Brent Donnellan, Editor, *Collabra: Psychology*

Dr. Donnellan,

We appreciate the feedback that was given for our manuscript entitled, “Smaller is better: Associations between personality and demographics are improved by examining narrower traits and regions.” It has been helpful in improving our paper. Below are the revision notes:

**RESPONSES TO DR. DONNELLAN’S COMMENTS**

1. Reviewer #1 suggested that some of the writing could be made more accessible. I think this is a valuable reaction and I think a cosmetic revision of the text could increase the accessibility of the work.

   *We agree. We have made a number of cosmetic improvements to the paper for greater accessibility. Many of those improvements are mentioned below.*

2. It might be useful to make a parallel between the approach of focusing on nuances and the ‘old school’ approach of what Block & Block and associates used to do by examining how different Q-sort items were associated with variables of interest. (Assuming you think this comparison has any merit.)

   *We think a more appropriate comparison is to the Minnesota Multiphasic Personality Inventory (MMPI); scales are built with self-reported Likert-like items that most highly correlate with a criterion, with no necessary underlying theory. Q-sort is similar in the sense that individual items are used, but is dissimilar regarding the sorting of cards into piles and the focus on a judge’s perspective on some subject matter (perhaps another person or oneself). We have updated the manuscript with a reference to the MMPI: “This type of scale has also been referred to as a ‘polyitem score’ (Möttus, Bates, et al., 2018) and is similar to the dustbowl empiricism of the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1942).” (pp. 5–6).*

3. Some reconciliation of the focus on nuances versus concerns about the amount of error in items might be worth discussing (Reviewer #2).

   *We have updated our literature review to make a clearer case concerning item reliability/error: “Personality researchers have recently delved into even narrower traits, or ‘nuances,’ which are the breadth of individual items (McCrae, 2015). Concerns about the low reliability of items have been answered with large-sample studies indicating that*
nuances are reliable and valid constructs; the unique variance of individual items have cross-rater agreement, are rank-order stable, and are heritable (MÖTTUS, McCRae, AllIK, & ReALO, 2014; MÖTTUS, KANDLER, BLEIDORN, RIEMANN, & McCRae, 2017; MÖTTUS, SINICK, et AL., 2018)” (p. 4).

4. I did not love the approach of moving to facet-level analyses if a domain was predictive (see also Reviewer #2). I recall some older Ashton and Paunonen papers that was pushed the value of facets and made a point that sometimes facets might have criterion-associations that differed in sign within a domain (see references below). If such a pattern happens, the overall domain effect could be close to zero. I also understand the need to control for Type 1 errors so I do not have a fool-proof way to navigate this issue, but I think you should consider the pros and cons of the approach in the paper with one that looks at facets regardless of domains.

We agree. In the first manuscript we sent to Collabra, we performed a post-hoc analysis to ensure that we didn’t miss any facets that may have been zeroed-out at the domain level due to opposing signs within a domain. Because those results were null (there were no interesting findings for those facets), we moved those analyses to a footnote and the supplementary materials. We have improved the language of the footnote to make it clear why we performed those analyses and have moved the footnote to the body of the text: “Because facets within a domain could have been differentially correlated with a demographic such that the opposing effects would have cancelled each other out at the level of the domain, we also found facet-level correlations for all domains that were not meaningfully correlated with a ZIP Code demographic: facets of Extraversion and Neuroticism with population density (Figure 5 in the supplemental materials) and income disparity (Figure 6 in the supplemental materials), as well as all 30 facets with median income (Figure 7 in the supplemental materials) and ethnic diversity (Figure 8 in the supplemental materials). None of these correlations reached the threshold for a notable effect size (|r| = .15)” (pp. 15–16).

5. Some of the screening decisions on pages 6 to 8 could use stronger justification. Or If this is standard lab practice, feel free to note this in the paper.

We have updated our manuscript as follows: “To limit the sample to those living in the U.S. who, more likely than not, had a choice concerning where they resided, we included participants only if they provided a valid U.S. ZIP Code of their residence and indicated they were 18 years of age or older. Participants were excluded if they were 18 years old and indicated they had not finished high school, due to our assumption that they would still be residing in their parents’ home (i.e., they did not yet have the choice concerning where to reside). As was standard lab practice, we removed participants who reported having previously taken the assessment and duplicate entries taken in the same internet browser session” (pp. 6–7).

