The Science Behind the Assessment

Version: 2023.04.25

Table of Contents

Document Overview

PrinciplesYou has several features that differentiate it from other personality instruments. First, it is based on contemporary developments in personality and organizational research, including that carried out by Professor Brian R. Little and Professor Adam Grant, who are a core part of our team. Second, it utilizes advanced psychometric techniques, including novel algorithms for creating archetypes. Third, PrinciplesYou is unique in measuring traits that Ray Dalio has identified over his decades of running a successful business, including his concept of the Shaper archetype. This document is designed to provide you with an overview of the science behind the PrinciplesYou assessment. It is aimed at the interested reader and those with technical expertise in psychometrics and personality assessment.

Excerpts of our analyses are included throughout the document, although what is shown represents a very small portion of all the analyses conducted. We also refer to the samples we have used to develop the scales. To clearly identify when we are referring to a sample, we underline it. These are the samples we will refer to, along with some of their characteristics:

- MTurk1 = A cross-sectional sample of 1,500 drawn from a predominantly US population.
- MTurk2 = A cross-sectional sample of 3,175 drawn from a predominantly US population.
- MTurk3 = A second wave of MTurk2, *n* = 190.
- PY = A sample of almost 300,000 PrinciplesYou test takers (45% US based, 55% non-US based).
- PYD = a subsample (*n* = 2,049) of the PY sample where we have collected additional information from participants with respect to their demographic information (e.g., gender, race, age).
- PYD2 = an extension of the PYD sample with additional test takers (total *n* = 2,696) to improve confidence in mean difference scores between race groups.
- BW = A cross-sectional sample of 405 Bridgewater employees.

Finally, this document is divided into six main sections, each centered around a key principle that underpins the entire PrinciplesYou development process:

1. Principle I: Traits are not Fates
2. Principle II: Look Beyond the Big Five
3. Principle III: Follow a Rational Process
4. Principle IV: Apply Psychometric Science
5. Principle V: Incorporate Typology
6. Principle VI: Continuously Improve

Principle I: Traits are not fates

The mission of the PrinciplesYou assessment offering is simple: *Understand yourself. Understand others. Work Better Together*. The benefits of our assessment flow to individual test takers, managers, and organizational leaders.

For individual test takers

An early goal of the PrinciplesYou initiative was to develop a tool that explores, identifies, and engages the test taker's preferences in the context of work and life. In this sense, test takers are empowered by a tool that facilitates rigorous self-understanding and reflection. It enables them to share with others who they are, what strengths they possess, and how they might better contribute to their team and organization. The insights that can be gained from such an activity – especially if done with team members – can improve the team's ability to meet organizational challenges.

Although the PrinciplesYou assessment measures stable personality traits, the Principles team – inspired by Professor Brian Little – has always maintained a key philosophy during development that *traits are not fates*. As we continue to build on the PrinciplesYou assessment, we believe that the insights an individual will gain from our assessment will help them to better achieve their personal projects and ultimately flourish.

For managers

Managers, dealing with a whole new world of challenges including the reality of a hybrid, remote, and distributed work environments, can use the assessment to be better attuned to their direct reports' preferences and styles. PrinciplesYou facilitates an openness for feedback and provides a foundation for ongoing personal development. The assessment also helps managers better understand and develop their own leadership capability.

For organizational leaders

The benefit of the PrinciplesYou assessment for organizational leaders is that they obtain deep insights into their teams and organization. These insights help them to develop strategies to leverage their people's strengths, build great teams, and ultimately create a high-performance culture.

Principle II: Look beyond the big five

First observed by Fiske in 1949 (see Fiske, 1949) after (a lot of) statistical number crunching, the big five has become the most ubiquitous model of personality. The big five model describes human personality along five theoretically independent dimensions. These five dimensions are:

- Openness, which describes how open-minded and intellectually curious one is.
- Conscientiousness, which describes how orderly, industrious, and detailed-oriented one is.
- Extraversion, which describes the degree to which one seeks out social interaction and rewarding stimuli.
- Agreeableness, which describes how caring, compassionate, and polite one is.
- Stability, which describes how sensitive to negative emotion one is.

The big five is a highly useful model of personality, however we wanted to build off the big five foundations and go beyond it. With this vision in mind, we used the insights from personality science and started with the big five to augment it with further insight from Ray Dalio's business experience, and Adam Grant and Brian Little's personality and organizational psychology research. PrinciplesYou is the outcome of this vision. It is a carefully constructed assessment instrument designed to help individuals understand their personality and reflect on important aspects of their work and personal lives. It is administered online and takes approximately 40 minutes to complete. On a seven-point scale (ranging from strongly disagree to strongly agree), individuals rate the extent to which a set of descriptive items (e.g., "I am very disorganized") applies to them. The output comprises a person's score on each of 12 major traits, 36 sub-traits (called facets), and 5 independent dimensions, all of which are further nested with three orientation which provide insight in terms of how one thinks, how they engage interpersonally, and how they apply themselves in the face of challenges (see Appendix 1: PrinciplesYou's 12 traits, 36 facets, and 5 independent dimensions). A person's score is also reported on 28 archetypes (see Appendix 2: Archetype archipelago) derived from an algorithm based on facet scores. An extensive narrative report provides details on the meaning of the person's scores on traits, facets, and archetypes. These narratives are unique to PrinciplesYou.

In the following sections we provide further detail in terms of how PrinciplesYou was derived from big five foundations, and how we have drawn from the personality science to create a truly unique and insightful tool.

Principle III: Follow a rational process

Back in 1967, Hase and Goldberg (1967) published a paper showing that using a rational approach during scale development can be just as effective as other traditional methods of scale development. What is a rational approach? It short, it is an approach to scale development that focuses on expert judgement and continuous refinement to arrive at a measure *that works*. It places more value on expertise and painstaking effort during the item selection process, as opposed to relying more exclusively on purely statistical techniques like factor analysis. This does not mean that statistical tools are not important, but it does mean that rational judgement is not subordinate to them.

