The Advantages of Item-level Assessment Across Research Questions in Personality Psychology

by

Kendall A. Mather

A dissertation accepted and approved in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

In Psychology

Dissertation Committee:

Dr. Sara Weston, Chair

Dr. David Condon, Core Member

Dr. Nash Unsworth, Core Member

Dr. Jake Searcy, Institutional Representative

University of Oregon

Spring 2026

© 2026 Kendall A. Mather

This work is openly licensed via CC BY-NC-ND 4.0



DISSERTATION ABSTRACT

Kendall A. Mather

Doctor of Philosophy in Psychology

Title: The Advantages of Item-level Assessment Across Research Questions in Personality Psychology

This dissertation addresses the growing need for item-level approaches to personality assessment and demonstrates a statistical technique better suited to item-level research questions—an area underrepresented among common techniques in psychology research. Chapter 1 discusses the advantages of a stronger focus on the item-level in personality research and explores why the value of item-level methods has been somewhat overlooked. Particular attention is given to the role of domain-level personality models in providing a shared framework and language for studying personality, as well as to prevailing test construction and validation practices that emphasize multi-scale measures and often treat individual items as uninformative beyond their role as indicators of broader constructs. Chapter 2 demonstrates the application of a statistical method for comparing groups at the item-level—the Dowd Test Statistic (DTS)—using data from the World Values Survey. This Empirical Cumulative Distribution Function (ECDF)-based test compares full distributions between two groups, offering improved sensitivity over *t*-tests and *F*-tests, which are limited to detecting differences in means and variances. Chapter 3 applies the DTS test to responses for 2,623 IPIP items across samples from 13 culturally and geographically diverse nations. Findings will be reported in terms of the items (i.e., IPIP items that elicit the greatest and fewest group differences across national pairs) and in terms of the nation/nation-pairs (i.e., the nations/pairs of nations that show the most and least similarity

across items). Chapter 4 summarizes key takeaways from these item-level findings and proposes directions for future research using item-focused methods.

This dissertation includes previously published coauthored material.

ACKNOWLEDGEMENTS

I am incredibly grateful to have had the opportunity to work with and learn from so many exceptional individuals during my time at University of Oregon. First, I would like to thank my advisors, Dr. David Condon and Dr. Sara Weston. The kindness, leadership, wisdom, constructive feedback, and encouragement you have offered over the years have made graduate school such an enriching experience for me.

During my first few years in the program, Dr. David Condon served as my primary mentor. His steadfast support and belief in me were invaluable to my growth both as a researcher and a person. His sense of humor, creativity, and curiosity always brightened my day, and the skills I developed during graduate school—as well as many of the projects I have completed—owe much to his mentorship and his generosity in sharing data from the SAPA project. This resource has been generative for countless others in the field, and stands as a reflection of David's intellect, vision, and passion for personality science and making high-quality data more accessible to all.

Over these final years, Dr. Sara Weston has been an unwavering source of support, kindness, and patience through the many recent bumps in the road I encountered along the way. Her organization, resourcefulness, and commitment to her students—evident in the many professional development resources she provides and the thoughtful feedback she offers—have continually amazed me. Her methodological expertise and creativity have contributed substantially to my development as a researcher, and I left every advising meeting with her feeling more confident in and more excited about my research than when I arrived. I am deeply grateful for her mentorship.

I would also like to thank my lab mates in the Personality Is Everywhere (PIE) lab from over the years—Andrew Castillo, Rachel Jacobson, Sarah McDougald, Ian Shryock, Amala Someshwar, and Dillon Welindt. I have learned so much from each of you, and I'm grateful to you all, along with David and Sara, for creating such a welcoming, supportive, and intellectually stimulating environment in the PIE

lab. In addition to my lab, I would like to thank Lori Olson for her continuous support with the more logistical side of progressing through program milestones, and whose quick responses to my emails were instrumental in keeping me on track.

I am also sincerely grateful to my dissertation committee members, Dr. Nash Unsworth and Dr. Jake Searcy, for the constructive feedback, expertise, and perspective they have brought to this project. Your willingness to invest your time and thoughtful attention into my work has meant a great deal to me. Additionally, I would like to thank Dr. Amber Gayle Thalmayer and Dr. Gerard Saucier, whom I feel incredibly fortunate to have had the chance to work with. Your research on cross-cultural personality assessment has shaped many of my own research interests, and I have greatly appreciated your thoughtful feedback on my writing and analyses over the years.

Finally, I would like to thank my friends and family for their support throughout this process. I am very grateful to Remi for our many walks around Hendricks Park and brunches at Morning Glory Café—which were always so restorative for me—and to the rest of the Smith family for the support you've offered my family and I in recent years. To my sisters, Margot and Lucy, and to the Crosby family—thank you for always cheering me on from afar and for the many laughter-filled facetime calls over the years. I am also immensely grateful to my dad, James. Although he is no longer with us, and will not see me finish graduate school, he was the first person to instill in me the belief that I could earn a PhD. I will always be thankful for his encouragement along the way—which often involved excitedly (though prematurely) referring to me as "Dr. Boo". Finally, I would like to extend my deepest gratitude to my mom, Karen. You have been the greatest and most constant source of strength, kindness, and joy in my life, and I would not be the researcher or the person I am today without your love and support.

TABLE OF CONTENTS

I. Review of item-level assessment *13*

The domain-level 15

Above the domain-level: Meta traits 19

Below the domain-level 22

Top-down and bottom-up personality frameworks 23

The Reliance on latent-variable modeling 26

Reconsidering the use of a latent variable model 28

The rise of latent variable modeling 30

How item-level assessment advances the three goals of prediction, description, and explanation 33

Item-level assessment and predicting outcomes 34

Item-level assessment and descriptive research 35

Item-level assessment and explanation-focused research 39

Conclusion 43

II. Describing and demonstrating an item-focused alternative: the DTS test *45*

Current Approaches Fall Short When Addressing Item-level Questions 45

ECDF-Based Approaches 47

Method 51

Analyses 54
Results 55
Cross-national Comparisons: Description of overall findings 55
Cross national differences detected by DTS that traditional approaches missed 67
Research Question 2: Examining changes within nations 73
Discussion 77
III. Applying DTS to IPIP items 82
Description and origins of the International Personality Item Pool 82
The Present Study 84
Method 85
Dataset and participants 85
Measures 86
Analyses 86
Results 86
Results by item 86
Results by nation pair and individual nation 91
Comparison of DTS to the T- and F-tests 97
Discussion 103
IV. Conclusion & Takeaways 112
Remaining challenges and future directions for item-focused research 113
Conclusion 115

References *117*

LIST OF FIGURES

Figure 1. 14

Figure 2. The two independently derived trait levels of the SAPA Personality framework (Condon, 2022) 25

Figure 3. A reflective measurement model including five items and five sources of item-specific (or unique) variance. Image from Markus and Borsboom (2011). 27

Figure 4. ECDF comparison between the Philippines and the United States for the item: I am quiet around strangers. 51

Figure 5. Histogram of DTS effect sizes for comparisons between nations across the 5 items ... 56

Figure 6. Self-reported Happiness: DTS effect size for each nation pair 58

Figure 7. Density distributions for three nation pairs on happiness 59

Figure 8. Self-reported Health: DTS effect size for each nation pair 60

Figure 9. Density distributions for three nation pairs on health 61

Figure 10. Self-reported Perceived Freedom and Control: DTS effect size for each nation pair. 62

Figure 11. Density distributions for three nation pairs on perceived freedom and control 63

Figure 12. Self-reported Life Satisfaction: DTS effect size for each nation pair 64

Figure 13. Density distributions for three nation pairs on overall life satisfaction..... 65

Figure 14. Self-reported Satisfaction with Financial Situation: DTS effect size for each nation pair 66

Figure 15. Density distributions for three nation pairs on satisfaction with one's financial situation..... 67

Figure 16. Density Distributions displaying the seven differences detected by the DTS test but missed by the T and F tests 72

Figure 17. DTS Effect Sizes Across Waves by Nation and Well-Being Item..... 74

Figure 18. Changes in means for each nation from wave 6 to wave 7 across items..... 75

Figure 19. Distribution of the proportions of significant DTS tests across IPIP items..... 88

Figure 20. Overall similarity in responding across nation pairs. 94

Figure 21. Distribution of DTS effect sizes by nation pair..... 96

Figure 22. Five nation comparisons for the item “I apologize a lot” that only DTS identified as statistically significant 100

Figure 23. Six nation comparisons for the item “I believe that most people would lie to get ahead” that only DTS identified as statistically significant. 101

Figure 24. Six nation comparisons for the item “I have no secrets” that only DTS identified as statistically significant 102

Figure 25. Six nation comparisons for the item “I disclose my intimate thoughts” that only DTS identified as statistically significant. 103

LIST OF TABLES

Table 1. Non-significant DTS comparisons 56

Table 2. All differences detected by DTS but not by the t-test, F-test, or both 68

Table 3. All differences detected by DTS but not by either the T-test or F-test. 76

Table 4. Items with 100% significant tests across nation pairs ($p < .05$), arranged in descending order of overall standard deviation ("Overall SD"). 88

Table 5. Items with the fewest significant tests across nation pairs, arranged in descending order of proportion of significant DTS tests. 91

Table 6. Proportion of significant tests per nation (across nation pairs and items) 92

Table 7. Items most frequently identified by DTS (but neither the T nor F test) 98

I. Review of item-level assessment

At its core, personality psychology is the scientific study of how individuals differ—an objective that is pursued in a remarkably broad range of ways. The field addresses topics such as characterizing and measuring personality differences (Wiggins & Pincus, 1992); exploring how they develop and change over time (Caspi et al., 2005); how they vary across cultures (Saucier et al., 2000); how well they predict outcomes (Roberts et al., 2007); the traces they leave in our life narratives (Dunlop et al., 2020), digital footprints (Azucar et al., 2018), and homes (Gosling et al., 2002); when they become maladaptive (Suzuki et al., 2017); and even their evolutionary origins (Penke et al., 2007). This variety of topics has made for a field that is rich with connections to other disciplines, leading some to propose that personality psychology is the "last refuge of the generalist" in psychology research (Revelle, 2008, p.2). However, a major challenge posed by the diversity of topics studied in personality psychology is the entailing diversity across studies in methodological priorities. This has made personality psychology resistant to the development of universal standards, guidelines, or best practices for assessing individual differences. One of the most widely recognized ways that studies differ in regard to their measurement needs—one that will be the focus of this review—is the appropriate level of abstraction[1] for assessing personality, which can range from very narrow to very broad (Hampson et al., 1986).

This idea that personality can be described at multiple levels of abstraction is captured by the *trait hierarchy*[2]—a conceptual tool for organizing personality traits into levels by breadth (one example is depicted in Figure 1). These levels include—from lowest to highest—

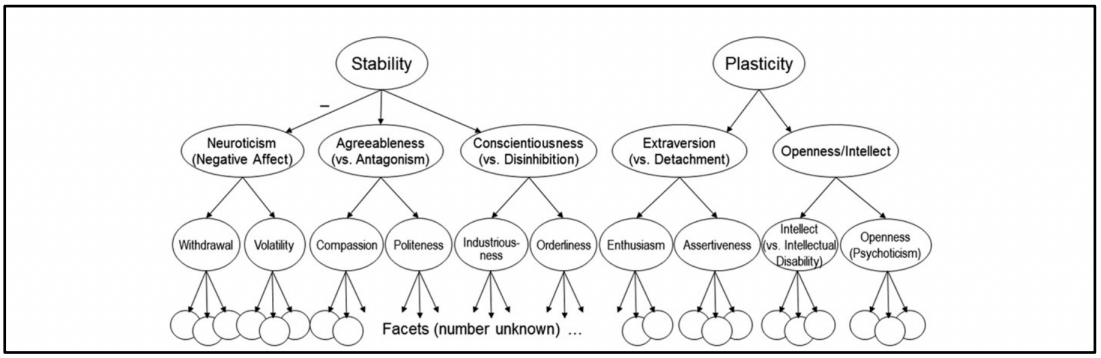
[1] Also referred to as *breadth*, *globality*, *inclusivity*, or *generality* (Hampson et al., 1986).

[2] Personality psychology has a long history of using hierarchies to describe trait space. See (Markon, 2009) for a review of the fundamentals of trait hierarchies and their use in personality psychology.

items/nuances, facets, aspects, domains, and meta traits. As an example, the item "[I] start conversations" is nested within the facet gregariousness", which is nested within the broader domain of extraversion, which is nested within the meta trait of Plasticity (DeYoung et al., 2002; also referred to as Alpha [Digman, 1997], or Dynamism [Thalmayer et al., 2024].

Figure 1.

The Trait Hierarchy (image from DeYoung & Kreuger, 2018)



Recognizing the need for more flexible personality assessment, Mõttus et al. (2017) advised against the adoption of prescribed norms and instead for tailoring approaches to personality assessment to the goals of a specific study. Mõttus et al. (2020) suggested that deciding on an appropriate personality measure for a study can be made more tractable by categorizing it under one of three broad goals of (a) description (describing personality), (b) prediction (using personality to predict outcomes), and (c) explanation (exploring the factors that shape personality). If a study is chiefly concerned with predicting outcomes, for example, the item-level appears to be optimal. This is because a single item can isolate the effect of a highly specific personality tendency (e.g., "[I] start conversations") on an outcome instead of grouping it with related attributes (e.g., expressing positive affect, high levels of activity) that together

form a broader trait (e.g., extraversion) but might relate differently to the outcome of interest. Studies that are more concerned with describing or explaining personality might use broader measures. Measures that capture personality at a broader level are ideal in many cases where a parsimonious but sufficiently comprehensive description of personality is needed (e.g., in panel surveys assessing many other variables alongside personality traits).

The utility of item-level assessment varies across studies, but the quality and psychometric properties of individual items are relevant to any study using personality measures. This importance does not appear to be reflected in the literature, as few prior studies have been done to develop individual items or focus on item-level research questions (Mõttus et al., 2017). This paper reviews several personality models and findings obtained at various levels of the trait hierarchy and discusses some potential benefits of conducting item-level analyses for a variety of research topics. In addition, we discuss reasons why the field developed a strong focus on domain-level measures, including the theories of test validity that were popular at the time of their development and practical constraints limiting the number of items researchers could realistically administer.

The domain-level

Despite its location in the center of the trait-hierarchy, the domain-level is a natural starting point for review since the majority of personality research relies on domain-level assessment materials—particularly those based on the Big Five (Goldberg, 1990) or Five Factor Model (McCrae & John, 1992), which describe personality using the five domains of Extraversion, Agreeableness, Conscientiousness, Neuroticism/Emotional Stability, and Openness to Experience. Domain-level measures have played a vital role in advancing personality science. At a time when there were seemingly "too many" traits to measure (Cattell, 1950, p.55), the

lexical approach[3] provided a relatively objective strategy for identifying a personality framework that was parsimonious but sufficiently comprehensive. The domain-level personality models rooted in the lexical approach established a common language for conceptualizing personality, which enabled coordination among researchers and replication of findings—some of the hallmarks of a cumulative science. Some even credit the Big Five/Five Factor Model with preventing personality psychology from vanishing in response to situationist critiques (e.g., Mischel, 1968) of the 1960s and 70s (Kristjánsson, 2012). Here, we briefly review several of the most popular domain-level models.

Big Five

Measures based on the Big Five model are by far the most popular for assessing personality, and the success of these measures is evident in the large body of findings they have generated. Scales such as the NEO-PI-R (Costa & McCrae, 1992) and the BFI (John, 1991) have been cited thousands to tens of thousands of times. They have been used in studies that contribute to our understanding of many topics in the realm of basic science but have also been used to obtain findings with clear practical utility. For example, the Big Five traits are fairly predictive of important life outcomes such as longevity, well-being, job success, and mental health (Lodi-Smith & Roberts, 2007; Ozer & Benet-Martínez, 2006; Roberts et al., 2007; Soto, 2019). In response to criticisms of the atheoretical roots of the Big Five model, some have also begun to develop theories that might explain why these traits emerge in lexical studies. For example, DeYoung (2015) has proposed the cybernetic big five theory, which draws on

[3] The lexical hypothesis claims that important individual differences including (but not limited to) personality and other psychological differences, will be encoded in various languages. This idea provides the foundation for lexical approaches to developing personality models, which involve obtaining a comprehensive list of psychologically-relevant adjectives and applying factor analysis techniques to identify a simplified model for describing personality using just a few dimensions. This approach is described in the foundational work of Allport (1937), Cattell (1943), Goldberg (1982), and Norman (1963). See Uher (2013) for an opposing viewpoint on lexical approaches.

neuroscientific findings and principles of control theory (Carver & Scheier, 1982) to explain how the Big Five traits arise as a result of complex networks of brain systems.

Big Five measures have been widely adopted within and outside of personality psychology, but some claim that the Big Five model is not the optimal or most universal domain-level model (Almagor et al., 1995; Ashton et al., 2004; De Raad et al., 2010; Saucier, 2009). Disagreements arise regarding the issue of how to optimally partition the variance of a dataset into a coherent set of personality factors (i.e., the factor solution that provides the best fit; Wiernik et al. [2020]) and what kinds of descriptive terms should be included in the lexical studies used to develop a model (e.g., whether to include nouns, or evaluative terms; e.g., De Raad & Barelds [2008]; Saucier [2009]). The most popular alternative framework is the Big Six or HEXACO model.

Big Six/HEXACO

Ashton and Lee (2001, 2007) suggested that a six-factor model—often referred to as the HEXACO model—is preferable to the five-factor model on both empirical and theoretical grounds. This model was also generated from the lexicon, using an approach similar to the one used to generate the Big Five model. The key difference in the process that led to the six-factor model was the inclusion of trait-descriptive adjectives that were originally deemed too evaluative by Allport and Odbert (1936) and Norman (1963) when developing the five-factor model. The six-factor model has three factors that are nearly identical to the traits given the same name in the five-factor model (Extraversion, Conscientiousness, and Openness to experience). The additional three factors—Honesty-Humility, Agreeableness, and Emotionality—do not map perfectly onto factors from the Big Five model. These three factors reflect a reconfiguration of the variance associated with Big Five Agreeableness and Neuroticism, in addition to the Honesty-Humility

variance that was not entirely captured in the Big Five model. Specifically, the Emotionality factor includes the non anger-related content from Big Five Neuroticism (e.g., sensitivity, anxiety) and the more emotional content from Big Five Agreeableness (e.g., sentimentality). The HEXACO Agreeableness factor excludes sentimentality and contains the anger-related traits associated with Big Five Neuroticism (which reflect low Agreeableness). The Honesty-Humility factor includes qualities like sincerity, fairness, modesty, and greed-avoidance—which are most closely associated with Big Five Agreeableness but are not core to any Big Five trait.

Ashton et al. (2014) argue that six factors is the maximum number that replicate across a variety of major languages. This is supported by Saucier's (2009) findings that a similar six factor model can be identified in English, Chinese, Greek, Filipino, Hebrew, Spanish, and Turkish. Dimensions from the six-factor model also appear to be more strongly aligned with concepts from evolutionary theory (e.g., reciprocal altruism and cooperation)—which helps establish a closer link between empirically derived personality models and well-supported theoretical frameworks. There is also evidence for the construct validity of the additional Honesty-Humility factor as it has demonstrated relationships with behavioral measures of dishonesty and non-cooperation, like the prisoner's dilemma task (Zettler et al., 2013) and the public goods game (Hilbig et al., 2012). The separation of Honesty-Humility and Agreeableness-related content has also been useful in the clinical domain for identifying the attributes that distinguish disorders such as psychopathy, narcissism, and borderline personality disorder (Hepp et al., 2014; Miller & Lynam, 2012).

Big seven and big eight

The Big Six/HEXACO model is generally viewed as the most viable alternative to the Big Five framework; However, Tellegen and Waller (1987) had already voiced criticisms of the

exclusion of morally-relevant or evaluative terms in previous lexical studies, which led to their proposal of a seven-factor model. Using a list of 400 descriptive adjectives, Tellegen and Waller (1987) derived seven factors. These include five factors that correspond to those from the big five model and two additional factors labeled Positive Valence and Negative Valence. The Positive Valence factor includes terms such as special, excellent, and impressive, and the Negative Valence factor includes terms such as immoral, evil, and wicked. Similar seven-factor structures have been found in Hebrew (Almagor et al., 1995), Spanish (Benet-Martínez & Waller, 1997), Filipino (Church et al., 1997), and in Greek (Saucier et al., 2005). De Raad and Barelds (2008) argued that the previous work using an expanded pool of adjectives to generate frameworks still failed to include many personality-relevant terms. When further expanding the set of terms to include nouns, verbs, and adverbs, De Raad and Barelds (2008) identified an eight-factor solution made up of the Big Five traits and three additional factors: Virtue, Competence, and Hedonism.

Above the domain-level: Meta traits

Although the reduction of trait space has previously been referred to as a "matter of unhappy necessity" (Cattell, 1945, p. 70), some have developed models of personality with even more simplicity than those at the domain-level that have made valuable contributions to personality research. Within the trait hierarchy, the domain-level traits from big few frameworks (i.e., the Big Five, HEXACO, and others described above) can also be organized into "meta" traits. These very broad traits (in addition to domain-level traits) are sometimes referred to as *source traits* or root traits, reflecting the view that these traits are the causal force behind the more specific personality tendencies captured by the domain-, aspect-, facet-, and nuance- levels of the trait hierarchy (Cattell, 1950; McCrae & Costa, 1995). Although this view of broad-level

traits as causal units has been challenged in recent work (Asendorpf, 2016), trait models at the broadest level possess several advantages including stronger cross-cultural generalizability (Saucier et al., 2014) and clearer links to various theoretical frameworks in other areas of psychology (Digman, 1997). The most popular models at the meta-trait level are the Big Two and Big Three.

Big two

There are several popular "big two" models that have a strong degree of overlap. Across models, it appears that trait content corresponding to big five Agreeableness and Extraversion are core to the two dimensions, though these "meta" traits are much broader and encompass many features of other domain-level traits. Big two models appear in the literature as early as the 1940s (H. Eysenck, 1947) but became more widely recognized with Bakan's (1966) two factor model of "Agency" (an individual's striving to master the environment and assert oneself) and "Communion" (an individual's motivation to cooperate with and relate to others).

Later, Digman (1997) identified two broader factors of virtue or "alpha" (Agreeableness, Conscientiousness, and Emotional Stability) and dynamism or "beta" (Extraversion and Openness) using second-order factor analysis, which was not commonly used at the time. This was partially due to the assumption that the big five factors were uncorrelated—providing a set of nonredundant dimensions for describing personality—which Digman (1997) recognized was a result of using varimax factor rotation (which imposes an orthogonal solution). As noted by Digman (1997), virtue and dynamism appear strongly linked with concepts from social psychology (warmth and competence; Fiske et al., 2007), clinical psychology (internalizing vs. externalizing symptoms; Achenbach, 1966), socioanalytic theory (getting along vs. getting ahead; Hogan & Roberts, 2004), and with theories of basic motivations (intimacy motivation and

power motivation; (McAdams, 1985) and motivational systems (approach and avoidance motivation; (Carver & White, 1994). DeYoung et al., 2002 propose that the terms "stability" and "plasticity" more accurately capture the essence of the big two dimensions and establish a connection for these concepts with neuroscientific phenomena (the serotonergic and dopaminergic systems).

Arguably, these big two models show the most promise due to their cross-cultural replicability, which has made them a strong focal point of research programs aimed at developing maximally universal models of personality. In the context of cross-cultural research, the big two dimensions are typically labeled "Social Self-Regulation" (propriety, socialization, community) and "Dynamism" (activity, potency, social boldness; Saucier et al., 2014). There is compelling evidence for the relative universality of the big two dimensions, which are found to emerge spontaneously in many native languages (Saucier et al., 2014; Thalmayer & Saucier, 2014). That is, they are found when using an "emic" approach, which involves identifying the optimal personality structure of a specific population using local personality concepts. This is a more rigorous alternative to the more common "etic" approach of evaluating whether a personality structure identified elsewhere—typically in Western contexts—adequately characterizes personality variation in a new context. Similar big two dimensions have emerged in "emic" psycholexical studies of many languages—Turkish, Filipino, Chinese, Korean, Hungarian, Greek, Polish, Maa, Supyire Senuofo—that represent a diverse range of language groups (Saucier et al., 2014).

Big three

Additional studies suggest that a "big three" structure comprising Extraversion/Dynamism, Agreeableness/Affiliation, and Conscientiousness/Order provides the

highest number of cross-culturally replicable dimensions. This model—sometimes referred to as the "pan-cultural three" (De Raad et al., 2010)— is less frequently discussed than the big two but is also strongly supported. The big three dimensions have been found in over a dozen psycholexical studies (De Raad et al., 1997, 2010; De Raad & Peabody, 2005; Saucier, 1997). Although Eysenck (1994) had proposed a big three model before, the particular dimensions included—psychoticism, extraversion, and neuroticism—are not as well supported by psycholexical findings and instead reflect the field's historically strong focus on emotional stability as an important individual difference (De Raad et al., 2010). Further support for a big three structure including Extraversion, Agreeableness, and Conscientiousness is found in recent studies which analyze text corpora using natural language processing (NLP) models instead of traditional personality datasets (e.g., responses to self-report questionnaires; Cutler & Condon, 2023).

Below the domain-level

Although the broadness of domain-level constructs like the big five traits has only become a more common point of criticism in recent work, the potential utility of more narrowly defined traits has long been recognized. Block (1995) discusses how an adequate understanding of personality demands that we think and measure at levels more specific than the domain-level, and describes previous efforts to develop frameworks with a larger set of more narrowly focused traits. The set of big five facets proposed by Costa and McCrae (1980)—which include six more narrowly defined traits per big five domain—remains a popular framework for examining personality at a lower level of the trait hierarchy. These facet-level scales have been useful for predicting life outcomes such as well-being (Siegler & Brummett, 2000) and job performance (Judge et al., 2013).

Later work proposed an intermediary level of narrower traits between the domain and facet levels called personality "aspects". DeYoung et al. (2007) conducted factor analysis on two sets of big five facet scales—the NEO-PI-R (Costa & McCrae, 1992) and the AB5C-IPIP (Goldberg, 1999)—and found two subordinate traits (or "aspects") for each domain-level trait. This level of the trait hierarchy also reflects an improvement beyond the domain level in the context of prediction. For instance, the "Openness" and "Intellect" aspects of the broader trait Openness to experience predict creative achievement in the arts and sciences, respectively.

The development of broad bandwidth personality frameworks—particularly the big five model—was a highly generative step for the field. However, while the question of how many distinct personality dimensions there are has been thoroughly addressed by the field, no one model is shown to be superior across contexts. Consequently, some suggest that further comparison of various low-dimensional personality structures no longer advances personality science (Wiernik et al., 2020). This has led to an increased focus on developing personality facets and items/nuances and a shift in priorities from reducing the complexity of trait space to mining it for many potentially useful trait configurations. However, exploring the value of lower-levels of the trait hierarchy will likely require a departure from the conventional "top-down" approach used to identify most personality frameworks. This "top-down" approach is described in the following section, and contrasted with the "bottom-up" approach, which foregrounds the item-level of the trait hierarchy.

Top-down and bottom-up personality frameworks

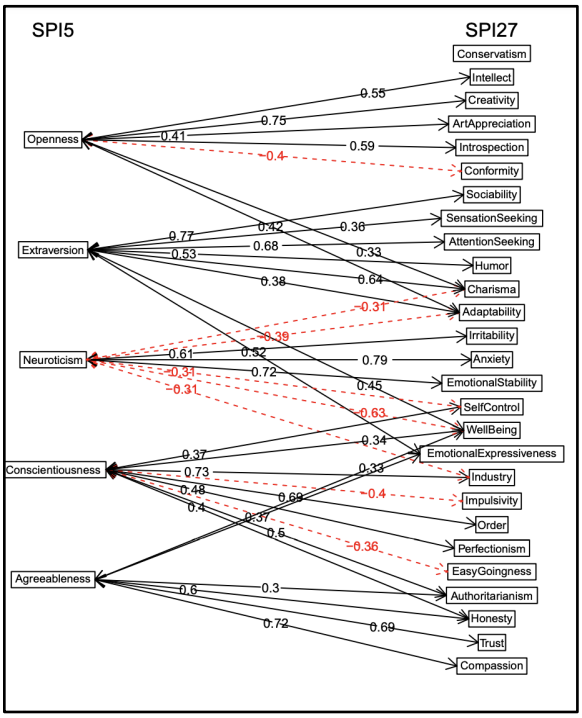
Stepping back from issues regarding the optimal level of abstraction, there is the choice of how to generate a personality framework in the first place—that is, whether to use a "top-down" or "bottom-up" approach (Condon et al., 2020). Top-down approaches begin with

identifying a set of traits at the highest level of interest (usually the domain level), most often through factor analysis. Once an interpretable and psychometrically sound factor solution is reached, lower-levels of the hierarchy are identified by decomposing the content of each broad trait into a set of narrower ones. This results in lower-order traits that are strictly nested within a single broader trait and reflect a narrower portion of its content. Factor solutions generated via a top-down approach (e.g., exploratory factor analysis) can have lower-levels that appear incomplete (i.e., missing important personality dimensions) or are otherwise misleading (Condon et al., 2020; Saucier & Iurino, 2019). Presumably, this is because top-down approaches prioritize variance decomposition at the broadest level, meaning that the content making up lower-level traits is dependent on the levels above. An alternative, "bottom-up" approach should be used to identify optimal lower-level solutions that are independent from the levels above.