6. Likewise, why the 10 participant per ZIP code rule? I don’t necessarily disagree, I just thought it could be better justified.

We have updated our manuscript as follows: “To lower the error associated with aggregated personality estimates of ZIP Codes, ZIP Codes were excluded if they contained less than 10 participants, as this cutoff improved the reliabilities of ZIP Code measures while retaining a large number of participants and ZIP Codes.[FOOTNOTE: See Tables 1, 2, and 3 in the supplemental materials for analyses related to different minimums of
participants per ZIP Code)” (p. 7).

7. When thinking about the amount of variance accounted for by ZIP code versus state, it might be useful to reference the neighborhood effects literature. I remember an older article from 2000 that I appended below.

We appreciate the citation regarding neighborhood effects. We have included a reference to it at the end of the section in the Introduction that discusses why ZIP Codes may better aggregate personality than states: “Additionally, characteristics of a neighborhood, such as its affluence or ethnic diversity, impact the individual differences of resident children and adolescents, such as their cognitive ability and behavioral problems (for a review, see Leventhal & Brooks-Gunn, 2000)” (p. 4).

8. It seems like political liberalism is the biggest effect so you might want to grapple with whether this is an attitude or trait per-se. (See also Reviewer #1).

We think that the question of whether Liberalism is a political attitude or a personality trait is best saved for a paper dedicated to the issue. Liberalism was a stated facet of Openness, and so was, by definition, a personality trait. Additionally, the typical definition of personality is that it is composed of stable patterns of thinking, feeling, and behaving, so Liberalism falls within this definition. A paper that would get to the bottom of the attitude-vs-trait question would need to critique the structure and item content of the NEO-PI-R/IPPI-NEO, as well as clearly distinguish attitudes from the aspect of personality that includes patterns of thinking. We believe this issue is outside the scope of our article. However, our paper is clear that Liberalism is a large driver of the demographic-personality relationships we found, which is a good place to start for researchers interested in making the case that the Liberalism facet is not in fact a personality trait.

9. I agree with Reviewer #3 who suggested offering a more stringent discussion of confounds.

Our goal in including the potential confounds was to preemptively answer reviewers’ questions about them (e.g., “but couldn’t X, Y, or Z account for those relationships?”). However, reviewers had the opposite reaction and felt there was no justification for including the confounds in the first place. We are happy to oblige reviewers by removing all confounds from our models in the multiple regression section of our analyses (Tables 4 and 5). The results are not substantially altered—domain-, facet-, and item-based models for a demographic account for similar amounts of variance, although it appears that nuances are slightly better than facets, which are slightly better than domains. We have updated all mentions of those analyses, as well. We have moved the original confound-filled models to the supplementary material and have referenced them with a small footnote: “ZIP Codes varied by participant demographics: educational attainment, parents’ education, average age, gender ratio, and percentage with minority status. For regression models that include those variables as covariates, see Tables 6 and 7 of the supplemental materials” (p. 19). If Dr. Donnellan would like for us to remove those analyses from the supplemental materials as well, we are happy to do so.

10. I like the constraints on generality statement paper from Simons, Shoda, and Lindsay (2017) and encourage you to include a dedicated section about such constraints in the Discussion.
We have included a dedicated section in the discussion concerning constraints on generality: “In accordance with recommendations by Simons, Shoda, & Lindsay (2017), we acknowledge the following constraints on generality. One constraint concerns sample representativeness. Although the sample was ethnically diverse, participants tended to be more educated than the U.S. population. Thus, it is likely that this study undersampled individuals of lower socioeconomic status. In general, online samples are not representative of the U.S. population, but tend to be more diverse than traditional samples from psychological research (Gosling, Vazire, Srivastava, & John, 2004). We expect for our findings to replicate in other large samples collected by personality psychologists, but a truly representative sample of the U.S. could show different results regarding ZIP Code personality-demographic relationships. We expect for personality to cluster by U.S. ZIP Codes, and for this effect to be stronger than for U.S. states, so long as the new sample is at least as diverse as the current one. Both clustering effects may be stronger in a more representative sample due to the current sample having a somewhat restricted range (specifically in terms of educational attainment). We expect for our results to generalize to a measure of neighborhoods (e.g., U.S. census tracts), but perhaps not geographic units as large as U.S. states, since a state-wide measure of population density would not accurately capture the density of cities or rural areas within the state. As our data were self-reported, we expect our results to generalize to self-reported measures of personality. Although we performed sensitivity analyses to determine the extent to which results were robust against including more or fewer ZIP Codes, substantially diverging from the number of ZIP Codes in this study (either more or fewer) could impact the extent to which these results will replicate in a new study. In particular, the results of this study may not generalize to an examination of ZIP Code personality-demographic relationships within a single metropolitan area. A replication of our nuance-level results would require a large pool of items from multiple inventories. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context” (pp. 25–26)

