The reliability and validity of psychometric instruments

May 17, 2024

This whitepaper is written in collaboration with Jüri Allik, a Professor of Experimental Psychology, Tartu University. He is a Foreign Member of the Finnish Academy of Science and Letters[1], Academia Europaea[2], and the Estonian Academy of Sciences[3]. His works have been cited in 29016 publications[4].

Evaluating the quality of psychometric instruments

All good instruments for measuring personality traits are long. For example, MMPI-3 (*Minnesota Multiphasic Personality Inventory* version 3) consists of 567 questions. The personality questionnaire NEO-PI-R/3 (*NEO Personality Inventory Revised* or version 3) includes 240 questions. The questionnaires measure several traits at the same time (for example, NEO-PI-R/3 measures 30 different personality traits that are divided into five clusters or dimensions. There are several parallel questions with approximately the same meaning for a reliable measurement of each trait. For example, in NEO-PI-R/3-s, each trait is measured with a sub-scale that has eight parallel questions. There is a good reason for the number of items because they are needed for the consistency and authenticity of these measurements.

Each measurement instrument—typically a scale consisting of multiple questions or items— is characterized by two main properties: **reliability** and **validity**.

Reliability

Reliability characterizes how consistently something is measured or, in other words, how small is the measurement error. For example, if we want to evaluate the reliability of a bathroom scale, we might weigh ourselves two or three times. If the result is different every time, then the scatter of the results shows the measurement error. If the variance - a measure of scatter - is small or doesn't exist at all, we can conclude that the scale is reliable, giving us the same results on each occasion.

Similarly, the best way to test the reliability of psychological instruments is repeated measurement. This is called ***test-retest reliability***. If the measurement is repeated in a couple of days or weeks, then we must assume that the trait we intended to measure has not changed during this relatively short time interval. For example, repeated testing with an average 2-year interval resulted in an average correlation of .66 for the 30 subscales of NEO-PI-R-i (Kallasmaa, Allik, Realo, & McCrae, 2000, Table 1). Another example, the Rosenberg's Self-Esteem Scale (RSES), which consists of 10 questions. The two-week test-retest reliability of this instrument was .84 (Pullmann & Allik, 2000, Table 2). Approximately the same test-retest reliability can be assumed for NEO-PI-R/3 subscales if the re-testing interval is no longer than one or two weeks.

As repeated testing is time-consuming, a more convenient method is to calculate a scale's ***internal consistency***. The best-known formula is ***Cronbach's alpha***, which is the mean correlation between questions or items (Cronbach, 1951). The higher the correlation between items, the higher reliability of this instrument. For example, the average Cronbach's alfa for the NEO-PI-3 subscales is .76 (Costa & McCrae, 1992, Table 5). There is the Spearman-Brown prediction (also prophesy) formula relating Cronbach alpha to test length used to predict the reliability of a test after changing the test length. This formula tells us that there is no better option for the increase of scale reliability than increasing the number of items.

If this alpha value of .76 looks modest for someone, then it is necessary to warn that higher reliability can be attained often by artificially bloating up the scale. For example, it can be done by including items with very similar meanings. Thus, it might be good advice to avoid both very low and high alphas.

Validity

Validity characterizes how well a scale measures what was ***intended*** to measure. It is possible to have a scale with high reliability (very small measurement error) but low validity. Thus, an exemplary Cronbach alpha is not a guarantee that we are measuring something that was required or planned.

For instance, if a bathroom scale shows the same measure every time, but the result is considerably different than the actual weight, then it is measuring something else with high precision. However, in the world of weighing machines, we can always buy a new scale which satisfies our demands. It is much more complicated with a psychological measurement where we do not have explicit units of measurement (think about the "unit" of neuroticism, for instance) or etalons. In practice, we are compelled to use incomplete solutions relying on the best scale that is currently available for a measure of a trait we are interested in. For example, if someone creates a new personality measure, the results can be compared to NEO-PI-R or its latest version, NEO-PI-3 (McCrae, Costa, & Martin, 2005). In the textbooks, it is typically called ***concurrent validity***.

