OPEN EDUCATIONAL RESOURCE

Validity and Reliability in Research: Types, Differences and Examples

Derek Jansen, MBA
Grad Coach
Reviewed by Kerryn Warren, PhD · November 2026

Abstract

Quantitative studies are expected to report the validity and reliability of their instruments, and that report is often the weakest part of the methods section: the instrument is described as valid and reliable, a textbook definition is cited, and no evidence follows. This article explains the difference between the two properties. Validity concerns whether an instrument measures what it is intended to measure, so that the interpretations drawn from its scores are justified; reliability concerns whether it measures consistently. The article describes the main types of validity evidence (face, content, construct and criterion), distinguishes them from the internal and external validity of a study, and sets out the main types of reliability (internal consistency, test-retest, inter-rater and parallel forms). It summarizes published guidance on interpreting Cronbach's alpha and the intraclass correlation coefficient. Three doctoral studies show how researchers have established, or openly failed to establish, the quality of their measures, including a scale that showed good internal consistency but poor test-retest reliability. The article closes with guidance on reporting and a note on qualitative research. It is intended for students and early-career researchers designing a quantitative study, and for the educators who teach research methods.

Keywords: validity and reliability in research; research validity; research reliability; reliability vs validity; types of validity; types of reliability; open educational resource

Suggested citation: Jansen, D. (2026). *Validity and reliability in research: Types, differences and examples*. Grad Coach. https://doi.org/10.5281/zenodo.24198889

About this article: This is the formal edition of an article first published at gradcoach.com/validity-reliability-research/, where the figure can also be downloaded for reuse. Further open educational resources on measurement and analysis are listed at gradcoach.com/open-educational-resources/.

© 2026 Grad Coach. Free to share and adapt under CC BY 4.0 with credit to Grad Coach and a link to gradcoach.com/validity-reliability-research/. Quoted thesis extracts remain their authors'; each thesis keeps its own license.

1. Introduction

Quantitative dissertations, theses and research reports usually include a passage on validity and reliability, and it is often the weakest passage in the methods section. A common pattern is to state that the instrument "is valid and reliable", cite a textbook definition and move on. Examiners and reviewers notice this, because a claim is not evidence, and every result in the study depends on the measurements being sound.

This article is written for students and researchers preparing that passage and for those who teach it. Section 2 compares validity and reliability. Section 3 describes validity and the main types of validity evidence, and Section 4 does the same for reliability, including the values commonly used to interpret Cronbach's alpha and the intraclass correlation coefficient. Section 5 uses a doctoral thesis to show that a measure can pass one test and fail another. Section 6 compares using an existing scale with building a new one, through two further studies. Section 7 sets out how the evidence is reported, and Section 8 considers qualitative research.

2. Validity and Reliability Compared

Both terms describe the quality of a measurement, whether the instrument is a survey scale, a test or a physical device. They ask different questions (Table 1).

Table 1. Validity and reliability compared.

	Validity	Reliability
The question	Is the instrument measuring what it is intended to measure?	Does it give the same result each time?
A failure	A "job satisfaction" scale that asks only about pay	A bathroom scale that shows a different weight at every reading
Main types	Face, content, construct, criterion	Internal consistency, test-retest, inter-rater, parallel forms
Typical evidence	Expert review, factor analysis, correlations with other measures	Cronbach's alpha, test-retest correlation or ICC, Cohen's kappa
Possible without the other?	No: an inconsistent measure cannot be valid	Yes: a measure can be consistently wrong

The last row is the most important. Tavakol and Dennick (2011) state the relationship directly: "An instrument cannot be valid unless it is reliable. However, the reliability of an instrument does not depend on its validity" (p. 53). **Reliability is necessary for validity, but not sufficient.** The relationship is conventionally pictured as a target, on which each dot is one reading and the bullseye is the true value (Figure 1).

