Scale Development in Research: A Methodological Framework

Amit Bag
Kolkata, India.
Email: amit18.sdi@gmail.com

Abstract

Scale development represents one of the most methodologically demanding tasks in quantitative research, requiring systematic attention to construct conceptualisation, item generation, factor structure, reliability, and validity. This narrative review synthesises foundational and contemporary methodological frameworks governing the development, validation, and reporting of research scales across the social, behavioural, health, and management sciences. Drawing on literature published between 1928 and 2024, the review traces the evolution of scaling methodology from classical approaches rooted in Thurstone's and Guttman's procedures to modern psychometric advances including item response theory, bifactor modelling, and structural equation modelling. Key procedural stages — domain specification, item generation, pilot testing, factor analysis, reliability assessment, and multidimensional validity evaluation — are examined in depth, with particular attention to the distinction between reflective and formative measurement models. The review also addresses measurement invariance testing, cross-cultural adaptation, and common method bias as contemporary challenges in scale validation research. Emerging reporting standards and best-practice recommendations are discussed to encourage methodological rigour and transparency in published scale development research. The review highlights persistent gaps, including inconsistent validity reporting, the over-reliance on Cronbach's alpha as a sole indicator of reliability, and the under-utilisation of item response theory in social science contexts. By synthesising these streams of inquiry, this review provides a structured methodological reference for researchers engaged in original scale development and those critically evaluating existing measurement instruments.

Keywords: *Scale development; psychometrics; construct validity; reliability; factor analysis; item response theory; measurement invariance; classical test theory.*

1. Introduction

1.1 Background and Significance

The measurement of abstract psychological, social, and behavioural constructs lies at the core of empirical inquiry in the social sciences, public health, education, management, and related disciplines. Researchers routinely seek to quantify phenomena such as attitudes, personality traits, perceptions, motivations, and subjective well-being — constructs that, by their nature, cannot be observed directly but must instead be inferred from observable indicators. The instrument through which this inference is made is the measurement scale, which aggregates multiple items into a composite score assumed to reflect an underlying latent variable (Churchill, 1979; Clark & Watson, 1995; Shukla, 2023). Measurement refers to the process of assigning numbers or labels to objects, individuals, or events according to specific rules. In social and behavioural sciences, many constructs of interest—such as attitudes, perceptions, satisfaction, or motivation—are abstract and cannot be directly observed. Scaling techniques provide systematic ways to represent these constructs numerically, allowing researchers to summarise data, test hypotheses, and draw inferences (Sarun, 2026). The

quality of scientific knowledge in these fields is therefore inextricably linked to the quality of the scales employed in data collection.

A poorly constructed scale can introduce systematic error into a study, bias parameter estimates, and ultimately generate misleading conclusions that propagate through the scientific literature (MacKenzie et al., 2011). Conversely, a rigorously developed and validated scale serves as a scientific asset — reusable across studies, populations, and contexts — and provides a reliable channel through which theoretical constructs can be operationalised and tested (Messick, 1995; Simms, 2008). Scales are data collection tools that can measure characteristics such as knowledge, emotion, interest, perception, attitude, belief, disposition, risk, quality of life and behavior. The scale development process includes determining the theoretical foundations, creating the items, pilot study, validity and reliability analysis, and final implementation (Gültürk et al., 2024). Given this foundational importance, scale development has attracted sustained methodological attention across disciplines, resulting in a rich body of guidance literature that has evolved considerably over the past century.

The early history of attitudinal measurement is marked by the contributions of Thurstone (1928), who proposed systematic procedures for constructing interval-level attitudinal scales using judges' ratings, and Guttman (1944), who developed the cumulative scaling model to assess whether items form a hierarchical continuum. For nearly a century, social researchers and psychologists employing psychometric scales have investigated attitude measurement, primarily deliberating between Thurstone scales of equal appearing intervals and Likert's summated rating scales (Abd ElHafeez et al., 2022; Symeonaki et al., 2024). The mid-twentieth century saw the formalisation of Classical Test Theory (CTT), which provided a conceptual and statistical basis for understanding measurement error and reliability (Cronbach, 1951). Later decades introduced Item Response Theory (IRT) as a more flexible psychometric framework, and exploratory and confirmatory factor analytic techniques emerged as central tools for evaluating the internal structure of scales (Fabrigar et al., 1999; Anderson & Gerbing, 1988).

Despite the availability of well-articulated methodological guidance, there remains substantial variability in the rigour with which scales are developed and reported in published research. Studies have documented widespread practices including insufficient item generation, failure to conduct confirmatory factor analysis, exclusive reliance on Cronbach's alpha, and inadequate reporting of validity evidence (Flake & Fried, 2020; Worthington & Whittaker, 2006). These practices carry substantive consequences for the quality and cumulativeness of empirical knowledge, and they underscore the continued need for comprehensive, accessible syntheses of scale development methodology.

1.2 Scope and Objectives of the Review

This review aims to provide a comprehensive and integrated overview of the methodological framework for scale development in research. It specifically addresses conceptual foundations, procedural stages, analytical methods, and validity considerations relevant to the development of new scales and the adaptation of existing instruments. The review is primarily directed at researchers in the social, behavioural, health, and management sciences who are engaged in measurement research or who wish to evaluate the methodological quality of published scales. The following specific objectives guide the review: (a) to trace the historical and conceptual evolution of scaling methodology; (b) to examine the major procedural phases of scale development from domain specification to validation; (c) to review classical and modern psychometric frameworks applicable to item and scale analysis; (d) to discuss reliability and validity evidence in the context of contemporary measurement standards; (e) to address cross-cultural adaptation, measurement invariance, and common method bias; and (f) to identify persistent methodological challenges and offer recommendations aligned with current best practices. This review does not attempt to evaluate the quality of any specific published scale; rather, it offers a methodological synthesis that cuts across disciplinary contexts and measurement traditions.