Bottom-up approaches differ from top-down approaches in that lower-level solutions (e.g., the nuance or item-level) are identified first and are then used as building blocks for traits at higher levels (Condon et al., 2020). Rather than imposing a strict nesting of each narrow trait within a single broader trait, nuances can be used to form multiple traits at the level above. Bottom-up approaches are particularly useful for developing more granular personality measures. This is because they identify the unique set of most meaningful trait clusters at lower-levels of abstraction, rather than the optimal set of trait clusters based on the partitioning of trait content at the level above. The SAPA personality framework (visualized in Figure 2; Condon, 2022) was developed using a bottom-up approach and has two independently identified levels of traits for assessing personality at a narrower-level (the little 27) and at the more familiar domain-level (the big five).

Figure 2.

The two independently derived trait levels of the SAPA Personality framework (Condon, 2022)



Nested factor structures, like the one shown in Figure 1, are more organized and interpretable than structures that do not nest lower-order traits strictly within a single broader trait. This is one likely reason why most personality frameworks—including all described up to this point—have been created using a top-down approach. However, identifying lower-order traits by simply decomposing the content of each higher-order trait, top-down approaches sometimes fail to capture relevant traits at lower levels. By focusing initially on the item-level, bottom-up approaches provide enough flexibility for multiple types of structures to emerge. The infrequent use of bottom-up approaches and limited focus on the individual items could be explained by current views on optimal data analysis techniques. For example, latent variable modeling techniques are strongly encouraged in personality research, but these techniques assume that a top-down approach will be taken and focus primarily on information reduction.

The Reliance on latent-variable modeling

Due to the near absence of frameworks developed using a "bottom-up" approach, most studies have taken a de-facto top-down approach to the selection and measurement of personality traits. The abundance of top-down frameworks has presumably resulted from the early need for a common framework with a manageable number of individual differences. Domain-level frameworks like the Big Five model—which depict personality structure as relatively simple and organized—addressed this need. However, one likely reason why top-down structures remain popular is the reliance on latent-variable modeling techniques—which are promoted as the gold-standard for developing models of personality and measuring personality attributes (e.g., Bandalos, 2018; Borsboom, 2008; Condon & Mõttus, 2021; Rhemtulla et al., 2020).

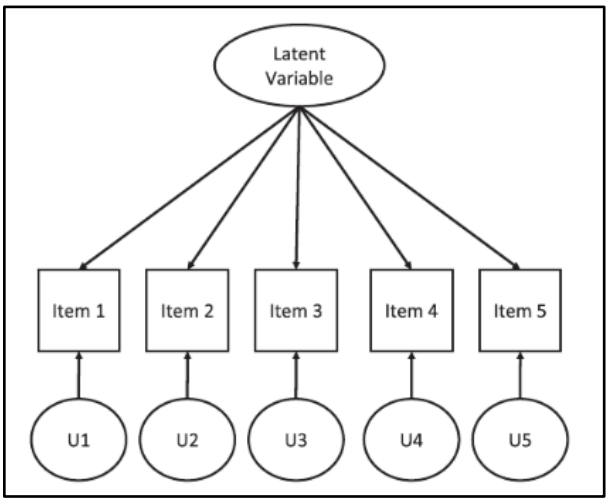
In this section, we begin with a review of latent variable modeling and the consequences of overreliance on these techniques for representing personality attributes (i.e., choosing a theoretical model) and measuring them. Representing and measuring personality attributes are highly connected but distinct activities that both frequently invoke the assumptions of the reflective measurement model (shown in Figure 3) and the statistical procedures associated with it (in particular, latent variable modeling (LVM) techniques). We describe how the popularity of LVM contributes to a disproportionate focus on broad levels of the trait hierarchy and shapes the way that individual items are viewed, selected, and analyzed in assessment contexts[4]. This section concludes with a review of how LVM became the dominant approach due to technical

[4] The choice to use LVM at all stems from the assumption that a reflective measurement model is theoretically optimal, but this possibility is not commonly acknowledged outside of a small area of the literature (e.g., Borsboom et al., [2003]. While it is important to consider alternative theoretical models (e.g., formative models, network models), we focus mostly on the disadvantages of LVM serving as the default statistical technique because of the clear ways in which this impedes knowledge production based on lower-level constructs.

limitations at the time of their development, and an alternative approach that could bring psychological assessment up to speed with technological advances.

Figure 3.

A reflective measurement model including five items and five sources of item-specific (or unique) variance. Image from Markus and Borsboom (2011).



The most common argument for using latent variable modeling relates to its advantages in the context of measuring personality attributes. When compared with the alternatives of sum or mean scoring, latent variable modeling is generally the superior method for generating scale-scores because these techniques can account for measurement error—the portion of variance that is not shared across items. The removal of this error component should result in scores that reflect one's standing on only the "true" construct component and not other influences (e.g., the phrasing of a particular item). This ability to account for irrelevant sources of variance has made

LVM techniques very useful. However, the use of LVM for this purpose implies that the attribute of interest is best represented as a single, reflective variable and that this dimension is accurately captured by the shared variance across items. While LVM techniques continue to be strongly encouraged (Borsboom, 2008), recent work demonstrates that latent variable modeling is not always the best tool for addressing several research questions and that improper application of these techniques can result in misleading conclusions (Rhemtulla et al., 2020).

Reconsidering the use of a latent variable model

The appropriateness of using a latent variable model is not frequently questioned in the context of measuring attributes (e.g., for purposes such as developing a scale or examining its relationship with other variables). This could be because LVM tends to be portrayed as merely a more rigorous alternative to sum/mean scoring rather than as a statistical technique associated with just one of many viable models with theoretical assumptions that need to be considered. However, the assumptions of latent variable models (e.g., tau equivalence, local independence) make for a variety of cases in which the consequences of using a latent variable model would be more severe than not accounting for measurement error at all (Rhemtulla et al., 2020). These include cases where the data are generated from a different model, such as a network model (e.g., in the case of depression, intelligence; Borsboom & Cramer, 2013; Van Der Maas et al., 2006) a formative/composite model (e.g., socioeconomic status), or a Multiple-Indicator Multiple Cause (MIMIC) model (Jöreskog & Goldberger, 1975). Model selection deserves careful consideration to avoid mischaracterizing the broad trait under consideration but it can also have important consequences for the role of the individual items, which are described as follows.

Identifying the most suitable model for an attribute can be challenging due to the availability of models that are mathematically equivalent (i.e., demonstrate equally good fit to

the data) but structurally distinct. For some attributes, model selection is a straightforward task because it is guided by clear theoretical arguments (e.g., modeling socioeconomic status as a formative variable; Borsboom et al., 2003). Adjudicating among equivalent models becomes difficult when there is a lack of consensus regarding the theoretically superior model, which appears to be the case when modeling broad personality traits. The two most frequently proposed models—latent variable models and network models (van Bork et al., 2021)—appeal to different but similarly plausible sets of assumptions regarding the nature of personality traits. For example, a domain-level personality trait such as Extraversion might be better represented as a network of mutually reinforcing indicators (e.g., talkativeness, positive affect; Cramer et al., 2012) than as a latent variable and common cause of its indicators. However, applying a latent variable model appears more justified from the perspective that the range of personality tendencies associated with extraversion are linked to fairly stable neurophysiological systems (i.e., behavioral activation/inhibition; Corr, 2004).

Some have proposed that the challenges of model selection might be best addressed by establishing a stronger understanding of the individual scale items—particularly, their causal relations with each other and other variables (Rhemtulla et al., 2020). This strategy appears sensible considering that model choice strongly determines whether the individual items themselves are viewed as relevant. Within a LVM framework, items are viewed primarily in terms of their relation to the broader latent variable rather than as potentially useful descriptive, predictive, or explanatory units. This is because LVMs assume that the only relevant portion of the variance is that which is shared across indicators (the “true score” variance), and any variance that is unique to an individual item is relegated to measurement error.

Removing unique item variance can lead to improved estimates of scale scores when it is mostly associated with aspects like the specific phrasing of an item. However, scales for measuring broad traits are typically composed of diverse content (e.g., designed to measure distinct aspects or facets), and focusing on solely the shared variance across indicators can result in a loss of meaningful personality-related information. This provides a strong reason for considering an alternative model that accords individual items a higher degree of relevance or causal force (e.g., a network model). However, it is uncommon for studies to begin with a thorough consideration of the individual items, perhaps because guidelines for modeling and measuring personality attributes center broad-levels of the trait hierarchy and suggest few alternatives to LVM.

The rise of latent variable modeling

The strong advocacy for LVM techniques can be better understood when considering the technological challenges they were developed to address and how emerging test validity theories at the time—which remain highly influential—were entangled with these practical concerns. The popularity of latent variable modeling can be traced back to the work of Cronbach (1960, 1955), who argued that psychological modeling techniques should be chiefly aimed at reducing information (the data) to a manageable set of constructs. This approach was suited for limitations such as the absence of statistical software tools and web-based recruitment methods, which demanded a stronger focus on parsimony with respect to data collection and analysis. Study designs that focused on relatively few, broad constructs (e.g., using brief self-report measures) and statistical techniques aimed at information reduction (e.g., factor analysis, the correlation coefficient, internal consistency metrics) became standard. Technological innovations of later decades cleared many of these barriers. Web-based recruitment methods and advances in data

collection and analysis techniques that do not require "full information" (e.g., random administration of items and analyses of pairwise combinations) made it possible to collect data on both a large number of participants and on a large number of items/variables.

Approaches to personality assessment have not evolved in step with these technological advances. Latent-variable modeling is still widely considered the gold standard technique for measuring personality attributes (Bandalos, 2018; Rhemtulla et al., 2020) and has only grown in popularity in response to growing awareness of the consequences of measurement error (e.g., attenuated correlations between variables; Cole & Preacher, 2014). The notion that LVM is nearly always a superior approach has spread to a variety of assessment contexts—educational and military settings, and occupational hiring and placement processes (Condon & Mõttus, 2021)—and it has become common to criticize approaches that depart from LVM techniques (Rhemtulla et al., 2020). This has resulted in a definition of methodological rigor that is somewhat narrow, yet has attained sweeping influence across diverse sectors, leading some to propose that psychological assessment is characterized by a "Latent-Variable Theory Psychological Testing complex" (Condon & Mõttus, 2021).

Dominant guidelines to psychological assessment are infused with many of Cronbach's concerns regarding the limitations on data collection and analysis techniques. However, alternative approaches to psychological assessment that do not prioritize information reduction have been proposed, which have potential to bring assessment practices up to speed with technological advances. Condon and Mõttus (2021) discuss emerging perspectives on psychological assessment at the beginning of the information age—contrasting Cronbach's (1955b) framework with a more "information-theoretic" approach focused on detection of the most important features of an individual instead of parsimonious data collection. This alternative

approach to testing was inspired by Shannon's information theory, which was originally developed as a new foundation for telecommunication systems (Shannon, 1948).

In the context of telecommunication, Shannon's information theory is concerned with efficient transmission of the information contained in any type of message (e.g., images, letters, numbers)—all of which can be represented using the common system of the 1s and 0s of computer binary code. This represents a departure from prior views, which focused more on the type of message being transmitted rather than on efficiency of transmission. As information theory grew in popularity, some began exploring its applicability in the realm of psychological assessment—which is characterized by the goal of retrieving accurate personality-related information from individuals. Although a more information-theoretic approach to assessment appears promising today, this proposal was swiftly dismissed by Cronbach (1953, 1955b) in its early stages of development due to its ostensible incompatibility with emerging standards for test validity. In retrospect, the incompatibility between information theory and psychological assessment appeared to lie in the practical issues of limited technology rather than more theoretical ones offered by Cronbach. These technological limitations precluded further development of an information-theoretic approach to personality assessment, which left the theoretical landscape clear for Cronbach's prescriptions to take root.

Emerging views on test validity also support a shift in focus from test analysis (e.g., examining unidimensionality at the scale-level) to the more item-centered task of test construction. Borsboom et al. (2004) express concerns regarding the ways that latent variable modeling techniques are used from the perspective of validity theory. Test validation research emphasizes broad constructs and their location in nomological networks, which Borsboom et al. (2004) claim has led to the inaccurate view that the validity of a measure consists in its structural

properties, test-retest reliability, and relation with other scales. Instead, Borsboom et al. (2004) claim that validity is a property of the scale that precedes any of these empirical qualities and concerns only (a) whether an attribute exists and (b) whether variations in the attribute causally produce variation in the measurement outcome[5]. Although it is important to evaluate the dimensionality of a measure and its relationships with other scales, Borsboom et al. (2004) suggest that validation research should be chiefly concerned with establishing a theory of response behavior (i.e., a theory of how an attribute is linked to responding at the item level). Similar views have been expressed in the work of Embretson (1983), which discusses the need to understand the processes underlying task performance.

How item-level assessment advances the three goals of prediction, description, and explanation

Despite the popularity of the big five framework, the value of more narrowly measured traits has long been acknowledged. The challenge, as described by Block (1995), lies in determining the optimal level of specificity and a method for identifying lower-level traits. Costa and McCrae's (1980) Big Five facets remain popular, but they were identified in a relatively subjective manner rather than through any empirically or theoretically informed strategy. Since the publication of this framework several others have been developed using various top down and bottom-up approaches. Findings based on a variety of these frameworks continuously demonstrate that traits below the domain-level contribute to each of the broad goals of prediction, description, and explanation.

[5] Note that this revised concept of validity offered by Borsboom et al. (2004) assumes a realist perspective on psychological attributes, which may not align with views of the more pragmatically inclined. However, arguments for a stronger focus on the item level can be found in views across the realist-antirealist spectrum.

Item-level assessment and predicting outcomes

The advantages of item-level assessment are most frequently discussed in the context of predicting outcomes. Models using the individual items to predict outcomes relating to educational attainment, health, and well-being routinely outperform facet- and broad-trait models (Mõttus et al., 2022). This is often true even after removing the variance associated with the broader domain from the items and predicting outcomes using only the items' unique variance (Seeboth & Mõttus, 2018). Revelle et al. (2021) suggest that focusing on the item-level could also improve understanding of the phenotypes associated with particular outcomes and demographic variables and, particularly, the degree of overlap between these phenotypes. For example, self-reported exercise frequency and self-reported wellness were only weakly correlated (r = .15), but the correlation between the two sets of items that predicted exercise frequency and wellness was much stronger (r = .67). This information could point to potential mechanisms linking exercise and wellness, which might reflect a more complex interaction between nuance-level traits than can be detected via domain-level measurement.

Studies that focus more on the predictive accuracy of standalone items, facets, and domains than on larger models also support a more item-level approach to predicting outcomes. For example, Arumäe et al. (2022) examined the relationship between personality traits and BMI using three 200-plus item questionnaires—NEO-PI-3, IPIP-NEO, 100NP—and demonstrated that the strongest correlations emerged at the "nuance" (item) level. While associations with broad traits were generally small (.10 or smaller) or non significant, clear relationships with BMI emerged for specific individual items related to orderliness (-), immoderation/impulsivity (+), yelling at others (+), conservative beliefs (+), and talkativeness (+). Additionally, the predictive accuracy of broader domains appeared to depend somewhat on inventory. For example, the

correlation between BMI and neuroticism ranged from .02 to .11 across inventories and correlations with openness ranged from -.09 to .01. This could reflect true differences across the culturally diverse samples used but also suggests that the differences in the particular items across inventories may have played a role.

These findings demonstrate the value of the unique variance found at the item-level, which cannot be explored when taking the conventional factor analysis approach and aggregating items into broader domains. We do not review additional examples demonstrating the advantages of item-level assessment for prediction-focused research here, as this has been thoroughly covered in previous work. See Achaa-Amankwaa et al. (2021), Mõttus et al. (2022), and Revelle et al. (2021) for examples of the benefits of item-level assessment in predicting many different life outcomes (e.g., educational attainment, health) and demographic variables (e.g., country of residence, age).

Item-level assessment and descriptive research

Descriptive personality research is concerned with establishing relationships between personality constructs and other variables such as demographic characteristics, behaviors, and life outcomes[6]. Descriptive findings are inherently valuable—if sometimes viewed as uninteresting (Baumeister, 2016)—but they also provide the foundation for prediction- and explanation-focused studies. Mõttus et al. (2020) propose that descriptive findings should be reported at the narrowest relevant level (i.e., usually the item/nuance level). This is because narrower traits can always be aggregated to examine the relationships of interest with broader traits but the reverse is not always possible, as some relevant traits at narrower levels might fall

[6] This differs from prediction-focused research, which uses individual difference variables to predict outcomes at a future time-point.

interstitially (between multiple broad traits) or outside of the trait space of the big five altogether (as demonstrated by Saucier & Iurino [2019]). However, it is most common for researchers to aggregate narrower traits a priori to examine relationships at the domain-level.

The prevalence of descriptive studies that only report relationships for "big few" traits might be explained by a few factors. First, because models like the Big Five are familiar. The big five and other "big few" models were developed against the backdrop of classical test theory, which emphasizes parsimony as an important model feature and implies that administering many items per construct is critical to addressing measurement error (Cronbach, 1951). Low-dimensional models appear to satisfy many of the measurement guidelines associated with classical test theory. Second, the tendency to report only on broad-level traits might also stem from premature claims that models like the big five or six are highly generalizable or even universal (McCrae & Costa, 1997). These claims have led to the suggestion that the big five model has "obtained the characteristics of an icon, with features that transcend the empirical facts" (De Raad, 2009, p.128).

Moreover, while few studies explicitly test whether these constructs are comparable across groups, even strong evidence for measurement invariance would not rule out the possibility that personality structure varies meaningfully across contexts (e.g., cultures, ethnicities, genders). It can be useful to know that a particular questionnaire demonstrates measurement invariance, but this can still only indicate that it can reasonably be used across groups—not that it is the optimal framework for describing personality in each of them. Consequently, selecting a framework like the big five to measure personality across contexts based on support for measurement invariance could still result in misleading descriptive findings. The potential heterogeneity across contexts in the optimal factor structure is important to

evaluate thoroughly, because findings on personality structure—which constitute a type of descriptive work in their own right—directly influence the development of the questionnaires used to establish many other findings in personality psychology[7].

The use of measurement invariance analyses alone for establishing generalizability has been an especially strong point of criticism in the area of personality assessment across cultures. More broadly, this is related to the disagreement regarding the relative importance of "etic" versus "emic" findings in cross-cultural research. Early studies comparing personality across cultures overwhelmingly adopted an "etic" strategy, which involves translating measures developed in one (typically Western) cultural context for administration in another. Findings from etic studies are useful but should be evaluated alongside findings from emic studies, which identify the locally optimal personality structure (e.g., for a nation or particular region) based on the most meaningful personality concepts in that specific context. Without findings from emic studies, it remains unclear whether important personality variation is obscured or missed entirely in some contexts, which compromises the quality and utility of descriptive findings. For example, the concept of "amae" represents an important trait dimension in Japanese lexical studies, capturing individual differences in the desire to be indulged by others and behaving in ways to evoke this reaction (Behrens, 2004). This individual difference can likely be translated and understood by others across many contexts (making it a "weak universal" [Thalmayer et al., 2022]), but it is not captured in the most popular, "universal" models of personality despite being an important individual difference in some contexts.

[7] Importantly, questionnaires are relied upon heavily in personality research but are not the only tool for establishing useful findings. More time and resource intensive approaches that use devices such as the Electronically Activated Recording (EAR) device, or that involve studying participants' life narratives, offer a lens through which to view personality that questionnaires do not.

It is not always possible to undertake an emic approach. A lack of time, resources, and the requisite cultural expertise are common barriers. In these cases, reporting findings at the item level in addition to, or perhaps in place of the broader domain level, could provide the most rigorous "etic" strategy for studying personality across contexts. This is particularly likely if "etic" studies administer items developed in a variety of contexts rather than a set of items from a single context, as is frequently done with items from a Western cultural context. Although reporting findings at the item-level would not overcome the inherent limitations of "etic" approaches, it would lessen the possibility that important cultural differences (or similarities) are masked in the process of aggregating items into broader traits. Examining relationships at the item level could also introduce new questions regarding personality expression across contexts and provide more direction when developing hypotheses and planning future studies.

Improving understanding and measurement of personality across cultures has been identified as a priority for the field (Syed & Kathawalla, 2022), and most research on the universality of personality structure across groups (or lack thereof) has involved cross-cultural comparison. However, the use of low-dimensional models like the big five is likely an issue when comparing many types of groups. For instance, recent findings indicate that personality structure also varies across age groups (Beck et al., 2023). While it is possible to uncover a big five structure in different age groups using factor analytic techniques, this does not indicate that a five factor structure provides the optimal number or nature of dimensions for characterizing personality within each age group. Beck et al. (2023) used network analyses—which do not presuppose a specific factor structure—to capture differences in personality structure across age groups. Items were found to demonstrate more cross-loadings and higher cross-loadings in older adult age groups, suggesting that personality structure is more complex at older ages. These

findings suggest that a closer evaluation of personality structure and the validity of the scales they generate is needed when conducting research in older populations, especially considering that some of this research focuses on important health-related outcomes.

Item-level assessment and explanation-focused research

The item-level of the trait hierarchy has clear value in advancing the broad goals of prediction (i.e., predicting outcomes) and description (i.e., describing personality at a more granular level). Item-level assessment is not discussed as frequently when considering the goal of explanation (i.e., explaining how stable personality differences arise). However, a stronger focus on personality at the item-level could further progress in this area as well. It is also timely to consider the potential contributions of item-level work to explaining individual differences as several have proposed explanation as the central task of personality research in coming years (Asendorpf, 2016; Baumert et al., 2017, 2019). This has been reflected in the uptick of publications focusing on personality dynamics (e.g., Revelle & Condon, 2015; Sosnowska et al., 2019, 2020), functioning (Wood et al., 2015), processes (Wrzus & Roberts, 2017), and mechanisms (Kuper et al., 2021).

To make progress on the goal of explanation, Baumert et al. (2019) argue that a stronger integration of structural, process-oriented, and developmental approaches to studying personality will be needed. This call for integration stems from recognition that trait dimensions found in structural models of personality should be understood more as descriptions of personality differences rather than as causes of behavior. Discussing causes of behavior without invoking factorial traits as explanations remains challenging despite acknowledgement that this reasoning becomes circular (i.e., when factorial traits are used to explain the behaviors that they are defined by; Blum et al., 2021; Cervone & Shoda, 1999). It is not clear whether (and when) broad,

domain-level traits should be viewed as causal units, or used to study developmental/process-related questions about personality. However, studies on these topics rely on the measures derived from structure-oriented work, which have almost exclusively been developed based on top-down models that center the domain-level. This is one potential barrier to bridging structural, process-oriented, and developmental approaches in personality research and could be addressed with a stronger understanding of personality and its measurement at the item-level.

The continued influence of big few models appears to be a double edged sword for explanation-oriented research. Although broad traits may not be the best units for modeling personality processes, the domain-level of the trait hierarchy offers a clear starting point for developing research questions. In this regard, big few models continue to fulfill the need for a common language among researchers that they did when personality psychology was first being established as a science. However, the potential consequences of relying on these models to shape research questions (e.g., uninformative or misleading findings) are unclear, and there appears to be little agreement even on the nature of the relationship between factorial traits and personality processes. This is evident in the variety of ways that researchers integrate broad traits into more dynamic models of personality processes. Frequently, broad traits are either (a) recast as processes themselves—as dynamic entities that emerge from the interactions between constituent lower-order traits or (b) incorporated as elements of processes which contain other individual difference constructs (e.g., values, beliefs, abilities).

Reconceptualizing broad traits as dynamic processes is appealing because it sidesteps the issue of circularity that arises when they are viewed as latent, psychobiological causes. It also appears plausible considering that the items used to measure domain-level traits typically reflect a mixture of affective, behavioral, cognitive, and desire/goal-related personality tendencies (i.e.,

Wilt and Revelle's [2015] ABCDs) which could be mutually reinforcing. As an example of this approach, Costantini et al. (2015) used network modeling to demonstrate that the stability of conscientiousness could be explained by the mutual reinforcement of conscientiousness indicators rather than by a latent, common cause of these behaviors. These findings suggest that a stronger focus on lower-levels of the trait hierarchy would be valuable for understanding the processes that underpin personality stability. However, studies that seek to explain stability at the broad-trait level using traits at lower levels are faced with new challenges when relying on top-down models.

In particular, it is unclear whether broad traits like Conscientiousness should be assumed to represent dynamic systems with most of their important components captured at lower-levels of the trait. This assumption is challenged by evidence that domain-level traits are non-ergodic[8] (Molenaar, 2004), meaning that for some individuals, the components making up conscientiousness at the population level are not as closely linked. For example, conscientious behaviors (e.g., industriousness) could be more strongly linked to motivations falling under another broad trait (e.g., an agreeableness-related desire to please others) rather than to motivations that are purely conscientiousness-related (e.g., a tendency to work hard regardless of interpersonal concerns). In sum, one consequence of using a domain-level trait to define the elements of a dynamic personality process is that potentially meaningful content from other traits is automatically excluded from consideration.

[8] In the context of personality structure, ergodicity concerns whether a given trait structure accurately depicts personality at both the population-level (e.g., trait structures derived via factor analysis of self-reported personality information from a large sample of individuals) and the individual-level (i.e., the number and nature of traits provide an accurate representation of an individual's personality).

Other process-oriented studies consider broad traits as elements of a process that interact with other individual differences such as values, goals, beliefs, and abilities. One benefit of this approach is that it involves a more explicit modeling of ABCDs as distinct categories of individual differences, which could differ meaningfully in their role or level of importance in personality processes. Studies taking this approach are often concerned with the relationship between traits and another domain like goals or values. One of the primary challenges of conducting this kind of research is determining how to study these constructs without blurring their conceptual or empirical boundaries. These distinct categories of traits, values, beliefs, goals, abilities, and others are useful at the conceptual level for discussing individual differences—most people likely have some sense of how one group of constructs differs from the others. However, it is only clear in this vague sense what is meant when contrasting traits with goals or values, making for a lack of precision that is not ideal for studying the relationship between these constructs scientifically.

As an example, Sagiv and Roccas (2021) review the relationship between values and other types of individual differences, and suggest that some distinguishing features of values are that they are inherently desirable (unlike traits) and trans-situational (unlike goals). According to Sagiv and Roccas (2021), traits are "descriptive in nature, defining what people are like, whereas values express motivations defining what people consider important. Unlike values, traits are not always desirable" (p. 296). However, it is unclear to what degree measures of similar trait constructs (e.g., Agreeableness) and value constructs (e.g., Benevolence) sufficiently uphold this distinction—either in terms of the items themselves or in the way these items are interpreted by participants. For instance, Agreeableness and Benevolence are both related to care or concern for others. Conceptually, these constructs might be distinguished as a tendency to display

care/concern for others (stemming from several possible motivations) versus an intrinsic motivation to show care/concern for others. Nonetheless, some measures of Agreeableness include items like “I pretend to be concerned for others” (reverse scored), implying that Agreeableness is defined to some extent by motivational or value-related content.

Conclusion

The development of the lexical approach in the early 20th century was a watershed for scientific progress. Many of the field’s achievements have resulted from the availability of lexically-based broad trait models, which offered a shared language and conceptual framework for studying individual differences systematically. The big five and similar models reflect a balance of parsimony and comprehensiveness that makes them relevant and practical tools for researchers studying a wide range of topics. However, the growing diversity in topics studied in personality psychology also demands greater flexibility in approaches to personality conceptualization, measurement, and statistical analyses. Considering this, the reliance on broad trait models has become somewhat limiting, and there is a clear need for a methodological paradigm shift to accommodate the unique measurement needs of the many topics addressed by the field.

In this review I discussed current views on personality assessment, described the historical context from which they arose, and explored the advantages of item-level approaches for addressing a variety of research questions. I began with a discussion of current views on methodological best practices—which center broad traits and scale-level statistical analyses—and ways in which these approaches do not meet the present needs of the field. Specifically, I focused on the limitations of latent variable modeling techniques serving as the default statistical approach in personality psychology due to the variety of assumptions they make about the nature

of personality traits—which might not be appropriate in all contexts. I end by discussing how an item-level focus could further progress on research topics across the three broad areas of prediction, description, and explanation.

The following two chapters focus on describing a demonstrating a more item-centric statistical technique for exploring cross-national differences in personality (Chapter 2), and applying this approach to a more comprehensive set of 2,623 personality items (Chapter 3). The research discussed in Chapter 3 was published in the Journal of *Current Research in Ecological and Social Psychology* and was conducted with my coauthors, Sara Weston and David Condon (see Mather et al., 2025).

