PSYCHTRIXWEB RESEARCH NOTE 005
Research Note · September 8, 2026

Validity in Psychological Assessment: Evidence, Interpretation, and Common Misconceptions

Enoch O. Oladunmoye *
Department of Applied Psychology, Kampala International University

PsychtrixWeb Research Note 005 · Version 1.0 · CC BY 4.0
https://www.psychtrixweb.online/research-notes/005-1-introduction

ABSTRACT

Validity is arguably the most fundamental consideration in psychological measurement, because psychological assessment ultimately involves making interpretations and decisions from observed responses or scores. Despite its centrality, validity is frequently misunderstood as a permanent characteristic of a test, a single statistical coefficient, or a property demonstrated simply by obtaining statistically significant results. Contemporary psychometric theory instead conceptualises validity as an evidence based argument concerning the interpretations and uses of scores, a distinction with substantial implications for questionnaire development, scale validation, psychological assessment, and empirical research more broadly. This Research Note provides a structured examination of validity in psychological assessment. Building on the distinction between constructs, observed scores, and measurement error established in PsychtrixWeb Research Note 001, and on the discussion of reliability in Research Note 002, the paper traces the conceptual foundations of validity, the historical transition from separate forms of validity toward a unified framework, and the major sources of validity evidence. Particular attention is given to content evidence, response processes, internal structure, relations with other variables, and the consequences of score use. The paper also examines convergent validity, discriminant validity, criterion related evidence, construct validity, structural validity, cross cultural validity, measurement invariance, and differential item functioning. Common misconceptions are critically examined, including the belief that Cronbach's alpha establishes validity, that factor analysis alone proves validity, and that a significant correlation automatically demonstrates construct validity. A practical, staged validity evaluation workflow is proposed, together with an indication of how such a workflow could be operationalised within PsychtrixWeb. Keywords: validity, psychological assessment, construct validity, content validity, convergent validity, discriminant validity, structural validity, measurement invariance, differential item functioning, psychometrics, PsychtrixWeb

Keywords: validity, psychological assessment, construct validity, content validity, convergent validity, discriminant validity, structural validity, measurement invariance, differential item functioning, psychometrics, PsychtrixWeb

1. Introduction

A psychological instrument does not possess validity in the same simplistic sense that a physical object possesses weight. Validity concerns the interpretations and uses of scores generated under specified conditions. Messick (1995) was particularly influential

in articulating validity as a unified concept concerned with the empirical justification of score meaning and the consequences associated with score interpretation and use. The Standards for Educational and Psychological Testing place validity at the centre of responsible testing and emphasise that evidence must support proposed score interpretations and uses (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

Consequently, validity should not be treated as a final statistical checkbox. It is better understood as an accumulating argument, built from multiple, mutually reinforcing lines of evidence rather than settled by any single test or coefficient.

2. Conceptual Foundations of Validity

2.1 What Is Validity?

Validity concerns the degree to which evidence and theory support the proposed interpretations of scores for their intended uses. This formulation carries several important implications. First, validity concerns interpretation rather than the instrument in isolation. Second, validity concerns evidence, which must be assembled and evaluated rather than assumed. Third, validity is tied to a specific purpose and population, so a conclusion that holds for one context cannot automatically be generalised to another. Fourth, and perhaps most importantly for practising researchers, validity is never established by a single statistical coefficient.

Consider a researcher who develops an Academic Resilience Scale and wishes to conclude that students scoring highly on the instrument possess greater academic resilience. The validity question is not simply "is the instrument valid?" but rather "what evidence supports the interpretation of these scores as indicators of academic resilience among this population?" That reframing changes how validation studies should be designed from the outset, because it requires the researcher to specify, in advance, what evidence would support or undermine the intended interpretation.

2.2 Validity Is Not a Fixed Property of an Instrument

One of the most persistent misconceptions in psychometrics is the flat statement that "this questionnaire is valid." Such a statement is incomplete. A more defensible formulation is that evidence supports the intended interpretation of scores from a questionnaire in a specified population and context. An instrument can function differently across populations, languages, cultures, age groups, educational levels, clinical groups, occupational groups, and administration modes. An instrument validated among university students in one country cannot automatically be assumed to possess equivalent validity when administered to adolescents, clinical patients, employees, or populations in another cultural setting.

This principle is particularly important for psychological research in Africa, where instruments developed elsewhere are frequently adapted for local populations without a corresponding re-examination of whether the original evidence base still applies.

3. Historical Evolution of Validity Theory

Traditional psychometric literature distinguished among content validity, criterion related validity, and construct validity, following the influential formulation of Cronbach and Meehl (1955). Although these categories remain useful for teaching purposes, contemporary validity theory treats them not as independent types of validity but as complementary sources of evidence feeding into a single argument. Messick (1995) argued that the traditional division into separate types was fragmented, and proposed

a unified conception of construct validity that incorporates multiple forms of evidence together with the consequences of score use. Table 1 summarises this historical progression.