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

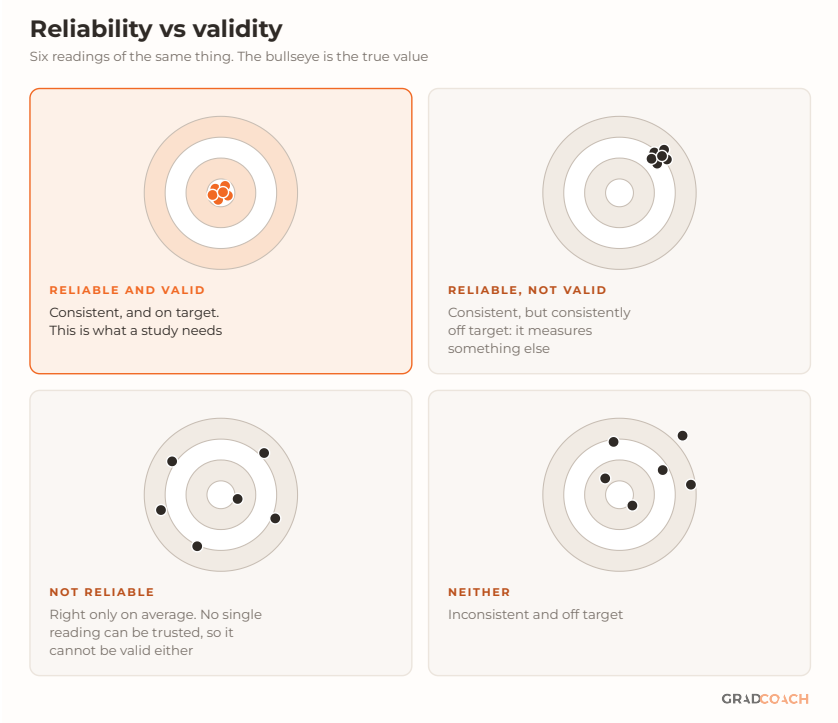


Figure 1. Reliability and validity as four targets; the bullseye is the true value.

Only the first target describes a usable measure. The second is the most misleading, because consistent numbers look trustworthy. A high Cronbach's alpha, the usual statistic for how closely a scale's items agree (Section 4.2), suggests that the questions move together; it does not show that they measure the concept the research questions are about.

3. Validity

In everyday use, validity is the extent to which an instrument measures what it is intended to measure (Tavakol & Dennick, 2011). The *Standards for Educational and Psychological Testing* define it more precisely as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014, p. 11). Two consequences follow from that definition. First, what is validated is a particular interpretation of the scores for a particular purpose, not the instrument in the abstract. Second, validity is not a property that an instrument simply has or lacks: the researcher builds a case for it from several kinds of evidence, and the strength of that case is what an examiner or reviewer judges.

Suppose a study measures job satisfaction with a set of Likert-scale items, all of which ask about pay. Table 2 shows the four items and one respondent's answers.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Table 2. A "job satisfaction" scale whose four items all concern pay (illustrative).

Item	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I am satisfied with my salary.				X	
My pay is fair for the work I do.					X
I am happy with my annual raises.				X	
My benefits package meets my needs.					X

These four items would probably agree with one another, so the scale could appear reliable. Job satisfaction, however, also covers the manager, colleagues, the work itself and the prospects of promotion. **A scale that captures only pay measures pay satisfaction**, and interpreting its scores as job satisfaction is not valid, however consistent the answers are.

The remedy begins before any item is written or chosen. The construct is defined from the theory the study uses, and the items are then checked against that definition: they should cover all of it and nothing outside it. If the theory holds that job satisfaction has five dimensions, the scale needs all five. The Job Descriptive Index, for example, divides job satisfaction into five facets and measures each separately: the work itself, pay, opportunities for promotion, supervision and coworkers (Bowling Green State University, n.d.). Table 3 shows what items covering those facets might address; the examples are illustrations written for this article, not the index's own items.

Table 3. The five facets of the Job Descriptive Index, with illustrative item topics.

Facet	An item might address
The work itself	Whether the work is interesting
Pay	Whether pay is fair for the work
Promotion	The chances of advancement
Supervision	The quality of the manager
Coworkers	The people the respondent works with

The pay-only scale in Table 2 covers one row of Table 3, and so cannot be a valid measure of job satisfaction.

3.1 Types of validity evidence for an instrument

Validity evidence is conventionally described in a few standard types. The current *Standards* treat validity as a unitary concept and speak of sources of validity evidence rather than distinct types of validity (AERA et al., 2014, pp. 13–14), but the traditional names remain in wide use in methods texts and theses. Four concern the **instrument**:

- **Face validity:** whether the items appear relevant, clear and sensible to the people who answer them. The term has been used in several senses. When the term means validity assumed from appearance alone, it has been heavily criticized. When it means test-takers' judgments of a measure's clarity, relevance, difficulty and sensitivity, it is a legitimate part

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

of scale construction and is distinct from content validity (Allen et al., 2023). Either way, it rests on people's judgments of the items rather than on how the scores behave, so on its own it cannot show that a scale measures its construct.

