Psychometrics

Susanna Fullmer & David Daniel

Psychometrics | Item Response Theory | Measurement | Reliability | Validity | Generalizability Theory

Measurement is the process of assigning numbers to attributes. As a society, we take measurements of almost everything—weight, height, temperature, the speed our car is going, the amount of time using our phones, and the amount of money in our bank accounts are a few examples. For tangible, physical attributes, measurement is easy and straightforward—a number is arrived at. But have you ever considered how we have come to measure happiness, intelligence, or confidence? These traits are what psychometricians call latent variables, or variables that cannot be physically or directly measured. The ability to measure these variables accurately and reliably is the focus of the field of psychometrics. As the field of psychometrics developed, tools like Item Response Theory (IRT) and Classical Test Theory (CTT) have increased our ability to measure latent variables.

History and Development of Psychometrics

Understanding the history of psychometrics helps us appreciate some of the difficulties in measuring these latent variables. The roots of psychometrics can be traced back to the late 1700s, but most modern psychometric methods were developed in the ensuing 100 years. The field of psychometrics largely developed from the fields of psychology and statistics. Eventually, psychometric principles were applied to education. We will explore a few of the major contributors and their contributions to modern-day psychometrics.

Francis Galton is considered the father (sometimes grandfather) of behavioral science (Clausen, 2007). In the late 1800s, Galton was one of the first people to begin measuring and investigating differences in human traits (Jones & Thissen, 2007). He primarily focused on measuring physical traits and took measurements of thousands of individuals' characteristics. Eventually, he attempted to measure latent variables, which he viewed as mental traits. However, his major contributions include bringing statistical analysis to behavioral science. He was the first to use methods such as correlation (how related two variables are) and apply the normal distribution (a common, bell-shaped, symmetrical distribution, which a number of variables follow naturally) to understand the characteristics he measured and their relationships.

In the early 1900s, there was a shift to specifically testing intelligence. Two scientists, Alfred Binet and Theophile Simon began testing cognitive abilities. Their goal was to assign someone a "mental age" based on the results of a test (Jones & Thissen, 2007). Inspired by their work, Charles Spearman wrote the paper "'General Intelligence', Objectively Determined and Measured", which some view as the beginning of modern-day psychometrics. Spearman (1904) took Binet and Simon's research a step further by analyzing cognitive tests and assigning a general intelligence factor, which he designated as "g". Within a few years, Lewis Terman also took Binet and Simon's work and developed what we know today as the intelligence quotient score (IQ; Jones & Thissen, 2007).

Around the 1920s, psychometrics began to look like what we know today. This shift happened largely as Louis Thurstone began applying psychometrics to education. He wrote tests like the Psychological Examination for High School Graduates and College Freshmen. This test assigned two scores, one for linguistic skills and another for quantitative skills. He wrote other tests for many years, allowing him opportunities to develop different methods used today in the field of psychometrics. These methods include multiple factor analysis and test theory. He also brought to light the ideas of reliability and validity, emphasizing their importance in psychometric testing (Jones & Thissen, 2007). Although it took time, these methods and theories are the tools that psychometricians use today.

The “How” of Psychometrics

Currently, the two renowned tools of psychometrics are generalized to classical test theory (CTT) and item response theory (IRT). However, the field is more nuanced than that. Other methods are intertwined with CTT and IRT, such as generalizability theory and factor analysis, which we will explore here. However, understanding CTT and IRT first requires an understanding of psychometric fundamentals: building a model, reliability, and validity.

Building a Model

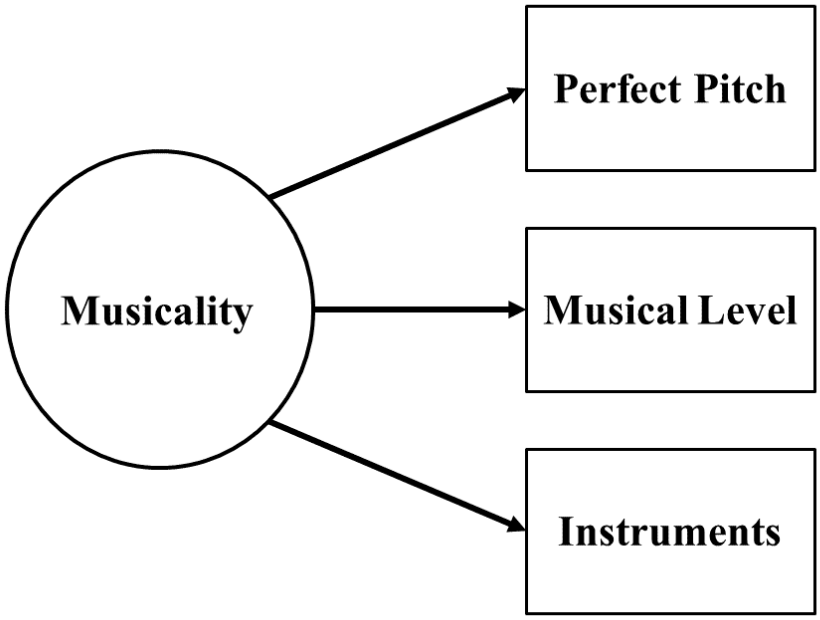
As previously mentioned, psychometrics studies the measurement of latent traits. Measuring the physically immeasurable can be done by assessing multiple related variables, called indicator variables, that can be objectively measured. Depending on the field of study, the composite of indicator variables is referred to as an instrument, model, scale, test, assessment, or questionnaire. Hence, we will use these terms interchangeably. To illustrate creating an instrument, we refer to a fictional instrument for measuring musicality that uses the following indicator variables:

