1. Scale development in marketing research

*Noorjahan Banon Teeluckdharry, Viraiyan
Teeroovengadam and Ashley Seebaluck*

Marketing is mostly what can be termed as an empiricism-driven discipline, whereby survey methods are often used for data collection and regression or structural equation modelling for data analysis (Kock et al., 2016). However, like in most social sciences, the phenomenon on its own cannot be directly observed and has to be assessed by proxy variables (Diamantopoulos et al., 2008). As such, a structured and comprehensive process of developing measurement models is of utmost importance since poorly conceived processes are likely to lead to highly inaccurate results.

Scale development practices mainly follow established theoretical and methodological guidelines (e.g., Churchill, 1979; DeVellis, 2016), but the emergence of new methods and software offers opportunities to develop and employ new approaches. This chapter provides an overview of the best practices for developing and validating scales. The aim is to provide a primer for scale development in a way that facilitates the development of new, valid and reliable scales, and to help improve existing ones. DeVellis (2016) is by far one of the most cited references when it comes to scale development, and here we have incorporated the guidelines suggested therein, among others. However, we found the flowchart in Boateng et al. (2018) to be more structured and comprehensive as well as more straightforward. The latter have identified three phases that span into nine steps. Therefore, building on their work (Figure 1.1), Phase 1 is about the item and domain development, whereby items are generated and their validity is assessed by the various stakeholders. Phase 2 showcases the steps in the scale construction and includes steps such as pre-testing questions, sampling and survey administration, item reduction and factor extraction. In Phase 3 we consider the test of dimensionality, reliability and validity.

Source: Adapted from Boateng et al. (2018).

Figure 1.1 Framework elaborating steps for scale development

PHASE 1: ITEM DEVELOPMENT

Step 1: Identification of the Domain(s) and Item Generation

Domain identification

Having a well-defined domain serves several purposes; but, most importantly, it provides us with the knowledge to specify the boundaries of the phenomenon under study, and to smooth out the process of item generation and content validation (Boateng et al., 2018). The identification of any domain calls for a thorough overview of existing literature. Having a well-formulated definition of the phenomenon to be measured helps avoid unintentional drifting and crossing over into other domains, especially when similar items are present that may tap into quite different constructs. Prior to providing the preliminary conceptual definition, apt focus should be given to reviewing appropriate related theories. In cases where the dimensions of the domain exist a priori, each dimension should be clearly defined. When there is a lack of any substantive theory that can be utilized as a guide to develop the scale, researchers should try to provide a tentative theoretical model while stipulating how the new construct relates to existing phenomena and their operationalization (DeVellis 2016).

When taking on a new scale development, it is important to specify the purpose for which the construct is being developed and whether there are existing measurement scales. If this is the case, justification is to be provided as to why the development of a new scale is being undertaken, and emphasis should be laid on the differences between existing scales and the new proposed one.

Furthermore, it is helpful to clarify the level of scale specificity, i.e., the extent to which the scale aims to measure a phenomenon – whether as a general global measure of a particular phenomenon or as a detailed and specific one. A good example is satisfaction with respect to a particular outcome, as opposed to overall satisfaction. Usually, the relative broadness or narrowness of the scale is context-specific, and will depend on the aim of the research.

Item generation

Item generation can also be referred to as question development. There are basically two ways to view this step: we can either apply a deductive approach, working from the domain to identify items from existing literature and existing scales; or we can use an inductive approach to generate items from the responses recorded during exploratory studies involving methodologies such as focus groups and interviews. However, it is best to combine both approaches into a hybrid one. This is because existing theory is not only a source of accumulated knowledge about the items; it also provides the theoretical underpinning for better pragmatic decisions related to the constructs. On the other hand, the use of a qualitative approach helps transmute occasionally abstract market perspectives into simple manifested forms of the phenomenon.

When it comes to what should be included in the pool of items, selection is rather subjective. For example, DeVellis (2016) advocated caution since items of a homogeneous nature would reflect the latent variable and, as such, they should be sensitive to the true score of the latent variable. They should not venture beyond the boundaries of the defining construct, yet they should exhaust all possibilities within those boundaries to accurately capture the essence of the construct. On the other hand, some authors, like Boateng et al. (2018), favour a more liberal approach and suggest incorporating all the items identified even though at first glance they do not appear to be a perfect fit to the domain, as successive evaluations at later stages will remove all undesirable items that lack clarity, show questionable relevance or display undesirable similarity to other items.

The number of items to be included will usually depend on the scale being developed – whether it is multidimensional as opposed to monodimensional. Since internal consistency reliability is dependent on the

strength of correlations among items, and by extension with the latent variable, the greater the number of items, the better will be internal consistency. However care is advocated to ensure that non-related items are not included since they are likely to induce error variance within the scale, leading to the latter's invalidation. Some authors suggest the initial item pool should be five times as big as the final pool so as to offer enough margin for the optimal grouping of items (Schinka et al., 2012); but the general rule of thumb is that the initial pool of items should be at least double the desired scale (Kline, 1993; Schinka et al., 2012).

As for the forms of the items, the wording and the types of responses being induced should be considered. Items should be straightforward and worded in simple language; they should follow the conventions of normal conversation and be as unambiguous as possible. Lengthy items should be avoided as they are likely to induce complexity and decrease clarity. Unnecessary wordiness should be avoided. However, the meaning of an item should never be sacrificed for the sake of brevity. Furthermore, the reading difficulty level of items should be well thought out. It is best to eliminate potential confusion associated with the use of multiple negatives within a sentence – for example: *It is not unlikely for me not to have a certain bias concerning a particular brand.*

Another important consideration is the avoidance of ambiguous items that convey two or more ideas simultaneously, such that an endorsement of the item might refer to either or both ideas. Items similar to a package deal that renders it impossible to endorse one part without agreeing to the other part should therefore be avoided. Sometimes it may be that the whole sentence describing an item is ambiguous and possibly open to numerous interpretations. This can get very perplexing for some respondents and can also create varying levels of bias within the responses given, and hence should be avoided.

The use of both positively and negatively worded items is rather common in scales. The proponents of such an approach suggest that its application is to reduce various forms of response bias, such as the tendency to select positive responses regardless of the content of the items or to select all high or all low ratings – either due to lack of engagement or confusion on behalf of the respondents. However, opponents are just as numerous. It has been demonstrated in several studies to induce another bias such as the negative items loading on one or more separate factors, that is, the inaccurate perception of factors being related to positively and/or negatively worded items. Henceforth such an approach should be used with caution, if not avoided. As reflected upon by DeVellis (2016), the disadvantages of using both positive and negative wording within a scale outweigh the benefits.

There is an additional aspect to consider when pooling items for the scale: the redundancy of items. It is recommended to have different ways of expressing a similar idea. This allows the most suitable version to be selected in the subsequent stages, thus yielding more reliable item sets. Furthermore, by using multiple and seemingly redundant items, the core common ground will summate across the items while cancelling out irrelevant idiosyncrasies (DeVellis, 2016). Risk of under-representation and narrow focus is also mitigated. However, when several construct-irrelevant statements display similarities in wording, this can cause respondents to react similarly to these items, and thus produces an inflated estimate of reliability when this isn't truly the case.

Lastly, in terms of wording, there should be due ethical consideration such that proposed items are not offensive or discriminatory towards any individual or social entity.

When it comes to questions with dichotomous response categories (e.g., true/false), no ambiguity should be present. When Likert scales are used the responses should be represented in an ordinal, meaningful manner which is clear to the respondents so that they are interpreted in the same way by each participant. Also, scales with 2–3 points tend to display lower reliability than Likert-type response scales with 5–7 points. When the items are reflecting the relative degrees of a single response quality – for example, *not satisfied to very satisfied* – it is advisable to use a five-point Likert scale. On the other hand, when the items reflect the relative degrees of two qualities, such as *completely disagree to completely agree*, a 7-point scale is a better option (Boateng et al., 2018).

Step 2: Content Validity

Content validity was first defined as: “the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his/her responses to a real or hypothetical universe of situations which together constitute the area of concern to the person interpreting the test” (Lennon, 1956, p. 295). Despite the numerous subsequent definitions, it is generally agreed that content validity refers to the extent to which the sampled items adequately reflect the domain and operational definition of the construct. In scale development, this step is known as theoretical analysis, whereby the initial pool of items is checked to see: whether they genuinely reflect the desired construct; whether the scale is measuring what it presumes to measure; and whether related items which are beyond the conceptual boundaries of the construct are being unintentionally added. Content validity is also about capturing the relevant experience of the target population.

