International Journal of Community Medicine and Public Health
Phalswal U et al. Int J Community Med Public Health. 2026 Feb;13(2):1065-1073
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20260329

Review Article

A systematic approach to the development and validation process of a research tool: an overview

Uma Phalswal[1]*, Raman Kalia[2]

[1]College of Nursing, All India Institute of Medical Sciences, Rishikesh, Dehradun, India
[2]Saraswati Nursing Institute, Kurali-Morinda road, Dhianpura (Roopnagar), Punjab, India

Received: 17 September 2025
Revised: 31 December 2025
Accepted: 02 January 2026

***Correspondence:**
Uma Phalswal,
E-mail: Phalswaluma2828@gmail.com

ABSTRACT

Research tool development and validation process is enigmatic and opaque due to the diverse range of techniques it requires. As a result, the main goal of this article is to provide an overview of the research tool development process as simply as possible in order to support the development of new, valid, and reliable research tools, as well as improvement of existing ones. We accomplish this work by presenting all of the required steps in the right sequence to develop a research tool. In light of our search, we suggest five phases of tool development and validation, each with its own sub-steps. The first step involves preparing the preliminary draft, which includes conceptualizing the construct, evaluating current practices, conducting focus group discussions, and creating the item pool. The second phase includes the tool draft's validation with internal and external review, as well as the tool's pre-testing. Phase three includes field testing of the finalized tool draft, sampling planning, and data collection. The fourth phase includes the analysis of research tool data with item analysis, exploratory factor analysis, internal consistency, and test-retest reliability. The fifth phase is all about scale interpretation using percentiles, standard scores, norms, and cut-off points. Additionally, we briefly discussed the key best practices indicated in each step. This review will help both scientists and practitioners to understand all the steps and methodologies of research tool development and validation.

Keywords: Tool development, Validation, Item generation, Factor analysis, Validity, Reliability

INTRODUCTION

Research tool is an instrument in the hand of researchers to measure what they want to measure in their study.[1] The selection of research tools is made based on the study as well - usage goals, methods and requirements.[2]

Sometimes researchers are unable to find an instrument to measure their construct. There can be two needs at the back of tool development, either the research construct is new or there are some limitations of the existing tools.[3] Tool should be applicable for selected information type. First, the Researcher selects from available tools and tests a hypothesis.[4]

A perfect measuring instrument should be a valid, reliable, objective sensitive, as well as efficient.[5] Developing and validating a research tool is very complex process. Creating a multi-item measure of a construct involves several steps.[6] Test development and standardization, then again, are a two-dimension related procedure where test advancement starts things out after which the norming takes over.[7]

Objectives

The objective of this review to provide accurate information regarding all the steps of research tool development and validation process in a single piece of paper.

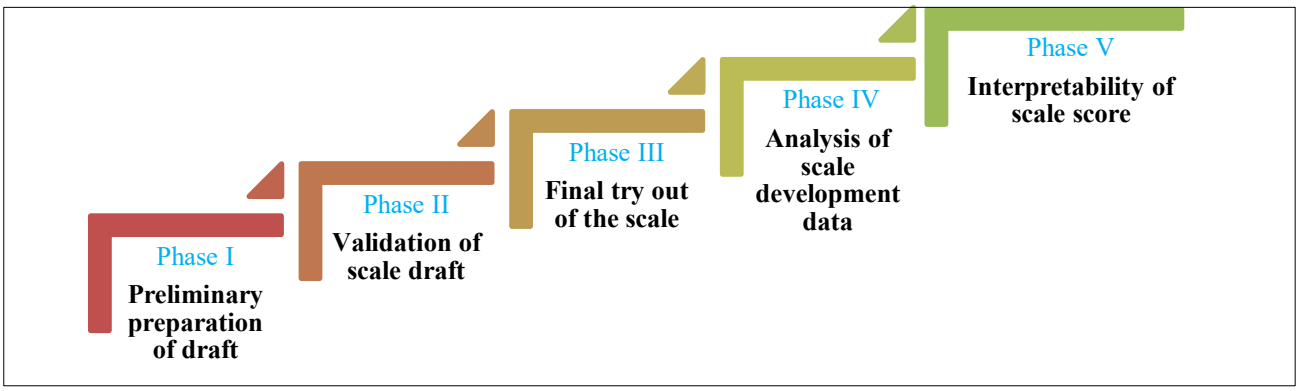

Figure 1: Phases of research tool development and validation.

