RESEARCH · QUANTITATIVE AND MIXED METHODS

Reliability and Validity: The Twin Pillars of Measurement

What reliability and validity mean in research measurement, the main types of each, how they are assessed and reported, how they relate to each other, and common mistakes in using them.

ModQu research article · 2,455 words · 23 references (APA 7th edition)

Every measurement involves two questions. Does it give consistent results? And does it measure what it is supposed to measure? The first is the question of **reliability**; the second is the question of **validity**. Together they determine whether numbers produced by a survey, test, rating scale or instrument can be trusted. A measure that fails either test produces data that look precise but may mislead. This article explains what reliability and validity mean, the main forms each takes, how they are assessed and how they relate to each other.

Definitions

Carmines and Zeller (1979) defined reliability as the extent to which an experiment, test or measuring procedure yields the same results on repeated trials. A reliable bathroom scale shows the same weight if you step on it twice in a minute. Validity, in the classic definition, is the extent to which a measure captures the concept it is intended to measure. A scale that consistently shows a weight five kilograms too light is reliable but not valid.

Modern measurement theory has refined the definition of validity. The *Standards for Educational and Psychological Testing* define validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses (American Educational Research Association et al., 2014). In this view, validity is not a property of a test itself but of the inferences drawn from it. The same questionnaire may support valid conclusions about one group and one purpose but not another.

The relationship between reliability and validity

Reliability and validity are related but distinct. The diagram below uses the familiar picture of shots at a target to show the four combinations.