II. Describing and demonstrating an item-focused alternative: the DTS test

In chapter 2 of this dissertation, I aim to address one of the primary barriers to an increased item-level focus: the limited availability of statistical techniques that center the item-level rather than treating items merely as indicators of broader constructs—even among techniques that appear more item-centric (e.g., Differential Item Functioning [DIF] analyses). I first discuss the limitations of using these common statistical approaches for examining findings at the item-level and then demonstrate an alternative statistical technique—the Dowd Test Statistic (DTS; Dowd, 2020)—for comparing groups at the item-level. Next, I apply the DTS test using data from the World Values Survey (WVS, https://www.worldvaluessurvey.org/wvs.jsp; Haerpfer et al., 2022; Inglehart et al., 2014) to compare nations on five distinct but overlapping features of well-being: happiness, health, perceived freedom and control, overall life satisfaction, and perceived financial security.

Current Approaches Fall Short When Addressing Item-level Questions

As reviewed in Chapter 1, individual items demonstrate many of the same trait-like properties as broader constructs—including heritability, stability over time, cross-rater agreement, and unique developmental trajectories (Mõttus et al., 2017; Mõttus et al., 2019; Seeboth & Mõttus, 2018; Henry et al., 2024). In addition, individual items often serve as more powerful predictors of important outcomes—even when relying only on their unique variance (i.e., the variance unrelated to a broader construct that is typically viewed as measurement error; Stewart et al., 2022). Taken together, this evidence suggests a need to reconsider the role of individual items in personality research. Here, I focus on a key methodological barrier that limits item-level work in practice: the limited availability of statistical techniques designed specifically for addressing item-level questions. More specifically, I describe and demonstrate a more item-

centric statistical approach to examining measurement invariance across groups that shows strong conceptual overlap with traditional approaches.

Traditional measurement invariance techniques face fundamental limitations when applied to item-level research questions because they are designed around the assumption of multiple indicators reflecting a common latent construct. For example, in multigroup confirmatory factor analysis (CFA), *configural* invariance establishes that the same factor structure emerges across groups, *metric* invariance demonstrates equal factor loadings across groups, and *scalar* invariance shows equal item intercepts across groups (Kline, 2023). Each of these steps requires that multiple items load onto a common latent factor—an impossibility when examining individual items as standalone constructs. When the item itself is the construct of interest rather than an indicator of a broader latent variable, the entire measurement invariance framework breaks down. There is no factor structure to compare, no loadings to constrain equal, and no meaningful way to establish scalar invariance.

This limitation extends to statistical techniques that appear to be more item focused, such as Differential Item Functioning (DIF) analysis, which aims to identify items that function differently across nations and other groups. However, DIF analysis fundamentally assumes an underlying latent trait and operates by matching individuals on their "true" level of this broader construct before comparing group responses for each item (Holland & Wainer, 1993; Millsap, 2011). Consequently, DIF findings derive their meaning primarily from their relationship to the broader scale rather than from independent item-level properties[9], making it suboptimal for

[9] DIF can be useful at the item-level independently from broader scales for assessing some non-substantive aspects of the psychometric quality of the items (e.g., detecting confusing wording, translation issues, response formatting problems, unnecessary linguistic complexity, and other possible phrasing issues).

exploring item-focused research questions. These constraints create a significant methodological gap: the statistical techniques designed to ensure cross-cultural validity cannot be meaningfully applied when researchers are primarily interested in comparing responses at the item-level—which, at times, might be the most informative level for understanding cultural differences in personality expression.

The prevalence of measurement invariance "failures" when attempting to develop generalizable scales may partly reflect this mismatch between statistical assumptions and research interests. The assumption of multigroup CFA and similar techniques that items are essentially interchangeable indicators of a latent construct is frequently violated in practice—leading many to conclude that measurement invariance standards are too stringent (Fischer & Karl, 2019). Instead, “failures” to demonstrate measurement invariance might signal a need to focus more strongly on the item-level earlier in the scale development process by exploring whether the items function similarly across contexts (through both qualitative and quantitative approaches[10]). Doing so could both improve our theoretical understanding of cross-cultural personality differences—some of which might be captured at the item-level—and improve the performance of scales when tested for measurement invariance.

ECDF-Based Approaches

One option for evaluating invariance in item response patterns across groups that does not assume the presence of a broader scale or latent construct could be to use an Empirical Cumulative Distribution Function (ECDF) based statistic, which are relatively common in fields such as environmental science, finance, and engineering but are not frequently used in

[10] Such as conducting more cross-cultural focus groups, cognitive interviewing procedures, ethnographic item development, and other qualitative techniques (Park et al., 2014; Willgerodt, 2003; Willis, 2005; Zhao et al. 2024).

psychology research. These statistical tests calculate the difference between ECDFs for two samples and provide an effect size (indicating the magnitude of the difference between groups) and a *p*-value (indicating whether the samples can be considered to represent the same broader population). Figure 4 illustrates this test: the distribution of responses to a single item is depicted for two countries, with differences in those distributions highlighted for each response option.

ECDF statistics offer several advantages over more familiar tests that could also be used for item-level comparisons, such as *t*-tests and *F*-tests. Specifically, ECDF-based statistics can detect a broader range of differences between group distributions, whereas t-tests and F-tests are primarily designed to detect differences in group means and variances, respectively. Using an ECDF-based statistic can thus prevent potentially meaningful differences in responding from being overlooked when groups do not show notable differences in means or variances. Additionally, ECDF-based tests allow for the detection of group differences using a single statistical test, which is advantageous when administering many items.

There are a variety of ECDF tests, including the Kolmogorov-Smirnov test (Smirnov, 1948), the Anderson-Darling statistic (Anderson & Darling, 1954), and the Wasserstein test (Villani, 2009). The present work will use the DTS test (Dowd, 2020), the most recently developed ECDF statistic which combines the strengths of the Anderson Darling statistic and the Wasserstein statistic. This test calculates the difference between ECDFs by taking the integral between the ECDFs while adjusting for variance (i.e., giving less weight to areas with higher variability, which are less stable). This approach improves statistical power by focusing on differences that are less likely to be due to sampling variability. The formula for DTS is as follows:

$$DTS = \int_{-\infty}^{\infty} \frac{|\hat{F}(x) - \hat{E}(x)|}{\hat{D}(x)(1-\hat{D}(x))} dx$$

Here, $\hat{F}(x)$and $\hat{E}(x)$represent the two samples being compared (e.g., United States and Philippines), and $\hat{D}(x)$ is the combined ECDF used to estimate the variance at each point *x.* To implement the test, a domain classifier is trained (on many bootstrapped samples of the data) to predict whether a particular response came from one sample or the other. For each response level x (i.e., Likert scale options 1 through 6), the model estimates the probability that the response originated from one group.

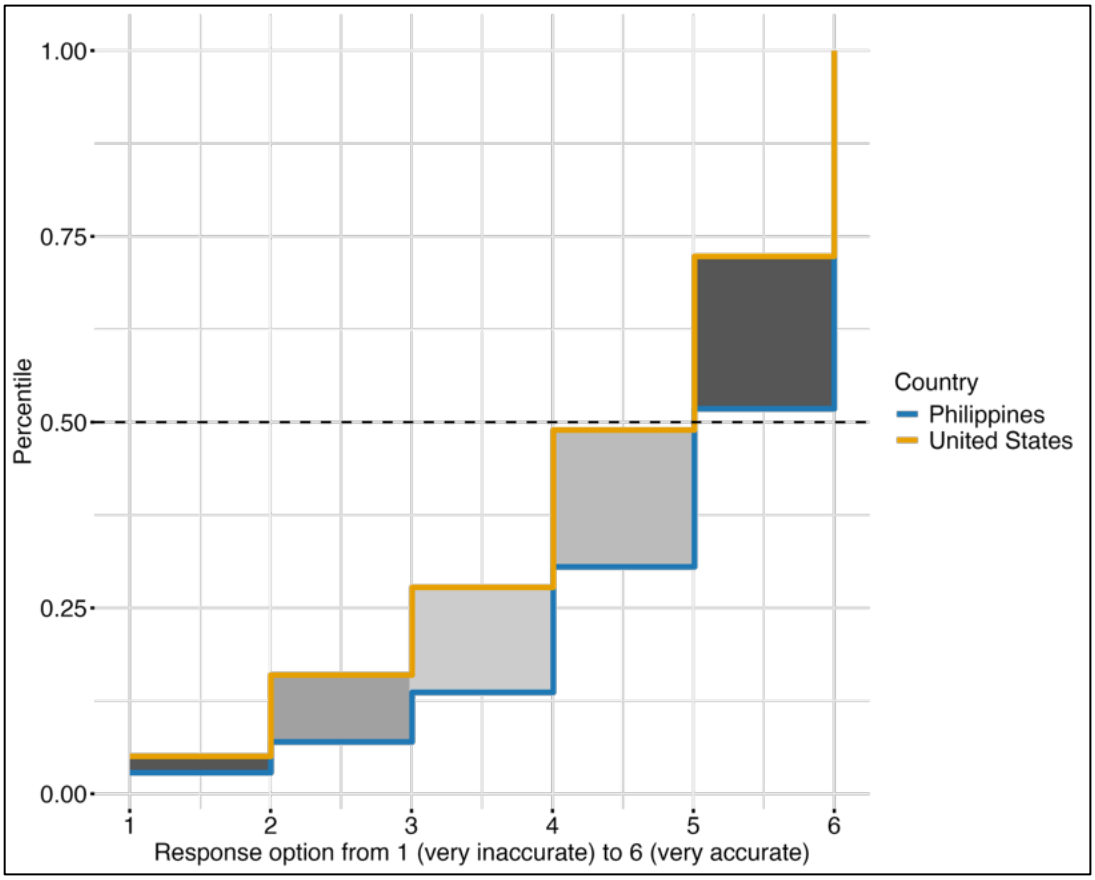
The function D(x)(1 - D(x)) appears in the denominator of the DTS formula, meaning that when D(x) is near 0 or 1 (i.e., when the classifier is confident about which group the response came from), the denominator is small and the contribution to the DTS statistic is larger. Conversely, when D(x) is near 0.5 (i.e., when the classifier is uncertain), the denominator is large and the contribution is smaller. This means that the DTS test gives greater influence to more stable, confident differences between groups and down-weights uncertain, noisier regions of the distribution. As an example, if $\hat{D}(3)$= 0.9, for the item *I am quiet around strangers*, the model predicts that the response of 3 (*somewhat inaccurate*) most likely came from the sample from the Philippines, whereas an output of $\hat{D}(3)$= 0.1 indicates that the model predicts the response came from the United States sample.

Effect sizes are unbounded in the DTS statistic and do not have a universal interpretive scale, like Cohen's d, so results are interpreted through relative comparisons. A permutation-based sampling approach is used to generate null distributions to estimate *p*-values, which reflect the probability of observing a DTS value as extreme as the one calculated under the assumption that the two samples come from the same population. This approach allows us to assess whether response distributions for individual personality items are likely to generalize across nations. In

the present study, I interpret statistically significant DTS tests ($p < .05$) as evidence what the distributions signal cultural differences, differential interpretation of item phrasing, response style variation, or other psychometric factors. Note that the DTS statistic itself is not explicitly standardized for sample size. However, the resampling procedure used to estimate the null distribution ensures that *p*-values are calibrated based on sample size. Figure 4 shows a visual depiction of a DTS test between the Philippines and the United States for the item *I am quiet around strangers* from the SAPA Personality Inventory (SPI; Condon, 2022).

Figure 4.

ECDF comparison between the Philippines and the United States for the item: I am quiet around strangers.



Note: This figure depicts the Empirical Cumulative Distribution Functions for samples from the United States and the Philippines on the item 'I am quiet around strangers'. Response options (1 through 6) are shown on the x-axis. The dashed horizontal line represents the 50th percentile. The DTS test indicated a significant difference between the group distributions (DTS = 94.67, p < .001).

Method

To demonstrate the use of the DTS test, I used data from Waves 6 and 7 of the World Values Survey (WVS, https://www.worldvaluessurvey.org/wvs.jsp; Haerpfer et al., 2022; Inglehart et al., 2014). The WVS is a project that collects data from a globally representative set

of nations on various topics such as social values and attitudes, happiness and well-being, economic values, science and technology, national security, ethical norms, and perceived corruption. Questionnaire responses are primarily collected through in-person interviews and web-based surveys. The analyses used the following five items from the Happiness and Well-Being subsection of the WVS to test two item-level research questions—one cross-sectional, and one longitudinal. These items were selected because they assess overlapping yet distinct features of well-being, including self-rated happiness, health, sense of control over one's life, overall subjective well-being, and financial security.

1. Taking all things together, would you say you are

1. Very happy
2. Rather happy
3. Not very happy
4. Not at all happy

2. All in all, how would you describe your state of health these days? Would you say it is…

1. Very good
2. Good
3. Fair
4. Poor
5. Very poor

3. Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means "no choice at all" and 10 means "a great deal of choice" to indicate how much freedom of choice and control you feel you have over the way your life turns out (code one number):

No choice at all A great deal of choice

1	2	3	4	5	6	7	8	9	10

4. All things considered, how satisfied are you with your life as a whole these days? Using this card on which 1 means you are "completely dissatisfied" and 10 means you are "completely satisfied" where would you put your satisfaction with your life as a whole?

Completely dissatisfied Completely satisfied

1	2	3	4	5	6	7	8	9	10

5. How satisfied are you with the financial situation of your household? Please use this card again to help with your answer

Completely dissatisfied Completely satisfied

1	2	3	4	5	6	7	8	9	10

Aim 1. First, I used the DTS test to examine cross-national differences in responding to these five items using data from Wave 7 of the WVS (Haerpfer et al., 2022). The Wave 7 dataset includes 64 surveys conducted across countries and societies around the world, with more than 80,000 respondents. Data for each nation were collected at a single time point between 2017 and 2023. To minimize potential confounding effects of the COVID-19 pandemic on global well-being, analyses included only the subset of nations with data collected between 2021 and 2023, thereby avoiding comparisons between pre- and post-pandemic samples. Participants under the age of 18 were also excluded. The final sample included 13,627 participants from 13 nations: Armenia (n = 1223), Czechia (n = 1200), India (n = 1692), Kenya (n = 1259), Libya (n = 1185), Maldives (n = 1036), Morocco (n = 1200), Netherlands (n = 2145), Slovakia (n = 1200), Great Britain (n = 2469), Uruguay (n = 1000), Uzbekistan (n = 1249), Venezuela (n = 1190). DTS comparisons were used to identify cross-national similarities and differences in the features of well-being assessed using the five items shown above.

Aim 2. Second, I explored within-nation differences in responses to these five items over time (e.g., the difference in distributions for India between Waves 6 and 7 on item Q49, which assesses overall life satisfaction). The nations included in these analyses were those with data collected at any point during Wave 6 and in 2021 or later (post COVID-19) during Wave 7. This subset includes 22,578 participants from seven nations: Armenia (n = 2,323), India (n = 5,743), Libya (n = 3,316), Morocco (n = 2,400), Netherlands (n = 4,047), Uruguay (n = 2,000), and Uzbekistan (n = 2,749).

One important consideration for this aim is the time span between data collection points, as most Wave 6 data for these nations were collected in 2011 or 2012, resulting in a roughly 10-year interval between Wave 6 and Wave 7 responses. While some changes in well-being may be attributable to the COVID-19 pandemic, each nation has experienced unique political, social, environmental, and economic developments during this period. Thus, pandemic-related challenges and disruptions may explain some observed differences, but a variety of other (potentially stronger and more lasting) factors could be driving any within-nation differences found in these analyses. This aim is mostly intended to demonstrate the range of questions that ECDF-based statistical tests can be used to address at the item-level.

Analyses

For both the cross-sectional and longitudinal research questions, I applied the DTS test to examine differences in empirical cumulative distribution functions—which identified differences in responding either (a) between nations during the same wave of data collection, or (b) within nations during two separate waves of data collection. For aim 1, this resulted in a total of 390 DTS tests for the 78 pairs of 13 nations on each of the 5 items. For aim 2, this resulted in a total of 35 DTS tests for the 7 nations on each of the 5 items. In addition to applying the DTS test for

both aims, I conducted *t* and *F* tests. The purpose of these analyses was to further explore whether the DTS statistic represents a more powerful alternative to more familiar analytical approaches (as demonstrated by Dowd [2020]), and can detect group differences that the *t* and *F* tests do not.

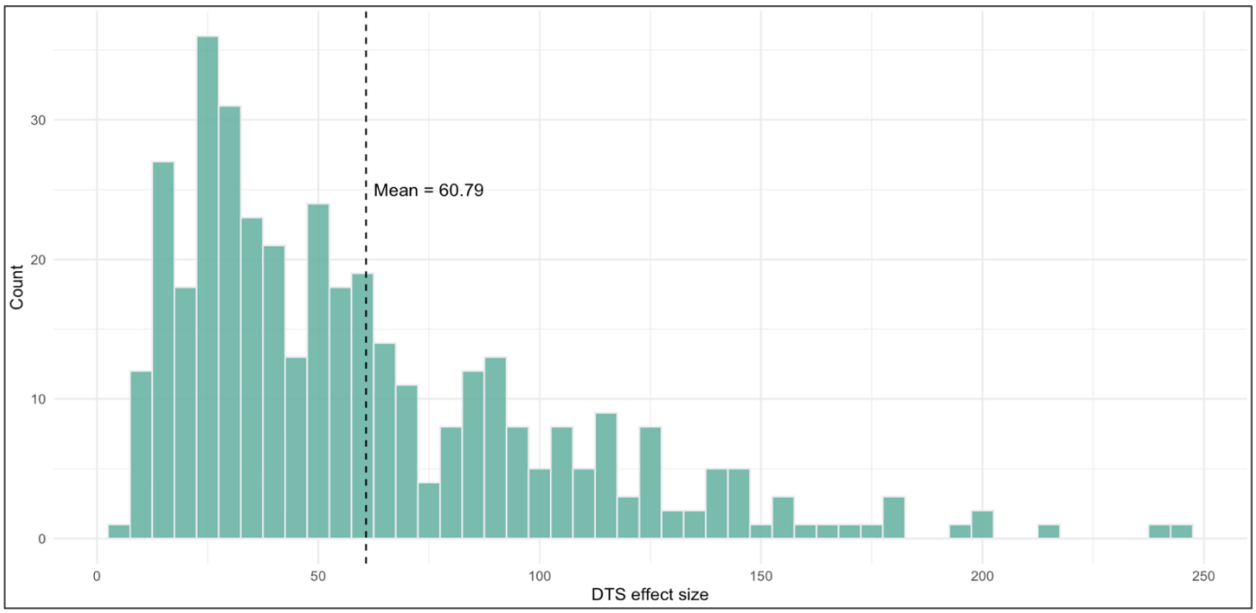
Results

Cross-national Comparisons: Description of overall findings

Overall, 383 out of the 390 DTS tests comparing each set of nations were statistically significant ($p < .05$). Non-significant DTS comparisons are shown in table 1. Significant DTS effect sizes ranged from 7.36 to 244.56 ($M = 60.79$; $SD = 43.51$; see figure 5 for the distribution of effect sizes found between nations). Because DTS effect sizes are interpreted in a relative way (rather than using an "objective" system with thresholds or cutoffs), the DTS effect size for any given comparison was considered larger or smaller in the present work based on where it ranks among the other DTS effect sizes identified. For example, considering the distribution of DTS effect sizes shown in figure 5, effect sizes from 0 to 10 might be considered relatively small, and effect sizes of 100 or greater would be relatively large.

Figure 5.

Histogram of DTS effect sizes for comparisons between nations across the 5 items



Note: only statistically significant DTS effects are included in this figure.

Table 1.

Non-significant DTS comparisons

Item	Content	Nation 1	Nation 2	DTS effect size	*P*-value
Q46	Happiness	Czech Rep	Morocco	4.11	0.27
Q47	Health	Kenya	Uzbekistan	7.39	0.07
Q47	Health	Uruguay	Venezuela	6.21	0.19
Q47	Health	Czech Rep	Maldives	5.05	0.33
Q48	Freedom	Czech Rep	Slovakia	10.42	0.40
Q49	Life Satisfaction	Uzbekistan	Venezuela	18.81	0.05

Q49	Life Satisfaction	India	Venezuela	16.74	0.07
Q50	Financial Situation	Maldives	Uruguay	13.80	0.17

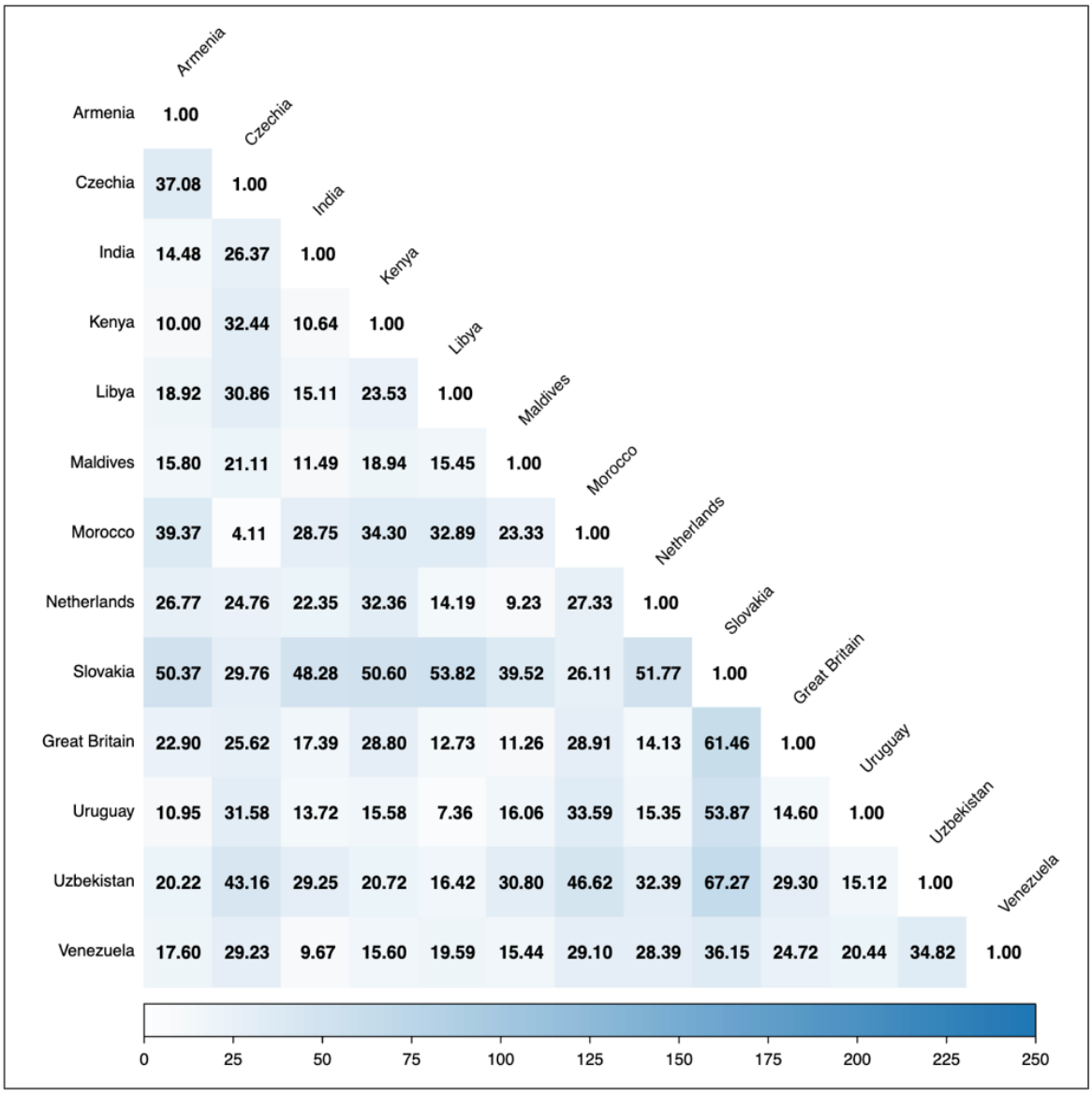
Note: The comparisons shown in this table reflect those that were not detected as statistically significant by the DTS test. The remaining 382 of 390 comparisons were identified as significant.

Cross national differences in happiness

Overall, the mean self-reported happiness—which ranged from 1 (Not at all happy) to 4 (Very happy) —was 3.17 (*SD*= 0.70). Nations differed somewhat regarding their self-reported happiness with Uzbekistan reporting the highest levels (*M*=3.47; *SD*=0.59), and Slovakia reporting the lowest levels (*M*=2.70; *SD*=0.68). Figure 6 displays DTS effect sizes for each pair of nations and reveals that the strongest differences in distributions were generally found between Slovakia and the other nations. All but one DTS test for self-reported happiness (comparing the Czech Republic and Morocco) were statistically significant. However, among all DTS tests conducted in the present work, effect sizes found for the self-reported happiness item were relatively small (*M* = 26.15; *SD* = 13.31). Figure 7 shows comparisons of the density distributions for three nation pairs on the happiness item—one that demonstrated a stronger difference in responding (Slovakia and Uzbekistan), one that demonstrated an average difference in responding (Kenya and Libya), and one that demonstrated virtually no difference in responding (Czechia and Morocco).

Figure 6.

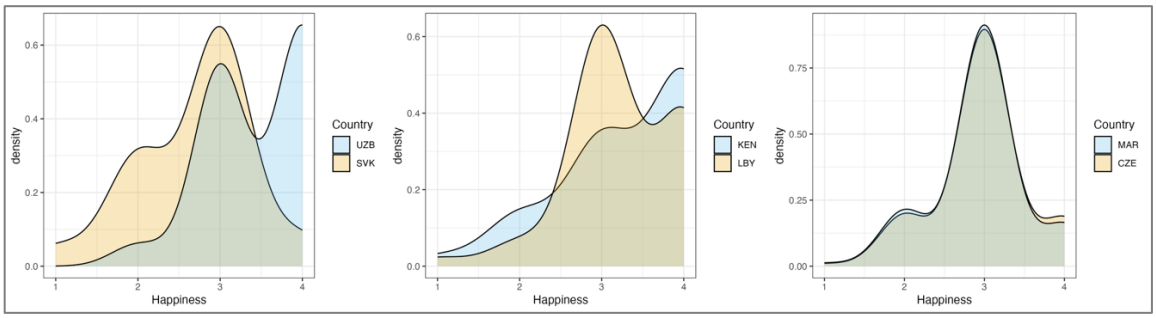
Self-reported Happiness: DTS effect size for each nation pair



Note: Each square represents a DTS effect size capturing the difference between two nations on the item "Taking all things together, would you say you are…(1) Not at all happy, (2) Not very happy, (3 Rather happy, (4)Very happy".

Figure 7.

Density distributions for three nation pairs on happiness



Note: These three plots reflect comparisons for the nation pair that demonstrated the strongest differences in responding (SLK and UZB), a nation pair that demonstrated differences in responding close to the average for this item (KEN and LBY), and a nation pair that demonstrated no meaningful differences in responding (CZE and MAR).

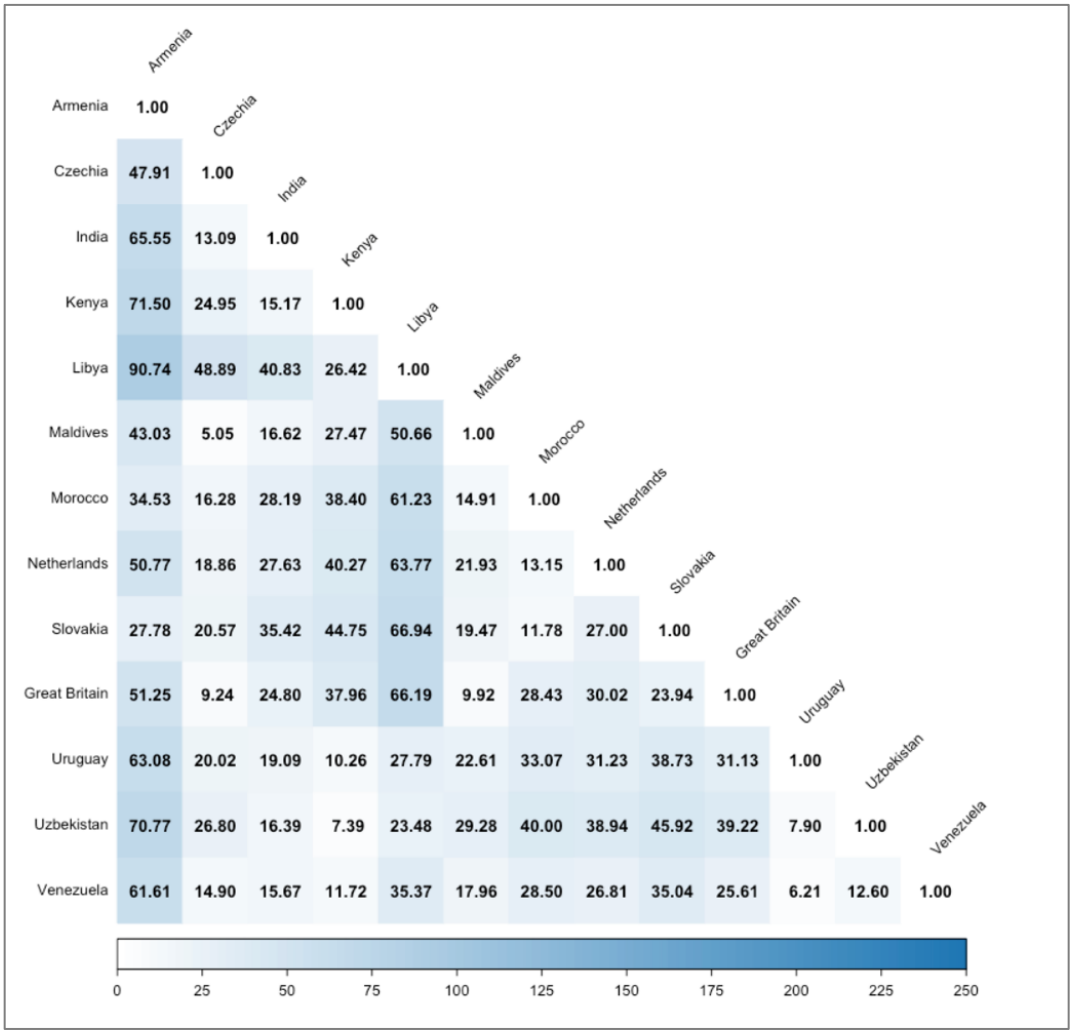
Cross national differences in health

Overall, the mean self-reported health—ranging from 1 (Very poor) to 5 (Very good)—was fairly high (M = 3.83; SD = 0.90). Nations showed some differences on average regarding their self-reported health, with Libya reporting the highest levels (M=4.34; SD=0.77), and Armenia reporting the lowest levels (M=3.26; SD=0.99). Figure 8 displays DTS effect sizes for each pair of nations and suggests that the strongest differences were generally found for Armenia with the other nations. However, similar to findings for self-reported happiness, the effect sizes found for the self-reported health item were relatively small (M = 31.90; SD = 18.30). In addition, four of the DTS tests for self-reported health were non-significant (tests for Uzbekistan with Uruguay and Kenya, the test between Uruguay and Venezuela, and the test between the Czech Republic and the Maldives). Figure 9 shows comparisons of the density distributions for three nation pairs on the health item—one that demonstrated a stronger difference in responding

(Armenia and Libya), one that demonstrated an average difference in responding (Netherlands and Uruguay), and one that demonstrated virtually no difference in responding (Maldives and the Czech Republic).