Phases of tool development and validation

There are five phases of tool development and validation (Figure 1).

PHASE I: PRELIMINARY PREPARATION OF DRAFT (BEGINNING STEPS)

It is illustrated in Figure 2.

Figure 2: Phase I: preliminary preparation of draft.

Conceptualization of construct

The first step in development of research tool is to become an expert in construct and understand what it measures. Complex construct has number of different dimensions so it is very important to understand all. Researcher should differentiate the main target construct from the related construct Ex: self-esteem and self-confidence. Also develop Clear understanding about the population for which tool is developing.[8]

Assessment of current practices

Evaluate the present ways of measuring a construct. Find out the available tools to measure the construct and detailed analysis of the available tools to understand the constraints of current available tools.[9]

Focus group discussion

Focus group discussion is "informal discussions among selected people on a specific topic" is used to get item formation ideas. Focus groups are usually more free-form, less structured and can bring a wider array of answers out.[9]

Creating item pool

One of the very early steps in construction of a measurement tool is to generate an item pool that could potentially be operationalized on scale. An item has to be constructed in a way that it reflects accurately the variable being measured. How one can create item pool?[10]

Existing instruments

The authors can make adaptations (add or remove items, re-word etc.) to an existing instrument so that the it is more culturally sensitive or for a population with low reading skills. Since scales are copyright protected, written permission from the original author of a published scale is required.

The literature

Items are drawn from the ideas created by previous research.

Concept analysis

Another source of ideas is concept analysis.

In-depth qualitative research

Deep investigation around the key construct is a particularly lush ground of items for scales. Qualitative research helps in providing an insight into the dimensions of a concept and also gives real words for items.

Clinical observations

Patients in clinical settings offers great source of items. Suggested items have been developed by direct observation of patients' behaviours in relevant environments, as well as, sometimes, from their comments or interactions.

Preparation of preliminary draft

Item attribute decision-making process

Number of items to create

The objective is to measure the construct with a set of items that represent its essence in slightly varied ways, such that unimportant quirks of individual items cancel each other out. There is no secret method for determining how many items to develop. Most of the time, making a lot of items is better because longer scales are more reliable.[11]

Number of response items and their categories

Items consist a stem (or in some cases, statement) as well as response options. Most of the tools have between 5 and seven options. Response options are best in odd numbers because this leaves respondents with the ability to not take a stance another answer. Some researchers prefer even numbers because they believe that it balances slight tendencies or avoid equivocation.

Frequently used words for response options are; - strongly disagree, disagree, agree, strongly agree; never, almost never, sometimes, often, almost always and very important, important, somewhat important, of little importance, unimportant.

Positive and negative word

Weather to include positively and negatively worded items.

Item intensity

The strengths should be more or less equivalent and also relatively strong.

Time frame

Product development should not lead to any inertia of a time frame. Coming from a place of understanding the construct, decide beforehand to play with time.

Wording the items

Each item in a tool must be inserted with complete meaning and clarity so that each respondent is responding to identical questions.[12] The following points are specific to tool items.

Clarity

Each of the elements in a tool should be clearly understandable. Words must be chosen with an eye toward the educational and reading level of those you are trying to reach. Researcher should pick out words that everyone knows, and everybody agrees they know exactly what those word means.

Jargon

Language should remain free of jargon. Avoid the use of medical terminology that is not widely used by most people.

Length

Prohibit long sentence or phrases. Read it aloud and see if you can cut even more - especially any extra words.

Double negatives

Positive is almost always preferable to negative, but double negatives are forbidden.

Double barrelled items

Do not contain two or more ideas in one item.

