Scientific Research Publishing

Psychology, 2018, 9, 2531-2560
http://www.scirp.org/journal/psych
ISSN Online: 2152-7199
ISSN Print: 2152-7180

Applied Psychometrics: The Steps of Scale Development and Standardization Process

Theodoros A. Kyriazos, Anastasios Stalikas

Department of Psychology, Panteion University, Athens, Greece
Email: th.kyriazos@gmail.com

How to cite this paper: Kyriazos, T. A., & Stalikas, A. (2018). Applied Psychometrics: The Steps of Scale Development and Standardization Process. *Psychology, 9,* 2531-2560.
https://doi.org/10.4236/psych.2018.911145

Received: September 10, 2018
Accepted: October 16, 2018
Published: October 23, 2018

Abstract

This work focuses on presenting the development process of a self-reporting measurement instrument. Numerous scale development procedures are reviewed. They are all summarized into an overall framework of consecutive steps. A concise description is contained in each step. Issues covered comprise the following. First, the theoretical underpinning of the scale construct is described, along with the response specifications and response formats available (most popular like Likert and some more elaborated). Then the item writing guidelines follow together with strategies for discarding poor items when finalizing the item pool. The item selection criteria described comprise an expert panel review, pretesting and item analysis. Finally, the dimensionality evaluation is summarized along with test scoring and standardizing (norming). Scale construction has implications on research conclusions, affecting reliability and the statistical significance of the effects obtained or stated differently the accuracy and sensitivity of the instruments.

Keywords

Test Constriction, Scale Development, Questionnaires, Self-Report Scales, Item-Writing, Scaling, Item Analysis, Factor Analysis, Expert Panel Review, Standardization, Norming

1. Introduction and Basic Concepts

Questionnaire (also called a test or a scale) is defined as a set of items designed to measure one or more underlying constructs, also called latent variables (Fabrigar & Ebel-Lam, 2007). In other words, it is a set of objective and standardized self-report questions whose responses are then summed up to yield a score. Item score is defined as the number assigned to performance on the item, task, or stimulus (Dorans, 2018: p. 578). The definition of a question-

naire or test is rather broad and encompasses everything from a scale, to measure life satisfaction (e.g. the SWLS Diener et al., 1985), to complete test batteries such as the Woodcock-Johnson IV battery by Schrank, Mather, and McGrew (2014) comprising cognitive tests, (Irwing & Hughes, 2018). The scale items are indicators of the measured construct and hence the score is also an indicator of the construct (Zumbo et al., 2002; Singh et al., 2016). Generally, there are attitude, trait, and ability scales (Irwing & Hughes, 2018). Attitude, ability and intellectual reasoning measures or personality measures are considered as technical tools, equivalent e.g. to a pressure gauge or a voltmeter (Coolican, 2014). Over the past decades, such instruments became popular in psychology mainly because they provide multiple related pieces of information on the latent construct been assessed (Raykov, 2012). Scale Development or construction, is the act of assembling or/and writing the most appropriate items that constitute test questions (Chadha, 2009) for a target population. The target population is as the group for whom the test is developed (Dorans, 2018). Test development and standardization (or norming) are two related processes where test development comes first and standardization follows. During test development, after item assembly and analysis, the items which are strongest indicators of the latent construct measured are selected and the final pool emerges, whereas in standardization, standard norms are specified (Chadha, 2009). Effective scale construction has important implications on research inferences, affecting first the quality and the size of the effects obtained and second the statistical significance of those effects (Furr, 2011), or in other words the accuracy and sensitivity of the instruments (Price, 2017). A set of standards for assessing standardized tests for psychology and education has been published jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA/APA/NCME, 1999, 2014; Streiner, Norman, & Cairney, 2015). Generally, successful tests are developed due to some combination of the three following conditions (Irwing & Hughes, 2018): 1) Theoretical advances (e.g. NEO PI-R by Costa & McCrae, 1995); 2) Empirical advances (e.g. MMPI by Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989); 3) A practical or market need (e.g. SAT by Coyle & Pillow, 2008).

The purpose of this work is to provide a review of the scale development and standardization process.

2. The Scale Development Process Overview

The scale development process as described by Trochim (2006) is completed in five steps (as quoted by Dimitrov, 2012): 1) Define the measured trait, assuming it is unidimensional. 2) Generate a pool of potential Likert items, (preferably 80-100) rated on a 5 or 7 disagree-agree response scale. 3) Have the items rated by a panel of experts on a 1 - 5 scale on how favorable the items measure the construct (from 1 = strongly unfavorable, to 5 = strongly favorable). 4) Select the

items to retain for the final scale. 5) Administer the scale and to some of the responses of all items (raw score of the scale), reversing items that measure something in the opposite direction of the rest of the scale. Because the overall assessment with an instrument is based on the respondent's scores on all items, the measurement quality of the total score is of particular interest (Dimitrov, 2012). In a similar vein, Furr (2011) also described it as a process completed in five steps: (a) Define the Construct measured and the Context, (b) Choose Response Format, (c) Assemble the initial item pool, (d) Select and revise items and (e) Evaluate the psychometric properties (see relevant section). Steps (d) and (e) are an iterative process of refinement of the initial pool until the properties of the scale are adequate. Test score then can be standardized (see relevant section).

There are several models of test development. In practice, steps within the different stages may actually be grouped and undertaken in different combinations and sequences, and crucially, many steps of the process are iterative (Irwing & Hughes, 2018). In **Table 1** the scale development process described by multiple different sources is presented as the steps suggested by different sources differ. Note that in **Table 1** an integrative approach to the scale development process combining steps by all sources is contained at the bottom of **Table 1**. The phases of the scale development process are presented in the sections below.

3. Phase A: Instrument Purpose and Construct Measured

When instruments are developed effectively, they show adequate reliability and validity supporting the use of resulting scores. To reach this goal, a systematic development approach is required (Price, 2017). However, the development of scales to assess subjective attributes is considered rather difficult and requires both mental and financial resources (Streiner et al., 2015). The prerequisite is to be aware of all existing scales that could suit the purpose of the measurement instrument you wish to develop, judging their use without any tendency to maximizing deficiencies before embark on any test construction adventure. Then, there is one more consideration: feasibility. Some feasibility dimensions need to be considered are time, cost, scoring, the method of administration, intrusiveness, the consequences of false-positive and false-negative decisions, and so forth (Streiner et al., 2015). After that, the scale development process can start with the definition of the purpose of the instrument within a specific domain, the instrument score and the constraints inherent in the development (Dimitrov, 2012; Price, 2017). As a rule, in the research field of psychology, the general purpose of a scale is to discriminate between individuals with high levels of the construct being measured from those with lower levels (Furr, 2011).

However, the test developed should first determine clearly the intended construct been measured. Defining the construct to be measured is a crucial step requiring clarity and specify (DeVellis, 2017; Price, 2017). Outlining a construct is possible by connecting ideas to a theory (e.g. the emotional intelligence;

Table 1. The scale development process described by multiple different sources.

Source	Steps of the Development process
Crocker & Algina, 1986	Set Primary test Purpose; Set Behaviors of the target construct; Set test specification; Prepare Item Pool; Review Pool items; Pilot test Pool; Revise Pool; Field test Pool; Item analysis; Validation studies; Norming-Manual
DeVellis, 2017	Set object of measurement; Generate Item Pool; Set Format for Measurement; Expert panel review of Pool; Consider Inclusion of Validation Items; Validation Study; Evaluate the Items; Optimize Scale Length
Furr, 2011: p. 6	Set construct and context; Set response format; Assemble item pool; Collect data; Examine psychometric properties
Streiner et al., 2016: p. 5	Gap identified; Generate items; Test items; Revise items; Reliability studies; Validity studies; Present results
Price, 2017: p. 167	Set theoretical foundation; Set the purpose; Select construct attributes; Define testing population; Set content of the items; Write items; Develop administration procedures; Pilot test; Revise test; Develop norms; Develop the technical manual
Irwing & Hughes, 2018: p. 4	Definition, Overall planning; Item writing Panel Review Piloting; Factor Analysis & IRT; Reliability Validation; Test scoring and norming; Test specification; Implementation and testing; Technical Manua
An integrative approach of all sources	Gap identified → Set theoretical foundation → Set the purpose → Select construct attributes → Overall planning → Generate Item Pool → Set test specifications → Set administration procedures → Expert Review Pool items → Revise Pool 1 → Pilot test Pool → Item analysis of Pool → Revise Pool 2 → Validation studies → Technical Manual; Iterative Steps of the Process

Goleman, 1995). However, constructs in psychology are not directly observable (Kline, 2009; Sawilowsky, 2007; Milfont & Fisher, 2010 among many others), thus developers have first to define a general philosophical foundation to connect the construct to a set of observable traits or behaviors (Price, 2017). For example, the Broaden and Build Theory of positive emotions by Fredrickson (Fredrickson, 1998, 2001, 2003, 2013) was postulated within the positive psy-

chology movement, initiated by Seligman (Seligman, 1998; Seligman & Csikszentmihalyi, 2000) that perceives psychology in a different perspective from "as usual" (Seligman & Pawelski, 2003). That is, the philosophical foundation of a test or instrument is a connector between the construct to be measured and a related body of a material called domain (Nunnally & Bernstein, 1994: p. 295 reproduced by Price, 2017). Dimitrov (2012) offers an illustrative example: various definitions of "self-efficacy" exist in models like the Social Cognitive Theory (Bandura, 1997), the Theory of Planned Behavior (Ajzen, 1991), the Transtheoretical Model (Prochaska, Norcross, Fowler, Follick, & Abrams, 1992), and the Health Action Process Approach (Schwarzer, 2001).

Then the construct can be operationalized. Deciding on the construct is usually based on a review of related literature, along with consultation with subject-matter experts. Then a concise, clear and precise definition of the construct is generated. Using this definition, the item content is specified with precision and clarity (Price 2017; DeVellis, 2017). An initial construct definition should be as clear as possible (DeVellis, 2017) but will often be somewhat broad. From this point, by systematic literature review, existing tests are identified and the nature of the target construct is studied. After this review, the test developer can refine the construct definition further (Irwing & Hughes, 2018). The construct operationalization specifies the following: (a) a model of internal structure; (b) a model of external relationships with other constructs; (c) potential relevant indicators, and (d) construct-related processes (Dimitrov, 2012). The next step is to link domain content with domain-related criteria. Then planning is necessary (Irwing & Hughes, 2018) to specify a wide range of options available pertaining to item specifications described next. Methods to identify the attributes that accurately represent the targeted construct (especially useful in ability and intelligence tests) by Price (2017) are presented in **Table 2** and **Figure 1**.