2. Methods for Literature Selection

2.1 Search Strategy and Databases

This review was conducted as a narrative literature review rather than a systematic review. A systematic review demands a pre-registered protocol, exhaustive search, and formal quality appraisal, and is primarily suited to answering narrowly defined empirical questions across a homogenous literature (Green et al., 2006). The

present topic — scale development methodology — spans multiple disciplines, evolves over decades, and encompasses methodological guidance that does not lend itself readily to formal pooling or meta-analysis. A narrative approach is therefore more appropriate, as it allows theoretical integration, historical contextualisation, and flexible synthesis across heterogeneous methodological traditions (Ferrari, 2015).

Literature was identified through searches conducted in the following academic databases: Web of Science, Scopus, PsycINFO, MEDLINE/PubMed, Google Scholar, ERIC (Education Resources Information Center), the Social Science Research Network (SSRN), CINAHL (Cumulative Index to Nursing and Allied Health Literature), and the Business Source Complete database. Search strings were constructed around the following core terms: "scale development," "instrument development," "psychometric validation," "construct validity," "item generation," "factor analysis measurement," "classical test theory," "item response theory," "Rasch model," "measurement invariance," "cross-cultural adaptation," and "reliability assessment." Boolean operators (AND, OR) were used to combine terms, and searches were restricted to peer-reviewed journal articles. The primary date range covered publications from 1980 to 2026, although seminal foundational works predating 1980 were deliberately included given their continuing theoretical significance to the field.

2.2 Inclusion and Exclusion Criteria

Articles were included if they presented original theoretical frameworks, empirical illustrations, methodological reviews, or critical commentaries pertaining to scale development, psychometric analysis, or measurement validation. Only peer-reviewed journal articles were eligible for inclusion. Books, conference papers, grey literature, theses, and technical reports were excluded to ensure the quality and verifiability of the cited evidence base. Articles were restricted to those published in English. Studies dealing exclusively with clinical diagnostic instruments or biomarker-based measures outside the psychometric tradition were excluded as tangential to the primary focus of the review. Articles that employed scale-type instruments solely as measurement tools, without contributing methodological insight into scale development or validation, were similarly excluded.

2.3 Study Screening and Selection Workflow

Initial searches generated several thousand candidate records across the databases consulted. Duplicates arising from the overlap of records across databases were identified through title-level comparisons and removed prior to further screening. Titles and abstracts of the remaining records were screened for relevance to scale development methodology. Articles retained at this stage were assessed at full-text level to confirm their eligibility against the stated inclusion and exclusion criteria. Priority was given to methodological review articles, empirically influential studies with high citation counts, and papers published in leading journals in psychometrics, applied psychology, organisational behaviour, health measurement, and marketing research. "Influential" studies were operationalised as those meeting the following criteria: a) published in a highly ranked peer-reviewed journal in the relevant field with high citation counts b) addressing a foundational theoretical or methodological question with broad applicability; or c) representing a methodological advance that has been incorporated into subsequent best-practice guidelines. The final reference set comprises peer-reviewed journal articles spanning classical and contemporary contributions, ensuring both historical depth and methodological currency. No language translation was performed, and non-English language publications were excluded.

3. Conceptual Foundations of Scale Development

3.1 Defining Scales and Constructs

A scale, in the psychometric sense, refers to a standardised measurement instrument comprising a set of items — usually questions or statements — that are aggregated to yield a composite score intended to represent a latent construct. A construct is a theoretical abstraction — such as intelligence, job satisfaction, or perceived service quality — that is not directly observable but whose existence is inferred from observable indicators (Borsboom et al., 2004). The adequacy of a scale depends not only on the quality of its individual items but on the degree to which the entire instrument faithfully represents the theoretical domain it purports to measure (Clark & Watson, 1995; Messick, 1995).

Constructs in the social sciences vary in their ontological character. Some are unidimensional, meaning that a single underlying factor accounts for the covariation among items. Others are multidimensional, comprising

distinct but related facets that together constitute the full construct domain. The distinction has important implications for scale development, because unidimensional and multidimensional scales require different analytical strategies and different interpretations of reliability indices (McDonald, 1970; Reise et al., 2013). Researchers must therefore specify the dimensionality of the target construct before embarking on item generation and must employ appropriate analytical methods to test dimensional assumptions empirically rather than merely assuming them.

3.2 Historical Evolution of Scaling Methods

The systematic development of attitude scales began in earnest in the late 1920s with Thurstone (1928), who introduced the conceptual basis for equal-appearing intervals measurement, in which attitudinal statements are evaluated by judges and ordered along a continuum to produce an interval-level scale. Guttman (1944) subsequently proposed cumulative scaling, wherein items are ordered by difficulty such that endorsement of a given item implies endorsement of all less extreme items — a hierarchical structure that permits evaluation of scalability through the coefficient of reproducibility. These classical approaches, while theoretically sophisticated, were resource-intensive and gave way to summated rating approaches, which became the dominant format for psychometric scales due to their practical simplicity and psychometric tractability (Simms, 2008).

By the mid-twentieth century, Loevinger (1957) had articulated a comprehensive framework for scale construction that emphasised the alignment between the substantive content of items, their structural relationships, and the external validity of the resulting scores. This framework anticipated many of the multi-stage validation procedures that became standard in subsequent decades. The seminal work of Churchill (1979), drawing on procedures in marketing research, codified a sequential paradigm — domain specification, item generation, data collection, purification, and validation — that has remained foundational across disciplines. The introduction of structural equation modelling and confirmatory factor analysis further transformed scale validation by enabling the simultaneous evaluation of measurement model fit, factor loadings, and construct interrelationships (Anderson & Gerbing, 1988). Simms (2008) provided a useful integrative review of both classical and modern scale construction methods, demonstrating that each era's methodological advances built upon — rather than entirely replaced — earlier contributions.

3.3 Reflective Versus Formative Measurement Models

A critical but frequently overlooked distinction in scale development is whether the measurement model is reflective or formative. In a reflective model, the latent construct is conceptualised as the common cause of its indicators: all items are manifestations of the underlying construct, they are expected to be internally consistent, and removing an item does not alter the construct being measured. Most scales in psychology and social science adopt a reflective model (Jarvis et al., 2003). In a formative model, by contrast, the indicators jointly define and constitute the construct: they are causes rather than effects of the construct, they need not be correlated with one another, and each item represents a distinct dimension of the construct such that removing an item changes the construct's definition (Coltman et al., 2008).

Jarvis et al. (2003) demonstrated through a systematic review of top marketing journals that a substantial proportion of published studies incorrectly specified reflective models for constructs that were conceptually formative, resulting in biased parameter estimates and misleading validity conclusions. Coltman et al. (2008) further illustrated the consequences of model misspecification by demonstrating that formative and reflective operationalisations of the same construct produce different empirical results. Researchers are advised to consider four criteria when distinguishing between models: the causal priority of construct versus indicators, covariation among indicators, interchangeability of indicators, and the expected nomological network of antecedents and consequences. Failure to resolve this question at the outset of scale development can compromise the entire subsequent validation process.

4. Item Generation and Content Specification

4.1 Theoretical Grounding and Domain Specification

The first substantive step in scale development is the precise specification of the construct's conceptual domain. This involves delineating the boundaries of the construct — what it includes and what it excludes — and establishing the theoretical basis for its dimensionality. Without a well-developed conceptual framework, item generation becomes arbitrary and the resulting scale may systematically underrepresent or misrepresent the target construct (Churchill, 1979; Clark & Watson, 1995). Researchers are advised to conduct a thorough review of the theoretical and empirical literature to identify the conceptual space the scale is intended to cover, the relationships between the construct and related concepts, and the specific manifestations or facets that constitute the construct's domain.

Hinkin (1998) identified three principal sources from which items may be generated: deductive methods, in which items are derived directly from the theoretical definition; inductive methods, in which items are generated from observations or interviews with members of the target population; and extrapolation from existing scales. Each source has strengths and limitations. Deductive methods yield items with strong theoretical grounding but may produce highly abstract or jargon-laden language that fails to resonate with respondents. Inductive methods yield ecologically valid items but may lack theoretical coherence. Extrapolation from existing instruments can accelerate development but risks importing the conceptual limitations of the source scale. A combination of approaches is generally recommended to maximise content coverage and respondent accessibility (Hinkin, 1998; Simms, 2008).

4.2 Item Writing and Format Considerations

Once the domain has been specified, a large initial pool of items should be generated — considerably more than will appear in the final scale. Churchill (1979) recommended generating a pool considerably larger than the intended final instrument, as a substantial proportion of items will typically be eliminated during pilot testing and statistical purification. Items should be written at an appropriate reading level for the target population, should avoid double-barrelled statements combining two distinct ideas in a single item, and should be unambiguous in their referent (Krosnick, 1999; Schwarz, 1999). The response format — most commonly a summated rating scale with anchors ranging from strong disagreement to strong agreement — should be chosen with attention to the number of response categories, the presence or absence of a midpoint, and the labelling of scale points (Krosnick, 1999).

The psychological processes underlying survey responding — comprehension, retrieval, judgement, and mapping onto a response scale — introduce potential sources of measurement error that item writers must seek to minimise (Schwarz, 1999). Cognitive interviewing with a small convenience sample is a recommended pre-pilot technique for identifying items that respondents find confusing, ambiguous, or emotionally charged. Content validity, which refers to the degree to which the item pool adequately represents the construct's full domain, is typically evaluated through a panel of subject-matter experts who rate items for relevance and representativeness (Clark & Watson, 1995; MacKenzie et al., 2011). Quantitative indices such as the Content Validity Ratio may be calculated to support item selection decisions at this stage, providing a transparent record of the content review process for the purposes of scientific reporting.

5. Classical Test Theory in Scale Development

5.1 Principles of Classical Test Theory

Classical Test Theory (CTT) constitutes the dominant psychometric framework in the social sciences and provides the statistical foundation for most scale development procedures. The core premise of CTT is that any observed score on a measurement instrument can be decomposed into a true score and a random error component (Cronbach, 1951; Loevinger, 1957). The true score represents the respondent's stable underlying standing on the construct, while the error component captures random fluctuations due to irrelevant factors such as fatigue, distraction, or temporary changes in motivational state. Reliability, within CTT, is defined as the proportion of observed score variance attributable to true score variance — a ratio that ranges from zero (entirely error) to one (perfectly reliable) (Cronbach, 1951).

A key property of CTT is that item and scale statistics are sample-dependent: item difficulty indices, item-total correlations, and reliability coefficients will vary across samples from different populations. This property limits the generalisability of CTT-based item statistics and motivates the use of more modern frameworks such as IRT,

which is discussed in Section 7. Nevertheless, CTT remains widely used because of its relative simplicity and the modest sample sizes it requires for basic item analysis. Within a CTT framework, item analysis typically involves examining item means, standard deviations, endorsement rates, and corrected item-total correlations to identify items that are too extreme in difficulty, that do not discriminate adequately between high and low scorers on the total scale, or that reduce rather than increase internal consistency (Hinkin, 1998; Simms, 2008).

5.2 Item Analysis under Classical Test Theory

Item purification under CTT proceeds by computing corrected item-total correlations — correlations between each item and the sum of remaining items — and examining Cronbach's alpha if each item is deleted. Items with very low or negative item-total correlations, or items whose deletion substantially increases alpha, are candidates for removal (Churchill, 1979). A corrected item-total correlation of 0.30 or above is commonly used as a minimum threshold, though this criterion is somewhat arbitrary and should be interpreted in the context of the construct's theoretical breadth and the intended coverage of item content (Clark & Watson, 1995).

It is important to recognise that item purification in CTT optimises for internal consistency, which can inadvertently narrow the content coverage of a broad construct by retaining only items that are highly intercorrelated. This trade-off between internal consistency and content representativeness is a well-recognised limitation of CTT-based item selection procedures (Loevinger, 1957; Clark & Watson, 1995). Researchers who develop scales for broad constructs with multiple facets should exercise caution in applying uniform item-total correlation thresholds and should anchor purification decisions in theory rather than statistical criteria alone. A theoretically motivated approach to item retention ensures that the final scale adequately samples the construct's full conceptual domain rather than merely maximising a statistical coefficient.

6. Factor Analytic Approaches

6.1 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) is typically conducted on a calibration sample to examine the latent factor structure underlying the item pool, without imposing a priori constraints on the pattern of factor loadings. The goal is to identify how many factors are needed to account for item intercorrelations and which items load substantially on each factor (Fabrigar et al., 1999). Best-practice recommendations for EFA include using principal axis factoring or maximum likelihood estimation as the extraction method, with oblique rotation — such as Promax or Direct Oblimin — when factors are expected to be correlated, as is typically the case in social science scales (Fabrigar et al., 1999; Simms, 2008).

The determination of the number of factors to retain is one of the most consequential and contested decisions in EFA. Common approaches include the eigenvalue-greater-than-one rule, visual inspection of the scree plot (Cattell, 1966), and Horn's (1965) parallel analysis, which compares obtained eigenvalues with those derived from random data of the same dimensions. Parallel analysis has been identified as superior to the Kaiser criterion and scree plot in simulation studies (Zwick & Velicer, 1986), yet it remains underutilised in published research. The minimum factor loading threshold for item retention is conventionally set at 0.40, though researchers should also attend to cross-loadings — items loading substantially on more than one factor — which may indicate conceptual ambiguity in item content (Fabrigar et al., 1999).

It is advisable to conduct EFA and Confirmatory Factor Analysis (CFA) on separate samples to avoid capitalising on sample-specific idiosyncrasies. When sample size does not permit sample splitting, researchers should clearly acknowledge this limitation and treat EFA results as provisional pending cross-validation (Fabrigar et al., 1999; MacKenzie et al., 2011). The appropriate sample size for EFA has been debated, with recommendations ranging from a minimum of five to ten respondents per item to absolute thresholds of 200 or more participants, though the adequacy of a given sample size depends on the communality of the items and the clarity of the factor structure.

6.2 Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is employed to test a specific, theoretically derived factor structure against empirical data. Unlike EFA, CFA requires researchers to specify in advance which items load on which factors and, where appropriate, which factors are correlated. The degree to which the specified model reproduces the observed covariance matrix is assessed via goodness-of-fit indices (Anderson & Gerbing, 1988; Hu & Bentler, 1999). Hu and Bentler (1999) established widely adopted cutoff values for common fit indices: a Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) of 0.95 or above, and a Root Mean Square Error of Approximation (RMSEA) of 0.06 or below, are suggestive of acceptable model fit, though these thresholds are heuristics rather than absolute standards and should be interpreted alongside theoretical considerations and the complexity of the model being evaluated.

Anderson and Gerbing (1988) recommended a two-step approach in which the measurement model is fully validated via CFA before the structural (path) model is estimated. This separation ensures that any observed relationships between constructs are not confounded by measurement model misspecification. CFA also provides estimates of Average Variance Extracted (AVE) and Composite Reliability (CR), which serve as indices of convergent and discriminant validity respectively (Fornell & Larcker, 1981). Convergent validity is supported when AVE exceeds 0.50, indicating that more variance in the indicators is explained by the latent construct than by measurement error. Discriminant validity is supported when the square root of AVE for each construct exceeds its correlations with all other constructs — a criterion that has remained influential despite subsequent criticisms (Henseler et al., 2015).

Henseler et al. (2015) proposed the Heterotrait-Monotrait (HTMT) ratio of correlations as an improved criterion for discriminant validity in variance-based structural equation modelling contexts, arguing that the Fornell-Larcker criterion lacks sensitivity when factor loadings are uniform across items. The HTMT criterion has gained traction in recent years as a more conservative and reliable indicator of discriminant validity (Henseler et al., 2015). Values below 0.85 or 0.90 are generally regarded as indicative of discriminant validity, depending on whether the constructs are conceptually distinct or more closely related. The adoption of HTMT alongside the Fornell-Larcker criterion provides a more comprehensive evaluation of discriminant validity than either indicator alone.

6.3 Bifactor Models

The bifactor model is an increasingly used extension of standard CFA that accommodates the simultaneous presence of a general factor and domain-specific factors. In a bifactor specification, each item loads on both a general latent factor — which captures the construct's common core — and one of several group factors representing distinct facets or dimensions. This structure is particularly appropriate when a scale is intended to be scored both as a total composite and as subscales (Reise, 2012; Reise et al., 2013). The bifactor model provides a means of evaluating the relative contributions of the general and specific factors to item variance, thereby informing whether total scores, subscale scores, or both are psychometrically defensible.

Reise (2012) demonstrated through simulation and empirical work that many scales commonly treated as unidimensional are in fact more adequately described by a bifactor structure, and that misspecification of dimensionality can distort reliability estimates and score interpretations. Indices derived from bifactor solutions — including omega hierarchical and omega total — are increasingly advocated as more informative reliability indicators than Cronbach's alpha for multidimensional scales (Reise et al., 2013; McNeish, 2018). These omega indices partition total scale variance into components attributable to the general factor, the group factors, and uniqueness, providing a more transparent account of what scale scores actually capture. The adoption of bifactor modelling in applied scale research remains uneven, however, and represents an area in which methodological practice lags substantially behind available statistical knowledge.

7. Item Response Theory and Rasch Modelling

7.1 Foundations of Item Response Theory

Item Response Theory (IRT) represents a paradigmatic shift from CTT in that it models the relationship between an individual's standing on a latent trait and the probability of a particular item response, rather than treating item statistics as properties of a specific sample (Reise et al., 1993; Simms, 2008). The family of IRT models — including the one-parameter (Rasch) model, the two-parameter logistic model, and the three-

parameter logistic model — each specify the item characteristic curve in terms of parameters such as item difficulty, discrimination, and guessing probability. Because these parameters are estimated independently of the sample's trait distribution given adequate sample size and model fit, IRT provides sample-invariant item statistics — a significant advantage over CTT for instrument development and item bank calibration.

A central application of IRT in scale development is computerised adaptive testing, in which items are selected dynamically based on the respondent's estimated trait level, yielding precise measurement with fewer items than fixed-length scales. Beyond adaptive testing, IRT is used in standard scale development to identify items with inadequate discrimination, to evaluate differential item functioning (DIF) — the possibility that an item performs differently for distinct demographic groups even after controlling for the latent trait — and to calibrate item banks for subsequent assembly of parallel forms (Reise et al., 1993). DIF analysis is particularly valuable in cross-group scale applications, as it identifies items whose functioning may be group-specific rather than universal, thereby supplementing the measurement invariance analyses conducted within a CFA framework.

7.2 Rasch Modelling as a Special Case

The Rasch model, which can be regarded as the one-parameter IRT model, occupies a distinctive position in psychometric theory in that it specifies a particular mathematical form for the item-trait relationship and tests whether the data conform to it (Reise et al., 1993). Rather than selecting a model that best fits the data, the Rasch approach constructs items to fit the model, thereby achieving the specific objectivity and measurement invariance that are considered prerequisite properties of a formal measurement system. In the Rasch framework, each item is characterised solely by its difficulty parameter, and items that exhibit differential discrimination are flagged as misfitting and subject to revision or removal.

Rasch modelling has been widely applied in health outcomes measurement, educational testing, and clinical assessment, and it offers powerful diagnostic tools for evaluating the dimensionality and ordering properties of scales. However, its application in social science research more broadly remains relatively limited compared to CTT-based methods, in part because of the stricter assumptions it imposes and the difficulty of finding empirical data that fully satisfy those assumptions. Researchers considering Rasch analysis should evaluate whether the construct of interest is amenable to the unidimensionality and local independence assumptions that the model requires, as violations of these assumptions produce misleading parameter estimates and incorrect model-fit conclusions (Simms, 2008).

8. Reliability Assessment

8.1 Internal Consistency Reliability

Reliability refers to the degree to which a measurement instrument yields consistent results across repeated administrations, raters, or items (Cronbach, 1951). Internal consistency reliability, which reflects the degree to which the items of a scale cohesively measure the same underlying construct, is the most commonly reported form of reliability in social science scale research. Cronbach's alpha — the most widely used internal consistency index — estimates reliability as a function of the average inter-item correlation and the number of items in the scale (Cronbach, 1951). Although alpha has achieved near-universal adoption, it has been subjected to sustained and increasingly compelling criticism in the psychometric literature.

Sijtsma (2009) provided a comprehensive critique of Cronbach's alpha, demonstrating that it is a lower bound to true reliability only under the strict assumption of essentially tau-equivalent measurement — a condition rarely met in practice — and that it is sensitive to the number of items irrespective of their quality. McNeish (2018) showed that alpha systematically underestimates reliability for congeneric scales where item loadings are unequal, and recommended that omega coefficients derived from factor models should replace alpha as the default internal consistency index. McDonald (1970) originally proposed omega total as a reliability estimate that does not assume tau-equivalence, and Flora (2020) provided a comprehensive tutorial on its computation, arguing that omega hierarchical is particularly appropriate for scales where a bifactor structure is plausible. Raykov (1997) proposed composite reliability as an index computed from CFA-based factor loadings and error variances, which provides a more accurate reliability estimate for congeneric measures. The persistence of Cronbach's alpha as the default reliability index in published research, despite these well-documented

alternatives, represents one of the most pervasive methodological deficiencies in contemporary scale research (Flake & Fried, 2020; McNeish, 2018).

8.2 Other Forms of Reliability

Beyond internal consistency, scale validation studies should attend to test-retest reliability, which assesses the stability of scores across two occasions for individuals who have not changed on the underlying construct. Test-retest reliability is particularly important for scales intended to measure stable traits, as opposed to state-like constructs that are inherently variable over time (Hinkin, 1998). The interval between test occasions should be chosen to be long enough to minimise memory-driven consistency but short enough to preclude genuine change on the construct — a judgement that requires knowledge of the construct's temporal stability and is therefore necessarily guided by theory.

Inter-rater reliability is relevant when scale responses require coding or rating by human evaluators. Agreement between independent raters is typically quantified using Cohen's kappa or intraclass correlation coefficients, depending on whether the scale of measurement is nominal or continuous. For self-report scales, the distinction between test-retest reliability and true score stability becomes important, as observed instability may reflect genuine change rather than measurement error. Researchers should therefore use appropriate longitudinal models that distinguish between these sources of variance when evaluating temporal stability (MacKenzie et al., 2011). The comprehensive assessment of reliability across multiple facets — rather than exclusive reliance on internal consistency — provides a more complete picture of an instrument's measurement quality.

9. Validity Evidence in Scale Development

9.1 Content and Face Validity

Validity, in contemporary psychometric theory, is understood not as a property of the scale itself but as the degree to which accumulated evidence supports specific interpretations of scale scores for particular purposes (Messick, 1995; Borsboom et al., 2004). This unified conceptualisation, articulated most influentially by Messick (1995), replaced earlier categorical distinctions between content, criterion, and construct validity with a single evaluative framework centred on the construct validity of score interpretations. Within this framework, content validity evidence — drawn from expert judgement about the representativeness of the item pool — constitutes an important but not sufficient basis for validity claims.