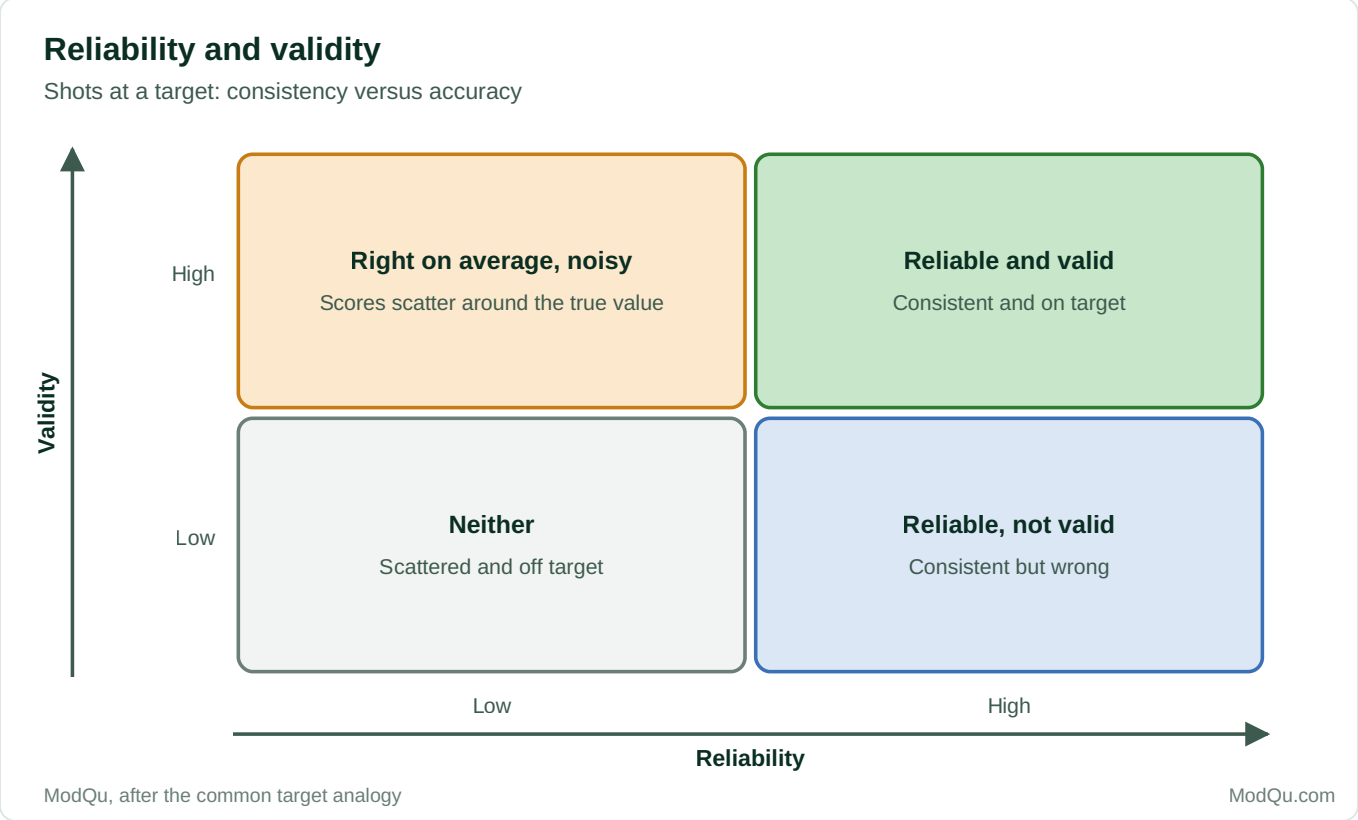


Figure 1. Reliability and validity shown as shots at a target. Strictly, high validity also requires reliability, so the top-left case describes a measure that is right on average but noisy. Diagram by ModQu.

A measure can be reliable without being valid: it consistently measures the wrong thing, or the right thing with a constant bias. A measure cannot be fully valid without being reliable, because random error limits how closely scores can reflect the underlying attribute. In classical test theory, reliability sets an upper limit on how strongly a measure can correlate with anything else (Nunnally & Bernstein, 1994). Spearman (1904) first showed that random measurement error weakens observed correlations, a problem now called **attenuation**, and proposed a correction for it.

Classical test theory

Much thinking about reliability rests on **classical test theory**, which treats every observed score as the sum of a true score and an error. The true score is the value a person would obtain if the measurement were perfect; the error is random noise. Reliability is the proportion of the variation in observed scores that reflects variation in true scores rather than error (Nunnally & Bernstein, 1994). A reliability of 0.80 means that 80 percent of the variation in scores reflects real differences between people, and 20 percent reflects error.

The theory assumes that errors are random, average to zero and are unrelated to true scores. These assumptions do not always hold. Systematic errors, such as a scale that always reads high or a question that some groups interpret differently, are not captured by reliability. They are threats to validity.

Types of reliability

Reliability can be assessed in several ways, each addressing a different source of error. The table below summarises the main types.

Types of reliability

Each addresses a different source of measurement error

Type	Question	Usual statistic
Test–retest	Same people, measured twice: same scores?	Correlation or intraclass correlation
Inter-rater	Different raters: same judgements?	Kappa; intraclass correlation
Internal consistency	Do the items measure the same thing?	Alpha; omega
Parallel forms	Two equivalent versions: same scores?	Correlation between forms

Based on Carmines & Zeller (1979); Cohen (1960); Shrout & Fleiss (1979); Cronbach (1951) ModQu.com

Figure 2. Four types of reliability and how each is assessed. Based on Carmines and Zeller (1979), Cohen (1960), Shrout and Fleiss (1979) and Cronbach (1951); table by ModQu.

Test–retest reliability compares scores from the same people measured twice, some time apart. High agreement shows that the measure is stable. The interval matters: too short, and people remember their answers; too long, and the attribute may genuinely change.

Inter-rater reliability compares judgements made by different observers or coders. It is essential when measurement depends on human judgement, such as coding interview transcripts or rating performance. Cohen (1960) proposed the **kappa** statistic, which measures agreement between two raters on categories while correcting for agreement expected by chance. Landis and Koch (1977) suggested benchmarks for interpreting kappa, describing values from 0.61 to 0.80 as substantial agreement and above 0.80 as almost perfect, though these labels are conventions rather than rules.

For continuous ratings, the **intraclass correlation coefficient** is used. Shrout and Fleiss (1979) described several forms, which differ depending on whether the same raters rate every case and whether the raters are considered a sample from a larger pool. Koo and Li (2016) provided guidance on choosing and reporting the correct form, noting that published studies often fail to state which was used.

