The ways of the world? Cross-sample replicability of personality trait-life outcome associations

Journal of Research in Personality 112 (2024) 104515

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Full Length Article

The ways of the world? Cross-sample replicability of personality trait-life outcome associations

Ross David Stewart [a,g,*], Alice Diaz [a], Xiangling Hou [b], Xingyu (Shirley) Liu [c], Uku Vainik [d,e,f], Wendy Johnson [a], René Mõttus [a,d]

[a] Department of Psychology, University of Edinburgh, Edinburgh, Scotland
[b] The Key Research Institute of Chongqing for Curriculum & Instruction, School of Education, Chongqing Normal University, Chongqing, China
[c] Department of Psychology, The University of California, San Diego, United States
[d] Institute of Psychology, University of Tartu, Tartu, Estonia
[e] Institute of Genomics, University of Tartu, Tartu, Estonia
[f] Montreal Neurological Institute, Mcgill University, Montreal, Canada
[g] Psychology Department, Wrexham University, Wrexham, Wales

A B S T R A C T

Research in (mostly) Western samples has indicated that personality domains' associations with life outcomes are replicable but often driven by their facets or nuances. Using three diverse samples (English-speaking, N=1,232; Russian-Speaking, N=1,604; Mandarin-speaking, N=1,216), we compared personality trait-outcome associations at domain, facet, and nuance levels, both within and among samples. Trait-outcome associations were at least moderately consistent among samples for all trait-hierarchy levels (average intraclass correlations = 0.64 to 0.74). Nuances provided the strongest predictive accuracy, both within and among samples. Trait-outcome associations were higher among English-speakers than Mandarin and Russian-speakers. Our observations suggested moderate generalizability among diverse samples, with nuances providing unique and replicable information. This offers potential to improve understanding of trait-outcome patterns.

1. The ways of the world? Cross-sample replicability of personality trait-life outcome associations

Personality traits have been linked to many life outcomes, including academic (Mammadov, 2022; Trapmann et al., 2007) and socio-economic (Jonassaint et al., 2011) achievement, relationship quality (O'Meara & South, 2019) and treatment success (Bucher et al., 2019). This highlights the traits' population-level implications that can inform policy development and implementation (Bleidorn et al., 2019). For example, identifying risk factors for negative outcomes such as substance use (Lackner et al., 2013) or disregard for the environment (Soutter & Mõttus, 2021) may facilitate designing interventions that address common psychological barriers to behavior change.

However, the theoretical and practical relevance of personality trait-outcome associations depends on how well they generalize among people and circumstances. Soto (2019) observed that 87 % of trait-outcome associations from mostly Western samples could be replicated in a large sample of US adults, although the associations' strengths were often weaker in the replication study than in the original research. Less clear, yet important, is whether trait-outcome associations would also replicate in samples with different cultural backgrounds (Klimstra & McLean, 2024). Only with evidence of the associations' replicability in relevant cultural circumstances should we think that Western-based interventions, policy suggestions or other research applications may succeed in other populations. For instance, based on prior Western evidence on personality trait-life satisfaction associations, Olaru et al. (2023) designed a personality intervention in a Western sample with the expectation that trait change would include increases in satisfaction. Before such designs are tested, let alone rolled out, in culturally more diverse settings, the associations' replicability should be tested first. Likewise, replicability in culturally diverse samples would indirectly support claims that personality traits have similar roles in many cultures (Allik et al., 2013).

Here, we investigated personality trait-life outcome associations in three different samples, representing diverse cultural backgrounds and speaking distinct languages: English-speakers (mostly UK residents), Russian-speakers from Russia or countries with substantial Russian-speaking minorities (e.g., Ukraine), and Mandarin-speakers (mostly Chinese residents).

* Corresponding author.
E-mail address: rstewa16@ed.ac.uk (R. David Stewart).

https://doi.org/10.1016/j.jrp.2024.104515
Received 6 July 2023; Received in revised form 10 July 2024; Accepted 12 July 2024
Available online 17 July 2024

1.1. Strengths of personality trait-life outcome associations

Besides replicability, the theoretical relevance of using personality traits to predict and intervene in life outcomes depends upon how strongly the traits track the outcomes. The Big Five domains – one of the broadest levels of the most commonly used personality trait hierarchy – provide broad summaries of individuals' personality traits and may each predict many life outcomes (Ozer & Benet-Martínez, 2006; Roberts et al., 2007). This offers a parsimonious descriptive approach, but it may not be optimal for every purpose, including prediction (Mõttus et al., 2020; Revelle, 2024).

Domains are presumed to consist of narrower traits such as facets that often track outcomes to greater and differing degrees within their presumed domains, so that individuals scoring similarly on a domain may vary in which facets associate with outcomes. For example, in one study, Conscientiousness facets explained 24 % more variance in job performance than their domain (Dudley et al., 2006), and a combination of multiple individual facets explained over 400 % more variance in body mass index (BMI) than a combination of their domains in another study (Vainik et al., 2019). In the latter study, only two domains, Conscientiousness and Neuroticism, showed (small) associations with BMI, but several comparatively stronger associations were found at the facet level, with some (but not all) facets from each of the five domains tracking BMI.

Moreover, outcomes might be even more strongly correlated with personality nuances. These are traits narrower still than facets, usually represented by single items that capture partly independent, yet often valid information about specific aspects of individuals' differences (Condon et al., 2020; McCrae, 2015). For instance, many items have unique variances that show cross-rater agreement (Mõttus et al., 2014), stability over time and heritability (Mõttus et al., 2017; Mõttus et al., 2019), and distinct developmental trends (Mõttus & Rozgonjuk, 2021). Importantly, nuances often provide stronger outcome predictions than domains and facets (Seeboth & Mõttus, 2018; Stewart et al., 2022). Occasionally, this may be because item content directly overlaps with the outcomes (which also inflates the domains' correlations with these outcomes!), but many nuance-level associations are meaningfully interpretable. For example, criminal behavior was most strongly ($r \approx 0.20$ to 0.30) associated with nuances from various domains/facets such as behaving irresponsibly, starting arguments, being unforgiving, cold, and uncaring, and not cleaning up after oneself, but other nuances of the same domains/facets had no or even opposite-direction links with the outcome (Stewart et al., 2022). Likewise, nuances referring to being lazy, disorganized, talkative, and full of energy out-predicted the Big Five domains for future BMI (Arumäe et al., 2023), and there are various other nuances with unique links with BMI (Arumäe et al., 2024). As another example, nuances about feeling misunderstood, unexcited, indecisive, envious, bored, used, incapable, and unrewarded were uniquely associated with low life satisfaction (Mõttus et al., 2024). Moreover, effects of personality trait change interventions are sometimes nuance-specific (Olaru et al., 2022). If trait interventions consistently change personality nuances that matter for important outcomes, it would be useful to understand more clearly how, to what degrees, and over what time spans they do so.

1.2. Examining Cross-Cultural consistency and diversity of Personality-Outcome associations

Some studies have already explored cross-cultural consistencies in trait-outcome associations in diverse samples. For example, Sutin et al. (2015) observed that, unlike in Western studies, Conscientiousness did not track with BMI in Asian samples, whereas Agreeableness and Extraversion did, positively. In contrast, Mezquita et al. (2019) reported that the Big Five domains had similar associations with multiple health outcomes in samples from Spain, Argentina, and the United States. In Mhlanga et al. (2019), four Big Five domains (Openness, Conscientiousness, Extraversion, and Agreeableness) tracked positively and Neuroticism negatively with job engagement in a South African sample. This contrasted with observations in a Polish sample, whereby only Neuroticism (negatively) and Conscientiousness (positively) were associated with job engagement (Mróz & Kaleta, 2016).

Associations can also potentially meaningfully differ cross-culturally in strength, even when consistent in direction. In Ma et al., (2021), Narcissism correlated more strongly with workplace engagement (positively) and boredom and exhaustion (negatively) in the US than in China. Conversely, Machiavellianism correlated more strongly with a range of workplace outcomes in China than in the US. Psychopathy, however, correlated similarly with outcomes in both cultures. Kim, et al. (2018) observed that extraversion was more strongly related to life satisfaction in North America (USA and Canada) than in Europe (UK, Japan, and Germany). In Mõttus et al. (2024), life satisfaction's association pattern with the Big Five domains and its overall predictability from the domains were similar among Estonian-, Russian- and English-speakers, but the strengths of some correlations varied slightly. For example, emotional stability was more strongly correlated with life satisfaction among Estonian-speakers than in the other two samples.

Fewer studies have investigated associations at multiple levels of the personality hierarchy. One study that did, Entringer et al. (2021), observed that, in a sample of over 2,200,000 participants from 96 countries facets did not predict individual religiosity to greater degrees than domains. However, in less religious cultures, facets accounted for 4.2 % of variance in individual religiosity, whereas in more religious countries, facets accounted for 19.5 % of variance. Furthermore, the specific facets most associated with religiosity differed among countries. In contrast, Mõttus et al. (2024) observed personality nuances' associations with life satisfaction generally replicated across their three samples, though with some variability in the correlations' strengths, and life satisfaction was equally predictable from personality nuances.

It is clear there is much more to learn about cross-cultural consistencies and diversities in personality-outcome associations, especially by addressing multiple levels of the trait hierarchy and diverse outcomes in the same study.[1]

1.3. Present study

We investigated the degrees to which personality trait-outcome associations were consistent in three culturally distinct samples, simultaneously considering domains, facets, and nuances. We concentrated on two aspects of these associations: a) the overall degrees to which outcomes were predictable from traits (multivariate associations) and b) associations between individual traits and outcomes. We anticipated that both might vary among samples, but we had no hypotheses about the extents of these variations due to the limited systematic research on this topic to date. Prior studies led us to expect that outcomes would be more strongly associated with nuances than domains and facets, both within samples and across samples (i.e., using models created in one sample to predict outcomes' values in another). Specifically, based on past findings from Seeboth and Mõttus, (2018) and Stewart et al. (2022), among others, we expected that, within samples, items would explain on average about 20 % more outcome variance than facets and about 40 % more than domains. However, we had no basis to hypothesize about the extents to which this advantage would be retained for cross-sample associations. We expected that individual trait-outcome correlations would replicate better for domains than facets and nuances, due to small cross-sample variations in nuance-outcome associations "washing out" with aggregation into domains.