- **Content validity:** whether the items represent every part of the construct. Evidence of this kind can come from expert judgments of how the items relate to the construct (AERA et al., 2014, p. 14); in practice, a panel of experts usually reviews the items.
- **Construct validity:** whether scores behave as the theory says they should. This includes **convergent** evidence (relationships with measures of the same or similar constructs) and **discriminant** evidence (weaker relationships with measures of different constructs) (AERA et al., 2014, pp. 16–17). Factor analysis is commonly used to test whether items group as the theory predicts.
- **Criterion validity:** whether scores relate to a relevant criterion, an outcome measured separately from the instrument. In a **predictive** study the criterion is measured later; in a **concurrent** study, at about the same time (AERA et al., 2014, p. 17). An aptitude test that predicts first-year grades is an example of predictive evidence.

3.2 Internal and external validity

Two further types concern the **study as a whole**, not the instrument.

Internal validity is whether the way a study was designed, conducted and analyzed allows trustworthy answers to its research questions, free of systematic error or bias (Andrade, 2018). It is the central concern in an experiment or quasi-experiment, where the question is usually whether one thing caused another, and the risk is that something other than the intervention explains the result. Suppose students who chose to join a tutoring program later scored higher than those who did not. If the volunteers were more motivated to begin with, motivation could explain the difference as well as the tutoring could. Random assignment, a comparison group and measuring both groups beforehand are the usual ways of ruling out explanations of this kind.

External validity is whether the findings generalize beyond the study to other contexts, such as other people, settings or times (Andrade, 2018). It depends heavily on sampling: findings from a random sample can be generalized to the population from which it was drawn (Andrade, 2018). A survey of first-year students at one university describes those students, and its findings may not hold for students elsewhere or in later years. A convenience sample does not make a study unusable, but it does limit how far the findings can be generalized, and the discussion should say so.

When an examiner or reviewer asks about "validity" in the methods section, they usually mean the instrument types. The internal and external validity of the design belong in the discussion of the design and its limitations.

4. Reliability

Reliability is whether an instrument **measures consistently** (Tavakol & Dennick, 2011, p. 53). The *Standards* use the broader term reliability/precision for the consistency of scores across replications of a testing procedure (AERA et al., 2014, p. 33). If a bathroom scale shows a different weight

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

at each reading, no single reading can be trusted, and nothing calculated from those readings can be trusted either.

Reliability also covers the people who use an instrument. If two researchers code the same interview, or rate the same classroom observation, and reach different answers, the data have a reliability problem, even if the coding scheme itself is sound.

4.1 Types of reliability

The *Standards* recognize three broad categories of reliability coefficient: those from alternate forms of a test, those from giving the same form more than once, and those from the items within a single administration (one occasion on which the instrument is completed). Where scoring involves judgment, indices of scorer consistency are added (AERA et al., 2014, p. 37). The type needed depends on how the data are collected:

- **Internal consistency:** whether the items in a scale agree with one another. It is estimated from a single administration, as a split-half coefficient, KR-20 or coefficient alpha (AERA et al., 2014, p. 35). Coefficient alpha, usually called **Cronbach's alpha**, is the coefficient most often reported for a multi-item survey scale.
- **Test-retest reliability:** whether the same people obtain similar scores when they complete the instrument again after an interval. It is usually reported as a Pearson correlation or an intraclass correlation coefficient (ICC) (Boateng et al., 2018).
- **Inter-rater reliability:** whether different coders or raters agree. Cohen's kappa is used for categorical judgments by two raters, and Fleiss' kappa, an adaptation of it, for three or more (McHugh, 2012); the ICC is used for scores (Koo & Li, 2016).
- **Parallel forms reliability:** whether two versions of the same test, built to the same specifications, give the same results (AERA et al., 2014, p. 35). It is used mainly with tests of knowledge or skill; survey instruments normally use the same questions on every administration (AERA et al., 2014, p. 35).

Survey-based studies typically report internal consistency for each scale. Test-retest reliability is added when the construct should be stable over time, and inter-rater reliability whenever human judgment turns data into codes or scores.