According to Guion (1977) there are five conditions of content validity:

1. There is consensus on the behavioural content.
2. There is unambiguity in the construct definition.
3. Content is construct-relevant.
4. Experts agree that domains have been adequately sampled.
5. Response content must be reliably observed and re-evaluated.

Content validity is a precondition to other forms of validity and, as such, must first be checked. Its purpose is to minimize the potential error associated with instrument operationalization in the initial stages and to increase the probability of obtaining supportive construct validity in the later stages (Shrotryia & Dhanda, 2019). Content validity assessment relies on the use of panel experts to evaluate the scale. When done correctly in a systemic and comprehensive manner invaluable information can be gathered (Rubio et al., 2003). Addressing content validity begins with the development of the instrument itself. It involves a two-stage process of instrument development and judgement. We agree with Carr et al. (1985) that the two stages (development and judgement/quantification) are crucial for any developed instrument. Furthermore, a third step of revising and reconstruction/reformation is also needed.

In most cases the opinions sought for content validation are those of expert judges (knowledgeable people in the content area or in the development of scales) or the target population, the lay experts, who are the potential users. Optimally the panel of experts should consist of both content experts and lay experts, although the opinions of expert judges are given more emphasis, especially in cases where resources are limited and only of the group being chosen (Boateng et al., 2018). The number of expert judges used for a study depends on the scale developer. Some studies, such as Shrotryia and Dhanda (2019), recommend a minimum of 3 judges, whereas others (for instance Haynes et al., 1995) recommend around 5 to 7 expert judges; but it is generally agreed that using more than 10 expert judges decreases the chance of consensual agreement (Polit & Beck, 2006).

The experts are likely to confirm or invalidate the definition proposed as well as provide recommendations with respect to grammar; the correct use of words and the position of key words are obtained from the panel of experts (Safikhani et al., 2013). The threats that typically arise during scale development are construct-irrelevant content – that is, the presence of unnecessary elements beyond the core intended concept of the construct – and content under-representation. To cater for construct-irrelevant content, the scale developers should apply both existing theory

and the judgements of content experts to validate excluding or including a certain element while considering the research objective. The same goes for under-representation of certain elements, the presence of which can jeopardize the internal consistency of the construct and interfere with the relationships and correlations between the items under certain factors or domains (Almanasreh et al., 2019).

Content validity procedures usually entail the experts rating how relevant each item is using techniques such as content validity ratio (CVR), content validity index (CVI) or Cohen's kappa coefficient (κ) for measuring inter-rater or expert agreement (Almanasreh et al., 2019). Figure 1.2 depicts a typical flowchart of the steps involved in content validity.

Content validity ratio is computed according to the Lawshe test. It provides confidence in selecting the most important and correct content in an instrument. Basically, CVR is a numeric value which varies from 1 to -1 , and is determined from expert ratings indicating the instrument's degree of validity. The experts are requested to rate the necessity of each

Source: Adapted from Zamanzadeh et al. (2015).

Figure 1.2 Flowchart illustrating typical steps in content validity

item for operating a construct using scores of 1–3, with a three-degree range of “not necessary”, “useful but not essential” and “essential” respectively. The formula for calculating CVR is $(N_e - N / 2) / (N / 2)$, where N_e is the number of panellists indicating “essential” and N is the total number of panellists. One rule of thumb suggests that a CVR of at least 0.78 on a scale is necessary to deem an item or scale as valid (Frey, 2018).

To cater for agreement among experts not due to chance, critical CVR (CVR critical) values have been proposed by Lowell Schipper (Lawshe, 1975). CVR critical is the minimum value of CVR such that the level of agreement exceeds that of chance for a given item, for a given alpha (Type I error probability, one tailed test $P = 0.05$) (Lawshe, 1975). Methods for calculating the critical values of CVR can be found in Ayre and Scally (2014). CVR critical values can be utilized to identify the number of panel experts required to agree on an item as essential and the items which should be retained or discarded from the final instrument.

Alternatively, the Delphi technique (or method) may be used for content validity. This is a distinct type of survey method for soliciting and obtaining group consensus on an important subject (Lindeman, 1975; Notter, 1983), and is described as the technique “for structuring group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem” (Linstone & Turoff, 1975, p. 3). This method is particularly convenient when content is either controversial or intangible, or when one of the expert judges potentially lacks sufficient knowledge or expertise of the subject matter (Tersine & Riggs, 1976). The key points of the Delphi technique are: (1) anonymity of panel members; (2) anonymity of responses; (3) numerous iterations; (4) group consensus; (5) controlled feedback of responses to panel members; and (6) statistical analysis of data (Lindeman, 1975; Notter, 1983). Its goal is to reach a consensus of opinions through a series of structured questionnaires sent to the experts which build upon their responses to previous rounds of questionnaires. The number of rounds is dependent upon several factors, namely: the level of consensus desired by the investigator; the degree of concreteness of items; unexpected findings; and the number of items on the questionnaire (Perroca, 2011). However, there are typically three rounds. The expert panel should preferably comprise stakeholders with relevant experience and facilitators in the field (Scheele, 1975). The size of the panel varies with the research objective; for instance, in Brockhoff’s study of Delphi performance (1975), groups with 11 participants were more accurate in their predictions than larger groups, whereas for fact-finding questions groups with seven participants performed better. On

the other hand, error decreases with larger Delphi panels (Linstone & Turoff, 1975).

Content validity index (CVI)

The CVI is the method used for quantifying the content validity of an instrument, based on expert ratings for each item in terms of its relevance to the underlying construct, generally applying on Likert scales ranging from 1 (not relevant) to 4 (extremely relevant). CVI can be computed at each item level (I-CVI) or for overall scale (S-CVI).

For each item, the I-CVI can be calculated by counting the number of experts who rated the item as 3 or 4 and dividing that number by the total number of experts – that is, the proportion of agreement about the content validity of an item. For example, if 8 out of 10 content experts rated an item as relevant (3 or 4), the CVI would be $8/10 = 0.80$. Zamanzadeh et al. (2015) suggest the following: if $I\text{-CVI} > 0.79$, the item is relevant; between 0.70 and 0.79, the item needs revision; and if the value is below 0.70 the item is eliminated.

Literature reports two methods for computing the overall content validity index. The first is universal agreement (UA-CVI), which is defined as the proportion of items on an instrument that achieved a rating of 3 or 4 (valid) by all the content experts. Not only is this approach rather conservative, but the increase in the number of experts made its achievability rather difficult. The second approach is Average CVI (Ave-CVI), which is defined as the average proportion of items rated as 3 or 4 (valid) across the various experts. There are basically three methods to calculate the Ave-CVI, though all will yield the same results:

1. Calculate the average CVI across the items by adding together the I-CVIs and dividing them by the number of items.
2. Count the total number of ratings multiplied by the number of items rated relevant by all experts combined, and divide by the total number of ratings.
3. Average the proportion of items rated relevant across experts.

Davis (1992) suggested a standard criterion of 0.8 as the lower limit of acceptability of a CVI of an instrument. Polit and Beck (2006) have criticized aspects of the CVI and recommended using Lynn's criteria for calculating the I-CVI ($I\text{-CVI} = 1$ with 3 or 5 experts and a minimum I-CVI of 0.78 for 6–10 experts), and an Ave-CVI of 0.90 or higher to have an excellent content validity of an instrument. It should be noted that Polit et al. (2007) slightly modified the above criteria and recommended perfect agreement only when there are three or four experts.

Table 1.1 Sample ratings of 10 items using a 4-point rating relevance scale (3,4 relevant), (1,2 not relevant)

Item	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Expert 7	No of agreement	Item-CVI*
1	3	4	4	3	3	4	4	7	1
2	3	2	4	3	3	3	4	6	0.72
3	2	2	4	4	3	3	1	4	0.58
4	4	3	2	4	3	3	4	6	0.86
5	4	3	3	3	1	2	4	5	0.72
6	3	4	1	3	3	3	3	6	0.86
7	3	4	4	1	3	3	2	5	0.72
8	3	4	4	3	3	4	4	7	1
9	3	4	3	4	4	3	3	7	1
10	1	3	4	3	3	4	3	6	0.86
AVE-CVI**									0.84
UA-CVI***									0.3
No of agreements	8	8	8	9	9	9	8	59	
Proportion of agreements with respect to total items	0.8	0.8	0.8	0.9	0.9	0.9	0.8	5.9	
Mean expert agreement proportion****									0.85

Notes: *Item-CVI = [number of items having 3 or 4 rating] / [number of experts] = content validity index of an item.