The first step in creating PrinciplesYou was to create an item pool of several hundred short questions such as "I enjoy parties," and "I seldom procrastinate." These questions/items were modeled on (and sometimes drawn directly from) existing public domain resources, primarily the International Personality Item Pool (IPIP). IPIP contains an open-source data bank of over 3,000 items that are organized into scales to measure dimensions of personality (e.g., "I enjoy parties" is one item that measures big five trait extraversion, and "I seldom procrastinate" is one item that measures big five conscientiousness). We also included several dozen items of our own invention, tapping traits beyond the big five that we judged to be relevant to organizational goals, including some uniquely identified with our own research, such as whether a person is a "Giver or Taker" or is "Person-Oriented". We also created new items that drew from Dalio's frameworks, designed to tap into traits such as humility and toughness. Many of these are unique to PrinciplesYou. The PrinciplesYou item pool currently includes approximately 600 items and is continually being revised and expanded.

These 600 items were administered to MTurk1, MTurk2, MTurk3, and the BW participants (although not all items were included in all samples). Through a highly iterative process involving statistical/psychometric analysis, expert judgement, and hours of debate (and occasional disagreement!), a final set of items were settled on. This process took over 16 months, and there is much detail that sits behind this process, which we discuss later. The final item set consists of 246 items, which are used in the PY sample.

We performed factor analyses with both items and facets and they partially confirmed the rational grouping of items into 12 trait dimensions, although there was moderate overlap between some of the facets. Based on these analyses, 12 trait dimensions each with 3 facets were chosen. An additional set of

5 independent scales were also created, for a total of 12 traits, 36 facets, and 5 independent dimensions. Archetypes were also developed, but the scoring of these is more complex and so we dedicate a section to this later. In addition to the factor analysis, the other psychometric tests consisted of an examination of internal consistency, construct validity, two-week test-retest reliability, network analysis, and, more recently, item behavior (e.g., inspecting category response curves), the details of which we now discuss.

Principle IV: Apply psychometric science

Psychometrics is about building measurement tools that measure psychological constructs like personality traits. Within psychometric analysis, *validity* is the attribute of a psychometric test that enables us to claim that "the test measures what we say it measures", *internal consistency* is a measure of how consistently each item of a facet or trait measures the facet or trait in question, and *test-retest reliability* is an indicator of the similarity of a construct's measurement across time. In this section, we detail how we have approached the psychometric tasks needed to build a valid assessment.

Although our scale development process was underpinned by a rational approach, we still conducted numerous psychometric tests and checks. We performed detailed statistical analyses on the rational scales to ensure the highest psychometric standards. The item pool was administered to MTurk1 and MTurk2 (MTurk2 was our primary study from which our reliabilities and norm group were drawn), and care was taken to screen out those who may have not been paying attention to their answers. The test was also administered to an internal BW sample.

Internal consistency

Our strategy initially emphasized the creation of traits and facets with high levels of internal consistency. Measures of internal consistency (omega total) are generally excellent across both MTurk2, PY, and PYD, averaging .87 for traits and .81 for facet/independent dimensions. Table 1 shows the omega estimates for the PY sample. These values are on par with those from the NEO-PIR, which is the gold standard for personality assessment. These big five trait scales, as reported in the NEO-PIR manual, had average alphas for traits of .88 and facets of .71. We also augmented internal consistency with data on test-retest reliability using the MTurk2 and MTurk3 samples. As shown in Table 1, two-week retest reliabilities were excellent (averaging .87 for facets and .92 for traits) and compare very well with other research-based personality tests.

Table 1. Omegas (ω) and Test-retest (*r*) correlations

	ω	*r*		ω	*r*		ω	*r*
Autonomous	**.80**	**.82**	**Composed**	**.92**	**.96**	**Determined**	**.89**	**.94**
Independent	.71	.72	Confident	.83	.92	Driven	.86	.82
Internally Motivated	.66	.76	Calm	.90	.93	Persistent	.83	.92
Self-accountable	.80	.87	Poised	.79	.94	Proactive	.82	.91
Flexible	**.78**	**.91**	**Humble**	**.80**	**.89**	**Extraverted**	**.91**	**.95**
Adaptable	.81	.91	Modest	.75	.83	Adventurous	.85	.90
Agile	.61	.81	Open-minded	.76	.81	Engaging	.85	.93
Growth-seeking	.76	.81	Receptive to Criticism	.74	.87	Gregarious	.87	.93
Leadership	**.91**	**.92**	**Nurturing**	**.87**	**.94**	**Tough**	**.89**	**.92**
Demanding	.79	.82	Empathetic	.81	.91	Critical	.85	.86
Inspiring	.87	.90	Helpful	.86	.93	Direct	.75	.87
Taking Charge	.88	.91	Person-oriented	.66	.81	Feisty	.81	.86
Creative	**.87**	**.93**	**Deliberative**	**.84**	**.92**	**Detailed and Reliable**	**.87**	.93
Curious	.75	.86	Impartial	.70	.77	Dependable	.81	.91
Original	.87	.92	Logical	.76	.86	Detail-oriented	.83	.88
Non-conforming	.83	.87	Systematic	.78	.87	Organized	.82	.87
			Independent Dimensions					
Status-seeking	.86	.82	Energetic	.89	.91	Practical	.78	
Conceptual	.78	.94	Humorous	.85	.91			

Note : Omega scores are based on the PY sample, and the test-retest correlations are based on the MTurk2 and MTurk3 samples. Practical does not have a test-retest correlation due to changes made to its constituent items after running the test-retest analysis.

Validity of the structure

Given that the assessment is built from big five inventories, the validity of the big five element is embedded within it. To illustrate this point, Table 2 shows a typical pattern of loadings when forcing a five-factor solution (Table 2 uses the PY sample).