Figure 8.

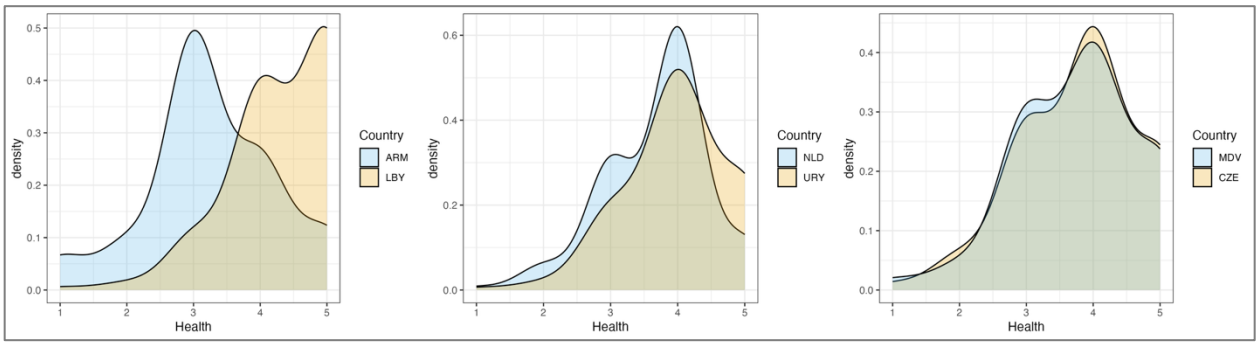
Self-reported Health: DTS effect size for each nation pair



Note: Each square represents a DTS effect size capturing the difference between two nations on the self-reported health item, which asks participants to indicate their perceived physical health from 1 (Very poor) to 5 (Very good).

Figure 9.

Density distributions for three nation pairs on health



Note: These three plots reflect comparisons for the nation pair that demonstrated the strongest differences in responding (ARM and LBY), a nation pair that demonstrated differences in responding close to the average for this item (NLD and URY), and a nation pair that demonstrated no meaningful differences in responding (CZE and MDV).

Cross national differences in perceived freedom and control

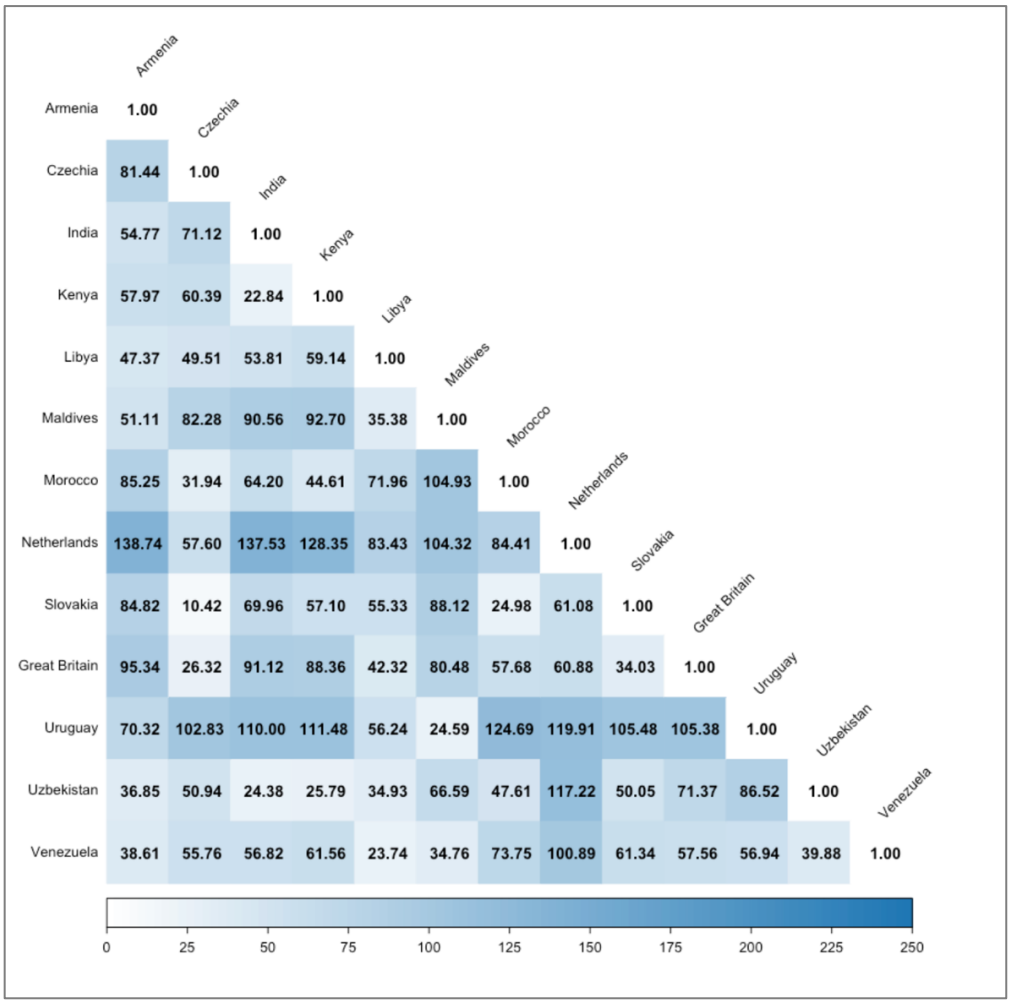
Overall, levels of perceived freedom and control were high across nations (M = 7.35, SD = 2.23), with Uruguay reporting the highest perceived freedom (M = 8.29, SD = 1.83) and Morocco reporting the lowest (M = 6.81, SD = 2.16). All DTS tests were statistically significant except for the comparison between the Czech Republic and Slovakia. Despite similar mean levels of perceived freedom across nations (as observed for self-reported happiness and health), DTS effect sizes for this item (see Figure 10) indicated larger differences between nations in their overall distributions (Mean DTS = 67.70, SD = 29.84).

These differences appear to be driven more by higher within-nation variance on this item, as well as differences in variances across nations. For example, participants from the Netherlands reported moderate mean levels of perceived freedom but showed some of the strongest distributional differences compared to other nations. This may be explained by the relatively

lower variance in responses among participants from the Netherlands. Figure 11 presents density distribution comparisons for three pairs of nations on the perceived freedom item: one pair showing a strong difference in responding (Armenia and the Netherlands), one showing an average difference (Uzbekistan and the Maldives), and one showing no difference (Slovakia and the Czech Republic).

Figure 10.

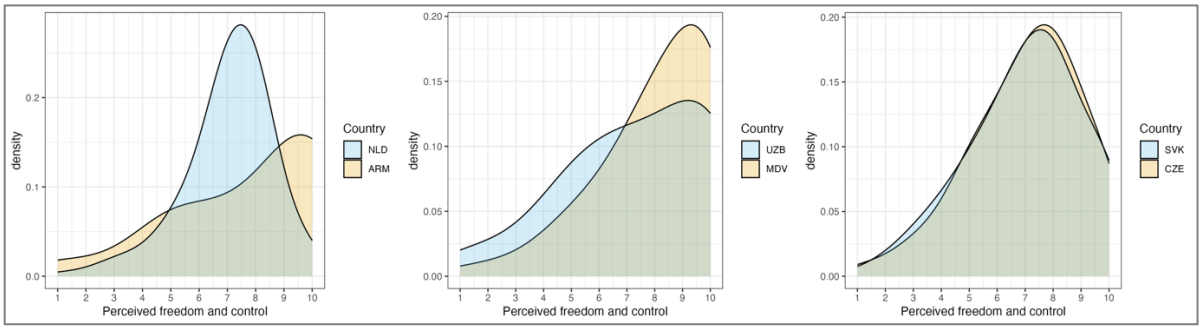
Self-reported Perceived Freedom and Control: DTS effect size for each nation pair



Note: Each square represents a DTS effect size capturing the difference between two nations on the self-reported freedom and control item, which asks participants to indicate the extent they feel a sense of freedom and control over their life and choices from 1 (No choice at all) to 10 (A great deal of choice).

Figure 11.

Density distributions for three nation pairs on perceived freedom and control



Note: These three plots reflect comparisons for the nation pair that demonstrated the strongest differences in responding (ARM and NLD), a nation pair that demonstrated differences in responding close to the average (MDV and UZB), and a nation pair that demonstrated minimal differences in responding (CZE and SVK).

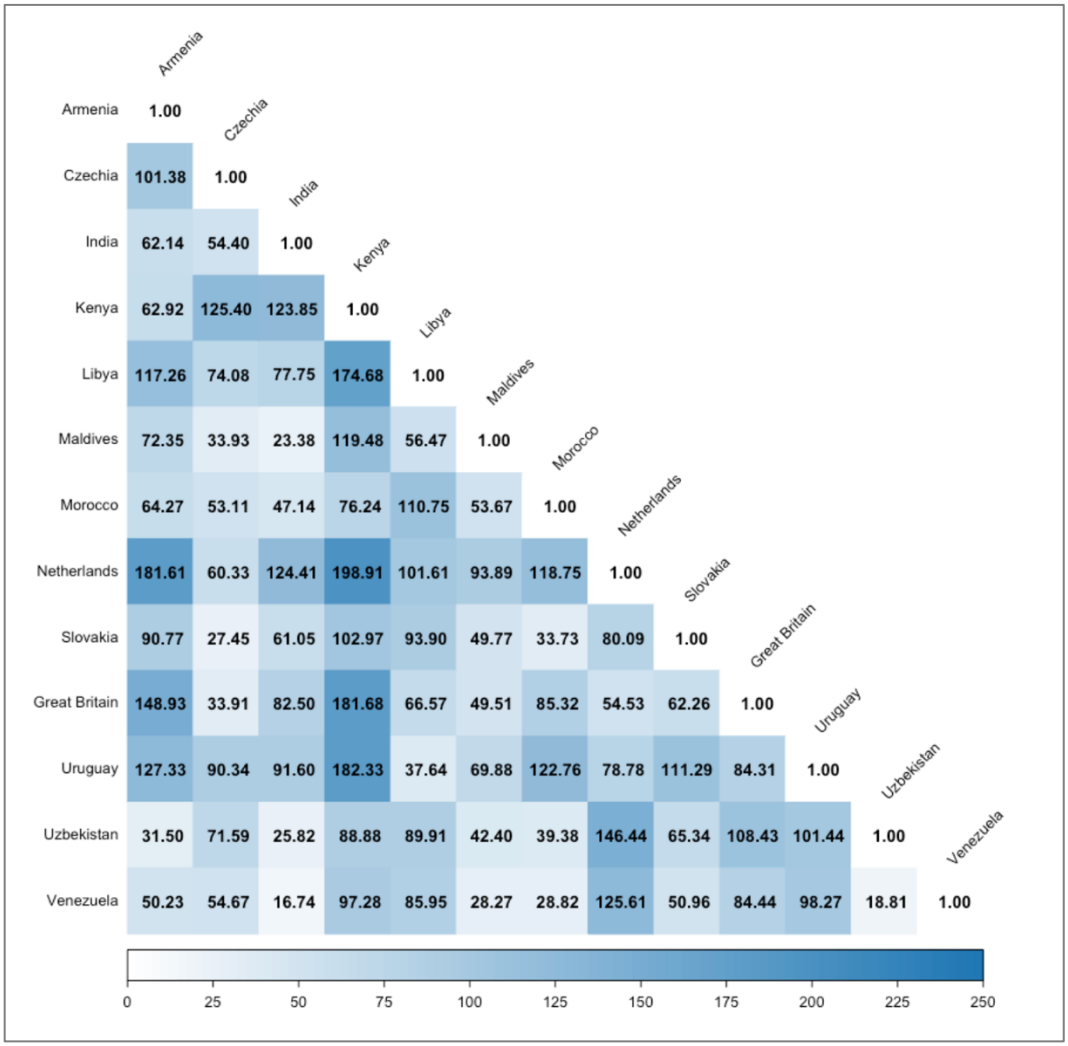
Cross national differences in overall Life Satisfaction

Nations varied more in mean levels of overall life satisfaction (M = 7.11, SD = 2.24), with Uruguay reporting the highest average life satisfaction (M = 8.13, SD = 1.84) and Kenya reporting the lowest (M = 5.86, SD = 2.68). All DTS tests were statistically significant except for two comparisons for Venezuela with Uzbekistan and India. Consistent with findings for the perceived freedom item, nations demonstrated stronger overall differences in distributions for life satisfaction (Mean DTS = 82.26, SD = 41.49). In general, nations with higher DTS effect sizes (see Figure 12) tended to exhibit either substantially higher variance (e.g., Kenya, Armenia) or lower variance (e.g., the Netherlands) in life satisfaction. Figure 13 presents density distribution comparisons for three pairs of nations on the overall life satisfaction item: one pair

showing a strong difference in responding (the Netherlands and Kenya), one showing an average difference (the United Kingdom and Uruguay), and one showing a smaller difference (Venezuela and the Czech Republic).

Figure 12.

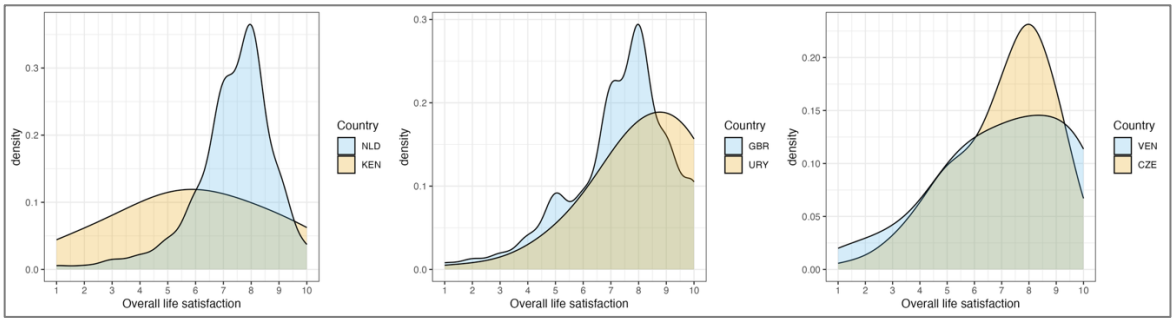
Self-reported Life Satisfaction: DTS effect size for each nation pair



Note: Each square represents a DTS effect size capturing the difference between two nations on the self-reported life satisfaction item, which asks participants to indicate the extent they feel satisfied with their lives overall from 1 (Completely dissatisfied) to 10 (Completely satisfied).

Figure 13.

Density distributions for three nation pairs on overall life satisfaction



Note: These three plots reflect comparisons for the nation pair that demonstrated the strongest differences in responding (KEN and NLD), a nation pair that demonstrated differences in responding close to the average for this item (GBR and URY), and a nation pair that demonstrated smaller differences in responding (CZE and VEN).

Cross national differences for satisfaction with financial Situation

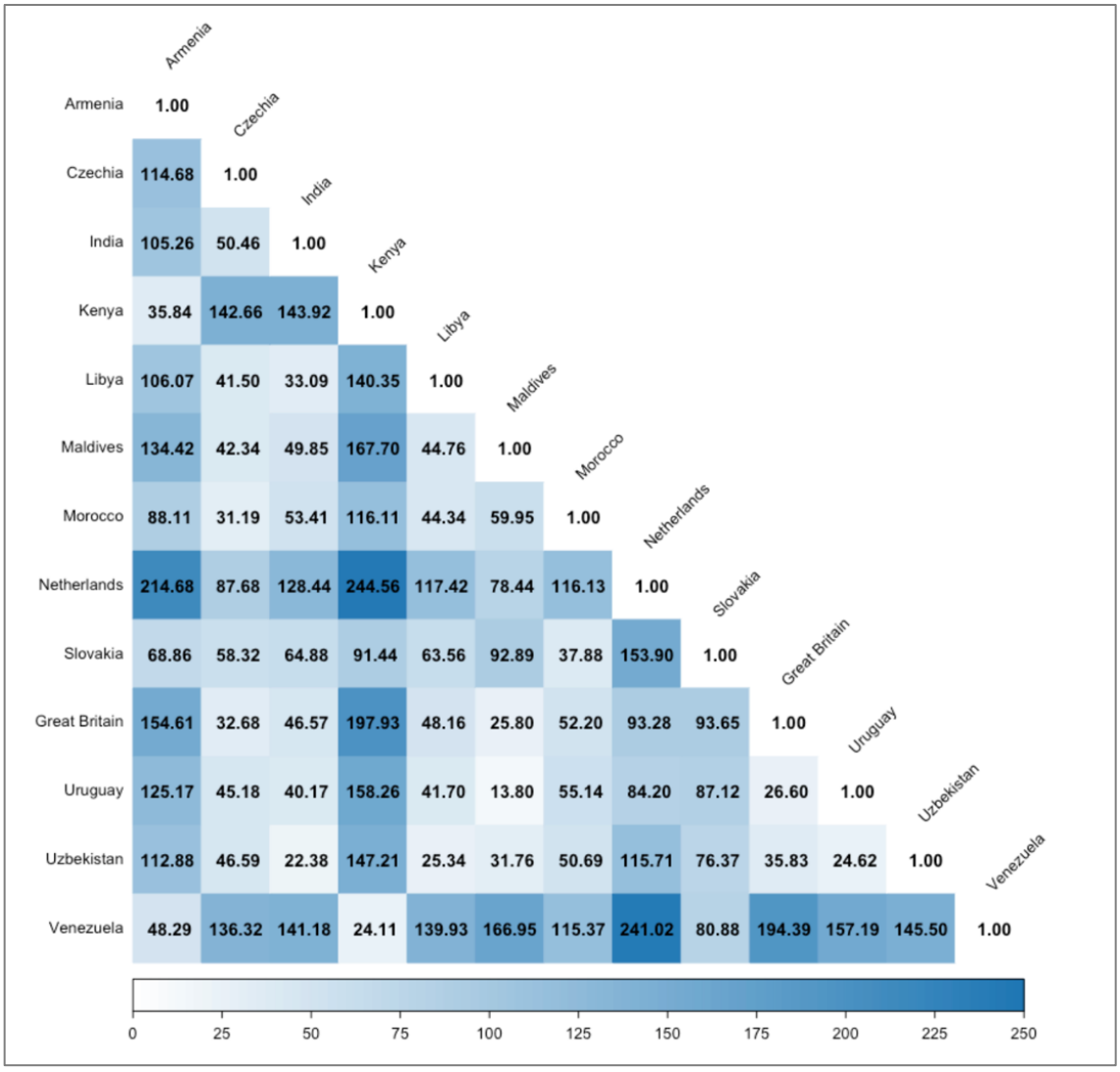
Some of the strongest differences identified between nations were in self-reported satisfaction with one's financial situation, with response options ranging from 1 (not at all satisfied) to 10 (completely satisfied). Overall, the mean level of financial satisfaction was 6.32 (*SD* = 2.45), with the Netherlands reporting the highest average levels (*M* = 7.28, *SD* = 1.69) and Kenya reporting the lowest (*M* = 4.86, *SD* = 2.69).

DTS effect sizes comparing nations were highest for this item (*M* = 90.61, *SD* = 55.21), and some of these values—particularly those comparing the Netherlands with Kenya (DTS = 244.56), Armenia (DTS = 214.68), and Venezuela (DTS = 241.02)—were especially large (i.e., more than two standard deviations above the overall mean; see Figure 14). All DTS tests were statistically significant except for the comparison between the Maldives and Uruguay. Figure 15 presents density distributions for three pairs of nations on the financial satisfaction item: one pair

showing a large difference (the Netherlands and Kenya), one showing an average difference (Kenya and Slovakia), and one showing virtually no difference (the Maldives and Uruguay).

Figure 14.

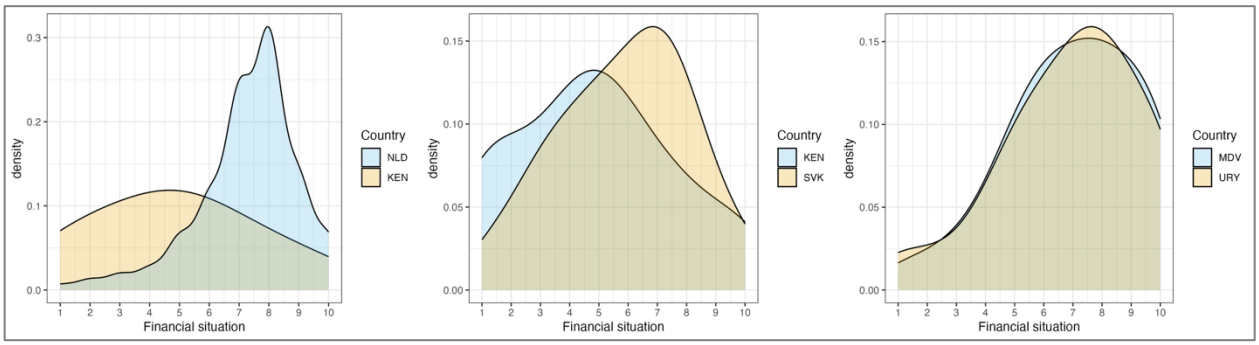
Self-reported Satisfaction with Financial Situation: DTS effect size for each nation pair



Note: Each square represents a DTS effect size capturing the difference between two nations on the satisfaction with financial situation item, which asks participants to indicate the extent they feel satisfied with their financial situation from 1 (Completely dissatisfied) to 10 (Completely satisfied).

Figure 15.

Density distributions for three nation pairs on satisfaction with one's financial situation



Note: These three plots reflect comparisons for the nation pair that demonstrated the strongest differences in responding (KEN and NLD), a nation pair that demonstrated differences in responding close to the average for this item (GBR and URY), and a nation pair that demonstrated smaller differences in responding (CZE and VEN).

Cross national differences detected by DTS that traditional approaches missed

In addition to the DTS comparisons, *t*-tests and *F*-tests were run for each nation pair to compare the DTS test with more familiar statistical approaches for comparing two groups. Of the 390 total pairs across the 13 nations, the DTS test detected the highest number of significant differences (383 nation pairs), followed by the *t*-test (332 nation pairs), and then the *F*-test (309). Table 2 displays all tests reflecting differences detected by the DTS test, but not the *t*-test, *F*-test, or both.

Table 2.

All differences detected by DTS but not by the t-test, F-test, or both

item	nation_1	nation_2	DTS stat	Tstat	Fstat	DTSpval	Tpval	Fpval
Differences missed by both the t and F test								
Q46	ARM	KEN	10.00	0.12	0.94	0.00	0.91	0.28
Q48	IND	UZB	24.38	-1.60	1.08	0.01	0.11	0.13
Q48	LBY	VEN	23.74	-1.47	0.96	0.01	0.14	0.51
Q48	IND	KEN	22.84	1.49	1.06	0.02	0.14	0.24
Q50	IND	LBY	33.09	-0.90	1.09	0.00	0.37	0.11
Q50	GBR	URY	26.60	-1.88	0.94	0.00	0.06	0.21
Q50	LBY	UZB	25.34	-1.56	0.90	0.00	0.12	0.08
Differences missed by F test only								
Q46	SVK	URY	53.87	21.98	1.02	0.00	0.00	0.72
Q46	LBY	SVK	53.82	-20.65	0.94	0.00	0.00	0.25
Q46	MAR	UZB	46.62	22.58	0.93	0.00	0.00	0.22
Q46	CZE	UZB	43.16	21.05	0.95	0.00	0.00	0.33
Q46	MDV	SVK	39.52	-15.75	1.00	0.00	0.00	0.97
Q46	NLD	UZB	32.39	14.41	1.09	0.00	0.00	0.09
Q46	ARM	VEN	17.60	-5.66	0.92	0.00	0.00	0.13
Q46	MDV	URY	16.06	6.13	1.02	0.00	0.00	0.70

Q46	KEN	VEN	15.60	-5.73	0.97	0.00	0.00	0.66
Q46	LBY	MDV	15.45	-3.90	0.93	0.00	0.00	0.25
Q46	GBR	URY	14.60	4.79	0.92	0.00	0.00	0.11
Q46	ARM	IND	14.48	-3.30	1.08	0.00	0.00	0.13
Q46	GBR	LBY	12.73	2.02	0.96	0.00	0.04	0.42
Q46	LBY	URY	7.36	2.56	0.96	0.01	0.01	0.47
Q47	LBY	NLD	63.77	-22.92	1.03	0.00	0.00	0.54
Q47	LBY	MAR	61.23	-21.91	0.90	0.00	0.00	0.07
Q47	ARM	GBR	51.25	14.57	1.04	0.00	0.00	0.38
Q47	MAR	UZB	40.00	13.23	0.99	0.00	0.00	0.84
Q47	KEN	MAR	38.40	-13.47	1.09	0.00	0.00	0.14
Q47	IND	SVK	35.42	-10.99	1.03	0.00	0.00	0.60
Q47	LBY	VEN	35.37	-11.91	0.92	0.00	0.00	0.13
Q47	MAR	URY	33.07	11.16	1.12	0.00	0.00	0.06
Q47	NLD	URY	31.23	10.41	0.98	0.00	0.00	0.71
Q47	MAR	VEN	28.50	9.82	1.02	0.00	0.00	0.79
Q47	LBY	URY	27.79	-10.12	1.01	0.00	0.00	0.86
Q47	GBR	IND	24.80	6.26	1.08	0.00	0.00	0.07
Q47	LBY	UZB	23.48	-8.52	0.89	0.00	0.00	0.05
Q47	CZE	SVK	20.57	-6.43	0.95	0.00	0.00	0.42
Q47	MDV	SVK	19.47	-5.49	0.99	0.00	0.00	0.81
Q47	IND	MDV	16.62	-4.70	1.04	0.00	0.00	0.45
Q47	CZE	IND	13.09	4.21	0.93	0.00	0.00	0.16
Q47	UZB	VEN	12.60	-3.37	1.03	0.00	0.00	0.64
Q47	KEN	VEN	11.72	-3.78	1.10	0.00	0.00	0.08
Q48	MAR	MDV	104.93	-14.64	1.10	0.00	0.00	0.12