** Ave-CVI = [summation of all item-CVI] / [no. of items] = content validity of scale.

*** UA-CVI = [number of items that achieved rating 3 or 4 by all experts] / [total number of items] = content validity of scale.

**** Mean expert agreement proportion = [agreements with respect to total items by each expert] / [number of experts].

Source: Adapted from Almanasreh et al. (2019).

It is important to know that while CVI is easy to compute and interpret, and provides content validity of each item and the instrument as a whole, it is the call of the scale developer which item(s) to include or exclude, as shown in Table 1.1 (Almanasreh et al., 2019).

Evaluation by lay experts (face validity)

Face validity can be defined as the “degree that respondents or end users or lay persons judge that the items of an assessment instrument are appropriate to the targeted construct and assessment objectives” (Haynes et al., 1995, p. 243n4). If a measure looks, on the face of it, as though it measures a particular attribute, then the existence of face validity can be assumed (McGartland Rubio & Kimberly, 2005). While face validity and content validity may sound similar, they are not. Content validity is established through structured and rigorous procedures; in comparison, face validity is rather subjective and established through cognitive interviews to establish whether the construct is a good measure of the domain. Since the target population are the lay experts for face validity, they can be asked to confirm the researchers’ ideas of “what looks as if” it might be correct, and to make adjustments to the scale as necessary so that the scale is truly crafted for the targeted audience and is context appropriate.

PHASE 2: SCALE DEVELOPMENT

Step 3: Pre-Testing of Questions

The primary purpose of pre-testing is to ensure that items are meaningful to the target population prior to administering the survey so as to minimize misinterpretation and subsequent measurement error. It also eliminates poorly worded items, reduces the cognitive burden on the surveyed population, eases out the rephrasing of statements for higher scale efficiency, and is an overall additional way for the target population to input further insights into the scale being developed (Boateng et al., 2018).

Qualitative methods chiefly identify items having certain wording concerns, such as being hyphenated, leading or confusing (Leech, 2002). While some of these tests are able to pinpoint those items with face validity concerns, they do not provide a direct numerical indicator. Hence, they may be able to remove items with large face validity concerns, but they cannot be used to solely retain the items with the greatest face validity. Howard (2018) provides a guideline of procedures to adopt depending on the main concerns at hand (Figure 1.3).

Source: Adapted from Howard (2018).

Figure 1.3 Flowchart of which procedure to use with regard to face validity and wording issues

Cognitive interviews

Cognitive interviewing incorporates a wide range of techniques for eliciting data about interpretation of the scale and how responses are formulated. Techniques may include verbal probing and thinking aloud (Collins, 2003). Verbal probing is about encouraging respondents to verbalize their interpretation of items and to paraphrase and/or comment on the wording of items in an effort to identify ambiguous or poorly worded questions as well as to comment on the perceived accuracy of their response. The think aloud technique is utilized when the scale developer wants to elicit data on participants' thought processes as they respond to each item, thus providing insights into the kinds of information participants retrieve from memory as they determine their response to an item. By eliciting thought processes associated with responses, investigators gain useful insights into participants' decision-making processes. Cognitive interviews ensure that the questions are producing the intended data, that they are clear to the respondents, that they are easy and simple to interpret, and that the options are appropriate and adequate. It is advocated that the outcomes of the interviews should be reported together with the solutions used to remedy the situation. Though there are some difficulties in determining whether the participants are interpreting items correctly or are completely off the mark, or to what extent a particular participant is struggling to

answer a question, it is up to the developers to know where to draw the line, especially since there is no hard and fast rules to coding participant responses (Howard, 2018). Likewise, researchers may choose whether to apply non-essential coding methods, such as behaviour coding or response latency, whereby items with longer latencies are believed to perform poorly.

Item-rating tasks

This is among the most common quantitative pre-test methods. In this method, participants are asked to evaluate the extent to which the item represents the principal construct, using scales such as “clearly representative”, “somewhat representative” and “not representative”, or “very good”, “good”, “fair” and “poor.” While there is no recommended sample size for item-rating tasks, typical sample sizes range from 10 to 30 (Goetz et al., 2013; Heene et al., 2014). There are three approaches that can be used for item-retention decisions:

1. Each item rating is allocated a number (e.g., very good 4, good 3, fair 2, poor 1), and the highest scoring items are retained.
2. Items that receive a certain percentage of highest or second highest responses (i.e., very good, good) are retained.
3. Items that receive any of the lowest responses (e.g., poor) are discarded.

Item-sort task

At the beginning of any item-sort task, the participants are presented with the items list and a list of constructs to which they have to match those items. Those with frequent assignments to the posited construct are retained. The typical number of participants for this step is 20–30, although Howard and Melloy (2016) demonstrated that sample sizes as small as five can be used. The researchers can choose a cut-off that would result in the desired number of items to be retained or can use the cut-off values provided by Howard and Melloy. Using this latter approach, which is based on traditional statistical significance testing, the results of item-sort tasks have sound statistical justification and have been shown to replicate exploratory factor analysis (EFA) results.

Step 4: Survey Administration and Sample Size

Survey administration

The aim at this stage is to allow for collection of data with the minimum of measurement errors from an adequate sample. We shall explore what

constitutes an adequate sample size shortly, but first let us consider some possible ways that data can be collected. These range from typical paper and pen/pencil interviewing (PAPI) to computer-assisted personal interviewing (CAPI), with each approach having its pros and cons.

For instance, in the case of CAPI, the use of technology can reduce errors associated with data entry, permit data collection from a larger sample at lower cost, enhance response rates, reduce enumerator errors, and allow for instant feedback along with providing space to get more confidential sensitive information (Goldstein et al., 2012; Fanning & McAuley, 2014; Dray et al., 2016). The fact that CAPI allows for the inclusion of photos and videos makes the survey even more engaging for the respondents, and at the same time allows researchers to convey their questions to the target audience with greater clarity. In audio computer self-assisted interviewing (A-CASI), questions may be recorded and read out loud to participants with low literacy levels. Furthermore, self-interviewing using CAPI allows participants to respond to the questions without the interviewer's involvement, which may give the responder a sense of comfort and privacy to answer sensitive questions with greater honesty and transparency. This is particularly true when participants are anonymized, allowing for less social desirability (Gosling et al., 2004) and fewer ethical problems (Nosek et al., 2002). While there may be a few start-up costs with the use of CAPI – such as the purchase of the software, along with more time and resources needing to be spent on configuring the questionnaire in the software and programming the consistency checks – there is a growing body of evidence supporting the idea that CAPI is a cheaper and faster form of data collection than PAPI. For example, Leisher (2014, p. 264) found that “the cost per completed interview for the tablet-based survey was 74% less than the paper-based survey average.” Also, the electronic data capture approach eliminates the need for post-interview data entry, and thus allows for immediate data retrieval. Hence the main benefits of the CAPI method can be summarized as the prevention of error in the data collected through its in-built automated routing, consistency checks and use of images for uniformity of data interpretation. This in turn leads to cost savings for the organizations involved and reduces the time between fieldwork and publication of final results.

As for the traditional PAPI method, the advantages lie in averting crisis in case data is lost: software crashes, devices can be lost or stolen; and it may be more suitable in areas with irregular electricity supplies and internet connections. Still, as the sample size increases, so the risk of human error increases as interviewing becomes more labour- and time-intensive (Fanning & McAuley, 2014; Greenlaw & Brown-Welty, 2009). Based on the merits of CAPI over PAPI, we recommend that researchers use CAPI in data collection for surveys when feasible.

Establishing sample size

The sample size to be used for scale development is often contentious, and the adequate size varies from study to study. However, the rule of thumb dictates at least ten participants for each scale item (Nunnally, 1978), though Hair et al. (2019b) recently suggested that a ratio of five respondents per item scale may be adequate. They also suggested that, in addition to the rule of thumb for calculating the minimum size of the sample population, the final sample size should be at least 20 per cent larger than the calculated one so as to cater for missing responses or any other issues related to data collection. Others have suggested sample sizes that are independent of the number of survey items. For example, Comrey (1988) recommended a range of 200–300 as appropriate for factor analysis, while Guadagnoli and Velicer (1988) suggested a minimum of 300–450 since replication might be needed for a sample below 300. Moreover, item reduction techniques such as bootstrapping may require larger data sets; but, generally, with larger samples sizes measurement errors are lower, factor loadings are more stable and the results can be better generalized to be applicable to the targeted population (Osborne & Costello, 2004). Furthermore, statistical issues play a central role when calculating the required sample size, which is often based on the confidence level (95 per cent), margin of error (3–5 per cent) and expected variability. Since the methodological rigour of the study and the generalizability of the results are heavily dependent on a sample size which is representative of the targeted population, there should be proper planning and execution for sampling design (Teeroovengadam & Nunkoo, 2018).