PHASE II: VALIDATION AND REFINEMENT OF SCALE DRAFT

It is illustrated in Figure 3.

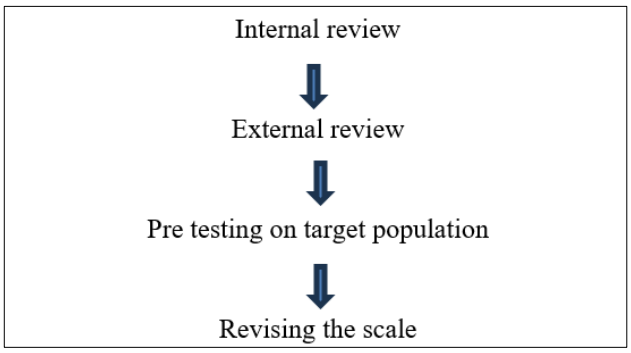

Figure 3: Phase II: validation and refinement of scale draft.

Internal review

A considerable item pool can then be internally reviewed. The specification of individual items should represent the construct and be phrased in well written, grammatically sound English.

Long sentences, and words with 4 or more syllables are also a no-no! Do not forget to evaluate the legibility of the scale unless it is for a very literate group.[13]

External review by experts

Once the researcher has made the first set of items, they should be looked over by a team of experts in the field to make ensure they are accurate. The content adequacy refers to the degree to which the items represent all product characteristics defined by the product, as explicitly stated in the product's definition. The experts should be provided with the construct definitions and instructed to sort the items according to these definitions to determine of their sort aligns with that of the scale developer. External review of the revised items by a panel of experts should be undertaken to assess the scale's content validity. Two rounds of review are advisable- one to refine the item and the second to assess the content validity of the tool.[14] Steps of external review: 3 steps included.

Expert selection and recruitment

The panel of experts comprises individuals who have strong credentials pertaining to the construct being measured. Further, in-depth knowledge regarding the target population will also be advantageous. The panel, consisting of 8 to 12 members, will have a mix of roles and disciplines. The expert panel will be sent a packet of materials comprising a strong cover letter, background information describing the construct and the target population, reviewer instructions, and a questionnaire seeking the opinion of the expert panel. The expert panel might be provided with a brief literature review and bibliography as well.

Preliminary expert evaluation: content validation of items

The expert's responsibility is to evaluate items as well as the whole scale and any subscales using the instructions given by the scale developer. The first expert panel can rate each item using certain criteria. For instance, each item can be rated as regards whether it is clear and unambiguous; relevant to the construct being measured, and appropriate for the target population. The questionnaire might also request for detailed comments regarding an item assessed to be not clear, relevant, or appropriate, such as how the item's wording could be made clearer, or why the given item is not appropriate. Moreover, besides rating items, the panel can be asked whether the items as a whole sufficiently cover the construct domain. Moreover, if the items also span a difficulty continuum or not? The expert agreement formula pertaining to each item is the number of experts agreeing, divided by the total number of experts. I-CVI recommended value is 0.78 or more.

Validation of the tool

The term validity refers to whether or not an instrument is measuring what it is supposed to. However, its significance is obvious but establishing it can be difficult. There are different types of validity. The initial one is the content validity, which is concerned with the expert opinion about whether or not the scale items represent the proposed domains or concepts that the tool is to measure. Convergent or discriminant validity are also taken into consideration. If we consider content validation, the second expert group rating could be whether the revised set of items is relevant or not, and the second parameter, which is also called I-CVI for the revised item set. I-CVI is computed as above. First-round data can be used to evaluate not just the items but also the performance of the experts. After the rating has been done, each revised item set could have another S-CVI computed. There is more than one way to compute an S-CVI. Calculation of S-CVI/UA: adding all items having I-CVI equal to 1 and dividing by the total number of items. Calculation of S-CVI/Ave: Taking the sum of the I-CVIs and dividing by the total number of items. Eg: 10-item scale in which the I-CVIs for the 5 items is 0.80 and the I-CVIs for the other 5 items are 1.00. S-CVI /Ave=0.90.

Recommendation

A S-CVI/UA $\geq$0.8 and a S-CVI/Ave $\geq$0.9 have excellent content validity

Hence, for a tool to have excellent content validity, it would be composed of items that had I-CVIs of 0.78 or higher and an S-CVI of 0.90 or higher. This in turn requires strong items, outstanding experts, and clear instructions to the experts regarding the underlying constructs and the rating task.[1,14,15]

Input from the target population

This is the first phase in which we pretest items from the pool. Generally, on any purposive pre-test of a new instrument, some number of people will be invited to fill out the items in question who should represent the target population. In analyzing the pre-test data, the researches look for items with non-response, items with limited variability, items with numerous mid-point responses. Such items are candidates for deletion or revision. Cognitive interviews are the type of pre-test.[16]

There are two types of cognitive interviews- think aloud method and verbal probing.

Think-aloud method

Here, the respondents are asked to state what they have in mind while working on a question by step and giving an answer. Ex: "please tell me what you are thinking as you answer this question" or "what steps are going through your head as you pick an option for this question"

Verbal probing

There are some targeted probes that the researcher uses, and when asked they have encouraged reflection on underlying cognitive processes. The verbal probes might be scripted or unscripted. Ex: "what do you think the question is asking you" and "please think aloud and tell me

how you would answer this question". Instead of pre-test, in this stage of development tool, focus groups can also be used.[17]

Revising the scale

Before the final field testing of the tool, we should get a rough feel for the face and content validity of the items in our scales. Revision of the scale should be based on external expert's content validity suggestions, and pretesting of the scale. Decisions about retaining or removing items is based not only on alpha coefficient but also on content validity considerations.[16]

PHASE 3: FIELD TESTING THE INSTRUMENT

Since the next step is to write the tool and to carry out a quantitative assessment of items, the next set of steps refers to the development of an instrument to assess the severity of psychopathology. The steps combine the development of a sampling and data collection plan and include the following elements.[18]

Developing a sampling plan

The sample to test the scale is representative of the population for which the scale is designed. The sample size depends on the complexity and precision of the analyses to be carried out. Since the entire population of scores from all possible persons who may master the scale cannot be obtained, the sample should be chosen as carefully as possible. The requirement for representativeness of a sample sometimes means that the sample for a scale should be chosen in different places in order to capture variation in responding to the items across geography.

Kline suggests that samples should range from 3-4 people per item to 40-50 people per item and is typically closer to 10 people per item. A common rule of thumb is to have at least 10 people per analyses to be estimated. Also, decide on whether to ask the sample to fill in the scale in a group or by themselves in order to have a general idea of the context, in which the scale will be normally administered.

Developing a data collection plan

It is time to decide how to administer the instrument and what to ask. In making a decision about the mode of administration, remember that the instrument should be administered in the way that is relevant to the final administration of the scale.

The instrument and its first version constitute the scale itself and should be composed of two parts: the scale items and the basic demographic information that allows interpreting the first part. If one is planning to assess test-retest reliability, ensure to ask for a method of calling the respondent to administer, and retrieve the first administration results.

Preparing for the data collection

Determine the instrument is likeable, looks professional and understandable before it gets into production. The instrument should have clear instructions for completion, and an assessment of the readability level of these instructions is helpful.[18,19]

PHASE IV: ANALYSIS OF SCALE DEVELOPMENT DATA

Basic item analysis

Item analysis is evaluation of each item in the preliminary scale. If each item is a measure of that construct, then the items should correlate with one another.

Inter item correlation

The degree of inter-item correlation can be assessed by inspection the correlation matrix of all the items with each other. Recommended value of inter-item correlation is between 0.30 and 0.70. Correlations lower than 0.30 suggesting little congruence with the underlying construct and ones higher than 0.70 suggesting over-redundancy.

Item total correlation

Calculate total scale score and then calculate correlations between items and total scores of the scale. If the item scores do not correlate well with the scale scores it is probably measuring something else and will lower the reliability of the scale. There are two types of item-scale correlations: one in which the total score includes the item under consideration and another in which the item is removed in calculating the total scale score.

The latter approach is preferable because the inclusion of the item on the scale inflates the correlation coefficients.

Recommendation

Eliminate the items whose item-scale correlation is less than 0.30.[20]

Exploratory factor analysis

A set of items do not form a scale; the items form a scale only if they have a common underlying construct. Factor analysis is data reduction technique and it identifies items that go together as a unified concept. Factor is a 'Weighted combination of items' that all are measuring the same dimensions. The items that are highly correlated or related to each other, measuring one construct should load on one factor and those measuring another construct should load on a different factor.

Analyses that yield no clear factors or one factor of a unidimensional scale are problematic. It explains amount of the variance in the scores. Based on these factor

loadings, the researcher needs to decide which items from the scale should be retained or deleted. Ferguson and Cox suggest that 100 respondents are the absolute minimum number to be able to undertake this analysis. However, others would suggest that this is insufficient and a rule of thumb would be five respondents per item. Exploratory factor analysis has two steps: Factor Extraction and factor rotation.[21]

Factor extraction

The goal of factor extraction is to extract clusters of highly interrelated items from a correlation matrix. A most widely used factor extraction method is principal components analysis and another is principal-axis factor analysis. Factor extraction yields an unrotated factor matrix, for all original items on each extracted factor. Factoring continues until no further meaningful variance is left. Two main methods are used to decide the number of emerging factors.

Eigen value

Eigen value is the amount of variance (information) explained by a factor. Eigen value should be more than 1 because less than 1 eigen value means factor is not explaining information that is equivalent to 1 question hence data reduction will not take place if factor is extracted less than 1 eigen value (Figure 4).[22]

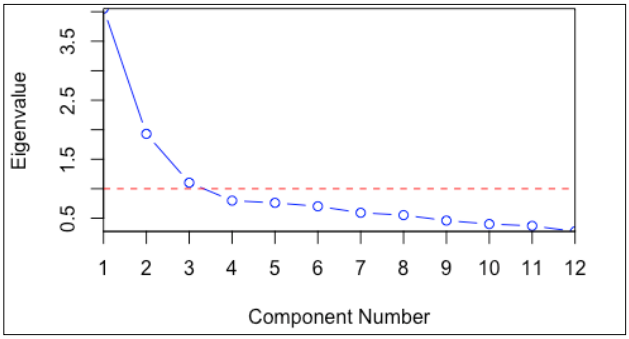

Figure 4: Eigen value versus component number.

Screen test

The screen test is another cutoff point. It is a test depending on the principle of discontinuity. Number of factors are identified from the break in the slope.

Factor rotation

Rotations minimize the complexity of the factor loading, to make the structure simpler to interpret. There are two types of factor rotation techniques.[21,22]

Orthogonal rotation

Orthogonal rotation does not allow the factors to be correlated pop by routinely restricting the angle between the axes to 90 degrees. The types of orthogonal rotation are Varimax, Equimax, Quartimax.

Oblique rotation

Oblique rotation allows the factors to be correlated by allowing the angle between the axes to be less than 90 degrees. The Promax and Direct Oblimin methods use Oblique rotation for factor analysis.

Factor loading

The entries under each factor are factor loading. The correlation between a factor and independent variable is called the factor loading. For orthogonally rotated factors, factor loading can range from -1.00 to +1.00 and can be interpreted like correlation coefficient, express the correlation between items and factors. Researchers work with rotated factor matrix in interpreting the factor analysis.[23]

Communality

Communality informs is the amount of variability in independent variable. Communality values range from 0 to 1.

Higher the value of the communality value, more useful is the variable in explaining the group characteristics. Low communality indicates that the identified factors in the model do not explain much variability in the variable and such variables be removed as a result of low values.[24]

Use of factor analysis

Factor analysis helps to identify the dimensionality of the construct. To make decision about item deletion and retention. If items have low loadings on all factors, they likely are good candidates for deletion. Items with fairly high loadings on multiple factors may also be candidates for deletion. Items with marginal loadings but that had good content validity could be retained for the internal consistency analysis.[22]

Reliability

Reliability refers to the accuracy and consistency of a measuring tool.

