СЕКЦІЯ 1
МЕТОДОЛОГІЯ ТА МЕТОДИ СОЦІОЛОГІЧНИХ ДОСЛІДЖЕНЬ

CORE PRINCIPLES AND BEST PRACTICES FOR MULTI-ITEM SCALE CONSTRUCTION IN THE SOCIAL SCIENCES

ЗАСАДНИЧІ ПРИНЦИПИ ТА РЕКОМЕНДОВАНІ ПРАКТИКИ РОЗРОБКИ БАГАТОПОЗИЦІЙНИХ ШКАЛ В СОЦІАЛЬНИХ НАУКАХ

UDC 303.023.22:316.3
DOI https://doi.org/10.32782/2663-5208.2025.73.1

Maltseva K.S.
DSc (Sociology), Associate Professor, Head of the Department of Sociology
National University of Kyiv-Mohyla Academy

Measurement is a fundamental component of scientific research. Data quality and validity of researcher's conclusions depend upon the quality of metrics and the efficacy of measurement. Although originally the researchers' main interest in measurement lay within the ability domain, it has gradually spread into different areas of social life. There are many ways of constructing measures in the social sciences, one of the most popular tools in psychology and sociology being verbally-mediated self-report method that is often used to construct items that are grouped into multi-item scales to measure some social or psychological construct of interest underlying them. Scales offer a more robust way of measuring dimensionality and exploring inter-item and inter-informant variation. One of the reasons for the appeal of multi-item scales is that their quality can be explicitly and directly assessed. Well-grounded scale construction is key to ensuring informative results and valid conclusions. In its turn, it necessitates understanding the psychometric premises underlying scale construction. This publication aims to provide an overview of core principles of multi-item scale construction and to emphasize some of the best practices recognized in the process of scale development. The theoretical premises of the scale construction process are articulated. By reviewing such facets of scale making as conceptualization of the target concepts and relationships among them, generalizability, meaningfulness, dimensionality, reliability, validity, factor loadings, dimension interpretation, subscales, scale labeling and scale length, this review seeks to outline the structure of the scale making process, articulate its theoretical assumptions, and emphasize the best practices that secure the development of reliable metrics.

Key words: *measurement; multi-item scales; scale construction; reliability; validity; quantitative methods.*

Вимірювання являє собою фундаментальну складову наукового дослідження. Якість даних та валідність висновків дослідника залежать від якості метрик та ефективності вимірювання. Хоча від початку дослідницький інтерес у галузі вимірювання стосувався оцінки людських здібностей, він поступово розширився на інші форми соціального життя. У соціальних науках є різні способи конструювання метрик, і однією з найбільш вживаних в психології та соціології є метод словесно сформульованих самозвітів, що часто використовується для конструювання індикаторів, які компонуються у багатопозиційні шкали для вимірювання певного соціального чи психологічного конструкту, що стоїть за ними. Такі шкали відкривають більш надійний спосіб підійти до оцінки розмірності та дослідити варіативність, що існує між змінними та респондентами. Одна з причин популярності вжитку шкал полягає у можливості безпосередньо і експліцитно оцінювати їхню якість. Побудоване на доказах конструювання шкал є ключовим елементом для забезпечення інформативних результатів та валідних висновків. Це, у свою чергу, передбачає врахування психометричних посилок, що стоять за процесом шкалоутворення. Ця публікація має на меті надати огляд засадничих принципів побудови багатопозиційних шкал і для цього наголошує на деяких рекомендованих практиках, що застосовуються у процесі шкалоутворення. Зазначаються теоретичні посилки процесу формування шкал. Розглядаючи різні аспекти процесу побудови шкал, як-от узагальненість, осмисленість, розмірність, надійність, валідність, факторні навантаження, інтерпретація вимірів, суб-шкали, назви для шкал та розмір шкали, цей огляд намагається окреслити структуру процесу побудови шкал, сформулювати його теоретичні засновки та наголосити на рекомендованих практиках, що забезпечують розробку надійних метрик.

Ключові слова: *вимірювання, багатопозиційні шкали, розробка шкал, надійність, валідність, кількісні методи.*

Relevance and research problem. All empirical sciences face a task of designing measurement procedures to acquire accurate information about material objects, cognitive states, individuals, or groups. Data quality and validity of researcher's conclusions depend upon the quality of metrics and the efficacy of measurement. Measurement is an essential component of scientific research, regardless of the disciplinary boundaries that divide scholarly efforts into different sciences. Measurement is furthermore fundamental to any scholarly activity and is significant in researching any social context [2; 7; 26]. As scientific knowledge is generated through systematic observation, the information obtained this way is often subjected to some form of quantification in order to be processed or made sense of. Rigorous measurement is the main avenue of communication between the researcher

and the real life phenomena of their research interest [26].

These issues highlight two major concerns with measurement in empirical research. The first one is *generalizability* – namely, does the metric work outside of the local context? If it cannot be equally efficiently applied with regard to other samples, points of time or number of observations, the metric is much less useful. The second concern is to do with *meaning*, as one's measurement has to be meaningful and tractable in order to allow making interferences and answering research questions. Finally, we have to assume that all measures are bound to have some amount of *error*, therefore employing multiple measurements are advisable in order to avoid the distortions in one's data and conclusions [13]. Scale usage is often favored in social sciences but there are issues connected to understanding the theoretical premises informing scale construction that ultimately lead to concerns in evaluation of scale quality and interpreting the scores. This publication seeks to address this problem by articulating both the requisites of scale making procedures and best practices in the process of scale development.

Review of current research and publications. A rise of interest in measurement coincided with the beginning of the systematic exploration of individual differences initiated by Galton in the late 19th century. Although originally the researchers' main interest lay within the ability domain, it has gradually spread into different areas of social life including values, temperament and vocational interests, and is steadily growing [3; 21]. The techniques of scale construction grew out of many years of factor-analyzing groups of self-report inventory items and are widely known among psychometricians [4; 19]. In anthropology, the technique of scales was effectively used by Roy D'Andrade in his analysis of American, Vietnamese, and Japanese values [6]. Scales are presently widely used in sociology and social epidemiology to address complex social phenomena and lived experience [24].

It is often the case for the quantitative surveying of social contexts that the connections between the research concepts and metrics are the relationships of interdependence; in the social sciences it is not uncommon that the arrival of newly developed metrics broadens epistemological horizons, leads to new insights, and ameliorates our understanding of the phenomena or processes in question [26]. As described in textbooks on methods, it often the case that when a 'tactical' approach to measurement changes, it alters the perspective on the studied constructs and their interrelationships, thus increasing the informativeness of the research results, especially in the case of interdisciplinary queries which tend to treat more complex relationships [18; 19; 20].

Embedding theory into measurement

There are theoretical and a-theoretical metrics. A-theoretical metrics service constructs that are not a product of any specific theoretical framework and can be easily accessed by a simple stimulus (i.e. What is your weight in kilos?). Theoretical metrics are developed for composite theoretical notions that are used to explain phenomena or behavior, and bear their marks in that they must correspond to them in their structure [16](for example, emotional vs. cognitive empathy) or level of complex abstraction (i.e. SES, neuroticism), so they require multiple stimuli to address their various facets.

Undoubtedly, not all metrics need to be theoretical (compound or complex) to be useful tools. There are some quite simple indicators (such as gender, age, your favorite ice cream flavor or your candidate in the forthcoming election) that can be collected empirically and require no additional explanation to be efficaciously made sense of. In their turn, theoretical constructs – those built on the foundation of a specific model – may not have a tangible status by themselves but there is an assumed theorized entitativity behind them. Alcoholism, abuse, optimism, individualism, depression, self-direction, extraversion, hedonism, anxiety are some of the examples. Therefore oftentimes such constructs are *latent variables* and require data reduction techniques (such as factor analysis or principal component analysis) for their extraction, and for their measurement *multi-item scales* are used.

Background of method of multi-item scale construction

Using methods of data reduction is a useful quantitative option for social scientists to parse large agglomerations of social or psychological information. Principal components analysis and factor analysis are popular tools among psychometricians, and they are sometimes used by anthropologists and sociologists to describe the culture-specific organization of beliefs in cognitive data. Data reduction is achieved by reconstructing the relationships between variables in a matrix and presenting them as a set of new latent variables summarizing variation present in the matrix; it results in condensation of a dimension (a principal component or a factor). Thus, data reduction permits accommodation of material of considerable complexity, without assuming unidimensional consensus. It also directly and explicitly tests the cohesiveness of the dimension rather than assuming that its sharing is homogenous among the informants [4; 6; 19]. Altogether it makes data reduction techniques a suitable tool to both produce a ''big picture'' and capture the internal organization of complex abstract entities such as cultural models, institutional lifeworlds, worldviews, ideologies etc.

One of the reasons for appeal of multi-item scales (compared to single-variable measures[1]) that is often quoted is that their quality can be explicitly and directly assessed. The conventional way to do it is by computing Cronbach's alpha (α) [21]. However, due to the lack of clarity of alpha's criteria that

[1] Having said that, it is important to acknowledge, however, that sometimes complex phenomena and/or states can be effectively represented by one well-chosen, effectively phrased item (such as seen with the questions about happiness). Single item measures have been shown to be effective, especially for volatile concepts such as subjective well-being.

is sometimes deemed not enough [22][2]. Furthermore, there is no clear agreement as to how the criteria of Cronbach's alpha should be interpreted [22], while the alpha between 0.6 and 0.8 is usually considered acceptable [10; 12; 17].

Scales offer a more robust way of measuring dimensionality and exploring inter-item and inter-informant variation [4; 15]. There is always a theoretically-informed dimension behind a scale which can be described accordingly and labeled in a meaningful, intersubjectively intuitive way. Such constructs often *influence* behavior rather than embody it. Latent variables are variables that collectively measure one construct that is behind a specific set of items – an agglomeration of indicators, which altogether measure a construct, that could not be as effectively captured by one item.

In an earlier article [14] I discuss how data reduction and cluster analysis techniques can be used to construct scales step-by-step and how to make use of the methodological possibilities of correspondence analysis of multi-item scales derived from cluster analysis and principal components analysis, to extract and explore large agglomerations of social information. Although there are different traditions in understanding the mechanics of correspondence analysis, the advantages of using scales are not disputed. Correspondence analysis reveals the structure of the data and provides a scaled model of that structure, summarizing complex relationships among many subjects and many sets of variables simultaneously [25]. Correspondence analysis of multi-item scales permits treating multiple values and norms dimensions (i.e., without assuming unidimensional structure of the data) and demographic categories simultaneously.

In the social sciences, well-grounded scale construction is key to ensuring informative results and valid conclusions. In its turn, it necessitates understanding the psychometric premises underlying scale construction. **This publication aims to** provide an overview of core principles of multi-item scale construction and to do so it emphasizes some of the best practices recognized in the process of scale development.

Assumptions, phases and facets of scale construction. There are multiple ways of constructing measures in the social sciences, including the use of biomarkers, collateral reports, behavioral observations (including online communities) etc. Despite the abundance of options, one of the most popular and most often employed tools in psychology and sociology is verbally-mediated self-report method that is often used to construct items that are grouped into scales to measure some social or psychological construct of interest [3].

Strategies of scale construction

Within the scale construction methods there traditionally have been three mainstream strategies: (1) rational-theoretical approach, (2) empirical criterion keying, and (3) internal consistency, or factor-analytic methods [23, p. 415]. The first strategy appears the simplest as it is grounded in the researcher's straightforward assessment of item's value based on its face validity. This way a scale is composed based on each item's perceived relevance to the scale, irrespective of its psychometric underpinnings, which is a major limitation of this approach.

In its turn, the empirical criterion keying method selects items for the scale primarily based on their effectiveness in distinguishing between the individuals on some continuum reflecting a trait of interest represented by the scale; the sole criterion for the item's inclusion in the test is its ability to discriminate between two groups where an attribute of interest (for example, depression) is either present or absent [23, p. 415–416]. The classic example of this mechanism is the Minnesota Multiphasic Personality Inventory (MMPI) [11; 1]. Measures obtained this way often show good convergent validity but lack discriminant validity[3] or internal coherence. This is a drawback preventing the empirical criterion keying from being recommended as the most advisable approach in scale construction.

Finally, the internal consistency (or factor-analytic) method works by distilling relatively homogenous groups of intercorrelated items (i.e. the ingredients for the future scales) demonstrating high discriminant validity. It is usually achieved by means of data reduction techniques (factor analysis or principal components analysis) to extract clusters of variables that form interpretable dimensions [23, p. 416]. However, the meaning of extracted dimensions is not derived in the process of their condensation and therefore cannot be assumed and can only be inferred, which calls for caution in labeling extracted scales. Therefore the most straightforward way to create scale names is by choosing a label invoking three scale items with the highest item-total correlations or, alternatively, three scale items with the highest factor loadings on the first factor (thus contributing to the contents of the scale the most).

As all the described approaches are not only mutually exclusive but also have their respective shortcomings making them all fall short of the universal solution for the problem, some integrative approach should better be assumed to balance them out. One such overarching, umbrella-like approach involves the concept of construct validity. Although often it is the case that scale developers consider construct validity only after the scale has been designed and not in the beginning of the process of making a scale, it is nonetheless important that in the absence of the readily available golden standard for measurement of non-tangible concepts their validation should be embedded into their respective theoretical networks

[2] Performing split-half test is usually recommended as an additional measure in such cases.

[3] Convergent validity implies high correlations with measures that are theoretically cognate (e.g. flourishing, meaningful life, and happiness); discriminant validity implies low correlations with theoretically unrelated measures or inverse relationships between entities that are theorized as conceptual opposites of each other (e.g. happiness and depression).

[23, p. 416–417]. In this manner, construct validity also furthers theoretical development [21].

As current literature privileges theory-based model of scale construction, the first premise of the scale development process is a theoretical one and its first step begins with a clear conceptualization of a target construct (fig. 1). The process of scale development therefore opens with (1) an articulation of the theoretical concepts of interest and their purported interrelationships, followed by (2) designing and developing ways to measure those constructs according to chosen theoretical framework, and finally (3) empirically testing the relations among those theoretical constructs through their observable manifestations [3]. The process of composing the initial pool of items is connected to a thorough literature review that allows not only to define the construct of interest and its boundaries clearly for a sufficient theoretical coverage (which constructs are relevant and which are tangential), but also to encompass the related constructs to ensure an adequately detailed nomological network through making all relevant content available for sampling. Thus conceptualization is a key first step in the process of making good scales.

Thoughtful formulation of items is another aspect of the process of scale development. Items need to be carefully phrased to avoid ambiguity, jargon, contamination with other phenomena or dispositions, etc. [23, p. 420; 3]. The choice of response format is also of vital import. The reason for such caution is the intrusiveness of some traits/features (for example, such as neuroticism) that intervene with phrasing and understanding an item quite easily (for example, by adding a proposition like '*I worry that…'* or similar phasing that introduces the neurotic flavor to the scale).

Dimensionality

A scale's dimensionality, or factorial structure, reflects the number and nature of variables assessed by its items [9, p.7]. Scale's dimensionality is usually assessed by some technique of data reduction (factor analysis or principal components analysis) and allows ascertaining exactly how much each scale item contributes to the underlying latent variable, judging by the magnitude of the item's factor loading. Exploratory factor analysis (EFA) is most often used to uncover the underlying scale structure (e.g. whether it is uni- or multi-dimensional) and to determine how items are grouped together. It is often helpful to contrast the results of EFA (the factorial structure) with the output of cluster analysis (the dendrogramme) to check for meaningful overlaps [14]. Confirmatory factor analysis (CFA) can be used to check for scale invariance, confirming if the scale has the hypothesized structure.

Validity and reliability

Validity (i.e. the condition when a metric measures exactly what it was meant to) and reliability (i.e. the condition when a measure measures what it was supposed to measure accurately) both relate to the link between the developed metric and the real life phenomena it attempts to measure. Both validity and reliability are encumbered with a series of issues that need to be acknowledged with respect to scale construction principles [18]. First of all, validity and reliability are interrelated. Second, high reliability is easier to ensure than high validity. At the same time, reliability does not guarantee validity nor is a sufficient condition for it. Moreover, the two parameters are connected to each other but they can clash – the increase in reliability of a metric (for example, by virtue of selecting more concrete and specific items to represent various facets of the phenomenon) can