Table 2. Varimax Rotation Forcing Five Factors (PY Sample)

	I	II	III	IV	V
Systematic	**.80**				
Organized	**.77**				
Driven	**.63**				
Dependable	**.61**			.31	
Persistent	**.58**			.40	
Practical	**.52**				
Logical	**.51**				
Self-accountable	**.51**				
Detail-oriented	**.50**				
Impartial	**.47**				

Curious		**.68**			
Original		**.62**			
Open-minded		**.59**			
Conceptual		**.58**			
Internally Motivated		**.55**			
Proactive	.38	**.50**	.31	.31	.35
Growth-seeking		**.47**		.46	
Person-oriented		**.47**			.37
Adventurous		**.41**	.38		.36
Feisty			**.72**		
Critical			**.68**		
Direct			**.67**		
Demanding	.42		**.54**		
Non-conforming		.44	**.53**		
Empathetic		.33	**-.52**		.33
Modest			**-.51**		
Independent			**.42**		
Helpful		.44	**-.34**		.34
Confident				**.78**	
Calm				**.72**	
Poised				**.65**	
Adaptable		.37		**.63**	
Receptive to Criticism		.37		**.51**	
Engaging			.30		**.74**
Gregarious					**.72**
Inspiring	.30	.39			**.56**
Energetic	.34			.45	**.53**
Taking Charge	.32		.46		**.50**
Humorous					**.39**
Status-seeking				-.45	**.32**
Agile					**.31**

In terms of the specific traits and facets, these were constructed to provide additional insight for users beyond what can be obtained with a big five model and are a unique feature of the PrinciplesYou assessment. Going beyond the big five was not without justification and indeed several factor analyses were conducted using the MTurk2, PY, and PYD samples as part of this process (scree plots, for example, provided evidence that there was much more than just the big five). Using the analysis conducted on English-speaking test takers in the PY sample as an example (a similar approach was first taken with the MTurk1 and MTurk 2 samples), we conducted four rounds of factor analysis (each using principal axis factoring extraction, varimax rotation, and maximum likelihood estimation). We assessed how the items loaded onto the facets as a comparison to the structure identified in the MTurk2 sample. The first round identified 17 clear factors, with good loadings against the expected facets. The remaining items were then subjected to a second round, which found 6 clear factors. Two additional rounds were then run on

the remaining items, which identified 4 factors. The full solution across four rounds included 27 factors, of which some factors included items from multiple facets (an example of this is shown in Table 3). The fact that items from multiple facets were captured under one factor makes sense given that those facets are theoretically related. Overall, the full solution provided excellent replication of the structure of the assessment.

Table 3. Example Factors from Item-level Factor Analysis (PY Sample)

Round 1: Factor 7		Round 1: Factor 8	
Item	**Loading**	**Item**	**Loading**
helpful4	.73	impartial6	.58
empathetic3	.62	logical1	.57
empathetic5	.62	impartial1	.53
helpful2	.62	impartial3	.52
helpful7	.61	logical8	.52
helpful1	.54	logical7	.51
helpful6	.52	logical3	.48
helpful5	.49	logical6	.43
empathetic8	.47	impartial5	.30
empathetic4	.40	impartial8	.19
empathetic2	.34	logical4	.16
empathetic7	.25	impartial7	.13

Note: The items are named here based on their coding convention and map back to the specific item (e.g., helpful4 is "Helping others is a core value for me").

Factor analysis was not the only technique used to develop the assessment. As noted previously, the development of the facets and traits was a highly iterative one that involved constant fine-tuning of the items in each facet/trait. The evidence we present in this document show that the traits and the facets i) differentially predict outcomes (that is, there is sufficient discrimination between traits/facets that justify not collapsing them), and ii) no trait pairs are correlated beyond $r = .70$ (the highest being Determined ~ Leadership, $r = .67$; Tough ~ Leadership, $r = .61$), and only seven (out of 784) between-trait facets are correlated higher than $r = .60$. Even most of the within-facet correlations are lower than $r = .70$, and none are correlated higher than $r = .74$.

Predicting outcomes

We have conducted several analyses predicting eight different outcomes in the MTurk2 sample (Life Satisfaction Overall, Working Life Satisfaction, Social Life Satisfaction, Homelife Satisfaction, Recreation

Satisfaction, Emotional Satisfaction, Physical Satisfaction, Pay Satisfaction, Supervisor-rated Work Performance, Self-rated Work Performance). The analysis shows that i) the traits/facets predict meaningful outcomes, and ii) the relationships between the traits/facets and the outcomes do not (for the most part) depend on sex, age, education. More specifically:

- For every outcome variable at least two traits are valid predictors, even after controlling for sex, age, and education. In many cases, as many as six traits show meaningful predictive value.
- At the facet level, again, for every outcome at least four facets (and in many cases 8 to 11) are valid predictors, even after controlling for sex, age, and education.
- The facets show differential prediction to the traits. That is, if a trait predicts an outcome, we don't necessarily see all three of the trait's constituent facets doing the work (and, conversely, sometimes we see that the trait is non-significant, but a constituent facet is). For example, Satisfaction with Working Life is predicted at the trait level by Composed (b = 0.767, SE = 0.082, p < .001) and yet at the facet level Confident is the only constituent significant facet (b = 0.691, SE = 0.108, p < .001) – Calm and Poised are both non-significant. Another example is Supervisor-rater Work Performance (which was reported by the participants based on their last performance review). This outcome measure is predicted at the trait level by Leadership (b = 0.113, SE = 0.035, p = .001) and not at all by Composed (b = 0.038, SE = 0.026, p = .147). Yet, at the facet level, Taking Charge is the key Leadership facet driving performance (b = 0.117, SE = 0.040, p = .003), and Poised (a facet of Composed) also is significant (b = 0.096, SE = 0.035, p = .006) even though Composed at the trait level isn't. Given our high internal consistency estimates, this suggests that although the facets do indeed form part of a higher-order trait, they are valid predictors in their own right.
- Even though we see these encouraging results, we also see that these results (with only a few minor exceptions) do not appear to depend on sex or education. There are some age dependencies, but these are sporadic. In general, these results suggest that the scales are working in the same way for all people (noting that language and race are commented on separately in the next subsection).