Period	Dominant view	Representative source	Key limitation addressed
1950s to 1970s	Three or four discrete types: content, criterion related (concurrent and predictive), and construct validity	Cronbach & Meehl (1955)	Treated validity as a set of separate certificates rather than a single argument
1980s to 1990s	Unified, construct centred validity incorporating multiple evidence sources and consequences of use	Messick (1995)	Fragmentation of evidence; neglect of the consequences of testing
2000s to present	Argument based validation; measurement invariance and fairness treated as core validity concerns	AERA, APA, & NCME (2014); Vandenberg & Lance (2000)	Group comparability and fairness were previously treated as optional extensions rather than central validity questions

Table 1. Historical progression of validity theory, from discrete types toward a unified, argument based framework.

Instead of asking "does my instrument have content validity?" and, separately, "does it have construct validity?", the contemporary researcher asks a single integrating question: what evidence supports the interpretation I intend to make from these scores? Different sources of evidence then contribute to that overall validity argument, rather than functioning as independent hurdles to be cleared.

4. The Unified Framework: Five Sources of Validity Evidence

Contemporary assessment practice commonly organises validity evidence around five interconnected sources: evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence concerning the consequences of testing. These sources are not five separate certificates to be collected independently; they are components of a single, integrated validity argument, as illustrated in Figure 1.

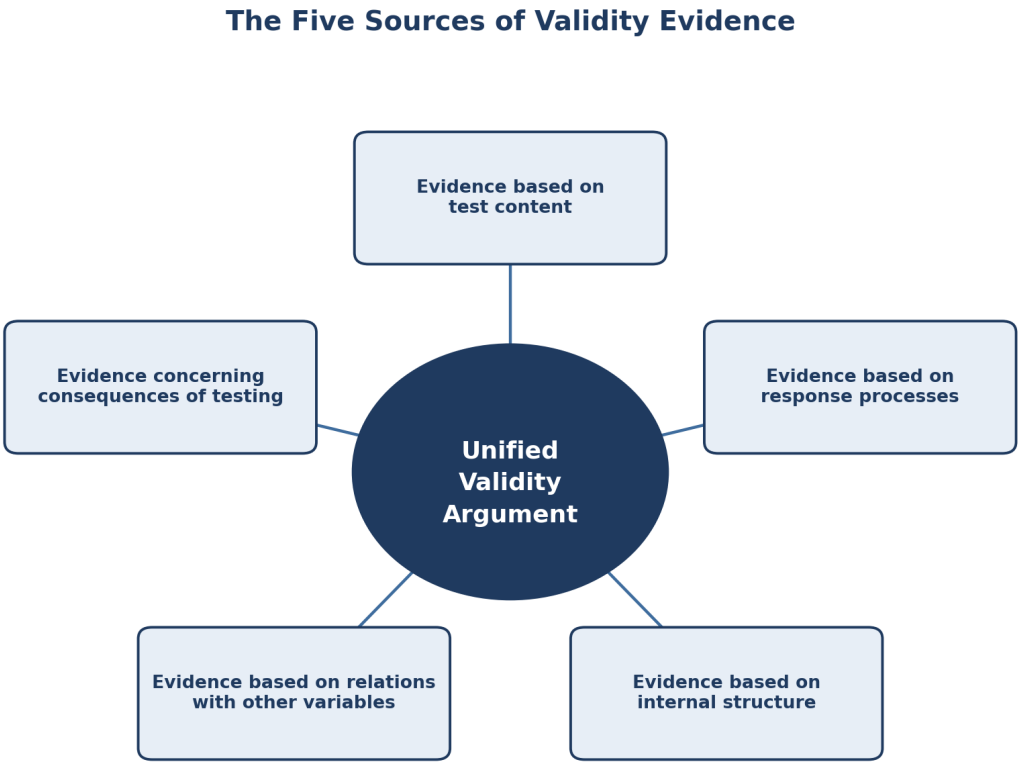

Figure 1. The five sources of validity evidence feed into a single, unified validity argument rather than standing as independent types of validity.

4.1 Evidence Based on Test Content

Content evidence concerns whether the content of an assessment adequately represents the construct domain. Suppose a researcher develops a scale measuring digital wellbeing, defined as encompassing healthy digital engagement, digital self regulation, psychological balance, adaptive technology use, and protection from harmful digital experiences. If the final scale contains only items concerning screen time management, it may not adequately cover the conceptual domain. The problem here is not primarily statistical; it is conceptual, and no amount of subsequent statistical analysis can fully repair a poorly specified item pool.

Content validity therefore begins with a sequential design process, shown in Figure 2.

Figure: Content Validation as a Sequential Design Process

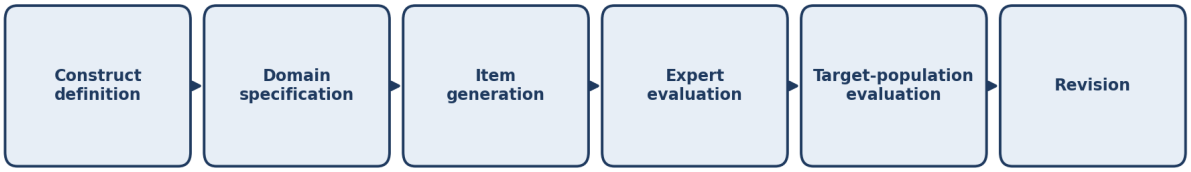

Figure 2. Content validation proceeds from construct definition through domain specification, item generation, expert and target population evaluation, and revision.