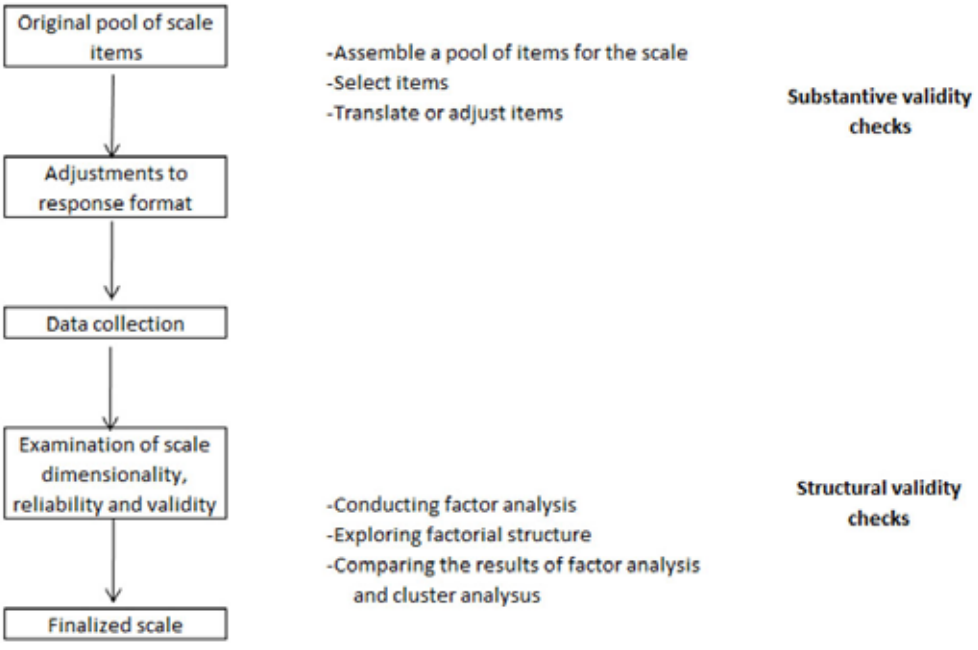

Fig. 1. Schematic representation of the phases of scale construction process

lead to decreased validity (which is often referred to as the problem of the forest and the trees).

Another problematic issue relates to the confusion between the *internal consistency* (i.e. reliability measured by Cronbach's alpha) of the scale, on the one hand, and *homogeneity*, on the other. Although these characteristics are sometimes assumed to be synonymous in the literature, they are not [3; 5; 13; 22]. Internal consistency conveys information about the degree of interrelatedness of a set of items; homogeneity refers to the unidimensionality of the set of items (i.e., whether or not they all tap into the same underlying construct) [22]. Alpha is not a measure of unidimensionality and multidimensional measures require other ways of establishing reliability, including creating subscales [3; 22]. Alpha supplies the details about the operationalization of the concept but it only yields accurate estimates under the condition that the scale is unidimensional. Alpha depends on homogeneity and the length of scale[4] [13].

It should be noted here that achieving a high alpha is not always possible. Sometimes a phenomenon being studied does not imply an inter-correlation of indicators or such inter-correlation cannot be expected. In cases such as these it is more important to capture the cumulative experience than intrecorrelations of scale items. For example, it is often the case with check-lists for traumatic events, war-induced stress or sexism. Sometimes there is no underlying theoretical dimension or expectation that these events would be intercorrelated (which is the premise for computing an alpha). The Chronbach's alpha, as we know, can only be legitimate for unidimensional entities [17]. When no such entity is theorized, the requisites for a high alpha as a condition for computing an index are not fulfilled. It can be however useful to know the composite score for the experiences which are constitutive parts of the scale contents, to be able to distinguish between the respondents along the continuum. Along the same lines, in a rubric "Can a reliability coefficient be too high?" Netemeyer, Bearden, & Sharma [17] point out the example of a sexism scale in which participants were supposed to report the extent to which they have experienced a number of different sexist situations:

We would not necessarily expect the experience of one event to be related to experiencing another event. In a case such as this, the reliability would be somewhat low, yet we may still want to sum the scores to give us an indication of how many events they experienced (pp. 55–56). […] So, although a rule of thumb cannot be provided for what a reasonable coefficient alpha may be, the mindless striving for homogeneity of tests or scales is often done at the expense of empirical usefulness of the resulting scales. Coefficients of homogeneity for any test or scale must be evaluated against the purpose of the test or scale, the construct being estimated, and the number of items in the test (p. 56).

Conclusions. The process of measurement can be described as bridging abstract research constructs that are not directly available for observation, with their empirical indicants that do lend themselves to empirical observation and direct measurement. Social research typically deals with theoretical metrics – latent parameters that are aggregate abstract notions behind a response to a specific survey question that are more informative than a response to a survey item itself [2; 26]. Development of valid metrics and construction of research instruments securing an adequate theoretical coverage of the researched constructs being measured allow claiming reliability of the collected data and draw conclusions from them. This way effective measurement is essential for maintaining the data-theory link in social sciences, thus legitimizing the research process. Moreover, an important aspect of organization of the measurement process is to do with the fact that mathematical tools that are currently available within the modern statistical toolkit, often carry the signatures of research problems they have been historically designed to serve or help solve [7; 8].