4. Phase B: Response Scale Specifications

One of the first decisions when designing a questionnaire is whether to include open (allowing answer in the respondents' own words) or closed questions (forcing responses from a set of choices). The vast majority of items are closed, although some open questions are used in survey research or items requiring a numerical input e.g. age, weight, (Krosnick & Presser, 2010). Nevertheless, items

Table 2. Methods for identifying the attributes that accurately represent the targeted construct.

Subject-matter experts decide on the attributes to be measured
Interviews of key elements through an iterative process
Review of the related literature
Content analysis to track dimensions or topic areas
Direct observation
(Price, 2017: pp. 190-191; Wolfe & Smith, 2007; Dimitrov, 2012).

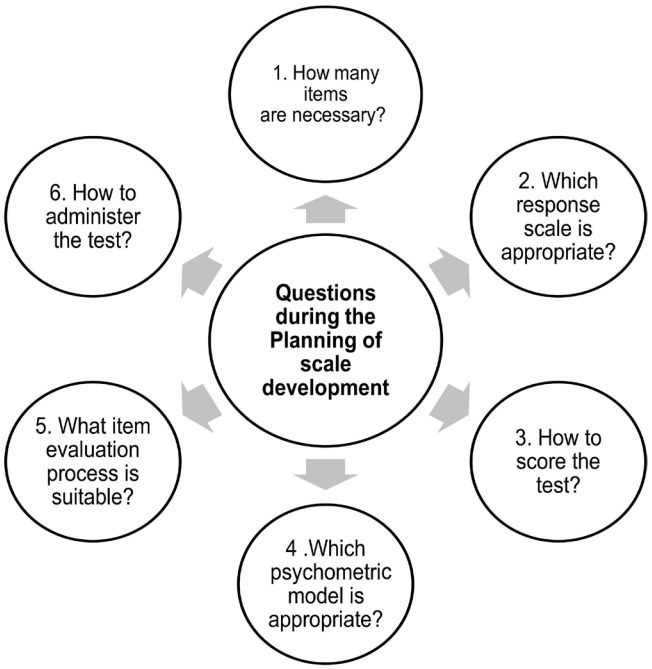


Figure 1. Questions to be answered during planning of the scale development (content by Irwing & Hughes, 2018: pp. 9-10).

used in questionnaires/tests of psychological research are closed-ended because this permits the generated data to be analyzed (Coolican, 2014; Furr, 2011). A third case is a combination of the open and closed-ended format by including an "other" option. This strategy, however, has been proven of imitated efficiency because respondents tend to ignore the other option (Krosnick & Presser, 2010; Lindzey & Guest, 1951; Schuman & Scott, 1987). Scaling in closed-ended items can be categorized as 1) categorical or continuous; 2) by their level of measurement, i.e. nominal, ordinal, interval and ratio (Streiner et al., 2015). In a *categorical scale* score is obtained by summing (or averaging) items receiving answers with binary values (i.e. 1 = true, 0 = false). In a *continuous scale,* the scores are summed (or averaged) based on items with numbers assigned to response categories, i.e. from 1 = *strongly disagree* to 5 = *strongly agree* for a five-point Likert scale item (Dimitrov, 2012; Barker, Pistrang, & Elliott, 2016). Regardless of ambiguities and disagreements, researchers generally treat Likert-type scales as an interval level of measurement (Furr, 2011). However, rating scales rated on a ≥ 5-point scale, are not considered an interval-level measurement but continuous (Streiner et al., 2015). The developer should decide what the response format will be on an early stage, simultaneously with the item generation so that these two have compatibility (DeVellis, 2017). Response scales come in different formats with several specifications to be considered by the developer (see **Figure 2**).

4.1. Response Scale Format

Roughly speaking, the response scale format denotes the way items are worded and responses are obtained and evaluated (Furr, 2011). Common scale formats

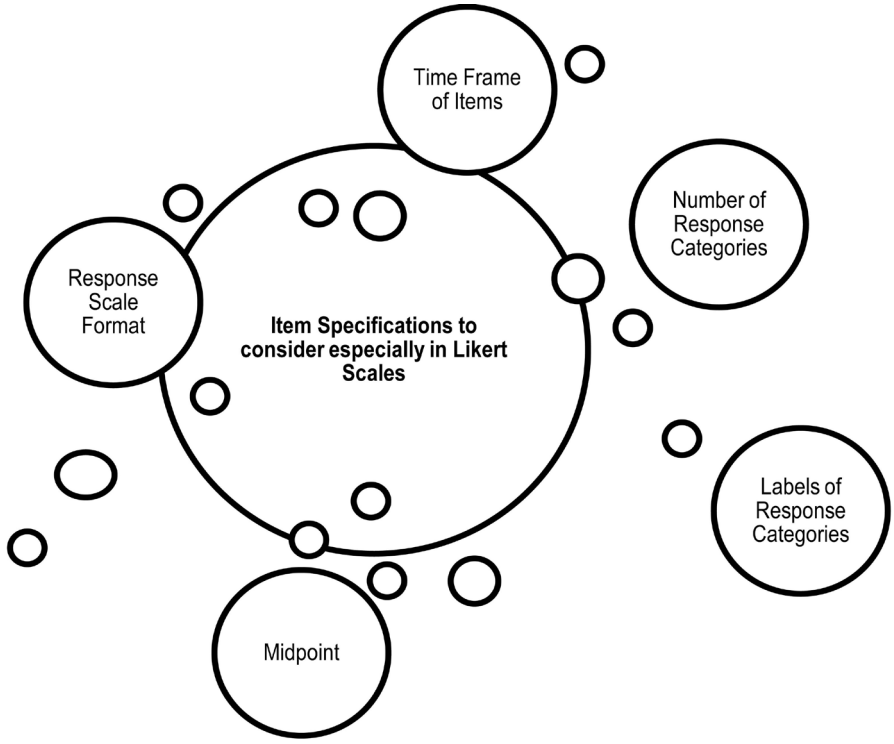


Figure 2. Item Specifications especially pertinent in Likert and Likert-type scales that should be decided along with item writing.

include (Nunnally & Bernstein, 1994; Dimitrov, 2012; Barker et al., 2016): (a) Guttman Scaling (Guttman, 1941, 1944, 1946); (b) Thurstone Scaling (Thurstone, 1928); (c) Likert Scaling (Likert, 1932, 1952). (A) and (B) are not equally weighted item scales while (c) is (DeVellis, 2017). The Classical Measurement Model is more suitable for scales with items being approximately equivalent sensors of the measured construct, like Likert (see also Price, 2017). Generally, scales made up of items that are scored on a continuum and then summed to generate the scale score are more compatible with the Classical Measurement Model (of latent variable measurement) postulating that items are comparable indicators of the underlying construct than with the Item Response Theory that is an alternative measurement perspective (DeVellis, 2017; Price, 2017) and cases(A) and (B) are more suited (DeVellis, 2017). For this reason, we only briefly describe Guttman and Thurstone Scaling and in more detail the Likert Scaling or generally all continuous and equally weighted scales (DeVellis, 2017) of direct estimation (Streiner et al., 2015).

Guttman Scaling

This is a comparative method (Streiner et al., 2015). A Guttman scaling (Guttman, 1941, 1944, 1946; Aiken, 2002) that consists of items tapping increasingly higher levels of an attribute (also called scalogram analysis, deterministic scaling, or cumulative scaling; Dimitrov, 2012). A respondent should select a group of items until the amount of the attribute measured exceeds the one possessed by the respondent. At that point, no other item by the group should be

selected. Purely descriptive data works well with a Guttman scale, e.g. Do you drink?—"Do you drink more than 2 glasses a day?" etc. A respondent's attribute level is showed by the highest affirmative response. Guttman scaling has rather limited applicability with disadvantages that often outweigh the advantages because the assumption of equally strong causal relationships between the latent variable and each of the items would not apply to Guttman scale items. Nunnally and Bernstein (1994) suggest conceptual models for this scale (DeVellis, 2017; Streiner et al., 2015). In practice, response patterns describing a perfect Guttman scale are rare (Price, 2017). See **Table 3** for an example.

Thurstone Scaling

Thurstone (1927) proposed three methods for developing a unidimensional scale: the method of *equal-appearing intervals*, the method of *successive intervals*, and the method of *paired comparisons* (Dimitrov, 2012). The central idea in all three methods is that the scale developer devises items that correspond to different levels of the measured attribute (DeVellis, 2017). Then a group of experts rates the degree the items are representative of the attribute on a scale of 1 (*least representative*) to 11 = *most representative* (Dimitrov, 2012). However, as a rule, the practical problems inherent in using the method with the Classical Measurement Model (DeVellis, 2017), its demanding development process in combination with comparable results to the Likert scale (Streiner et al., 2015) often minimizes its advantages.

Likert Scaling

The Likert Scaling—or Likert normative scale (Saville & MacIver, 2017)—developed by Likert (1932, 1952)—is perhaps the most common response format in psychology (Furr, 2011; Dimitrov, 2012; Barker et al., 2016) and it is versatile and effective for discriminating levels of ability or achievement (Haladyna, 2004;

Table 3. Popular Scaling formats.

<table>
<tr><td>Guttman Scaling</td><td colspan="8">I am able to (select one):
1. Run 200 meters
2. Run 400 meters
3. Run 600 meters
4. Run one kilometer</td><td>True__
True__
True__
True__</td><td>False__
False__
False__
False__</td></tr>
<tr><td>Thurstone Scaling</td><td colspan="8">1. Success is for me a prerequisite for happiness
2. Getting a good job is important but not necessary
3. Happiness has nothing to do with material or work achieves
4. Achieving success gets in the way of being happy</td><td>Agree__
Agree__
Agree__
Agree__</td><td>Disagree __
Disagree __
Disagree __
Disagree __</td></tr>
<tr><td rowspan="3">Semantic Differential</td><td colspan="10">Video games are:</td></tr>
<tr><td>Easy</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Hard</td></tr>
<tr><td>Good</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Bad</td></tr>
<tr><td rowspan="3">Visual Analogue</td><td colspan="10">How severe was your headache the last 24 hours?</td></tr>
<tr><td rowspan="2">No pain</td><td colspan="7" rowspan="2"></td><td colspan="2" rowspan="2">Most severe headache ever experienced</td></tr>
<tr></tr>
</table>

Price, 2017). It contains two parts: (1) the item and (2) a response scale containing a set of alternatives of growing intensity indicated by an integer numerical value and verbal descriptors called anchors (Barker et al., 2016). Each response is rated with a particular integer value (e.g., 1 = Strongly Disagree; 5 = Strongly Agree), summed or averaged across all items of a scale dimension (Furr, 2011). Examples are presented in **Table 4**.