Internal consistency measures whether the items in a multi-item scale measure the same thing. If a ten-question scale measures job satisfaction, people who are satisfied should tend to score high on most items. Cronbach (1951) introduced **coefficient alpha**, which became the most widely reported reliability statistic in social research. Tavakol and Dennick (2011) explained that alpha depends both on how strongly items correlate and on how many items there are, so a long scale can have a high alpha even when items are only weakly related.

Parallel forms reliability compares scores on two versions of a measure designed to be equivalent. It is used when the same test cannot be repeated, for example because people would remember the answers.

The limits of alpha

Cronbach's alpha is often treated as the definitive measure of reliability, but it has well-known limitations. Sijtsma (2009) argued that alpha is neither a measure of whether a scale is unidimensional nor a good estimate of reliability in many cases, because it assumes that every item contributes equally to the total. When items differ in how strongly they relate to the underlying attribute, alpha underestimates reliability. McNeish (2018) reviewed the alternatives and recommended **coefficient omega**, which allows items to differ in their contribution and is now easy to compute in standard software.

Streiner (2003) warned against two common misuses: treating high alpha as proof that a scale measures a single construct, and seeking ever-higher alpha by adding redundant items. He noted that very high values, above about 0.90, may indicate that items repeat each other rather than covering the full range of the construct.

Improving reliability

Reliability can usually be improved by design. Adding well-written items to a scale increases reliability, because random errors on individual items tend to cancel out; the Spearman–Brown formula, described by Nunnally and Bernstein (1994), estimates how much reliability will rise as a scale is lengthened. Clear instructions and unambiguous wording reduce the chance that respondents interpret questions differently. For ratings made by people, training raters, giving them written definitions and examples, and checking agreement on a sample of cases before the main study all raise inter-rater reliability. Standardising conditions, such as the time of day or the mode of administration, removes avoidable sources of variation. Each of these steps costs time, so the effort should match the use to which scores will be put: a measure used to make decisions about individuals needs higher reliability than one used to compare group averages.

Types of validity

Validity has traditionally been divided into several types. Modern theory treats them as different kinds of evidence for a single, unified concept, but the traditional labels are still widely used.

Content validity concerns whether a measure covers all relevant aspects of the concept. A test of arithmetic that includes only addition lacks content validity. Content validity is usually assessed by experts who judge whether items represent the domain.

Criterion validity concerns whether scores relate to an external standard. **Concurrent validity** compares scores with a criterion measured at the same time, such as a short depression screen compared with a full clinical diagnosis. **Predictive validity** asks whether scores predict a future outcome, such as whether a hiring test predicts later job performance.

Construct validity concerns whether a measure behaves as the theory of the construct says it should. Cronbach and Meehl (1955) introduced the idea, arguing that when there is no single criterion against which a measure can be checked, its validity must be established by showing that it fits into a network of expected relationships with other variables.

Campbell and Fiske (1959) proposed the **multitrait–multimethod matrix**, which assesses **convergent validity**, whether measures of the same construct by different methods agree, and **discriminant validity**, whether measures of different constructs are distinct.

Face validity is whether a measure appears, on the surface, to measure what it claims. It is not strong evidence, but it can affect whether respondents take a measure seriously.

Validity as a unified concept

Messick (1995) argued that all forms of validity are aspects of construct validity, and that validation should consider six kinds of evidence: content, the response processes of test-takers, internal structure, relationships with other variables, generalisation across groups and settings, and the consequences of using the scores. The emphasis on **consequences** was new: if a test used for selection systematically disadvantages a group for reasons unrelated to the construct, that is a validity problem.

Kane (2013) developed an **argument-based approach**, in which researchers set out the chain of inferences from observed scores to intended conclusions and uses, and then gather evidence for each link. A score on a customer satisfaction scale, for example, might be used to infer that customers are satisfied, that satisfaction will predict loyalty, and that a drop in scores calls for action. Each inference needs its own evidence.

Borsboom et al. (2004) offered a different view. They argued that a test is valid if the attribute it aims to measure exists and variations in the attribute cause variations in the measurement outcomes. On this account, validity is about the relationship between the attribute and the test, not about the network of evidence. The debate continues, but both sides agree that validity requires more than a single correlation.