Q48	MDV	SVK	88.12	12.88	1.04	0.00	0.00	0.52
Q48	CZE	MDV	82.28	-12.50	0.93	0.00	0.00	0.19
Q48	MAR	VEN	73.75	-10.14	0.92	0.00	0.00	0.14
Q48	LBY	MAR	71.96	8.67	1.05	0.00	0.00	0.43
Q48	ARM	KEN	57.97	7.08	1.02	0.00	0.00	0.70
Q48	ARM	IND	54.77	6.00	0.96	0.00	0.00	0.46
Q48	ARM	UZB	36.85	4.14	1.04	0.00	0.00	0.47
Q48	GBR	SVK	34.03	4.05	0.92	0.00	0.00	0.11
Q48	CZE	GBR	26.32	-3.48	1.04	0.00	0.00	0.41
Q48	KEN	UZB	25.79	-2.92	1.02	0.00	0.00	0.73
Q49	SVK	URY	111.29	-16.71	1.11	0.00	0.00	0.10
Q49	LBY	MAR	110.75	13.75	1.02	0.00	0.00	0.74
Q49	CZE	URY	90.34	-13.40	1.03	0.00	0.00	0.62
Q49	KEN	UZB	88.88	-10.23	1.09	0.00	0.00	0.14
Q49	GBR	URY	84.31	-11.49	0.99	0.00	0.00	0.78
Q49	LBY	MDV	56.47	6.69	1.02	0.00	0.00	0.73
Q49	MAR	MDV	53.67	-6.65	1.00	0.00	0.00	0.97
Q49	CZE	GBR	33.91	-4.16	1.05	0.00	0.00	0.36
Q49	CZE	SVK	27.45	3.67	0.93	0.00	0.00	0.23
Q50	URY	VEN	157.19	19.61	0.93	0.00	0.00	0.21
Q50	KEN	UZB	147.21	-17.47	1.09	0.00	0.00	0.15
Q50	UZB	VEN	145.50	17.84	1.12	0.00	0.00	0.06
Q50	IND	KEN	143.92	16.32	0.91	0.00	0.00	0.07
Q50	IND	VEN	141.18	16.73	1.10	0.00	0.00	0.08
Q50	LBY	VEN	139.93	16.29	1.01	0.00	0.00	0.91
Q50	GBR	SVK	93.65	11.37	1.03	0.00	0.00	0.53

Q50	MDV	SVK	92.89	11.94	1.01	0.00	0.00	0.82
Q50	SVK	URY	87.12	-10.81	0.91	0.00	0.00	0.11
Q50	LBY	URY	41.70	-3.34	1.09	0.00	0.00	0.18
Q50	MAR	SVK	37.88	4.54	0.89	0.00	0.00	0.05
Q50	ARM	KEN	35.84	3.70	1.06	0.00	0.00	0.33
Q50	CZE	MAR	31.19	3.17	0.96	0.00	0.00	0.50
Q50	GBR	MDV	25.80	-2.82	1.02	0.01	0.00	0.74
Q50	IND	UZB	22.38	-2.58	0.99	0.02	0.01	0.78

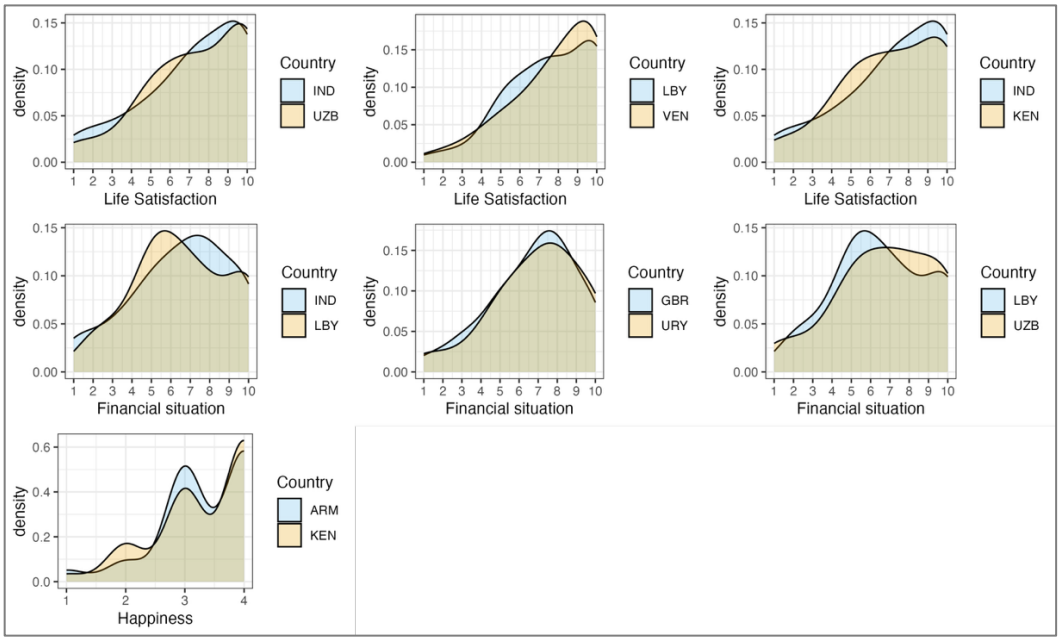
Differences missed by t test only

Q46	NLD	VEN	28.39	-1.19	0.56	0.00	0.23	0.00
Q46	KEN	LBY	23.53	-1.82	1.50	0.00	0.07	0.00
Q46	ARM	LBY	18.92	-1.71	1.41	0.00	0.09	0.00
Q46	GBR	IND	17.39	0.07	0.74	0.00	0.94	0.00
Q46	KEN	URY	15.58	0.62	1.44	0.00	0.54	0.00
Q46	MDV	VEN	15.44	-0.71	0.70	0.00	0.48	0.00
Q46	IND	LBY	15.11	1.70	1.30	0.00	0.09	0.00
Q46	ARM	URY	10.95	0.74	1.35	0.00	0.46	0.00
Q46	MDV	NLD	9.23	0.39	1.24	0.01	0.69	0.00
Q47	IND	URY	19.09	1.84	1.48	0.00	0.07	0.00
Q47	IND	VEN	15.67	0.31	1.34	0.00	0.76	0.00
Q47	MAR	SVK	11.78	-1.96	0.78	0.00	0.05	0.00
Q47	GBR	MDV	9.92	0.49	1.13	0.03	0.63	0.02
Q47	CZE	GBR	9.24	-1.30	0.86	0.03	0.19	0.00
Q48	IND	NLD	137.53	1.57	3.01	0.00	0.12	0.00
Q48	KEN	NLD	128.35	-0.37	2.83	0.00	0.71	0.00

Q48	GBR	IND	91.12	1.54	0.53	0.00	0.12	0.00
Q48	GBR	UZB	71.37	-0.49	0.57	0.00	0.62	0.00
Q48	CZE	IND	71.12	-1.44	0.55	0.00	0.15	0.00
Q48	IND	SVK	69.96	1.92	1.75	0.00	0.06	0.00

Figure 16.

Density Distributions displaying the seven differences detected by the DTS test but missed by the T and F tests



Notes: Each plot depicts the density distributions for two nations on one of the five items that the DTS test (but neither the T test or F test) detected as showing a statistically significantly difference. The three plots in the first row include comparisons for the life satisfaction item (for India and Uzbekistan, Libya and Venezuela, and India and Kenya). The three plots in the second row include comparisons for the financial satisfaction item (for India and Libya, United Kingdom and Uruguay, and Libya and Uzbekistan), and the plot in the third row shows the comparison between Armenia and Kenya on the Happiness item.

Research Question 2: Examining changes within nations

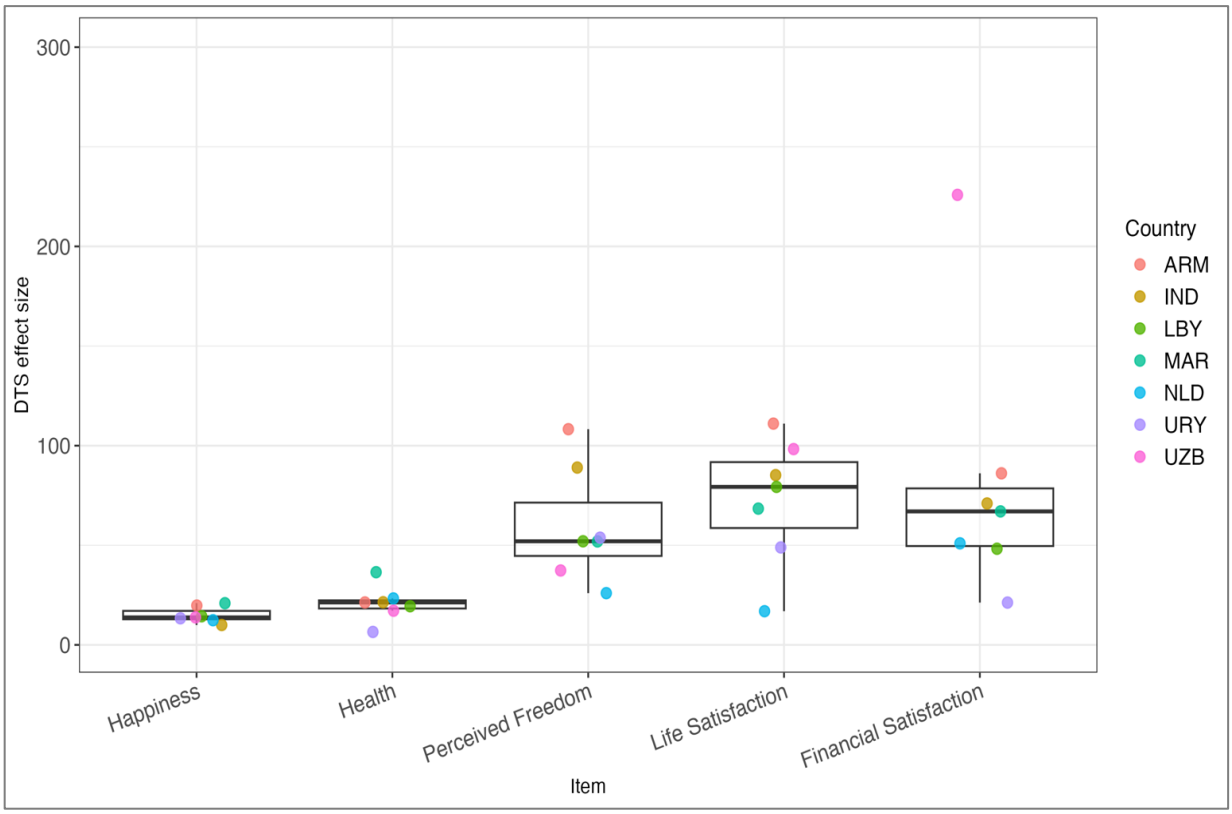
A total of 35 DTS tests were conducted to examine changes in the five aspects of well-being over time for each of the seven nations. All DTS tests were statistically significant ($p < .05$) except for the test between time points for Uruguay on self-reported health (DTS = 6.53, $p = .14$). Significant DTS effect sizes ranged from 9.98 to 225.87 (M=51.19; SD=43.44).

Figure 17 shows DTS effect sizes for the 35 within-nation comparisons, grouped by well-being item. The pattern of within-nation differences over time aligns with the between-nation comparisons. That is, for some items (e.g., Happiness and Health), DTS effect sizes are generally smaller and relatively consistent across nations, whereas for others (e.g., Perceived Freedom and Control, Life Satisfaction, and Financial Satisfaction), effect sizes are larger and more variable across nations. This pattern suggests that certain aspects of well-being are comparatively stable over time within nations, whereas others exhibit greater temporal variability.

Although most within-nation effect sizes were smaller than those observed for between-nation comparisons, some were notably large. For example, particularly strong effects were observed for Uzbekistan on financial satisfaction (DTS = 225.97, $p < .001$), and for Armenia on overall life satisfaction (DTS = 111.06, $p < .001$) and perceived freedom and control (DTS = 108.26, $p < .001$). Figure 18 shows corresponding changes in mean levels for each nation across items from Wave 6 to Wave 7.

Figure 17.

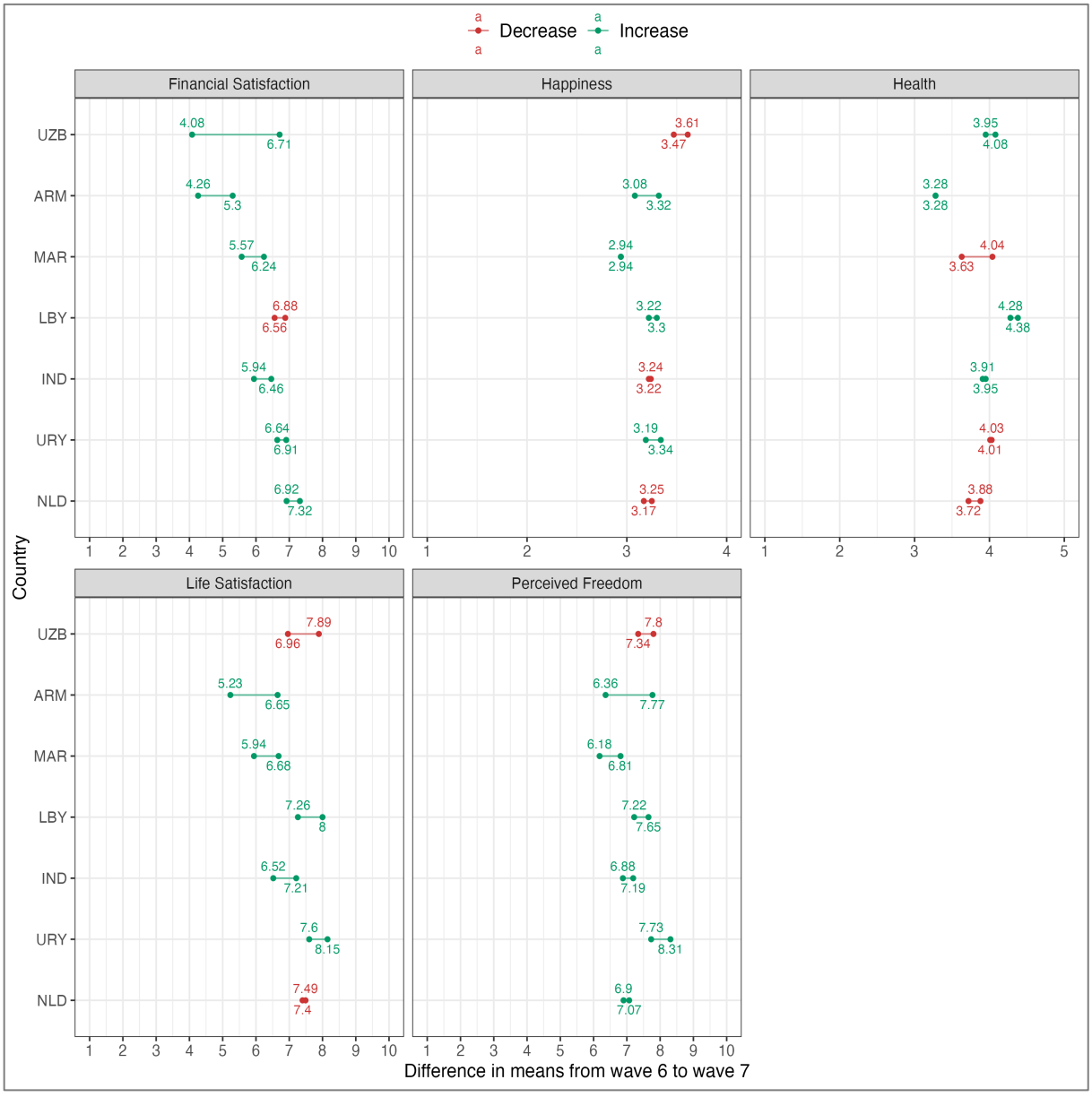
DTS Effect Sizes Across Waves by Nation and Well-Being Item



Note: Each point represents a DTS effect size for a given nation (indicated by color), comparing responses to an item between Wave 6 and Wave 7. Effect sizes are grouped by well-being item, with each boxplot representing the distribution of DTS effect sizes across nations for that item. All effect sizes are statistically significant ($p < .05$), with the exception of the Uruguay comparison for self-reported health. ARM = Armenia, IND = India, LBY = Libya, MAR = Morocco, NLD = Netherlands, URY = Uruguay, and UZB = Uzbekistan.

Figure 18.

Changes in means for each nation from wave 6 to wave 7 across items.



Note: Each point in this figure represents the mean of responses for each nation—shown on the x axis—during Wave 6 or Wave 7. Results for each of the five items are shown in different panels of the plot. Changes shown in green reflect an increase in the mean from wave 6 to wave 7. Changes shown in red reflect a decrease in the mean from wave 6 to wave 7.

Cross national differences detected by DTS that traditional approaches missed

Findings from the *t*-tests and *F*-tests revealed that 13 of the 35 total comparisons were detected as statistically significant by the DTS test but missed by one of the more traditional approaches (though none were missed by both). These 13 comparisons are shown in table 3.

Table 3.

All differences detected by DTS but not by either the T-test or F-test.

Item	Item content	Nation	dts_stat	t_stat	f_stat	dts_pval	tpval	fpval
Q46	Happiness	MAR	20.92	0.13	1.84	0.00	0.90	0.00
Q46	Happiness	ARM	18.66	7.00	1.04	0.00	0.00	0.46
Q46	Happiness	UZB	14.28	-6.32	0.92	0.00	0.00	0.11
Q46	Happiness	NLD	13.19	-4.58	0.92	0.00	0.00	0.08
Q46	Happiness	URY	12.47	4.79	1.12	0.00	0.00	0.07
Q46	Happiness	LBY	10.45	1.80	1.28	0.00	0.07	0.00
Q46	Happiness	IND	9.94	-0.95	0.86	0.00	0.34	0.00
Q47	Health	ARM	21.39	-0.40	0.81	0.00	0.69	0.00
Q47	Health	IND	21.38	1.34	0.88	0.00	0.18	0.00
Q47	Health	UZB	15.58	3.80	0.97	0.00	0.00	0.53
Q48	Freedom	ARM	108.72	-12.58	1.06	0.00	0.00	0.36
Q50	Financial	NLD	46.95	-6.59	1.06	0.00	0.00	0.21
Q50	Financial	URY	21.03	-2.52	0.97	0.02	0.01	0.62

Discussion

Findings from the present analyses reveal several cross-national differences, as well as within nation differences over time, on five items assessing distinct but overlapping features of well-being: happiness, health, perceived freedom and control over one's life, overall life satisfaction, and financial security. Several of these results clearly highlight the value of item-level assessment. For example, cross-national comparisons indicated relatively small differences between nations in reported happiness. However, differences between nations on reported overall life satisfaction—a construct that is often defined similarly to happiness—were considerably larger.

In addition, findings from the present work offer further support for the DTS test as a more powerful alternative to more familiar statistical options (i.e., *t*-tests and *F*-tests) for comparing groups, consistent with suggestions by Dowd (2020). This may be especially useful in contexts in which a large number of comparisons are conducted, as a single omnibus test of distributional differences may reduce the likelihood of type I errors from multiple comparisons. Moreover, due to its sensitivity, DTS, and perhaps similar ECDF-based tests, may be particularly valuable in contexts where avoiding false negatives (i.e., failing to detect meaningful group differences) is a greater priority than avoiding false positives (i.e., detecting small differences of limited practical importance). For example, this may be useful in research comparing health outcomes across demographic groups, where subtle but meaningful disparities could be overlooked if only means or variances are examined.

Cross-national findings

The cross-national findings from the present work suggest that distinct—though in some cases closely related—features of well-being elicit differences in responding across nations and cultures. While DTS effect sizes were relatively similar across nation pairs for some of items (i.e., the Happiness and Health items), findings suggested that other constructs—particularly Overall life satisfaction and Financial security—showed substantially greater differences across nations. Taken together, these findings indicate that even constructs that appear to have a high degree of conceptual overlap (e.g., Happiness and Life Satisfaction) may capture meaningfully different aspects of well-being across cultural contexts.

One possible explanation for the stronger cross-national differences is that some items are more closely tied to cultural norms and structural conditions (e.g., political, economic, and social factors). For example, Financial Satisfaction is likely shaped by factors such as economic inequality, employment opportunities, and cost of living, all of which vary widely across nations. Similarly, the stronger cross-national differences identified for Overall Life Satisfaction align with prior work suggesting that broad evaluations of one's life—often assessed using similar constructs such as Subjective Well-Being—are associated with variables such as income and economic opportunity (Diener et al., 2013). By contrast, Happiness is often defined more narrowly in terms of positive and negative emotions and sometimes viewed as a relatively basic feature of temperament (Caprara et al., 2017; Lyubomirsky et al., 2005). Consistent with this view, happiness is often found to be less strongly linked to more external structural factors (as reflected in perspectives such as the hedonic treadmill model [Brickman & Campbell, 1971] and findings that most people report above-neutral levels of happiness across diverse cultural contexts [Diener & Diener, 1996; Biswas-Diener et al., 2005]).

Within nation findings

Several within-nation findings reported in the present work align closely with political and economic changes seen in the nations represented in these analyses. For example, Participants from Uzbekistan reported considerably higher satisfaction with their financial situation from Wave 6 to Wave 7. This could plausibly reflect the election of a new president in 2016, which triggered wide ranging economic and social reforms leading to decreased poverty, higher and more stable incomes, more employment opportunities, and greater access to housing (World Bank, 2024; International Monetary Fund, 2024). Similarly, participants from Armenia reported marked increases in overall life satisfaction, satisfaction with one's financial situation, and perceived freedom and control over their lives between waves 6 and 7—which may reflect anti-corruption reforms, wage growth, expanding employment opportunities, increased living standards, and other changes during this period (Makaryan et al., 2024; Marie, 2025).

However, these findings should be interpreted cautiously given the considerable time gap between data collection waves, which spans a period of roughly 10 years and includes the COVID-19 pandemic. The pandemic had widespread and wide-ranging effects on well-being, including disruptions to economic and political conditions, physical and mental health, perceived autonomy, and overall life satisfaction. As such, within-nation changes identified in the present work may reflect not only nation-specific political and economic developments, but also broader global impacts and their consequences downstream. Because the present analyses rely on only two cross-sectional time points and do not include measures that could help begin to disentangle these influences, it is not possible to isolate pandemic-related influences from other structural or policy-related changes within nations.

Although all within-nation differences identified by the DTS test were also detected by at least one of the two more familiar statistical approaches, 13 of the 35 comparisons were missed

by either the *t*-test or the *F*-test. This further demonstrates the utility of the DTS statistic, which appears to identify subtle differences in distributions using a single test—this could, in some cases, make DTS a preferable alternative to approaches designed to identify only specific types of group differences (i.e., means or variances).

Limitations and future directions

There are several limitations to consider regarding interpretation of the present findings and evaluating the broader use of the DTS test. First, although the DTS test appears to offer a more powerful method for detecting group differences, the DTS effect size is not a standardized metric. As a result, it lacks the intuitive interpretability of more familiar standardized effect sizes (e.g., Cohen's d), and conclusions regarding the magnitude of effect sizes must be made based on relative comparisons within a given sample rather than absolute benchmarks.

A limitation DTS shares with the more familiar statistical approaches used in the present analyses is the lack of direct insight regarding underlying causes of differences identified between groups. For instance, the differences found in the present work could reflect a variety of factors, including psychometric challenges (e.g., phrasing interpretation across cultures, response styles), genuine cultural or national differences in these aspects of well-being, or some combination of both. Additional, more qualitative or mixed-method work such as cognitive interviewing or cross-cultural focus groups would help to disentangle these influences on responding. A related limitation pertains specifically to the within-nation findings. For each nation, many changes likely occurred across the substantial gap of roughly 10 years between Waves 6 and 7, including both direct and indirect effects of the COVID-19 pandemic. Any of these changes may have influenced features of well-being examined in the present work. Accordingly, these findings should be interpreted as descriptive and exploratory, and viewed as a

more of a potential starting point for future research aimed at identifying the underlying explanations for any within-nation changes observed over time.

Finally, the present study was designed primarily to describe and demonstrate the DTS approach using a limited set of five items assessing features of well-being. Although the relative statistical power of the DTS test was demonstrated, the extent to which DTS captured substantively meaningful cross-national differences—rather than minor distributional differences with limited importance—that more familiar approaches missed remains unclear. Further work is needed to determine whether there are domains of individual differences (e.g., other personality traits, values, behavioral tendencies, etc.) for which DTS would offer an especially meaningful advantage.

Conclusions

The present study demonstrates the value of adopting an item-level approach to examining differences in various aspects of well-being both cross-nationally and within nations over time. In addition, application of the DTS test suggests that meaningful differences between groups may exist beyond the mean- and variance-level differences captured by more familiar statistical approaches. Overall, this work contributes to both the growing literature emphasizing the value of more item-focused research and highlights the importance of expanding the repertoire of statistical approaches that are well-suited for item-level analysis.

III. Applying DTS to IPIP items

From Mather, K., Weston, S. J., & Condon, D. M. (2025). Exploring Generalizability from the Bottom up: Applying DTS to demonstrate an item-level approach to personality assessment. *Current Research in Ecological and Social Psychology*, 100256

In the present work, I apply the DTS statistic to examine cross-national differences in an expanded pool of 2,623 personality items from the International Personality Item Pool (IPIP). First, I discuss the IPIP item pool, and how a stronger item-level focus could be useful in cross-cultural personality assessment—particularly given the generalizability challenges involved when measuring personality constructs at higher levels of the trait hierarchy (i.e., facets and domains)[11]. Next, I apply the DTS statistic to conduct the cross-national comparisons. To conclude, I discuss how the item-level findings from these statistical tests could contribute to both substantive questions in cross-cultural personality research and the development of more generalizable personality scales.

Description and origins of the International Personality Item Pool

The International Personality Item Pool (IPIP; Goldberg, 1999; Goldberg et al., 2006) is a public-domain resource for personality items. This resource was developed to address several barriers to personality research, particularly the proprietary and copyrighted status of the most common and well-established personality inventories (e.g., the NEO-PI-R; Costa & McCrae, 1992), which can discourage the use of larger samples due to higher costs and limit researchers' access to, and ability to revise and improve, these measures. The IPIP represents a freely

[11] There is stronger evidence for generalizability at the "Meta-trait" level, for models such as the Big Two (Thalmayer et al., 2024; Saucier et al., 2014) and the Big Three (De Raad & Peabody, 2005; Peabody, 1987; Cutler & Condon, 2023).

available alternative for researchers and others interested in administering personality items and includes a comprehensive set of items assessing a diverse range of individual differences.

The 3,000-plus items currently available through the IPIP were developed from an initial pool of 909 items from the AB5C model (Hofstee et al., 1992), 1,557 personality-descriptive terms generated by de Raad et al. (1988), and 266 items written to assess the "Intellect" factor identified in English and German lexical studies. These items were further examined and refined using a sample of Oregon homeowners now known as the Eugene–Springfield Community Sample (Hendricks, 1997). In addition to addressing proprietary constraints, the IPIP is useful because its items can be easily organized into broader scales assessing constructs at multiple levels of the trait hierarchy, such as domains, facets, aspects, and meta-traits. In addition, many items assess individual differences that fall outside of the most common models of personality (e.g., the Big Five or HEXACO models) and focus on interests (e.g., artistic interests), more clinically relevant attributes (e.g., grandiosity), and other characteristics that are less frequently assessed in personality research.

The accessibility and versatility of the IPIP have led to its widespread use in studying topics both within and beyond personality psychology—as reflected in the nearly 5,000 citations of the IPIP tracked by google scholar. IPIP items and scales have been translated into more than 25 languages (Goldberg et al., 2006) and used to study topics such as job and academic performance (Lounsbury et al., 2003; Avey et al., 2010), online behavior (Kosinski et al., 2013), leadership (Liu et al., 2023), morality (Curry et al., 2019), health behaviors (Willroth et al., 2023), and well-being (Miller, 2009; Kuper et al., 2023). The IPIP has also been used in several studies focusing on cross-cultural comparison (e.g., Vinson et al., 2020; Klar et al. 2022; Rocha, 2025).

The Present Study

Measuring personality consistently across cultures remains a challenge, partially because constructs at higher levels of the trait hierarchy (e.g., facets and domains) may vary in meaning, salience, and structure across cultural contexts. One way to address this challenge is to examine personality at the item level, where individual differences can be evaluated without relying on assumptions about the invariance of broader constructs. An item-level approach may be especially useful for identifying patterns of similarities and differences that are obscured when items are aggregated into broader scales, and for evaluating the cross-cultural generalizability of specific personality attributes. Building on this perspective, the present study applies the DTS test to a large and diverse pool of personality items to examine cross-national differences in response patterns without imposing a predefined trait structure

In the present study, I apply the DTS test to explore invariance in response patterns across 13 nations using 2,623 IPIP items. These items represent a diverse collection of personality measures (see https://ipip.ori.org/ for the complete list), with some designed to assess the Big Five personality traits and their facets, while others measure additional personality constructs beyond the Five-Factor Model or that fall interstitially between constructs. The current study is exploratory in nature and is not designed to test specific hypotheses. However, certain groups of items are expected to elicit more differences in responding across cultures/nations than others. For example, I anticipate greater differences across nations in items used to assess Openness to Experience based on previous findings suggesting that Openness is not a particularly salient individual difference in many nations (Schwaba & Thalmayer, 2024). The present work aims to highlight potentially meaningful similarities and differences in personality expression across nations but does not attempt to determine the sources of the observed differences. Further

research will be required to determine whether any cross-national differences in responding primarily reflect true cultural variations (e.g., norms around behavior or emotional expression) or psychometric challenges (e.g., item phrasing, concept familiarity, or response styles).

Method

Dataset and participants

Data were collected from the SAPA Project, an online data collection platform where individuals respond to personality items in exchange for information about their personalities based on the SAPA Personality Inventory (SPI; Condon, 2018). The data used here were collected between February 2017 and January 2022. All SAPA Project participants begin by optionally reporting demographic information and then are administered a 250-item subset of personality questions drawn from a large pool of items including those in the SPI and the remaining items in the IPIP. Responding was optional for all questions, but participants were informed that responding to at least 25 questions was necessary to receive feedback on their personality and that responding to more questions would result in more accurate feedback.