We also collected data on Bridgewater's 360-degree employee competency ratings in the BW sample (which is an extremely comprehensive dataset) and compared this to employee personality. We found numerous correlations between the competency scores and the personality measures. For example, "lateral thinking" is predicted by Creative (r = .47), Non-conforming (r = .47), and Original (r = .45).

"Holding people accountable" is predicted by Tough (*r* = .28), Demanding (*r* =.32) and the facet level and Leadership at the trait level (*r* = .32). "Willing to touch the nerve" is predicted by Tough (*r* = .56), "shaping change" is predicted by Demanding (*r* = .34), Inspiring (*r* = .36), and Taking Charge (*r* = .39) as well as the overall trait of Leadership (*r* = .38). "Organized and reliable" is predicted by Detailed/Reliable (*r* = .49). "Learning from mistakes" is predicted by Humble at the trait level (*r* = .64), Modest (*r* = .66), Open-minded (*r* = .57), and Receptive to Criticism (*r* = .67). "Composed" is predicted by Composed (*r* = .61). There are still many others across a range of competencies. In fact, all 17 traits were at least moderately predictive of a related competency measure. These results compare favorably with other research-based assessments in terms of predicting job performance.

Third, we conducted occupation analysis. For this analysis, we obtained occupation data based on O*Net classifications in the MTurk2 sample along with the personality items, and we predicted the odds (using logistic regression) of occupying a role in each job family as predicted by traits and facets. Job family is a very blunt measure of occupation because it aggregates up more specific occupation groups and so we did not expect to see large effects. However, there are several relationships that are notable. For example, in the "Computer and Mathematical" job family (which is one of the narrower job families in that the constituent occupations are more closely related), we find that people who are Original, Logical, Confident, *not* Person-oriented, and Driven are more likely to be found in these roles. Another example is "Life, Physical, and Social Science" (which consists mostly of the life and physical sciences), where people who are Feisty, Poised, low Empathetic at the facet level, and Deliberative at the trait level, are more likely to be found. We suspect that analysis at more fine-grained occupation level would yield clearer results (and would likely be better predicted by facets than traits), but such a study would require a very large *n*.

Sex, race, location, and language

We ran many analyses based on language, country, race and sex. Overall, the evidence indicates that the assessment works consistently across demographic groups, subject to the language of the test taker.

Table 4 summarizes the whole-sample means and standard deviations for every trait and facet based on the PY sample and also shows Cohen's *d* effect sizes (where the pooled standard deviation is weighted by group size) comparing the differences between men and women using the PYD sample. These effects are in the expected direction. For example, women are higher in Nurturing, men are higher in Tough and

Composed. Other comparisons have been run based on other grouping variables (e.g., age, education). These differences are also in the expected direction. For example, people aged 45 and over are lower in Extraverted (*d* = -0.22) and Creative (*d* = -0.18), and higher in Composed (*d* = 0.36) and Nurturing (*d* = 0.21) compared to people aged 30 and under, which is consistent with research on personality maturation (e.g., McCrae et al., 1999).

Table 4. Means, Standard Deviations, and Effect Size Differences Between Men and Women

	M	SD	*d*		M	SD	*d*		M	SD	*d*
Autonomous	**5.25**	**0.58**	**.02**	**Composed**	**4.94**	**0.97**	**-.36**	**Determined**	**4.91**	**0.83**	**-.09**
Independent	4.53	0.78	-.03	Confident	4.72	1.15	-.39	Driven	4.79	1.01	-.10
Internally Motivated	5.56	0.73	.09	Calm	4.93	1.16	-.23	Persistent	4.66	1.10	-.04
Self-accountable	5.65	0.83	-.01	Poised	5.18	0.97	-.33	Proactive	5.28	0.94	-.08
Flexible	**4.56**	**0.60**	**-.15**	**Humble**	**4.87**	**0.64**	**-.08**	**Extraverted**	**3.71**	**1.08**	**-.25**
Adaptable	4.77	0.85	-.17	Modest	4.60	1.13	.31	Adventurous	3.95	1.09	-.23
Agile	4.11	0.96	-.13	Open-minded	5.29	0.85	-.24	Engaging	3.73	1.33	-.25
Growth-seeking	4.82	0.83	.01	Receptive to Criticism	4.73	0.93	-.27	Gregarious	3.44	1.37	-.16
Leadership	**4.52**	**1.01**	**-.12**	**Nurturing**	**4.86**	**0.83**	**.44**	**Tough**	**4.21**	**0.88**	**-.42**
Demanding	4.41	0.92	-.19	Empathetic	4.58	1.02	.52	Critical	4.51	1.04	-.43
Inspiring	4.67	1.18	-.05	Helpful	5.18	1.03	.22	Direct	4.28	0.99	-.28
Taking Charge	4.48	1.30	-.10	Person-oriented	4.81	0.87	.39	Feisty	3.84	1.02	-.38
Creative	**4.55**	**0.77**	**-.20**	**Deliberative**	**5.17**	**0.66**	**-.21**	**Detailed and Reliable**	**5.06**	**0.79**	**.06**
Curious	5.18	0.88	-.01	Impartial	5.23	0.81	-.06	Dependable	5.49	0.94	.05
Original	4.78	0.96	-.22	Logical	5.08	0.84	-.21	Detail-oriented	4.83	1.01	.10
Non-conforming	3.70	1.11	-.22	Systematic	5.19	0.72	-.27	Organized	4.86	1.03	.00
				Independent Dimensions							
Status-seeking	4.23	1.19	.05	Energetic	4.83	1.23	-.20	Practical	5.43	1.19	-.08
Conceptual	4.82	1.14	-.19	Humorous	5.10	1.01	.02				

Note: Means and standard deviations are based on the PY sample. Effect sizes are Cohen's *d* with pooled standard deviation weighted by group size comparing women to men (+ve shows that women are higher than men), and are based on the PYD sample.