Ratings shown on **Table 4** are mapped onto a bipolar continuum of equal points ranging from strongly approving the statement to strongly disproving. The response options should be worded to have equal intervals with respect to agreement/disagreement forming a continuum (DeVellis, 2017). A neutral point on the scale offers the "middle of the road" response option (Price, 2017). An efficient Likert item could rate opinions, attitudes, beliefs in clear terms but it is more compatible with strongly worded statements because mild items elicit general agreement (DeVellis, 2017). Although it enables direct comparison between people it has received some criticism because of abstract quantification of measurement levels (Saville & MacIver, 2017). Another variation of ordered categorical scale like the Likert is the behavior rating scale. For example, a student's classroom behavior with an item like "Student misbehaves in class" is rated as Always = 5 Never = 1 (Price, 2017, example adapted from Price).

The Likert rating scales and the summated rating scales do not follow a measurement model (Torgerson, 1958) however, the following assumptions are made: 1) category intervals have approximately equal length, 2) category labels are subjectively set, and 3) a pretest phase during item development is followed by an item analysis of the responses (Price, 2017). It is not necessary to span the range of weak to strong assertions in this type of scale because the response options offer the possibility of gradations of the measured construct (DeVellis, 2017).

Just as the form of the question can influence the response, so can the form of the response scale (Barker et al., 2016; Saris & Gallhofer, 2007; Schwartz, 1999). Other response scales alternatives to the Likert-type are briefly the presented in **Table 5**.

Semantic Differential

The semantic differential scale (Osgood & Tannenbaum, 1955; Osgood, Tannenbaum, & Suci, 1957) yields ratings on a bipolar scale with opposite adjective

Table 4. Likert Scales with 5 and 7 points.

I have so much in life to be thankful for	Positive	I am searching for meaning in my life
1 = strongly disagree 2 = disagree 3 = slightly disagree 4 = neutral 5 = slightly agree 6 = agree 7 = strongly agree	1 = Very rarely or never 2 = Rarely 3 = Sometimes 4 = Often 5 = Very often or always	1 = Absolutely Untrue 2 = Mostly Untrue 3 = Somewhat Untrue 4 = Can't Say True or False 5 = Somewhat True 6 = Mostly True 7 = Absolutely True
The Gratitude Questionnaire-Six Item Form (GQ-6) by (McCullough, Emmons, & Tsang, 2002)	Scale of Positive and Negative Experience (SPANE) by (Diener et al., 2009, 2010)	Meaning in Life Questionnaire (MLQ) by Steger et al. (2006)

Table 5. Different rating scales formats.

Adjectival rating scale	How much role should the school principal have in deciding if twins will attend separate class?									
	☐ No role at all		☐ A minor role			☐ A major role			☐ Be the sole deciders	
	Cheating in university entrance exams ifyou have the chance is:									
Summated rating scale	1 Always justified	2	3	4	5	6	7	8	9	10 Never justified
Juster rating scale (Hoek & Gendall, 1993 quoted by Streiner et al., 2015)	10	Certain, practically certain					99 in 100 chance			
	9	Almost sure					9 in 10 chance			
	8	Very probable					8 in 10 chance			
	7	Probable					7 in 10 chance			
	6	Good possibility					6 in 10 chance			
	5	Fairly good possibility					5 in 10 chance			
	4	Fair possibility					4 in 10 chance			
	3	Some possibility					3 in 10 chance			
	2	Slight possibility					2 in 10 chance			
	1	Very slight possibility					1 in 10 chance			
	0	No chance, almost no chance					1 in 100 chance			
Harter rating scale (Harter, 1982 quoted in Streiner et al., 2015)	Really true for me ☐	Sort of true for me ☐	Some kids like doing homework		But	Others dislike doing homework		Sort of true for me ☐	Really true for me ☐	

pairs on each end (Heise, 1970; Price, 2017; DeVellis, 2017). Response values are aggregated across all adjective pairs to calculate the participant's score (Furr, 2017). See **Table 3** for an example.

Visual Analog

The Visual Analog Scale (VAS; Hayes & Patterson, 1921) is marked by a straight line with labels at both ends representing the boundaries of the target construct (Dimitrov, 2012). The line has a fixed length of usually 100 mm (Streiner et al., 2015). Like the Likert scale, the semantic differential and the Visual Analogue response formats can be highly compatible with the theoretical model of Classical Measurement (Latent variable; DeVellis, 2017). This scaling is widely in medicine to assess e.g. pain (Huskisson, 1974), mood (Aitken, 1969), or functional capacity (Scott & Huskisson, 1978), Streiner et al. (2015) comments. See **Table 3** for an example.

4.2. Response Formatting Considerations

There are many considerations in constructing response scales (Barker et al., 2016). The first consideration is the number of response categories and their

labels, whether to offer a midpoint or a "no opinion" option and other details like the time frame (Dimitrov, 2012; DeVellis, 2017; Price, 2017; Barker et al., 2016; Furr, 2011). These considerations are especially relevant to the Likert scale—by far the most commonly used (Furr, 2011; Dimitrov, 2012; Barker et al., 2016).

Number of Response Options

The minimum required is two, i.e. in binary scales (e.g., Agree/Disagree, True/False), but a larger number has benefits and costs (Furr, 2011). Likert (1932, 1952) scales most often uses 5 points; semantic differential (Osgood, Suci, & Tannenbaum, 1957) 7 points, and Thurstone's (1928) 11 points (Krosnick & Presser, 2010). Other sources suggest 5 points for unipolar and 7 points for bipolar as optimal scale length (Fabrigar & Ebel-Lam, 2007). Five to nine points are suited for most occasions and in any case (Streiner at al., 2015; Krosnick & Presser, 2010) and are the most frequently used (Furr, 2011). However, there are really no standards (Krosnick & Presser, 2010: p. 268). Binary item scoring is mostly used in settings where nonresponse is not a possible option, or/and it is treated as incorrect (Dorans, 2018) otherwise may result in information loss and (Streiner et al., 2015) and may be unappealing to respondents (Streiner et al., 2015; also quoting Jones, 1968; Carp, 1989).

A potential benefit is that a relatively large number of options allows for finer gradations (Furr, 2011), just like increasing the accuracy of a microscope. If a response scale is unable to discriminate differences in the target construct, its utility will be limited (DeVellis, 2017). Additionally, reliability is lower for scales with only two or three points in comparison to scales with more points, this reliability increase disappears after 7 points (Krosnick & Presser, 2010 also quoting Lissitz & Green, 1975; Jenkins & Taber, 1977; Martin, 1978; Srinivasan & Basu, 1989) and the same is generally true for validity (Krosnick & Presser, 2010; Green & Rao, 1970; Lehmann & Hulbert, 1972; Lissitz & Green, 1975; Martin, 1973, 1978; Ramsay, 1973).

The potential cost of having many response options is the increase in random error, rather than the systematic portion of the increase in the target construct (Furr, 2011; DeVellis, 2017). Another issue to consider is the respondents' capability to discriminate meaningfully among multiple options. Sometimes too many options cause respondents to use only options that are multiples of 5 or 10 (DeVellis, 2017). Finally, empirical some evidence showed that people in many tasks cannot discriminate easily beyond seven points (Streiner at al., 2015 also quoting Miller, 1956; Hawthorne et al., 2006).

Labels of response options (anchoring)

The descriptors most often tap agreement (Strongly agree to Strongly disagree), but it is possible to construct a Likert scale can be constructed to measure almost any attribute, like agreement (Strongly agree to Strongly disagree), acceptance (Most agreeable - Least agreeable), similarity (Most like me - Least like me), or probability e.g. Most likely - Least likely (Streiner et al., 2015).

Generally, empirical research deems the use of fully-labeled response options

more effective i.e., labeling generate measures with better psychometric quality than does labeling only the endpoints (Krosnick et al., 2005; Furr, 2011; Fabrigar & Ebel-Lam, 2007; Streiner et al., 2015) or every other point and the endpoints (Streiner at al., 2015). More specifically, respondents seem to be more influenced by the adjectives on the scale ends than those located in-between. They also tend to be more satisfied when all of the scale points are labeled (Streiner et al., 2015; Dickinson & Zellinger 1980) and tend to choose them more often than non-labeled points (Streiner et al., 2015).

However, when labeling several practical matters need to be considered. First, labels should differentiate meaningfully the levels of measurement offered. Additionally, they should represent psychologically-equal differences among the response options, as much as possible (DeVellis, 2017; Furr, 2011). The third consideration is the ranking of the response options should be meaningful for all items, logical and consistent (Furr, 2011).

Mid-points

A neutral midpoint can also be added to dichotomous/bipolar rating scales selecting an even point number of response options (Furr, 2011), e.g., a strong positive vs. a strong negative attitude. This can be accomplished by specifying an odd number of points, allowing equivocation ("neither agree nor disagree") or uncertainty ("not sure"). In a unipolar scale, the odd or even number of points issue is probably of little consequence (Streiner et al., 2015). Common choices for a midpoint include "neither agree nor disagree", "agree and disagree equally" (DeVellis, 2017), "neutral" (Furr, 2011; Streiner et al., 2015), or "undecided" (Price, 2017).

Krosnick and Schuman (1988) and Bishop (1990) suggested that those with less intense attitudes or with limited interest were more prone to select mid-points (O'Muircheartaigh et al., 1999; Krosnick & Presser, 2010). O'Muircheartaigh et al. (1999) also noticed that adding midpoints the reliability and validity of ratings were improved. Also, Structural Equation Modeling on error structures showed that the omission of a middle point resulted in the random selection of one of the closer (and moderate) scale point alternative. This suggests that offering a midpoint choice is probably more appropriate than excluding it (Krosnick & Presser, 2010). However, a "Don't know" response option has been empirically proven inefficient (even when offered separately from a mid-point) (Krosnick et al., 2005; Furr, 2011).