Before testing the associations, we checked the domain and facet scales' measurement invariance (MI) in the three samples. Sufficient levels of MI support the case that domain and facet scales can be used in

[1] This study was not preregistered.

research spanning languages and cultural backgrounds. Conversely, lack of sufficient MI suggests that (a) personality scales behave differently among contexts and, (b) by implication, their nuance-specific correlations with outcomes may also vary from sample to sample, at least to some degrees. To date, no study has shown full MI across samples from different cultures (Dong & Dumas, 2020), but the extents and implications of MI violations vary with scales and samples, as do the reasons for them (Funder & Gardiner, 2024).

2. Methods

2.1. Participants

We recruited 4,127 participants, and then removed cases where individuals gave a body weight above three standard deviations over the mean or an age above 90 (leaving 4,052 participants; we suspected removed data may have been unreliable), putting participants into three samples: Russian speakers (N=1,604; 69 % female; age range = 18–86, M=41.35, SD=15.49), Mandarin speakers (N=1,216; 73 % female; age range = 18–60, M=28.41, SD=7.50), and English speakers (N=1,232; 56 % female; age range = 18–76, M=35.94, SD=11.81). Missing values for participants were replaced with the variable mean. The English-speaking (EN) participants were recruited through social media platforms and the Prolific participant sourcing platform that compensates its participants monetarily. The Mandarin-speaking Chinese (CH) sample was recruited entirely through social media and participated without compensation. Finally, the Russian-speaking (RU) sample was recruited via Google Ads targeting individuals in Russia and other countries with Russian-speaking minorities (e.g., former Soviet Union members); these participants were not monetarily compensated. All completed the same survey in the recruitment languages on the formr.org platform (Arslan et al., 2020).

2.2. Measures

2.2.1. Personality traits

We developed a 90-item personality questionnaire for this study, using an existing pool from other ongoing research projects (100NP; Henry & Mõttus, 2023 [https://osf.io/tcfgz/]). This item pool was intended to be comprehensive while prioritizing items' retest reliability, variance, and cross-rater agreement, and minimizing social desirability and redundancy (Condon et al., 2020). This pool's 198 items were mostly selected from the International Personality Item Pool (IPIP; Goldberg et al., 2006) and the Synthetic Aperture Personality Assessment (SAPA; Condon, 2018), but some new items were generated to cover all domains and facets of the Five Factor Model (FFM) and HEXACO, otherwise known as "the Big Few" (Mõttus et al., 2020), and some traits beyond them (e.g., competitiveness, envy, religiosity, sexuality, humor). For details on item selection steps and their psychometric properties, see Henry and Mõttus (2023). For this study, we focused on the FFM domains and facets as assessed by the NEO Personality Inventories (NEO-PI-R; Costa Jr & McCrae, 2008) because it is among the best-established measures of personality traits. From the 198 items, we created the shorter, 90-item questionnaire paralleling the NEO-PI-R scales, with 18 items per domain, and 3 per facet. To do so, we used data from an independent sample of mostly UK residents and English-speaking residents of other European countries who had previously completed the survey (N=1,436, 59 % female).

In this test development sample, using Jamovi (The Jamovi Project, 2022), we ran an exploratory factor analysis (EFA) on the 198 items using oblimin rotation, forcing the items into five domain-factors, and selecting 18 items for each domain-factor that would cover the domain's six facets. Two authors (RS and WJ) allocated items into the facets through a two-step process. For each domain, we ran the 18 items with the highest loadings on one of the five factors in the initial EFA through another EFA and extracted six correlated factors, generating a basis for each facet. Then we ran a maximum likelihood confirmatory factor analysis (CFA) on each identified facet, using the UK sample gathered for this cross-sample study as a replication sample independent of the initial test development sample. For the CFA, we fixed latent trait variances to 1. To evaluate model fit, we considered Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI). Given CFI's vulnerability to larger sample sizes, we prioritized RMSEA values in determining fit. We observed borderline acceptable RMSEA<.08 for all domains (RMSEA range = 0.062 – 0.079), but none of our models demonstrated close fit (<.05) (Browne & Cudeck, 1993). However, CFI (range = 0.81 – 0.85) did fall below the accepted cut-off value of 0.95 (Hu & Bentler, 1999; McDonald & Ho, 2002). Because suboptimal model fit is a rule rather than an exception for FFM domains (Hopwood et al., 2011), we decided to proceed with the partial evidence supporting our models.

RS and WJ then read each item at face value, assessing whether its content was consistent with the targeted facet's definition. If they deemed items to, by dictionary definition, fit better elsewhere, they moved them into the better-fitting facets, and replaced them with other items loading on the same factors, based on items' loadings in the original EFA in the test development sample. To check that new items 'worked' in their new facets, they ran new CFAs for each facet and domain, with the same fit criteria. They repeated this process until all items in the new 90-item measure had minimum loadings of 0.30 on both facets and domains, with acceptable RMSEAs used as the main indicator of fit (due to its priority in the steps above) and both RS and WJ were satisfied that the facets' item content was suitable. AIC, BIC and SABIC were not given by Jamovi for this stage. The scales' internal consistencies (MacDonald's omegas) are reported in Supplementary Table S17 for data quality assessment purposes. They were high for domains (usually, $\omega > 0.80$), but lower for facets (averaging 0.58 for the UK, 0.48 for RU, and 0.47 for CH) because the facets were only based on three items each, and they deliberately did not aim for highly intercorrelating items because this would reduce the scales' associations with outcomes (Revelle, 2024). Notably, facets' internal consistencies were lower in the RU and CH samples, potentially indicating higher data quality among the UK participants or somewhat poorer construct validity in the RU and CH samples.

2.2.2. Outcomes

We assessed 34 outcomes, some with pre-existing scales and others using items we wrote. The outcomes spanned a wide range, but they were all either social norms (such as holding a driver's license) or rather broadly defined to be generally applicable (e.g., crime(s) committed measured with a single item) We translated and back translated these and the existing scales for which translations were lacking, with no changes required. XL and XH, native Mandarin speakers, provided Mandarin translations, and an independent translator provided Russian translations. More information on the project and the full list of items can be found at the Open Science Network (OSF; https://osf.io/d2vpt/).

2.2.3. Pre-existing scales

Life-satisfaction. We used the Satisfaction with Life Scale (Diener et al., 1985). Each of the five items is rated using a 7-point Likert scale. XH and XL translated the scale to Mandarin; a Russian translation was available for download at https://eddiener.com/scales/.

Health. We used the Short-form General Health Survey (version 2; SF-12v2) (Ware et al., 1996) to measure health outcomes. This survey assesses eight health aspects: Physical Functioning, Role-Physical (how often daily activities were impaired by physical health limitations), Bodily Pain, General Health, Vitality (feeling energetic), Social Functioning, Role-Emotional (how often daily activities were impaired by emotional difficulties), and Mental Health with 12 items. The first four aspects are considered features of physical health, and the rest mental health. Items are scored on either 3- or 5-point Likert scales. There was some uncertainty about existing translations (e.g., Hoffmann et al.,

2005), so XH and XL translated and back translated the English version for a more directly comparable Mandarin version, and an independent translator did this for the Russian translation, with RS confirming the back translations.

Perceived Social Support. To measure social support outcomes, we used the Multidimensional Scale of Perceived Social Support (MSPSS; Zimet et al., 1988)(Zimet et al., 1988), which has three subscales: perceived support from Family, Friends, and Significant Others. All three scales have good test–retest reliability and internal consistency, as well as strong factorial validity (Cartwright et al., 2022; Wang et al., 2021). The scale has 12 items, with four items measuring each subscale. Each item is scored on a 7-point Likert scale. For the Mandarin version, Zimet et al. provided a translated copy, and we located a Russian version online (Pushkarev et al., 2020).

2.3. Other outcomes

We measured the following outcomes with single items rated on Likert scales of varying lengths; the items and scoring keys can be found in the Supplementary Material at the OSF (https://osf.io/d2vpt/):

Crime (ever convicted of a crime), frequency of donating to causes (from never donated to donate four or more times a week), driving license status (was the participant a qualified driver), previous driving fines (number of traffic citations), duration of holding a driving license, educational attainment, exercise frequency, physical fight history (ever been involved in one), number of hobbies, number of holidays per year, income, current romantic relationship status, current relationship duration (in years), career satisfaction, financial satisfaction, home satisfaction, living area satisfaction, work satisfaction, smoker status, time spent with others, volunteering history (how often participants volunteered for any activities). More information on each outcome can be found in the online technical document at the OSF (https://osf.io/d2vpt/).

2.4. Analyses

2.4.1. Measurement invariance

To compare personality trait-outcome associations robustly among samples, the assessments ideally should show MI (Meade & Lautenschlager, 2004). To test this, there is a standard procedure for multi-item scales (here, scales for domains and facets, and the outcomes assessed with multiple items). Using this procedure, we tested MI (a) in the English-speaking sample used for item development and our English-speaking sample to be compared to the samples in other languages, and then (b) in the three samples assessed in different languages. Currently, no procedure exists for testing MI for single items, so we could not test it for personality nuances and single-item outcomes. MI for culturally diverse samples does have its doubters: Funder and Gardiner (2024) have claimed that it is unattainable and perhaps even unnecessary.

