Open Peer Commentary and Author's Response

Author's Rejoinder

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, for that matter), 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.

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 want 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 both the target article and the comments: ***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 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 the author of the target article 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 might suffice in some cases; rather, it means supplementing scale-level with item-level analyses. 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 of Them

1. 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 the

interpretations of observed associations on the level of analysis that yields most consistent results and not to invoke hypothetical higher-order constructs when there is no consistent evidence pointing to their possible relevance. Regardless of whether we think we have evidence for the veracity of the higher-order traits in the first place, 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 behaviours, 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 fictitious, whose relevance is not given 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.

2. 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 from traits to outcomes. 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.

3. 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 enough 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, Mõttus, Allik, Esko, & Realo, 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).

4. Investigating item-level correlations can help to identify possible measurement/construct overlaps (Sher; Wood & Harms). 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 each item often reflects specific personality characteristics over and above these broader traits. McCrae (2015) has called these specific personality characteristics "nuances". Mõttus and colleagues (in press; 2014) reported that the unique variances of most of the 240 NEO-PI items (after being residualized for the variances of facets and domains) had significant rank-order stability and cross-rater agreement, and a large proportion of them also displayed 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 description of individual differences and predictions of outcomes. Perhaps many of them do only 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., 2014; 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 into the mix 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 are 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 easily up on self-improvement. Period.

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 and colleagues (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 the "bad apples" are removed (if there are any). Also, the associations of single items with outcomes are plotted (see Figure 1). Admittedly, this procedure does not take into account the variability 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 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.

== Insert Figure X here ==

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 the associations can be tested in multiple partitions of a sample to test their robustness. Also, for parsimonious models that also 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 intended 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 the 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.

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 with age. 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 amount of how personality characteristics relate to age may pertain to items' unique rather than shared variance.

Examination at the item level 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 differences 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, the 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 the manifestations of the trait. 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 correlating 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 too 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, the unambiguous readability of items is paramount. Second, an important formal property of items is their 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 responses 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 item. Such problems could be fixed by writing items with reasonable (medium) levels of "difficulty" and by reducing social desirability in 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 beyond those thought to be most representative or otherwise relevant), have been omitted from them in the first place (for a good example of such 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 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, Sherman, & Funder, 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 usually), 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 plausibility of causal inferences because it allows us to temporally separate ostensible causes and outcomes. 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 and colleagues (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 et al., 2013, Dundon 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 the associations at the two levels of explanation do not have to be similar. 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 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 revision).

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 causes have to vary within individuals (cf. Borsboom, Mellenbergh, & van Heerden, 2003). We may or may not agree with this general principle (I am not sure Freese did), but I suspect 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 & Perguinin) or something else? Can composites be causal even when they do not reflect underlying traits or emergent properties? Perhaps in some specific cases they do (van Bork et al.). These are all important questions, but we may not be able to solve them at this point, or reach a consensus on them. 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 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

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

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. http://doi.org/10.1037/a0032640

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. http://doi.org/10.1037/0022-3514.80.6.1011

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. http://doi.org/10.1037/a0025918

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. http://doi.org/10.3389/fpsyg.2013.00456

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. http://doi.org/10.1093/ije/dyt268

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

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

Mõttus, R., Epskamp, S., & Francis, A. (in revision). Within- and between individual variability of personality characteristics and physical exercise. *Journal of Research in Personality.*

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. http://doi.org/10.1016/j.jrp.2014.07.005

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

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

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. http://doi.org/10.1136/bjsm.2006.028860

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. http://doi.org/10.1037/pspp0000036

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*. http://doi.org/10.1111/jopy.12168

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. http://doi.org/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. http://doi.org/10.1001/archgenpsychiatry.2009.166

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. http://doi.org/10.1111/j.1467-9868.2011.00771.x

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 personality facets and BMI. *European Journal of Personality*, *29*, 622–634. http://doi.org/10.1002/per.2009

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*,