However, dependent on the target construct, there may be reasons to exclude equivocation if respondents most likely will use the midpoint choice to avoid answering (Fabrigar & Ebel-Lam, 2007; DeVellis, 2017). There is no criterion other than the needs of the particular research (Streiner et al., 2015). Empirical analysis of mid-points responses suggests that considering mid-point responses as being the halfway between two opposite ends of the target construct compromises the psychometric properties of the scale (Furr, 2011 also quoting O'Muircheartaigh et al., 2000).

5. Phase C: Item Generation (Item Pool)

Along with specifying the response format, a parallel step in developing a questionnaire is assembling and/or devising items for the initial pool (DeVellis, 2017; Furr, 2011). The content specification of an instrument requires that the developer: 1) operationalizes the construct by specifying an exhaustive list of potential indicators (items) of the target construct, 2) select from this list the representative sample of indicators (Dimitrov, 2012). This is perhaps one of the most important steps of the process (Price, 2017), since no subsequent statistical operation could counterbalance poorly stated or absent items (Streiner et al., 2015).

Number of items to include

The initial item pool is larger than the final scale set. As a rule, it can be 3 or 4 times larger (DeVellis, 2017; Streiner et al., 2015), or if the construct is rather narrow 2 times larger (DeVellis, 2017). Writing more good items than required permits selection of the best items, i.e. those which best estimate the target construct and that work well with other items in the scale based on research (Saville & MacIver, 2017). Content redundancy is an asset during the pool construction because it boosts internal-consistency reliability which, in turn, supports validity (Devellis, 2017).

Sources of potential items

The first source of information is to examine what others have done (Furr, 2011; Streiner et al., 2015; Wechsler (1958), for example, incorporated into his IQ tests 11 subtests (see also Taylor, 1953; Hathaway & McKinley, 1951 for similar strategies). There are a number of reasons for item adaption from previous instruments. First, it saves work. Second, existing items have usually proven to be psychometrically sound and third, as a rule, there are not unlimited ways to ask about a specific problem (Streiner et al., 2015). Additionally, when writing items there are five different potential sources of ideas (Streiner et al., 2015): a) the target population (focus group), b) theory, c) existing research, d) expert opinion and/or key informant interviews and e) clinical observation, if applicable. These item sources are not mutually exclusive and a scale developer may use items generated from some or all of these sources (Streiner et al., 2015). Focus groups are a group of carefully selected people (six to twelve, Willms & Johnson, 1993; p. 61) talking freely and spontaneously about the target construct in the presence of a facilitator (Streiner et al., 2015; Willms & Johnson, 1993). Usually, two or three groups suffice. Conditions that make focus groups ineffective is when the target population is difficult to interact publicly (i.e. because of a certain phobia) or because the construct taps embarrassing behaviors or perceived inadequacies (Streiner et al., 2015). Theory on the other hand (broadly defined), may include both formal models or vaguely formed ideas of behaviors, especially if the construct belongs to a relatively narrow domain. Additionally, research findings can be a rich source of potential items and subscales either through a literature review of existing studies in the area or an ad hoc research. However, when the construct taps a new area, previous research may be unavailable. Next,

the expert opinion practice has no rules on how many experts to use, how to choose them, or how differences among their views can be reconciled. Key informant interviews are interviews with a small number of people who are chosen because of their unique knowledge. Generally, the less that is known about the area under study, the less structured is the interview. There is no set number of people who should be interviewed. Clinical observation is perhaps one of the most fruitful sources of items for scales targeting a clinical population (Streiner et al., 2015). The information collected from the above procedures (e.g. expert review) should be used for supporting the content aspect of construct validity (Dimitrov, 2012; Streiner et al., 2015; DeVellis, 2017).

Item Wording

The item wording is important because the way a question is phrased can determine the response (Sudman & Bradburn, 1982; Bradburn et al., 2004; Saris & Gallhofer, 2007; Schwartz, 1999). During item-writing, issues such as language clarity, content relevancy, and the use of balanced scales (i.e. with items worded both positively and negatively) are usually considered (Furr, 2011). Balancing a scale means to word some (e.g. half of them; see BRS by Smith et al., 2008) items positively and other negatively towards the target construct to minimize the response set effect, that is series of similar responses (Anastasi, 1982; Likert, 1932; Cronbach, 1950). However, research generally suggests that is inefficient (Streiner et al., 2015; DeVellis, 2017).

The following suggestions were made for item construction of attitude scales (Gable & Wolfe, 1993: pp. 40-60; reproduced by Price, 2017: p. 178): 1) Avoid items in the past tense; 2) Constructing items that include a single thought; 3) Avoid double-negatives; 4) Prefer items with simple sentence structure; 5) Avoid words denoting absoluteness such as only or just, always, none; 6) Avoid items likely to be endorsed by everyone; 7) Avoid items with multiple interpretations; 8) Use simple and clear language; 9) Keep items under 20 words. This means approximating the reading ability of a child aged 11 - 13 years, a reading level used by most newspapers (DeVellis, 2017; Streiner et al., 2015). Specifically, the reading ability of children attending fifth-grade is 14 words and 18 syllables per sentence, i.e., an item (based on continuous text research (Dale & Chall, 1948; Fry, 1977; DeVellis, 2017; Streiner et al., 2015), thus questionable (see Streiner et al., 2015). Sentences sixth-grade level children can handle contain 15 - 16 words and about 20 syllables. A general rule for efficient implementation of reading ability rules is common sense (DeVellis, 2017), and the same is true for the item writing rules (Krosnick & Presser, 2010).

Generally, the personalized wording is more involving and is preferable by most developers. However, this may not be an asset in a sensitive context. Finally, the tense used in all items should be consistent pointing to a clear time frame (Irwing & Hughes, 2018). Moreover, whether or not positively and negatively worded items are both included in the pool must be considered. Anyhow, the grammar rules must be followed. This will help avoid some ambiguity often

emerging from a pool of items containing both positively and negatively worded items (Devellis, 2017) since scholars are in debate on this issue. To include or not filler items is also another consideration (see DeVellis, 2017 for details). See a summary of key principles of writing good items in **Figure 3** and some examples of unsuccessfully worded items in **Table 6**.

6. Phase D: Item Evaluation

The item generation phase is completed when an expert panel reviews the item pool (DeVellis, 2017). The items generated are reviewed for quality and relevance by the expert panel (Morrison & Embretson, 2018) or /and by pilot testing (Price, 2017). Generally, after reviewing items by expert groups it is also a common practice to pilot test items to acquire data for a first item analysis (Irwing & Hughes, 2018 also quoting DeMaio & Landreth, 2004; Presser & Blair, 1994; Willis, Schechter, & Whitaker, 2000). Alternatively, four additional methods can be used to provide feedback on the relevance, clarity, and unambiguousness: Field pretests, cognitive interviews, randomized experiments and focus groups

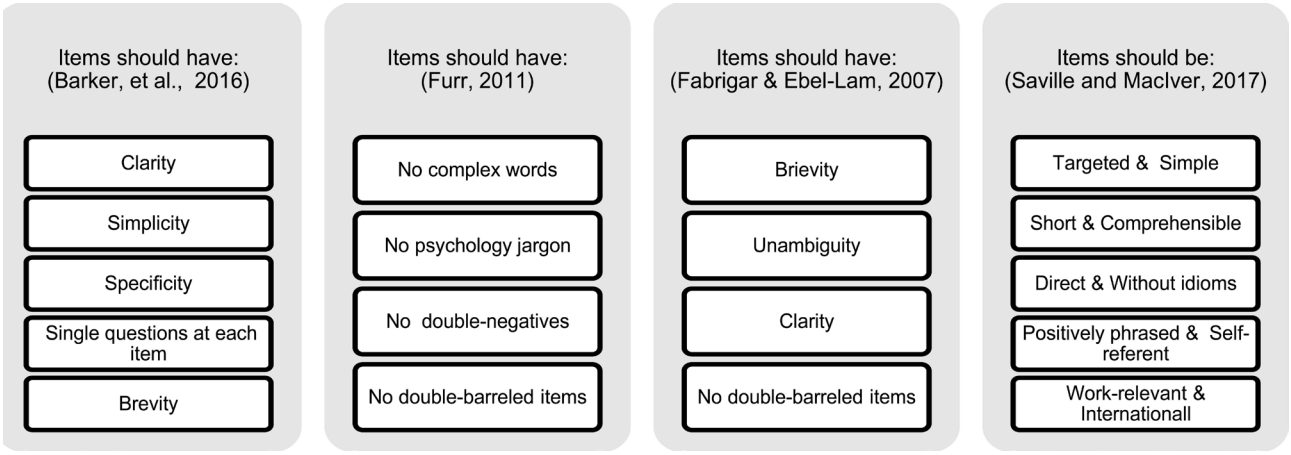


Figure 3. Key principles for successful item writing as suggested by four different sources in scale development literature.

Table 6. Some examples of unsuccessfully item wording.

Item	Problem
Don't you think that smoking should be banned in public buildings?	Leading question—it favors a yes answer
How often do you refer to a psychologist?	Implicit assumption—it assumes the respondent referred to a psychologist
How often did you break down and burst into tears?	Non neutrality—"Break down" gives a negative undertone to crying
Do you ever suffer from back pains?	Ambiguous and unclear—Does not specify the problem and the time frame
Are you satisfied with your job or there were some problems?	Double barreled question (asks two different things at the same time)
Did you notice any motor conversion symptoms over the last 4 *weeks*?	Complicated—Uses professional jargon
It is true that one of the things I seem to have a problem with is making a point when discussing with other people	Lack of brevity/economy—"I often have difficulty in making a point" conveys the same meaning in fewer words

Content adapted by Barker et al., 2016: pp. 111-112; DeVellis, 2017: p. 101.

(Irwing & Hughes, 2018; Streiner et al., 2015). The item validity is complemented by item analysis to estimate the psychometric quality of each item in measuring the target construct (e.g., Ackerman, 1992; Allen & Yen, 1979; Anastasi & Urbina, 1997; Clauser, 2000; Crocker & Algina, 1986; Haladyna, 1999; Janda, 1998; Wilson, 2005; Wright & Masters, 1982 as quoted by Dimitrov, 2012). Item analysis results from support construct validity (Streiner et al., 2015).