We carried out the MI tests using Jeffrey's Amazing Statistics Programme's (JASP; JASP Team, 2023) SEM feature, looking at invariance at four cumulative levels: configural (consistent baseline factor structures), metric (items'/facets' factor loadings can be constrained equal without disrupting model fit), residual (items/facets' residual variances can also be constrained equal) and, further, strict (items'/facets' intercepts can be constrained equal). Historically, the most widely accepted indicators of MI are χ^2 and CFI (Putnick & Bornstein, 2016), with changes in CFI<.01 indicating acceptable fit. However, this has been challenged in recent research (e.g Chen, 2007; Meade et al., 2008; Rutkowski & Svetina, 2014). As such, given our sample sizes, we looked as well at a range of measures such as AIC and SABIC, with lower values indicating better fitting models (Sen & Bradshaw, 2017). To date, we are unaware of any cross-cultural studies that have shown evidence of strict MI for personality scales (Dong & Dumas, 2020). Our first comparison was not cross-cultural: not all participants in the inventory development sample spoke English as a first language, but they had good command of it. Thus, we expected to see little non-invariance at that stage.

For both sets of tests, we first estimated item parameters freely in each sample, then added factor loading, residual variance, and intercept constraints, one at a time. Metric invariance (equal factor loadings) indicated that constituents' relative contributions to the construct were similar, allowing comparison of latent traits' structural relations, such as their correlations among themselves and with other variables. Residual invariance constraint indicated that constituents' absolute contributions were similar, allowing for comparing structural relations at observed trait score levels (e.g., items' sum scores). Failure to attain it indicates differences in measurement reliability and/or validity. Strict invariance additionally allows comparing traits' observed mean scores among groups. This was not of interest in our study, but failure to attain it involves group differences in items beyond the latent traits.

We then also ran the same MI tests for the multi-item outcome scales (Supplementary Table S10). The test development sample had not been assessed for these outcomes.

The code for testing all four MI stages is available at the OSF (https://osf.io/d2vpt/).

2.4.2. Within-sample analyses

Our main analyses focused on the extents to which trait-outcome associations replicated across the three samples with diverse cultural backgrounds. For this, however, we first needed to examine the associations within each sample. We focused on two kinds of trait-outcome associations: first, outcomes' average overall predictability from personality traits (domains, facets, and nuances), best assessed with a prediction-oriented modeling strategy; second, individual traits' correlations with individual outcomes. Due to age and gender splits in our samples, we residualised the outcomes for both, separately in each sample.

To estimate each outcome's overall predictability within each sample, we ran a series of elastic net regressions (ENR; Zhou & Hastie, 2005), with domains, facets, and items in turn as independent variables. The ENR shrinks coefficients towards 0, mitigating the possibility of inflated associations due to over-fitting (Yarkoni & Westfall, 2017). The ENR includes or excludes highly correlated variables from the models by either co-shrinking their coefficients to 0 or keeping them non-zero, and estimates which coefficient combination provides the greatest predictive accuracy across different data 'folds' (Waldmann et al., 2013). Before running each ENR, we split the sample randomly into two subsamples, for model training (70 %) and validation (30 %). Within the training sample, we ran an ENR with a 10-fold cross-validation and chose a shrinkage parameter from among many random ones that minimized prediction error across the folds. We then transferred the model to the independent validation sample to predict the outcome from personality traits and Pearson-correlate its predicted values with its observed values to measure predictive accuracy. Such complete separation of model training and validation precluded over-fitting because the sample idiosyncrasies the model could capitalize on in the training sample would not be present in the validation sample. We repeated this training-cross-validation procedure 100 times for every personality trait-outcome combination with different random sample splits, averaging the predictive accuracies over the repeats but also reporting their standard deviations to indicate results' robustness.

Next, to test individual associations, we Pearson-correlated each outcome with personality domains, facets, and items within each sample.

2.4.3. Cross-sample analyses

Our next step was to test the trait-outcome associations' cross-sample replicabilities. To do so, we first compared the extents to which personality traits predicted outcomes in each sample. For example, did personality traits predict outcomes better among Mandarin-speakers than among English-speakers? Next, separately for domains, facets, and nuances and each outcome, we estimated outcomes in one ("target")

sample from models that were trained in the combined data of the other two samples, going through all three combinations. For example, we used a stratified combination of the UK and RU data to train models for predicting outcomes in the CH data, using an equal number of participants from each of the UK and RU samples to create one combined sample equal to the size of the CH test sample. That is, one stratified sample consisted of 608 participants from the UK sample, and 608 participants from the RU sample, hence equaling the CH sample (N=1,216). We took even samples from the training samples to minimize the possibility of sample similarity influencing results (e.g., if the less "WEIRD" cultures were more similar, then this could influence cross-sample predictions). We constrained the combined training sample size to avoid comparisons being confounded by sample size (larger samples may allow training better models, thus giving the cross-sample prediction an advantage over within-sample predictions). We used the same procedures for the rest of this step as outlined in the within-sample analysis, including averaging the predictive accuracies across 100 random training samples and using 70 % of the training sample.

Comparing the extents to which models trained in combined samples predicted outcomes in the target samples to the within-sample prediction accuracies indicated the cross-sample generalizability of the models' parameters (i.e., the trait-outcome associations). Cross-sample and within-sample predictive accuracies being equal would indicate good replicability; the former being much lower would indicate poor replicability. Often, it may be the cross-sample predictive accuracy that concerns researchers most. For example, this is a standard approach in genome-wide association studies (GWAS) where *meta*-analytic allele-phenotype associations trained across many samples (prediction models) are used to create polygenic scores (usually treated as predictions) in an independent sample and the predictive accuracy of these scores is tested against the phenotype's observed values in this sample. This could also become a standard approach for personality traits-based predictions.

To compare the degrees to which trait-outcome associations replicated further, we correlated (among samples) the correlation profiles of respective outcomes with items, facets, and domains in turns. Next, we calculated single-profile absolute intra-class correlations (ICCs) among these correlation profiles for each outcome, at each trait level among the three samples to quantify their consistencies. High ICCs would indicate that traits similarly correlated with the outcomes among the samples.

We controlled age and gender in all analyses by residualizing the outcomes for these first, separately each sample. The code can be found at https://osf.io/d2vpt/.

3. Results

3.1. Measurement invariance

Within the two English-speaking samples (test development sample and the English-speaking sample used for cross-language sample comparisons) at the domain level, we observed strict invariance (ΔCFI between more and less constrained models < 0.01) for Openness, Neuroticism, and Extraversion, as well as residual invariance for Agreeableness and Conscientiousness (see Supplementary Table S1). We observed up to residual invariance for Competence, Order, Achievement Striving, Deliberate, Values, Modesty, Trust, Altruism, Tender-Mindedness, Depression, Anxiety, Self-Consciousness, Impulsivity, Positive Emotion, and Warmth facets, as well as strict invariance for Assertiveness, Activity, Angry Hostility, Vulnerability, Aesthetics, and Self–discipline facets. We also observed metric invariance for Duty, Fantasy, Feelings, Straightforwardness, Compliance, and Gregariousness, but only configural invariance for Seek Excitement, Actions, and Ideas facets (see Supplementary Table S2). Although far from perfect MI, we considered this evidence sufficient to justify using the same items we had selected to reflect the FFM domains and facets in the development sample to reflect them in the cross-sample comparisons, especially because when we could not proceed beyond metric invariance (equal residual variance was not met), the largest ΔCFI value was 0.06, but usually, they were notably smaller.

However, when we investigated MI among our three different-language samples, all domains but Neuroticism and most facets met only configural MI criteria, with ΔCFI>.01 after imposing cross-sample equality constraints on factor loadings (Supplementary Tables S3 to S4). Only the Neuroticism domain and Deliberation, Ideas, Modesty, Compliance, Depression, Self-Consciousness, Positive Emotion, and Activity facets met metric MI criteria, and no scale met the criteria for more stringent MI levels. This was slightly different when looking at AIC (no domains, and only Modesty, Depression, Positive Emotion and Activity reaching metric invariance) and SABIC (Agreeableness, Ideas, Actions, Modesty, Straightforwardness, Compliance, Depression, Self-Consciousness, Impulsive, Positive Emotion and Activity reaching metric invariance, and Competence meeting residual invariance), but the overall picture was the same: MI was generally poor. Items' loadings on their domains and facets in all three data sets are shown in Supplementary Table S5. Likewise, all outcome scales failed to meet anything but configural MI (aside from Perceived Friend Support, which met metric invariance using SABIC), suggesting that, as for the personality traits in general, the items measuring the outcomes were the same among the samples, but they did so to different degrees (Supplementary Table S6).

Though we acknowledge the possible interpretative limitations that such a pervasive lack of MI placed on the merits of our cross-sample comparisons, we proceeded with our analyses.

3.2. Associations within samples

Table 1 shows the accuracy of domain, facet, and item-based prediction models for each outcome and each sample, as well as the mean and median accuracy among the outcomes. In all samples, there was a wide range of predictive accuracy across the outcomes, but items consistently displayed greater accuracy.

Within the UK sample, domains were least predictive ($r = 0.25$, SD=.04), followed by facets ($r = 0.30$, SD=.04), then by items ($r = 0.35$, SD=.04). A similar pattern was observed in the RU sample, with domain-based models being less accurate ($r = 0.20$, SD=.03) than both facets ($r = 0.21$, SD=.03) and items ($r = 0.24$, SD=.03). In the CH sample, similar observations were made. Items ($r = 0.26$, SD=.04) out predicted both facets ($r = 0.23$, SD=.04) and domains ($r = 0.22$, SD=.04). Expressed in relative terms, in the UK sample, items were more accurate than domains by 40.0 %, and facets by 16.7 %, while facets exceeded domains by 20.0 %. In the RU sample, items were more accurate than domains and facets by 20 % and 14.3 %, respectively, while facets were more accurate than domains by 5.0 %. With the CH sample, the respective percentages were 18.2 %, 13.0 %, and 4.5 %.