COSMIN similarly emphasises relevance, comprehensiveness, and comprehensibility when evaluating content validity, while noting that item relevance should be considered in relation to the construct, the target population, and the intended purpose (Mokkink et al., 2010a; Terwee et al., 2018).

Content Validity Is More Than Expert Agreement

Researchers sometimes report only that "five experts reviewed the questionnaire and agreed that it was valid." This is insufficiently informative. A rigorous content validation process should document how the construct was defined, how domains were identified, how experts were selected, what criteria experts used, whether items were relevant, whether important content was missing, whether items were understandable, whether response options were appropriate, and what revisions were made as a result.

Researchers may also use quantitative indices such as:

- the item level content validity index (I-CVI);
- the scale level content validity index (S-CVI);
- the content validity ratio (CVR);
- inter-rater agreement measures.

However, numerical indices should supplement rather than replace substantive judgement; content validity is partly a design and reasoning problem, not simply a numerical exercise.

4.2 Evidence Based on Response Processes

Response process evidence asks whether the psychological processes involved in responding correspond to the processes assumed by the measurement model. Suppose an item states: "I can regulate my emotions when using social media." What does the respondent understand by "regulate"? Does the respondent interpret "social media" as Facebook, TikTok, WhatsApp, Instagram, or all digital communication generally? Does the respondent interpret the item as emotional regulation in general, or emotional regulation specifically triggered by social media use? If respondents systematically interpret an item differently from what the researcher intended, the measurement process is compromised regardless of how the resulting scores subsequently behave in statistical analyses.

Response process evidence can therefore involve:

- cognitive interviews;
- think-aloud procedures;
- respondent debriefing;
- qualitative probing;
- usability testing;

- examination of response patterns.

This form of evidence is especially important during scale development and cross cultural adaptation, when the researcher cannot assume that an item translates its intended meaning without distortion.

4.3 Evidence Based on Internal Structure

Internal structure evidence concerns whether the relationships among items are consistent with the theoretical structure of the construct. This is where many familiar psychometric techniques, including exploratory factor analysis, confirmatory factor analysis, bifactor modelling, multidimensional item response theory, and structural equation modelling, become central to the validity argument.

Suppose a researcher theorises that Psychological Wellbeing consists of three dimensions: Emotional Wellbeing, Psychological Functioning, and Social Wellbeing. If confirmatory factor analysis (CFA) demonstrates that the proposed three factor structure fits the data reasonably well, this contributes evidence supporting the intended interpretation. However, good model fit does not, by itself, prove validity; it provides evidence concerning only one component of the broader measurement argument. COSMIN similarly distinguishes structural validity from other measurement properties and treats internal structure as one component of a wider framework (Mokkink et al., 2010b).

Exploratory Factor Analysis and Validity

Exploratory factor analysis (EFA) is useful when the researcher is uncertain about the underlying structure of a construct. It can help investigate the number of factors, item-factor relationships, cross-loadings, poorly functioning items, and potential subconstructs. However, EFA does not automatically establish construct validity. A factor structure should always be interpreted in relation to theory, item content, sample characteristics, extraction method, rotation, and factor retention decisions, because an empirically generated factor does not automatically represent a theoretically meaningful construct.

Confirmatory Factor Analysis and Validity

CFA allows researchers to evaluate an explicitly specified measurement model. For example, a researcher might specify that Resilience is indicated by Persistence, Adaptability, Recovery, and Goal Orientation, and estimate the model to see whether observed relationships among items are compatible with this hypothesised structure before, rather than after, seeing the data. Researchers should nevertheless avoid the common shortcut of concluding that "CFA model fit was acceptable, therefore the scale is valid." Model fit is evidence about internal structure specifically, not a complete validation certificate for the instrument as a whole.

4.4 Evidence Based on Relations With Other Variables

A particularly important component of validity concerns the relationship between scores and theoretically relevant external variables. This includes convergent validity, discriminant validity, criterion related evidence, known-groups comparisons, predictive relationships, and theoretically expected correlations. The fundamental principle underlying all of these is that a valid measure should behave in ways predicted by theory.

Suppose a researcher develops a measure of academic self efficacy. Theory may predict that self efficacy should correlate positively with academic engagement, correlate positively with persistence, correlate negatively with academic anxiety, and differ between groups with substantially different academic experiences. Crucially, the researcher can and should formulate these expectations before examining the results.

COSMIN explicitly emphasises the importance of a priori hypotheses regarding the expected direction and magnitude of relationships when evaluating construct validity (Mokkink et al., 2010a).