Determining the type of data to use

Normally scale development would require data from a single point in time; but to fully test the reliability of a scale, as will be demonstrated in subsequent steps, an independent data set is required. Data from longitudinal studies may be used for initial scale development (e.g., initial data), to conduct confirmatory factor analysis (follow-up data) and to assess test–retest reliability (using baseline and follow-up data). However, the use of longitudinal data may lead to common error variance because the same participants are involved (Boateng et al., 2018). Hence, ideally the scale development should be developed on sample A, whether it is longitudinal or cross-sectional data collection, and then use an independent sample B for confirmatory factor analysis (CFA). For example, Hill et al. (2019) conducted two studies to develop a scale to measure the components for meaning in life. In Study 1, they conducted an exploratory factor analysis of the items and assessed the test–retest reliability of the new measure,

called the Meaning in Life Measure (MILM). In Study 2, they tested the replicability of the factor structure of the MILM with a new sample using CFA and tested the MILM's validity and reliability.

Step 5: Factor Reduction

The purpose of item reduction is to ensure that only parsimonious, functional and internally consistent items are ultimately included (Thurstone, 1947). It is about identifying items that are the least related to the domain under study for deletion or modification. Functional items can be described as those that are correlated with each other, discriminate between individual cases, underscore a single or multidimensional domain and contribute significantly to the construct (Boateng et al., 2018).

The measurement theories that underpin scale development are the Classical Test Theory (CTT) and the Item Response Theory (IRT) (Fan, 1998; DeVellis, 2016). Each theory can be used on its own or in conjunction with another (Glockner-Rist & Hoijtink, 2003). The basic idea underlying CTT is that an observed score is simply the result of the respondent's true score plus error, which is differentiated into different categories. By contrast, IRT differentiates errors with respect to characteristics of individual items. Therefore, with IRT methods the researcher is able to examine the item information and standard error functions for the item pool and determine the effect of adding or deleting a given item (Harvey & Hammer, 1999). One of the objectives of IRT is to establish certain characteristics of items that are independent of the respondent assessing the items, and the emphasis is placed on items as opposed to the scale as a whole. For guidelines on how to perform IRT, these sources are recommended: Alaya (2009) and Baker and Kim (2017).

Depending on which test theory is driving the scale, several techniques are used to reduce the item pool, including:

- item difficulty and item discrimination indices (primarily for binary responses); and
- inter-item and item-total correlations (categorical items).

Item difficulty index

The item difficulty index pertains to both CTT and IRT, and assesses the relative difficulties of test items. It is important to note that these difficulties are defined in terms of the likelihood of correct response, not in terms of perceived difficulty or amount of effort required. Knowing the item difficulty is useful when building a scale to match the levels of a target population. For example, an instrument aimed at measuring wellbeing within an

organization should not be constituted solely of items that resonate only with those who have clinical depression. When used under the CTT framework, the item difficulty index refers to the proportion of participants who answer an item correctly: the higher the score, the greater the proportion of the sample who understood and answered the question correctly. Low scores would suggest a need to modify items or delete them from the pool of items. When it comes to the IRT framework, the item difficulty index refers to the probability of a particular test taker correctly answering any given item. This enables identification of levels of individual performance on specific questions, as well as allowing for the selection of questions specific to a segment of the targeted population (Hambleton & Jones, 1993). Depending on the type of scale being developed, the researcher must determine whether they need items with low, medium, or high difficulty. For instance, researchers interested in general-purpose scales will focus on items with medium difficulty (0.4–0.6) (Raykov, 2011).

Item discrimination index

The item discrimination index, also called the item-effectiveness test, is a measure of how well an item is able to distinguish between participants who are familiar with the context being studied and those who are not. It basically measures the difference in performance between groups. On an attitude scale, it would discriminate between those with positive attitudes and those with negative attitudes.

In CTT, the item discrimination index can be computed through correlational analysis between the performance on an item and an overall criterion using either the point biserial correlation coefficient or the phi coefficient (Sim & Rasiah, 2006). This index considers the relationship between an examinee's performance on the given item, whether it is correct or incorrect, and the examinee's score on the overall test. Under the IRT framework, item discrimination is a slope parameter that determines how sharply the probability of a correct response changes as the trait being investigated increases (DeMars, 2010).

For an item that is highly discriminating, in general the examinees who responded to the item correctly also did well on the test, while in general the examinees who responded to the item incorrectly also tended to do poorly on the overall test. While the range of the discrimination index is within -1.0 to 1.0 , an item with discrimination below 0.0 would suggest a problem. Likewise, an item discriminating negatively may indicate that the item is measuring something other than what the rest of the test is measuring. Moreover, when interpreting the value of a discrimination it is important to be aware that there is a relationship between an item's difficulty index and its discrimination index. If an item has a very high (or very

low) item difficulty index, the potential value of the discrimination index will be much less than if the item has a mid-range p-value.

Inter-item and item-total correlations

Inter-item correlations consider the degree to which the scores on one item are related to the scores on all other items in a scale. They provide an assessment of item redundancy by examining the extent to which items on a scale are assessing the same content (Cohen & Swerdlik, 2005; Raykov, 2011). When the inter-item correlations are too low, this indicates that the items may not be representative of the same domain. When the value is too high, this may suggest that the bandwidth of the construct captured is too narrow or that each of the items is not contributing something unique to the construct. According to Piedmont (2014), the average inter-item correlation for a set of items should be between 0.20 and 0.40. The reason stipulated is that, while being reasonably homogenous, the items do contain sufficiently unique variance, and thus are not isomorphic. On the other hand, Clark and Watson (2016) endorsed a mean inter-item correlation range of 0.15 to 0.20 for scales that measure broad characteristics and 0.40 to 0.50 for those tapping narrower ones.

Item-total correlations examine the relationship between each item and the total score of scale items. There are two types of item-scale correlation – corrected and uncorrected. For example, if there are ten items, in the corrected version the item being assessed is correlated with all the remaining nine other items excluding itself, while in the uncorrected correlation the item under consideration is correlated with all ten items. Since the uncorrected version is likely to inflate the correlation coefficient, the use of the corrected version is advisable (DeVellis, 2016). The higher the correlation coefficient, the more clearly an item belongs to the scale. As a rule of thumb, if it is less than 0.3 then the item is dropped from the scale. However various cut-off points are adopted. For example, Cristobal et al. (2007) recommend 0.30; Loiacono et al. (2002) and Ladhari (2010) suggest 0.40; and Francis and White (2002) and Kim and Stoel (2004) 0.50. Calculating an item-to-scale coefficient for each of the initial items and dropping the inappropriate ones should result in a unidimensional scale.

Examine the quality of data

For the findings to be accurate and reliable, missing data, absence of outliers, linearity and extreme multicollinearity should be given due importance by researchers (Beavers et al., 2013). When it comes to missing data, patterns should be inspected. Usually, it is advisable to consider deleting cases when the majority of responses (50 per cent or more) contain missing data (De Vaus, 2002). There are several ways to deal with missing data,

including ad hoc procedures such as listwise deletion and pairwise deletion (Marsh, 1998; Roth, 1994), but these do not have any theoretical justification. We also have theory-based maximum likelihood (ML) approaches and similar response pattern imputation. While going into detail about these specific approaches is beyond the scope of this chapter, there are numerous manuals about data-cleaning best practices (e.g., Hair et al., 2019a; Meyers et al., 2016) that may be referred to on how to deal with missing data, outliers, or other potential issues.

DeVellis (2016) also suggests examining the item performance in terms of reversing scoring if needed, item variance and item means. High difference within the item variances may indicate different sources of error among the items. Furthermore, a mean close to the mid-range of possible scores is also desirable since items with means too near to an extreme of the response range will have low variances, and those that vary over a narrow range will correlate poorly with other items.

Step 6: Extraction of Factors

Factor analysis examines whether, based on the patterns in the responses collected in a survey, we can summarize the variables into fewer, more general factors. A simple example is provided in Table 1.2, whereby the variables are grouped under different factors.

According to De Vaus (2002) there are four main steps in forming scales using factor analysis, namely: selecting the variables to be analysed; extracting an initial set of factors, extracting a final set of factors by “rotation”; and constructing scales based on the results at step 3 and using these in further analysis.