In no sample did the average standard deviation of trait-outcome correlations exceed 0.04, demonstrating consistency and robustness of results. In the CH sample, for the crime outcome, the standard deviation was 0.07 for domains and items. No other outcome exceeded a standard deviation of 0.05.

Domains', facets', and items' correlations with each outcome in each sample are shown in Supplementary Tables S7 to S15.

3.3. Cross-sample outcome analysis

Table 2 shows the average accuracy of domain, facet and item-based models for each outcome when predicting them from models trained in samples speaking different languages.

When predicting the UK participants' outcomes from models trained in the combined RU and CH (RUCH) samples, the models were less accurate than corresponding predictions trained on the UK data, despite training samples being equal in size and training and validation samples never overlapping. Yet, the general pattern of items ($r = 0.28$, SD=.03)

Table 1
Mean accuracy of prediction models within-samples (r).

	UK						RU						CH					
Outcome	Domains	SD	Facets	SD	Items	SD	Domains	SD	Facet	SD	Items	SD	Domains	SD	Facets	SD	Items	SD
Crime	0.09	0.05	0.17	0.04	0.19	0.05	−0.04	0.02	0.02	0.04	0.01	0.05	0.05	0.07	0.01	0.05	0.05	0.07
Donating	0.21	0.04	0.22	0.04	0.23	0.04	0.20	0.03	0.20	0.04	0.27	0.04	0.11	0.04	0.13	0.04	0.13	0.04
Driving Status	0.19	0.04	0.24	0.04	0.23	0.04	0.13	0.03	0.12	0.03	0.17	0.03	0.06	0.04	0.03	0.03	0.05	0.04
Driving Fines	0.11	0.03	0.09	0.04	0.11	0.03	0.05	0.03	0.02	0.03	0.06	0.03	0.00	0.04	0.01	0.04	−0.03	0.04
Driving Time	−0.02	0.04	0.02	0.04	0.07	0.04	0.01	0.02	0.06	0.04	0.07	0.03	−0.09	0.02	−0.06	0.02	−0.07	0.04
Education	0.20	0.04	0.26	0.04	0.33	0.04	0.13	0.03	0.14	0.04	0.18	0.03	0.12	0.04	0.19	0.05	0.27	0.04
Exercise	0.19	0.05	0.23	0.05	0.30	0.04	0.21	0.04	0.21	0.04	0.23	0.04	0.22	0.05	0.25	0.04	0.25	0.04
Fight History	0.14	0.05	0.26	0.04	0.28	0.04	0.11	0.03	0.12	0.04	0.12	0.03	0.16	0.03	0.16	0.04	0.19	0.04
Number of Hobbies	0.24	0.04	0.30	0.04	0.34	0.04	0.21	0.04	0.30	0.04	0.33	0.03	0.31	0.04	0.37	0.04	0.39	0.04
Holidays	0.25	0.04	0.31	0.04	0.36	0.03	0.19	0.04	0.16	0.04	0.20	0.04	0.22	0.05	0.20	0.05	0.17	0.04
Income	0.25	0.04	0.31	0.03	0.30	0.03	0.18	0.03	0.17	0.03	0.26	0.04	0.13	0.04	0.14	0.04	0.12	0.05
Relationship History	0.11	0.04	0.15	0.04	0.23	0.04	0.04	0.04	0.05	0.03	0.03	0.03	0.09	0.04	0.16	0.04	0.20	0.04
Relationship Time	−0.01	0.04	0.08	0.04	0.08	0.05	0.03	0.04	0.05	0.04	0.04	0.03	0.05	0.04	−0.02	0.03	0.04	0.03
Career Satisfaction	0.39	0.04	0.43	0.04	0.52	0.03	0.16	0.03	0.15	0.04	0.15	0.03	0.40	0.04	0.39	0.04	0.45	0.04
Financial Satisfaction	0.32	0.04	0.47	0.03	0.56	0.03	0.11	0.04	0.10	0.04	0.15	0.04	0.25	0.04	0.27	0.04	0.35	0.04
Home Satisfaction	0.32	0.04	0.37	0.03	0.45	0.04	0.08	0.04	0.08	0.03	0.08	0.04	0.24	0.05	0.26	0.05	0.27	0.04
Life Satisfaction	0.45	0.03	0.58	0.03	0.74	0.02	0.46	0.03	0.52	0.03	0.63	0.02	0.45	0.04	0.49	0.04	0.58	0.04
Living Area Satisfaction	0.21	0.04	0.22	0.04	0.23	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.23	0.04	0.21	0.04	0.25	0.04
Work Satisfaction	0.29	0.04	0.37	0.05	0.42	0.04	0.13	0.04	0.10	0.03	0.14	0.04	0.33	0.04	0.33	0.04	0.38	0.04
Smoke	−0.03	0.04	−0.03	0.05	0.01	0.04	0.00	0.03	0.04	0.03	0.07	0.03	0.03	0.03	0.05	0.04	0.11	0.04
Time with others	0.29	0.04	0.33	0.04	0.38	0.04	0.09	0.03	0.11	0.03	0.04	0.03	0.22	0.04	0.23	0.04	0.24	0.04
Volunteering	0.16	0.05	0.19	0.03	0.22	0.04	0.22	0.04	0.25	0.03	0.26	0.04	0.11	0.04	0.08	0.03	0.11	0.04
Weight	0.12	0.04	0.27	0.05	0.33	0.04	0.04	0.04	0.07	0.03	0.10	0.04	0.13	0.04	0.09	0.05	0.10	0.04
General Health	0.40	0.04	0.47	0.03	0.53	0.03	0.40	0.03	0.40	0.03	0.45	0.03	0.31	0.04	0.32	0.04	0.34	0.04
Physical Functioning	0.15	0.05	0.19	0.05	0.33	0.04	0.21	0.04	0.21	0.04	0.25	0.04	0.17	0.04	0.19	0.05	0.23	0.05
Role Physical	0.26	0.05	0.28	0.05	0.35	0.04	0.34	0.04	0.35	0.04	0.39	0.04	0.25	0.04	0.25	0.04	0.30	0.04
Role Emotional	0.52	0.04	0.55	0.03	0.58	0.03	0.45	0.03	0.47	0.03	0.49	0.03	0.47	0.04	0.47	0.04	0.50	0.03
Bodily Pain	0.22	0.04	0.27	0.03	0.34	0.04	0.33	0.03	0.34	0.04	0.36	0.04	0.22	0.04	0.20	0.04	0.22	0.04
Mental Health	0.15	0.04	0.20	0.04	0.20	0.04	0.08	0.04	0.09	0.03	0.07	0.04	0.07	0.05	0.03	0.04	−0.01	0.04
Vitality	0.52	0.03	0.61	0.02	0.68	0.02	0.48	0.03	0.52	0.03	0.56	0.03	0.46	0.03	0.53	0.03	0.54	0.03
Social Functioning	0.48	0.04	0.55	0.03	0.56	0.03	0.45	0.04	0.49	0.03	0.50	0.03	0.39	0.04	0.40	0.04	0.44	0.04
Significant Other Social Support	0.30	0.05	0.35	0.05	0.47	0.04	0.36	0.04	0.38	0.04	0.41	0.04	0.48	0.03	0.49	0.03	0.54	0.03
Family Social Support	0.36	0.04	0.39	0.04	0.43	0.04	0.36	0.04	0.40	0.03	0.44	0.03	0.40	0.04	0.44	0.04	0.47	0.04
Friends Social Support	0.52	0.04	0.56	0.04	0.62	0.03	0.43	0.04	0.46	0.03	0.49	0.03	0.55	0.04	0.56	0.03	0.60	0.03
Median	0.23	0.04	0.27	0.04	0.33	0.04	0.17	0.04	0.15	0.03	0.19	0.03	0.22	0.04	0.20	0.04	0.24	0.04
Mean	0.25	0.04	0.30	0.04	0.35	0.04	0.20	0.03	0.21	0.03	0.24	0.03	0.22	0.04	0.23	0.04	0.26	0.04

NOTE: UK=English Speaking Sample, RU=Russian Speaking Sample, CH=Chinese Speaking Sample.

Table 2
Mean accuracy of prediction models among-samples (r).

