PERSONALYSIS®

Personalysis Research Report #3: Reliability Estimates of Personalysis Scales

Authors:
Robert W. Loy, MA
(Industrial and Organizational Psychology),
Loy Consulting

Lawrence G. Weiss, PhD
(Industrial and Organizational Psychology),
Test Development Consultant

Linda Layton, PhD
(Organizational Psychology),
PentaFold Development

February 1st, 2023

Table of Contents

Abstract

The purpose of this study is to evaluate the reliability estimates of each Personalysis scale. This technical report describes the reliability results using test-retest reliability and classification consistency. Data were collected on 718 MTurkers who completed the Personalysis assessment across two time points. 10 of the 12 (83%) Personalysis scales demonstrated adequate test-retest reliability. Further, all 12 Personalysis scales demonstrated adequate classification consistency scores. Overall, the results suggest that all 12 Personalysis scales demonstrate adequate reliability.

Introduction

Reliability is an essential component of psychological assessments. It is the measure of a psychological scale's consistency. Reliability is the interpretation of the error associated with a psychological scale. As error decreases, a scale's reliability increases. Furthermore, as reliability increases the confidence in the scale scores also increases.

The reason why reliability is a large component of the effectiveness of a test is that psychological measures must have reliability to predict the construct they purport to measure. If a psychological measure lacks reliability, then it will be unable to predict what it purports to measure (Livingston, 2018). A test that produces scores that are dissimilar across or within testing conditions lacks consistency. Thus, if scale scores are inconsistent, then there is no way to know if scale scores are predictive of what they purport to measure.

Alpha reliability is the most commonly used estimate to examine the internal consistency of a psychological assessment. Although internal consistency is typically reported for personality assessments, ipsative personality measures such as Personalysis do not allow for interpretation of internal consistency estimates (Bartam, 1996; Hicks, 1970; Johnson et al., 1988). While alpha is not an appropriate reliability estimate when evaluating forced-choice personality assessments using ipsative scoring methods, in the interest of completeness, we calculated alpha reliability coefficients for each scale. The median alpha was .72, and seven scales had alphas above the minimum threshold of .70.

For the purposes of Personalysis, two types of reliability are reported to evaluate the overall reliability of the assessment. First, test-retest reliability measures the temporal stability of a psychological assessments across two time points. A test-retest reliability statistic correlates scale scores from time one to time two. Second, classification consistency is used in addition to the test-retest reliability statistic. Similar to test-retest reliability, classification consistency compares scores across two time points, though classification consistency interprets a dichotomous coding of whether an individual is either a style (e.g., Preferred Red) or not a style (e.g., Not Preferred Red). Personalysis uses a dichotomous scoring system to indicate whether a person is a style or not. Thus, classification consistency is the most appropriate statistic to interpret for reliability, though test-retest reliability is also reported.

Test Retest Reliability

Test-retest reliability is commonly employed to estimate the temporal stability of a measure (Coaley, 2010). Consider the example of a scale to measure

one's weight. One day the individual steps onto a scale at time one and the scale reads 200 pounds, one minute later the same individual steps onto the scale and the scales reads 100 pounds. This would be a clear indication that this scale is unreliable and it is difficult to interpret the scale's results because of the inconsistency across test administrations. Test-retest reliability is similar in the way that it measures psychological constructs rather than weight. It is calculated by correlating scale scores from time one to time two. If scores on the assessment from time one to time two stay relatively the same, this provides evidence the scale demonstrates high reliability.

A threshold of .60 is the standard to determine adequate test-retest reliability coefficients (EFPA, 2013). It is common for personality scales, as opposed to other psychological assessments (e.g., cognitive ability), to have reliabilities in the range of .60 to .80 due to their broad bandwidth. In addition to personality scales having lower reliabilities generally, ipsative personality measures tend to demonstrate lower reliabilities than their Likert-type counterparts (Bartram, 1996; Seybert & Becker, 2019; Viswesvaran & Ones, 2000). Test-retest reliability estimates within the current study may underestimate Personalysis' reliability estimates.