Coefficient of stability

Refers to consistent performance on a test over a period of time.

Coefficient of equivalence

Refers to a relationship between the scores of the participants on two forms of the same test.

Coefficient of internal consistency

Refers to the consistency of the performance of individual on different parts or items of the test taken at a single sitting.[25]

Internal consistency analysis

Internal consistency refers to the homogeneity of items in a scale. This coefficient should be as high as possible. If not, the items contributing to low reliability, these items should be dropped and other new items are generated. It is calculated by deleting a specific item from the scale and run the computer item-wise reliability with Cronbach alpha with the total. This statistic uses inter-item correlations to determine if constitutive items measure the same domain. If the items show good internal consistency, the alpha value should exceed 0.70 for a holistic tool or 0.80 for a refined tool.[26]

Test-retest analysis

Test–retest reliability can assess the stability of a measure over time and this should be included in the process of any tool development. An issue in retest is timing of retest relative to initial administration. When timing is too brief, carryover effect can lead to high reliability. Some experts advised time interval between measurement is between 1 to 2 weeks. Test-retest reliability can be calculated by applying Pearson's correlation between the test and retest score of each participant.[25,26]

PHASE V: INTERPRETABILITY OF SCALE SCORE

Interpretability refers to understanding what a scale or raw score on a scale means? Seldom is a raw score on a scale directly interpretable. If you expect that the scale will be used by others, it is prudent to create a manual for its use. In addition, scale developers may consider registering a copyright, even if they never plan to publish the scale commercially. Ways to facilitate greater interpretability of scale scores include the following.

Percentiles

Raw score values from a scale can be made more interpretable by converting them into percentiles. A percentile is a function that returns the percentage of people who scored below a particular value. Despite the fact that the obtained characteristic is very understandable and easy to interpret for most people, it should not be used for the purposes of serious analytical work and conducting tests.

The value of a percentile can vary between 0 and 99, where the 50th percentile is a median. It should be noted that a percentile is most useful when its characteristic is based on a sufficiently large and representative sample.

Standard scores

The standard score is the number of standard deviations by which the value of a raw score (i.e., an observed value or data point) is above or below the mean value of what is being observed or measured. Raw scores above the mean have positive standard scores, while those below the mean have negative standard scores. The most common formula is. x is the raw score, m is the mean, s is the standard deviation, and z is the standard score.

Norms

Norms is the average or typical score on a particular test made by specified population. It provides data for comparison and interpretation of test scores. There are different types of norms including percentiles, age/grade norms, and national/local norms. The normative sample used to develop norms should be large, representative of the target population, and clearly defined.

Cutoff points

Cutoff points are typically used as the base for making decisions about needed treatments or further assessments. For example, children whose weight are below the 5th percentile are usually thought to be underweight, while children whose weights are above the 95th percentile are usually thought to be overweight. In other cases, the cutoff points are designated with standard scores. Cutoff points that are linked to the measure's distribution are considered norm-referenced.[28]

CRITIQUING SCALE DEVELOPMENT STUDIES

The criteria for evaluating scale development and assessment reports are as follows: according to the report, was the definition of the construct clear? Has the report rightly situated the study by reviewing the literature on the question and discussing theoretical issues? Has the report provided a clear description of the kinds of people for whom the scale is intended? Was there any information in the report on how the items were generated? Were the procedures of the scale development or validation reports used in the study at least plausible or sensible? Was any information provided on the reading level of the scale items? Was there any information from the report on the description of efforts of content validation and was such a description adequate? Was there any indication from the report about good content validity? Were proper efforts made to revise the scales (for example, pre-tests and item analysis)? Was the development and/or validation sample of participants used in the study in terms of representativeness, size, and homogeneity appropriate%. was factor analysis used in examining or validating the number of scales. If "yes," was there any evidence in the report on such use and was the report supported the factor structure and naming of factors? For the purposes of evaluating the scale on internal consistency and reliability, were a proper method used in the study, and did the

methods and internal consistency and reliability estimates were very high? Was there any evidence in the report that the method used in the assessment of the criterion or construct validity was used in the study? Was evidence available in the report on the validity of the scale? What other kinds of procedures would most increase the ability of these results to justify use of the scale? Did efforts made in the report to evaluate test-retest reliability of the change score and the responsiveness of the new measure? Does the report give information on how to score the scale and how to interpret scale scores% For example, does the report provide the means and standard deviations or cut-off scores or norms%.[29,30]