Validity across groups and settings

A measure that works well in one group may not work in another. Questions can be understood differently across languages, cultures, age groups or occupations, so that the same score means different things. Messick (1995) included generalisation across groups and settings among the aspects of validity for this reason. Before comparing groups, researchers should check that the measure functions the same way in each, a property known as **measurement invariance**. If a satisfaction item is interpreted more harshly in one country than another, apparent differences in satisfaction may reflect the item rather than real differences between customers. Translation needs particular care: a literal translation may change the meaning or difficulty of an item, and back-translation and testing with respondents are usually needed.

Validity of research designs

The terms are also used at the level of whole studies. Shadish et al. (2002) distinguished four kinds of validity for research designs: **statistical conclusion validity**, whether the statistical inferences are sound; **internal validity**, whether an observed relationship is causal; **construct validity**, whether the study's operations represent the constructs of interest; and **external validity**, whether findings generalise to other people, settings and times. These design-level concepts build on measurement validity: a study cannot reach valid conclusions if its measures are unreliable or invalid.

Agreement between methods

In applied fields, a common question is whether a new, cheaper or faster method can replace an established one. A high correlation between the two is not enough, because two methods can be highly correlated while one consistently gives higher values than the other. Bland and Altman (1986) proposed plotting the differences between methods against their average, and calculating **limits of agreement**, the range within which most differences fall. If the limits are narrow enough for practical purposes, the methods can be used interchangeably. Their approach is now standard in medicine and is useful whenever two measurement methods are compared.

Assessing and reporting

Good practice in reliability and validity involves several steps:

- **Use established measures where possible**, and check that they have been validated for a population and purpose similar to yours.
- **Pilot new measures**, checking how respondents interpret items and whether scores behave as expected.
- **Report reliability for your own data**, not only from earlier studies, because reliability depends on the sample.
- **State which statistic was used**, including the form of any intraclass correlation and whether alpha or omega was computed.
- **Present validity evidence** appropriate to the intended use of scores.

Mokkink et al. (2010) developed the COSMIN taxonomy, an international consensus on terms and definitions for measurement properties of health outcome measures, which provides a useful checklist for evaluating instruments. Kimberlin and Winterstein (2008) summarised the main concepts for researchers in applied health fields, emphasising that reliability and validity must be shown for each new context.

Common mistakes

Flake and Fried (2020) described **questionable measurement practices**, such as creating new scales without validation, modifying existing scales without reporting the changes, and choosing among alternative measures after seeing the results. They argued that these practices undermine research as seriously as better-known problems with statistical analysis, because conclusions cannot be stronger than the measures behind them.

Other common mistakes include:

- assuming that a measure valid in one population is valid in another, for example after translation;
- reporting a single reliability coefficient as if it settled the question of quality;
- treating a high correlation between two methods as proof that they agree;
- ignoring systematic bias because random error is low.

An applied example

A hypothetical example: a company develops a five-item scale to measure employee engagement. Internal consistency is checked using omega, which is 0.84. A test–retest study over two weeks shows high stability. For content validity, managers and staff review the items and suggest adding one on recognition. For construct validity, the scale correlates moderately with job satisfaction, as expected, and weakly with a measure of commute time, which theory says should be unrelated. For predictive validity, departments with lower scores show higher staff turnover in the following year. The figures are illustrative, but together they form an argument that the scale can be used to compare departments and to track change over time. A separate argument would be needed before using it to assess individual employees, because that use carries different consequences.

Conclusion

Reliability asks whether a measure is consistent; validity asks whether it supports the conclusions drawn from it. Reliability can be assessed in several ways, each addressing a different source of error, and the choice of statistic matters. Validity is now understood as an argument built from many kinds of evidence, including the consequences of using scores. Both must be established for each new population and purpose. Without them, even the most sophisticated analysis rests on uncertain ground.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. SAGE. https://doi.org/10.4135/9781412985642

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy, 65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., Bouter, L. M., & de Vet, H. C. W. (2010). The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *Journal of Clinical Epidemiology, 63*(7), 737–745. https://doi.org/10.1016/j.jclinepi.2010.02.006

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd