The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2531-0488.htm

Received 10 May 2019
Accepted 8 August 2019

Development and validation of attitudes measurement scales: fundamental and practical aspects

Joseph F. Hair, Jr
University of South Alabama, Mobile, Alabama, USA

Marcelo L.D.S. Gabriel
Graduate Program in Management, Universidade Ibirapuera, Sao Paulo, Brazil

Dirceu da Silva
Universidade Estadual de Campinas, Campinas, Brazil, and

Sergio Braga Junior
Faculdade de Ciências e Engenharia de Tupã, Universidade Estadual Paulista, Tupã, Brazil

Abstract

Purpose – This paper aims to present the fundamental aspects for the development and validation (D&V) of attitudes' measurement scale, as well as its practical aspects that are not deeply explored in books and manuals. These aspects are the results of a long experience of the authors and arduous learning with errors and mistakes.

Design/methodology/approach – The nature of this paper is methodological and can be very useful for an initial reading on the theme that it rests. This paper presents four D&V stages: literature review or interviews with experts; theoretical or face validation; semantic validation or validation with possible respondents; and statistical validation.

Findings – This is a methodological paper, and its main finding is the usefulness for researchers.

Research limitations/implications – The main implication of this paper is to support researchers on the process of D&V of measurement scales.

Practical implications – Became a step-by-step guide to researchers on the D&V of measurement scales.

Social implications – Support researchers on their data collection and analysis.

Originality/value – This is a practical guide, with tips from seasoned scholars to help researchers on the D&V of measurement scales.

Keywords Business research, Scale development, Scale validation, Quantitative methods, Attitude measurement, Attitudes scale

Paper type Technical paper

Author Contributions: All the authors have contributed equally to the manuscript.

RAUSP Management Journal
Vol. 54 No. 4, 2019
pp. 490-507
Emerald Publishing Limited
2531-0488
DOI 10.1108/RAUSP-05-2019-0098

1. Introduction

Measurement of attitudes is a topic of keen interest in the social sciences and related fields. The origins of measurement stem from the work of social psychologists in the 1920s and 1930s that raised the topic to a higher level of empiricism and established the foundations of current research in this area.

The most widely used, and misused, approaches to measuring attitudes are the Likert-type scales, named after Rensis Likert. He created the summed or attitudes scales in 1932, in his seminal article entitled *The Technique for the Measurement of Attitudes* (Likert, 1932).

In the paper, Likert addressed a problem presented in 1929, by psychologist Gardner Murphy on five areas of attitudes, and his suggestions led to the simplification of the Thurstone measurement technique (Thurstone, 1928), which was the basis for the psychometric theory of attitude measurement. Since its proposition, it has been used in numerous fields, especially in psychology, sociology, education, business administration, anthropology, among many other fields in the social sciences and humanities.

Intended to be a very practical guide for the non-expert researcher, this article is an extension of psychometric theory as proposed by Likert (1932), Nunnally and Bernstein (1994) and Stevens (1946), as well as on applied usage of such theories in scale development as found in De Vellis (2003) and Netemeyer, Bearden, and Sharma (2003) and on management literature concerning the topic and practical examples.

2. Background

Before proceeding with the article, some recommendations are essential for establishing a common understanding. First, a clear and concise definition is required. Two terms are often used as synonyms – measure and measurement. However, the term measure refers to the evaluation of physical quantities, or measurable phenomena such as mass, temperature, length, time, while the term measurement is more appropriate for attitudes, perceptions, opinions, behavior or non-direct measurable phenomena.

Second, the process of development and validation (D&V) of attitude measurement scales is presented in four stages: a literature review or interviews with experts; theoretical or face validation; semantic validation or validation with possible respondents; and statistical (empirical) validation (Netemeyer *et al.*, 2003). However, there is a step zero, which is the definition and understanding, with the highest possible precision, of the knowledge domain, or in other words, what attitude is supposed to be measured (De Vellis, 2003).

Third, there are situations in which measurement scales have been developed and are established in other languages. These cases are not within the scope of this paper, but two steps are useful in such cases. First, existing scales should have a new cross-cultural translation and adaptation process. As an example, Beaton, Bombardier, Guillemin, and Ferraz (2000) provided straightforward and practical recommendations; second, after the first step, the adapted scale must be validated again in the new context. Behling and Law (2000) are also a recommended source for additional information.

Finally, the minimum sample size for exploratory factor analysis (EFA) is has been debated and is not consistent. As in any statistical sampling, more is always better. The rule of thumb adopted by many researchers, to generate a potential validity criterion of probability, is that the number of respondents for each assertion is equal to the number of response options in each assertion. From a practical perspective, a scale composed of 14 items (or assertions) with five options to be chosen by the respondent on each assertion, will require a minimum sample of 70 respondents. Others propose the ideal sample would be ten respondents per item on the scale, but Hair, Black, Babin, and Anderson (2019b) indicate a ratio of five respondents per scale item is sufficient.

As a recommendation, besides the rule of thumb guidelines, the minimum sample should comprise at least 20 per cent more respondents, owing to the customary issues of missing responses, straightlining or responses and other similar issues in data collection and cleaning. For more details, see Hair *et al.* (2019b).

3. Literature review or interviews with experts

There is a clear distinction between the two processes. A literature review is performed when the information and scientific knowledge about the object to be measured is abundant and available. On the other hand, interviews with knowledgeable experts are chosen when no or limited information or knowledge exists about the objectives of the research.

In other words, interviews are completed when the object of the survey is new or "virgin" where not much is known yet. As an example, a researcher in 2007 wants to perform a survey on the perceptions of smartphones users. The available scientific knowledge about this phenomenon (a brand new product at that point in time) would be hard to find in journals, so in-depth interviews (qualitative research approach) are a suitable solution.

This is perhaps the most challenging step and can lead to errors and mistakes, as the novice or inexperienced researcher often believes in this as the most natural step. So here is the advice from seasoned researchers: this is not an easy task. If the scale starts with the wrong premises, the consequences are likely to be of little value, and the results obtained frustrating, to say it gently. Remember the old saying from the statistical analysis – "garbage in, garbage out"? As a golden rule in scale development, you should understand that a scale cannot be developed based only on an individual's previous experience, particularly when the field is new and emerging.

Like many other aspects of scientific exploration, developing a scale has a recommended method and a well-founded procedure. The first step is, as already noted, a sound literature review, based on relevant papers from recognized journals when the subject is already known. This concern is a quality criterion for two reasons:

(1) relying on a relevant paper in a recognized journal is the natural focus of proper research; and

(2) the selection of such references ensures the state-of-the-art research in any scientific field is relied upon.

The proper literature review also enables the researcher to build a reference table with, for example, three columns. On the first column, the name of the construct; on the second column, list the items on the scale that will measure the construct; and on the third column, identify the references that supported the definition of the construct and the items to measure it, as seen in Table I.

Table I.
Table for the literature review organization and scale development support

Constructs	Items	References
(Green) Attitudes	I even forget that there are green products on the market	Wu and Chen (2014)
	The consumption of green products is viable for everyone	Wu and Chen (2014)
	Consuming green products is of great importance for people	Wu and Chen (2014)
	Consuming green products is fundamentally important	Blackwell *et al.* (2005)
	Consuming green products guarantees our future	Blackwell *et al.*, 2005

Source: Braga, Martínez, Correa, Moura-Leite, & Da Silva (2019)

The reference table is critical to provide an overview for the reader on how the scale was developed. It also will make it possible for other researchers to replicate the process, ensure transparency and establish credibility about the scale development process.

As mentioned, the D&V of a measurement scale are based on a series of steps. But the construction of this reference table is a preliminary step that does not need to be included in the final version of an article describing the measurement scale.

A clear definition is required before proceeding to the next steps on D&V of a measurement scale and is related to the nature of the constructs that will be developed.

Constructs are mental creations and do not exist as such, which implies some difficulties to accurately define and measure what they are. Consider beauty, for example. A precise evaluation of beauty, as a human mental creation, is very difficult but beautiful things can be assessed indirectly through a specific set of elements. Understanding this while developing measurement scales leads to an initial conclusion: constructs (concepts) cannot be defined or evaluated (measured) by using a single item. Valid and reliable measurement of constructs requires multiple indicators to be used.

If a single item is used to measure a construct, it is not considered a construct, but rather a measured variable. Similarly, other human mental creations such as education, satisfaction, quality, brand engagement or brand trust and loyalty must be measured with several indicators to be valid and reliable. Thus, no single item is capable of accurately measuring a complex concept such as loyalty, satisfaction and similar abstract concepts.

Some synonyms to constructs found in the literature are latent variable, subscale, unobserved variable, unmeasured variable, factor, component, composite and so forth.

If a construct is composed of several items, also referred to as indicators, extra care should be taken when reviewing the literature to avoid generic phrases to define the items of a scale. In short, researchers should create an operational set of items that accurately reflect the concept being measured. Bevilacqua and Petroni (2002) proposed "after-sales technical support" as a construct, but no items were generated to measure the construct. In such situations, researchers must identify relevant indicators measuring the meaning like "I always try to evaluate the quality of after-sales technical support when choosing a supplier".

As with constructs, items also have their synonyms – observed variable, measured variable, indicator, etc.

To produce operational items on a scale, avoid lukewarm statements about the concept. Items that are neither hot nor cold in terms of attitude tend to be answered in a neutral position, and they will not reflect the "underlying dispositions toward overt actions" or even "verbal substitutes for overt action" as defined by Likert (1932). At the same time, items should not be considered as leading respondents to provide a particular answer.

4. Theoretical or face validation

After the reference table's development, it is time to move on to the second step, called theoretical or face validation. A clear distinction is required here between the theoretical and the empirical validation, as the latter is performed through statistics procedures and will be explained later in this paper.

In this step, the researcher developing the measurement scale should ask the experts or specialists (preferably other researchers, PhD level, who have experience in the construct domain of the scale) to evaluate the items and their relationship to the measured construct.

The quantity of these specialists is varied. Lawshe (1975) proposed an approach called content validity ratio (CVR), Lynn (1986) further explored the index of content validity and Ayre and Scally (2014) reproduced Lawshe's CVR to build a critical CVR values' table, which is very useful as a guide to the freshman researcher. The crucial point here is: does the

researcher have enough time to completely and correctly evaluate the proposed construct? Often the answer is no, unfortunately.

So the rule of thumb is – "the more, the merrier" – but researchers do not have infinite time or money to perform this activity. Something around four to six experts should solve the problem.

The warning here is: pay attention to the quality of expert responses. Beware of experts who suggest little changes in the scale, or indicate the scale is perfect, which may not be the case, or the expert did not pay much attention when reviewing the constructs.

The suggested questions to be made to the experts are:

- Is the wording of the items correct for the respondent audience (sample)?
- Should some item(s) be removed that do not apply to the construct domain?

> The reason for this deletion could be a tiny detail or something that, in practice, is not performed by the respondent. For example, a measurement scale was developed to evaluate the crowding effect on shopping in street markets and open markets, and there were two items out of context: (a) one affirmed that the lighting was adequate, and (b) another if the ambient music motivates purchases or not. In street markets with street vendors, neither of these items are relevant. This is an example of overlooking the fact that the situation this scale would be applied in would not consider these two dimensions.

- Are there other items missing from the scale?
- Are there (technical or specific) terms that can be misunderstood by the respondents?
- If the researcher already has a factor model (the constructs with your items), the experts should evaluate if each group of items belongs to each construct.

A second warning: most of the experts have many tasks to do, and performing a proper analysis takes much time, which can cause a delay in the research project.

5. Semantic validation or validation with possible respondents

Semantic validation is a confirmatory step to gauge the effectiveness of the developed scale if applied to the respondents who are the focus of the research, the target sample.

This step is not a simple application such as a pilot sample, but rather a "meta-analysis" process. That is, with scale development, one must consider the difficulties of the respondents in understanding the statement (the semantics) of the items. The respondents are not supposed to really "answer" your scale, but evaluate the meaning and understanding of each statement and respond accordingly.

For example, if the scale has a technical term, such as "downsizing" in the management field which means the reduction in the company's size to improve organizational efficiency, and this scale will be applied to general employees, composed of different functions, varied educational background and unbalanced work experience, it is highly likely that many respondents do not know the terms used by managers and people from management areas of the company.

There are also regional or colloquial terms that must be considered, so they do not cause problems with the interpretation of the items by the respondents.

Between 20 to 40 respondents should be obtained to complete this step (the rule is still the same – the more, the merrier). Parsimony is also welcomed: get the maximum information with minimum cost.

The ideal situation to perform a semantic validation session is obtained by putting together the respondents in the same room, but most of the time, this is impossible to accomplish.

As an alternative, individual interviews should be held. In both cases, the researcher should explain that this is an initial application of the scale and ask for the help of the respondents to verify if the instrument is easy to understand.

To do this, the researcher asks the respondent to highlight by making a graphic mark (dash, circle, etc.) on words or phrases misunderstood. The best way to do this procedure is via a printed questionnaire, as respondents will mark it through. One can also send a Word file to individuals and ask them to comment, but it is often difficult to obtain adequate responses.

After collecting the completed and annotated questionnaire, the researcher will ask the respondent to further explain the notes made.

Once the researcher has the information about the respondents' questions, an in-depth analysis of what the questions represent, and how it must be changed/rephrased to be better understand is required. It is always beneficial to show the changes to other researchers if the experts are no longer available because of time or money constraints.

In an ideal world, the second round of semantic validation would be required after the changes, but two things generally happen at this time:

(1) A new group of respondents is not available.

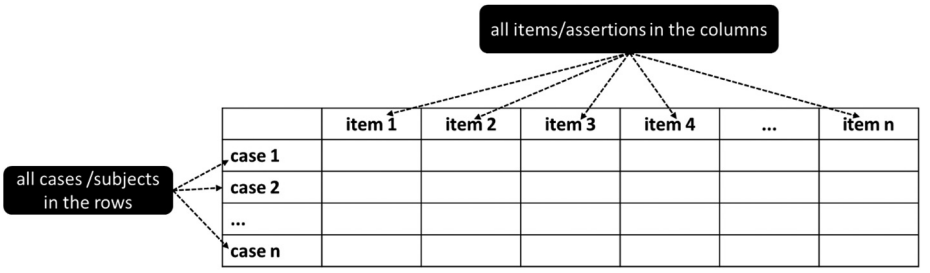
(2) Resources are scarce. If the researcher can overcome the two barriers presented, the results will be better.

A final comment on semantic validation – the respondents that evaluated the scale in this phase must not be recruited during the data collection, and the reason is quite simple – they already know the scale and probably had time to reflect and "build" an ideal set of answers, many of them socially desirable.

6. Statistical or empirical validation

Statistical validation is the easiest part to accomplish scale validation if the researcher masters some quantitative data analysis techniques. After applying the questionnaires (a.k.a.: data collection), the researcher must create a spreadsheet with the assertions or items in the columns and the respondent subjects (or cases) in the rows, as shown in Figure 1.

As pointed by Hair *et al.* (2019b), missing data are often a problem in social sciences research when data are obtained through survey research, which is the case of measurement scales. Also, deciding how to handle missing data is always complicated. There are several



Source: The authors

Figure 1.
A suggested scheme for data tabulation

procedures when facing the issue of missing data: listwise, pairwise, replace with mean and impute a response.

The listwise procedure deletes the subject (the entire line). Pairwise eliminates the subject when the variable with the missing data enters the analysis. Finally, as the name says, the latter is the substitution by the mean value of the responses of the variables. Inserting the sample mean is controversial and should be avoided, as it reduces the variability in the data. In most instances, it is best if possible to use some method to estimate an accurate response for the missing data, or to code the data as missing so that can be considered in the analysis. See Hair *et al.* (2019b) for a summary of how to deal with missing data.

If the size of the sample is sufficient, then the best solution is to not include the subjects with missing answers in the spreadsheet.

An almost philosophical remark regarding the mean value is required here. Different from the mode and the median, the mean is a calculation that is valid only if the data is normally distributed, an assumption hard to find in social science research, so avoid mean replacement on missing data, as this procedure decreases variability in the data and reduces the possibility of finding relationships in the data.

From the data collected, there are four techniques for the statistical validation of the scale:

(1) exploratory factor analysis (EFA);

(2) confirmatory factor analysis (CFA);

(3) confirmatory composite analysis (CCA); and

(4) item response theory (IRT).

Note that CCA is a recently proposed scale validation process that is associated with partial least squares structural equation modeling (PLS-SEM) (Hair, Page, & Brunsveld, 2019a). CCA is increasingly being used to replace traditional scale development procedures, particularly EFA and when appropriate CFA.

EFA is used initially to understand and clarify new scales, as it enables the researcher to identify consistent constructs. This paper will not address the other techniques (CFA, CCA and IRT), but make some suggested guidelines.

To perform the CFA and CCA, the following sources are recommended: Hair *et al.* (2019b) and Brown (2015); for IRT, Alaya (2009) and Baker and Kim (2017) are useful; and for CCA, Hair *et al.* (2019a).

Some scholars apply EFA and CFA or CCA in the same survey. To do this, the researcher must collect either data from two separate samples, or randomly divide a single sample into halves. With one of the samples, the EFA is executed, and with the other sample, the measurement model is examined in an effort to "confirm" it with the CFA or CCA. Another approach is to execute EFA on a pilot study sample, and then execute CFA on the final sample data.

6.1 Exploratory factor analysis

EFA is the multivariate technique used when one does not apply CFA or CCA. That is, exploratory research is undertaken, and it is not known which items were added in the data analysis process to form the constructs. As the name implies, it is used to explore the data. The analysis is guided by the data (Nunnally & Bernstein, 1994).

The most commonly used technique is performed by calculating the Pearson correlations among all items (or observed variables). After this calculation, Pearson's correlation values are compared and then grouped, forming the factors or composites.

In EFA, a "mathematical art" idealized by Kaiser (1958) in 1958 is employed. It involves the rotation of the axes of the Cartesian system of reference (graph).

This procedure is executed by multiplying the correlation matrix by another one made up of sine and cosine. So that the values of correlations are not arrayed, but rather "the point of view that each value sees the others" (for a better understanding of this process, see Pett, Lackey & Sullivan, 2003; Hair *et al.*, 2019b, 2019a).

The most widely used rotation method is the varimax (orthogonal; the axes rotate remaining at 90 degrees). There are six rotation methods, including orthogonal (varimax, quartimax and equamax – produce uncorrelated factors) and oblique (Direct oblimin, quartimin and promax – cause the reference system to rotate with different angles of 90 degrees and produce correlated factors to each other). Orthogonal rotations produce more interpretable factors (Figure 2).

There are many statistical packages available for EFA calculation. For example, SPSS, SAS, Stata, Statistica, R Project, etc. By using SPSS, the researcher has to go through six stages (a very elaborate systematization with all the steps that can be found in Moretti *et al.*, 2019).

The sequence to calculate the EFA in SPSS is "Analyze → Dimension Reduction → Factor . . .".

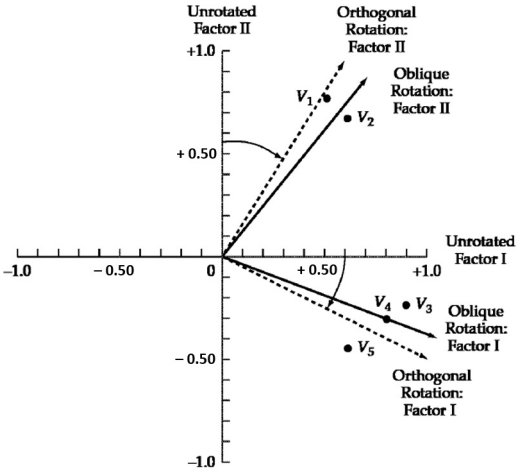
A dialog box will open, and the researcher will be ready to model the EFA. For this, five dialog boxes will be explained in details. After each selection, the button "Continue" must be selected to advance to the next dialog boxes.

6.2 Factor analysis: descriptives

There are many possibilities for selection. Select all (Figure 3).

6.3 Factor analysis: extraction

- *Method*: There are seven methods. Select "main components."
- *Analyze*: Select "correlation matrix" (this option allows the researcher to use data without a "strong" adherence to the normal distribution because it will use an array



Source: Hair *et al*. (2010)

Figure 2.
Comparison between orthogonal and oblique rotations

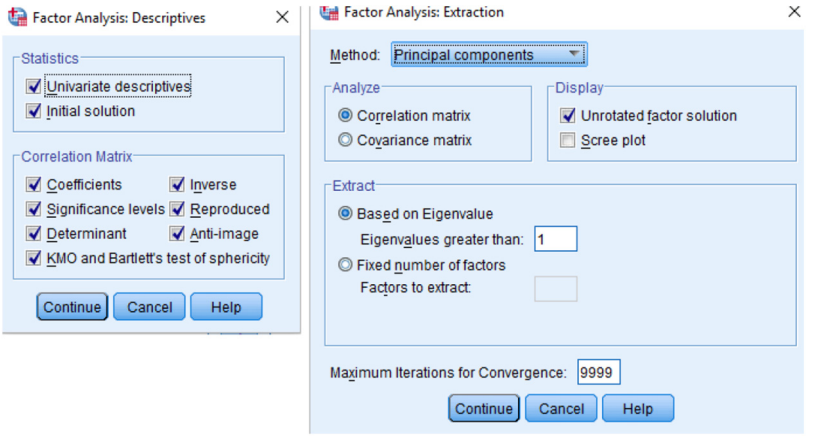


Figure 3.
Dialog boxes of descriptive statistics and extraction method in SPSSS software

of correlations). The display is optional, in the case shown in Figure 3, the "unrotated factor solution" option is the default SPSS option.

- *Extract*: There are two options "based on eigenvalue greater than 1" (free rotation) (this is the Kayser normalization. Eigenvalues smaller than 1 have no physical interpretation) or "fixed number of factors." Here, the researcher determines how many factors or constructs the software will generate (also called forced rotation). The best suggestion is always free rotation.
- *Maximum iterations for convergence*: This is the last definition in this step. The default value 25 must be changed to 9999. This feature was used when the computers had few resources. It was a limiter of computer processor usage. (Figure 3).

6.4 Factor analysis: rotation

An option is required. The varimax method is the most used; in "display" pick "rotation solution" to see the rotated matrix. As extraction, the last option is "maximum iterations for convergence." Change the value 25 to 9999 (same reason as explained on maximum iterations for convergence).

6.4.1 Factor analysis: factor scores. This dialog box is used to generate the adjusted Y values (from a regression line, which describes the factors). When the option "save as variable" is chosen, the software saves the adjusted Y values ($\hat{Y}$) of each factor ($\hat{Y} = bx + a$). These values allow classifying each respondent in the respective factors of the factorial model. The option "factor display factor coefficient matrix" must be checked as well to see the factorial solution.

Finally, the last box has two groups of options:

(1) "missing values" (check the listwise option, as explained above); and

(2) "coefficient display format."

Check both options and set to "absolute value below," 0.40 (or 40 per cent). These procedures do not modify data or analyze but make the presentation of information clearer and easier to interpret (Figure 4).

Returning to the six stages discussed by Moretti *et al.* (2019):

(1) The choice of the rotation method and the extraction method (there are seven methods of extraction, i.e., how the software will proceed to form combinations that allow the evaluation of the extracted variance, that is, the performance that the factor model explains the data).

Among the seven methods of extraction, the most used is principal component analysis, because it creates a model more organized and more accessible to be interpreted.

(2) The values of two tests should be observed: one is the Kaiser–Meyer–Olkin (KMO), which evaluates if the size of the sample as a whole is adequate to calculate the EFA. Also, the other is Bartlett's test of sphericity, which compares the data with an identity matrix, that is, assess whether the data are free of "single response bias."

In the first, (KMO) values above 0.60 or 60 per cent show that EFA can be used. In the second one, the *p*-value must be observed. The *p*-value should be less than 0.05 (test must be significant) (an example of SPSS output is in Table II).

(3) Evaluation of the KMO test for each variable.

This result, in SPSS, is in the matrix of anti-image correlations, in the main diagonal (Table III). Values equal to or above 0.5 or 50 per cent are adequate (Hair *et al.*, 2019b).

Variables with values below 0.5 should be eliminated from the factor/composite model. It is common for researchers to ask whether the increase in the sample size can "save" the eliminated variable, as the KMO (general and individual) refers to the adequacy of the variable correlations for the sample to calculate the EFA. That is a good question. There is no guarantee that additional observations will work. Thus, elimination may be the best option. In Table III, Variables 1 and 4 are appropriate, and Variables 2 and 3 are not suitable.

(4) Communality. Here are the measures of how much each variable is explained by the model. Values below 0.5 or 50 per cent should be eliminated (Hair *et al.*, 2019b).

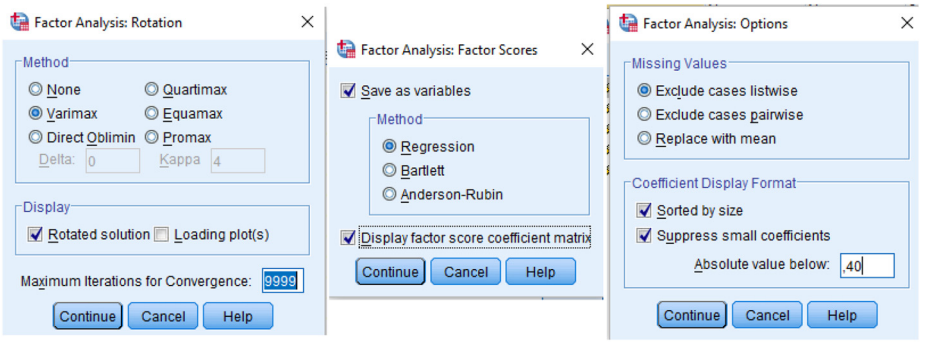


Figure 4.
Dialog boxes of rotation, factor scores and options in SPSSS software

Kaiser–Meyer–Olkin Measure of Sampling Adequacy (MSA)	0.738
Bartlett's test of sphericity	
Approx. chi square	466.814
df	91
Significance	0.000

Table II.
Output SPSS for tests KMO and Bartlett

Anti-image correlation					
VAR00014	0.045	0.173	−0.092	−0.029	0.140
VAR00001	*0.634*[a]	0.048	−0.427	0.039	0.258
VAR00002	0.046	*0.343*[a]	−,435	0.120	0.166
VAR00003	−0.427	−0.435	*0.475*[a]	−0.149	−0.304
VAR00004	0.039	0.120	0.149	*0.774*[a]	0.103
VAR00005	0.258	0.166	−0.304	0.103	0.643[a]
VAR00006	−0.276	−0.057	−0.030	−0.279	−0.130
VAR00007	0.109	−0.056	−0.026	−0.037	−0.163
VAR00008	0.194	0.152	−0.173	0.096	−0.014
VAR00009	−0.036	0.179	−0.024	0.064	0.026
VAR00010	−0.298	−0.297	0.352	−0.209	−0.318
VAR00011	−0.286	0.109	0.060	0.055	−0.095
VAR00012	−0.133	−0.159	0.168	−0.104	−0.038
VAR00013	−0.043	0.052	−0.136	−0.061	−0.016
VAR00014	0.081	0.271	−0.162	−0.042	0.221

Note: [a]Measures of MSA

Table III.
"Fragment of the output" of the EFA's correlation matrix

In SPSS output variables, 4 and 11 are not appropriate and should be eliminated (Table IV).

(5) Total variance explained (TVE) by the model. Like the commonalities, the total variance explained is the part of the data that the model can explain. Values are at or above 0.60 or 60 per cent as appropriate. Table V shows that the TVE value is 63.148, indicating that the factorial model is adequate. As a detail, it would take 14 factors to explain 100 per cent of the data, but as already mentioned initial eigenvalues below 1 have no physical explanation and thus were eliminated.

(6) After Variables 2, 3 4, and 11 were removed, the model was rerun. The overall KMO was 0.789; KMOs per variable were above 0.50; the commonality was above 0.50; the TVE was 64,753, and the rotated matrix was left with three factors or constructs (Table VI).

	Initial	Extraction
VAR00001	1.000	0.645
VAR00002	1.000	0.698
VAR00003	1.000	0.719
VAR00004	1.000	*0.414*
VAR00005	1.000	0.597
VAR00006	1.000	0.739
VAR00007	1.000	0.722
VAR00008	1.000	0.683
VAR00009	1.000	0.564
VAR00010	1.000	0.698
VAR00011	1.000	*0.470*
VAR00012	1.000	0.025
VAR00013	1.000	0.674
VAR00014	1.000	0.593

Note: Extraction method: principal component analysis

Table IV.
SPSS output communality

Component	Initial Eigenvalues Total	% of Variance	Cumulative (%)	Extraction Total
1	*4.527*	*32.333*	*32.333*	4.527
2	*1.631*	*11.649*	*43.981*	1.631
3	*1.545*	*11.032*	*55.014*	1.545
4	*1.139*	*8.134*	*63.148*	1.139
5	*0.864*	*6.170*	*69.318*	
6	0.785	5.610	74.929	
7	0.723	5.166	80.094	
8	0.616	4.401	84.496	
9	0.513	3.667	88.162	
10	0.431	3.080	91.242	
11	0.404	2.886	94.128	
12	0.343	2.447	96.575	
13	0.272	1.944	98.519	
14	0.207	1.481	100.000	

Note: Extraction method: principal component analysis

Table V.
SPSS output total variance explained

	Rotated component matrix[a] Component		
	1	2	3
VAR00013	0.845		
VAR00010	0.724		
VAR00012	0.723		
VAR00014	0.601		
VAR00005		0.768	
VAR00008		0.716	
VAR00007		0.705	
VAR00006			0.842
VAR00001			0.584
VAR00009			0.553

Notes: Extraction method: principal component analysis; Rotation method: varimax with Kaiser normalization; [a]Rotation converged in eight iterations

Table VI.
SPSS output rotated component (factors or constructs) matrix

After these adjustments, the factors will be named, that is, the researcher should find the name that represents the set of variables that it supports. There is no example to be shown here, but it is strongly recommended that the researcher discusses the findings with colleagues to check the adequateness of each choice.

The best practices suggest the researcher calculate the other available options even after adjusting the factorial model with varimax and major components. For example, quartimax, equamax and promax with the extraction of the "principal axis factoring" and "alpha factoring." After the calculation of each model, the researcher should evaluate if there were changes in the models and decide which model better explains the data.

Finally, it is time to calculate Cronbach's alpha test. It refers to the internal consistency reliability of the data.

A third warning here – Cronbach's alpha should always be calculated for the construct. When calculated for full scale, the result may be greater than 0.90 (it gets inflated). This test indicates whether the data are free of bias (highest value close to 1.0). In exploratory studies, values above 0.60 and up to 0.70 are acceptable (Hair *et al.*, 2019a; Peterson, 1994).

Another situation that may occur with the Cronbach's alpha test values is a small number of indicators. When a construct has only a few items, the alpha test value tends to be small but that does not imply bias.

This test indicates whether the data are free of bias (highest value close to 1.0). A different approach, not covered in this paper, is the Dillon–Goldstein rho test (or composite reliability), which is more stable and does not depend on the number of items in the constructs (for more details see Hair *et al.*, 2019b). In addition, composite reliability weight the individual indicators while the Cronbach's alpha considers all indicators to be equally weighted.

The path to calculate the Cronbach's alpha test values for each construct in the SPSS software is: Analyze → Scale → Reliability Analysis.

Pick the measured variables (scale item) that form a construct and then choose "statistics" → scale if item deleted.

This procedure will show which variables (or items) should be scaled or removed from the model (Table VII for a hypothetical analysis). Analyzing Table VII, if the variable VAR00002 were deleted from the construct, the value of the alpha test would be 0.730. The "full" value, with the eight variables in the hypothesized construct, is 0.681.

Despite the evidence that removes VAR00002 will increase the Cronbach's alpha coefficient, the decision must be based on the subjacent theory of the measured construct, which requires the researcher to invest time in evaluating each assertion in the scale to avoid deletion based only on statistics.

7. Practical suggestions

The following suggestions are supposed to be used as a checklist to the researcher when developing a measurement scale. The suggestions are ranked by numbers and comments, when required, and are presented in italic:

(1) The indicators must express the desired behavior and not the fact to be measured.

(2) The indicators must be clear, concise and direct.

Table VII. Output SPSS for a decision on the Cronbach's alpha test

Reliability statistics

Cronbach's alpha	No. of items
0.681	8

Item-total statistics

	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Cronbach's alpha if item deleted
VAR00001	31.9900	23.788	0.337	0.659
VAR00002	32.1600	27.388	0.015	0.730
VAR00003	31.5500	23.220	0.477	0.630
VAR00004	32.2600	23.750	0.305	0.667
VAR00005	32.1500	24.391	0.335	0.659
VAR00006	31.4100	21.194	0.574	0.600
VAR00007	32.2100	20.794	0.527	0.608
VAR00008	31.9100	21.719	0.454	0.629

(3) It is recommended that 10 per cent of the indicators be negatively worded to minimize bias in the responses.

(4) It is desirable that the indicators be placed in a random order so that they can minimize the learning or repetition response trends.

Experts express different opinions about this guideline, but the practice has shown that it may be useful. That is, it may prevent responses from being invalid by a response tendency, a bias or a halo effect (a type of the "contamination" from one response to another – see for example Wirtz & Bateson, 1995).

(5) Each item should measure only one proposition or conceptualization; it should not be double barreled with two issues embedded in one question. Avoid the usage of "and" (conjunctive) and "or" (disjunctive) in the indicators.

For example, if the scale requests a score from 0 to 10 for the respondent's perception of the item, "I always buy ***and*** *recommend the Nice Place store." If the respondent only buys, would the answer be a 5.5 note?*
To eliminate this problem, the statement should be split in two: "I always shop at the Nice Place store" and then "I always recommend the Nice Place store."

(6) Number of answer categories in a scale.

On his seminal article, Likert (1932) proposed the famous five-point scale, and since then, many scales were developed following this rule. The reason for odd points in a scale (3, 5, 7, 9, etc.) is related to the construction of each assertion that "seem desirable to have each statement so worded that the ***modal*** *reaction to it is approximately in the middle of the possible responses." (Likert, 1932).*

Also, as stated in Item 1 above, what is intended to be measured is the desired behavior or present attitude of each respondent. As such, Likert (1932) also recommend the usage of the term "should" on each statement.

Finally, in almost all instances, scales that were developed with five response categories should be converted to a minimum of 7 points to respond to, and in some instances, a 0 to 10 point scale would be most appropriate. More scale points increase the variability in the responses, and accurate statistical analysis requires variability in the data.

There are many ways to present the indicators and the possibilities for answers (or choices). Considering the following item: "***I always seek to buy products from environmentally correct companies,***" here are some possibilities of presentation format:

A) Please mark an X in the parentheses that indicates your answer:

( ) totally disagree	***( ) disagree***	***( ) indifferent***	***( ) agree***	***( ) totally agree***

B) Please mark a X, from 1 to 7, where 1 represents the total disagreement and 7 represents the total agreement

( ) 1	( ) 2	( ) 3	( ) 4	( ) 5	( ) 6	( ) 7

C) Please give a note of 0 to 10, where 0 represents the total disagreement and 10 represents he total agreement. NOTE_____

D) Please mark an X in the parentheses that indicates your answer:

( ) Agree	( ) Neither agree nor disagree	( ) Disagree

There are many other variations (Hair, 2011) . To decide which type to use, consider two aspects:

- The higher the number of categories, the broader the spectrum of responses and the data analysis may be more revealing, as it encourages more variability in responses.
- College people do not have problems with Type C, but for children and older adults, the Type D may be better. Type A is a widely used form of research with the general public. Type B is also widely used, but it can create problems for people who are not experienced in responding to Likert scales.
- As defined by Stevens (1946), except Type C, all the other types are ordinal scales which means that the numbers presented (as on Type B) are used as "positional mark" to help respondents to position the answer in a continuum from total disagreement to complete agreement. As such, the numeric properties of such a scale does not allow the researcher to calculate the mean and the standard deviation of responses. In these cases, the only valid and allowed statistics are mode and frequencies.

(7) A Likert scale should have several items covering all possible construct responses.

Attention to the fact that scales with many items can distract the respondent leading to less attention in answering the items, particularly at the end of the scale. Therefore, the researcher should evaluate the time required for the answers. Most recently, Web-based research platforms have developed algorithms to estimate the time required to complete the survey.

In addition, to convert the original Likert scale format with labels on all categories from ordinal to interval data, you must remove the category labels in the middle of the scale. When using Likert scales, category labels should be used only on the ends of the scales and removed from the middle categories. Also, the scale points should be labelled below with numbers so the respondent perceives them as being equally distant apart. The result will be intervally scaled data and not ordinal, which extends the flexibility of using the data with many other statistical methods. Note that in most situations, scales should have at least 7 points to respond to, and in some situations, 9 or 11 points (Hair, Celsi, Money, Samouel, & Page, 2016) to increase the variability of the data. Finally, if existing scales are being adapted in almost all situations, the traditional five-point scale should be converted to a seven-point or higher number of response options.

(8) Self-evident truths or self-evident lies:

A phrase that contains a concept that the vast majority or the totality of respondents will agree or disagree. As an example, suppose that an item on a scale is:

"The supreme head of the Catholic Church is the Pope."

( ) totally disagree	***( ) disagree***	***( ) indifferent***	***( ) agree***	***( ) totally agree***

Or

"Who really rules is the Queen in England"

( ) totally disagree	***( ) disagree***	***( ) indifferent***	***( ) agree***	***( ) totally agree***

How would the respondent answer to these two items?
The self-evident truths (or lies) end up being excluded in any statistical treatment since it has a variance close to zero (or zero, that is, the answers are equal).
The researcher should keep in mind that the best definition for any variable is that which varies! If it does not varies, it is constant!

(9) Double negative.

For example, "I do not eat any broccoli." That means "I always eat broccoli." The correct is "I eat no broccoli". This kind of situation can create different responses for different people. Some will not observe the negative combination, and others will interpret and give a different response from the first group.

(10) Social desirability (SD) evaluation.

*As already stated, a measurement scale is an indirect approach to measure latent variables or constructs and as such, depends upon the respondents' sincerity on each response. If well instructed about the confidentiality of data collected and the proper use of the results, people are more likely to cooperate with researchers, and the results are beneficial to the advancement of knowledge. In cases where the topics covered on the measurement scale are controversial or could negatively expose some group of respondents, it is quite possible that respondents will choose socially desirable answers, leading the researcher to wrong conclusions. One possible approach to avoid respondents who are likely to provide socially desirable answers is to use the Social Desirability (SD) scale, developed by Crowne and Marlone (*Crowne, 1960*) after Allen* Edward's (1957*) studies. The scale is composed of 33 items, and the answers can be evaluated being as* ***socially acceptable but not probable or socially unacceptable but probable****. The results from the SD scale can be used to choose which cases are valid in such situations.*

8. Final considerations

This paper is intended to be a practical guide to the novice researcher on D&V of attitudes measurement scales. To achieve this goal, a prototype of a framework was built, starting from the literature review and interviews with experts to define the construct domain, the writing of each statement and its caveats, the empirical (statistical) validation using EFA and finally several practical suggestions with methodological comments on each topic.

The complexity involved in the D&V of a measurement scale is treated in various reference manuals, many of them cited in this paper. Owing to the nature of a methodological article aimed at a broader audience, several concepts were not fully developed, but the authors strongly recommend the interested researcher to study the topic with the references presented here.

The topic of validity was summarized in the article and included theoretical and empirical validation. The interested researcher should read the Cronbach and Meehl's (1955) seminal article on the topic, as well CFA described by Hair *et al.* (2019b), and CCA summarized by Hair *et al.* (2019a).

As a last advice, researchers should keep in mind that a good scale should be:

- reliable, which means that repeated applications of the same scale must lead to consistent scores; and
- valid, the confirmation that a scale measures the construct (or constructs) that is (are) intended to be measured.

D&V of measurement scales is a vast field of knowledge, and the authors hope the present article can "enlighten" possible novice researchers who wish to initiate the construction and validation of scales and bring new practitioners to this field.

References

Alaya, R. J. (2009). *The theory and practice of item response theory*, New York, NY: The Guilford Press.

Ayre, C., & Scally, A. J. (2014). Critical values for Lawshe's content validity ratio: Revisiting the original methods of calculation. *Measurement and Evaluation in Counseling and Development*, *47*, 79–86.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R*, Cham, Switzerland: Springer International Publishing AG.

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*, 3186–3191.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions* (Vol. 133). Thousand Oaks, CA: Sage.

Bevilacqua, M., & Petroni, A. (2002). From traditional purchasing to supplier management: A fuzzy logic-based approach to supplier selection. *International Journal of Logistics Research and Applications*, *5*, 235–255.

Braga, S. Jr, Martínez, M. P., Correa, C. M., Moura-Leite, R. C., & Da Silva, D. (2019). Greenwashing effect, attitudes, and beliefs in green consumption. *RAUSP Management Journal*, *54*, 226–241.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*, 2nd ed., New York, NY: The Guilford Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Crowne, D. P. (1960). Marlowe-Crowne social desirability scale. *Journal of Consulting Psychology*, *24*, 349–354.

De Vellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed., Vol. 26). Thousand Oaks, CA: Sage Publications.

Edwards, A. L. (1957). *The social desirability variable in personality assessment and research*, Worth, TX: Dryden Press.

Hair, J. F., Page, M., & Brunsveld, N. (2019a). *Essentials of business research methods* (4th ed.). New York, NY: Routledge.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019b). *Multivariate data analysis* (8th ed.). London, Unites Kingdom: Cengage Learning.

Hair, J. F., Celsi, M., Money, A., Samouel, P., & Page, M. (2016). *Essentials of business research methods* (3rd ed.). New York, NY: Routledge.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5–55.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, *35*, 382–385.

Moretti, E. A., Anholon, R., Rampasso, I. S., Silva, D., Santa-Eulalia, L. A., & Ignácio, P. S. A. (2019). Main difficulties during RFID implementation: An exploratory factor analysis approach. *Technology Analysis and Strategic Management*, *31*, doi: 10.1080/09537325.2019. 75351.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*, London, United Kingdom: Sage Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*, New York, NY: McGraw-Hill.

Peterson, R. A. (1994). A meta-analysis of Cronbach´s coefficient alpha. *Journal of Consumer Research*, *21*, 381–391.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*, Thousand Oaks, CA: Sage Publications.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677–680.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554.

Wirtz, J., & Bateson, J. E. (1995). An experimental investigation of halo effects in satisfaction measures of service attributes. *International Journal of Service Industry Management*, *6*, 84–102.

Corresponding author
Marcelo L.D.S. Gabriel can be contacted at: mgabriel.br@gmail.com