4.4.1 Convergent Validity

Convergent validity concerns whether a measure relates to other measures with which it should theoretically be associated. For example, a new anxiety scale should correlate positively with established anxiety measures, relevant physiological indicators, and theoretically related distress measures. A strong relationship can support the argument that the new scale captures the intended construct, but the magnitude of the expected relationship should itself be theoretically justified rather than judged solely by whether p is below .05. As discussed further in Section 6, with a sufficiently large sample even a very small correlation can become statistically significant, so statistical significance alone does not amount to substantive validity evidence.

4.4.2 Discriminant Validity

Discriminant validity concerns whether a measure can be distinguished from theoretically different constructs. Suppose a researcher develops a measure of Academic Resilience. The researcher should demonstrate that the measure is not simply another measure of general optimism, self esteem, academic motivation, or conscientiousness. This does not require that the constructs be completely unrelated, since psychological constructs frequently overlap; the relevant question is whether they demonstrate sufficient empirical and conceptual distinctiveness. In CFA and structural equation modelling research, discriminant validity is commonly examined through factor correlations, competing measurement models, average variance extracted (AVE) based approaches, cross-loadings, theoretically informed model comparisons, and the heterotrait-monotrait ratio of correlations, or HTMT, proposed by Henseler, Ringle, and Sarstedt (2015) as a more reliable alternative to the traditional Fornell-Larcker criterion.

Feature

Convergent validity

Discriminant validity

Core question

Does the measure relate to constructs it should theoretically resemble?

Does the measure remain distinct from constructs it should theoretically differ from?

Expected relationship

Moderate to strong association

Weak association, or a demonstrably distinct factor structure

Typical evidence

Correlations with established measures; shared variance in SEM

HTMT ratios, factor correlations, competing model comparisons, AVE based indices

Common pitfall

Treating any significant correlation as sufficient evidence

Assuming constructs must be uncorrelated rather than merely distinguishable

Table 2. Convergent and discriminant validity compared: complementary rather than opposing forms of relational evidence.

4.4.3 Criterion-Related Evidence

Criterion related validity concerns the relationship between scores and an external criterion, for example the relationship between a psychological test score and clinical diagnosis, an aptitude score and subsequent performance, a selection test and job

performance, or a screening instrument and an established diagnostic outcome. Researchers should, however, be cautious about the term "gold standard": in many areas of psychology a genuinely error free criterion does not exist (Oladunmoye 2015). COSMIN notes that researchers sometimes incorrectly treat established instruments as gold standards when they should instead be evaluated through explicit hypotheses about the relationships among constructs (Mokkink et al., 2010a). The researcher should therefore justify the chosen criterion rather than simply labelling it a gold standard.

4.4.4 Construct Validity

Construct validity is broader than any single statistical test; it concerns whether the observed evidence behaves as expected if the proposed construct interpretation is correct. Imagine that a researcher proposes that psychological resilience protects students against academic stress. The validity argument might predict that resilience relates positively to adaptive coping and academic persistence, while relating negatively to psychological distress. These hypotheses, following the nomological network logic first articulated by Cronbach and Meehl (1955), provide opportunities to test whether observed relationships conform to theory. Construct validity is consequently cumulative: no single correlation establishes it.

4.5 Evidence Concerning Consequences of Testing

Messick's unified validity framework also emphasises the consequences associated with interpretation and use. This does not mean that every undesirable outcome automatically invalidates an instrument; rather, researchers should consider whether the proposed interpretation and use produce consequences that are consistent with the construct, the purpose, and the intended application. This becomes particularly important in clinical assessment, employment selection, educational placement, high-stakes examinations, diagnostic screening, and algorithmic decision-making, where scores can materially influence people's opportunities and treatment. Psychological assessment is therefore not merely a statistical exercise.

5. Extending the Validity Argument Across Groups

5.1 Validity and Measurement Invariance

An instrument may demonstrate an acceptable factor structure in the total sample while functioning differently across groups. Suppose a Psychological Wellbeing Scale is administered to males and females, younger and older participants, urban and rural participants, different language groups, and different countries. If the meaning of the latent construct differs across groups, comparisons of group means or structural relationships may be misleading (Oladunmoye & Muhammad, 2024). Measurement invariance analysis, most commonly conducted through multigroup confirmatory factor analysis, investigates whether the measurement model operates comparably across groups, typically proceeding through the stages summarised in Figure 3.

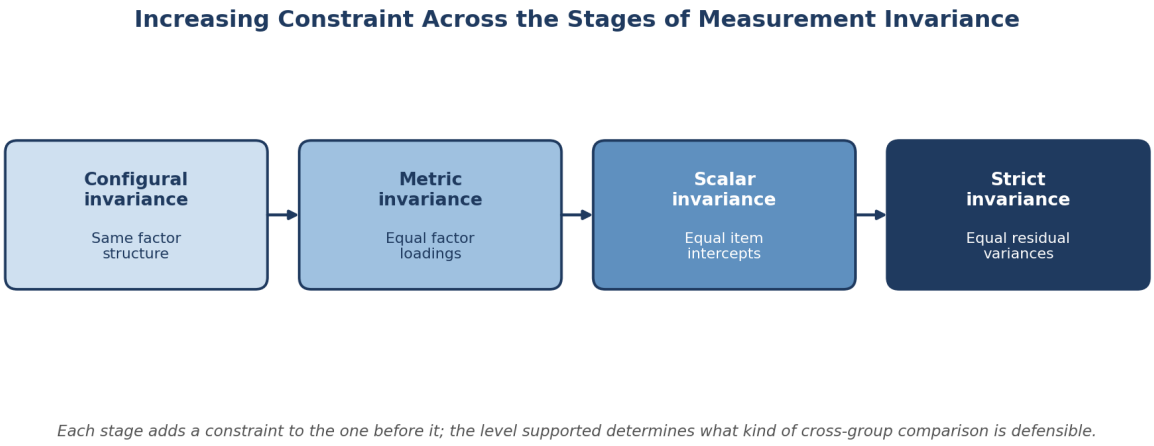

Figure 3. Configural, metric, scalar, and strict invariance represent progressively stronger constraints; the highest level supported determines which cross-group comparisons are defensible.

The interpretation of group comparisons depends on which level of invariance is supported, a point developed at length in the widely cited review by Vandenberg and Lance (2000). This is why measurement invariance should be considered an important extension of the validity argument rather than merely an optional advanced statistical technique.

5.2 Validity and Differential Item Functioning

Differential item functioning (DIF) provides a complementary, item level perspective on measurement comparability. An item demonstrates potential DIF when individuals from different groups with comparable levels of the underlying trait have different probabilities of endorsing the item. For example, two students may have equivalent levels of academic resilience, yet an item might systematically favour one cultural or linguistic group over another. The total scale may appear reliable overall while specific items operate differently beneath the surface. Zumbo (2007) traces the evolution of DIF methodology across three successive generations of theorising and demonstrates why psychometric evaluation should move beyond reliability alone. PsychtrixWeb can therefore position DIF and measurement invariance as complementary fairness and validity analyses within a single workflow.

5.3 Cross-Cultural Validity

Psychological constructs are not always expressed identically across cultural contexts. A questionnaire developed in one setting may contain culturally specific assumptions, idiomatic expressions, unfamiliar situations, inappropriate examples, or behavioural norms that do not transfer. Consequently, translating a questionnaire word for word is not equivalent to validating it in another culture. Cross cultural validation should consider linguistic equivalence, conceptual equivalence, cultural relevance, response processes, internal structure, reliability, measurement invariance, and DIF (Van de Vijver & Leung, 1997). COSMIN treats cross cultural validity and measurement invariance as important aspects of evaluating whether an instrument operates comparably across groups (Mokkink et al., 2010a).

6. Why Statistical Shortcuts Do Not Establish Validity

6.1 Why Reliability Does Not Establish Validity

This principle directly connects Research Note 003 with Research Note 002. Suppose a scale achieves an internal consistency coefficient of alpha = .93. This indicates high

internal consistency under the conditions in which alpha was estimated, but it does not, by itself, demonstrate content validity, structural validity, convergent validity, discriminant validity, cross cultural validity, or measurement invariance. As discussed in Research Note 002, reliability is evidence concerning measurement consistency or precision; validity concerns the defensibility of score interpretations. Reliability is important for validity, but reliability is not validity: a measure can be consistently wrong.

6.2 Why Factor Analysis Does Not "Prove" Validity

Another common misconception holds that "the CFA fit indices were good, therefore the questionnaire is valid." This claim is too strong. A CFA model can fit well for several reasons unrelated to substantive validity. Researchers must also consider whether the construct is well defined, whether items represent the domain, whether the model was theoretically justified in advance, whether the sample is appropriate, whether alternative models were considered, whether relations with external variables are theoretically consistent, and whether the scale functions across relevant populations. Structural evidence is important, but it remains only one component of the broader validity argument.

6.3 Why Statistical Significance Is Not Enough

Consider a correlation of r = .12 with p < .001. A researcher might write that "the significant correlation demonstrates convergent validity," a conclusion that is potentially misleading. A correlation of .12 may be statistically significant in a very large sample while providing relatively weak evidence for the proposed relationship. Figure 4 illustrates this mathematically: as sample size increases, even a small effect (r = .12) eventually crosses the conventional alpha = .05 threshold, while a moderate to large effect (r = .50) crosses it almost immediately. The p-value therefore reflects sample size as much as it reflects the strength of the underlying relationship.