Classification Consistency

Classification consistency measures how often an individual placed into a style (e.g., Preferred Red) at time one is classified as the same style at time two. Classification consistency is the most appropriate method to estimate the reliability of Personalysis scales. This is because, in applied practice, interpretation and feedback to Personalysis test-takers centers on their style classifications rather than the specific scores.

If scores were produced randomly for each Personalysis scale we would expect a classification consistency score around 50%. Any score above 50% indicates some degree of consistency where an individual is more likely to be categorized as the same style from time one to time two than chance alone.

Although there is no single standard to interpret reliability estimates, it is difficult to understand reliability without a baseline. Table 1 presents a guideline to interpret each reliability estimate to provide readers an idea what estimates are acceptable for Personalysis, an ipsative personality assessment. Test-retest standards are taken from the European Federation of Professional Psychologists Associations (EFPA, 2013). Classification consistency standards are taken from the Altman (1991) by converting kappa agreement into percentages.

Table 1

Guidelines to Interpret Reliability Estimates

	Test-retest	Classification Consistency
Excellent	.80	90%
Good	.70	80%
Adequate	.60	70%

Method

Sample

The sample consisted of 718 participants (from MTurk) selected to represent the population of working adults by gender, ethnicity, race, age, and educational level. Data were collected using MTurk, a widely used crowdsourcing tool. Data collection procedures are described in detail in Technical Report #2: Methodology of Personalysis Studies including inclusion/exclusion criteria, data quality controls, and data cleaning procedures. The demographics of the final sample provide a reasonable representation of the Personalysis target population. Study demographics are as follows:

Table 2

Demographics of MTurk Sample

Gender	*n*	%	Census estimates
Female	332	46%	47%
Male	370	52%	53%
Other*	16	2%	n/a
Ethnicity			
Hispanic	83	12%	12%
Non-Hispanic	634	88%	88%
Race			
American Indian or Alaska Native	2	<1%	n/a
Another race	3	<1%	n/a
Asian or Asian American	58	9%	8%
Black or African American	78	12%	12%
Two or more races	27	4%	n/a
White or Caucasian	476	74%	77%
Age			
20-24	42	6%	9%
25-54	554	77%	66%
55-64	88	12%	18%
over 65	34	5%	7%
Educational Attainment			
Less than high school or other	4	1%	8%
High school graduates	44	6%	23%
Some college, associate's degree, or vocational training	250	35%	29%
Bachelor's degree or higher	420	58%	41%

Notes. *Other denotes non-binary/non-conforming (*n* = 12), transgender man (*n* = 3), and transgender woman (*n* = 1). *N* = 718.

Instruments

All participants completed the Personalysis assessment instrument twice after an average retest interval of 13 days. The characteristics of this instrument are described in Technical Report #2: Methodology of Personalysis Studies.

Procedures

All analyses were conducted using the R Statistical Software (v4.2.2; R Core Team 2022). There were three approaches to conduct the following analysis. First, test-retest was calculated by correlating time one Personalysis scale scores with time two Personalysis scale scores. Second, each Personalysis user was given a style designation using Personalysis scoring procedure. Third, Personalysis users that fell within a certain range were classified as a style, and then examined to determine whether they were the same style at time two. The package "stats" (v2.2.9; Revelle, 2022) was used to conduct the test-retest reliability analysis.

RESULTS

Test-Retest Estimates of Personalysis Scales

The current section reports the test-retest reliability estimates of all 12 Personalysis scales. A within-subjects approach was used for test-retest between scores at time one to scores at time two. As stated in Table 1, a .60 test-retest reliability coefficient is the minimum standard to demonstrate adequate test-retest reliability for personality scales.

Red Scales

All three Red scales demonstrate adequate test-retest reliability estimates above .60. The Preferred Red scale demonstrated good test-retest reliability at .70.