4.2 Interpreting Cronbach's alpha

Alpha normally lies between 0 and 1 (Tavakol & Dennick, 2011, p. 53). Alpha falls below zero when the average covariance between items is negative, which can happen when an item worded in the opposite direction has not been reverse-scored (scored so that a high value means the same as on the other items). There is no single agreed threshold. Tavakol and Dennick (2011) note that reported acceptable values range "from 0.70 to 0.95" (p. 54). Taber (2018) surveyed articles in leading science education journals. Where authors stated a threshold of acceptability, it was normally .70, but the verbal labels attached to alpha values (such as "acceptable", "good" and "excellent") varied widely between studies (p. 1278). Bands of .70 for acceptable, .80 for good and .90 for excellent are often cited (e.g., George & Mallery, 2019, as cited in Tunstall, 2026, p. 75), but they are conventions rather than fixed standards.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Two points are frequently missed:

- **Higher is not always better.** An alpha above .90 may indicate that some items are redundant, "testing the same question but in a different guise", and a maximum of .90 has been recommended (Tavakol & Dennick, 2011, p. 54).
- **Alpha is calculated per scale, not for the whole questionnaire.** Alpha assumes that the items measure one concept. Calculated across several concepts at once, it is inflated by the larger number of items, so it should be calculated for each concept, which in practice means each scale and subscale (Tavakol & Dennick, 2011, p. 54).

When alpha is unexpectedly low, reverse-worded items are the first thing to check, for example "I rarely feel stressed" in a stress scale. **An item that runs in the opposite direction to the rest lowers alpha** until it is reverse-scored. This cause is also the easiest to correct.

4.3 Interpreting the intraclass correlation coefficient

For test-retest and inter-rater ICCs, Koo and Li (2016) propose that values below .5 indicate poor reliability, .5 to .75 moderate, .75 to .9 good and above .9 excellent. They recommend judging the level by the 95% confidence interval of the estimate rather than by the single value, and reporting the interval with the estimate.

5. Passing One Test and Failing Another

A measure can be reliable without being valid. Internal consistency shows only whether items move together, and items can move together around the wrong concept; the pay-only scale in Section 3 is the standard example.

A related problem arises within reliability itself: **a scale can pass one reliability test and fail another**. A clinical psychology doctorate from the University of Lincoln shows this clearly. Tunstall (2026) tested the Trent Measure of Vicarious Trauma (TMVT), a 13-item measure, with 418 helping professionals in the United Kingdom. Vicarious trauma is the cumulative toll that exposure to clients' trauma takes on such professionals.

On internal consistency the scale performed well, with an alpha of .86 (Tunstall, 2026, p. 75). It also correlated with an established measure of psychological distress, the DASS-21 total score (r = .65; p. 78), which is convergent evidence of validity. Then 318 of the participants completed it again, on average 19.3 days later (p. 80):

> Test-retest reliability was poor for the TMVT full scale and subscales despite strong internal consistency, indicating substantial short-term variability in individual vicarious trauma responses.
>
> Tunstall (2026, p. 82)

The test-retest ICC, reported in its average-measures form (the reliability of the mean of the two occasions), was .208 (95% CI [.013, .365]; Tunstall, 2026, p. 82). Even the upper limit lies in the "poor" band of Koo and Li (2016). The single-measurement form (the reliability of a score from one occasion), which Koo and Li advise when an instrument will be used on one occasion, would be lower still. The thesis is instructive because it does not conceal the result. It **reports the alpha**

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

and the test-retest ICC side by side and discusses what might explain the difference. One possibility is that vicarious trauma itself changes over a period of weeks, in which case fluctuating scores may reflect real change rather than measurement error (p. 90). The thesis also notes the opposite case: if vicarious trauma is an enduring change, the instability may count against the measure.

Whether the construct changes over weeks matters. Test-retest reliability **counts against a measure only when the construct should be stable**. The *Standards* note that test-retest coefficients are not always preferable to internal-consistency coefficients, and that where scores are not expected to change, stability over occasions may reasonably be assumed (AERA et al., 2014, p. 38). If the experience genuinely changes over two or three weeks, a low retest coefficient may say more about the experience than about the scale. A disappointing reliability result should therefore be reported and interpreted, not omitted.

6. Using an Existing Scale or Building a New One

As a general rule, **an existing, validated scale should be used wherever a suitable one exists.** A published scale arrives with validity and reliability evidence already in the literature. A newly written scale arrives with none, and building that evidence is a study in its own right. Boateng et al. (2018) describe scale development as costly and time-consuming, and their first step includes confirming that no existing instrument will serve the purpose, or justifying why a new one is needed.

Suitable scales can usually be found in the following places:

- the methods sections and appendices of studies on the topic, which name the scales they used;
- the paper that originally developed and validated the scale, which will be cited in any case;
- databases to which many libraries subscribe, such as PsycTESTS and the Mental Measurements Yearbook;
- the scale's authors, who may need to be contacted for permission and the full item list.