BIBLIOGRAPHY:

1. Butcher J. N., Dahlstrom W. G., Graham J. R., Tellegen A., Kaemmer B. *Minnesota Multiphasic Personality Inventory (MMPI-2). Manual for Administration and Scoring*. Minneapolis, MN: University of Minnesota Press, 1989.

2. Carmines E. G., Zeller R. A. *Reliability and validity assessment*. Newbury Park, CA: Sage, 1979.

3. Clark L. A., Watson D. Constructing validity: Basic issues in objective scale development. *Psychological Assessment.* 1995. Vol. 7. P. 309–319.

4. Comrey A. L., Lee H. B. *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum. 1992.

5. Cortina J. M. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology.* 1993. Vol. 78. P. 98–104.

6. D'Andrade R. *Study of personal and cultural values: American, Japanese and Vietnamese*. New York, NY: Palgrave Macmillan, 2008.

7. DeVellis R. T. *Scale development. Theory and applications*. Second Edition. Applied Social Science Research Methods Series (Vol. 26). Thousand Oaks: SAGE Publications, 2003.

8. Duncan O. D. *Notes on social measurement: Historical and critical*. New York, NY: Russel Sage, 1984.

9. Furr R. M. *Scale construction and psychometrics for social and personality psychology.* London: SAGE Publications, 2011.

10. Graham J. Congeneric and (essentially) Tau-equivalent estimates of score reliability: what they are and how to use them. *Educational and Psychological Measurement,* 2006. Vol. 66. № 6. P. 930–944.

11. Hathaway S. R., McKinley J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation, 1943.

[4] The length of scale also taps into the question of item redundancy (Alpha depends on homogeneity and the length of scale [13, p. 471].

12. Hulin C., Cudeck R., Netemeyer R., Dillon W. R., McDonald R., Bearden W. Measurement. *Journal of Consumer Psychology.* 2001. Vol. 10. № 1-2. P. 55–69. doi: 10.1207/s15327663jcp1001&

13. John O. P., Soto C. J. The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), Handbook of Research Methods in Personality Psychology (pp. 461– 494). New York: Guilford, 2007.

14. Maltseva K. Using correspondence analysis of scales as part of mixed methods design to access cultural models in ethnographic fieldwork: Prosocial cooperation in Sweden. *Journal of Mixed Methods Research.* 2016. Vol. 10. № 1. P. 82–111. https://doi.org/10.1177/1558689814525262

15. Maltseva K., D'Andrade R. Multi-item scales and cognitive ethnography. In D. B. Kronenfeld, G. Bennardo, V. C. de Munck, & M. Fischer (Eds.), *A companion to cognitive anthropology* (pp. 153–170). Oxford, England: Blackwell, 2011.

16. Messick S. Validity of psychological assessment. *American Psychologist.* 1995. Vol. 50, P. 741–749.

17. Netemeyer R. G., Bearden W. O., Sharma S. *Scaling procedures: Issues and applications*. Thousand Oaks, California: Sage, 2003.

18. Neuman W. L. *Social research methods. Qualitative and quantitative approaches* (7th ed.). Boston, MA: Allyn and Bacon, 2011.

19. Nunnally J. C. *Psychometric theory*. New York, NY: McGraw-Hill, 1978.

20. Nunnally J. C., Bernstein I. H. *Psychometric theory*. Third Edition. McGraw Hill Series in Psychology. New York, NY: McGraw Hill Inc, 1994.

21. Revelle W., Garne K. M. Measurement: reliability, construct validation, and scale construction. In Harry T. Reis, Tessa West, Charles M. Judd (editors) *Handbook of Research Methods in Social and Personality Psychology* (Cambridge Handbooks in Psychology). Ch. 20, pp. 471–501. Cambridge: Cambridge University Press, 2025.

22. Schmitt N. Uses and abuses of coefficient alpha. *Psychological Assessment.* 1996. Vol. 8. P. 350–353.

23. Simms L. J. Classical and modern methods of psychological scale construction. Social and Personality Psychology Compass. 2008. Vol. 2. № 1. P. 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

24. Streiner D.L., Norman G.R. Health measurement scales: A practical guide to their development and use. 4th Edition, Oxford University Press, Oxford, 2008.

25. Weller S. C., Romney A. K. *Metric scaling. Correspondence analysis*. Newbury Park, CA: Sage, 1990.

26. Zeller R. A., Carmines E. G. *Measurement in the social sciences: The link between theory and data*. Cambridge: Cambridge University Press, 1980.