DISCUSSION

According to experts, measurement is the most important factor in scientific research.[31] The purpose of scaling is to make a scale that suits measurements and has certain characteristics for measuring a construct. Scaling types and response formats, such as Likert type, forced-choice, and the multiple-choice response formats, are used universally in all psychology.[32] In a way, scaling type and response format effects item writing and scale development. The item pool should be as rich as possible for the developing scale. The scale should have many items on the construct to be measured. Steps in the instrument development process include: defining the purpose of the instrument, the field of the thing, and the construct to be measured, deciding on the response scale format, generating items to create item pools 2 to 4 times the desired final extent, selecting items with a review of the expert group description and/or pre-posting to maximize the instrument's reliability in conjunction with an item analysis, and conducting large-scale evidence study to establish construct validity, supplementary item analysis, and factor analysis to standardize scale scores. After the pool was made presentable by experts and/or pretesting and the reliability was controlled by item analysis, a construct validation study to measure the dimensionality of the scale and standardization should be done.[33] The measurement reliability value of how much the score is valid, reliable, and repeatable. The construct validation value of the measurement is based mainly on the correlation, consistency of the measurement and the items constituting the thing to be measured, and this is mainly revealed by the factor analysis of the scales.[34,35] If a scale is developed thoughtfully and precisely, have a good chance of growing into an examiner that measures real world phenomena quite accurately.

CONCLUSION

This review consolidates the complex process of research tool development and validation into a clear, systematic framework. By outlining five sequential phases—from construct conceptualization to score interpretation—it highlights the importance of methodological rigor at every stage. Emphasis on expert review, empirical testing, and robust psychometric analysis ensures that tools are both valid and reliable. The structured approach presented can guide researchers in developing new instruments and refining existing ones. Overall, the article serves as a practical reference for producing scientifically sound measurement tools that accurately capture real-world constructs.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Andersen RD, Jylli L, Ambuel B. Cultural adaptation of patient and observational outcome measures: a methodological example using the COMFORT behavioral rating scale. Int J Nurs Stud. 2014;51(6):934-42.
2. Anthony M, Yastik J, MacDonald DA, Marshall KA. Development and validation of a tool to measure incivility in clinical nursing education. J Prof Nurs. 2014;30(1):48-55.
3. Barker C, Pistrang N, Elliott R. Research Methods in Clinical Psychology: An Introduction for Students and Practitioners. 2016. Available at: https://www.al-edu.com/wp-content/uploads/2014/05/Barker-et-al-Research-Methods-in-Clinical-Psychology.pdf. Accessed on 20 October 2025.
4. Beck CT, Gable RK. Postpartum Depression Screening Scale: Development and Psychometric Testing. Nurs Res. 2000;49(5):272-82.
5. Beck CT, Gable RK. Further validation of the Postpartum Depression Screening Scale. Nurs Res. 2001;50(3):155-64.
6. Choe K. Development and preliminary testing of the Schizophrenia Hope Scale, a brief scale to measure hope in people with schizophrenia. Int J Nurs Stud. 2014;51(6):927-33.
7. DeVellis RF. Scale Development: Theory and Applications. Thousand Oaks, CA: Sage. 2017;4.
8. Ferguson E, Cox T. Exploratory factor analysis: a user's guide. Int J Selection Asses. 1993;1:84-94.
9. Ferketich S. Focus on psychometrics. Aspects of item analysis. Res Nurs Health. 1991;14(2):165-8.
10. Furr RM. Scale Construction and Psychometrics for Social and Personality Psychology. New Delhi, IN: Sage Publications. 2011.
11. Gaugler JE, Hobday JV, Savik K. The CARES(®) Observational Tool: a valid and reliable instrument to assess person-centered dementia care. Geriatr Nurs. 2013;34(3):194-8.
12. Grassley JS, Spencer BS, Bryson D. The development and psychometric testing of the Supportive Needs of Adolescents Breastfeeding Scale. J Adv Nurs. 2013;69(3):708-16.
13. Hinkin TR. A Review of Scale Development Practices in the Study of Organizations. J Management. 1995;21(5):967-88.