In terms of observations with respect to differences in average facet/trait levels between race groups (PYD2 sample), Asian test takers in Asia (but not Asian test takers in The West) are lower in agentic and extraverted facets (e.g., Taking Charge, *d* = -0.28, Growth Seeking, *d* = -0.11, Gregarious, *d* = -0.25, etc.) in comparison to white test takers in the West, which can be expected as per the literature (see McCrae & Terracciano, 2005). We did not hypothesize other race differences, however compared to White test takers in the West, Black test takers are higher in Person-oriented (*d* = 0.37), Helpful (*d* = 0.31), Practical (*d* = .30), Inspiring (*d* = 0.28), Detail-oriented (*d* = .24), and Humorous (*d* = .24), and lower in Conceptual

(*d* = -.31). Finally, Latino test takers are characterized by higher Flexible (e.g., Growth-seeking, *d* = 0.42), Humble (e.g., Open-minded, *d* = 0.12, Receptive to Criticism, *d* = 0.20) and Determined (e.g., Driven, *d* = 0.27, Persistent, *d* = 0.27, Proactive, *d* = .18) facets. These effect sizes are based on the PYD2 sample, noting that due to the small sample size for some groups (especially the Black test takers), 95% confidence intervals effect sizes are wide (a confidence interval can be thought of as a range within which we are 'reasonably' confident that the true effect size lies). More specifically, the effect sizes reported for Black test takers can be expected to fall +/- .22 of the reported effect size, for Asians in the West and Latino test takers it is somewhere between +/- .17, and for Asians in Asia it is somewhere between +/- .14. This range means that for many facets, there is a reasonable probability that the difference between Black test takers and white Western test takers is actually zero (e.g., we can be reasonably confident that the true effect size for Driven lies somewhere between -.02 and .43). The full set of effects can be seen in Table 5.

Table 5. Effect Size Differences by Race Compared to White Western Test Takers

	AA	AW	B	L		AA	AW	B	L
Autonomous	**-.11**	**.09**	**.12**	**.17**	**Composed**	**-.21**	**.14**	**.10**	**.02**
Independent	-.25	.00	-.03	-.10	Confident	-.16	.08	-.01	.07
Internally Motivated	.26	.26	.27	.30	Calm	-.16	.11	.16	-.02
Self-accountable	-.17	-.02	.06	.23	Poised	-.24	.19	.09	.01
Determined	**-.09**	**.12**	**.18**	**.31**	**Flexible**	**-.16**	**.29**	**.06**	**.28**
Driven	.02	.14	.20	.27	Adaptable	-.35	.18	-.12	.15
Persistent	-.06	.02	.20	.27	Agile	.12	.28	.07	.06
Proactive	-.20	.13	.00	.18	Growth-seeking	-.11	.13	.20	.42
Humble	**-.16**	**.23**	**-.01**	**.15**	**Extraverted**	**-.22**	**-.02**	**-.06**	**.02**
Modest	-.29	.12	-.17	-.01	Adventurous	-.15	.03	-.06	.06
Open-minded	.00	.26	.10	.12	Engaging	-.14	-.04	-.10	.03
Receptive to Criticism	-.02	.13	.09	.20	Gregarious	-.25	-.03	.02	-.05
Leadership	**-.17**	**.04**	**.09**	**.16**	**Nurturing**	**-.03**	**.26**	**.30**	**.11**
Demanding	-.02	.00	.00	.09	Empathetic	-.16	.10	.09	.00
Inspiring	-.12	.14	.28	.21	Helpful	.02	.22	.31	.19
Taking Charge	-.28	-.03	-.02	.12	Person-oriented	.08	.33	.37	.09
Tough	**.00**	**.03**	**.03**	**.12**	**Creative**	**-.15**	**.10**	**-.04**	**.00**
Critical	.10	.12	.06	.20	Curious	-.29	.17	.05	.10
Direct	-.05	-.04	-.09	.15	Original	-.10	.00	-.09	.02
Feisty	-.05	.00	.12	-.02	Non-conforming	-.02	.08	-.04	-.10
Deliberative	**-.02**	**.16**	**-.06**	**-.02**	**Detailed and Reliable**	**-.11**	**.07**	**.00**	**.07**
Impartial	.08	.25	-.15	.03	Dependable	-.29	.04	-.19	.02
Logical	-.07	.03	-.13	-.06	Detail-oriented	-.01	.06	.24	.00
Systematic	-.06	.13	.11	-.02	Organized	.03	.05	-.07	.13

	Independent Dimensions								
Status-seeking	-.13	.07	-.16	-.20	Energetic	-.29	.04	-.09	.18
Conceptual	-.35	-.10	-.31	-.11	Humorous	-.06	.20	.24	.03
Practical	.24	.24	.30	.13					

Note : Based on the PYD2 sample. Effect sizes are Cohen's *d* with pooled standard deviation weighted by group size comparing each group to white Western test takers where *n* = 1,679 (+ve shows that the group scores higher than the white Western test takers). AA = Asia-based Asian (*n* = 222), AW = Western-based Asian (*n* = 150), B = Black (*n* = 80), L = Latino (*n* = 157). Non-Western White (*n* = 150) and Other (*n* = 258) not shown to reduce clutter.