Participants were excluded from analyses if they reported being under the age of 18, if they did not report speaking English either “well” or “very well/fluent”, or if the nation they resided in did not have a minimum of 50 administrations across all 2,623 IPIP items. The final sample included 1,025,954 participants (67.13 % Female) who ranged in age from 18 to 90 years old (M = 28.82, SD = 12.61). All participants resided in one of the following thirteen nations which had at least 50 administrations per IPIP item: India (n = 212,931), Pakistan (n = 123,099), the United States (n = 330,121), Nigeria (n = 90,340), the Philippines (n= 93,224), South Africa

(n = 43,669), Bangladesh (n = 43,312), Brazil (n = 12,718), the United Kingdom (n = 25,032), Zimbabwe (n = 11,397), Argentina (n = 8,867), Mexico (n = 10,109), and Egypt (n = 21,135).

Measures

Personality was assessed using 2,623 open access items from the IPIP website. These items assess many aspects of personality and are often aggregated into broader scales to measure traits such as the Big Five domains and facets (Goldberg et al., 2006) and other personality constructs that are located interstitially between Big Five traits or that fall outside of the Big Five trait space (e.g., locus of control, resourcefulness, leadership). See the IPIP website for a comprehensive list of constructs that can be measured using items from this project. However, findings from the present work will be reported at the item-level.

Analyses

I applied the DTS test to examine differences in empirical cumulative distribution functions across all possible pairs of the 13 nations for each of the 2,623 IPIP items, resulting in a total of 204,594 two-sample tests. In addition to applying the DTS test, I conducted t-tests and F-tests and compared the relative numbers of statistically significant comparisons identified by each approach. These analyses were conducted to further examine whether the DTS statistic can identify subtle but meaningful differences between groups that one or both of these more familiar approaches might otherwise miss.

Results

Results by item

Overall, 56% of the 204,594 DTS tests were statistically significant, suggesting the distribution of responses to personality items varies considerably across nations. Findings can be

summarized in terms of either items or nations/nation-pairs (i.e., those with the most and the fewest significant DTS tests). I begin by reporting on items for which a relatively high or low proportion of the statistical tests of that item were significant. The distribution of the proportion of significant tests across items was roughly normally distributed (ranging from 0.08 to 1.00, with an average of 0.56; see Figure 19). However, there were more items than expected for which almost all tests across nation pairs were statistically significant (37 items, compared to the expected 1-2; see Table 4). These items ask about a variety of personality attributes but many related to impulsivity/sensation-seeking (e.g., *Jump into things without thinking*, *Rarely overindulge*, *Take risks*) or features of openness to experience (e.g., *Am full of ideas*, *Prefer variety to routine*, *Think quickly).* Many additional openness-related items beyond those shown in Table 4 ranked among those with the highest proportion of significant tests across nation pairs (i.e., more than 80-90% significant tests).

Notably, the DTS-derived item metrics—proportion of significant DTS tests and mean DTS effect size—were positively correlated with item standard deviation (which ranged from .66 to 2.19; $M = 1.51$, $SD = 0.19$). Items with greater response dispersion tended to produce a greater proportion of statistically significant DTS tests ($r = .26$), and somewhat larger average DTS effect sizes ($r = .14$), suggesting that DTS test outcomes are not entirely independent of item variance. Items with greater response distributions may provide more opportunity for ECDFs to diverge across nations, increasing sensitivity to cross-national differences partially for statistical reasons. Accordingly, items that show particularly high levels of cross-national differences likely reflect a blend of substantive differences in responding and, to some degree, the greater response variability inherent to those items.

Figure 19.

Distribution of the proportions of significant DTS tests across IPIP items.

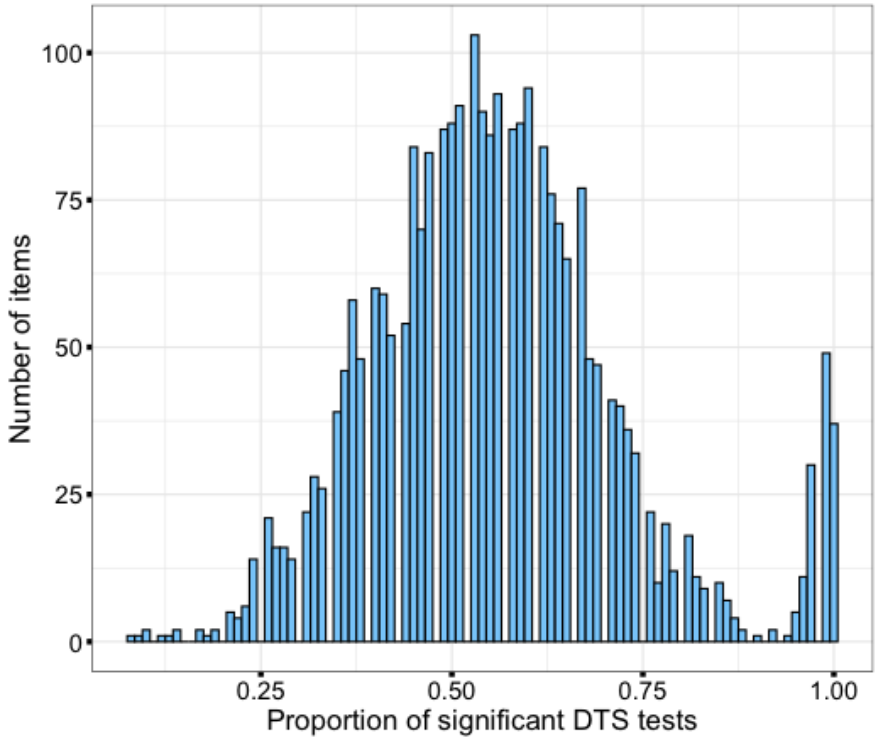


Table 4.

Items with 100% significant tests across nation pairs ($p < .05$), arranged in descending order of overall standard deviation ("Overall SD").

Item	Overall mean	Overall SD
Believe in one true religion.	3.37	2.19
Don't consider myself religious.	3.58	1.92

Item	Overall mean	Overall SD
Like to stand during the national anthem.	4.38	1.77
Often forget to put things back in their proper place.	3.47	1.76
Never spend more than I can afford.	4.01	1.72
Panic easily.	3.76	1.72
Dislike myself.	3.05	1.72
Tend to vote for conservative political candidates.	2.61	1.70
Get angry easily.	3.60	1.70
Think that my moods don't change more than most people's do.	3.56	1.65
Love dangerous situations.	2.79	1.64
Put on a show to impress people.	2.97	1.64
Am not bothered by disorder.	3.20	1.63
Jump into things without thinking.	3.15	1.61
Seek danger.	2.55	1.56
Rarely get irritated.	2.96	1.54
Start tasks right away.	3.63	1.54
Have a slow pace to my life.	3.81	1.53
Like to be thought of as a normal kind of person.	4.16	1.52
Prefer variety to routine.	4.05	1.51
Trust people to mainly tell the truth.	4.12	1.51
Am skilled in handling social situations.	3.93	1.50
Often make decisions on the spur of the moment.	3.87	1.50
Rarely overindulge.	3.44	1.49
Feel comfortable with myself.	4.43	1.49
Never splurge.	3.17	1.48
Take risks.	3.86	1.47
Trust what people say.	3.72	1.46

Item	Overall mean	Overall SD
Like order.	4.41	1.44
Use others for my own ends.	2.32	1.43
Have a natural talent for influencing people.	4.00	1.42
Believe that people are basically moral.	3.86	1.42
Like to take it easy.	4.43	1.36
Let things proceed at their own pace.	4.17	1.32
Think quickly.	4.50	1.31
Am full of ideas.	4.64	1.29
Am able to come up with new and different ideas.	4.64	1.23

Another subset of items had a notably small proportion of significant tests across nation pairs. Table 5 displays the 13 items with fewer than 20% significant tests. Several of these items ask about more specific behaviors and life experiences—which may be thought of as characteristic adaptations—that have relatively low variability both between and within nations (e.g., *Know how to fly a helicopter*, *Have lived in all 50 states in the U.S.A., Think a lot before complaining in a restaurant if my meal is not well-prepared).* However, similar to the set of items with the most significant tests, several of these items relate to various features of openness to experience (e.g., *Believe there are many sides to most issues, Want to understand things, Enjoy the beauty of nature*). Several other openness-related items were represented among the items with relatively few significant tests and can be viewed in the supplemental file along with findings for all 2,623 IPIP items regarding the proportion of significant tests, mean DTS effect size, and overall item means and variances.

Table 5.

Items with the fewest significant tests across nation pairs, arranged in descending order of proportion of significant DTS tests.

Item	Proportion < .05	Mean	SD
Want to know the reasons why.	0.19	5.43	0.93
Think a lot before complaining in a restaurant if my meal is not well-prepared.	0.19	4.65	1.58
Look at the facts.	0.18	5.09	1.02
Have lived in all 50 states in the U.S.A.	0.17	1.17	0.67
Love to chat.	0.17	4.10	1.54
Don't have a soft side.	0.14	1.90	1.24
Am able to see how other peoples' emotions can affect their behaviors.	0.14	5.03	1.13
Am committed to principles of justice and equality.	0.13	5.06	1.13
Have little patience for slackers.	0.12	4.32	1.39
Know how to fly a helicopter.	0.10	1.16	0.69
Enjoy the beauty of nature.	0.10	5.45	0.94
Want to understand things.	0.09	5.56	0.76
Believe that there are many sides to most issues.	0.08	5.25	0.99

Results by nation pair and individual nation

Nation pairs varied considerably in the percentage of significant tests across the 2,623 items (M = 55.8%, SD = 21.2%), ranging from 16.4% (Argentina and Mexico) to 98.6% (United States and Pakistan). Nation pairs with the most similar response patterns (i.e., 30% or fewer

significant tests) were typically geographically proximate countries, such as Latin American pairings among Argentina, Mexico, and Brazil, as well as African nation pairs (e.g., Nigeria and Zimbabwe, South Africa and Zimbabwe). Egypt also showed high similarity with several nations including Mexico, India, Argentina, Bangladesh, and Pakistan. Conversely, nation pairs with the most different response patterns (highest proportion of significant tests) frequently involved either the United States or the Philippines paired with other countries.

The proportion of significant tests for each of the 13 individual nations is shown in Table 6. Figure 20 visualizes the overall response similarity between nations for each of the 78 possible pairs. Similarity scores were calculated as 1 minus the proportion of significant DTS tests (e.g., if 19% of tests between Argentina and Brazil were significant, their similarity score would be 1 - 0.19 = 0.81, indicating high similarity in response patterns). Figure 21 displays the distribution of DTS effect sizes for each of the 78 nation pairs, highlighting outliers for each nation pair that represent items with particularly large differences in response between the two groups.

Table 6.

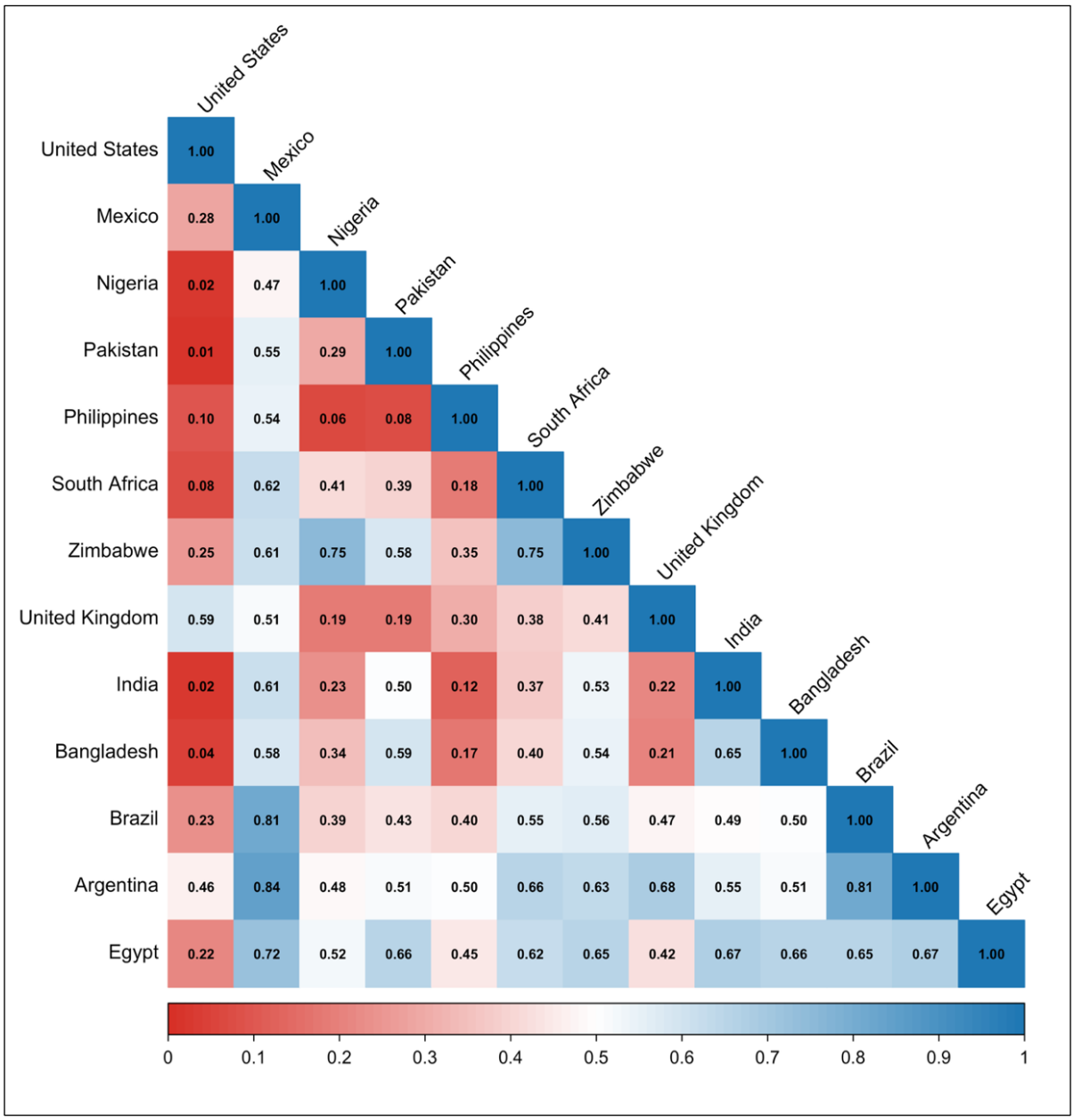
Proportion of significant tests per nation (across nation pairs and items)

Nation	Proportion of significant tests ($p < .05$)
Argentina	0.39
Mexico	0.41
Egypt	0.42
Zimbabwe	0.45
Brazil	0.48
South Africa	0.55

Bangladesh	0.57
India	0.59
Pakistan	0.60
United Kingdom	0.62
Nigeria	0.65
Philippines	0.72
United States	0.81

Figure 20.

Overall similarity in responding across nation pairs.

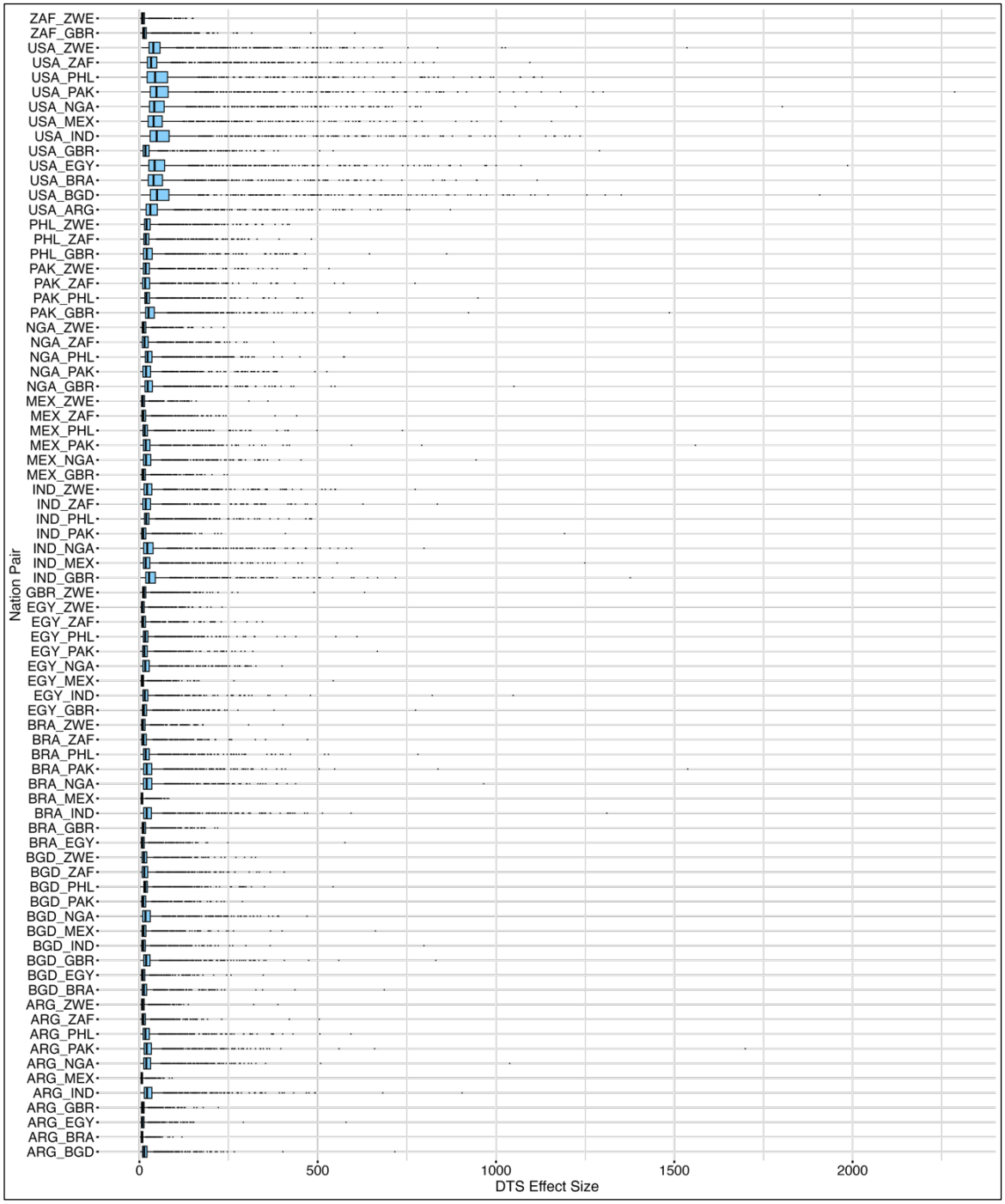


Note: This heatmap displays the similarity in responding across all IPIP items for each nation pair. Each cell represents the degree of similarity between nations, calculated as 1 minus the proportion of statistically significant DTS tests between those nation pairs. Values closer to 1 (i.e., more than .5), shown

in blue, indicate greater similarity, and values closer to 0 (less than .5), shown in red, indicate less similarity.

Figure 21.

Distribution of DTS effect sizes by nation pair



Note: ISO country codes are used on the y-axis in place of the full names of nations. ZAF = South Africa, ZWE = Zimbabwe, GBR = United Kingdom, USA = United State, PHL = Philippines, PAK = Pakistan, NGA =

Nigeria, MEX = Mexico, IND = India, EGY = Egypt, BRA = Brazil, BGD = Bangladesh, ARG = Argentina. Nation pairs are approximately grouped and arranged by nation.

Comparison of DTS to the T- and F-tests

In addition to using the DTS test for cross-national comparisons, I applied two more familiar statistical approaches—the *t*-test and *F*-test—to compare nation means and variances. While the DTS test detected 56% of the 204,594 comparisons as statistically significant, the *t*-test identified 50%, and the *F*-test identified 30.7%. Only a small number of nation comparisons—2,691 (1.3% of the total)—were identified as statistically significant uniquely by the DTS test (i.e., not by the *t* or *F* test). This small proportional gain suggests that DTS does not greatly increase the number of detected differences beyond more familiar approaches. However, these comparisons detected uniquely by DTS appear to reflect a specific type of cross-national difference—that is, differences in the overall *shape* of the response distributions (e.g., skew or multimodality) that are not captured by tests which focus solely on means and variances.

Most items represented among the comparisons detected only by DTS had three or fewer significant tests (1,507 of the 2,623 items). However, a small subset of items were more consistently identified only by DTS (approximately six to seven times each) as eliciting statistically significant differences between nations. These items and specific nation comparisons are shown in table 7. Inspection of these items suggests that DTS is particularly sensitive to differences in certain interpersonal norms, such as apologizing frequently, disclosing personal thoughts, or beliefs about whether others tend to be honest. Figures 22-25 show the distributions for nation comparisons on several of these items that were identified as statistically significant only by the DTS test.

Table 7.

Items most frequently identified by DTS (but neither the T nor F test).

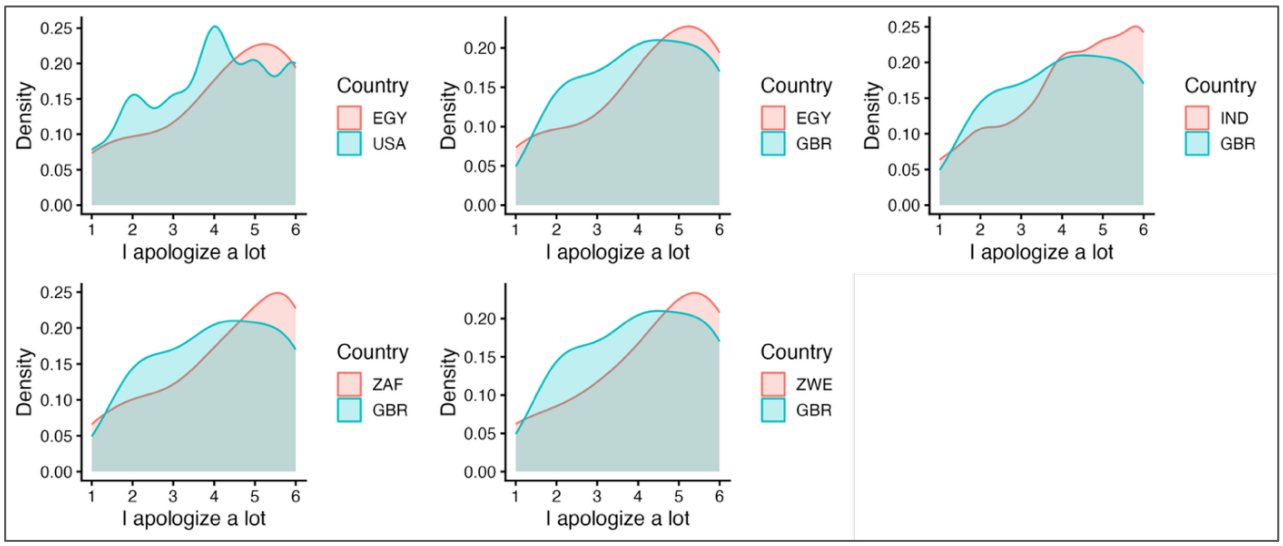
Item	Item_Content	Country1	Country2	DTS_pval	T_pval	F_pval
q_299	Apologize a lot.	BGD	USA	0.017	0.081	0.464
q_299	Apologize a lot.	EGY	USA	0.028	0.064	0.265
q_299	Apologize a lot.	ARG	NGA	0.031	0.063	0.629
q_299	Apologize a lot.	EGY	GBR	0.04	0.418	0.131
q_299	Apologize a lot.	GBR	IND	0.021	0.145	0.345
q_299	Apologize a lot.	GBR	ZAF	0.015	0.106	0.155
q_299	Apologize a lot.	GBR	ZWE	0.011	0.164	0.202
q_373	Believe that most people would lie to get ahead.	ARG	MEX	0.029	0.15	0.166
q_373	Believe that most people would lie to get ahead.	BGD	BRA	0.028	0.139	0.227
q_373	Believe that most people would lie to get ahead.	BRA	PAK	0.04	0.147	0.406
q_373	Believe that most people would lie to get ahead.	BRA	ZAF	0.044	0.145	0.077
q_373	Believe that most people would lie to get ahead.	EGY	ZAF	0.047	0.09	0.116
q_373	Believe that most people would lie to get ahead.	PHL	ZWE	0.048	0.238	0.059
q_561	Disclose my intimate thoughts.	BGD	EGY	0.032	0.161	0.127
q_561	Disclose my intimate thoughts.	BRA	PHL	0.036	0.347	0.087
q_561	Disclose my intimate thoughts.	EGY	PAK	0.049	0.105	0.363
q_561	Disclose my intimate thoughts.	MEX	NGA	0.048	0.054	0.739
q_561	Disclose my intimate thoughts.	GBR	PAK	0.038	0.095	0.051
q_561	Disclose my intimate thoughts.	PHL	ZAF	0.044	0.124	0.171
q_991	Get to the point quickly.	MEX	USA	0.009	0.321	0.065
q_991	Get to the point quickly.	EGY	PHL	0.032	0.071	0.193
q_991	Get to the point quickly.	IND	ZWE	0.034	0.078	0.071
q_991	Get to the point quickly.	GBR	MEX	0.006	0.084	0.12
q_991	Get to the point quickly.	NGA	ZWE	0.046	0.072	0.126
q_991	Get to the point quickly.	GBR	ZWE	0.024	0.26	0.083
q_1049	Have a quiet laugh.	ARG	BRA	0.024	0.084	0.257
q_1049	Have a quiet laugh.	ARG	GBR	0.02	0.072	0.134
q_1049	Have a quiet laugh.	IND	NGA	0.029	0.079	0.238
q_1049	Have a quiet laugh.	IND	ZAF	0.042	0.138	0.103
q_1049	Have a quiet laugh.	GBR	IND	0.043	0.135	0.071
q_1049	Have a quiet laugh.	GBR	MEX	0.044	0.119	0.146

q_1123	Have no secrets.	BGD	BRA	0.035	0.053	0.698
q_1123	Have no secrets.	BGD	MEX	0.045	0.135	0.114
q_1123	Have no secrets.	IND	ZAF	0.025	0.097	0.636
q_1123	Have no secrets.	GBR	IND	0.021	0.17	0.086
q_1123	Have no secrets.	ZAF	ZWE	0.045	0.065	0.848
q_1123	Have no secrets.	GBR	ZWE	0.016	0.081	0.337
q_1798	Talk even when I know I shouldn't.	ARG	USA	0.034	0.058	0.204
q_1798	Talk even when I know I shouldn't.	BGD	BRA	0.036	0.509	0.054
q_1798	Talk even when I know I shouldn't.	BGD	MEX	0.034	0.073	0.236
q_1798	Talk even when I know I shouldn't.	BRA	GBR	0.028	0.066	0.06
q_1798	Talk even when I know I shouldn't.	BRA	ZWE	0.044	0.112	0.303
q_1798	Talk even when I know I shouldn't.	MEX	PAK	0.048	0.205	0.131
q_1984	Work too much.	BRA	USA	0.034	0.28	0.055
q_1984	Work too much.	ARG	ZWE	0.042	0.05	0.923
q_1984	Work too much.	BGD	BRA	0.044	0.084	0.201
q_1984	Work too much.	BRA	EGY	0.033	0.075	0.132
q_1984	Work too much.	BRA	NGA	0.009	0.078	0.071
q_1984	Work too much.	MEX	ZAF	0.048	0.054	0.541

Note: The 8 items represented in this table were most often detected by DTS as eliciting statistically significant differences between nations, and missed by the T or the F test.

Figure 22.

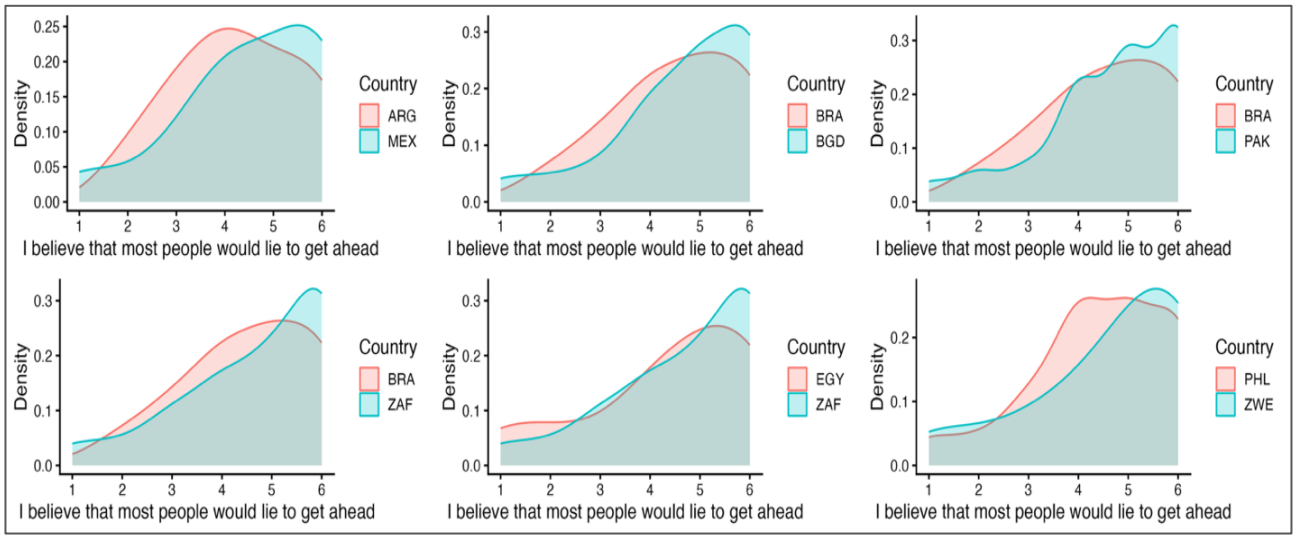
Five nation comparisons for the item "I apologize a lot" that only DTS identified as statistically significant



Note: Each of these five plots depicts the density distributions for two nations on the item "I apologize a lot". The distributions for *each of these nation pairs were identified as significantly different by the DTS test but not by the T test or the F test. EGY= Egypt, USA = United States of America, GBR = United Kingdom, IND = India, ZAF = South Africa, and ZWE = Zimbabwe. Response options range from 1 (Very Inaccurate) to 6 (Very Accurate).*

Figure 23.