Yellow Scales

Two of the three (67%) Yellow scales demonstrated at least adequate test-retest reliability. Preferred Yellow demonstrated good test-retest reliability at .79, but, almost met the threshold for excellent test-retest reliability. Social Yellow demonstrated excellent test-retest reliability at .80. Yellow Instinctive was just short of the threshold for adequate reliability at .57.

Blue Scales

All three Blue scales demonstrated good test-retest reliability. Social Blue demonstrated good test-retest reliability, but was close to demonstrating excellent test-retest reliability at .79.

Green Scales

Two of the three (67%) Green scales demonstrated at least good test-retest reliability. Social Green demonstrated excellent test-retest reliability at .80. Social Green demonstrated good test-retest reliability (.79), but was close to meeting the threshold to demonstrate excellent test-retest reliability. Green Instinctive was below the threshold at .54.

Overall, 10 of the 12 (83%) Personalysis scales demonstrated adequate test-retest reliability. 8 of the 12 (67%) Personalysis scales demonstrated good test-retest reliability. 2 of the 12 (17%) scales demonstrated excellent test-retest reliability, while three scales were close to meeting the threshold to demonstrate excellent test-retest reliability. Two scales (17%) failed to meet the minimum threshold.

Table 3
Test-Retest Reliability of Personalysis Colors by Dimension

	Test-Retest
Red	
Preferred	.70
Social	.62
Instinctive	.68
Yellow	
Preferred	.79
Social	.80
Instinctive	.57
Blue	
Preferred	.77
Social	.79
Instinctive	.71
Green	
Preferred	.80
Social	.79
Instinctive	.54
Median	.74

Note. *N* = 718.

Standard Classification Consistency Estimates of Personalysis Scales

The current section reports the standard classification consistency estimates across two time points. Standard classification consistency is based upon individuals who score 3.0 or higher being labeled as a style. We based classifications on scores of 3.0 or higher because this is the threshold used in the computer generated Personalysis feedback reports. If an individual is labeled the same style across both time points, then this would result in a consistent classification, thereby increasing the classification consistency score.

As stated in Table 1, a classification consistency score of 70% is considered adequate.

Red Scales

All three Red scales demonstrated adequate or better classification consistency. Red Social demonstrated a good classification consistency estimate at 86%.

Yellow Scales

All three Yellow scales demonstrated at least adequate classification consistency. Both Social and Instinctive Yellow demonstrated good classification consistency estimates, while Preferred Yellow was close to meeting the threshold to demonstrate good classification consistency at 79%.

Blue Scales

All three Blue scales demonstrated good classification consistency.

Green Scales

All three Green scales demonstrated at least adequate classification consistency. Both Preferred and Social Green demonstrated good classification consistency, while Instinctive Green was close to meeting the threshold of good classification consistency at 79%.

Overall, all 12 Personalysis scales demonstrated adequate or higher classification consistency. 8 of the 12 (67%) Personalysis scales demonstrated good classification consistency, and two additional scales (17%) were close to meeting the threshold of good classification consistency.

Table 4
Classification Consistency of Personalysis Colors by Dimension

Red	
Preferred	78%
Social	86%
Instinctive	76%
Yellow	
Preferred	79%
Social	86%
Instinctive	84%
Blue	
Preferred	80%
Social	81%
Instinctive	85%
Green	
Preferred	81%
Social	81%
Instinctive	79%
Median	81%

Note. *N* = 718.

Classification Consistency Estimates of Personalysis Scales Using Strong Scores

The current section reports the classification consistency estimates across two time points using a higher score threshold to determine participants style. Anyone who scored 4.0 or higher (max score of 6) was labeled as a style, anyone who scored at 1.5 or lower was labeled as an absence of the style. This threshold was examined because the computer generated Personalysis feedback reports provide stronger feedback about each personality style when scores are higher.

Overall, all 12 Personalysis scales demonstrated at least good classification consistency. 11 of the 12 (92%) Personalysis scales demonstrated excellent classification consistency when using a higher score threshold.