14. Hilton A, Skrutkowski M. Translating instruments into other languages: development and testing processes. Cancer Nurs. 2002;25(1):1-7.
15. Karaçam Z, Kitiş Y. The Postpartum Depression Screening Scale: its reliability and validity for the Turkish population. Turk Psikiyatri Derg. 2008;19(2):187-96.
16. Konrath S, Meier BP, Bushman BJ. Development and Validation of the Single Item Trait Empathy Scale (SITES). J Res Personality. 2017;73:111-22.
17. Kumar A. Review of the Steps for Development of Quantitative Research Tools. J Adv Pract Nurs. 2015;1:103.
18. Kyriazos T, Stalikas A. Applied Psychometrics: The Steps of Scale Development and Standardization Process. Psychology. 2018;9:2531-60.
19. Kyriazos TA. Applied Psychometrics: The 3-Faced Construct Validation Method, a Routine for Evaluating a Factor Structure. Psychology. 2018;9:2044-72.
20. Kyriazos TA. Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General. Psychology. 2018;9:2207-30.
21. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.
22. Morrison KM, Embretson S. Item Generation. Wiley. 2018.
23. Navabi N, Ghaffari F, Shamsalinia A, Faghani S. Development and validation of evaluation tools of nursing students' clinical pharmacology unit. Drug Healthc Patient Saf. 2016;8:101-9.
24. Netemeyer RG, Bearden WO, Sharma S. Scaling Procedures: Issues and Applications. Thousand Oaks, CA. Sage Publications. 2003.
25. Oppenheim AN. Questionnaire Design, Interviewing and Attitude Measurement. Pinter, London. 1992. Available at: https://www.scirp.org/reference/referencespapers?referenceid=435162. Accessed on 20 October 2025.
26. Peng Q, Wu W. Development and validation of oral chemotherapy self-management scale. BMC Cancer. 2020;20(1):890.
27. Pinto MD, Hickman R, Logsdon MC, Burant C. Psychometric evaluation of the revised attribution questionnaire (r-AQ) to measure mental illness stigma in adolescents. J Nurs Meas. 2012;20(1):47-58.
28. Polit DF. Getting serious about test-retest reliability: a critique of retest research and some recommendations. Qual Life Res. 2014;23(6):1713-20.
29. Polit DF, Beck C, Owen S. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health. 2007;30(4):459-67.
30. Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Wolters Kluwer. 2017;10:331-54.
31. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-57.
32. Rattray J, Jones MC. Essential elements of questionnaire design and development. J Clin Nurs. 2007;16:234-43.
33. Rowe H, Sperlich M, Cameron H, Seng J. A Quasi-experimental outcomes analysis of a psychoeducation intervention for pregnant women with abuse-related posttraumatic stress. J Obstet Gynecol Neonatal Nurs. 2014;43(3):282-93.
34. Ruiz RJ, Gennaro S, O'Connor C, Marti CN, Lulloff A, Keshinover T et al. Measuring coping in pregnant minority women. West J Nurs Res. 2015;37(2):257-75.
35. Schriesheim, CA, Eisenbach RJ. An exploratory and confirmatory factor analytic investigation of item wording effects on the obtained factor structures of survey questionnaire methods. J Manag. 1995;21:1177-93.

Cite this article as: Phalswal U, Kalia R. A systematic approach to the development and validation process of a research tool: an overview. Int J Community Med Public Health 2026;13:1065-73.