Expert Panel Review of Items

Expert reviews may include: 1) content reviews, which provide input about the initial pool of items regarding their relevance to the content domain, accuracy, and completeness; 2) sensitivity reviews, evaluating potential item bias; and 3) standard setting, a process in which experts identify cutoff scores for criterion-referenced decisions on levels of performance or diagnostic classifications (Dimitrov, 2012).

The review serves multiple purposes related to maximizing the content validity. The review process is especially useful when developing an instrument comprising separate scales to measure multiple constructs. The procedure generally involves rating the relevance of each item to the construct according to a definition provided. The definition can be can also confirm or not. Reviewers can also judge the clarity and conciseness of each item. The expert reviewers can also judge the completeness of the content. The developer can accept or reject the experts' advice because content experts might not be familiar with the scale construction principles (DeVellis, 2017). Criteria for items to discarded are summarized in **Table 7**.

A more sophisticated guide to select the most valuable items is to use the content validity ratio (CVR) (Lawshe, 1975; Waltz & Bausell, 1981; Lynn, 1986). Each expert panel member (may contain both scholars and general population), is given a list of the items along with the content dimension they belong. Their job is to evaluate each item on a 4-point scale (4 = Highly Relevant; 3 = Quite Relevant/Highly Relevant but Needs Rewording; 2 = Somewhat Relevant; and 1 = Not Relevant). Then the CVR is calculated using the following formula to evaluate the ratings:

Table 7. Proposed Criteria for retaining and discarding items before or/and after expert reviewing

Highest Interpretability
Lowest Ambiguity
Reject Double-barrelled items (checking two things in one item) like "I feel dizziness and trembling of the hands"
Reject items using Jargon language
Do not mix positively and negatively items
Avoid lengthy items

Content is based on Streiner et al., 2015.

Formula 1: The content validity ratio (CVR)

$$\text{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

Source: Streiner et al. (2015: p. 27) based on the work of Lawshe (1975), Waltz & Bausell, 1981 and Lynn, 1986.

where n_e is the number of raters with a rating of 3 or 4 (i.e. an essential item rating) and N is the total number of raters. The CVR can range from −1 to +1, and a zero value means that half of the panel rated the item as essential. Lawshe (1975) suggested a CVR value of 0.99 for five or six raters (the minimum number), 0.85 for eight raters, and 0.62 for 10 raters. Items with lower values should be rejected (Streiner et al., 2015).

Pilot testing the Items (Pretesting)

So far, the test construction depends on theory, prior empirical evidence, and subjective judgments based on expert knowledge. The next stages include administration to an appropriate sample(s) (Irwing & Hughes, 2018). These are considered probably the quintessence of the scale development process perhaps after the item development (DeVellis, 2017). Pilot testing involves testing the scale to a representative sample from the target population to obtain statistical information on the items, comments, and suggestions (Streiner et al., 2015). Descriptive statistics then will go through item analysis providing important information for each item (Price, 2017). Item analysis is used for selecting the best items. An item analysis allows detection of items that are: 1) ambiguous, 2) incorrectly keyed or scored, 3) too easy or too hard, and 4) not discriminative enough (Price, 2017). This phase generally comprises the following statistical techniques: a) Examine the intercorrelations between all item pairs based both on panel expert ratings and pilot testing; b) Remove items with low correlation with the total score; c) Track the differences between the item means and the 25% of the expert ratings. Items that have higher values are potentially better discriminators of the target construct; and d) Take into account the characteristics of each item and practical considerations retain items with high item-total correlations and high discrimination (Dimitrov, 2012; Trochim, 2006).

Note, however, that some scholars suggest a large development sample of e.g. N = 300 for a 20 item scale after expert review (DeVellis, 2017), while others propose an item review (like panel review) in 1 - 3 small groups. Group sample suggestions vary from N = 100 (Singh et al., 2016) to 6 - 10 (see Streiner et al., 2015) or 20 - 30 (Barker et al., 2016) to evaluate item clarity, reliability, and item characteristics (means and standard deviations) and check dimensionality before large-scale research in order to plan large-scale research better (Muthén & Muthén, 2009; Barker et al., 2016; Singh et al., 2016). This is due to lack of general consensus on all the steps of the scale development process. See the comparison of numerous alternative processes in **Table 1**. Pilot testing is part of an iter-

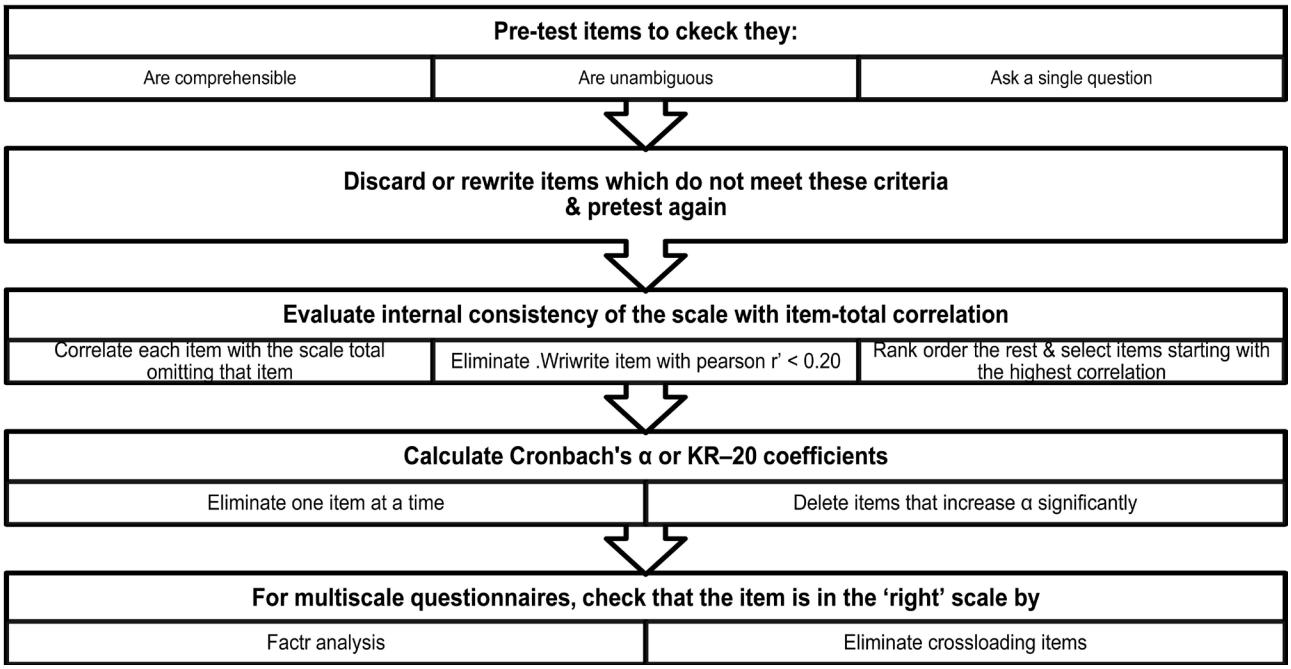
ative process that can be repeated as many times required to ensure desired item properties (Furr, 2011; Price, 2017). The sample size issue is generally part the construct validation sample debating and it is beyond the scope of this work. For details refer to Kyriazos (2018a, 2018b).

Criteria for Item Analysis

Items that are similar insofar as they share relevance to the target construct and not with regards to any other aspect can be good items and not be discarded (DeVellis, 2017). The item quality criterion is a high correlation with the true score of the latent variable. So, the highest intercorrelated items indicated by inspecting the correlation matrix are preferable. If items with negative correlations with other items occur, then reverse scoring may be considered. Items positively correlated with some and negatively correlated with others should be eliminated in a homogeneous set if reverse scoring items do not eliminate negative correlations (DeVellis, 2017). See **Figure 4** for an overview of the pilot testing criteria proposed by Streiner et al. (2015: p. 94). Note also that Item analysis can be carried-out within the SEM context, however this approach is beyond the scopes of this work. Refer to Raykov (2012) for details.

Response Bias

An additional consideration when selecting items is whether items cause response sets which either bias responses or generate response artifacts. Generally, this is mainly attributed to the sequence of items. The most common response sets are: yeah-saying (acquiescence bias—respondents agree with the statements), nay-saying (respondents reject the statements), consistency and availability artifacts, halo (Thorndike, 1920; Campbell & Fiske, 1959: p. 84), and



Content is based on Streiner et al. (2015: p. 94).

Figure 4. Overview of the pilot testing procedure and item analysis procedure.

social desirability artifacts, i.e. respondents try to present themselves in a favorable light Likert scales may also present a central tendency bias—respondents avoid selection of extreme scale categories (Irwing & Hughes, 2018; Dimitrov, 2012).

7. Phase E: Testing the Psychometric Properties of the Scale

In the final phase of the test development process, a validation study is always carried out in a large and representative development sample (DeVellis, 2017) to estimate further the psychometric properties of the scale (Dimitrov, 2012). That is, after an initial pool of items has been developed and pilot tested (pre-tested) in a representative sample, the performance of the individual items to select the most appropriate to include in the final scale and to examine scale dimensionality (DeVellis, 2017). The statistical techniques used for these purposes is item analysis (like during pretesting) and factor analysis (Price, 2017). Criteria for item selection regarding item analysis in this phase are the same as in pretesting (Singh et al., 2016). Dimensionality of a scale is examined with Exploratory Factor Analysis and Confirmatory Factor Analysis (Furr, 2011; Singh et al., 2016). Usually, scales are administered, analyzed, revised, and readministered a number of times before their psychometric properties are acceptable (Irwing & Hughes, 2018; Furr, 2011).

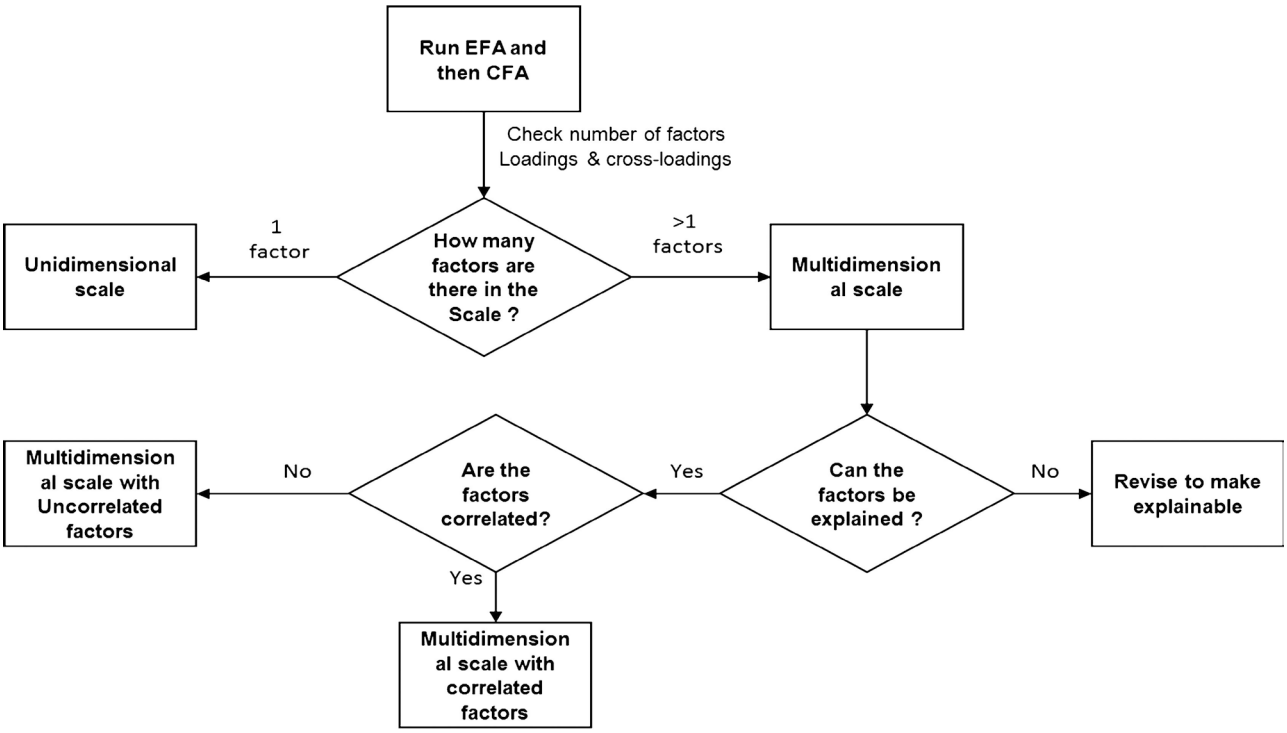
7.1. Dimensionality

A scale's dimensionality, or factor structure, refers to the number and nature of the variables reflected in its items (Furr, 2011). A scale measuring a single construct (e.g. property or ability) is called unidimensional. This means there is a single latent variable (factor) underlies the scale items. In contrast, a scale measuring two or more constructs (latent variables) is multidimensional (Dimitrov, 2012).

Developers examine several issues regarding a scale's dimensionality in this phase of the scale development process. First, they seek to define the number of dimensions underneath the construct. These are called latent variables (factors) and are measured by scale items. A scale is unidimensional when all items tap a single construct (e.g. self-esteem). On the other hand, a scale is multidimensional when scale items tap two or more latent variables, e.g. personality tests (Dimitrov, 2012). If a scale is multidimensional, the developer also examines whether the dimensions are correlated with each other. Finally, in a multidimensional scale, the latent variables must be interpreted according to the theoretical background to see what dimensions they tap, identifying the nature of the construct the dimensions reflect (Furr, 2011) demonstrating construct validity (Streiner et al. (2015) and calculate the reliability of each one. Factor analysis has the answers to dimensionality questions (see **Figure 5**).

7.2. Factor Analysis

"Factor analysis is a statistical technique that provides a rigorous approach for



Source: Adapted by Furr, 2011: p. 26.

Figure 5. The process of dimensionality evaluation of the scale under development and issues related with it.

confirming whether the set of test items comprises a test function in a way that is congruent with the underlying G theory of the test" (Price, 2017: p. 180), based on the classical measurement theory, also termed Classical Test Theory (DeVellis, 2017). Factor analysis is an integral part of scale development. It permits data to be analyzed to determine the number of underlying factors bet heath a group of items called factor so that analytic procedures of the psychometric properties like Cronbach's alpha (Cronbach, 1951) correlations with other constructs can be performed properly. Eventually, through factor identification insights into the latent variable nature underlying the scale items is gained (DeVellis, 2017). A factor is defined as an unobserved or latent variable representative of a construct (Price, 2017: p. 236).

The detailed description of these techniques is beyond the scope of this work but you can refer to Kyriazos (2018a, 2018b) for a complete description of the construct validation process. For scale validation studies refer to Howard et al. (2016), El Akremi, Gond, Swaen, De Roeck, and Igalens (2015), Konrath, Meier, Bushman (2017). Pavot (2018) also suggest reviewing Lyubomirsky and Lepper (1999), Seligson, Huebner, and Valois (2003) and Diener et al. (2010).

7.3. Item Response Theory (IRT)

There is also an alternative to the classical test theory model called Item response theory (IRT). IRT is often presented as a superior alternative to CTT (see De Boeck & Wilson, 2004; Embretson & Reise, 2010; Nering & Ostini, 2010; Reise &

Revicki, 2015 quoted by DeVellis, 2017). IRT is a model-based measurement approach using item response patterns and a person's abilities. In IRT, personal responses to each scale item are explainable based on his or her ability level. The respondent's ability is represented by a monotonically increasing function, based on response patterns (Price, 2017).

According to IRT, several factors affect a person's responses. Along with the person's perceived level of the construct being measured by each scale item, other item properties potentially affecting responses are: (a) item difficulty, (b) item discrimination, and (c) guessing. In most IRT applications in the context of psychology, researchers estimate both psychometric properties at the item level and at the scale level. IRT includes many specific measurement models as a function of different factors potentially affecting individual responses. However, all IRT models are framed according to the probability of a respondent to respond in a specific manner to an item, as a result of a specific level of the underlying behavior. The simplest IRT measurement models comprise only item difficulty while more complex models also comprise two or more item parameters, such as item discrimination and guessing. There are different models for dichotomous items and different for polytomous items (Furr, 2011). IRT models also vary according to the number of item response options.

The effectiveness of a technique is a function of the theoretical framework of the target construct. IRT scoring is used in tests of cognitive ability, however, in other situations, this type of scoring may not be desirable (Irwing & Hughes, 2018). A combination of CTT and TRT was suggested as an alternative option (Embretson & Hershberger, 1999; DeVellis, 2017; Irwing & Hughes, 2018). In most cases a common practice in test development involves a combination either of confirmatory factor analysis (CFA) and IRT (Irwing & Hughes, 2018) or more commonly a combination of EFA and CFA (Steger et al., 2006; Fabrigar & Wegener, 2012; Kyriazos, 2018a).

7.4. Test Scoring and Standardization (Norming)

Raw scale scores can either be based on a unit-weighted sum of item scores or on factor scores. Unit weighted scoring schemas, generate standardized scores using an appropriate standardization sample, or normative sample (Dimitrov, 2012), for example, stanine, sten, and t scores (Smith & Smith, 2005). Unit weighted sums of item scores without standardization may be considered at some research frameworks. Box-Cox procedures (Box & Cox, 1964) to estimate the power to which the scale score should be raised to follow normality. Subsequently, the scale score is also raised to the previously estimated power and standardized. Standardization (or norming) is carried out by subtracting the mean transformed score from the transformed scale scores and dividing by the standard deviation of the transformed scores (Irwing & Hughes, 2018). A standardized score denotes the relative position of each respondent in the target population (Dimitrov, 2012).

Streiner et al. (2015) note the following: (A) Variable weighting on scale items is effective only under certain conditions. (B) if a test is constructed for local/limited use only the sum of the items is probably sufficient. To enable comparison of the results with other instruments, scores is suggested to transformed into percentiles, into z-scores or T-scores. (C) For measurement of attributes that are not the same in males and females, or for attributes that show development changes then separate age and/or age-sex norms can be considered (Streiner et al., 2015).

8. Summary & Conclusions

Experts suggest that effective measurement is the cornerstone of scientific research (DeVellis, 2017; Netemeyer, Bearden, & Sharma, 2003) and it is an integral part of the latent variable model (Slavec & Novsek, 2012). Generally, there are attitude, trait, and ability measures. The purpose of scaling is to construct a scale with specific measurement characteristics for the construct measured. The most commonly employed response formats in all psychology are the Likert type, multiple choice, or forced-choice items. Scaling generally is divided into the types established by Thurstone (1927, 1928), Likert (1932, 1952), or Guttman (1941, 1944, 1946). In Likert scaling the response levels are anchored with consecutive integer values, each corresponding to verbal labels indicating approximately evenly spaced intervals and it is the most popular scale in measures of psychology (Dimitrov, 2012; Furr, 2011, Barker et al., 2016). To a degree, the scaling type and the response format, have an impact on item writing and on the scale development as a whole (Irwing & Hughes, 2018). An item pool should be as rich as possible for the developing scale. It should contain numerous items pertinent to the target construct (DeVellis, 2017). Steps of an instrument development process involves the following: 1) the definition of instrument purpose, domain and construct; 2) defining the response scale format; 3) item generation to construct an item pool 2 - 4 times larger than the desired length of the final scale version; 4) item selection based on expert panel reviews and/or pretesting to maximize instrument reliability with item analysis; 5) large-scale validation study(s) to establish construct validity with supplementary item analysis, factor analysis and to standardize the scale scores.

Construct validation studies to evaluate scale dimensionality and norming is a necessary step in scale development after the pool is examined by experts and/or pretesting. The reliability of measurements signifies the degree to which a score shows accuracy, consistency, and replicability. Construct validity is mainly evidenced by the correlational and measurement consistency of the target construct and its items (indicators) mainly by carving out a factor analysis (Dimitrov, 2012). Scales which are developed thoughtfully and precisely have a greater potential of growing into questionnaires that measure real-world criteria more accurately (Saville & MacIver, 2017).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Ackerman, T. A. (1992). A Didactic Explanation of Item Bias, Item Impact, and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement, 29,* 67-91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x

Aiken, L. R. (2002). *Attitudes and Related Psychosocial Constructs: Theories, Assessment and Research*. Thousand Oaks, CA: Sage.

Aitken, R. C. B. (1969). A Growing Edge of Measurement of Feelings. *Proceedings of the Royal Society of Medicine, 62,* 989-92.

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes, 50,* 179-211. https://doi.org/10.1016/0749-5978(91)90020-T

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME) (1999). *Standards for Educational and Psychological Testing* (2nd ed.). Washington DC: Authors.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA/APA/NCME) (2014). *Standards for Educational and Psychological Testing* (3rd ed.). Washington DC: Authors.

Anastasi, A. (1982). *Psychological Testing* (5th ed.). Macmillan, New York.

Anastasi, A., & Urbina, S. (1996). *Psychological Testing* (7th ed.). New York, NY: Pearson.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. New York, NY: Freeman.

Barker, C., Pistrang, N., & Elliott, R. (2016). *Research Methods in Clinical Psychology: An Introduction for Students and Practitioners* (3rd ed.). Oxford, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119154082

Bishop, G. F. (1990). Issue Involvement and Response Effects in Public Opinion Surveys. *Public Opinion Quarterly, 54,* 209-218. https://doi.org/10.1086/269198

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26,* 211-254.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design—For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco, CA: Jossey-Bass.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56,* 81-105. https://doi.org/10.1037/h0046016

Carp, F. M. (1989). Maximizing Data Quality in Community Studies of Older People. In M. P. Lawton, & A. R. Herzog (Eds.), *Special Research Methods for Gerontology* (pp. 93-122). Amityville, NY: Baywood Publishing.

Chadha, N. K. (2009). *Applied Psychometry*. New Delhi, IN: Sage Publications.

Clauser, B. E. (2000). Recurrent Issues and Recent Advances in Scoring Performance Assessments. *Applied Psychological Measurement, 24,* 310-324. https://doi.org/10.1177/01466210022031778

Coolican, H. (2014). *Research Methods and Statistics in Psychology* (6th ed.). New York: Psychology Press.

Costa, P. T., & McCrae, R. R. (1995). Domains and Facets: Hierarchical Personality Assessment Using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64,* 21-50. https://doi.org/10.1207/s15327752jpa6401_2

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT Predict College GPA after Removing *g*. *Intelligence, 36,* 719-729. https://doi.org/10.1016/j.intell.2008.05.001

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1950). Further Evidence on Response Sets and Test Design. *Educational and Psychological Measurement, 10,* 3-31. https://doi.org/10.1177/001316445001000101

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika, 16,* 297-334. https://doi.org/10.1007/BF02310555

Dale, F., & Chall, J. E. (1948). A Formula for Predicting Readability: Instructions. *Education Research Bulletin, 27,* 37-54.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer. https://doi.org/10.1007/978-1-4757-3990-9

Demaio, T., & Landreth, A. (2004). Do Different Cognitive Interview Methods Produce Different Results. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Questionnaire Development and Testing Methods*. Hoboken, NJ: Wiley.

DeVellis, R. F. (2017). *Scale Development: Theory and Applications* (4th ed.). Thousand Oaks, CA: Sage.

Dickinson, T. L., & Zellinger, P. M. (1980). A Comparison of the Behaviorally Anchored Rating Mixed Standard Scale Formats. *Journal of Applied Psychology, 65,* 147-154. https://doi.org/10.1037/0021-9010.65.2.147

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49,* 71-75. https://doi.org/10.1207/s15327752jpa4901_13

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S. et al. (2009). New Well-Being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research, 97,* 143-156. https://doi.org/10.1007/s11205-009-9493-y

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New Wellbeing Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. *Social Indicators Research, 97,* 143-156. https://doi.org/10.1007/s11205-009-9493-y

Dimitrov, D. M. (2012). *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields.* Alexandria, VA: American Counseling Association.

Dorans, N. J. (2018). Scores, Scales, and Score Linking. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, V.II* (pp. 573-606). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118489772.ch19

El Akremi, A., Gond, J.-P., Swaen, V., De Roeck, K., & Igalens, J. (2015). How Do Employees Perceive Corporate Responsibility? Development and Validation of a Multidimensional Corporate Stakeholder Responsibility Scale. *Journal of Management, 44,* 619-657. https://doi.org/10.1177/0149206315569311

Embretson, S. E., & Hershberger, S. L. (1999). Summary and Future of Psychometric Models in Testing. In S. E. Embretson, & S. L. Hershberger (Eds.), *The New Rules of Measurement* (pp. 243-254). Mahwah, NJ: Lawrence Erlbaum. https://doi.org/10.4324/9781410603593

Embretson, S. E., & Reise, S. P. (2010). *Item Response Theory* (2nd ed.). New York, NY: Routledge Academic.

Fabrigar, L. R., & Ebel-Lam, A. (2007). Questionnaires. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 808-812). Thousand Oaks, CA: Sage.

Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory Factor Analysis.* New York, NY: Oxford University Press, Inc.

Fredrickson, B. L. (1998). Cultivated Emotions: Parental Socialization of Positive Emotions and Self-Conscious Emotions. *Psychological Inquiry, 9,* 279-281. https://doi.org/10.1207/s15327965pli0904_4

Fredrickson, B. L. (2001). The Role of Positive Emotions in Positive Psychology: The Broaden-and-Build Theory of Positive Emotions. *American Psychologist, 56,* 218-226. https://doi.org/10.1037/0003-066X.56.3.218

Fredrickson, B. L. (2003). The Value of Positive Emotions: The Emerging Science of Positive Psychology Is Coming to Understand Why It's Good to Feel Good. *American Scientist, 91,* 330-335.

Fredrickson, B. L. (2013). Positive Emotions Broaden and Build. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 1-53). Cambridge, MA: Academic Press.

Fry, E. (1977). Fry's Readability Graph: Clarifications, Validity, and Extension to Level 17. *Journal of Reading, 21,* 249.

Furr, R. M. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. New Delhi, IN: Sage Publications. https://doi.org/10.4135/9781446287866

Gable, R. K., & Wolfe, M. B. (1993). *Instrument Development in the Affective Domain: Measuring Attitudes and Values in Corporate and School Settings* (2nd ed.). Boston, MA: Kluwer. https://doi.org/10.1007/978-94-011-1400-4

Green, P. E., & Rao, V. R. (1970). Rating Scales and Information Recovery—How Many Scales and Response Categories to Use? *Journal of Marketing, 34,* 33-39. https://doi.org/10.2307/1249817

Guttman, L. (1941). The Quantification of a Class of Attributes: A Theory and Method for Scale Construction. In P. Horst (Ed.), *The Prediction of Personal Adjustment* (pp. 321-348). New York: Social Science Research Council.

Guttman, L. (1946). An Approach for Quantifying Paired Comparisons and Rank Order. *Annals of Mathematical Statistics, 17,* 144-163. https://doi.org/10.1214/aoms/1177730977

Guttman, L. A. (1944). A Basis for Scaling Qualitative Data. *American Sociological Review, 9,* 139-150. https://doi.org/10.2307/2086306

Haladyna, T. M. (1999). *Developing and Validating Multiple-choice Items* (2nd ed.). Mahwah, NJ: Erlbaum.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items*. Mahwah, NJ: Erlbaum.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development, 53,* 87-97. https://doi.org/10.2307/1129640

Hathaway, S. R., & McKinley, J. C. (1951). *Manual for the Minnesota Multiphasic Personality Inventory* (Rev. ed.). New York: Psychological Corporation.

Hawthorne, G., Mouthaan, J., Forbes, D., & Novaco, R. W. (2006). Response Categories and Anger Measurement: Do Fewer Categories Result in Poorer Measurement? Development of the DAR5. *Social Psychiatry and Psychiatric Epidemiology, 41,* 164-172. https://doi.org/10.1007/s00127-005-0986-y

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental Development of the Graphic Rating Method. *Psychological Bulletin, 18,* 98-99.

Heise, D. R. (1970). Chapter 14. The Semantic Differential and Attitude Research. In G. F. Summers (Ed.), *Attitude Measurement* (pp. 235-253). Chicago, IL: Rand McNally.

Hoek, J. A., & Gendall, P. J. (1993). A New Method of Predicting Voting Behavior. *International Journal of Market Research, 35,* 1-14. https://doi.org/10.1177/147078539303500405

Howard, J., Gagné, M., Morin, A. J., & Van den Broeck, A. (2016). Motivation Profiles at Work: A Self-Determination Theory Approach. *Journal of Vocational Behavior, 95-96,* 74-89. https://doi.org/10.1016/j.jvb.2016.07.004

Huskisson, E. C. (1974). Measurement of Pain. *The Lancet, 304,* 1127-1131. https://doi.org/10.1016/S0140-6736(74)90884-8

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, V.I* (pp. 4-47). Hoboken, NJ: Wiley.

Janda, L. H. (1998). *Psychological Testing: Theory and Applications*. Boston, MA: Allyn & Bacon.

Jenkins, G. D., & Taber, T. D. (1977). A Monte Carlo Study of Factors Affecting Three Indices of Composite Scale Reliability. *Journal of Applied Psychology, 62,* 392-398. https://doi.org/10.1037/0021-9010.62.4.392

Jones, R. R. (1968). Differences in Response Consistency and Subjects' Preferences for Three Personality Response Formats. In *Proceedings of the 76th Annual Convention of the American Psychological Association* (pp. 247-248) Washington DC.

Kline, R. B. (2009). *Becoming a Behavioral Science Researcher: A Guide to Producing Research That Matters*. New York: Guilford Publications.

Konrath, S., Meier, B. P., & Bushman, B. J. (2017). Development and Validation of the Single Item Trait Empathy Scale (SITES). *Journal of Research in Personality, 73,* 111-122. https://doi.org/10.1016/j.jrp.2017.11.009

Krosnick, J. A., & Presser, A. (2010). Question and Questionnaire Design. In P. V. Marsden, & J. D. Wright (Eds.), *Handbook of Survey Research* (2nd ed., pp. 264-313). Bingley, UK: Emerald.