Table 5

Classification Consistency by Strong Personalysis Styles

High scores	4.0
Low scores	1.5
Red	
Preferred	90%
Social	96%
Instinctive	85%
Yellow	
Preferred	90%
Social	93%
Instinctive	97%
Blue	
Preferred	90%
Social	92%
Instinctive	93%
Green	
Preferred	91%
Social	91%
Instinctive	90%
Median	91%

Note. Average *N* = 404.

Discussion

Two metrics were used to evaluate the reliability of each Personalysis scale. First, each scale was evaluated for test-retest reliability. 10 of the 12 (83%) Personalysis scales demonstrated adequate test-retest reliability. 8 of the 12 (67%) Personalysis scales demonstrated good test-retest reliability. These findings are encouraging in view of the literature noting generally lower retest reliabilities with ipsatized personality scales (Bartram, 1996; Seybert & Becker, 2019; Viswesvaran & Ones, 2000).

Second, classification consistency was used as this method best captures the way the Personalysis scales are interpreted in practice. All 12 Personalysis scales demonstrated adequate classification consistency. 8 of the 12 (67%) Personalysis scales demonstrated good classification consistency, while an additional two Personalysis scales demonstrated marginally good consistency.

Another classification analysis examining stronger Personalysis scores was used to examine classification consistency. 11 of the 12 (92%) Personalysis scales demonstrated excellent classification consistency when using a stronger score as the threshold for classification. These findings further support that Personalysis score interpretations are reliable, and the reliability of Personalysis scales increases as scores become more extreme at the high or low end of the scale. This is important because the feedback given to test takers is stronger when they obtain higher scores.

In conclusion, there was adequate evidence to establish reliability for all 12 Personalysis scales when considering test-retest and classification consistency together.

References

Altman, D.G. (1991). *Practical Statistics for Medical Research*. Chapman & Hall.

Bartam, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology, 69*(1), 25-39. https://doi.org/ 10.1111/j.2044-8325.1996.tb00597.x

Coaley, K. (2010). *An Introduction to Psychological Assessment and Psychometrics*. Sage.

European Federation of Psychologists' Associations. (2013). *EFPA review model for the description and evaluation of psychological tests*. EFPA Board of Assessment.

Hicks L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-184. https://doi.org/10.1037/h0029780

Johnson, C. E., Wood, R. & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153-162.

Livingston, S. A. (2018). *Test Reliability-Basic Concepts*. Educational Testing Service.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Revelle, W. (2021). *psych: Procedures for psychological, psycho- metric, and personality research* (Version 2.2.9) [Computer software]. https://cran.r-project.org/web/packages/psych/psych.pdf

Seybert J., Becker D. (2019). Examination of the test-retest reliability of a forced-choice personality measure. *ETS Research Report Series,* 1, 1–17. https://doi.org/10.1002/ets2.12273

Viswesvaran, C. & Ones, D. S. (2000). Measurement Error in "Big Five Factors" Personality Assessment: Reliability Generalization across Studies and Measures. *Educational and Psychological Measurement, 60*(2), 224–235. https://doi.org/10.1177/00131640021970475

Appendix

Table 6

Personalysis Scales Supported by Adequate Interpretation of Reliability Estimates

	Test-Retest*	Classification Consistency	
Red		Standard*	Strong**
Preferred	Supported	Supported	Supported
Social	Supported	Supported	Supported
Instinctive	Supported	Supported	Supported
Yellow			
Preferred	Supported	Supported	Supported
Social	Supported	Supported	Supported
Instinctive	Unsupported	Supported	Supported
Blue			
Preferred	Supported	Supported	Supported
Social	Supported	Supported	Supported
Instinctive	Supported	Supported	Supported
Green			
Preferred	Supported	Supported	Supported
Social	Supported	Supported	Supported
Instinctive	Unsupported	Supported	Supported

Notes. * denotes *N* = 718. ** denotes average *N* = 404.