Content validity refers to the degree to which the items of a scale adequately sample from the full conceptual domain of the target construct. It is evaluated through expert panel review, in which subject-matter experts assess the relevance and representativeness of items. Face validity, by contrast, refers to the superficial appearance of items as measuring the intended construct from the perspective of respondents — a weaker form of evidence that does not substitute for empirical validity evaluation. Both content and face validity are assessed qualitatively or through simple quantitative indices prior to empirical data collection, and their adequacy is a precondition for further psychometric analysis (MacKenzie et al., 2011; Clark & Watson, 1995). The use of multiple expert reviewers with clearly defined rating criteria strengthens the rigour and reproducibility of content validation procedures.

9.2 Construct Validity

Construct validity encompasses the evidence that a scale measures the theoretical construct it is designed to measure, as distinct from other related or unrelated constructs. It is evaluated through multiple lines of convergent and discriminant evidence (Anderson & Gerbing, 1988; Fornell & Larcker, 1981). Convergent validity is demonstrated when a scale's scores correlate substantially with scores from other established measures of the same or closely related constructs. Discriminant validity is demonstrated when the scale's scores are relatively weakly correlated with measures of theoretically distinct constructs (MacKenzie et al., 2011). The simultaneous evaluation of convergent and discriminant validity within a multi-trait, multi-method matrix or a CFA framework provides the most compelling construct validity evidence.

Flake et al. (2017) found that construct validation practices in social psychological research are frequently incomplete: most studies reported only limited validity evidence, with convergent and discriminant validity

often tested through a single set of analyses rather than systematically across samples and conditions. Borsboom et al. (2004) argued that for construct validity to be properly established, a causal connection must exist between the latent construct and its indicators — a philosophical position with implications for how validity evidence is interpreted and what kinds of theoretical claims a validated scale can support. Known-groups validity, in which scores are compared between groups known to differ on the construct, provides an additional and often underutilised source of construct validity evidence.

9.3 Criterion-Related Validity

Criterion-related validity refers to the degree to which scale scores predict or correlate with an external criterion that is conceptually related to the construct. Concurrent validity is assessed when scale scores and criterion scores are collected simultaneously, while predictive validity involves the criterion being assessed at a later time point. The selection of an appropriate criterion is theoretically guided — the criterion should be a theoretically expected outcome, antecedent, or correlate of the construct, with the expected magnitude and direction of the relationship specified in advance (Messick, 1995; MacKenzie et al., 2011). Criterion-related validity evidence is strengthened when the obtained relationships are replicated across independent samples and when competing explanations for the observed associations — such as shared method variance — can be ruled out through appropriate design and analytical controls.

10. Measurement Invariance and Cross-Cultural Adaptation

10.1 Measurement Invariance Testing

A scale that performs differently across demographic groups, cultures, or time points violates the principle of measurement invariance, meaning that apparent differences in construct scores between groups may reflect differential functioning of the measurement instrument rather than genuine differences in the underlying construct (Vandenberg & Lance, 2000; Cheung & Rensvold, 2002). Testing for measurement invariance has become a standard requirement in comparative research and is typically conducted within a multi-group CFA framework that evaluates progressively constrained models of the factor structure.

Vandenberg and Lance (2000) outlined a hierarchical sequence of measurement invariance models: configural invariance (the same factor structure holds across groups), metric invariance (equal factor loadings across groups), scalar invariance (equal item intercepts across groups), and strict invariance (equal residual variances across groups). Each level imposes progressively stronger constraints, and each is associated with a specific comparison question. For meaningful comparison of factor means across groups, at minimum scalar invariance must be established, although partial scalar invariance — where at least two items per factor demonstrate invariant loadings and intercepts — may be sufficient under some conditions (Millsap & Kwok, 2004; Putnick & Bornstein, 2016). Cheung and Rensvold (2002) proposed that a change in CFI of less than or equal to 0.01 between nested invariance models is an acceptable criterion for affirming the more constrained model, a guideline that has been widely adopted in applied research.

Dimitrov (2010) provided a practical framework for testing factorial invariance in the context of construct validation, arguing that invariance testing should be regarded as an integral component of the scale validation process rather than an optional supplementary analysis. Putnick and Bornstein (2016) surveyed developmental psychology literature and found that while measurement invariance testing was increasingly reported, a substantial proportion of comparative studies still proceeded without it — an oversight that may render cross-group comparisons uninterpretable. The systematic integration of invariance testing into scale validation protocols is therefore advocated as a minimum standard for research that employs scales across distinct groups or time points.

10.2 Cross-Cultural Adaptation Procedures

When scales developed in one linguistic or cultural context are adapted for use in another, a structured adaptation process is required to preserve the conceptual equivalence of the instrument across cultures (Beaton et al., 2000). The widely adopted guidelines proposed by Beaton et al. (2000) specify a five-stage procedure: forward translation by two independent bilingual translators, synthesis of the two translations, backward

translation by two translators blind to the original, expert committee review, and pre-testing with a small sample from the target population. This procedure addresses both linguistic equivalence — ensuring that translated items carry the same literal meaning — and conceptual equivalence — ensuring that the construct operationalised by the original scale is culturally valid in the target context.

It is important to note that linguistic adaptation alone does not guarantee conceptual equivalence: constructs may carry different cultural meanings, response styles may differ systematically across cultures, and the nomological network of a construct may vary across cultural contexts (Beaton et al., 2000; Cheung & Rensvold, 2002). After linguistic adaptation, the cross-cultural equivalence of the adapted scale should be evaluated through measurement invariance testing, ideally with samples drawn from both the source and target cultures. Differential item functioning analysis using IRT provides a complementary method for identifying specific items that behave non-equivalently across cultural groups (Reise et al., 1993). A complete cross-cultural adaptation project therefore integrates translation procedures, expert review, cognitive interviewing, and psychometric testing within a unified validation framework.