Selecting the variables prior to factor analysis

One of the issues of factor analysis is that, regardless of the variables used, a set of “underlying” factors will be produced, whether they make sense or not. When selecting variables to be factor analysed, it is important to be able to assume that correlations between the variables will not be caused by other variables in the analysis. It is also necessary to ensure that the

Table 1.2 Example of variables grouped under different factors

Factor 1: Location	Factor 2: Value for money	Factor 3: Efficiency
X ₁ : Convenient location	X ₃ : Low prices	X ₅ : Ease of parking
X ₂ : Near home	X ₄ : Attractive promotions	X ₆ : Great customer service
		X ₇ : Ease of locating items

variables to be analysed have at least reasonable correlations with some other variables. Hence the factorability of the data is verified through inspection of the correlation matrix, Bartlett's test of sphericity and the Kaiser-Meyer-Olkin (KMO) measure. The correlation matrix should include numbers at levels of 0.30 or higher; Bartlett's chi-square should be significant at a probability of 0.05 or less; and a KMO value of 0.60 or higher is recommended before proceeding with factor analysis (Pett et al., 2006; Tabachnick et al., 2007, pp. 481–498).

Initial set of factors

Extraction of factors is usually done using exploratory factor analysis (EFA), whereby latent factors that represent the shared variance in responses among the multiple items are extracted. Prior to extracting factors there are two decisions to consider: the method of extraction and the number of factors to be extracted. Based on a number of statistical theories, there are many factor-extraction methods, but the two most common are principal component(s) analysis (PCA) and common factor analysis (FA). The purpose of PCA is to reduce the number of items while retaining as much of the original item variance as possible, while that of FA is to understand the latent factors or constructs that account for the shared variance among items – and thus is more appropriate for scale development (Worthington & Whittaker, 2006). Furthermore, error variance included in PCA tends to inflate the size of components, causing researchers to retain too many components (Goldberg & Velicer, 2006). On the other hand, common factor analysis results are more generalizable for confirmatory factor analysis (Haig, 2005; Worthington & Whittaker, 2006), as illustrated in Box 1.1.

BOX 1.1

An example of the use of FA versus PCA in a simulated data set might illustrate the differences between these two approaches. Imagine that a researcher at a public university is interested in measuring campus climate for diversity. The researcher created 12 items to measure three different aspects of campus climate (each using four items): (a) general comfort or safety; (b) openness to diversity; and (c) perceptions of the learning environment. In a sample of 500 respondents, correlations among the 12 variables indicated that one item from each subset did not correlate with any other items on the scale (e.g., no higher than $r = 0.12$ for any bivariate pair containing these items). In FA, the three uncorrelated

items appropriately drop out of the solution because of low factor loadings (loadings < 0.23), resulting in a three-factor solution (each factor retaining three items). In PCA, the three uncorrelated items load together on a fourth factor (loadings > 0.45). This example demonstrates that, under certain conditions, PCA may overestimate factor loadings and result in erroneous decisions about the number of factors or items to retain.

Source: Worthington & Whittaker (2006).

In scale development, the methods to determine the number of factors to retain include a scree plot (Cattell, 1966), the variance explained by the factor model, and the pattern of factor loadings. The eigenvalue is a measure indicating the amount of variance in the entire set of items accounted for by a given factor. Kaiser (1958) believed that eigenvalues of less than 1.0 reflect potentially unstable factors, while the scree test is considered more accurate than the eigenvalue rule (Pett et al., 2006; Preacher & MacCallum, 2003; Reise et al., 2007), whereby the number of factors to be retained is determined by when a line elbows off from a somewhat subjectively straight dotted line. The amount of the total variance in the original set of variables that is explained by the factors is an indication of good factor analysis. Since to increase the total variance explained we have to increase the number of factors, it becomes important to retain stable factors with an eigenvalue of 1 or higher used (De Vaus, 2002). The number of factors to be retained can also be computed using methods such as parallel analysis (Horn, 1965), minimum average partial procedure (Velicer, 1976) or the Hull method (Lorenzo-Seva et al., 2011).

The rotation method

Rotation is necessary in order to more clearly identify the scale's factors (or dimensions) since the unrotated results from a factor analysis are not easy to interpret. The term "rotation" is used because, historically and conceptually, the axes are being rotated so that the clusters of items fall as closely as possible to them. There are two broad categories of rotation, namely oblique and orthogonal, which refer to the angle maintained between the X and Y axes. Therefore, orthogonal rotations produce factors that are uncorrelated (i.e., maintain a 90-degree angle between axes), while oblique methods allow the factors to correlate (i.e., allow the X and Y axes to assume an angle other than 90 degrees). However, in social sciences like marketing, it is very common for factors to correlate with one another (DeVellis, 2016). Hence, if factors are substantially correlated, an oblique

rotation provides a better and more accurate solution. Oblique rotation options include Direct Oblimin and Promax. Though both allow factors to correlate, Promax begins with an orthogonal solution and then transforms it to an oblique solution (Hendrickson & White, 1964), and it has been argued to be more robust (Thompson, 2004). If factors do not correlate, it is advisable to assess whether two separate constructs are being measured rather than one.

Retention or deletion of items

It is suggested that items for retention should be evaluated based on theory, communalities, item loadings, no significant cross-loadings, minimum of three salient loadings, factor reliability levels and parsimony (Carpenter, 2018).

One example of how theory can affect the item to retain is to consider “price”. Often when service quality scale is being conceptualized customers are likely to mention price and value as dimensions of the service quality, though the consensus in existing marketing literature is that price is often considered as a separate construct such as perceived value (Dabholkar et al., 1996; Brady & Cronin, 2001; Martínez & Martínez, 2010). Henceforth items to be retained should be backed by theory.

While there is a fixed level for what degree of communality qualifies a variable as having sufficient common variation to retain in a factor solution, there are various suggestions for the communalities cut-off. For example, Osborne et al. (2008) suggest that communality above 0.4 is acceptable. Items loading highly on more than one factor should be removed (Hair et al., 2010). As for factor item, there are various suggested cut-off points, which normally range from 0.3 to 0.4 or above (Worthington & Whittaker, 2006; Hair et al., 2010).

When it comes to the number of items per subscale, many methodologists would recommend being over-inclusive with the number (Kline, 1993), but usually a minimum of three is recommended in order to capture the true centre of each dimension (Carpenter, 2018). Two-item scales are plausible only if items are highly correlated (i.e., $r < 0.70$) (Worthington & Whittaker, 2006).

PHASE 3: SCALE EVALUATION

Step 7: Tests of Dimensionality

The term “test dimensionality” does not only refer to the particular set of items comprising a test; it is also the function of the interaction between

the set of items and the group of examinees responding to those items (Gessaroli & de Champlain, 2005). Earlier definitions related to testing homogeneity and reliability, while the current definition extends to incorporate some form of local independence. According to Brown (2015), the test of dimensionality is one in which extracted factor structure is tested at a different time point in a longitudinal study or in a new sample using CFA, bifactor modelling or measurement invariance.

Confirmatory factor analysis

Confirmatory factor analysis is a multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs. CFA borrows similar concepts from exploratory factor analysis (EFA). However, instead of letting the data tell us the factor structure, we specify an a priori hypothesized model to which systematic fit assessment procedures are applied and the relationship between latent constructs are estimated. Wieland et al. (2018) suggest that item removal should be considered when the Goodness of Fit (GFI) index indicates insufficient model fit based on the standardized residuals, estimated improvements in the χ^2 value with corresponding degrees of freedom, the magnitude of modification indices, the normed fit index, the value of the comparative GFI index and the overall interpretability. The techniques for testing dimensionality include the chi-square test of exact fit, Root Mean Square Error of Approximation ($RMSEA \leq 0.06$), the Tucker–Lewis Index ($TLI \geq 0.95$), Comparative Fit Index ($CFI \geq 0.95$), Standardized Root Mean Square Residual ($SRMR \leq 0.08$) and Weighted Root Mean Square Residual ($WRMR \leq 1.0$) (Morin et al., 2016; Yu, 2002).

CFA is a well-established method in the literature for confirmatory studies (Hair et al., 2019b) and has often been utilized in the past for scale reliability, identification of items whose removal may enhance content validity, and testing of convergent and divergent reliability (Hair et al., 2020). However, confirmatory composite analysis (CCA) has recently been proposed as a scale validation process and as a replacement for traditional scale development procedures. CCA as an alternative means to confirm linear composite constructs in measurement models is just emerging (Hair et al., 2019b, 2020; Henseler et al., 2014). While both CCA and CFA aim at confirming the measurement models, CCA is different from CFA based on its statistical objective, which is to maximize the variance extracted from dependable variables, and facilitating the prediction of these exogenous variables (Hair et al., 2020).