	RUCH-UK						UKCH-RU						RUUK-CH					
Outcome	Domains	SD	Facets	SD	Items	SD	Domains	SD	Facets	SD	Items	SD	Domains	SD	Facets	SD	Items	SD
Crime	0.03	0.04	0.05	0.06	0.02	0.04	0.00	0.01	0.02	0.02	0.02	0.01	0.04	0.03	0.02	0.03	0.00	0.02
Donating	0.19	0.01	0.19	0.03	0.21	0.02	0.18	0.02	0.19	0.01	0.22	0.02	0.09	0.01	0.10	0.01	0.12	0.01
Driving Status	0.13	0.03	0.13	0.04	0.14	0.03	0.12	0.02	0.11	0.02	0.13	0.03	0.06	0.02	0.08	0.02	0.09	0.02
Driving Fines	0.08	0.02	0.08	0.03	0.08	0.04	0.06	0.02	0.07	0.02	0.06	0.02	0.05	0.02	0.05	0.02	0.07	0.02
Driving Time	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.03	0.02	0.01	0.02	0.02	0.01	0.01	0.00	0.02	0.02	0.02
Education	0.12	0.04	0.15	0.04	0.18	0.04	0.11	0.01	0.13	0.01	0.15	0.02	0.04	0.03	0.08	0.03	0.13	0.03
Exercise	0.19	0.01	0.21	0.02	0.24	0.03	0.21	0.01	0.21	0.01	0.23	0.01	0.21	0.01	0.23	0.01	0.24	0.01
Fight History	0.12	0.02	0.18	0.04	0.17	0.03	0.09	0.02	0.11	0.01	0.13	0.01	0.10	0.03	0.12	0.03	0.13	0.03
Number of Hobbies	0.18	0.02	0.27	0.02	0.28	0.02	0.19	0.01	0.28	0.01	0.31	0.01	0.29	0.01	0.35	0.02	0.34	0.02
Holidays	0.23	0.02	0.24	0.03	0.25	0.03	0.19	0.01	0.18	0.01	0.20	0.01	0.19	0.02	0.18	0.02	0.18	0.01
Income	0.19	0.03	0.19	0.03	0.22	0.03	0.18	0.01	0.17	0.01	0.18	0.02	0.06	0.02	0.05	0.02	0.06	0.01
Relationship History	0.12	0.02	0.12	0.03	0.11	0.04	0.08	0.01	0.09	0.01	0.09	0.01	0.10	0.01	0.10	0.03	0.13	0.02
Relationship Time	0.01	0.03	0.02	0.04	0.00	0.03	−0.04	0.02	−0.01	0.02	−0.01	0.02	0.01	0.04	0.03	0.03	0.02	0.03
Career Satisfaction	0.38	0.01	0.38	0.02	0.46	0.03	0.15	0.00	0.13	0.01	0.15	0.01	0.39	0.01	0.39	0.01	0.46	0.01
Financial Satisfaction	0.30	0.02	0.37	0.03	0.47	0.03	0.11	0.01	0.13	0.01	0.15	0.01	0.25	0.01	0.25	0.01	0.32	0.01
Home Satisfaction	0.30	0.02	0.30	0.05	0.34	0.05	0.09	0.00	0.07	0.01	0.07	0.01	0.25	0.01	0.25	0.01	0.28	0.01
Life Satisfaction	0.46	0.00	0.56	0.01	0.72	0.01	0.47	0.00	0.51	0.01	0.62	0.00	0.45	0.00	0.48	0.01	0.58	0.00
Living Area Satisfaction	0.19	0.03	0.17	0.03	0.17	0.03	0.05	0.01	0.03	0.01	0.02	0.01	0.22	0.01	0.19	0.02	0.19	0.03
Work Satisfaction	0.30	0.01	0.33	0.02	0.38	0.02	0.13	0.00	0.11	0.01	0.13	0.01	0.32	0.01	0.31	0.02	0.38	0.01
Smoke	0.01	0.02	0.02	0.02	0.05	0.03	−0.01	0.02	0.02	0.02	0.06	0.03	0.00	0.03	0.01	0.03	0.05	0.03
Time with others	0.28	0.02	0.27	0.04	0.25	0.05	0.09	0.00	0.07	0.01	0.04	0.01	0.20	0.02	0.18	0.02	0.20	0.02
Volunteering	0.16	0.01	0.17	0.01	0.21	0.03	0.19	0.02	0.21	0.02	0.21	0.03	0.12	0.01	0.11	0.01	0.13	0.01
Weight	0.02	0.03	0.07	0.06	0.10	0.07	0.06	0.02	0.10	0.02	0.10	0.02	0.02	0.04	0.06	0.02	0.07	0.02
General Health	0.39	0.01	0.42	0.01	0.49	0.01	0.40	0.00	0.39	0.01	0.43	0.01	0.31	0.01	0.32	0.01	0.35	0.01
Physical Functioning	0.14	0.02	0.16	0.02	0.21	0.04	0.19	0.01	0.19	0.02	0.25	0.02	0.13	0.02	0.16	0.02	0.13	0.02
Role Physical	0.26	0.01	0.27	0.01	0.31	0.02	0.33	0.01	0.32	0.01	0.34	0.02	0.24	0.01	0.23	0.02	0.25	0.01
Role Emotional	0.52	0.00	0.54	0.01	0.55	0.01	0.44	0.00	0.46	0.01	0.45	0.01	0.46	0.01	0.47	0.01	0.48	0.01
Bodily Pain	0.23	0.00	0.24	0.01	0.27	0.02	0.31	0.01	0.31	0.02	0.31	0.02	0.19	0.01	0.19	0.01	0.21	0.02
Mental Health	0.12	0.03	0.12	0.05	0.10	0.04	0.11	0.01	0.11	0.01	0.11	0.01	0.06	0.02	0.07	0.02	0.06	0.02
Vitality	0.51	0.01	0.59	0.01	0.66	0.01	0.48	0.00	0.52	0.01	0.57	0.00	0.45	0.00	0.52	0.01	0.54	0.01
Social Functioning	0.48	0.00	0.52	0.01	0.54	0.01	0.46	0.00	0.49	0.01	0.50	0.01	0.40	0.00	0.41	0.01	0.45	0.01
Significant Other Social Support	0.30	0.00	0.33	0.01	0.40	0.02	0.35	0.01	0.38	0.01	0.40	0.01	0.48	0.01	0.48	0.02	0.51	0.01
Family Social Support	0.36	0.00	0.39	0.01	0.43	0.01	0.36	0.00	0.40	0.01	0.44	0.01	0.40	0.01	0.44	0.01	0.48	0.01
Friends Social Support	0.52	0.01	0.55	0.01	0.60	0.01	0.42	0.00	0.45	0.01	0.48	0.01	0.55	0.00	0.56	0.01	0.60	0.01
Median	0.19	0.02	0.22	0.02	0.25	0.03	0.16	0.01	0.15	0.01	0.17	0.01	0.20	0.01	0.18	0.02	0.20	0.01
Mean	0.23	0.02	0.25	0.03	0.28	0.03	0.19	0.01	0.21	0.01	0.22	0.01	0.21	0.01	0.22	0.02	0.24	0.02

NOTE: UK=English Speaking Sample, RU=Russian Speaking Sample, CH=Chinese Speaking Sample.

being more accurate than facets (r = 0.25, SD=.03) or domains (r = 0.23, SD=.02) persisted, suggesting that item's predictive advantage was not sample-specific. Put differently, cross-sample domain-based predictions were 8.0 % less accurate than within-sample predictions, whilst predictive accuracy for facets and items were 16.7 % and 20.0 % lower, respectively. For the RU and CH samples, however, the difference in the predictive accuracies of models trained in the same versus different languages were smaller. For RU domains' and items', predictive accuracy was lower by 5 % (r = 0.19, SD=.01) and 8.3 % (r = 0.22, SD=.01), respectively, with no difference for facets. Finally, for CH outcomes, predictive accuracy of domains (r = 0.21, SD=.01), facets (r = 0.22, SD=.02) and items (r = 0.24, SD=.02) was lower by 4.5 %, 4.3 % and 7.7 %, respectively.

Overall, although cross-sample predictions tended to be less accurate than those within samples, especially for the UK sample, trait-outcome models trained in the combined data of two samples did predict outcomes in independent samples tested in a different language. So, most trait-outcome associations were at least partly replicable among samples and testing languages. Despite the somewhat lower cross-sample associations, items predicted outcomes best even in cross-sample models, so items' advantages were only partly sample/culture/language-specific. Furthermore, in both within- and among- sample analyses, the UK outcomes were most accurately predicted.

3.4. Trait-outcome association agreement

The single-profile absolute ICCs (see Table 3) indicated that cross-sample similarities in trait-outcome correlations were somewhat higher for domains (mean ICC=.74) than for facets (mean ICC=.70); however, the 95 % confidence intervals of the domains' and facets' ICCs' overlapped for all outcomes. The mean ICC for items, 0.64, was lower than those of domains and facets, but again, confidence intervals overlapped for all outcomes. Moreover, ICCs for domains, facets and items were highly correlated, r > 0.92, indicating that outcomes with more similar personality correlates among samples tended to display this pattern for all assessed levels of the trait hierarchy – some outcomes were more accurately predictable from personality traits than others, regardless of whether domains, facets or items were used.

The ICCs medians were higher than means, between 0.69 and 0.83, and skews negative, between −0.99 to −1.81. This indicates poor replicability for some outcomes' correlations with personality traits but good replicability for most. Outcomes with the most different personality trait correlates among countries, across all trait hierarchy levels were, current relationship duration, smoking status, duration of holding a driving license, body weight and having ever committed a crime. For all other outcomes, ICCs were at least 0.45.

Overall, these findings are consistent with the mostly UK-driven differences between the models' accuracies in the within- and between-sample analyses, showing somewhat lower cross-sample generalizability for item-outcome correlations than for domain-outcome correlations, with facets in the middle. Yet, the associations tended to replicate at least moderately for most outcomes.

4. Discussion

We set out to investigate three main features of trait-outcome associations. First and second, we investigated the extents to which (a) outcomes were similarly predictable from personality traits among samples with diverse cultural backgrounds and (b) specific trait-outcome associations replicated among the samples. Third, we asked

Table 3
Cross-Sample Consistency in How Outcomes Correlated with Traits.