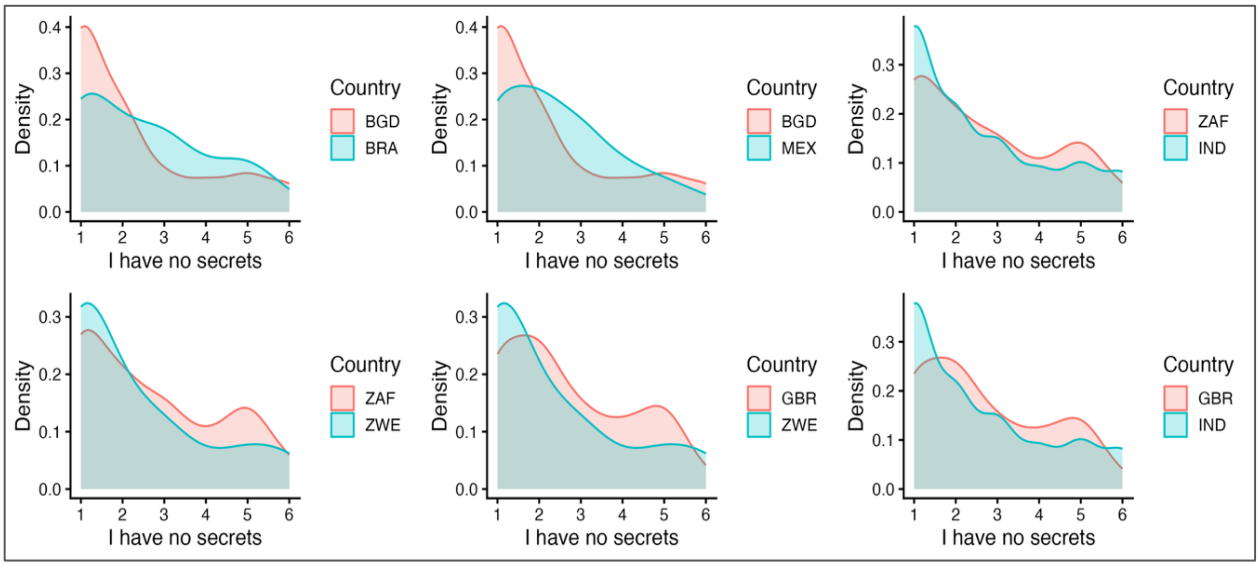
Six nation comparisons for the item "I believe that most people would lie to get ahead" that only DTS identified as statistically significant.



Note: Each of these five plots depicts the density distributions for two nations on the item "I believe that most people would lie to get ahead". The distributions for each of these nation pairs were identified as significantly different by the DTS test but not by the T test or the F test. ARG= Argentina, MEX = Mexico, BRA = Brazil, BGD = Bangladesh, PAK = Pakistan, ZAF = South Africa, EGY = Egypt, PHL = The Philippines, and ZWE = Zimbabwe.Response options range from 1 (Very Inaccurate) to 6 (Very Accurate).

Figure 24.

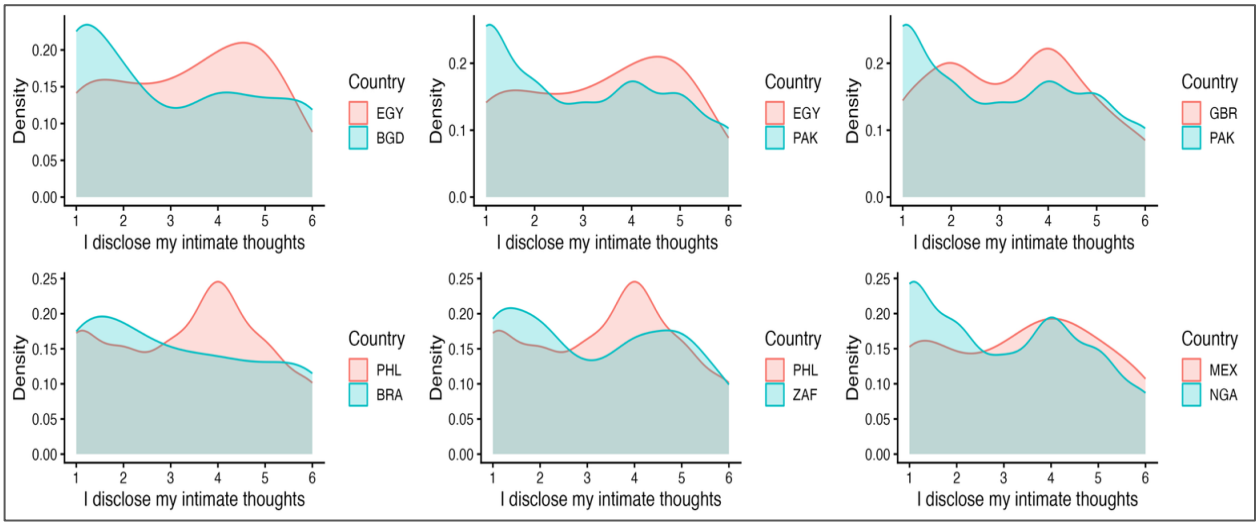
Six nation comparisons for the item "I have no secrets" that only DTS identified as statistically significant



Note: Each of these five plots depicts the density distributions for two nations on the item "I have no secrets". The distributions for each of these nation pairs were identified as significantly different by the DTS test but not by the T test or the F test. BGD = Bangladesh, BRA = Brazil, MEX = Mexico, ZAF = South Africa, IND = India, ZWE = Zimbabwe, and GBR = United Kingdom. Response options range from 1 (Very Inaccurate) to 6 (Very Accurate).

Figure 25.

Six nation comparisons for the item "I disclose my intimate thoughts" that only DTS identified as statistically significant.



Note: Each of these five plots depicts the density distributions for two nations on the item "I disclose my intimate thoughts". The distributions for each of these nation pairs were identified as significantly different by the DTS test but not by the T test or the F test. BGD = Bangladesh, EGY = Egypt, PAK = Pakistan, GBR = United Kingdom, PHL = The Philippines, BRA = Brazil, MEX = Mexico, ZAF = South Africa, and NGA = Nigeria. Response options range from 1 (Very Inaccurate) to 6 (Very Accurate).

Discussion

The current study demonstrates that focusing on the item level can be a valuable approach to addressing questions of invariance, which are more commonly considered at the broader construct level after developing an initial scale. The present findings align with those of previous studies, which generally report greater differences in responding to personality measures between geographically and/or culturally diverse nations. These results are also consistent with prior research on the personality tendencies that show the most (or least)

generalizability across cultural and geographical contexts. For instance, items reflecting features of openness to experience and emotional expression (e.g., *Get angry easily*, *Panic easily*, *Rarely get irritated*, *Think that my moods don't change much more than most people's do*) elicited notable differences across nations, reflecting the influence of cultural norms -- e.g., display rules -- on these traits (Eid & Diener, 2001; Gross & John, 1997). However, the present findings also highlight the importance of evaluating differences between groups at more granular levels. While many openness-related items showed the strongest differences in responding across nations, others demonstrated some of the smallest differences, and several themes emerged among openness items that distinguished those with greater versus fewer cross-national differences.

Openness items with relatively high proportions of significant differences often addressed (a) preferences and beliefs about topics that are highly influenced by local culture, such as politics, religion, and art (e.g., *I believe in one true religion, I tend to vote for conservative political candidates)*; (b) standing out or behaving unconventionally (e.g., *I like to be thought of as a normal kind of person)*; (c) being "quick" to understand or catch on (e.g., *I think quickly)*; and (d) having a vivid imagination or generating new ideas (e.g., *I am full of ideas, I am able to come up with new and different ideas)*. Conversely, openness items with smaller proportions of significant differences tended to focus on (a) appreciation for learning about other cultures; (b) perspective-taking and considering multiple viewpoints (e.g., *I believe that there are many sides to most issues, I can see different points of view)*; (c) enjoyment of learning for its own sake rather than quickness of understanding (e.g., *I want to understand things, I love to learn new things)*; and (d) practical problem-solving or generating ideas for improvement rather than ideation without a specific purpose (e.g., *I see through problems, I have new ideas for improving things)* . These findings support the notion that narrower personality tendencies associated with

broad traits may differ in relevance across contexts, even if the broader traits themselves show limited generalizability (Schwaba & Thalmayer, 2024).

The observed variation in item-level invariance further underscores the limitations of traditional approaches to measurement invariance, which assume item interchangeability (tau equivalence) within latent constructs (i.e., that items have similar relationships to the underlying trait, and can be substituted for one another; Lord & Novick, 1968). This assumption likely contributes to the difficulty in establishing invariance at the domain level (Dong & Dumas, 2020). The present findings also align with those of Achaa-Amankwaa et al. (2021), which demonstrated that item-level models—composed of the most predictive items, regardless of their source measures—outperform both domain- and facet-level models in predicting participants' nationality across 10 countries. Based on a similar set of items from the IPIP-NEO scales (Goldberg, 1999), these results reinforce the notion that individual personality items are meaningful units of analysis. Taken together with the present findings, this suggests that item-level approaches may offer more sensitive detection of cross-cultural personality differences that might be obscured in analyses focused on broader constructs.

Nation pairs demonstrating the most and fewest differences

Regarding cross-national differences, the United States stood out as particularly distinct from many of the other countries included in the analyses. This pattern was likely influenced by the fact that the International Personality Item Pool (IPIP) items were developed using Western—and primarily U.S.—samples. The United States consistently appeared in nation pairings with the highest proportion of significant differences and showed only moderate similarity in responses to culturally similar nations, such as the United Kingdom. This suggests that many items may reflect cultural assumptions, phrasing conventions, or references to specific

activities that are more familiar or accessible to U.S. respondents. Similar concerns have been raised in prior work, such as Church et al. (2011), which identified strong differential item functioning (DIF) between the United States and the Philippines for an extraversion item related to the enjoyment of riding roller coasters—likely due to the lower prevalence of roller coasters and amusement parks in the Philippines. In contrast, nation pairs with greater response similarity tended to cluster geographically, as observed among Latin American nations (Brazil, Argentina, and Mexico) and Southern African nations (Zimbabwe and South Africa).

Differences in dominant religions and the centrality of religion across nations—one of the most divergent cultural attributes (Goodwin et al., 2020)—may also help explain the observed patterns of similarity and difference. Although many IPIP items assess personality traits that vary at least moderately within nations, some target religious or spiritual beliefs, which often show limited variability within certain cultural contexts. Several of these items ask directly about belief (e.g., *I am devoted to religion*, *I believe in God*, *I guide my life using religious scriptures*), and others may refer to behaviors that vary in acceptability across religious denominations and sects. Two items that directly reference religious belief—*I believe in one true religion* and *I don't consider myself religious*—were among those with the highest proportion of significant DTS tests. Additionally, two nations with relatively high levels of religiosity—the Philippines and Nigeria (Finke & Adamczyk, 2014)—were among the most distinct from other countries in the current study, reinforcing prior findings on religion's central role in shaping personality-relevant norms. These results support the idea that both geographic and cultural proximity contribute to patterns of response similarity and difference.

Comparison of DTS to more familiar statistical options

Although the number of cross-national differences uniquely identified by DTS was small, these comparisons suggest that the advantages of the DTS test lie more in its ability to detect qualitative differences in the shape of response distributions that are not captured by tests that focus solely on means or variances (e.g., differences in skew or multimodality). In addition, the IPIP items more likely to be detected only by DTS tend to assess interpersonal beliefs and tendencies, such as apologizing frequently, believing that most people would lie to get ahead, keeping secrets or disclosing one's intimate thoughts, and how often or loudly one expresses oneself. Visual inspection of the density distributions for nation pairs on these items suggest several possible explanations. Some differences appear to be driven by extreme responding or greater skew in one nation than another, while others appear to reflect clustering of responses at different points along the scale.

For example, on the item *I have no secrets*, most nations' distributions were positively skewed, but for some nations (e.g., Bangladesh), the skew was more extreme. This may reflect item phrasing (i.e., use of the term *secrets*, including variation in its cultural meaning and its positive or negative connotations), or genuine cross-cultural differences in norms and expectations regarding when, whether, and with whom secrets are shared. Interestingly, for a conceptually similar item—*I disclose my intimate thoughts*—nation distributions were generally less skewed.

Another item, *I apologize a lot*, also demonstrated potentially meaningful differences across groups. Notably, the United Kingdom was represented in four of the seven comparisons that were identified as statistically significant only by DTS. Moreover, the distribution of responses for participants from the United Kingdom suggested that they were more likely to disagree with the statement *I apologize a lot*. This is somewhat surprising given evidence that

politeness and etiquette are especially valued in the United Kingdom, as well as linguistic studies documenting more frequent apologetic statements in British English (Leech, 2014; House, 2021; Taguchi, 2019; Kádár & Haugh, 2013; Kaloi et al., 2025). One possible explanation is reference group effects—the tendency for individuals reporting on their personality to evaluate themselves relative to those around them (often members of their culture, age group, gender, or other salient social groups; Lenhausen et al., 2023). Accordingly, participants from the United Kingdom may view themselves as less likely to apologize because they are making judgments within a cultural context in which apologies—both casual and formal—are used more frequently than in other cultures. As a result, participants from the United Kingdom who rated themselves lower in frequency of apologizing may nevertheless apologize more often in absolute terms than participants from other nations included in the present study.

Limitations and future directions

There are several limitations of the present study that could be addressed in future work. First, due to the descriptive nature of the present findings, it is unclear how similarities and differences in responding across nations that emerged in the present work should be interpreted. Differences could reflect true cultural variations or arise from psychometric challenges, such as item phrasing or concept familiarity. As such, it is premature to make definitive claims about the invariance or non-invariance of items across groups, or about the underlying explanations for differences in responding. This will be a task for future, more explanation-focused studies. Another limitation is the lack of established standards for evaluating invariance at the item level, in contrast to traditional (often scale-focused) techniques, which have more clearly defined benchmarks (e.g., indices of configural, metric, and scalar invariance). However, the complexity of standards for metrics at the item-level is likely to be much lower than needed for the scale-

level, which could make the item-level an easier starting point for developing generalizable personality measures. Additionally, more generalizable measures of broader traits could be developed based on item-level findings regarding the (non) invariance of items. The use of more invariant items appears likely to lead to more invariant scales, though this remains an open question for future research.

The representativeness of the nations and participants included in this study also warrants consideration. The specific nations sampled could have influenced the results, as the items showing the greatest and fewest differences may depend on the collective similarity or dissimilarity of the nations involved. While this study included culturally diverse nations, the data were clearly not representative of the global population. It is also possible that any differences in sample sizes across nations contributed to the differences observed in effect sizes across nation pairs (though only nations with a minimum of 50 participants per item were included in analyses). In addition, cultural diversity in the present sample was inferred from geographical diversity, which may only indirectly capture cultural differences. Participant characteristics also pose a limitation, as the study included only individuals with high English proficiency and who presumably had an interest in personality. These characteristics reduced the cultural representativeness of the sample to an unknown extent.

A related limitation of this study is the representativeness of the item bank. While the creation of the International Personality Item Pool was a significant contribution to the field and has remained a valuable resource for researchers, its items were developed using Western (and primarily U.S.) samples. This likely contributed to the particularly strong overall difference observed between the United States and the other 12 nations included in this study. A more diverse item bank that is not as skewed toward Western (and perhaps specifically U.S.) ways of

thinking might have resulted in a smaller overall difference between U.S. participants and those from other nations. Additionally, IPIP items mostly assess personality using short, concrete phrases that reference observable behaviors. This approach has several psychometric advantages, as their specificity minimizes interpretational differences across participants (unlike more abstract adjectives) without being too context-specific (since personality traits are generally considered thought of as decontextualized patterns of behavior). However, using items developed based on alternative phrasing approaches (e.g., single abstract adjectives or longer sentence-based items) could provide insight into the effects of wording and other semantic features.

There are also interpretational challenges associated with the findings of the present study, as I did not apply a correction for multiple testing. This omission likely did not compromise the main conclusions in the present work, given that the primary goal was to detect broad patterns in the relative number of statistically significant tests across nations and items—of which there were many. However, the potential presence of false positives may limit the usefulness of these results for some researchers who are interested in making specific comparisons between individual nations on particular items. Further investigation will be needed to assess whether any given DTS comparison (i.e., between two nations on a specific item) is statistically robust.

Another methodological caveat is the finding that metrics such as the proportion of significant DTS tests per item and the average DTS effect size per item were positively associated with item standard deviation. Because DTS is designed to detect differences in distributions, items with broader response distributions may be more likely to yield larger DTS statistics for mechanical, structural reasons, independent of whether they reflect stronger substantive cross-national differences. Although correlations for these DTS metrics with item

standard deviation were somewhat modest, they suggest that item variance contributes to some extent to DTS sensitivity. Accordingly, one direction for future work could be to evaluate DTS effect sizes after adjusting for item variance.

Chapter 3 summary

The present study used an item-level approach to explore measurement invariance across 13 nations, using a large bank of personality items from the International Personality Item Pool. The findings revealed that many personality items do not exhibit the expected level of (non) invariance based on the broader construct they typically measure. For example, while Openness to Experience measures often show poor invariance across diverse nations, some individual Openness items appear to show very strong invariance—which aligns with previous work suggesting that at least some features of Openness likely show stronger universality (Schwaba &Thalmayer, 2025; Fetvadjiev et al., 2015). These results highlight the value of assessing personality at more granular levels within the trait hierarchy. They also suggest that an item-level approach could facilitate development of more generalizable measures of broader traits. Future work will be needed to establish clearer standards for interpreting effect sizes that reflect item-level differences between groups using the DTS statistic and similar ECDF-based tests.

IV. Conclusion & Takeaways

Broadly, the goal of the present work was to further explore the advantages and potential of item-level assessment to advance a variety of research topics within personality psychology. I began with a review of existing findings that demonstrate the value of adopting an item-level focus, discussed explanations for the field's strong focus on broader constructs and scales, and explored the utility of item-level findings across research questions falling under the three primary goals of prediction, description, and explanation. Next, I described and demonstrated a more item-centric alternative to familiar statistical tests—the DTS test—that detects group differences in the shape of response distributions, using data from the World Values Survey. Then, I applied the DTS test to compare nations in an expanded pool of 2,623 items from the International Personality Item Pool—a public domain resource containing a highly comprehensive set of personality items designed to assess a wide range of individual differences at multiple levels of analysis (i.e., from items to "meta traits"). These comparisons identified cross-national differences in the shape of item distributions that would have remained undetected using only more familiar tests of means (*t*-tests) and variances (*F*-tests), particularly for items assessing a variety of interpersonal norms.

Several findings from this work reflect those of previous studies, including the limited generalizability of Openness-related individual differences (Schwaba & Thalmayer, 2024), as well as cross-national patterns of similarity and difference that align with factors such as cultural similarity and geographical proximity. However, the present work also yielded several surprising findings, most notably the presence of some Openness-related constructs among the items showing the fewest differences in responding across nations. These individual differences—which focus on attributes such as intellectual curiosity, perspective taking, and problem

solving—appeared to demonstrate greater universality and relevance across nations and cultural contexts. These findings are consistent with more recent theorizing that Openness/Intellect likely has a "universal core" (Schwaba & Thalmayer, 2025, p. 18), one that may emphasize knowledge and wisdom more than other features commonly associated with this construct in Western contexts, such as non-traditionalism and eccentricity.

These findings could be useful in future scale development efforts aimed at cross-cultural generalizability. In particular, they suggest that Openness measures may be improved by focusing on items that capture these components of the construct that appear to be more universal (i.e., intellectual curiosity, perspective-taking, and problem-solving), while placing less emphasis on aspects that are more culturally variable (e.g., non-traditionalism, religiosity, or eccentricity). Accordingly, an item-level approach could offer a useful starting point for further establishing which components of Openness and other broad trait constructs are more likely to generalize across cultural contexts.

Remaining challenges and future directions for item-focused research

While the present work was able to capture potentially meaningful relationships at the item-level, interpretation of findings at the item level remains a challenge. The increased parsimony of frameworks of broader constructs like the Big Two, Three, Five, or Six makes them a more tractable starting point for interpretation and enables researchers to explore findings more thoroughly (i.e., by focusing on only a few broad constructs). A focus on individual items, in contrast, and particularly when administering large sets of items, requires more adjudication regarding the most central or potentially important findings. For instance, the present work involved 204,594 tests between nation pairs across IPIP items, and many of the similarities and

differences between nations that were not presented in the results of this manuscript could nonetheless be useful to those focusing on more specific populations or individual differences

In some contexts, such as predictive modeling, where explanation is not the primary goal, the variety of item-level findings may pose less of a challenge. However, researchers focused on description- or explanation-related questions—including those regarding invariance across cultural contexts—are likely engaged in more theory-building efforts and interpretation of relationships between variables. One strategy for addressing this challenge could be to develop a publicly available central repository for descriptive findings in personality research (as proposed by Mõttus et al. [2020]). Similar repositories are more common in psychology-adjacent fields, like neuroscience and genetics, and could facilitate hypothesis-generation, theory-building, and collaboration among personality researchers based on item-level findings. Since many findings between nation pairs and personality variables could not be explored in depth in the present work, but might be of interest to other researchers, a dataset containing all DTS test results will be made available on the Open Science Framework. This dataset includes item-level DTS statistics, *p*-values, descriptive statistics, and other metrics for all nation-pair comparisons, providing a foundation that other researchers can build on in future work.

In future work, it could also be useful to report findings at both the item-level and one or more broader levels of the trait hierarchy (e.g., the domain level). This could be a stronger approach for several reasons, which include providing a more tractable starting point for interpreting findings. Generally, it is recommended that descriptive findings like those reported in the present work be reported at the nuance- or item-level because lower-dimensional structures like the big five can be derived from higher-dimensional structures (but the reverse is not possible; Condon & Mroczek, 2016; Saucier & Iurino, 2019). However, reporting findings at the

level of broader constructs in addition to the item/nuance-level could be useful for situating findings within the broader literature—especially considering that the majority of past studies focus on broader constructs such as the big five.

Discussing findings in terms of broader constructs might also be necessary, to some extent, considering that constructs in psychology are "fuzzy" (Ziegler et al., 2015) and often discussed using relatively broad or imprecise terms (e.g., "success", "happiness") and interpersonal descriptors (e.g., "an introvert", "funny", "organized"). While more precise measurement of personality constructs may be important for advancing scientific understanding of many topics, discussing research questions and findings only in terms of the more granular, specific tendencies may be challenging, and create more of a disconnect between the way personality traits are defined and measured scientifically, and the way they are discussed colloquially, or in more practical settings. Thus, reporting on item-level and broader-level constructs could be viewed as serving complementary purposes, with item- or nuance-level findings being used to clarify whether any narrower, more specific personality tendencies seem to play a central role in producing an outcome, behavior, or other topic of interest.

Conclusion

Beginning in the early development of personality psychology as a formal scientific field, researchers worked to identify parsimonious models capable of describing personality in a comprehensive and scientifically useful way. This effort emerged at a time when personality psychology was characterized by a "babel of concepts" (John & Srivastava, 1999, p. 102), and the development of Big Few models such as the Big Five and Big Six provided researchers with a much-needed common language for studying and discussing individual differences. These

models represent some of the field's most important achievements and have generated many influential findings regarding the relevance of personality across a wide range of contexts, including health and longevity, educational and occupational outcomes, relationship functioning, and well-being. These Big Few models will continue to serve the field by fulfilling the purposes for which they were originally developed and, potentially, by helping usher personality psychology into a period of greater methodological flexibility and diversity that can accommodate the variety of topics studied in the field.

The present work contributes to a growing literature suggesting that greater attention to the item level offers a valuable approach to personality assessment. This approach can be used alongside—or, in some cases, in place of—broad trait assessment models to reveal patterns of similarity and difference. The DTS test, in particular, provides a useful item-focused method for identifying qualitative differences in the shape of response distributions that might otherwise remain obscured when relying on comparisons of means or variances. More broadly, greater adoption of item-centric approaches could improve sensitivity to culturally meaningful variation and provide a more flexible framework for studying personality across diverse contexts. Moving forward, progress in personality science will likely depend not on choosing between broad and narrow approaches to assessment, but on strengthening and integrating both—recognizing that understanding personality from the "bottom up" may be essential for building more valid, more predictively powerful, more generalizable, and more theoretically informative models of individual differences.

References

Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2021). Coffee or tea? Examining cross-cultural differences in personality nuances across former colonies of the British Empire. *European Journal of Personality*, *35*(3), 383–397. https://doi.org/10.1177/0890207020962327

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs: General and Applied*, *80*(7), 1–37. https://doi.org/10.1037/h0093906

Allport, G. W. (1937). *Personality: A psychological interpretation* (pp. xiv, 588). Holt.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. https://doi.org/10.1037/h0093360

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality and Social Psychology*, *69*(2), 300–307. https://doi.org/10.1037/0022-3514.69.2.300

Anderson, T. W., & Darling, D. A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association*, *49*(268), 765–769. https://doi.org/10.1080/01621459.1954.10501232

Arumäe, K., Vainik, U., & Mõttus, R. (2022). Body Mass is Linked With a Broad Range of Personality Nuances, but Especially Those With Behavioral Content: A Multi-Sample Exploration. *Personality Science*, *3*, 1–26. https://doi.org/10.5964/ps.7583

Asendorpf, J. B. (2016). Causal Unity of Broader Traits is an Illusion. *European Journal of Personality*, *30*(4), 304–340. https://doi.org/10.1002/per.2060

Ashton, M. C., & Lee, K. (2001). A theoretical basis for the major dimensions of personality. *European Journal of Personality*, *15*(5), 327–353. https://doi.org/10.1002/per.417

Ashton, M. C., & Lee, K. (2007). Empirical, Theoretical, and Practical Advantages of the HEXACO Model of Personality Structure. *Personality and Social Psychology Review*, *11*(2), 150–166. https://doi.org/10.1177/1088868306294907

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A Review of Research and Theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Ashton, M. C., Lee, K., Perugini, M., Szarota, P., De Vries, R. E., Di Blas, L., Boies, K., & De Raad, B. (2004). A Six-Factor Structure of Personality-Descriptive Adjectives: Solutions From Psycholexical Studies in Seven Languages. *Journal of Personality and Social Psychology*, *86*(2), 356–366. https://doi.org/10.1037/0022-3514.86.2.356

Avey, J. B., Luthans, F., & Youssef, C. M. (2010). The additive value of positive psychological capital in predicting work attitudes and behaviors. *Journal of management*, *36*(2), 430-452.

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150–159. https://doi.org/10.1016/j.paid.2017.12.018

Bakan, D. (1966). *The duality of human existence: An essay on psychology and religion* (p. 242). Rand Mcnally.

Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. Guilford Publications.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003

Baumert, A., Schmitt, M., & Perugini, M. (2019). Towards an explanatory personality psychology: Integrating personality structure, personality process, and personality development. *Personality and Individual Differences*, *147*, 18–27. https://doi.org/10.1016/j.paid.2019.04.016

Baumert, A., Schmitt, M., Perugini, M., Johnson, W., Blum, G., Borkenau, P., Costantini, G., Denissen, J. J. A., Fleeson, W., Grafton, B., Jayawickreme, E., Kurzius, E., MacLeod, C., Miller, L. C., Read, S. J., Roberts, B., Robinson, M. D., Wood, D., & Wrzus, C. (2017). Integrating Personality Structure, Personality Process, and Personality Development. *European Journal of Personality*, *31*(5), 503–528. https://doi.org/10.1002/per.2115

Beck, E. D., Condon, D., & Jackson, J. (2023). Interindividual age differences in personality structure. *European Journal of Personality*, *37*(3), 257–275. https://doi.org/10.1177/08902070221084862

Behrens, K. Y. (2004). A Multifaceted View of the Concept of Amae: Reconsidering the Indigenous Japanese Concept of Relatedness. *Human Development*, *47*(1), 1–27. https://doi.org/10.1159/000075366

Benet-Martínez, V., & Waller, N. G. (1997). Further Evidence for the Cross-Cultural Generality of the Big Seven Factor Model: Indigenous and Imported Spanish Personality Constructs. *Journal of Personality*, *65*(3), 567–598. https://doi.org/10.1111/j.1467-6494.1997.tb00327.x

Biswas-Diener, R., Vittersø, J., & Diener, E. (2005). Most people are pretty happy, but there is cultural variation: The Inughuit, the Amish, and the Maasai. Journal of Happiness Studies, 6, 205–226.

