thiery, E., … van Os, J. (2012). A time-lagged momentary assessment study on daily life physical activity and affect. *Health Psychology*, *31*, 135–144. doi:10.1037/a0025688.

Wiernik, B. M., Kostal, J. W., Wilmot, M. P., & Ones, D. S. (2016). *Variance decomposition of typical other-rated Conscientiousness facet measures* [Technical report]. doi: 10.6084/m9.figshare.3100795

Wiernik, B. M., Wilmot, M. P., & Kostal, J. W. (2015). How data analysis can dominate interpretations of dominant general factors. *Industrial and Organizational Psychology*, *8*(03), 438–445. doi:10.1017/iop.2015.60.

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. 505–560). New York, NY: Plenum Press.

Wood, D. (2007). Using the PRISM to compare the explanatory value of general and role-contextualized trait ratings. *Journal of Personality*, *75*(6), 1103–1126.

Wood, D., Gardner, M. H., & Harms, P. D. (2015). How functionalist and process approaches to behavior can explain trait covariation. *Psychological Review*, *122*(1), 84–111. doi:10.1037/a0038423.

Wood, D., & Wortman, J. (2012). Trait means and desirabilities as artifactual and real sources of differential stability of personality traits. *Journal of Personality*, *80*(3), 665–701.

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality*, *19*(5), 373–390. doi:10.1002/per.542.

Ziegler, M. (2014). Stop and state your intentions!: Let's not forget the ABC of test construction. *European Journal of Psychological Assessment*, *30*, 239–242. doi:10.1027/1015-5759/a000228.

Ziegler, M., Bensch, D., Maaß, U., Schult, V., Vogel, M., & Bühner, M. (2014). Big Five facets as predictor of job training performance: The role of specific job demands. *Learning and Individual Differences*, *29*, 1–7. doi:10.1016/j.lindif.2013.10.008.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*, 231–237. doi:10.1027/1015-5759/a000309.