Outcome	Domain ICC	Lower Bound	Upper Bound	Facet ICC	Lower Bound	Upper Bound	Item ICC	Lower Bound	Upper Bound
Crime	0.35	−0.16	0.88	0.22	0.00	0.47	0.11	−0.01	0.25
Donate	0.83	0.46	0.98	0.81	0.69	0.90	0.74	0.66	0.81
Driving Status	0.71	0.23	0.96	0.64	0.45	0.79	0.54	0.42	0.65
Driving Fines	0.74	0.28	0.97	0.58	0.38	0.75	0.46	0.33	0.58
Driving Time	0.19	−0.26	0.82	0.14	−0.06	0.39	0.16	0.03	0.30
Education	0.57	0.04	0.94	0.61	0.41	0.77	0.49	0.37	0.61
Exercise	0.97	0.86	1.00	0.92	0.86	0.96	0.85	0.79	0.89
Fight History	0.61	0.08	0.94	0.59	0.39	0.76	0.53	0.41	0.64
Number of Hobbies	0.81	0.41	0.98	0.88	0.80	0.94	0.83	0.76	0.88
Holidays	0.86	0.53	0.98	0.81	0.69	0.90	0.72	0.63	0.80
Income	0.78	0.36	0.97	0.71	0.55	0.84	0.62	0.52	0.72
Relationship History	0.90	0.63	0.99	0.69	0.52	0.83	0.54	0.43	0.65
Relationship Time	−0.25	−0.44	0.45	−0.04	−0.21	0.20	−0.05	−0.15	0.07
Career Satisfaction	0.84	0.47	0.98	0.79	0.66	0.88	0.73	0.65	0.81
Financial Satisfaction	0.80	0.38	0.97	0.73	0.57	0.85	0.67	0.57	0.75
Home Satisfaction	0.77	0.33	0.97	0.71	0.54	0.83	0.64	0.54	0.73
Life Satisfaction	0.98	0.93	1.00	0.97	0.94	0.98	0.92	0.89	0.94
Living Area Satisfaction	0.75	0.28	0.97	0.68	0.50	0.81	0.60	0.48	0.70
Work Satisfaction	0.85	0.50	0.98	0.77	0.63	0.88	0.72	0.63	0.79
Smoke	−0.12	−0.39	0.63	0.04	−0.14	0.29	0.10	−0.02	0.24
Time with others	0.74	0.28	0.97	0.65	0.46	0.80	0.54	0.42	0.65
Volunteering	0.85	0.50	0.98	0.84	0.73	0.91	0.74	0.66	0.81
Weight	0.20	−0.26	0.83	0.25	0.03	0.49	0.28	0.15	0.42
General Health	0.97	0.86	1.00	0.93	0.88	0.96	0.89	0.85	0.92
Physical Functioning	0.83	0.46	0.98	0.80	0.67	0.89	0.66	0.56	0.75
Role Physical	0.94	0.77	0.99	0.92	0.86	0.96	0.86	0.81	0.90
Role Emotional	0.97	0.88	1.00	0.96	0.92	0.98	0.90	0.87	0.93
Bodily Pain	0.87	0.57	0.98	0.86	0.76	0.92	0.81	0.74	0.86
Mental Health	0.84	0.49	0.98	0.68	0.50	0.82	0.53	0.41	0.64
Vitality	0.99	0.94	1.00	0.97	0.94	0.98	0.93	0.90	0.95
Social Functioning	0.97	0.88	1.00	0.95	0.91	0.97	0.90	0.87	0.93
Significant Other Social Support	0.93	0.74	0.99	0.90	0.82	0.95	0.85	0.79	0.89
Family Social Support	0.98	0.93	1.00	0.95	0.92	0.98	0.90	0.87	0.93
Friends Social Support	0.96	0.85	1.00	0.94	0.89	0.97	0.90	0.86	0.93
Mean	0.74	0.42	0.94	0.70	0.57	0.81	0.64	0.55	0.72
Median	0.83	0.47	0.98	0.78	0.64	0.88	0.69	0.60	0.77

whether previously observed patterns of lower-level traits (nuances and facets) out-predicting domains would replicate not only within but also among samples.

Our results indicated that outcomes were generally more predictable from domains, facets, and nuances in the English-speaking sample of UK residents than among Russian-speakers from mostly Eastern Europe and Mandarin-speakers mostly from China at all levels of personality assessment. Next, among samples, trait-outcome associations tended to replicate at least moderately and, for most outcomes, even well, with domains' associations the most and items' associations the least replicable, by relatively small margins. That is, for most outcomes, traits associated with the outcome in one sample were likely to be similarly associated with the same outcome in the other two, slightly more so for domains than for facets and nuances. Yet, nuances predicted most outcomes more accurately than facets and domains not only within but also among samples. So, nuances' general predictive advantage over domains and facets was sufficiently strong to be present in cross-sample estimates even when individual nuance-outcome associations varied among the samples.

4.1. (Lack of) measurement invariance (MI)

As we outlined in the Results, our personality, and outcome scales rarely met even minimal MI criteria (see Supplementary Tables S3, S4, and S5), often considered necessary for cross-sample comparisons (e.g., Van de Schoot et al., 2015). It would have been desirable if both the personality and outcome scales had met at least cross-sample metric and preferably residual MI which indicates that items define the constructs with similar relative and absolute strengths; since we did not plan to compare trait and outcome levels among samples, strict invariance indicating intercept equality was less important. However, lack of MI is almost invariably observed in cross-cultural personality studies (Dong & Dumas, 2020).

But how much is this a real problem rather than a reflection of a psychological reality that is more complex than the simplistic latent trait models underlying MI tests can capture? Robitzsch and Lüdtke (2023) argued that, though MI does help to think about and model latent variables, it is not a prerequisite for group comparison. Items functioning differently in different groups does not mean that comparisons between the groups cannot be valuable. Where lack of MI occurs, and how, can indicate specific cultural differences, and recognizing those differences can be used to improve measures and understandings of the social implications of cultural features. Similarly, Funder and Gardiner (2024) argue that achieving MI is unrealistic in culturally diverse samples. They stress that, while items may be interpreted differently among groups, what matters is the ability to predict outcomes researchers or practitioners care about (Revelle, 2024). Lack of MI, however, does indicate that the predictive pathways partly differ among samples.

Moreover, we should note that the Big Five content in the 100NP, which we used as our item pool, closely correlates with the Big Five scales of multiple other questionnaires (Anni et al., 2024). Therefore, it is unlikely that our scales were unusual in their content and the findings are scale-specific.

As a result, rather than concluding that our results – and thereby the results of virtually all studies using self-report scales to make comparisons among culturally diverse samples tested in different languages – are meaningless, we attempt to interpret our observations in more nuanced ways.

4.2. Personality traits as nuanced constructs

It is plausible that translation difficulties that we were unable to address despite our best efforts and/or undetected saturation in English language/UK culture contributed to both (a) some cross-sample inconsistencies in trait-outcome association and (b) poor measurement invariance. Sampling differences may also have been involved, including possible differences in data quality. For example, the UK participants were "professional" participants used to being compensated for high-quality work, whereas the RU and CH participants were internet volunteers.

Besides translation difficulties and possible data quality and/or sampling differences, however, both poor MI and some inconsistencies in trait-outcome correlations, especially at lower levels of the trait hierarchy, can be interpreted substantively as personality being an inherently nuanced phenomenon. That is, trait constructs consist of partly, and often even largely (McCrae & Mõttus, 2019), distinct narrow nuance-traits that often relate to each other and other variables in different ways than the domains and even facets they "belong" to. This explains why personality assessment items rarely, if ever, neatly coalesce into scales despite researchers' efforts (Hopwood et al., 2011), and why items often associate more than domains with many outcomes (Mõttus et al., 2020). Likewise, nuance-traits are likely to vary slightly in meaning and association patterns with each other and other variables in diverse contexts, explaining pervasively poor MI among cultures (Dong & Dumas, 2020). In aggregated scale scores, however, nuances' uniquenesses often offset each other. This makes aggregates appear somewhat more robust – at the cost of blunting their usefulness for purposes such as outcome prediction. This, of course, is an instantiation of the widely recognized bandwidth-fidelity dilemma (Ones & Viswesvaran, 1996). Therefore, the potential sources of underlying poor MI and, relatedly, cross-sample variability in lower-level trait-outcome links may be understood as a fact of nature and an opportunity to learn more about underlying developmental pathways rather than a methodological problem. Researchers always must choose whether they want greater accuracy in their estimates, having to recognize the reality of greater complexity in return, or whether they prefer simpler and (somewhat) more replicable but less accurate findings. It assessment tradeoff also applies to aggregate life outcomes.

4.3. Weirdly higher predictability

One notable inconsistency among samples was that personality traits were generally more associated with outcomes in the WEIRDEst sample, the UK, than in the RU and CH samples. One reason for this could have been that monetarily compensated participants systematically may have paid more attention to the answers they gave or spent longer on the survey. After checking the scales' internal consistencies (measured with McDonald's omega; Table S17) within all samples at both the domain and facet levels as possible indicators of data quality, we judged that this may well have been the case. The omega values were higher in the UK sample than in the other two, suggesting greater reliability there.

This may also explain why, when predicting UK outcomes from RUCH data, the observed decrease in predictive accuracy was greater than when predicting outcomes in the CH or RU samples. Given that models trained in a combined sample of possibly 'poorer-quality' data were predicting outcomes which had previously been predicted from models trained in possibly higher quality data, a greater decrease in accuracy was expected. Furthermore, both the CH and RU outcomes were being predicted from a sample containing some higher quality data, perhaps explaining why these two samples saw a smaller decrease in outcome prediction accuracy.

Another reason could have been that outcomes were generally more dependent on people's characteristics than on outside circumstances in the UK than in the other samples. For example, Chinese culture is generally considered to be interdependent, and UK culture independent. Strong interfamily support is one prominent feature of interdependent cultures, but not particularly noted in independent ones. This may make interfamily support more variable, more dependent on individual family members' personalities. As well, many outcomes (e.g. ease of accessing a car or obtaining a driver's license, volunteering traditions, or attitudes regarding divorce) may have differed in accessibility among the samples. If this were systematic, however, we could have seen cross-sample

differences in outcome means and/or variances, but there were no such systematic patterns (outcome median standard deviation was lower in CH, but similar in the UK and RU, but the median of outcome means was lowest in RU, but similar for the UK and CH; Supplementary Table S16). Another possible reason is that English language/UK cultural saturation may have created content overlap between the traits and outcomes.

In conclusion, it remains an open question why personality traits were more associated with outcomes in comparatively more Western-like samples.

4.4. The Big picture

Our observations indicated some inconsistencies in how personality traits were linked with outcomes in culturally diverse samples, especially for outcomes such as current relationship and smoking statuses, duration of holding a driving license, body weight and having ever committed a crime. These inconsistencies may provide valuable information for better understanding the various traits (reflecting affects, behaviors, cognitions, desires, etc.) that contribute to specific outcomes in particular circumstances. For example, often being bored tracked with educational attainment similarly in all three samples. However, in the Mandarin speaking sample, getting angry easily negatively correlated with educational attainment much more strongly than in either of the other samples (see Tables S7 to S15, and Tables S18 to S20). As such, interventions that focus on anger management may work better in CH-like circumstances, whereas interventions that focus on ways to make individuals more engaged may work better in other circumstances. Furthermore, it is likely that some outcomes impact one another, such as life satisfaction furthering various health behaviors and feeling healthier making life simply feel better (e.g., Grant et al., 2009). These associations may well vary with and be underpinned by circumstances such as economic opportunities and access to health information that are associated with other outcomes such as educational attainment and socio-economic status (Klimstra & McLean, 2024).