However, the lack of objective standards or units of measurement can be compensated in several ways. For example, knowing that the measurable trait causes certain phenomena, we can check if these consequences co-exist with measured traits. This is called ***predictive validity***. For example, it is known that depression (the neuroticism subscale, N3: Depression in NEO-PI-R) predicts suicidal thoughts and attempts. Consequently, the predictive validity of N3-e would assume a correlation between the intensity of suicidal thoughts and the high number of suicidal attempts among people who rate high in N3: Depression.

Another form of predictive validity is an ***agreement between the ratings*** of two judges, usually a person and somebody who knows her or him well. If a person's self-ratings allow us to accurately predict the ratings of another informant—a relative, partner, colleague, or friend—then this is one of the best validity indicators. The agreement between self-other ratings is the necessary condition for assuming that we are dealing with the same existing characteristic, not with phantasies in somebody's head. Agreement or consensus between judges is one of the best criteria for the validity of personality judgements (Allik, De Vries, & Realo, 2016). Although the agreement between two opinions is not guaranteed that they are true - it is possible to agree on what is totally untrue—agreement in the description of personality is very often a sign of their reality.

If the situation with the validity of psychological measurements looks desperate, it is useful to remind the history of the length measurement. The first measures of length were body parts like the foot or forearm. To level the individual differences, the average of measures of twenty randomly selected persons was typically used. An important step to eliminating subjectivity was defining the meter as a part of the meridian (one-tenth of the millionth part from the length of a meridian from the equator to the North Pole). The standard platinum rod of the
equivalent length is still held in Paris. The latest way for defining a unit of length is by measuring the time it takes for light to cover this distance in the vacuum. As we still do not have the standards for neuroticism or extraversion, human psychology is still somewhere on the level of the length measurement when feet (about 0.3 m) and cubits (0.533 m) were the main units of length.

Until we haven't invented the standards yet, it is most reasonable to examine all measurement results with common sense and decide if these are in accord with our understanding of the phenomena we want to measure. If the results are consistent and they fit well to our best knowledge, then we may have a modest level of confidence that we succeeded in measuring what was desired and planned.

Quality of instruments used on the Wisnio platform

The science of measuring different psychometric characteristics of individuals and evaluating the quality of such measures has not changed significantly in decades, and well-documented and comprehensively researched instruments are publicly available.

In spite of that, the usefulness of such instruments in predicting work-related outcomes has been modest at best (Zell, Lesick et al., 2021). In our experience, the main limiting factor is not that the available instruments are of low quality (although this is often the case) but rather how these instruments are used.

Leadership assessment is often the final element of the hiring process, and in most cases, only the chosen candidate is asked to participate in an assessment centre. The

outcome of that assessment centre describes certain characteristics of the candidate and sometimes points out a few potential risk factors or personal development areas. In most cases, those findings are in no way connected to the job context or the characteristics of the team they will be working with. As the hiring decision is typically already made by that time, this information might help with onboarding and development in the future but typically has little to no impact on the quality of the hiring decision.

Considering the above, our focus has not been on trying to incrementally improve the quality of psychometric instruments but rather on making the data useful and actionable for practical decision-making purposes. Therefore, Wisnio uses publicly available and extensively researched and validated psychometric instruments - Shalom Shwartz's Questionnaire of Universal Human Values and the IPIP120 Personality Inventory.

Reliability of the instruments

The values inventory (based on Schwartz's refined values theory) has a mean Cronbach's alpha reliability coefficient of 0.70 (SD = 0.08) measured across 49 cultural groups. For the four higher-order values, the mean Cronbach's alpha reliability coefficient is 0.84 (SD = 0.03) (Schwartz & Cieciuch, 2022).

The personality inventory is based on IPIP120, which is a shortened and simplified version of the NEO PI-R. Cronbach's alpha reliability coefficients of the 30 personality facets range from 0.63 to 0.88, with three facets with a reliability coefficient of less than 0.7 (Johnson, 2014). For the five main dimensions, alphas range from 0.87 to 0.90, with a median of 0.88 (Maples, Miller, Carter, 2014).

Validity of the instruments

IPIP120 is designed to approximate NEO PI-R and measure largely similar constructs with a simplified instrument. Correlations between IPIP120 and NEO PI-R facets are over 0.66(Johnson, 2014). Correlations between self-other ratings of IPIP-NEO scales (0.38 to 0.58) indicate that IPIP-NEO scales are working as well as the original, longer scales (Johnson, 2014). Detailed comparison tables between IPIP-NEO and other popular instruments based on the Five Factor Model (including Hogan HPI, HEXACO, MPQ, NEO PI-R) can be found on the International Personality Item Pool database (https://ipip.ori.org/, 21.02.2022).