11. Advanced Considerations in Scale Development

11.1 Common Method Bias

Common method bias (CMB) refers to spurious variance attributable to the measurement method rather than to the constructs the measures purport to represent. It is a particularly salient concern in single-source, self-report studies where both predictor and criterion variables are collected from the same respondent at the same time using the same response format (Podsakoff et al., 2003). CMB can artificially inflate observed correlations between constructs, thereby producing spuriously significant relationships and threatening the validity of study conclusions. Podsakoff et al. (2003) provided a comprehensive review of CMB sources and remedies, distinguishing between procedural and statistical approaches to its management.

Procedural remedies include temporal separation of predictor and criterion measurement, use of multiple data sources, and variation in scale format and length to reduce respondent fatigue and response sets. Statistical remedies include Harman's single-factor test, the unmeasured latent factor method, and the correlated uniqueness model, though these approaches have limitations and should be regarded as partial safeguards rather than definitive solutions (Podsakoff et al., 2003). MacKenzie et al. (2011) noted that many scale development studies fail to adequately address the possibility of CMB, particularly when concurrent validity is established using a single-occasion design. Researchers are therefore advised to incorporate CMB controls into the study design phase of scale development rather than addressing them only post hoc.

11.2 Structural Equation Modelling and Nomological Validity

Structural Equation Modelling (SEM) plays a dual role in scale validation: as the analytical framework for CFA and as the vehicle for testing the scale's nomological validity by examining its hypothesised relationships with antecedents, consequences, and related constructs. Nomological validity — the degree to which the scale conforms to theoretical predictions about its relationships with other constructs — is a critical component of construct validity evidence (Messick, 1995; MacKenzie et al., 2011). SEM allows researchers to estimate measurement and structural relationships simultaneously while accounting for measurement error, thereby producing more accurate estimates of construct interrelationships than methods that ignore latent variable structure.

In the context of formative scales, Partial Least Squares SEM (PLS-SEM) has been advocated as an alternative to covariance-based SEM due to its ability to handle non-normal data and smaller samples (Hair et al., 2011). However, PLS-SEM has also attracted substantial methodological criticism: Rönkkö and Evermann (2013) argued that many claims made on behalf of PLS-SEM are empirically unsupported and that the method's ability to assess construct validity is fundamentally limited relative to covariance-based approaches. Researchers should carefully consider the assumptions and limitations of each SEM variant in selecting the most appropriate method for their validation context, and should justify their choice with reference to the theoretical and empirical properties of the constructs and data at hand.

11.3 Reporting Standards and Best Practices

Transparent and comprehensive reporting is essential for the scientific community's ability to evaluate the quality of published scales and to replicate their use. Worthington and Whittaker (2006) conducted a systematic content analysis of scale development studies published in the Journal of Counseling Psychology over a ten-year period (1995–2004) and identified widespread deficiencies: many studies reported insufficient detail about item generation procedures, failed to test for model fit in CFA, reported alpha as the sole reliability indicator, and provided incomplete validity evidence. These findings have been replicated in subsequent analyses across other disciplines, and Flake and Fried (2020) proposed the concept of "questionable measurement practices" — a suite of under-examined decision points in measurement research that, if not transparently reported, undermine the reproducibility and interpretability of empirical findings.

Best-practice guidelines — including those synthesised by MacKenzie et al. (2011) — recommend that scale development reports include: (a) an explicit articulation of the construct domain and its dimensionality; (b) documentation of item generation procedures and item pool characteristics; (c) evidence of content validity from expert review; (d) item-level descriptive statistics and item-total correlations; (e) EFA results with extraction and rotation method, fit indices, and factor loadings; (f) CFA results including fit indices, factor loadings, and error variances; (g) reliability estimates including omega coefficients in addition to alpha; (h) convergent and discriminant validity evidence; (i) criterion-related validity evidence; (j) measurement invariance tests where cross-group comparison is intended; and (k) information on sample size, sampling strategy, and characteristics of the development samples. The adoption of these reporting standards would substantially improve the transparency, reproducibility, and cumulative scientific value of scale development research.

12. Conclusions

Scale development constitutes a foundational methodological enterprise that conditions the quality of empirical knowledge across the social, behavioural, health, and management sciences. This review has examined the principal conceptual, procedural, and analytical dimensions of scale development, tracing the field's evolution from classical scaling methods through classical test theory to modern psychometric frameworks including confirmatory factor analysis, item response theory, bifactor modelling, and structural equation modelling. It is evident from this synthesis that a robust scale development process requires a theoretically grounded, multi-stage approach in which construct definition, item generation, pilot testing, factor analysis, reliability assessment, and multi-faceted validity evaluation are conducted in a systematic and interrelated fashion.

Several consistent themes emerge from this review. First, the persistent reliance on Cronbach's alpha as the primary reliability index, despite well-established alternatives such as McDonald's omega, represents an ongoing limitation in scale development practice. Second, the under-utilisation of confirmatory factor analysis, measurement invariance testing, and bifactor models reflects a gap between available psychometric methods and their adoption in applied research. Third, the reflective-versus-formative distinction remains inadequately appreciated, leading to measurement model misspecification in a non-trivial proportion of published studies. Fourth, common method bias continues to be insufficiently addressed, particularly in single-source self-report studies. Collectively, these observations point to the need for enhanced methodological education, more stringent editorial standards, and greater uptake of established best-practice guidelines in scale development research.