Once chosen, a scale should be used as published. If the wording is changed, items are dropped or the scale is translated, the change should be stated and explained, and the altered version treated as needing fresh evidence. Even an unchanged scale does not carry its original evidence over automatically: alpha is a property of the scores from a specific sample, so it should be calculated each time the instrument is administered rather than taken from published estimates (Tavakol & Dennick, 2011, p. 53), and reliability coefficients from one population may differ from those obtained in another (AERA et al., 2014, p. 37).

6.1 An adapted scale, re-tested

A doctoral dissertation from Purdue University shows what that fresh evidence looks like. Huang (2019) surveyed undergraduates about culturally responsive teaching, their sense of belonging and their academic self-efficacy, with 1,061 usable responses (p. 82).

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

No published scale fitted the study's setting (undergraduate mathematics classes) exactly, so Huang adapted existing ones. The culturally responsive teaching scale, for example, began from a published 18-item scale for middle school students, dropped its foreign language affirmation subscale, and added relationship-building items adapted from a second instrument, giving 23 items (Huang, 2019, pp. 66–67, 98). Because the scales had changed, the dissertation treats them as "three new scales" (p. 89) and tests them accordingly.

The study did three things that other researchers can follow. First, an expert panel of four faculty members, three doctoral students and one master's student reviewed the questionnaire (p. 73), and items were reworded on their advice. Second, a pilot with 39 responses tested the survey before the main data collection (p. 75). Third, the reliability table reports **the alpha of the published scale each one started from beside the alpha in the study's own sample** (Table 4).

Table 4. Cronbach's alpha for the original scales and in the study's sample (Huang, 2019, p. 89).

Scale	Original alpha	Alpha in this study
Students' perceptions of culturally responsive teaching	.90	.95
Sense of belonging	.88	.82
Academic self-efficacy	.85	.89

The last column is the one to read. **A scale that was reliable for its original authors may not be reliable in a new sample**, and an adapted scale is even less certain to be. Examiners and reviewers therefore expect both values. The first scale's .95 also lies above the .90 mark discussed in Section 4.2, which is the kind of value a reviewer may ask about.

Huang also ran factor analyses to check that the adapted items grouped as the theory predicted: exploratory and confirmatory analyses for the culturally responsive teaching scale, and a confirmatory analysis for the sense of belonging scale (pp. 98–111). The self-efficacy scale was not analyzed separately; its items entered the study's structural equation model as indicators of one latent variable (p. 114). Factor analysis of this kind is construct validity evidence, and it is a step an adapted scale needs.

6.2 A self-built questionnaire, reported honestly

Sometimes no suitable scale exists. A Doctor of Nursing Practice project from the University of Massachusetts Amherst shows the difficulty this creates. De Villier (2025) set out to measure what primary care providers knew about diet and heart disease before and after an education session sent by email. Finding no reliable existing instrument, the author wrote a 17-item questionnaire on the providers' views and knowledge, based on the literature (p. 12). The methods section reports this in a single sentence whose two halves are worth reading separately:

> Although reliability and precision were not confirmed in this study, a Likert Scale has been found to be 88% reliable and 89% precise when used in other research studies (Belita et al., 2022).
>
> de Villier (2025, p. 12)

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

The first half of the sentence reflects the right instinct: **stating what was not established is far better than implying that it was.** The second half illustrates what to avoid. Reliability belongs to a specific instrument used with a specific group, not to a response format, so a figure from a different study cannot stand in for evidence from the study's own data (Tavakol & Dennick, 2011).

With six providers taking part (de Villier, 2025, p. 13), a meaningful reliability analysis was probably not realistic, which makes the open admission the right decision. In a larger study, the minimum for a newly written scale is an expert review for content validity, a pilot and an alpha calculated from the study's own data.

7. Reporting Validity and Reliability

For each instrument, the validity and reliability passage should answer three questions: **where the instrument came from, what evidence supports it, and what was found in the study's own data.** Box 1 shows one way to answer all three, using the figures from Huang's (2019) study; each sentence supplies either a source, a piece of evidence or a result from the study's own data.

Box 1. An example validity and reliability paragraph, written for this article from the figures in Huang (2019).

> Students' perceptions of culturally responsive teaching were measured with a 23-item scale adapted from two published instruments. The main one, an 18-item scale for middle school students, had a reported Cronbach's alpha of .90. To establish face and content validity, a panel of eight experts reviewed the questionnaire, and items were reworded on their advice. A pilot study (n = 39) tested the survey's clarity and completion time. Because the scale had been adapted, its structure was tested with exploratory and confirmatory factor analysis. In the main study, internal consistency was high (α = .95).