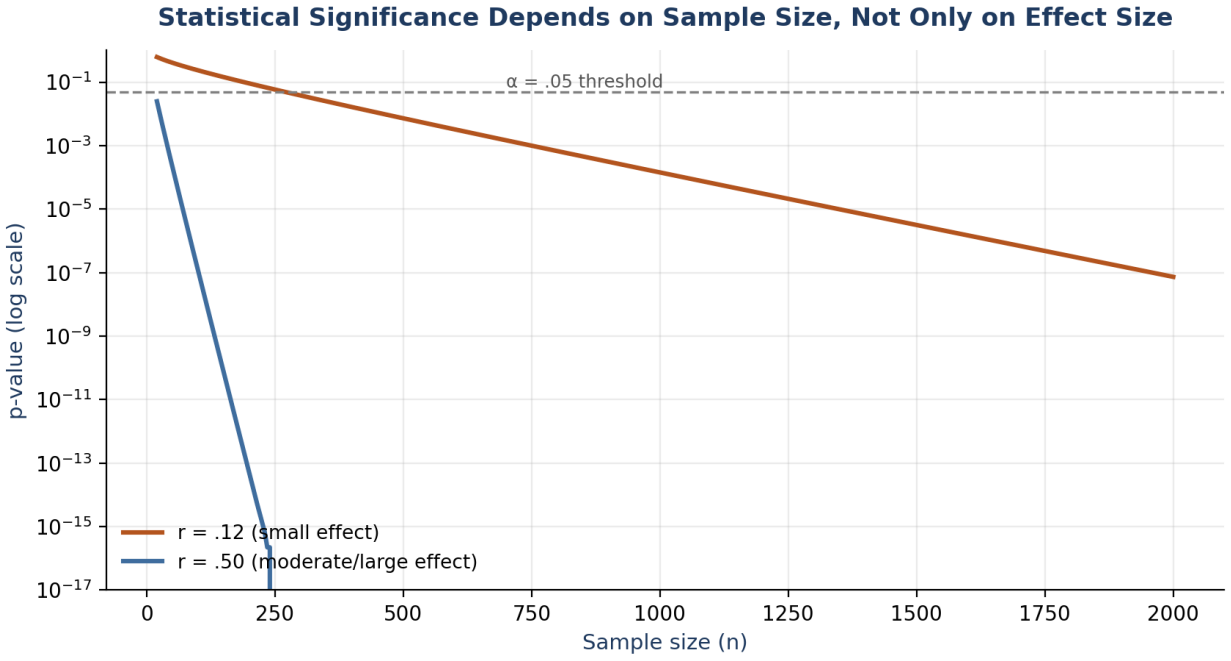

Illustration based on the standard formula relating a Pearson correlation to its t-statistic (t = r√(n-2) / √(1-r²)). A correlation as small as r = .12 becomes statistically significant once n exceeds roughly 270, underscoring why significance alone cannot substitute for a theoretically grounded validity argument.

Figure 4. Illustrative relationship between sample size and statistical significance for a

small (r = .12) and a moderate to large (r = .50) correlation. Significance is reached far sooner for the larger effect, and eventually for the smaller one as well once the sample is large enough.

Validity research should therefore consider effect size, theoretical expectation, direction, magnitude, confidence intervals, competing explanations, and measurement quality. The question is not merely whether a relationship was statistically significant, but whether the observed relationship was consistent with what theory predicted.

6.4 A Summary of Common Misconceptions

Misconception

More accurate statement

"Validity is a property of the questionnaire."

Validity concerns interpretations and uses of scores, not the instrument in isolation.

"Cronbach's alpha proves validity."

Alpha is reliability evidence; it says nothing directly about content, structural, or external validity.

"CFA proves validity."

CFA provides structural evidence, one source among several needed for a full argument.

"A significant correlation proves convergent validity."

The relationship must also be theoretically predicted and substantively meaningful, not merely significant.

"Expert agreement proves content validity."

Expert judgement is important, but must be systematic, criterion based, and documented.

"A translated questionnaire remains valid automatically."

Translation alone does not establish cross cultural validity; equivalence must be tested.

"Validity is either present or absent."

Validity evidence lies on a continuum and depends on the specific interpretation and use under consideration.

"One validation study validates an instrument forever."

Evidence should be accumulated across populations, contexts, and intended uses over time.

Table 3. Eight recurring misconceptions about validity and the more defensible statements that should replace them.

7. A Comprehensive Validity Evaluation Workflow

A rigorous validation project can be organised into eleven stages, moving from conceptual groundwork through to the integration of accumulated evidence. Table 4 sets out each stage together with its primary purpose.

Stage

Focus

Typical activities

1. Define the construct

Conceptual groundwork

Specify conceptual definition, boundaries, dimensions, and theoretical relationships

2. Map the content domain

Domain specification

Identify dimensions, facets, behavioural manifestations, and relevant contexts

3. Develop or select items
Item design
Ensure items adequately represent the conceptual domain
4. Evaluate response processes
Cognitive validity
Cognitive interviews, pilot testing, qualitative probing
5. Evaluate internal structure
Structural evidence
EFA, CFA, item response theory, bifactor modelling
6. Evaluate reliability
Precision
Internal consistency, temporal stability, inter-rater reliability, measurement error
7. Formulate validity hypotheses
Planning
Specify expected direction, magnitude, comparison groups, and related constructs in advance
8. Test relations with other variables
External evidence
Convergence, discrimination, criterion relationships, predictive relationships
9. Evaluate subgroup comparability
Fairness
Measurement invariance, DIF, cross cultural validity
10. Evaluate interpretation and use
Consequences
Assess whether evidence supports the intended application of scores
11. Integrate the evidence
Synthesis
Build a single, coherent argument rather than reporting isolated statistics

Table 4. An eleven-stage workflow for organising a validation project from construct definition to evidence integration.