RESPONSES TO THE FIRST REVIEWER’S COMMENTS

1. I would encourage the authors to condense and improve their writing to make it less technical. In many parts the text reads almost as legal or administrative text. But you want the text to be easy to follow with minimal effort. For example, “For each demographic, an empirical personality scale (also called a “polyitem score”) was created with each set of items, and we contrasted this list of items with the applicable facets.” The reader is not yet familiar with the concept of ‘empirical personality scales’ and it is unclear how ‘applicable facets’ can be ‘contrasted’ with ‘each set of items’.

We have cleaned up this section to as follows: “For each demographic, we created a scale composed of these best items (a ‘best-items scale’). This type of scale has also been referred to as a ‘polyitem score’ (Mõttus, Bates, et al., 2018) and is similar to the dustbowl empiricism of the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1942)” (pp. 5–6).

Or “Each empirical personality scale was input into a series of multiple regression models to determine if the scale accounted for unique variance of the respective demographic...” It is unclear what the ‘series of multiple regression models’ is, how does ‘unique variance’ differ from just variance, or what the ‘respective demographic’ refers to. Or “Additionally,
each final multiple regression model that used an empirical personality scale accounted for a substantial amount of variance in the corresponding demographic…” It is difficult to locate who is doing what to whom.

We have cleaned up this section to as follows: “To compare how strongly ZIP Code population density and income disparity were related to personality at the nuance, facet, and domain levels, we created three linear regression models for each criterion, with each model using a different level of personality” (p. 19).

The text can be simplified by removing technical language and by focusing on clear, active sentences.

We agree, and we appreciate your making us aware of these issues.

Also, there is no need to repeat all the technical information after it has been described once. For example, I wouldn’t repeat the “r ≥ 0.11 and being in the top 34 items in at least 99% of iterations” (pp. 17-18) each time when describing the results for different outcomes. These numbers burden the reader’s attention without providing new information, so the reader just has to spend energy on ignoring those numbers. Small is better, and less is more.

We have cleaned up this section. For example: “For population density, the average training multiple correlation was R = .38 and the average validation correlation was R = .33. There were 21 personality items selected” (p. 16).

2. On the positive side, the discussion is structured nicely to emphasize the main conclusions!

Thank you.

3. The term “empirical personality scale” is perhaps not the best description because all of your data are empirical, and it is not clear how a personality scale could be non-empirical.

We have updated our manuscript to refer to this kind of scale as a “best-items scale” and have changed the text in the Introduction as follows: “To determine personality-demographic relationships at the nuance level, we first found the personality items that were most highly correlated with each demographic. For each demographic, we created a scale composed of these best items (a “best-items scale’)” (p. 5).

4. The subtitle on page 22 says “Compared to domains, facets provided more information concerning the relationships between demographics and personality.” But the R2 estimates were almost the same with facets and domains, so it is unclear how the facets provided more information, unless you mean that the domain-level associations could be specified with facet-level measures?

We were making a distinction between (a) the greater specificity of the personality-criterion relationship (i.e., information) that was revealed using narrower traits and (b) the predictive accuracy of the different models. For clarity, we have changed the language of the paper to refer to a as “specificity”: “The study’s second aim was to determine whether greater specificity (i.e., a more specific relationship that could not be attributable to a broader trait) could be gleaned from geographical psychology findings with the use of progressively narrower personality traits (domains, facets, and nuances, respectively)” (p.
5. I wonder how the results from the nuance-level analysis should be interpreted. On page 24 you note that the nuances correlated with population density came from different facets, but they were all related to aversion to rule-following. Does this support a nuance-level analysis, or does it suggest that population density is related to a higher-level personality trait of rule following? That is, you would move back from the nuance-level to a trait-level after you have identified the most relevant items? Or if you find that the correlations with openness were largely due to political orientation, would you conclude that the association is not about personality differences per se but instead about political attitudes?