On the flip side, however, our observations suggested substantial aggregate robustness in personality-outcome links, especially at the domain level. We had no *a priori* reason to preclude the possibility that trait-outcome associations might be very different in diverse samples, preventing using observations from one sample to make useful guesses about people in different circumstances (Klimstra & McLean, 2024). Offering support to those who argue for universality in personality and life outcome associations (e.g., Allik et al., 2013), this was not the case – at least in aggregate. To an extent, our observations suggest that the Western studies that dominate the personality traits-life outcomes literature have a substantial degree of relevance in other parts of the world. This conclusion is consistent with other studies indicating cross-cultural consistency in diverse psychological research findings (e.g., Klein et al., 2018), and with overall trends and national goals in non-WEIRD regions toward 'parity' with WEIRD regions.

At the same time, however, some of this consistency was inevitable: personality item content overlapped with outcome content, as has been the case in many other studies (e.g., Soto, 2019). For example, the item/nuance 'I am happy with my life' had some of the highest correlations, positively as would be expected, with life satisfaction. 'Am usually active and full of energy' was similarly correlated with 'Role Emotional' and 'Vitality', and 'Have a dark outlook on the future' was notably positively correlated with 'Mental Health'. These three examples held in all three samples, as would be expected from existing research in areas not usually characterizing themselves as 'personality psychology', beyond being intuitively 'obvious'. They were less strong in CH than in the other two samples. Potentially more interesting were consistencies such as that 'My feelings are easily hurt' was notably negatively correlated with the outcome 'Vitality', but more so in the CH sample. Tables S18-S20 show trait-associations for outcomes within cultures for outcomes.

4.5. Limitations and future research

Our study, however, has limitations which may be addressed in future research. Whilst we did try to control age and gender, our samples were not entirely representative of their recruitment populations. For example, we recruited our CH sample through social media – excluding participants not using social media and our UK sample from the for-pay Prolific panel that tends to attract people with time on their hands. As well, all three samples contained overly high proportions of women. Secondly, items used in the 100NP go beyond the FFM. For example, some items from HEXACO are also contained, as well as some looking at envy, which is distinct from the FFM (Dragostinov et al., 2024). However, we forced items into the FFM structure, possibly producing distorted trait-outcome associations. As mentioned above, some items, such as "being active", clearly overlap with outcome features (exercise) – highlighting the potential for content overlap within not only our study, put personality-outcome research in general. Many of our outcomes were extremely broadly defined to be applicable to most participants, so that the same responses were relevant to huge behavioral ranges, underlying motivations, circumstantial conditions, etc.

Psychological research's often poor replicability has been well documented (Open Science Collaboration, 2015; Wiggins & Christopherson, 2019). Although initial evidence (e.g., Soto, 2019, and ours) suggests that personality trait-outcome associations may be at least moderately replicable, and often even well replicable, far more research is needed. First, this research should address broader ranges of cultural backgrounds than we were able to cover. Possibly, observations may replicate worse among more diverse circumstances – but perhaps not, too. Second, future research should consider broader outcome ranges, especially more "objective" life-course variables such as formal records of grades, income, health, antisocial behavior, among others. Third, research should include assessments beyond self-reports of personality traits and characteristics devoid of social desirability (McCrae & Mõttus, 2019). Combining self-reports with informant-ratings may help to (a) overcome single-method biases (e.g., social desirability distortions) that could operate differently in different circumstances and (b) measure single-item nuances more reliably. Fourth, research could be based on alternative, established personality questionnaires, especially those covering broader ranges of traits than the NEO-PI-R facets (Mõttus et al., 2020). Moreover, such work could use assessments developed in less WEIRD cultural circumstances (e.g., Cheung, 2020; Fetvadjiev et al., 2015). Given that our measure was designed purely for this study, replications of our study should be performed using other measures, preferably those with foci on lower-order traits (e.g. "The facetMAP"; Irwing et al., 2023).

5. Conclusion

In conclusion, our results suggested that there was at least moderate cross-sample consistency in trait-outcome associations. So, it appeared that individuals with similar trait scores often tended to experience similar outcomes despite their different circumstances. We also added to growing evidence that nuances tend to be more highly associated with outcomes than facets and domains. So far, this has been observed for models created and tested in largely similar samples; we observed that this pattern remained present even when the models were created and tested in samples with diverse backgrounds, assessed in different languages.

Financial disclosure.

Estonian Research Council: PRG2190, PSG656, and PSG759.

The Key Research Base Project of Humanities and Social Sciences of Chongqing Municipal Education Commission: Grant 23SKJD046.

CRediT authorship contribution statement

Ross David Stewart: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alice Diaz:** Project administration, Investigation, Data curation. **Xiangling Hou:** Writing – original draft, Data curation. **Xingyu (Shirley) Liu:** Writing – original draft, Data curation. **Uku Vainik:** Writing – original draft, Investigation, Data curation. **Wendy Johnson:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **René Mõttus:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2024.104515.

References

Allik, J., Realo, A., & McCrae, R. R. (2013). The universality of the five-factor model of personality. In *Personality disorders and the five-factor model of personality*, (3rd ed. pp. 61-74). American Psychological Association. https://doi.org/10.1037/13939-005.

Anni, K., Vainik, U., & Mõttus, R. (2024). Personality Profiles of 263 Occupations. https://doi.org/10.31234/osf.io/ajvg2.

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods, 52*(1), 376–387. https://doi.org/10.3758/s13428-019-01236-y

Arumäe, K., Mõttus, R., & Vainik, U. (2023). Body mass predicts personality development across 18 years in middle to older adulthood. *Journal of Personality, 00*, 1–15. https://doi.org/10.1111/jopy.12816

Arumäe, K., Vainik, U., & Mõttus, R. (2024). A bottom-up approach dramatically increases the predictability of body mass from personality traits. *PLOS ONE, 19*(1), e0295326.

Bleidorn, W., Hill, P. L., Back, M. D., Denissen, J. J. A., Hennecke, M., Hopwood, C. J., Jokela, M., Kandler, C., Lucas, R. E., Luhmann, M., Orth, U., Wagner, J., Wrzus, C., Zimmermann, J., & Roberts, B. (2019). The policy relevance of personality traits. *American Psychologist, 74*(9), 1056–1067. https://doi.org/10.1037/amp0000503

Browne, M. W., & Cudeck, R. (1993). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bucher, M. A., Suzuki, T., & Samuel, D. B. (2019). A meta-analytic review of personality traits and their associations with mental health treatment outcomes. *Clinical Psychology Review, 70*, 51–63. https://doi.org/10.1016/j.cpr.2019.04.002

Cartwright, A. V., Pione, R. D., Stoner, C. R., & Spector, A. (2022). Validation of the multidimensional scale of perceived social support (MSPSS) for family caregivers of people with dementia. *Aging Mental Health, 26*(2), 286–293. https://doi.org/10.1080/13607863.2020.1857699

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, F. M. (2020). Chinese Personality Assessment Inventory. In V. Zeigler-Hill, & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences*. Cham: Springer. https://doi.org/10.1007/978-3-319-24612-3_17.

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Ziegler, M., & Zimmermann, J. (2020). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment, 36*(6), 923–934. https://doi.org/10.1027/1015-5759/a000626

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

Costa, P. T., Jr., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), The SAGE handbook of personality theory and assessment, Vol. 2. *Personality measurement and testing* (pp. 179–198). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49*(1), 71–75. https://doi.org/10.1207/s15327752jpa4901_13

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences, 160*, Article 109956. https://doi.org/10.1016/j.paid.2020.109956

Dragostinov, Y., Henry, S., Hofmann, R., Stewart, R. D., Xu, l., Zhang, Y., … Mõttus, R. (2024). Is Envy Redundant with Big Five? 'True' Correlations and Associations with Age, Sex, Education, and Income in Multi-Rater Data. https://doi.org/10.31234/osf.io/x5k9e.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57. https://doi.org/10.1037/0021-9010.91.1.40

Entringer, T. M., Gebauer, J. E., Eck, J., Bleidorn, W., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2021). Big Five facets and religiosity: Three large-scale, cross-cultural, theory-driven, and process-attentive tests. *Journal of personality and social psychology, 120*(6), 1662–1695. https://doi.org/10.1037/pspp0000364

Fetvadjiev, V. H., Meiring, D., van de Vijver, F. J. R., Nel, J. A., & Hill, C. (2015). The South African Personality Inventory (SAPI): A culture-informed instrument for the country's main ethnocultural groups. *Psychological Assessment, 27*(3), 827–837. https://doi.org/10.1037/pas0000078

Funder, D. C., & Gardiner, G. (2024). MIsgivings about measurement invariance. *European Journal of Personality, 0*(0). https://doi.org/10.1177/08902070241228338.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96. https://doi.org/10.1016/j.jrp.2005.08.007

Grant, N., Wardle, J., & Steptoe, A. (2009). The Relationship Between Life Satisfaction and Health Behavior: A Cross-cultural Analysis of Young Adults. *International Journal of Behavioral Medicine, 16*, 259–268. https://doi.org/10.1007/s12529-009-9032-x

Henry, S., & Mõttus, R. (2023, May 24). The 100 Nuances of Personality: Development of a Comprehensive, Non-Redundant Personality Item Pool. https://doi.org/10.17605/OSF.IO/TCFGZ.