The cumulative logic underlying this workflow is summarised visually in Figure 5.

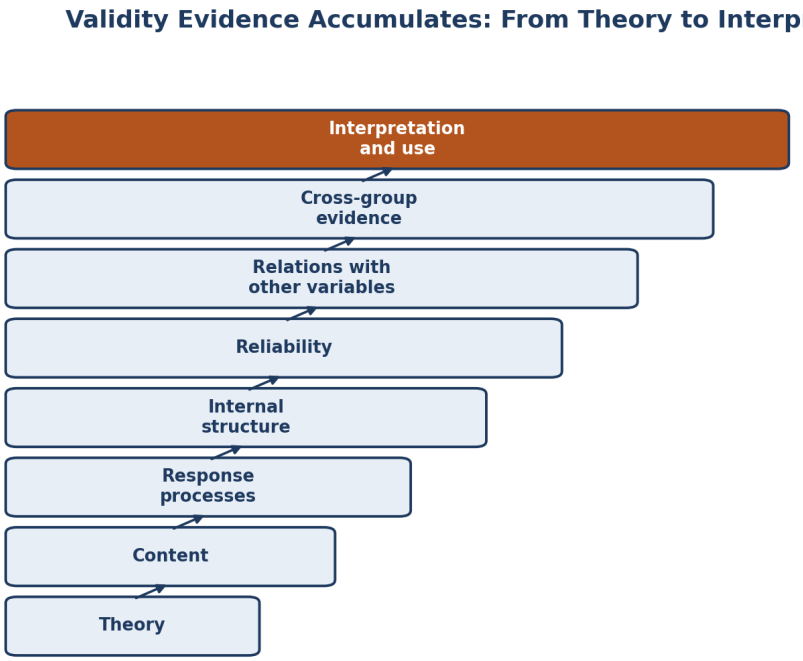

Figure 5. Validity evidence accumulates across successive stages, from theory through content, response processes, internal structure, reliability, and external relations, to cross-group evidence and, finally, interpretation and use.

8. Organising Validity Evidence: An Evidence Matrix

Researchers can organise accumulated evidence within a single matrix that links each validity question to its corresponding evidence type and typical analysis, as shown in Table 5. Presenting the evidence in this form makes explicit why validity is better understood as an evidence system than as a single pass or fail verdict.

Validity question

Evidence type

Possible analysis

Does content represent the construct?

Content evidence

Expert review, I-CVI, S-CVI, CVR

Do respondents understand items as intended?

Response process evidence

Cognitive interviews, think-aloud protocols

Does the internal structure fit theory?

Structural evidence

EFA, CFA, item response theory

Are scores consistent?

Reliability evidence

Cronbach's alpha, omega, intraclass correlation

Does the scale relate to similar constructs?

Convergent evidence

Correlation, structural equation modelling

Does it remain distinct from different constructs?

Discriminant evidence
HTMT, factor correlations, competing models
Does it predict relevant outcomes?
Criterion or predictive evidence
Regression, structural equation modelling
Does it work similarly across groups?
Invariance evidence
Multigroup CFA
Do individual items behave differently across groups?
DIF evidence
IRT based or logistic regression DIF models
Can scores be interpreted appropriately?
Interpretability evidence
Norms, thresholds, and score meaning studies

Table 5. A validity evidence matrix linking substantive validity questions to evidence types and typical analyses.

9. Reporting Validity in a Research Article

A weak report might state simply that "the instrument was valid and reliable." A considerably stronger report would instead state something closer to the following:

"Evidence supporting the intended interpretation of the Academic Resilience Scale was obtained from multiple sources. Content evaluation supported the relevance and comprehensiveness of the item pool. Confirmatory factor analysis supported the hypothesised three factor structure. Internal consistency estimates were satisfactory for the three dimensions. Correlations with academic self efficacy and psychological distress were consistent with pre-specified hypotheses, providing evidence based on relations with other variables. Multigroup confirmatory factor analysis further indicated whether the measurement structure was comparable across relevant demographic groups."

The second formulation demonstrates how the validity argument was constructed, rather than simply asserting its conclusion. A strong psychometric instrument rarely becomes defensible on the strength of a single analysis; evidence accumulates, as set out in Section 7, and a validation study should therefore be read as producing a body of evidence that supports, qualifies, or challenges specific interpretations rather than a bare "valid or invalid" verdict.

10. Implications for Psychological Researchers and for PsychtrixWeb

10.1 Implications for Researchers

Researchers should design validation studies as evidence building projects. Before collecting data, they should identify:

- What construct is being measured?
- What theoretical model defines it?
- What population is being studied?
- What interpretation will be made from the scores?
- What decisions will depend on the scores?
- What evidence is required to support those interpretations?
- What alternative explanations could challenge the interpretation?

This approach encourages researchers to move away from statistical ritual and toward a genuinely scientific programme of measurement.