When we ran validity checks (correlation structure comparisons, omega internal consistency, factor analyses, etc.) based on language, country, and race, the evidence supports the view that the assessment works consistently across countries and race subject to the language of the test taker. More specifically, when we analyzed the language and country of test takers in the PY sample, we found that test takers in English-speaking countries (e.g., US, UK, Australia, Canada, Singapore) produce remarkably similar correlation structures, which also is reflected in similar factor loadings when forcing a five-factor model in factor analyses. When looking at the pattern of omega scores, these too were very similar across English-speaking countries. Although we found that omega scores for facets and traits were very good for English-speaking test takers, there was some expected breakdown of the omega scores in non-English-speaking sub-samples. This is perhaps not surprising given that test takers' browsers would be required to either translate the items, or the test taker would be required to translate the item themselves. In either case, it is likely that some items were 'lost in translation'. That said, most facets (and all traits) fared well across language groups. In terms of the factor analysis, the big five framework was extracted using factor analysis with both oblique and varimax rotation across all major language groups. Although the correlation patterns and factor structures were broadly in line with the English-speaking patterns/structures, there were more facets (rather than traits) that failed to reach adequate levels of internal consistency ($\omega \geq .70$). This problem was most pronounced for Chinese speaking test takers in China. For Chinese-speaking test takers in China, 12 out of 41 facets did not reach $\omega = .70$, however 10 of these were close (scoring $\omega \geq .60$). Overall, these language-based trends are very encouraging given that the assessment was built for English-speaking test takers. We are therefore confident that with relatively minimal translation work, the assessment can be adapted to other languages.

With respect to race using the PYD sample, omegas are generally very good. The Latino group omega score is notably low for Receptive to Criticism, ω = .65 (and to a much lesser extent Systematic, ω = .67), although detailed inspection of the underlying items shows that the low score is caused by isolated items. For example, the item "If it's true that I'm lousy at something, I'd want to know it" does not appear to be interpreted in the same way as for non-Latino groups. Simply dropping this item alone increases the omega score to .69. Although not as problematic, the item "When it comes to negative feedback I'm pretty thick-skinned about it" is also interpreted differently for Latino groups compared to non-Latino groups. Dropping both items increases the omega to .72, which indicates that the scale is likely measuring the same construct. It is a question for future research to better understand why, precisely, the Latino group interprets these items differently. Despite this, we still find that a five-factor solution can be found for the Latino group, in line with the larger sample (this is also true for other race groups). An example of this extraction is shown in Table 6 for the Latino group (*n* = 111).

Table 6. Varimax Rotation Forcing Five Factors (PYD Sample, Latino Group)

	I	II	III	IV	V
Systematic	**.82**				
Organized	**.83**				
Driven	**.57**				
Dependable	**.45**				
Persistent	**.45**			.53	
Practical	**.47**				
Logical	**.60**				
Self-accountable	**.47**	.34			
Detail-oriented	**.35**			-.31	
Impartial	**.49**				
Curious		**.56**			
Original		**.54**			
Open-minded	.47	**.37**			
Conceptual		**.52**			
Internally Motivated		**.64**			
Proactive		**.48**	.31	.31	.42
Growth-seeking		**.67**		.30	
Person-oriented					.59
Adventurous		**.41**	.44	.33	.34

Feisty	.32		**.74**		
Critical	.31		**.69**		
Direct			**.69**		
Demanding	.39		**.58**		
Non-conforming			**.60**		
Empathetic			**-.36**		.57
Modest			**-.43**		
Independent			**.45**		
Helpful		.37			.52
Confident				**.72**	
Calm				**.76**	
Poised				**.62**	
Adaptable		.46		**.64**	
Receptive to Criticism		.43		**.39**	
Engaging					**.74**
Gregarious					**.78**
Inspiring		.34			**.58**
Energetic	.36		.30	.50	**.39**
Taking Charge		.32	.39		**.52**
Humorous					**.50**
Status-seeking		-.34			**.35**
Agile					**.30**

Overall, when considering all of the psychometric evidence, our conclusion is that the PrinciplesYou assessment appears to be working well across different sex, race, and location groups. The evidence suggests that the more important psychometric consideration is the language of the test taker.

Recruitment, selection, and adverse impact

Because our tools have not been designed explicitly for selection purposes (although they could be adapted for such "high-stakes" applications in partnership with the Principles team), we have not conducted specific adverse impact tests (e.g., applying the four-fifths rule).

Principle V: Incorporate typology

There has long been a tension between the scientific aversion to personality typology and the layperson desire for clear types. Rather take a narrow view of this issue, in developing the archetypes and the associated algorithm, we opted to focus first on getting the facet/trait scales right and then on using these as a basis for developing our own types that sit alongside the main assessment. Once we were broadly satisfied with the facet/trait structure, the process for developing the archetypes was i) highly iterative, in the same spirit as the rest of the assessment development process, ii) involved many of us

debating the merits of archetypes and how they should be scored (the Q-sort methodology became a key method that we adapted for this purpose), iii) governed by a principle of having a within-person type (again, the Q-sort methodology inspired this decision) but that was in some way anchored to the population (see Step 2 of the algorithm below), and iv) would sit alongside the rest of the assessment (facets/traits) to provide test takers with a unique way to understand their key strengths. Both the algorithm and thematic descriptions of the archetypes are distinctive to our assessment.