A few sentences of evidence, each with a number or a named procedure, carry more weight than a page of definitions. Some departments expect the study's own values in the methods section, as here, and others in the results section beside the descriptive statistics, so the program handbook or a recent thesis from the same program should be checked.

Four problems recur in this part of a thesis or report:

- **A limitations discussion headed "validity and reliability".** Limitations are what the design could not do; validity and reliability evidence shows that the measures work. They belong in different sections.
- **A high response rate offered as evidence of validity.** A response rate bears on the risk of non-response bias, a form of selection bias, not on whether the instrument measures the right thing.
- **Software output headings used as thesis headings.** SPSS labels its output "Reliability"; the results should be organized around the research questions instead.
- **The original authors' alpha reported with nothing from the study's own data.** Alpha should always be calculated for the sample at hand.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

Avoiding these four problems anticipates the questions an examiner or reviewer is likely to ask about this section.

8. Qualitative Research

Validity and reliability come from measurement, so they fit qualitative research awkwardly. A qualitative study does not measure a construct with an instrument, and another researcher working from the same interviews would not be expected to produce identical themes.

Many qualitative researchers use **trustworthiness** instead. Lincoln and Guba introduced four criteria intended to parallel the conventional quantitative ones: credibility, transferability, dependability and confirmability (Lincoln & Guba, 1985, as cited in Nowell et al., 2017). Each is supported by named procedures, such as member checking, an audit trail and a reflexive journal (Nowell et al., 2017).

Where coding involves more than one coder working from a codebook, a study may still report intercoder agreement. Whether it should depends on the analytic approach: intercoder reliability is contested in qualitative research, with some arguing that it is inappropriate or unnecessary for the goals of qualitative analysis and others that it improves the systematicity and transparency of coding (O'Connor & Joffe, 2020).

9. Conclusion

Validity is whether an instrument measures what it is intended to measure; reliability is whether it measures consistently. A measure can be reliable without being valid, but not valid without being reliable. Each property is established through specific evidence: expert review, factor analysis and correlations with other measures for validity, and Cronbach's alpha, test-retest coefficients and inter-rater statistics for reliability. Published thresholds for these statistics are conventions, to be applied with judgment and reported with the values from the study's own sample. Where a suitable validated scale exists, it should be used and its reliability re-tested in the new sample; where a scale is adapted or newly written, it needs fresh evidence. Where a result disappoints, or a test was not possible, the strongest course is to report it and explain it, as the studies discussed here did.

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com

References

Allen, M. S., Robson, D. A., & Iliescu, D. (2023). Face validity: A critical but ignored component of scale construction in psychological assessment. *European Journal of Psychological Assessment*, 39(3), 153–156. https://doi.org/10.1027/1015-5759/a000777

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Andrade, C. (2018). Internal, external, and ecological validity in research design, conduct, and evaluation. *Indian Journal of Psychological Medicine*, 40(5), 498–499. https://doi.org/10.4103/IJPSYM.IJPSYM_334_18

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, Article 149. https://doi.org/10.3389/fpubh.2018.00149

Bowling Green State University. (n.d.). *Job Descriptive Index*. Retrieved November 11, 2026, from https://www.bgsu.edu/arts-and-sciences/psychology/services/job-descriptive-index.html

de Villier, A. (2025). *The Mediterranean diet & heart disease: Measuring knowledge through provider education* [Doctor of Nursing Practice project, University of Massachusetts Amherst]. ScholarWorks@UMass Amherst. https://doi.org/10.7275/9f48-dq07

Huang, Y. (2019). *Undergraduate students' perceptions of culturally responsive teaching and their sense of belonging and academic self-efficacy in higher education* [Doctoral dissertation, Purdue University]. Purdue University Graduate School. https://doi.org/10.25394/PGS.9116885.v1

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

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282. https://doi.org/10.11613/BM.2012.031

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, 16(1), Article 1609406917733847. https://doi.org/10.1177/1609406917733847

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19, Article 1609406919899220. https://doi.org/10.1177/1609406919899220

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

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

Tunstall, C. M. (2026). *Confirmatory factor analysis and further validation of the Trent Measure of Vicarious Trauma (TMVT)* [Doctoral thesis, University of Lincoln]. University of Lincoln Research Repository. https://doi.org/10.24385/lincoln.33113120

© 2026 Grad Coach · CC BY 4.0 · gradcoach.com