In addition to its ability similarly to CFA to carry out the checks mentioned above in terms of scale reliability, composite reliability, content validity, and convergent and discriminant validity, there is an emerging

literature suggesting the superiority of CCA over CFA. Its advantages are given in terms of the following: the number of items retained being on the higher end with CCA leading to improved construct validity with more items available to serve as proxies (Hair et al., 2020); determinant construct scores are available (Rigdon et al., 2019); and the potential of CCA to assess both reflective and formative measurement models (Hair et al., 2020; Motamarri et al., 2020). More details on how to perform a CCA can be found in Hair et al. (2020).

Bifactor modelling

Bifactor modelling, also referred to as nested factor modelling, is a form of item response theory used in testing dimensionality of a scale (Reise et al., 2007). In a bifactor model, items are influenced by what can be referred to as a global factor, and also display evidence of sets of items tapping into coherent, homogeneous subcomponents referred to as local factors. It is important to note that bifactor models are not the same as hierarchical models as displayed in Figure 1.4, since they imply different causalities.

By carrying out the factor analysis we can identify whether there is inherent multidimensionality and clusters within items essentially previously identified as unidimensional. The bifactor model partitions the total covariance among the items into a global factor component underlying all items and local factor components explaining the residual covariance not explained by the global factor (Morin et al., 2016).

To assess whether to retain a construct as unidimensional or multidimensional, the factor loadings from the general global factor are

Source: Authors.

Figure 1.4 Example of bifactor model versus hierarchical model

then compared to those from the group local factors. Understanding of residual properties of specific factor score estimates is important for their interpretation. Whether the specific factor scores or weighted combinations of the general and specific factors should be reported will depend on how the scores will be used (DeMars, 2013). Where the factor loadings on the general factor are significantly larger than the group factors, a unidimensional scale is implied (Reise et al., 2007). While these analyses can be rather complex, and require specific software packages, DeMars (2013) and Morin et al. (2016) provide some guidelines on how to interpret the bifactor modelling scores.

Measurement of invariance

Chan (2011) states that “we cannot assume the same construct is being assessed across groups by the same measure” without tests of measurement invariance (p. 108). Multigroup confirmatory factor analysis is often used to test measurement invariance across participants from various groups (Milfont & Fischer, 2010). It permits the examination of whether the same measure is interpreted in conceptually similar ways across different groups (Bialosiewicz et al., 2013). There are three typical phases of measurement invariance testing, namely configural invariance, metric invariance and scalar invariance. Though there are stricter tests of measurement invariance, researchers agree that these three sequential steps are sufficient for establishing measurement invariance (Bialosiewicz et al., 2013; Milfont & Fischer, 2010). For example, if we use age to check invariance, for configural invariance we will need to examine whether the overall factor structure fits well for all age groups in the sample. Metric invariance will examine whether factor loadings are similar across groups, and scalar invariance will examine whether the item intercepts are equivalent across groups.

Step 8: Tests of Reliability

Reliability is the degree to which the measure of a construct is consistent or dependable when a measurement is repeated under identical conditions. The tests developed to assess reliability of a scale includes Cronbach’s alpha (Vandenberg & Lance, 2000), test–retest reliability (coefficient of stability) (DeVellis, 2016), McDonald’s Omega (McDonald, 1999), Raykov’s rho (Raykov, 2011), split-half estimates, the Spearman-Brown formula, alternate form method (coefficient of equivalence) and inter-observer reliability (DeVellis, 2016). However Cronbach’s alpha and test–retest reliability are the mostly utilized to assess reliability of scales (Raykov, 2011).

Cronbach’s alpha considers the degree to which a set of items in the scale co-vary, relative to their sum score (DeVellis, 2016). Gardner (1995)

noted that alpha is maximized when every item in a scale shares common variance with at least some other items in the scale. Coefficients of internal consistency increase as the number of items goes up, to a certain point. For instance, a five-item test might correlate 0.45 with true scores, and a 12-item test might correlate 0.85 with true scores. Hence there should be a trade-off between the reliability and the brevity of the scale.

An alpha coefficient of 0.70 has often been regarded as an acceptable threshold for reliability (Nunnally, 1978). DeVellis (2016) suggests that, below 0.65, the reliability level is undesirable; between 0.65 and 0.70 it is minimally acceptable; between 0.70 and 0.80 it is respectable; between 0.80 and 0.90 it is very good; and much above 0.90 one should consider shortening the scale. Likewise, dropping items with poor inter-item correlation is likely to increase the internal consistency.

Test-retest reliability refers to the temporal stability of a test from one measurement session to another: that is, the test is administered to the same group of respondents at two different times whereby the correlation between scores on the identical tests operationally defines the test-retest reliability. Rosnow and Rosenthal (1991) identify a few limitations to test-retest: namely, shorter intervals between tests might cause the answers to be affected by memory; or, conversely, a longer interval might witness the maturation of the respondents which is reflected in their opinions. Pearson's correlation is often used to estimate the theoretical reliability coefficient between parallel tests, while some authors prefer the intra-class correlation coefficient (Weir, 2005).

Step 9: Tests of Validity

Scale validity is defined as the extent to which a measurement instrument is measuring the construct it was developed to evaluate (Raykov, 2011). It is important to understand that throughout scale development validation is an ongoing process which starts right from Step 1 with proper identification of the items and construct under study to its generalizability in this last step.

There are numerous ways of examining validity, but the most common validity tests for scale development are content validity, criterion validity and construct validity. Content validity has already been discussed in Step 2.

Criterion validity

This type of validity has two subcategories, namely concurrent validity and predictive validity. In concurrent criterion validity there is usually a known standardized instrument, or even a known criterion that measures the same underlying concept of interest. It is defined as the extent to which

test scores have a stronger relationship with the criterion, which is also known as the gold standard.

Concurrent criterion validity is the extent to which test scores have a stronger relationship with the criterion (gold standard), which is in most cases a more valid and established measurement. The validity is appraised using Pearson product moment correlation or latent variable modelling (Boateng et al., 2018). Establishing concurrent validity is particularly beneficial when a new measure is created that claims to be better than the prior instrument. In marketing, criterion validity can be crucial – as seen in the infamous case when Coca-Cola decided to alter the flavour of their trademark drink. Surveys were carried out to assess whether or not people liked the new flavour, which they did; however the researchers did not consider using the original taste as a benchmark to question the target audience on which one they preferred. Consequently, the company incurred heavy losses.

Nunnally (1978, p. 87) provides a useful definition of predictive validity as: “when the purpose is to use an instrument to estimate some important form of behaviour that is external to the measuring instrument itself, the latter being referred to as criterion [predictive] validity”. In predictive validity we assess the operationalization’s ability to predict something it should theoretically be able to predict. For example, we can consider the predictive validity for measure of attitude towards the advertisement (predictor) and the brand attitude (criterion). As rightly stipulated by Becker et al. (2013), it is important to look beyond any given set of phenomena for additional connections and for new opportunities to challenge the existing knowledge; and this is where predictive validity is important. Worthington and Whittaker (2006) firmly recommend cross-validation of re-specified structural equation models to establish predictive validity. Predictive validity can be estimated by examining the association between the scale scores and the criterion in question.

Construct validity

Construct validity is the extent to which the identified items are tapping into the underlying theory or model of behaviour. Boateng et al. (2018) suggest four indicators of construct validity relevant to scale development: convergent validity, discriminant validity, differentiation by known groups and correlation analysis.

Convergent validity is determined by examining the average variance extracted, which indicates the amount of variance explained by a latent variable. An average variance extracted value of 0.50 or higher indicates that convergent validity has been achieved (Hair et al., 2010). Convergent validity states that tests having the same or similar constructs should be

highly correlated. Two methods are often applied to test convergent validity. One is to correlate the scores between two assessment tools or tool sub-domains that are considered to measure the same construct. The other method is the multitrait-multimethod matrix (MTMM) approach.

Discriminant validity, on the other hand, ensures that a construct measure is empirically unique and represents phenomena in a way that does not merely reflect some other construct (Hair et al., 2010; Raykov, 2011). It requires that a test does not correlate too highly with measures from which it is supposed to differ (Campbell, 1960). Traditionally, in marketing it is established using the Fornell and Larcker (1981) criterion and cross-loadings (Chin, 1998), though recent research suggests that the Fornell-Larcker criterion is not effective under certain circumstances (Henseler et al., 2015). For example, it overestimates indicator loadings (Hui & Wold, 1982; Lohmöller, 1989). Henseler et al. (2015) carried out a simulation and their study revealed that the Fornell-Larcker criterion and assessment of the cross-loadings are insufficiently sensitive to detect discriminant validity problems. Henceforth the heterotrait-monotrait ratio of correlations (HTMT) is proposed as a better approach to assess discriminant validity.

Known-groups validation or differentiation between known groups involves examining the distribution of a newly developed scale score and showing that some scales can differentiate between different groups based on their score (DeVellis, 2016). The validity of the measure is established if the measure is able to discriminate between the groups through statistically significant findings.

Bivariate correlation can also be used to establish convergent and discriminant validity by providing evidence based on the relationship between a test score and a conceptually related construct. Convergent validity is identified by strong relationships, whereas discriminant validity was indicated by weak relationships (Swank & Mullen, 2017).

CONCLUSION

A rigorous scale development process is vital to increase the trustworthiness of research results. As rightly put by DeVellis (2016), the cost of using a poorly conceptualized scale, even if it is the only one available, might be greater than any benefits attained. Accurately defining and operationalizing the phenomenon helps tackle resulting challenges of poor validity and reliability or limited the scale's generalizability. The use of a systematic approach along with multiple methods increases the likelihood of constructing a valid, reliable and applicable scale (Streiner et al., 2015).

We support Wieland et al.'s (2017) dualistic approach to scale purification: that the scales should take into account not only statistical criteria usually related to statistical heuristics or tests but also judgemental criteria that focus both on theoretical and empirical meanings. While statistical criteria are essential to evaluate the quality of scales, judgemental criteria are what give theoretical meaning to the construct (Borsboom et al., 2004). Henceforth, throughout scale development, judgemental criteria should also be applied, especially when considering scale purification. Another important side of scale development is ensuring that the scale is parsimonious. It is about minimizing the amount of data required while ensuring that all relevant aspects of the construct are captured (Wieland et al., 2017). This calls for a balance between the statistical and judgemental criteria.

We acknowledge that there are a number of issues not addressed here, such as interpretation of the tests or the different software to use, and therefore we recommend reading on each suggested test to assess its suitability to the measures being developed. Likewise, the necessity of the nine steps that we have outlined in this chapter is very likely to vary from study to study, and it is at the researcher's discretion which method is most suited for their studies. While some of the steps may have been labelled differently, on a global level the essence of scale development remains the same.

REFERENCES

- Alaya, R.J. (2009). *The theory and practice of item response theory*. New York: Guilford.
- Almanasreh, E., Moles, R., & Chen, T.F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, 15(2), 214–221.
- 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: Springer.
- Beavers, A.S., Lounsbury, J.W., Richards, J.K., & Huck, S.W. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research, and Evaluation*, 18(6), 1–13.
- Becker, J.M., Rai, A., & Rigdon, E. (2013). Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance. Proceedings of the 34th International Conference on Information Systems (ICIS), Milan.
- Bialosiewicz, S., Murphy, K., & Berry, T. (2013). An introduction to measurement invariance testing: Resource packet for participants. Demonstration Session, American Evaluation Association, October, Washington, DC, 1–37.
- Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quinonez, H.R., & Young, S.L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149.
- Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061–1071.

- Brady, M.K., & Cronin Jr, J.J. (2001). Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *Journal of Marketing*, 65(3), 34–49.
- Brockhoff, K. (1975). The performance of forecasting groups in computer dialogue and face-to-face discussion. In H.A. Linstone and M. Turoff (Eds.), *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Campbell, D.T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15(8), 546–553.
- Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, 12(1), 25–44.
- Carr, J.H., Shepherd, R.B., Nordholm, L., & Lynne, D. (1985). Investigation of a new motor assessment scale for stroke patients. *Physical Therapy*, 65(2), 175–180.
- Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2), 245–276.
- Chan, D. (2011). Advances in analytical strategies. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology* (Vol. 1, pp. 85–113). Washington, DC: American Psychological Association.
- Chin, W.W. (1998). The partial least squares approach to structural equation modeling. *Modern Methods for Business Research*, 295(2), 295–336.
- Churchill, G.A. Jr (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1), 64–73.
- Clark, L.A., & Watson, D. (2016). Constructing validity: Basic issues in objective scale development. In A.E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 187–203). Washington, DC: American Psychological Association.
- Cohen, R.J., & Swerdlik, M.E. (2005). *Psychological testing and assessment: An introduction to tests and measurement* (6th ed.). Boston: McGraw-Hill.
- Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, 12(3), 229–238.
- Comrey, A.L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56(5), 754–761.
- Cristobal, E., Flavian, C., & Guinaliu, M. (2007). Perceived e-service quality (PeSQ): Measurement validation and effects on consumer satisfaction and web site loyalty. *Managing Service Quality: An International Journal*, 17(3), 317–340.
- Dabholkar, P.A., Thorpe, D.I., & Rentz, J.O. (1996). A measure of service quality for retail stores: Scale development and validation. *Journal of the Academy of Marketing Science*, 24(1), 3–16.
- Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, 5(4), 194–197.
- De Vaus, D.A. (2002). *Surveys in social research*. Crows Nest, NSW: Allen and Unwin.
- DeMars, C. (2010). *Item response theory*. Oxford: Oxford University Press.
- DeMars, C.E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13(4), 354–378.
- DeVellis, R.F. (2016). *Scale development: Theory and applications* (4th ed.). Los Angeles: Sage.
- Diamantopoulos, A., Riefler, P., & Roth, K.P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12), 1203–1218.
- Dray, S., Dunsch, F., & Holmlund, M. (2016). Electronic versus paper-based data collection: Reviewing the debate. World Bank Development Impact blog. <https://blogs.worldbank.org/impacetevaluations/electronic-versus-paper-based-data-collection-reviewing-debate> (Accessed 15 June 2020).
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381.
- Fanning, J., & McAuley, E. (2014). A comparison of tablet computer and paper-based questionnaires in healthy aging research. *JMIR Research Protocols*, 3(3), 38.
- Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39–50.

- Francis, J.E., & White, L. (2002). PIRQUAL: A scale for measuring customer expectations and perceptions of quality in internet retailing. In K. Evans & L. Scheer (Eds.), *2002 AMA Winter Educators' Conference* (pp. 263–270). Austin, TX: American Marketing Association.
- Frey, B.B., ed. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Los Angeles: Sage.
- Gardner, W. (1995). On the reliability of sequential data: Measurement, meaning, and correction. In J.M. Gottman (Ed.), *The analysis of change* (pp. 339–359). Mahwah, NJ: Erlbaum.
- Gessaroli, M.E., & de Champlain, A.F. (2005). Test dimensionality: Assessment of. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: Wiley.
- Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, 10(4), 544–565.
- Goetz, C., Coste, J., Lemetayer, F., Rat, A.C., Montel, S., Recchia, S., Debouverie, M., Pouchot, J., Spitz, E., & Guillemin, F. (2013). Item reduction based on rigorous methodological guidelines is necessary to maintain validity when shortening composite measurement scales. *Journal of Clinical Epidemiology*, 66(7), 710–718.
- Goldberg, L.R., & Velicer, W.F. (2006). Principles of exploratory factor analysis. In S. Strack (Ed.), *Differentiating normal and abnormal personality* (pp. 209–237). New York: Springer.
- Goldstein, M., Benerjee, R., & Kilic, T. (2012). Paper v plastic part 1: The survey revolution is in progress. World Bank Development Impact blog. <https://blogs.worldbank.org/impactevaluations/paper-v-plastic-part-i-the-survey-revolution-is-in-progress> (Accessed 15 June 2020).
- Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59(2), 93–104.
- Greenlaw, C., & Brown-Welty, S. (2009). A comparison of web-based and paper-based survey methods: Testing assumptions of survey mode and response cost. *Evaluation Review*, 33(5), 464–480.
- Guadagnoli, E., & Velicer, W.F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265–275.
- Guion, R.M. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1(1), 1–10.
- Haig, B.D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research*, 40(3), 303–329.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2019a). *Multivariate data analysis* (8th ed.). Andover: Cengage Learning.
- Hair, J.F., Gabriel, M.L.D.S., da Silva, D., & Braga Jr., S. (2019b). Development and validation of attitudes measurement scales: Fundamental and practical aspects. *RAUSP Management Journal*, 54(4), 490–507.
- Hair, J.F., Howard, M.C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research*, 109, 101–110.
- Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
- Harvey, R.J., & Hammer, A.L. (1999). Item response theory. *The Counseling Psychologist*, 27(3), 353–383.
- Haynes, S.N., Richard, D., & Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7(3), 238–247.