The archetypes themselves were developed by the Principles team and are built on the PrinciplesYou assessment. Ray Dalio recognized the need for a typology system because many users of personality tests find that having a 'type' is an intuitive and easy way of obtaining a personality snapshot. The challenge with such an approach is that if it is used exclusively, then the assessment is likely to produce invalid results (i.e., the type you are assigned to is too crude or vague to be of much practical use). As mentioned, our approach has instead been to start with a valid personality assessment, and then to build typology on top of that assessment framework. The team started with an initial list of archetypes that they could easily identify based on their combined expertise and experience in personality science and business. This initial list broadly aligns with the islands in the archipelago shown in Appendix 2. From there (after much iterative legwork), additional nuance was added, resulting in the 28 archetypes. The general logic is that the archetypes in each island have certain qualities in common but are subtly different to each other in small ways. Take the 'Architects' island for example, which includes the Orchestrator, Strategist, Planner archetypes. Architects tend to be Detail-oriented, Organized, Dependable, and Systematic, and to a lesser extent also Logical, Practical, and Persistent. Where Planners align well with these facets, Orchestrators have a slight tilt towards Taking Charge, Person-oriented, and Gregarious, whereas Strategists are slightly less Detail-oriented but are also slightly more Conceptual. Moreover, as detailed below, the individual report provides the test taker's top three, and bottom two, archetypes to provide more nuanced personality feedback than would be available with just a single type.

Development of the archetypes and testing of the scoring algorithm was done primarily on the MTurk2 sample and was then validated with Bridgewater employees (the BW sample) and more recently in a small sample of PrinciplesYou test takers (the PY sample). We have found the output of this strategy to be well received by our users. For example, when we asked a smaller sample of the PrinciplesYou test

takers to rate the value of their archetype matches and descriptions, 87% said that they found them to be either “extremely valuable” or “quite valuable”.

With the above in mind, the archetypes are calculated with three steps:

1. First, a *raw score* for each archetype (*j*) is calculated for each test taker (*i*). This *raw score* is calculated by taking the *raw scores* of each facet and multiplying them by a weighting matrix (a truncated version of this is shown in Table 7). The weighting matrix contains approximately normally distributed weightings of the importance of each facet to each archetype (this is inspired by the Q-sort methodology). The result of the multiplication is a contribution to the archetype *raw score* by each of the facets (where, in many cases, the facet’s contribution will be zero if it is irrelevant to the archetype). Each facet’s contribution is summed to calculate the archetype *raw score*. The preceding calculations are designed to result in a *raw score* that sits between 1 and 7 (the same as for the traits and facets).

Table 7. Archetype Weighting Matrix Example

	Original	Curious	Feisty	Critical	Direct
Commander	-1	-1	0	2	0
Shaper	3	2	3	2	0
Quiet Leader	1	1	1	1	2
Promoter	0	1	1	-2	-1
Campaigner	0	0	-2	0	0

Note: The importance of each facet to each archetype ranges from -3 to +3.

2. Second, the *raw score* for each archetype (for each test taker) is adjusted to create an *adjusted score* for each archetype (for each test taker). The adjustment process is designed to generate a score that better lends itself to within-person rank ordering of archetypes. In essence, the calculation takes the *raw score* and subtracts it from a *compressed benchmark score*, where the *compressed benchmark score* modifies the population norms slightly. The formula for calculating the adjusted score for this step is given by:

$$Adjusted\ score_{ij} = x_{ij} - \left[\frac{(c_j - R_{min})\cdot(T_{max} - T_{min})}{R_{max} - R_{min}} + T_{min}\right]$$

where x is the ith test taker’s *raw score* for archetype j, everything inside the square brackets is the *compressed benchmark score*, c is the mean score in the normed data for archetype j, R_{min} (

R_{max}) is the lowest (highest) mean score out of all 28 archetype average scores in the normed data, and T_{min} (T_{max}) is the lowest (highest) score that the archetype average scores are compressed to (currently set to the mean +/- 1 SD of all archetype average scores in the normed data). This compression step may not be intuitive, but it was the result of many weeks of fine-tuning and validating the archetype outputs against individuals at Bridgewater. Without this step we found that certain archetypes were overrepresented and did not quite reflect the best fit with the individuals at Bridgewater.

3. Third, the adjusted score is used to rank order the archetype scores within-person. The highest three adjusted scores are then used to determine the test taker's top three archetypes. One important note here is that we also check for a "level of match" based on population data, with a "Good Match" noted if the score is 80th percentile or above, and a moderate match if between 60th and 80th percentile. If the individual doesn't have any "Good Match" scores for any archetypes (which is the case for approximately 10% of the population), we then include a note stating that they don't match strongly to any types. This occurs largely if the individual has close-to-average results across most traits.

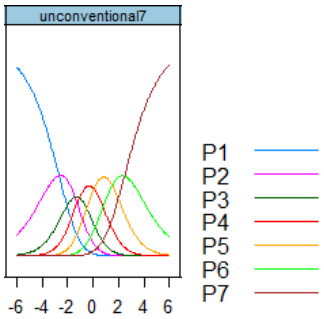
When looking at the prevalence of archetypes in various job families, we have run analyses using the MTurk2 sample and found complementary relationships for many job families. For example, within the "Computer and Mathematical" job family, Investigator, Quiet Leader, Critic, Growth Seeker, and Artisan archetypes are relatively more prevalent than expected, and Adventurer, Peacekeeper, Problem Solver, Helper, Protector Sentinel, Coach archetypes are less prevalent than expected. These relationships reflect the systematic, introverted, inquisitive, and impersonal demands typical of such roles. Within the "Healthcare Practitioner and Technical" job family, Helper, Peacekeeper, Entertainer, and Impresario archetypes are more prevalent than expected, and Growth Seeker, and Commander archetypes are less prevalent than expected, consistent with the person-centered character of healthcare roles. Within the "Architecture and Engineering" job family, Orchestrator, and Planner archetypes are more prevalent, and Growth Seeker archetypes are relatively weakly represented, consistent with such roles that typically require conscientiousness, logical problem solving, and systematizing.