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

Blum, G. S., Baumert, A., & Schmitt, M. (2021). Chapter 2—Personality processes—From description to explanation. In J. F. Rauthmann (Ed.), *The Handbook of Personality Dynamics and Processes* (pp. 33–55). Academic Press. https://doi.org/10.1016/B978-0-12-813995-0.00002-9

Boag, S. (2015). Personality assessment,'construct validity', and the significance of theory. *Personality and Individual Differences*, *84*, 36-44.

Borsboom, D. (2008). Latent Variable Theory. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., J Mellenbergh, G., & Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*, 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

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

Caprara, G. V., Eisenberg, N., & Alessandri, G. (2017). Positivity: The dispositional basis of happiness. *Journal of Happiness Studies*, *18*(2), 353-371.

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological Bulletin*, *92*(1), 111–135. https://doi.org/10.1037/0033-2909.92.1.111

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*(2), 319–333. https://doi.org/10.1037/0022-3514.67.2.319

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality Development: Stability and Change. *Annual Review of Psychology*, *56*(1), Article 1. https://doi.org/10.1146/annurev.psych.55.090902.141913

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

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, *38*(4), 476–506. https://doi.org/10.1037/h0054116

Cattell, R. B. (1950). Describing personality through traits. In *Personality: A systematic theoretical and factual study, 1st ed* (pp. 21–44). McGraw-Hill. https://doi.org/10.1037/10773-002

Cervone, D., & Shoda, Y. (1999). Beyond Traits in the Study of Personality Coherence. *Current Directions in Psychological Science*, *8*(1), 27–32. https://doi.org/10.1111/1467-8721.00007

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the study of personality in culture. *American Psychologist, 66*(7), 593–603. https://doi.org/10.1037/a0022389

Church, A. T., Reyes, J. A. S., Katigbak, M. S., & Grimm, S. D. (1997). Filipino Personality Structure and the Big Five Model: A Lexical Approach. *Journal of Personality*, *65*(3), 477–528. https://doi.org/10.1111/j.1467-6494.1997.tb00325.x

Church, A. T. (2001). Personality measurement in cross-cultural perspective. *Journal of*

Personality, *69*(6), 979-1006.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, *101*(5), 1068–1089. https://doi.org/10.1037/a0025290

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Condon, D. M. (2022). *The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model* [Working Paper]. https://doi.org/10.31234/osf.io/sc4p9

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., & Mõttus, R. (2021). A role for information theory in personality modeling, assessment, and judgment☆. In D. Wood, S. J. Read, P. D. Harms, & A. Slaughter (Eds.), *Measuring and Modeling Persons and Situations* (pp. 1–31). Academic Press. https://doi.org/10.1016/B978-0-12-819200-9.00018-1

Condon, D. M., & Mroczek, D. K. (2016). Time to move beyond the Big Five?. *European journal of personality*, *30*(4), 311.

Corr, P. J. (2004). Reinforcement sensitivity theory and personality. *Neuroscience and Biobehavioral Reviews*, *28*(3), 317–332. https://doi.org/10.1016/j.neubiorev.2004.01.005

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, *38*(4), Article 4. https://doi.org/10.1037//0022-3514.38.4.668

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

Costantini, G., Richetin, J., Borsboom, D., Fried, E. I., Rhemtulla, M., & Perugini, M. (2015). Development of Indirect Measures of Conscientiousness: Combining a Facets Approach and Network Analysis. *European Journal of Personality*, *29*(5), 548–567. https://doi.org/10.1002/per.2014

Cramer, A. O., Van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... & Borsboom, D. (2012). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414-431.

Cramer, A. O., Van Der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... & Borsboom, D. (2012). Measurable like temperature or mereological like flocking? On the

nature of personality traits. *European Journal of Personality*, *26*(4), 451-459.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), Article 4. https://doi.org/10.1037/h0040957

Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation'with a new questionnaire. *Journal of Research in Personality*, *78*, 106-124.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*, *125*(1), 173–197. https://doi.org/10.1037/pspp0000443

De Raad, B. (2009). Structural models of personality. In *The Cambridge handbook of personality psychology* (pp. 127–147). Cambridge University Press. https://doi.org/10.1017/CBO9780511596544.011

De Raad, B., & Barelds, D. P. H. (2008). A new taxonomy of Dutch personality traits based on a comprehensive and unrestricted list of descriptors. *Journal of Personality and Social Psychology*, *94*(2), 347–364. https://doi.org/10.1037/0022-3514.94.2.347

De Raad, B., Barelds, D. P. H., Levert, E., Ostendorf, F., Mlačić, B., Blas, L. D., Hřebíčková, M., Szirmák, Z., Szarota, P., Perugini, M., Church, A. T., & Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology*, *98*(1), 160–173. https://doi.org/10.1037/a0017184

De Raad, B., & Peabody, D. (2005). Cross-culturally recurrent personality factors: Analyses of three factors. *European Journal of Personality*, *19*(6), 451–474. https://doi.org/10.1002/per.550

De Raad, B., Perugini, M., & Szirmák, Z. (1997). In pursuit of a cross-lingual reference stucture of personality traits: Comparisons among five languages. *European Journal of Personality*, *11*(3), 167–185. https://doi.org/10.1002/(SICI)1099-0984(199709)11:3<167::AID-PER286>3.0.CO;2-B

De Raad, B., Barelds, D. P. H., Timmerman, M. E., De Roover, K., Mlačić, B., & Church, A. T. (2014a). Towards A Pan–Cultural Personality Structure: Input from 11 Psycholexical Studies. *European Journal of Personality*, *28*(5), 497–510. https://doi.org/10.1002/per.1953

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

DeYoung, C. G. (2015). Cybernetic Big Five Theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/j.jrp.2014.07.004

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? *Personality and Individual Differences*,

33(4), 533–552. https://doi.org/10.1016/S0191-8869(01)00171-4

Diener, E., & Diener, C. (1996). Most people are happy. Psychological Science, 7, 181–185.

Diener, E., Tay, L., & Oishi, S. (2013). Rising income and the subjective well-being of nations. *Journal of personality and social psychology*, *104*(2), 267.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246. https://doi.org/10.1037/0022-3514.73.6.1246

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*, 109956.

Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Pre.

Dowd, C. (2020). *A new ECDF Two-Sample Test Statistic*.

Dunlop, W. L., Harake, N., Wilkinson, D., & Graham, L. E. (2020). On rhetoric and ratings: Assessing narrative identity via conceptual coding and self-ratings. *Journal of Research in Personality*, *85*, 103922. https://doi.org/10.1016/j.jrp.2020.103922

Durkee, P. K., Lukaszewski, A. W., von Rueden, C. R., Gurven, M. D., Buss, D. M., & Tucker-Drob, E. M. (2022). Niche Diversity Predicts Personality Structure Across 115 Nations. *Psychological Science*, *33*(2), 285–298. https://doi.org/10.1177/09567976211031571

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: inter-and intranational differences. *Journal of personality and social psychology*, *81*(5), 869.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Eysenck, H. (1947). *Dimensions of Personality*. Kegan Paul.

Eysenck, H. J. (1994). Normality–abnormality and the three-factor model of personality. In *Differentiating normal and abnormal personality* (pp. 3–25). Springer Publishing Company.

Finke R, Adamczyk A. Cross-National Moral Beliefs: The Influence of National Religious Context. Sociol Q. 2008 Fall;49(4):617-652. doi: 10.1111/j.1533-8525.2008.00130.x. PMID: 25097270; PMCID: PMC4119762.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fleeson, W., & Gallagher, P. (2009). The implications of Big Five standing for the distribution of trait manifestation in behavior: fifteen experience-sampling studies and a meta-analysis. *Journal of personality and social psychology*, *97*(6), 1097.

Franić, S., Borsboom, D., Dolan, C. V., & Boomsma, D. I. (2014). The big five personality traits: psychological entities or statistical constructs?. *Behavior genetics*, *44*, 591-604.

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality.

Advances in Personality Assessment, *1*, 203–234.

Goldberg, L. R. (1990). An alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, *59*(6), 1216. https://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1999). *A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several Five-Factor models. 7*, 7–28.

Goldberg, L. R., & Somer, O. (2000). The hierarchical structure of common Turkish person-descriptive adjectives. European Journal of Personality, 14, 497–531

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology*, *82*(3), 379–398. https://doi.org/10.1037/0022-3514.82.3.379

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2012). *Ten-Item Personality Inventory* [Dataset]. https://doi.org/10.1037/t07016-000

Gross, J. J., & John, O. P. (1997). Revealing feelings: facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of personality and social psychology*, *72*(2), 435.

Guyon, H., Kop, J.-L., Juhel, J., & Falissard, B. (2018). Measurement, ontology, and epistemology: Psychology needs pragmatism-realism: *Theory & Psychology*. https://doi.org/10.1177/0959354318761606

Hampson, S. E., John, O. P., & Goldberg, L. R. (1986). Category breadth and hierarchical structure in personality: Studies of asymmetries in judgments of trait implications. *Journal of Personality and Social Psychology*, *51*(1), 37. https://doi.org/10.1037/0022-3514.51.1.37

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2021. World Values Survey Time-Series (1981-2020) Cross-National Data-Set. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. Data File Version 2.0.0, doi:10.14281/18241.15.

Hahn, D. W., Lee, K., & Ashton, M. C. (1999). A factor analysis of the most frequently used Korean personality trait adjectives. European Journal of Personality, 13, 261–282.

Heine, S. J. (2020). *Cultural psychology: Fourth international student edition*. WW Norton & company.

Hendriks, A. A. J. (1997). *The construction of the five-factor personality inventory (FFPI)*. Academisch proefschrift, Rijksuniversiteit Groningen, The Netherlands.

Henry, A., & Liu, M. (2024). Jingle–jangle fallacies in L2 Motivational Self System research: A response to Al-Hoorie et al.(2024). *Applied Linguistics*, *45*(4), 738-746.Hofstede, G., & McCrae, R. R. (2004). Personality and Culture Revisited: Linking Traits and Dimensions

of Culture. *Cross-Cultural Research*, *38*(1), 52-88. https://doi.org/10.1177/1069397103259443

Hepp, J., Hilbig, B. E., Moshagen, M., Zettler, I., Schmahl, C., & Niedtfeld, I. (2014). Active versus reactive cooperativeness in Borderline Psychopathology: A dissection based on the HEXACO model of personality. *Personality and Individual Differences*, *56*, 19–23. https://doi.org/10.1016/j.paid.2013.08.013

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality*, *26*(3), 245–254. https://doi.org/10.1002/per.830

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

Hogan, R., & Roberts, B. W. (2004). A Socioanalytic Model of Maturity. *Journal of Career Assessment*, *12*(2), 207–217. https://doi.org/10.1177/1069072703255882

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

International Monetary Fund. (2024, May 14). *Uzbekistan: Staff concluding statement of the 2024 Article IV mission*. https://www.imf.org/en/news/articles/2024/05/14/mcs-uzbekistan-staff-concluding-statement-of-the-2024-article-iv-mission

John, O. P. (1991). *The big five inventory—Versions 4a and 54*. https://cir.nii.ac.jp/crid/1370004237628112900

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, *98*(6), 875–925. https://doi.org/10.1037/a0033901

Kádár, D. Z., & Haugh, M. (2013). Understanding politeness. Cambridge University Press.

Kaloi, M. A., Ali, N., & Abbas, Q. (2025). "I'm Sorry" vs."My Bad": A Cross-Cultural Pragmatic Analysis of Apology Strategies in British and American English. *The Critical Review of Social Sciences Studies*, *3*(1), 3674-3690.

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, *110*(15), 5802-5805.

Kristjánsson, K. (2012). Situationism and the Concept of a Situation. *European Journal of Philosophy*, *20*(S1), E52–E72. https://doi.org/10.1111/j.1468-0378.2011.00474.x

Kuper, N., Modersitzki, N., Phan, L. V., & Rauthmann, J. F. (2021). The dynamics, processes, mechanisms, and functioning of personality: An overview of the field. *British Journal of Psychology*, *112*(1), Article 1. https://doi.org/10.1111/bjop.12486

Kuper, N., Kroencke, L., Harari, G. M., & Denissen, J. J. (2023). Who benefits from which activity? On the relations between personality traits, leisure activities, and well-being. *Journal of Personality and Social Psychology*, *125*(1), 141.

Laajaj, R., Macours, K., Pinzon Hernandez, D. A., Arias, O., Gosling, S. D., Potter, J., Rubio-Codina, M., & Vakis, R. (2019). Challenges to capture the big five personality traits in non-WEIRD populations. *Science Advances*, *5*(7), eaaw5226. https://doi.org/10.1126/sciadv.aaw5226

Leech, G. N. (2014). *The pragmatics of politeness*. Oxford University Press.

Lenhausen, M. R., Hopwood, C. J., & Bleidorn, W. (2023). Nature and impact of reference group effects in personality assessment data. *Journal of Personality Assessment*, *105*(5), 581-589.

Liu, X., Huang, Y., Kim, J., & Na, S. (2023). How ethical leadership cultivates innovative work behaviors in employees? Psychological safety, work engagement and openness to experience. *Sustainability*, *15*(4), 3452.

Lodi-Smith, J., & Roberts, B. W. (2007). Social Investment and Personality: A Meta-Analysis of the Relationship of Personality Traits to Investment in Work, Family, Religion, and Volunteerism. *Personality and Social Psychology Review*, *11*(1), 68–86. https://doi.org/10.1177/1088868306294590

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Lounsbury, J. W., Loveland, J. M., Sundstrom, E. D., Gibson, L. W., Drost, A. W., & Hamrick, F. L. (2003). An investigation of personality traits in relation to career satisfaction. *Journal of career assessment*, *11*(3), 287-307.

Lukaszewski, A.W., Gurven, M., von Rueden, C.R., Schmitt, D.P., 2017. What explains personality covariation? A test of the socioecological complexity hypothesis. Soc. Psychol. Personal. Sci 8 (8), 943–952. doi:10.1177/1948550617697175.

Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? Psychological Bulletin, 131, 803–855. doi:10.1037/0033-2909.131.6.803.

Makaryan, A., Matevosyan, D., Mkrtchyan, H., & Isanians, V. (2024, May 15). *Armenia's economic growth: Myth or reality?* Caucasus Watch. https://caucasuswatch.de/en/insights/armenias-economic-growth-myth-or-reality.html

Marie, L. (2025). *Armenia: Democratic impulse, economic growth and the challenges of transformation for an economy under pressure*. Cairn.info. https://shs.cairn.info/armeniademocratic-impulseeconomic-growth-and-the-challenges-of-transformation-for-an-economy-under-pressure--1000025620865-page-1

Markon, K. E. (2009). Hierarchies in the Structure of Personality Traits. *Social and Personality Psychology Compass*, *3*(5), 812–826. https://doi.org/10.1111/j.1751-9004.2009.00213.x

Marsella, A. J., Dubanoski, J., Hamada, W. C., & Morse, H. (2000). The Measurement of Personality across Cultures: Historical, Conceptual, and Methodological Issues and Considerations. *American Behavioral Scientist*, *44*(1), 41–62. https://doi.org/10.1177/00027640021956080

Mastor, K. A., Thalmayer, A. G., Swami, V., & Saucier, G. (n.d.). 'Attracting-Heart' or 'Long-Handed'?: The structure of Malay personality terms [Unpublished manuscript]. School of Liberal Studies, National University of Malaysia.

Mather, K., Weston, S. J., & Condon, D. M. (2025). Exploring Generalizability from the Bottom up: Applying DTS to demonstrate an item-level approach to personality assessment. *Current Research in Ecological and Social Psychology*, 100256

Maul, A. (2013). On the ontology of psychological attributes. *Theory & Psychology*, *23*(6), 752–769. https://doi.org/10.1177/0959354313506273

McAdams, D. P. (1985). Motivation and friendship. In *Understanding personal relationships: An interdisciplinary approach* (pp. 85–105). Sage Publications, Inc.

McCrae, R. R., & Costa Jr, P. T. (1995). Trait explanations in personality psychology. *European Journal of Personality*, *9*(4), 231–252. https://doi.org/10.1002/per.2410090402

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

McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

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

Miller, J. D., & Lynam, D. R. (2012). An examination of the Psychopathic Personality Inventory's nomological network: A meta-analytic review. *Personality Disorders: Theory, Research, and Treatment*, *3*(3), 305–326. https://doi.org/10.1037/a0024567

Miller, S. S. (2009). Materialism and psychosocial maladjustment: What accounts for the relation? (Unpublished doctoral thesis). University of Oregon.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Mischel, W. (1968). *Personality and Assessment*. Wiley.

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research and Perspectives*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D. K., & Revelle, W. (2022). *Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes* [Working Paper]. https://doi.org/10.31234/osf.io/4q9gv

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. https://doi.org/10.1037/pspp0000100

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), Article 6. https://doi.org/10.1002/per.2311

Mõttus, R., Sinick, J., Terracciano, A., Hrebícková, 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

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology*, *66*(6), 574–583. https://doi.org/10.1037/h0040291

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the Prediction of Consequential Outcomes. *Annual Review of Psychology*, *57*(1), 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Park, H., Sha, M., & Pan, YL. (2014). Investigating validity and effectiveness of cognitive interviewing as a pretesting method for non-English questionnaires: Findings from Korean cognitive interviews. International Journal of Social Research Methodology, 17(6), 643-658. https://doi.org/10.1080/13645579.2013.823002

Peabody, D. (1987). Selecting representative trait adjectives. *Journal of Personality and Social Psychology, 52*(1), 59–71. https://doi.org/10.1037/0022-3514.52.1.59

Penke, L., Denissen, J. J. A., & Miller, G. F. (2007). The evolutionary genetics of personality. *European Journal of Personality*, *21*(5), Article 5. https://doi.org/10.1002/per.629

Revelle, W. (2008). *Association for Research in Personality: The home for psychological generalists*.

Revelle, W., & Condon, D. M. (2015). A model for personality at three levels. *Journal of Research in Personality*, *56*, 70–81. https://doi.org/10.1016/j.jrp.2014.12.006

Revelle, W., Dworak, E. M., & Condon, D. M. (2021). Exploring the persome: The power of the item in understanding personality structure. *Personality and Individual Differences*, *169*, 109905. https://doi.org/10.1016/j.paid.2020.109905

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*, 112552. https://doi.org/10.1016/j.paid.2024.112552.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and

Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*, *2*(4), Article 4. https://doi.org/10.1111/j.1745-6916.2007.00047.x

Rocha, P. (2024). Cultural Correlates of Personality: A Global Perspective With Insights From 22 Nations. *Cross-Cultural Research*, *0*(0). https://doi.org/10.1177/10693971241264363

Sagiv, L., & Roccas, S. (2021). How Do Values Affect Behavior? Let Me Count the Ways. *Personality and Social Psychology Review*, *25*(4), 295–316. https://doi.org/10.1177/10888683211015975

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, *73*(6), 1296–1312. https://doi.org/10.1037/0022-3514.73.6.1296

Saucier, G. (2009). Recurrent Personality Dimensions in Inclusive Lexical Studies: Indications for a Big Six Structure. *Journal of Personality*, *77*(5), 1577–1614. https://doi.org/10.1111/j.1467-6494.2009.00593.x

Saucier, G., Georgiades, S., Tsaousis, I., & Goldberg, L. R. (2005). The Factor Structure of Greek Personality Adjectives. *Journal of Personality and Social Psychology*, *88*(5), 856–875. https://doi.org/10.1037/0022-3514.88.5.856

Saucier, G., Hampson, S. E., & Goldberg, L. R. (2000). Cross-language studies of lexical personality factors. In *Advances in Personality Psychology*. Psychology Press.

Saucier, G., & Iurino, K. (2019). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000273

Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, *69*(6), 847-879.

Saucier, G., Thalmayer, A. G., Payne, D. L., Carlson, R., Sanogo, L., Ole-Kotikash, L., Church, A. T., Katigbak, M. S., Somer, O., Szarota, P., Szirmák, Z., & Zhou, X. (2014). A Basic Bivariate Structure of Personality Attributes Evident Across Nine Languages. *Journal of Personality*, *82*(1), Article 1. https://doi.org/10.1111/jopy.12028

Schwaba, T., & Thalmayer, A. G. (2024). *Openness/intellect: The weirdest trait in the Big Five world?* OSF. https://doi.org/10.31234/osf.io/wnzy4

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

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. The Bell System Technical Journal. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Siegler, I. C., & Brummett, B. H. (2000). Associations among NEO personality assessments and well-being at midlife: Facet-level analyses. *Psychology and Aging*, *15*(4), 710–714. https://doi.org/10.1037//0882-7974.15.4.710

Smaldino, P. E., Lukaszewski, A., Rueden, C. von, & Gurven, M. (2019). Niche diversity can explain cross-cultural differences in personality structure. *Nature Human Behaviour*, 1–8. https://doi.org/10.1038/s41562-019-0730-3

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281. https://doi.org/10.1214/aoms/1177730256

Sosnowska, J., Kuppens, P., De Fruyt, F., & Hofmans, J. (2019). A dynamic systems approach to personality: The Personality Dynamics (PersDyn) model. *Personality and Individual Differences*, *144*, 11–18. https://doi.org/10.1016/j.paid.2019.02.013

Sosnowska, J., Kuppens, P., Fruyt, F. D., & Hofmans, J. (2020). New Directions in the Conceptualization and Assessment of Personality—A Dynamic Systems Approach. *European Journal of Personality*, *n/a*(n/a), Article n/a. https://doi.org/10.1002/per.2233

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

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

Suzuki, T., Griffin, S. A., & Samuel, D. B. (2017). Capturing the DSM-5 Alternative Personality Disorder Model Traits in the Five-Factor Model's Nomological Net. *Journal of Personality*, *85*(2), 220–231. https://doi.org/10.1111/jopy.12235

Syed, M., & Kathawalla, U.-K. (2022). Cultural psychology, diversity, and representation in open science. In *Cultural methods in psychology: Describing and transforming cultures* (pp. 427–454). Oxford University Press. https://doi.org/10.1093/oso/9780190095949.003.0015

Szarota, P. (1996). Taxonomy of the Polish personality-descriptive adjectives of the highest frequency of use. Polish Psychological Bulletin, 27, 342–351.

Szirmák, Z., & De Raad, B. (1994). Taxonomy and structure of Hungarian personality traits. European Journal of Personality, 8, 95–118.

Taguchi, N. (2019). Second language acquisition and pragmatics: An overview. *The Routledge handbook of second language acquisition and pragmatics*, 1-14.

Tellegen, A., & Waller, N. G. (1987). *Re-examining basic dimensions of natural language trait descriptors*. 95th Annual Convention of the American Psychological Association, New York.

Thalmayer, A. G., & Saucier, G. (2014). The Questionnaire Big Six in 26 Nations: Developing Cross-Culturally Applicable Big Six, Big Five and Big Two Inventories. *European Journal of Personality*, *28*(5), 482–496. https://doi.org/10.1002/per.1969

Thalmayer, A. G., Saucier, G., & Rotzinger, J. S. (2022). Absolutism, Relativism, and Universalism in Personality Traits Across Cultures: The Case of the Big Five. *Journal of*

Cross-Cultural Psychology, *53*(7–8), 935–956. https://doi.org/10.1177/00220221221111813

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American?. *American Psychologist*, *76*(1), 116

Thalmayer, A. G., Job, S., Shino, E. N., Robinson, S. L., & Saucier, G. (2021). ǂŪsigu: A mixed-method lexical study of character description in Khoekhoegowab. *Journal of Personality and Social Psychology*, *121*(6), 1258–1283. https://doi.org/10.1037/pspp0000372

Thalmayer, A. G., Mather, K. A., Saucier, G., Naudé, L., Florence, M., Adonis, T.-A., Shino, E. N., Asatsa, S., Witzlack-Makarevich, A., Bächlin, L. Z. M., & Condon, D. M. (2024). The cross-cultural big two: A culturally decentered theoretical and measurement model for personality traits. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000528

Thielmann, I., Akrami, N., Babarović, T., Belloch, A., Bergh, R., Chirumbolo, A., Čolović, P., de Vries, R. E., Dostál, D., Egorova, M., Gnisci, A., Heydasch, T., Hilbig, B. E., Hsu, K.-Y., Izdebski, P., Leone, L., Marcus, B., Međedović, J., Nagy, J., … Lee, K. (2020). The HEXACO–100 Across 16 Languages: A Large-Scale Test of Measurement Invariance. *Journal of Personality Assessment*, *102*(5), 714–726. https://doi.org/10.1080/00223891.2019.1614011

Triandis, H. C., & Suh, E. M. (2002). Cultural influences on personality. *Annual Review of Psychology*, *53*(1), 133–160. https://doi.org/10.1146/annurev.psych.53.100901.135200

Uher, J. (2013). Personality Psychology: Lexical Approaches, Assessment Methods, and Trait Concepts Reveal Only Half of the Story—Why it is Time for a Paradigm Shift. *Integrative Psychological and Behavioral Science*, *47*(1), 1–55. https://doi.org/10.1007/s12124-013-9230-6

van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2021). Latent variable models and networks: Statistical equivalence and testability. *Multivariate Behavioral Research*, *56*(2), 175–198. https://doi.org/10.1080/00273171.2019.1672515

van Emmerik, H., Gardner, W. L., Wendt, H., & Fischer, D. (2010). Associations of Culture and Personality With McClelland's Motives: A Cross-Cultural Study of Managers in 24 Countries. *Group & Organization Management*, *35*(3), 329-367. https://doi.org/10.1177/1059601110370782

Villani, C. (2009). The Wasserstein distances. In C. Villani (Ed.), *Optimal Transport: Old and New* (pp. 93–111). Springer. https://doi.org/10.1007/978-3-540-71050-9_6

Wiernik, B. M., Yarkoni, T., Giordano, C. A., & Raghavan, M. (2020). Two, five, six, eight (thousand): Time to end the dimension reduction debate! *European Journal of Personality*, *34*(4), 559–560.

Wiggins, J. S., & Pincus, A. L. (1992). Personality: Structure and Assessment. *Annual Review of Psychology*, *43*(1), 473–504. https://doi.org/10.1146/annurev.ps.43.020192.002353

Willgerodt, M. A. (2003). Using focus groups to develop culturally relevant instruments. *Western Journal of Nursing Research*, *25*(7), 798-814.

Willis, G. B. (2005). Cognitive interviewing. SAGE Publications, Inc., https://doi.org/10.4135/9781412983655

Willroth, E. C., Luo, J., Atherton, O. E., Weston, S. J., Drewelies, J., Batterham, P. J., ... & Graham, E. K. (2023). Personality traits and health care use: A coordinated analysis of 15 international samples. *Journal of personality and social psychology*, *125*(3), 629.

Wilt, J., & Revelle, W. (2015). Affect, Behavior, Cognition, and Desire in the Big Five: An Analysis of Item Content and Structure. *European Journal of Personality*, *29*(4), Article 4. https://doi.org/10.1002/per.2002

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

World Bank. (2024, July 16). *Helping Uzbekistan undertake a historic social and economic transformation*. https://www.worldbank.org/en/results/2024/07/16/helping-uzbekistan-undertake-a-historic-social-and-economic-transformation

Wrzus, C., & Roberts, B. W. (2017). Processes of Personality Development in Adulthood: The TESSERA Framework. *Personality and Social Psychology Review*, *21*(3), 253–277. https://doi.org/10.1177/1088868316652279

Vinson, J. M., Curtis, M. B., Conover, T. L., & Chui, L. (2020). Ethical relativism in accounting: A cross-cultural examination of the influence of culture and risk taking propensity on ethical decision-making. *Journal of International Accounting, Auditing and Taxation*, *41*, 100350.

Zettler, I., Hilbig, B. E., & Heydasch, T. (2013). Two sides of one coin: Honesty–humility and situational factors mutually shape social dilemma decision making. *Journal of Research in Personality*, *47*(4), 286–295. https://doi.org/10.1016/j.jrp.2013.01.012

Zhao, Y., Summers, R., Gathara, D., & English, M. (2024). Conducting cross-cultural, multi-lingual or multi-country scale development and validation in health care research: A 10-step framework based on a scoping review. *Journal of Global Health*, *14*, 04151.

Zhou, X., Saucier, G., Gao, D., & Liu, J. (2009). The factor structure of Chinese personality terms. Journal of Personality, 77, 363– 400.

Ziegler, M., Kemper, C. J., & Lenzner, T. (2015). The issue of fuzzy concepts in test construction and possible remedies. *European Journal of Psychological Assessment*, *31*(1), 1–4. https://doi.org/10.1027/1015-5759/a000255