Hoffmann, C., McFarland, B. H., Kinzie, J. D., Bresler, L., Rakhlin, D., Wolf, S., & Kovas, A. E. (2005). Psychometric properties of a Russian version of the SF-12 Health Survey in a refugee population. *Comprehensive Psychiatry, 46*(5), 390–397. https://doi.org/10.1016/j.comppsych.2004.12.002

Hopwood, C. J., Malone, J. C., Ansell, E. B., Sanislow, C. A., Grilo, C. M., McGlashan, T. H., Pinto, A., Markowitz, J. C., Shea, M. T., & Skodol, A. E. (2011). Personality assessment in DSM-5: Empirical support for rating severity, style, and traits. *Journal of Personality Disorders, 25*(3), 305–320.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Irwing, P., Hughes, D. J., Tokarev, A., & Booth, T. (2023). Towards a taxonomy of personality facets. *European Journal of Personality*. https://doi.org/10.1177/08902070231200919

JASP Team (2023). JASP (Version 0.17.2)[Computer software].

Jonassaint, C. R., Siegler, I. C., Barefoot, J. C., Edwards, C. L., & Williams, R. B. (2011). Low life course socioeconomic status (SES) is associated with negative NEO PI-R personality patterns. *International Journal of Behavioral Medicine, 18*(1), 13–21. https://doi.org/10.1007/s12529-009-9069-x

Kim, H., Schimmack, U., Oishi, S., & Tsutsui, Y. (2018). Extraversion and life satisfaction: A cross-cultural examination of student and nationally representative samples. *Journal of Personality, 86*(4), 604–618. https://doi.org/10.1111/jopy.12339

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., & Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Klimstra, T. A., & McLean, K. C. (2024). Reconsidering normative interpretations in personality research. *European Journal of Personality*. https://doi.org/10.1177/08902070241238788

Lackner, N., Unterrainer, H.-F., & Neubauer, A. C. (2013). Differences in big five personality traits between alcohol and polydrug abusers: implications for treatment in the therapeutic community. *International Journal of Mental Health and Addiction, 11*(6), 682–692. https://doi.org/10.1007/s11469-013-9445-2

Ma, G. X., Born, M. P., Petrou, P., & Bakker, A. B. (2021). Bright sides of dark personality? A cross-cultural study on the dark triad and work outcomes. *International Journal of Selection and Assessment, 29*(3–4), 510–518. https://doi.org/10.1111/ijsa.12342

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality, 90*, 222–255. https://doi.org/10.1111/jopy.12663

McCrae, R. R., & Mõttus, R. (2019). What Personality Scales Measure: A New Psychometrics and Its Implications for Theory and Assessment. *Current Directions in Psychological Science, 28*(4), 415–420. https://doi.org/10.1177/0963721419849559

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

McDonald, R. P., & Ho, M.-H.-R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

Meade, A. W., & Lautenschlager, G. J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods, 7*(4), 361–388. https://doi.org/10.1177/1094428104268027

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Mezquita, L., Bravo, A. J., Morizot, J., Pilatti, A., Pearson, M. R., Ibáñez, M. I., … & Cross-Cultural Addictions Study Team. (2019). Cross-cultural examination of the Big Five Personality Trait Short Questionnaire: Measurement invariance testing and associations with mental health. *PLOS ONE, 14*(12), e0226223.https://doi.org/10.1371/journal.pone.0226223.

Mhlanga, T. S., Mjoli, T. Q., & Chamisa, S. F. (2019). Personality and job engagement among municipal workers in the Eastern Cape province, South Africa. *South African Journal of Human Resource Management, 17*(1), 1–11.

Mróz, J., & Kaleta, K. (2016). Relationships between personality, emotional labor, work engagement and job satisfaction in service professions. *International Journal of Occupational Medicine and Environmental Health, 29*(5), 767–782. https://doi.org/10.13075/ijomeh.1896.00578

Mõttus, R., & Rozgonjuk, D. (2021). Development is in the details: Age differences in the Big Five domains, facets, and nuances. *Journal of Personality and Social Psychology, 120*, 1035–1048. https://doi.org/10.1037/pspp0000276

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

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 112*(3), 474–490. https://doi.org/10.1037/pspp0000100

Mõttus, R., Sinick, J., Terracciano, A., Hřebíčková, M., Kandler, C., Ando, J., Mortensen, E. L., Colodro-Conde, L., & Jang, K. L. (2019). Personality characteristics below facets: A replication and meta-analysis of cross-rater agreement, rank-order stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology, 117*(4), e35–e50. https://doi.org/10.1037/pspp0000202

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., Epskamp, S., Greiff, S., Johnson, W., Lukaszewski, A., Murray, A., Revelle, W., Wright, A. G. C., Yarkoni, T., Ziegler, M., & Zimmermann, J. (2020). Descriptive, Predictive and Explanatory Personality Research: Different Goals, Different Approaches, but a Shared Need to Move Beyond the Big Few Traits. *European Journal of Personality, 34*(6), 1175–1201. https://doi.org/10.1002/per.2311

Mõttus, R., Realo, A., Allik, J., Ausmees, L., Henry, S., McCrae, R. R., & Vainik, U. (2024). Most people's life satisfaction matches their personality traits: True correlations in multitrait, multirater, multisample data. *Journal of Personality and Social Psychology, 126*(4), 676–693. https://doi.org/10.1037/pspp0000501

O'Meara, M. S., & South, S. C. (2019). Big five personality domains and relationship satisfaction: Direct effects and correlated change over time. *Journal of Personality, 87*, 1206–1220. https://doi.org/10.1111/jopy.12468

Olaru, G., Stieger, M., Rüegger, D., Kowatsch, T., Flückiger, C., Roberts, B. W., & Allemand, M. (2022). Personality change through a digital-coaching intervention: Using measurement invariance testing to distinguish between trait domain, facet, and nuance change. *European Journal of Personality*. https://doi.org/10.1177/08902070221145088

Olaru, G., van Scheppingen, M. A., Stieger, M., Kowatsch, T., Flückiger, C., & Allemand, M. (2023). The effects of a personality intervention on satisfaction in 10 domains of life: Evidence for increases and correlated change with personality traits. *Journal of Personality and Social Psychology, 125*(4), 902–924. https://doi.org/10.1037/pspp0000474

Ones, D. S., & Viswesvaran, C. (1996). Bandwidth-fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17*, 609–626. https://doi.org/10.1002/(SICI)1099-1379(199611)17:6<609::AID-JOB1828>3.0.CO;2-K

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science,* 349(6251), aac4716. https://doi.org/doi:10.1126/science.aac4716.

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

Pushkarev, G. S., Zimet, G. D., Kuznetsov, V. A., & Yaroslavskaya, E. I. (2020). The Multidimensional Scale of Perceived Social Support (MSPSS): Reliability and Validity of Russian Version. *Clinical Gerontologist, 43*(3), 331–339. https://doi.org/10.1080/07317115.2018.1558325

Putnick, D. L., & Bornstein, M. H. (2016). Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences, 221*, Article 112552. https://doi.org/10.1016/j.paid.2024.112552

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 in Psychological Science, 2*(4), 313–345. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 1–12*. https://doi.org/10.1080/10705511.2023.2191292

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Seeboth, A., & Mõttus, R. (2018). Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions. *European Journal of Personality, 32*(3), 186–201. https://doi.org/10.1002/per.2147

Sen, S., & Bradshaw, L. (2017). Comparison of Relative Fit Indices for Diagnostic Model Selection. *Applied Psychological Measurement, 41*(6), 422–438. https://doi.org/10.1177/0146621617695521

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science, 30*(5), 711–727. https://doi.org/10.1177/0956797619831612

Soutter, A. R. B., & Mõttus, R. (2021). Big Five facets' associations with pro-environmental attitudes and behaviors. *Journal of Personality, 89*(2), 203–215. https://doi.org/10.1111/jopy.12576

Stewart, R. D., Mõttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of Personality, 90*(2), 167–182. https://doi.org/10.1111/jopy.12660

Sutin, A. R., Stephan, Y., Wang, L., Gao, S., Wang, P., & Terracciano, A. (2015). Personality traits and body mass index in Asian populations. *Journal of Research in Personality, 58*, 137–142. https://doi.org/10.1016/j.jrp.2015.07.006

The jamovi project (2022). jamovi (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

Trapmann, S., Hell, B., Hirn, J.-O., & Schuler, H. (2007). Meta-Analysis of the Relationship Between the Big Five and Academic Success at University. *Journal of Psychology, 215*, 132–151. https://doi.org/10.1027/0044-3409.215.2.132

Vainik, U., Dagher, A., Realo, A., Colodro-Conde, L., Mortensen, E. L., Jang, K., Juko, A., Kandler, C., Sørensen, T. I. A., & Mõttus, R. (2019). Personality-obesity associations are driven by narrow traits: A meta-analysis. *Obesity Review, 20*(8), 1121–1131. https://doi.org/10.1111/obr.12856

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.01064

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

Wang, D., Zhu, F., Xi, S., Niu, L., Tebes, J. K., Xiao, S., & Yu, Y. (2021). Psychometric Properties of the Multidimensional Scale of Perceived Social Support (MSPSS) Among Family Caregivers of People with Schizophrenia in China. *Psychological Research and Behavior Management, 14*, 1201–1209. https://doi.org/10.2147/prbm.S320126

Ware, J., Jr, Kosinski, M., & Keller, S. D. (1996). A 12-Item Short-Form Health Survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220–233. https://doi.org/10.1097/00005650-199603000-00003

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology, 39*, 202–217. https://doi.org/10.1037/teo0000137

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives in Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zimet, G. D., Dahlem, N. W., Zimet, S. G., & Farley, G. K. (1988). The Multidimensional Scale of Perceived Social Support. *Journal of Personality Assessment, 52*(1), 30–41. https://doi.org/10.1207/s15327752jpa5201_2

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