This review also highlights notable advances in the field. The growing adoption of bifactor models, the development of the HTMT ratio for discriminant validity assessment, improved reporting guidelines, and the formalisation of cross-cultural adaptation procedures collectively represent meaningful progress. The integration of IRT and Rasch modelling into social science scale development, while still limited, offers considerable promise for achieving sample-invariant item calibration and more precise measurement. Future directions in the field may include the application of network psychometrics as an alternative to latent variable models, the development of dynamic measurement instruments adapted to ecological momentary assessment designs, and the use of machine learning methods for item selection and construct discovery. These emerging approaches will require careful evaluation through the lens of established measurement theory if they are to yield defensible and interpretable psychometric instruments.

13. Limitations

This review is subject to several limitations that should be acknowledged. As a narrative review, it does not employ the systematic and exhaustive search procedures of a meta-analysis or systematic review, and the selection of literature — while guided by explicit criteria — inevitably reflects some degree of editorial judgement. Seminal works may have been omitted, and the balance of coverage across disciplines may not perfectly reflect the relative volume of scale development research in each field. Additionally, the focus of the review is primarily on reflective latent variable models, which dominate the scale development literature but do not represent all measurement approaches. Network psychometric models, Bayesian estimation approaches, and dynamic latent variable models receive only limited attention due to space constraints and the broader integrative scope of the review.

The exclusion of non-English language publications may also have introduced a degree of language bias in the coverage of cross-cultural adaptation and measurement equivalence research, as important contributions from non-Anglophone scholarly traditions may not be represented. The restriction to peer-reviewed journal articles, while appropriate for maintaining quality standards, means that important methodological developments documented in technical reports or book chapters — though excluded per the review protocol — are not systematically addressed. These limitations notwithstanding, the review provides a comprehensive and structured synthesis of the core methodological framework for scale development that is intended to be useful to both novice and experienced researchers across the disciplines in which quantitative measurement plays a central role.

Disclaimer (Artificial Intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

References

Abd ElHafeez, S., Salem, M., & Silverman, H. J. (2022). Reliability and validation of an attitude scale regarding responsible conduct in research. *PloS one*, *17*(3), e0265392. https://doi.org/10.1371/journal.pone.0265392

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. Psychological Bulletin, 103(3), 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. Spine, 25(24), 3186–3191. https://doi.org/10.1097/00007632-200012150-00014

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

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. Journal of Marketing Research, 16(1), 64–73. https://doi.org/10.1177/002224377901600110

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. Journal of Business Research, 61(12), 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

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

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. Measurement and Evaluation in Counseling and Development, 43(2), 121–149. https://doi.org/10.1177/0748175610373459

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Ferrari, R. (2015). Writing narrative style literature reviews. Medical Writing, 24(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

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

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. Social Psychological and Personality Science, 8(4), 370–378. https://doi.org/10.1177/1948550617693063

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R for all things omega. Advances in Methods and Practices in Psychological Science, 3(4), 484–501. https://doi.org/10.1177/2515245920951747

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. Journal of Marketing Research, 18(1), 39–50. https://doi.org/10.1177/002224378101800104

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. Journal of Chiropractic Medicine, 5(3), 101–117. https://doi.org/10.1016/S0899-3467(07)60142-6

Gültürk, E. A. (2024). Scale adaptation and redevelopment: A review on validity and reliability. Journal of Cellular and Molecular Immunology, 3(1), 26-32.

Guttman, L. (1944). A basis for scaling qualitative data. American Sociological Review, 9(2), 139–150. https://doi.org/10.2307/2086306

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. Journal of Marketing Theory and Practice, 19(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

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

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. Organizational Research Methods, 1(1), 104–121. https://doi.org/10.1177/109442819800100106

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55. https://doi.org/10.1080/10705519909540118

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical examination of construct indicators and measurement model misspecification in marketing and consumer research. Journal of Consumer Research, 30(2), 199–218. https://doi.org/10.1086/376806

Krosnick, J. A. (1999). Survey research. Annual Review of Psychology, 50(1), 537–567. https://doi.org/10.1146/annurev.psych.50.1.537

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. MIS Quarterly, 35(2), 293–334. https://doi.org/10.2307/23044045

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. British Journal of Mathematical and Statistical Psychology, 23(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

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

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. Psychological Methods, 9(1), 93–115. https://doi.org/10.1037/1082-989X.9.1.93

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, 88(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. Developmental Review, 41, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. Applied Psychological Measurement, 21(2), 173–184. https://doi.org/10.1177/01466216970212006

Reise, S. P. (2012). The rediscovery of bifactor measurement models. Multivariate Behavioral Research, 47(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment, 95(2), 129–140. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rönkkö, M., & Evermann, J. (2013). A critical examination of common beliefs about partial least squares path modeling. Organizational Research Methods, 16(3), 425–448. https://doi.org/10.1177/1094428112474693

Sarun, H. (2026). Measurement and scaling. *Journal of Agriculture and Environment, 3*(2), 326–330. https://doi.org/10.5281/zenodo.18823340

Schwarz, N. (1999). Self-reports: How the questions shape the answers. American Psychologist, 54(2), 93–105. https://doi.org/10.1037/0003-066X.54.2.93

Shukla, D. (2023). A narrative review on types of data and scales of measurement: An initial step in the statistical analysis of medical data. *Cancer Research, Statistics, and Treatment*, *6*(2), 279–283. https://doi.org/10.4103/crst.crst_1_23

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

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. Social and Personality Psychology Compass, 2(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Symeonaki, M., Stamou, G., Kazani, A., Tsouparopoulou, E., & Stamatopoulou, G. (2024). Examining the development of attitude scales using Large Language Models (LLMs). *arXiv preprint arXiv:2405.19011*.

Thurstone, L. L. (1928). Attitudes can be measured. American Journal of Sociology, 33(4), 529–554. https://doi.org/10.1086/214483

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

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. The Counseling Psychologist, 34(6), 806–838. https://doi.org/10.1177/0011000006288127

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99(3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432