Principle VI: Continuously improve

Although we have conducted numerous psychometric tests, we continue to seek improvements. As part of our advanced process of improvements, we have drawn from item response theory (IRT) to more precisely assess how each of the PrinciplesYou items behave. As noted by Emberton and Reise (2000), IRT can have benefits when used for personality assessment, however it is not always the case that such an approach will yield a better result than classical test theory. Indeed, IRT is generally better suited to performance-oriented measurements such as IQ, spelling tests, examinations, etc. Notwithstanding these points, we have analysed category response curves for every item in every facet in both the MTurk2 and PY samples. What follows is a brief introduction to the approach we have taken, but the important point is that our analysis shows that only 10 items (out of 246) require some form of amendment in the future. Four of these 10 are in the Agile facet, which is perhaps not surprising given the relatively lower internal consistency for Agile (see Table 1 above).

An overview of category response curves

Category response curves provide a visual method for assessing how well each anchor point on an item's measurement scale works. In our case, "category" means an anchor point on the 1-7 Likert scales used in the assessment, where 1 = "strongly disagree" and 7 = "strongly agree". The underlying theoretical assumption is that the test taker's trait produces their scores on the specific items. Here is what a "good" plot looks like:



On the x-axis we have the standardised score on the facet/trait (in this example, the facet of Non-conforming). Each curve is a probability curve for each of the 7 categories (scale anchors). What we should see – if everything is working well – is that if someone scores high on Non-conforming, they should be far more likely to score high on the item's upper anchor points (the item in this example is unconventional7, or "I love to break with convention"). In the above example, this is indeed the case because we can see that someone who has a Non-conforming standardised score of +4 (which is

extreme) has a probability of virtually zero in terms of choosing a 1 (P1) on the scale. All items (six in our case) that are used to measure Non-conforming are checked individually.

The second thing we should see if everything is working well, is a reasonably consistent order of the bell curves with the peaks of each curve sitting nicely in between each other in the expected order. When they are ordered nicely – as they are in the above example – it means that as someone scores higher on a trait, they are also more likely to score higher on the item's scale (by choosing a higher anchor point). If the order is wrong, it suggests that the scale it not working as expected. For example, why would someone who is very high on the trait, be more likely to select a 5 instead of a 6 on the scale? A problem of this sort suggests that there must be an interpretation problem with the item/anchors.

The third point – although a relatively minor one – is that the peaks of each curve should "pop out", rather than be subsumed by the other curves. In the above example, this point is pretty much true, although only just, because scale anchor 3's (P3) peak is almost subsumed by scale anchor 2 (P2) and scale anchor 4 (P4). It's not a major problem if a scale anchor is subsumed by the others, it just means that the scale is probably not as efficient as it could be.

The IRT analyses we have performed are based on the use of a *generalised partial credit model*, which is a form of dominance IRT model. A dominance IRT model assumes that as the underlying trait increases, so too does the probability of selecting a higher scale anchor point. This assumption might be too strict because it does not allow for the possibility that the probability of endorsing a higher scale anchor point might both *increase and decrease* as the trait score increases. This possibility can occur, for example, where the wording of an item includes an explicit or implicit condition. To draw from an example in Broadfoot (2008, p. 11), "an item measuring extroversion might say "I sometimes like to go to parties." A person that is extremely extroverted might not endorse this item because he *always* likes to go to parties. This person is disagreeing with the item from above the item. Alternatively, an extremely introverted person might not endorse this item because she *never* likes to go to parties." An alternative approach, therefore, is to use a *generalized graded unfolding model*. Analysis based on this alternative approach is planned for future work.

Final comments

Going well beyond its big five foundations, PrinciplesYou is a personality assessment tool built by a team of psychologists and personality scientists, augmented by the insights and business experience of Ray Dalio. The development of the tool was driven by well-established personality science and psychometric principles underpinned by an appreciation that *traits are not fates*. We believe that PrinciplesYou exemplifies an industry best-practice tool for personality assessment that delivers immense insight to individuals, managers, and organizational leaders so that they may create value for themselves and the organizations they serve.

References

Broadfoot, A. A. (2008). *Comparing the dominance approach to the ideal-point approach in the measurement and predictability of personality.* (Unpublished doctoral thesis). Bowling Green State University.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* New Jersey: Lawrence Erlbaum Associates.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology, 44*(3), 329-344.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67*, 231-248.

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*, 407-425.

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

Appendix 1: PrinciplesYou's 12 traits, 36 facets, and 5 independent dimensions

APPLY	**Autonomous**	**Composed**	**Determined**	**Flexible**	**Humble**	***INDEPENDENT DIMENSIONS***
	Independent	*Confident*	*Driven*	*Adaptable*	*Modest*	
	Internally Motivated	*Calm*	*Persistent*	*Agile*	*Open-minded*	***Energetic***
	Self-accountable	*Poised*	*Proactive*	*Growth-seeking*	*Receptive to Criticism*	***Status-seeking***
ENGAGE	**Extraverted**	**Leadership**	**Nurturing**	**Tough**		
	Adventurous	*Demanding*	*Empathetic*	*Critical*		
	Engaging	*Inspiring*	*Helpful*	*Direct*		***Humorous***
	Gregarious	*Taking Charge*	*Person-oriented*	*Feisty*		
THINK	**Creative**	**Deliberative**	**Detailed and Reliable**			
	Curious	*Impartial*	*Dependable*			***Conceptual***
	Original	*Logical*	*Detail-oriented*			***Practical***
	Non-conforming	*Systematic*	*Organized*			

APPLY: Your *motivational orientation* describes how you manage and apply yourself as challenges are faced, how ambitiously you set goals for yourself, how you cope with setbacks and failure, and how you leverage these experiences to learn, develop, and grow.

ENGAGE: Your *interpersonal orientation* reflects how you engage with others. Understanding your natural inclinations can help you get the most out of your relationships with people.

THINK: Your *cognitive orientation* describes your approach to thinking. Your approach to thinking can reveal what type of work you might prefer, at which aspects of a job you are more likely to excel, and how you tend to approach and solve problems.

Appendix 2: Archetype archipelago