The relationship between Big Five personality characteristics and different aspects of job performance and work-related behaviours has been studied extensively, and meta-analysis consistently indicate that each Big Five trait is a valid correlate of performance (Zell & Lesick, 2021). Similarly, hundreds of primary studies and several meta-analyses have been conducted on team effectiveness and how personality diversity and values

similarity impacts team processes, engagement, and outcomes. A comprehensive literature review on the topic is published by Mathieu et al. (2008).

Salom Shwartz's theory of Universal Human values has been repeatedly validated in numerous controlled studies in over 80 countries (Schwartz, 2017). Over the 40 years since the publication of the original theory, it has been cited in more than seventeen thousand scholarly articles. Various validation methods are used for this, such as importance ratings of values, direct similarity judgment tasks, pile sorting, and spatial arrangement and even for how the values of other people, such as family members, are perceived.

The value theory makes two claims to universality. First, people in all cultures recognize the same set of basic values. Second, these values form the same circular motivational continuum in all cultures (Schwartz, 2017). It has also been shown that people can assess other people's values whom they know well accurately (Dobewall et al., 2014). Self-other agreement in four higher-order values (median r = .47) and six culture-specific value factors (median r = .50) was proved substantial. Therefore, other ratings of personal values can be used to validate and complement self-report value measures.

Summary

The psychometric qualities of instruments incorporated on the Wisnio platform are of high quality - their validity and reliability are proven in numerous controlled studies, which are publicly available. Based on that, we can ensure that the instruments measure the characteristics they intend to measure and that they do so consistently. Based on numerous meta-analyses, we know that the characteristics evaluated by Wisnio help to predict performance, tenure, work engagement, and several aspects of team cohesiveness and collaboration. The strength of those relationships - meaning the predictive validity of such instruments - depends mainly on how the data is used. Measuring irrelevant details with a high degree of accuracy is less useful than developing a robust understanding of relevant data. Due to that, Wisnio is designed to ensure that psychometric data is viewed in the context of a specific team and job and that the information is used constructively.

References:

1. Allik, J., De Vries, R. E., & Realo, A. (2016). Why are moderators of self-other agreement difficult to establish? *Journal of Research in Personality, 63*, 72-83.

2. Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, Fl.: Psychological Assessment Resources.

3. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

4. Dobewall, H. et al (2014). A comparison of self-other agreement in personal values versus the Big Five personality traits. Journal of Research in Personality, 50.

5. International Personality Item Pool, https://ipip.ori.org/, 21.02.2022.

6. Johnson, J. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. Journal of Research in Personality, 51.

7. Kallasmaa, T., Allik, J., Realo, A., & McCrae, R. R. (2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the Five-Factor Model. *European Journal of Personality, 14*, 265-278.

8. Maples, J.; Miller, J.; Carter, T. (2014). A Test of the International Personality Item Pool Representation of the Revised NEO Personality Inventory and Development of a 120-Item IPIP-Based Measure of the Five-Factor Model. Psychological Assessment, 26-4.

9. Mathieu et al (2008). Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future. Journal of Management, 34-3.

10. McCrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO-PI-3: A More Readable Revised Neo Personality Inventory. *Journal of Personality Assessment, 84*(3), 261-270.

11. Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences, 28*, 701-715.

12. Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. Online Readings in Psychology and Culture, 2(1).

13. Schwartz, S.H. (2017). The Refined Theory of Basic Values. In: Roccas, S., Sagiv, L. (eds) Values and Behavior. Springer.

14. Schwartz, S., Cieciuch, J (2022). Measuring the Refined Theory of Individual Values in 49 Cultural Groups: Psychometrics of the Revised Portrait Value Questionnaire. Assessment 29-5.

15. Zell, E.; Lesick, T. (2021). Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. Journal of Personality, 2021 1-15.

---

[1] https://en.wikipedia.org/wiki/Finnish_Academy_of_Science_and_Letters

[2] https://en.wikipedia.org/wiki/Academia_Europaea

[3] https://en.wikipedia.org/wiki/Estonian_Academy_of_Sciences

[4] https://scholar.google.com/citations?user=xaxV8_8AAAAJ&hl=en