Whether to stick with nuances or summarize up to the facet level is a challenging question, and currently there aren’t any hard-and-fast rules. We took a face-valid approach of examining the item content to find obvious patterns. The case of rule-following was perhaps the most interesting finding, in that it appeared to us as that it was a facet-sized trait that spanned across five different facets. We have added a few words to clarify this point: “Thus, a more precise description of the relationship between ZIP Code population density and personality would ignore the broader facets of Dutifulness, Morality, and Orderliness and instead focus on nuances related to aversion to rules, which themselves may indicate a novel facet-sized trait” (p. 23). Concerning Liberalism, we believe this issue is outside the scope of our current paper (please see our response to Dr. Donnellan’s point #8.)

6. Also, it might be worth emphasizing that the nuance-level or even facet-level analysis may not be feasible in many research contexts in which it is not possible to administer a sample of hundreds of personality items to the research participants.

We have updated our manuscript as follows: “Future studies in all branches of personality psychology could benefit from examining the extent to which facet- and nuance-level analyses provide greater specificity to domain-level findings, as long as researchers have the resources to collect a large sample of participants and randomly sample from an item pool of hundreds of items” (p. 26).

RESPONSES TO THE SECOND REVIEWER’S COMMENTS

1. I found the result that some facet-correlations were in part driven by the same outcome-relevant item content being spread across them very informative.

Thank you.

2. There is a highly relevant pre-print (https://psyarxiv.com/dpqrx/).

Thank you for mentioning this interesting article. It is indeed highly relevant to the current study. We have cited it: “Lastly, one recent study found that a participant’s country of origin was better predicted with personality items than facets or domains (Achaa-Amankwaa, Olaru, & Schroeders, 2020)” (p. 5).

3. The degrees to which traits varied across ZIP codes seems comparable to the typical degrees to which traits vary across countries. Perhaps this can be an instructive parallel.
We are hesitant to make that comparison without first confirming it in the same data set that our U.S. analyses came from; we are not confident in assuming that the effect size that other studies have founds would generalize to our data. As we have already wandered off on a few tangents with this paper, we feel a comparison of ZIP and state ICC1s to that of countries would be outside the scope of the current project. However, we believe that this paper is a first step in considering whether there is a geographic unit that maximally aggregates personality. We hypothesize that the smaller the unit the better (down to the neighborhood), and that countries do not actually account for much variance in comparison to neighborhoods. We would be surprised if countries had an effect of similar magnitude to that of neighborhoods, but we definitely think this question should be explored in future research.

4. It was not entirely clear from the methods that the 30 IPIP-NEO facets would be used for analyses. Perhaps a couple of additional words could clarify this.

We have updated our manuscript to clarify this issue: “Big Five domains and facets were measured with the 300-item IPIP-NEO inventory from the International Personality Item Pool (IPIP; http://ipip.ori.org/), an online repository for public domain items and scales (Goldberg, 1999; Goldberg et al., 2006). Each of the IPIP-NEO Big Five domains consisted of six facets and 60 items; each of the 30 IPIP-NEO facets consisted of ten items” (p. 9).

5. It was not quite clear what substantive purpose the ICC2s served. Perhaps this could be either elaborated – or dropped.

As our manuscript states, “ICC2 measures group mean reliability and estimates the extent to which one would expect aggregated differences (i.e., aggregated ZIP Code personality differences) to replicate in a new sample of a similar size” (pp. 11–12). We believe this is an important estimate to report for full disclosure of our results, especially because the reliabilities do not look good. Additionally, geographical psychology papers have typically reported ICC2 reliability values (while not reporting ICC1 values, which we disagree with), so it is also a useful metric for comparison across papers. However, there’s not much more of interest to say about them once we have reported them, so we do not believe that additional text would be useful.

6. I am not sure why facet-level analyses were not carried out for facets whose domains were not correlated with outcomes (the same for items and facets). It could be that some domains did not correlate with some outcomes because only some of their facets did (perhaps even in opposite directions; the same applies for items within facets). This may have lead to underestimating the incremental information in facets (and items).