Krosnick, J. A., & Schuman, H. (1988). Attitude Intensity, Importance, and Certainty and Susceptibility to Response Effects. *Journal of Personality and Social Psychology, 54,* 940-952. https://doi.org/10.1037/0022-3514.54.6.940

Kyriazos, T. A. (2018a). Applied Psychometrics: The 3-Faced Construct Validation Method, a Routine for Evaluating a Factor Structure. *Psychology, 9,* 2044-2072. https://doi.org/10.4236/psych.2018.98117

Kyriazos, T. A. (2018b). Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General. *Psychology, 9,* 2207-2230.

https://doi.org/10.4236/psych.2018.98126

Lawshe, C. H. (1975). A Quantitative Approach to Content Validity. *Personnel Psychology, 28,* 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lehmann, D. R., & Hulbert, J. (1972). Are Three-Point Scales Always Good Enough? *Journal of Marketing Research, 9,* 444-446. https://doi.org/10.2307/3149313

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology, 140,* 1-55.

Likert, R. A. (1952). A Technique for the Development of Attitude Scales. *Educational and Psychological Measurement, 12,* 313-315.

Lindzey, G. G., & Guest, L. (1951). To Repeat—Check Lists Can Be Dangerous. *Public Opinion Quarterly, 15,* 355-358. https://doi.org/10.1086/266315

Lissitz, R. W., & Green, S. B. (1975). Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach. *Journal of Applied Psychology, 60,* 10-13. https://doi.org/10.1037/h0076268

Lynn, M. R. (1986). Determination and Quantification of Content Validity. *Nursing Research, 35,* 382-386. https://doi.org/10.1097/00006199-198611000-00017

Lyubomirsky, S., & Lepper, H. S. (1999). A Measure of Subjective Happiness: Preliminary Reliability and Construct Validation. *Social Indicators Research, 46,* 137-155. https://doi.org/10.1023/A:1006824100041

Martin, W. S. (1973). The Effects of Scaling on the Correlation Coefficient: A Test of Validity. *Journal of Marketing Research, 10,* 316-318. https://doi.org/10.2307/3149702

Martin, W. S. (1978). Effects of Scaling on the Correlation Coefficient: Additional Considerations. *Journal of Marketing Research, 15,* 304-308. https://doi.org/10.2307/3151268

McCullough, M. E., Emmons, R. A., & Tsang, J. (2002). The Grateful Disposition: A Conceptual and Empirical Topography. *Journal of Personality and Social Psychology, 82,* 112-127. https://doi.org/10.1037/0022-3514.82.1.112

Milfont, T. L., & Fischer, R. (2010). Testing Measurement Invariance across Groups: Applications in Cross-Cultural Research. *International Journal of Psychological Research, 3,* 111-121. https://doi.org/10.21500/20112084.857

Miller, G. A. (1956). The Magic Number Seven plus or minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Bulletin, 63,* 81-97.

Morrison, K. M., & Embretson, S. (2018). Item Generation. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development, V.I* (pp. 46-96). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118489772.ch3

Muthén, L. K., & Muthén, B. O. (2009). *Mplus Short Courses, Topic 1: Exploratory Factor Analysis, Confirmatory Factor Analysis, and Structural Equation Modeling for Continuous Outcomes*. Los Angeles, CA: Muthén & Muthén.

Nering, M. L., & Ostini, R. (2010). *Handbook of Polytomous Item Response Theory Models*. New York: Routledge.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling Procedures: Issues and Applications.* Thousand Oaks, CA: Sage Publications.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw-Hill.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (1999). Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data. *Paper presented at the American Asso-*

ciation for Public Opinion Research Annual Meeting, St. Petersburg, FL.

O'Muircheartaigh, C., Krosnick, J. A., & Helic, A. (2000). *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. http://harrisschool.uchicago.edu/About/publications/working-papers/pdf/wp_01_3.pdf

Osgood, C. E., & Tannenbaum, P. H. (1955). The Principle of Congruence in the Prediction of Attitude Change. *Psychological Bulletin, 62,* 42-55.

Osgood, C. E., Tannenbaum, P. H., & Suci, G. J. (1957). *The Measurement of Meaning*. Urbana, IL: University of Illinois Press.

Pavot, W. (2018). The Cornerstone of Research on Subjective Well-Being: Valid Assessment Methodology. In E. Diener, S. Oishi, & L. Tay (Eds.), *Handbook of Well-Being*. Salt Lake City, UT: DEF Publishers.

Presser, S., & Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? In P. Marsden (Ed.), *Sociology Methodology* (Vol. 24, pp. 73-104). Washington DC: American Sociological Association.

Price, L. R. (2017). *Psychometric Methods: Theory into Practice*. New York: The Guilford Press

Prochaska, J. O., Norcross, J. C., Fowler, J., Follick, M. J., & Abrams, D. B. (1992). Attendance and Outcome in a Worksite Weight Control Program: Processes and Stages of Change as Process and Predictor Variables. *Addictive Behaviors, 17,* 35-45. https://doi.org/10.1016/0306-4603(92)90051-V

Ramsay, J. O. (1973). The Effect of Number Categories in Rating Scales on Precision of Estimation of Scale Values. *Psychometrika, 38,* 513-532. https://doi.org/10.1007/BF02291492

Raykov, T. (2012). Scale Construction and Development Using Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling* (pp. 472-492). New York: Guilford Press.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, Evaluation, and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470165195

Saville, P., & MacIver, R. (2017). A Very Good Question? In B. Cripps (Ed.), *Psychometric Testing: Critical Perspectives* (pp. 29-42). West Sussex, UK: John Wiley & Sons, Ltd.

Sawilowsky, S. S. (2007). Construct Validity. In N. J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (pp. 178-180). Thousand Oaks, CA: Sage.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside.

Schuman, H., & Scott, J. (1987). Problems in the Use of Survey Questions to Measure Public Opinion. *Science, 236,* 957-959. https://doi.org/10.1126/science.236.4804.957

Schwartz, N. (1999). Self-Reports: How the Questions Shape the Answers. *American Psychologist, 54,* 93-105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarzer, R. (2001). Social-Cognitive Factors in Changing Health-Related Behavior. *Current Directions in Psychological Science, 10,* 47-51. https://doi.org/10.1111/1467-8721.00112

Scott, P. J., & Huskisson, E. C. (1978). Measurement of Functional Capacity with Visual Analog Scales. *Rheumatology and Rehabilitation, 16,* 257-259. https://doi.org/10.1093/rheumatology/16.4.257

Seligman, M. E. (1998). What Is the Good Life? *APA Monitor, 29,* 2.

Seligman, M. E., & Csikszentmihalyi, M. (2000). Positive Psychology: An Introduction. *American Psychologist, 55,* 5-14. https://doi.org/10.1037/0003-066X.55.1.5

Seligman, M. E., & Pawelski, J. O. (2003). Positive Psychology: FAQS. *Psychological Inquiry, 14,* 159-163.

Seligson, J. L., Huebner, E. S., & Valois, R. F. (2003). Preliminary Validation of the Brief Multidimensional Students' Life Satisfaction Scale (BMSLSS). *Social Indicators Research, 61,* 121-145. https://doi.org/10.1023/A:1021326822957

Singh, K., Junnarkar, M., & Kaur, J. (2016). *Measures of Positive Psychology: Development and Validation*. Berlin: Springer. https://doi.org/10.1007/978-81-322-3631-3

Slavec, A., & Drnovsek, M. (2012). A Perspective on Scale Development in Entrepreneurship. *Economic and Business Review, 14,* 39-62.

Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The Brief Resilience Scale: Assessing the Ability to Bounce Back. *International Journal of Behavioral Medicine, 15,* 194-200. https://doi.org/10.1080/10705500802222972

Smith, M., & Smith, P. (2005). *Testing People at Work: Competencies in Psychometric Testing*. London: Blackwell.

Srinivasan, V., & Basu, A. K. (1989). The Metric Quality of Ordered Categorical Data. *Marketing Science, 8,* 205-230. https://doi.org/10.1287/mksc.8.3.205

Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The Meaning in Life Questionnaire. Assessing the Presence of and Search for Meaning in Life. *Journal of Counseling Psychology, 53,* 80-93. https://doi.org/10.1037/0022-0167.53.1.80

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health Measurement Scales: A Practical Guide to Their Development and Use* (5th ed.). Oxford, UK: Oxford University Press. https://doi.org/10.1093/med/9780199685219.001.0001

Sudman, S., & Bradburn, N. M. (1982). *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco, CA: Jossey-Bass.

Taylor, J. A. (1953). A Personality Scale of Manifest Anxiety. *Journal of Abnormal and Social Psychology, 48,* 285-290. https://doi.org/10.1037/h0056264

Thorndike, E. L. (1920). A Constant Error in Psychological Ratings. *Journal of Applied Psychology, 4,* 25-29. https://doi.org/10.1037/h0071663

Thurstone, L. L. (1927). Three Psychophysical Laws. *Psychological Review, 34,* 424-432. https://doi.org/10.1037/h0073028

Thurstone, L. L. (1928). Attitudes Can Be Measured. *American Journal of Sociology, 33,* 529-554. https://doi.org/10.1086/214483

Torgerson, W. (1958). *Theory and Methods of Scaling.* New York: Wiley.

Trochim, W. M. (2006). *The Research Methods Knowledge Base* (2nd ed.). http://www.socialresearchmethods.net/kb

Waltz, C. W., & Bausell, R. B. (1981). *Nursing Research: Design, Statistics and Computer Analysis*. Philadelphia, PA: F.A. Davis.

Wechsler, D. (1958). *The Measurement and Appraisal of Adult Intelligence* (4th ed.). Baltimore, MD: Williams and Wilkins. https://doi.org/10.1037/11167-000

Willis, G., Schechter, S., & Whitaker, K. (2000). A Comparison of Cognitive Interviewing, Expert Review and Behavior Coding: What Do They Tell Us? In *Proceedings of the Section on Survey Methods* (pp. 28-37). Alexandria, VA: American Statistical Association.

Willms, D. G., & Johnson, N. A. (1993). *Essentials in Qualitative Research: A Notebook for the Field*. Unpublished Manuscript, Hamilton, ON: McMaster University.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith Jr., E. V. (2007). Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part I Instrument Development Tools. *Journal of Applied Measurement, 8,* 97-123.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press.

Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The Construction and Use of Psychological Tests and Measures. In *Encyclopedia of Life Support Systems*. France: United Nations Educational, Scientific, and Cultural Organization Publishing (UNESCO-EOLSS Publishing).