10.2 Implications for PsychtrixWeb

The conceptual framework developed in this Research Note suggests that PsychtrixWeb should treat validity as a workflow rather than a button. A future psychometric analysis environment should allow researchers to specify the construct being measured, its subconstructs, the indicators representing each dimension, the population being assessed, the demographic variables requiring comparison, the theoretical hypotheses linking scores to other variables, and the psychometric evidence already gathered. This creates the foundation for a more advanced Psychometric Validity Workspace that could eventually integrate content validity, response-process evidence, EFA, CFA, reliability, convergent validity, discriminant validity, criterion related evidence, measurement invariance, DIF, item response theory, structural equation modelling, and automated psychometric reporting within a single, coherent interface.

Toward Evidence-Based Automated Psychometric Reporting

One of the most useful applications of PsychtrixWeb would be to transform statistical output into an evidence based validity narrative. Instead of simply producing a line of fit indices such as CFI = .96, RMSEA = .052, and SRMR = .041, the system could generate a structured interpretation along the following lines:

"The specified measurement model demonstrated acceptable global fit based on the selected indices. These findings provide evidence concerning the internal structure of the proposed instrument. Structural evidence should nevertheless be interpreted alongside content evidence, reliability estimates, and relationships with theoretically relevant external variables."

This distinction is critical: statistical output is not the same as psychometric interpretation, and a validity aware reporting engine should make that distinction explicit rather than leaving it to the reader to infer.

11. Conclusion

Validity is the central organising principle of psychological measurement, because psychological assessment involves making inferences from observed responses and scores. A rigorous validity argument cannot be reduced to Cronbach's alpha, a factor analysis, a correlation coefficient, or a single expert review exercise. Instead, researchers should integrate evidence concerning construct definition, content, response processes, internal structure, relations with other variables, reliability and measurement precision, cross cultural comparability, measurement invariance, differential item functioning, and the intended interpretation and use of scores.

As established in Research Note 001, psychological measurement begins with the relationship between theoretical constructs and observable indicators. Research Note 002 demonstrated that reliability concerns the consistency and precision of resulting scores. This Research Note extends that foundation by emphasising that consistent scores are not sufficient unless the interpretation made from those scores is adequately supported by evidence. The central principle is therefore that a psychometric instrument is not validated by one statistic: its intended score interpretations are supported by a cumulative body of theoretical and empirical evidence.

For researchers, this means that validation should be planned from the beginning of instrument development rather than added as an afterthought once statistical analysis is complete. For PsychtrixWeb, it means that psychometric analysis should evolve beyond isolated calculations toward an integrated evidence architecture for

psychological measurement. The next Research Note will translate these principles into practice by examining the complete process of developing and validating a psychological scale, from construct conceptualisation and item generation through pilot testing, factor analysis, reliability, validity, invariance, DIF, and final instrument evaluation.

Recommended Citation

Oladunmoye, E. O. (2026). Validity in psychological assessment: Evidence, interpretation, and common misconceptions (PsychtrixWeb Research Note No. 003). Psychtrix Initiative Limited.

References

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

2. 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

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

4. Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, 43(1), 115-135. https://doi.org/10.1007/s11747-014-0403-8

5. 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

6. Mokkink, L. B., Terwee, C. B., Knol, D. L., Stratford, P. W., Alonso, J., Patrick, D. L., Bouter, L. M., & de Vet, H. C. W. (2010a). The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: A clarification of its content. BMC Medical Research Methodology, 10, 22. https://doi.org/10.1186/1471-2288-10-22

7. 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. (2010b). The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. Quality of Life Research, 19, 539-549. https://doi.org/10.1007/s11136-010-9606-8

8. Oladunmoye, E. O. (2026). Reliability in psychological measurement: Concepts, estimation, interpretation, and reporting (PsychtrixWeb Research Note No. 002). Psychtrix Initiative Limited.

9. Oladunmoye, E. O. (2026). Understanding psychometric measurement: Constructs, variables, scores, and measurement error (PsychtrixWeb Research Note No. 001). Psychtrix Initiative Limited.

10. Oladunmoye, E. O., Muhammad T. S., (2024). Development and Validation of Multiple Intelligence Test among emerging adults in the United Kingdom. ISAR Journal of Arts, Humanities and Social Sciences, 2(4), 18-24.

11. Oladunmoye, E. O., (2015). Development and validation of social provision scale on first year undergraduate psychological adjustment. Journal of Education and Practice, 6 (28), 78-90.

12. Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. W., & Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: A Delphi study. Quality of Life Research, 27, 1159-1170. https://doi.org/10.1007/s11136-018-1829-0

13. Van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. Sage Publications.

14. Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1), 4-70. https://doi.org/10.1177/109442810031002

15. Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4(2), 223-233. https://doi.org/10.1080/15434300701375832

SUGGESTED CITATION

Oladunmoye, E. O. (2026). Validity in Psychological Assessment: Evidence, Interpretation, and Common Misconceptions. PsychtrixWeb Research Note, 005. Psychtrix Initiative Limited. https://www.psychtrixweb.online/research-notes/005-1-introduction