As we mentioned in response to Dr. Donnellan’s point #4, the facet-level analysis you suggested was carried out and described in a footnote, but was perhaps easily missed because it was in a footnote. We updated the language of the footnote to make that reason for our analyses clearer and moved it to the body of the paper (see our response to Dr. Donnellan’s point #4). We are not certain we understand your point about items, but we think you have the same concern about items. We included all items, not just items within domains that were associated with a demographic. We have added the word “all” to make our description clearer: “In order to gain even greater specificity in the relationship between the four ZIP Code demographics and aggregated personality, we correlated each demographic with all 696 personality items.” (p. 16).
7. As authors correctly note, the predictive value of nuances is most likely underestimated for only a small selection of thematically overlapping items, selected on the basis of what were undeniably harsh criteria, was used in polyitem scores. Is there a way of relaxing the criteria a little and seeing what happens (that is, having more content in polyitem scores)?

We chose such harsh criteria because we were concerned about capitalizing on chance, and we wanted to only publish results that had the best shot of being replicated in future studies. We believe that our conservative approach is warranted considering the hundreds of comparisons in our exploratory analyses and the fact that our effective N was 2,074 ZIP Codes.

8. I wonder if authors could comment the implications of their findings for (what is often thought to be notoriously low) reliability of single items?

As we mentioned in response to Dr. Donnellan’s point #3, we have updated our literature review to make a clearer case concerning item reliability/error: “Concerns about the low reliability of items have been answered with large-sample studies indicating that nuances are reliable and valid constructs; the unique variance of individual items have cross-rater agreement, are rank-order stable, and are heritable (Möttus, McCrae, Allik, & Realo, 2014; Möttus, Kandler, Bleidorn, Riemann, & McCrae, 2017; Möttus, Sinick, et al., 2018)” (p. 4).

9. It would be interesting to see how the polyitem scores can predict across outcomes (scores created for one outcome correlating with another). This can be a metric for the specificity vs generality in how the outcomes track with personality traits (one may even call it discriminant validity). Perhaps this (discriminant validity) could even be somehow compared across items, facets and the Big Five.

We agree that this is an interesting idea. The proposed comparison would need a large number of criteria and personality items in order to determine the extent to which this discriminant validity would vary, and whether there would be any general rules that would emerge (e.g., item-based models have less discriminant validity than facet-based models, which have more discriminant validity than domain-based models). While we find this question interesting, we believe it is outside the scope of the current paper.

RESPONSES TO THE THIRD REVIEWER’S COMMENTS

Some minor places in possible need of improvement:

1. Explain confounds better—on p. 19 (and Footnote 9) where confounds are briefly mentioned, I thought a more detailed and clearer justification would be useful. I’m thinking of the Rohrer, 2019 paper (Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data), which provides a clear way to decide when a variable can be considered a confound and needs to be controlled for (the footnote sort of does that, but too briefly to be clear enough, and should probably be moved to the text, because I missed it on the first read).

As we mentioned in response to Dr. Donnellan’s point #9, we have removed the confounds from the regression models for the main paper, and moved the regression models with the
potential confounds to the supplemental materials.

2. Justify missing data treatment further—on p. 8, the authors briefly mention that the missing data mechanism was MMCAR due to planned missingness and then say missing scores were not imputed. I was confused by this, because I thought, according to Little (2013) that the whole point of planned missing data was that modern procedures were really good at recovering the missing data and producing unbiased estimates. I'm not saying the authors should use multiple imputation, as I understand that the amount of missing data might make that unfeasible, I just wanted to see a more detailed explanation for this decision.

A recent paper indicates that imputing 90% MMCAR data, specifically SAPA data, resulted in a loss of predictive accuracy that was greater than simply using the 10% of data that was available. We now reference that paper: “Due to high rates of missingness, missing scores were not imputed; aggregated personality scores for ZIP Codes and all analyses were based only on non-missing item scores. Although imputation can allow for unbiased estimates of missing data (Little, Jorgensen, Lang, & Moore, 2013), a recent study of SAPA data with 90% data missingness indicated that models trained on imputed data were less predictive than models trained only on the 10% actual data (Elleman et al., in press)” (p. 8).

This concludes the revision notes for our paper. The coauthors and I are in agreement with the current draft’s content and the order of author names. I will continue to serve as the corresponding author for this manuscript. Thank you for your time and consideration.

Sincerely,

Lorien G. Elleman