1. Do you have perfect pitch? (yes/no)
2. Which word best classifies your musical level? (amateur, casual, experienced, professional)
3. How many instruments do you play?

We could diagram our model as shown in Figure 1. In the basic format for drawing models, referred to as path diagrams, latent traits are always represented by circles, and indicator variables are represented in box shapes (Wang & Wang, 2012). Note that the arrows point from the latent trait to the indicator variables. The theory is that the latent trait is what determines how subjects respond to the indicator variables.

Figure 1

Path Diagram of a Fictional Musicality Instrument



Reliability and Validity

Reliability is a specific term in test theory that means the results are consistent. A respondent taking a reliable test multiple times would result in scores that are similar. Reliability can refer to different aspects, such as “consistency over a period of time, over different forms of the assessment, within the assessment itself, and over different raters” (Miller et al., 2013, p. 110). In this chapter, we focus on the internal consistency aspect, in which case, reliability is measured using correlation. Correlation is the assignment of a value, called the correlation coefficient, that measures the relationship between two variables.

The ideal calculation for reliability requires correlating two scores from the same respondent under the exact same conditions, including time. Since replicating conditions is implausible, we use estimations instead (Miller et al., 2013). For instance, a commonly-used estimation is Cronbach’s alpha coefficient, which comes from splitting the test in half, resulting in two tests taken under the same conditions. The scores of one half are then correlated with the scores of the other half. That process is then repeated for all possible test-half combinations, and the average of all those correlation coefficients is the Cronbach’s alpha coefficient (Wu et al., 2016).

While reliability is vital, it is not sufficient without validity. Validity assesses how adequately the test measures the intended latent trait. Imagine throwing darts. It is not enough for darts to land within centimeters of each other if the darts land far from the target. In psychometrics, darts landing close to each other are akin to reliability, while landing on the target represents validity; both are needed to succeed. Like reliability, validity exists on a spectrum of high and low. Unlike reliability, validity is determined more by evaluative judgments rather than calculated values. According to Miller et al. (2013), validity is evaluated based on the consideration of content, construct, criterion, and consequences.

- Tests include only a sample of all possible indicator variables. Content validity measures the representativeness of that sample. In the musicality instrument from Figure 1, removing the first two questions would decrease content validity. Asking about perfect pitch, musical level, and instruments provides a fuller picture of musicality than just instruments alone.
- Construct validity refers to the relevance of the indicators to the latent trait. For instance, a psychological questionnaire measuring depression probably does not need a question about ice cream preference. Likewise, a math test with word problems can unintentionally measure English prowess. Indicator variables should be well thought out to ensure high construct validity.
- Criterion validity requires comparing test results to a standard. A school teacher might calculate the correlation between their students' test results and the national average. The higher the correlation with a trusted source, the higher the criterion validity.
- The last consideration is consequence validity, which is a subjective judgment on whether the test's consequences are overall beneficial or harmful. For example, a high stakes standardized test might lead to student burn out but appropriately measure students' compatibility with prospective colleges. In this case, consequence validity would be the judgment made of whether the benefits of compatibility outweigh the disadvantages of burn out.

Test Theory

A common misconception among researchers is that IRT and CTT are interchangeable, when they actually "provide complementary results" (Wu et al., 2016, p. 74). While CTT focuses on the reliability of the results, IRT focuses on the relationship between the items and the latent trait. With the knowledge of the fundamentals, these differences can be explored in some depth.

Classical Test Theory

The foundation of CTT is that the observed score from the instrument, that measures the latent trait, is made up of a respondent's true score and random errors. Written as an equation, that is:

$$X = T + E$$

where X is the observed score, T is the true score, and E represents random error. Random error refers to controllable errors and errors due to chance. Since the true score will always be unknown, instruments can only estimate true scores, and the accuracy of the estimates can be evaluated using reliability measures.

Wu et al. (2016) showed mathematically that measuring reliability estimates the correlation of observed scores and true scores. Ideally, reliability should have a high, positive correlation meaning that as an observed score increases, the true score also increases.