- Heene, M., Bollmann, S., & Bühner, M. (2014). Much ado about nothing, or much to do about something? Effects of scale shortening on criterion validity and mean differences. *Journal of Individual Differences*, 35(4), 245–249.
- Hendrickson, A.E., & White, P.O. (1964). Promax: A quick method for rotation to oblique simple structure. *British Journal of Statistical Psychology*, 17(1), 65–70.
- Henseler, J., Dijkstra, T.K., Sarstedt, M., Ringle, C.M., Diamantopoulos, A., Straub, D.W., . . . & Calantone, R.J. (2014). Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013). *Organizational Research Methods*, 17(2), 182–209.
- Henseler, J., Ringle, C.M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135.
- Hill, C.E., Kline, K.V., Miller, M., Marks, E., Pinto-Coelho, K., & Zetzer, H. (2019). Development of the meaning in life measure. *Counselling Psychology Quarterly*, 32(2), 205–226.
- Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
- Howard, M.C. (2018). Scale pretesting. *Practical Assessment, Research, and Evaluation*, 23(1), Art. 5.
- Howard, M.C., & Melloy, R.C. (2016). Evaluating item-sort task methods: The presentation of a new statistical significance formula and methodological best practices. *Journal of Business and Psychology*, 31(1), 173–186.
- Hui, B.S., & Wold, H. (1982). Consistency and consistency at large of partial least squares estimates. In K.G. Joreskog & H.O.A. Wold (Eds.), *Systems under indirect observation, part II* (pp. 119–130). Amsterdam: North-Holland.
- Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.
- Kim, S., & Stoel, L. (2004). Apparel retailers: Website quality dimensions and satisfaction. *Journal of Retailing and Consumer Services*, 11(2), 109–117.
- Kline, P. (1993). *A handbook of psychological testing* (2nd ed.). London: Routledge.
- Kock, F., Josiassen, A., & Assaf, A.G. (2016). Advancing destination image: The destination content model. *Annals of Tourism Research*, 61, 28–44.
- Ladhari, R. (2010). Developing e-service quality scales: A literature review. *Journal of Retailing and Consumer Services*, 17(6), 464–477.
- Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575.
- Leech, B.L. (2002). Asking questions: Techniques for semistructured interviews. *PS: Political Science and Politics*, 35(4), 665–668.
- Leisher, C. (2014). A comparison of tablet-based and paper-based survey data collection in conservation projects. *Social Sciences*, 3, 264–271.
- Lennon, R.T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16(3), 294–304.
- Lindeman, C.A. (1975). Delphi survey of priorities in clinical nursing research. *Nursing Research*, 24(6), 434–441.
- Linstone, H.A., & Turoff, M., eds. (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.
- Lohmöller, J.-B. (1989). Basic principles of model building: Specification, estimation, evaluation. In H.O.A. Wold (Ed.), *Theoretical empiricism: A general rationale for scientific model-building* (pp. 1–26). New York: Paragon House.
- Loiacono, E.T., Watson, R.T., & Goodhue, D.L. (2002). WebQual: A measure of website quality. *Marketing Theory and Applications*, 13(3), 432–438.
- Lorenzo-Seva, U., Timmerman, M.E., & Kiers, H.A. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, 46(2), 340–364.
- Marsh, H.W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, 5(1), 22–36.

- Martínez, J.A., & Martínez, L. (2010). Some insights on conceptualizing and measuring service quality. *Journal of Retailing and Consumer Services*, 17(1), 29–42.
- McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- McGartland Rubio, D., & Kimberly, K.L. (2005). Content validity. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*. Pittsburgh: Elsevier.
- Meyers, L.S., Gamst, G., & Guarino, A.J. (2016). *Applied multivariate research: Design and interpretation* (3rd ed.). Thousand Oaks, CA: Sage.
- Milfont, T.L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111–130.
- Morin, A.J., Arens, A.K., & Marsh, H.W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1), 116–139.
- Motamarri, S., Akter, S., & Yanamandram, V. (2020). Frontline employee empowerment: Scale development and validation using confirmatory composite analysis. *International Journal of Information Management*, 54, 102177.
- Nosek, B.A., Banaji, M.R., & Greenwald, A.G. (2002). E-research: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues*, 58(1), 161–176.
- Notter, L.E. (1983). *Essentials of nursing research* (3rd ed.). New York: Springer.
- Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Osborne, J.W., & Costello, A.B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research, and Evaluation*, 9(1), Art. 11.
- Osborne, J.W., Costello, A.B., & Kellow, T.J. (2008). Best practices in exploratory factor analysis. In J.W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 86–99). Los Angeles: Sage.
- Perroca, M.G. (2011). Development and content validity of the new version of a patient classification instrument. *Revista latino-americana de enfermagem*, 19(1), 58–66.
- Pett, M., Lackey, N., & Sullivan, J. (2006). *Making sense of factor analysis in health care research: A practical guide*. London: Sage.
- Piedmont, R.L. (2014). Inter-item correlations. In A.C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 3303–3304). Dordrecht: Springer.
- Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29(5), 489–497.
- Polit, D.F., Beck, C.T., & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459–467.
- Preacher, K.J., & MacCallum, R.C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2(1), 13–43.
- Raykov, T. (2011). Evaluation of convergent and discriminant validity with multitrait-multimethod correlations. *British Journal of Mathematical and Statistical Psychology*, 64(1), 38–52.
- Reise, S.P., Morizot, J., & Hays, R.D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(1), 19–31.
- Rigdon, E.E., Becker, J.M., & Sarstedt, M. (2019). Parceling cannot reduce factor indeterminacy in factor analysis: A research note. *Psychometrika*, 84(3), 772–780.
- Rosnow, R.L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin*, 110(3), 574–576.
- Roth, P.L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47(3), 537–560.
- Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S., & Rauch, S. (2003). Objectifying

- content validity: Conducting a content validity study in social work research. *Social Work Research*, 27(2), 94–104.
- Safikhani, S., Sundaram, M., Bao, Y., Mulani, P., & Revicki, D.A. (2013). Qualitative assessment of the content validity of the Dermatology Life Quality Index in patients with moderate to severe psoriasis. *Journal of Dermatological Treatment*, 24(1), 50–59.
- Scheele, D. (1975). Reality construction as a product of Delphi interaction. In H.A. Linstone & M. Turoff (Eds.), *The Delphi method: Techniques and applications* (p. 37). Reading, MA: Addison-Wesley.
- Schinka, J.A., Velicer, W.F., & Weiner, I.R. (2012). *Handbook of psychology, vol. 2: Research methods in psychology*. Hoboken, NJ: Wiley.
- Shrotriyia, V.K., & Dhanda, U. (2019). Content validity of assessment instrument for employee engagement. *SAGE Open*, 9(1).
- Sim, S.M., & Rasiah, R.I. (2006). Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper. *Annals of the Academy of Medicine Singapore*, 35(2), 67–71.
- Streiner, D.L., Norman, G.R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press.
- Swank, J.M., & Mullen, P.R. (2017). Evaluating evidence for conceptually related constructs using bivariate correlations. *Measurement and Evaluation in Counselling and Development*, 50(4), 270–274.
- Tabachnick, B.G., Fidell, L.S., & Ullman, J.B. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.
- Teeroovengadam, V., & Nunkoo, R. (2018). Sampling design in tourism and hospitality research. In R. Nunkoo (Ed.), *Handbook of research methods for tourism and hospitality management* (pp. 477–488). Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.
- Tersine, R.J., & Riggs, W.E. (1976). The Delphi technique: A long-range planning tool. *Business Horizons*, 19(2), 51–56.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association.
- Thurstone, L.L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind*. Chicago: University of Chicago Press.
- Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4–70.
- Velicer, W.F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement*, 36(1), 149–159.
- Weir, J.P. (2005). Quantifying test–retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength & Conditioning Research*, 19(1), 231–240.
- Wieland, A., Durach, C.F., Kembro, J., & Treiblmaier, H. (2017). Statistical and judgmental criteria for scale purification. *Supply Chain Management: An International Journal*, 22(4), 321–328.
- Wieland, A., Kock, F., & Josiassen, A. (2018). Scale purification: State-of-the-art review and guidelines. *International Journal of Contemporary Hospitality Management*, 30(11), 3346–3362.
- Worthington, R.L., & Whittaker, T.A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, 34(6), 806–838.
- Yu, C.Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Unpublished doctoral dissertation). University of California Los Angeles.
- Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A.R., (2015). Design and implementation content validity study: Development of an instrument for measuring patient-centered communication. *Journal of Caring Sciences*, 4(2), 165–178.