Generalizability Theory

Reliability and validity are central components to CTT. However, generalizability theory, referred to as the daughter of CTT, offers an improved conceptualization of reliability and validity. Consequently, more modern approaches are shifting to the generalizability theory (Prion et al., 2016). Generalizability theory expands on Equation 1 by adapting how the errors are used. While CTT lumps all errors together, generalizability theory isolates each error transforming the equation to be more like the following (Prion et al., 2016):

$$X = T + E_1 + E_2 + E_3$$

For example, E1 could represent spelling errors, E2 could represent biased grading, while E3 could represent respondents misreading answer choices. In Equation 2, only three errors are listed for simplicity but there can be any number of errors.

Item Response Theory and Factor Analysis

Recall that instruments are made up of indicator variables. Because the relevance and priority of indicator variables can be subjective, IRT and factor analysis are used to quantify each indicator's contribution. These methods of analysis are closely related to each other. In fact, some researchers consider IRT a subcategory of factor analysis while others see them as two separate forms of analysis that merely intersect like a Venn diagram. Focusing on the Venn diagram analogy, the two methods intersect in their purpose, but what falls outside the intersection (i.e., their differences) is in their calculations and data restrictions (Jones & Thissen, 2007; Groenen & van der Ark, 2006).

The complexity of the calculations is beyond the scope of this chapter; however, an overview is that IRT calculations are based on probabilities, whereas factor analysis calculations are based on covariance, or a measurement of how much the items are related to each other (Jones & Thissen, 2007; Wang & Wang, 2012). Factor analysis and IRT also differ in the type of data that can be used. Notice that the questions in the musicality model have different answer options.

1. The first question has "yes" or "no" answer options, which produces dichotomous data.
2. The second question gives limited categories, which produces categorical data, also referred to as nominal or polytomous data. When the order of categories matters (e.g., categories of "low", "medium", and "high"), the data is ordinal.
3. The third question is not limited by categories. Instead, the possible values are endless, which produces continuous data.

While IRT is limited to models with dichotomous and categorical data, factor analysis can use all three types of data mentioned. However, the verdict is still out on which method produces more accurate estimations. Maydeu-Olivares et al. (2011) noted that the accuracy of estimations is indistinguishable between IRT and factor analysis when using dichotomous data. However, by most standards, IRT is more accurate than factor analysis with ordinal data.

Conclusion

The measurement of physical characteristics is, for the most part, straightforward. The difficulty comes in measuring latent variables. Many psychologists and statisticians attempted to measure these variables leading to the formation of the field of psychometrics. In the field of psychometrics, factor analysis and IRT are two tools that have been developed to help determine which indicator variables influence the latent variables. In addition, classical test theory and generalizability theory help to confirm the reliability of a test. These basic concepts help build the foundation for the field of psychometrics, but the field has much more depth and nuance, which we invite you to explore on your own.

References

Clauser, B. E. (2007). The life and labors of Francis Galton: A review of four recent books about the father of behavioral statistics. Journal of Educational and Behavioral Statistics, 32(4), 440–444. doi:10.3102/1076998607307449.

Groenen, P. J. F., & van der Ark, L. A. (2006). Visions of 70 years of psychometrics: The past, present, and future. Statistica Neerlandica, 60(2), 135–144. doi:10.1111/j.1467-9574.2006.00318.x

Jones, L. V., & Thissen, D. (2007). A history and overview of psychometrics. Psychometrics, 26, 1–27. doi:10.1016/s0169-7161(06)26001-2

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. Structural Equation Modeling: A Multidisciplinary Journal, 18(3), 333–356. doi:10.1080/10705511.2011.581993

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). Reliability. In J.W. Johnston, L. Carlson, & P.D. Bennett (Eds.), Measurement and assessment in teaching (11th ed., pp. 108-137). Pearson.

Miller, M. D., Linn, R. L., & Gronlund, N. (2013). Validity. In J.W. Johnston, L. Carlson, & P.D. Bennett (Eds.), Measurement and assessment in teaching (11th ed., pp. 70-107). Pearson.

Prion, S. K., Gilbert, G. E., & Haerling, K. A. (2016). Generalizability theory: An introduction with application to simulation evaluation. Clinical Simulation in Nursing, 12(12), 546–554. doi:10.1016/j.ecns.2016.08.006

Spearman, C. (1904). "General intelligence," objectively determined and measured. The American Journal of Psychology, 15(2), 201-292. doi:10.2307/1412107

Wang, J., & Wang, X. (2012). Structural equation modeling. doi:10.1002/9781118356258

Wu, M., Tam, H. P., & Jen, T.H. (2016). Classical test theory. Educational Measurement for Applied Researchers, (pp. 73–90). Springer. doi:10.1007/978-981-10-3302-5_5.

Susanna Fullmer

Brigham Young University

David Daniel

Brigham Young University