Chapter 3: Measurement Theory and Psychological Scaling

Daniel P. Hinton and Tracey Platt

Introduction

Quantitative research in every branch of psychology involves the measurement of psychological constructs, and consumer psychology is no exception. The use of tools to measure psychological constructs is known as **psychometrics**. This chapter will outline the use of psychometric measures within consumer psychology and related fields – both in academic and practice settings – and discuss the theory underlying psychological measurement, before exploring the process by which these measures are developed by psychologists.

Why use psychometrics in Consumer Psychology?

Over the past 50 years and beyond, the field of consumer psychology has seen a widespread adoption of psychometric tools for the measurement of psychological constructs of interest. In practices such as marketing, in the past these variables tended to be measured poorly (Churchill, 1979). However, companies are becoming increasingly aware that the need to properly understand consumer behaviour necessitates the use of robust measurement tools in order to understand the psychological processes that drive this behaviour. Psychometrics have been developed and deployed to understand consumer preferences and attitudes (e.g. Kidwell, Hardesty & Childers, 2007; Gattol, Sääksjärvi & Carbon, 2011), how brand loyalty is created and maintained through the development of affection for a brand (Albert & Valette-Florence, 2010), how consumers react to advertising through new forms of media (Bauer Reichardt, Barnes, & Neumann, 2005), the process by which consumers make purchasing decisions (Schwartz et al., 2002), and even more negative consumer behaviours such as compulsive purchasing (Maraz et al., 2015). Overall, psychometric measures equip the consumer psychologist with a set of tools with which they can better understand customers and their behaviour.

Learning Objectives

By the end of this chapter, you should be able to:

- understand the key differences between psychometric and non-psychometric measures
- understand the theoretical underpinnings of psychological measurement, and compare and contrast Classical Test Theory with Item Response Theory
- understand how psychometrics are typically used within the field of consumer psychology
- understand the process of scale design and scale development, from initial conception through to finalising a scale's item set
- critically evaluate modern alternatives to the psychometric approach to scale development, such as Rossiter's C-OAR-SE method

Psychometrics and their Key Characteristics

At the broadest level of abstraction, a psychometric tool is one that is designed to measure some kind of psychological construct. The constructs measured by psychometric instruments are many and varied, encompassing personality traits, mental ability, and a host of other individual differences such as attitudes, values, and beliefs. Psychometrics are typically comprised of a number of questions or statements (known as **items**) to which the test taker must respond from a finite number of response options. Psychometric tools may be comprised only of a single scale that has been designed to measure a single construct (in which case they are referred to as **unifactorial** measures). Equally, they may be made up of a number of conceptually related scales, all of which are designed to measure distinct constructs, such as personality questionnaires.

To the casual observer, it may seem as though there is no real difference between a psychometric tool and any other form of questionnaire or survey. However, there are a number of key differences between psychometric and non-psychometric tools. Firstly, published psychometrics should have all been through a rigorous process of design and development, similar to the one described later on in this chapter. Secondly, and related to the previous point, psychometric measures should all demonstrate good **psychometric properties**. When we talk about psychometric properties, we are referring to a tool having an evidence base to support its **reliability** and **validity**. A measure's reliability refers to the consistency or accuracy with which is measures the construct of interest. A measure shows validity if it measures what it is supposed to. That is to say, it measures the construct that it was designed to. In psychological research, there are a great many forms of reliability and validity, only some of which are relevant for the scale development process. A brief summary of the most relevant types of reliability and validity for the design and development of psychometric measures is shown in Box 3.1.

Finally, psychometric measures are standardised in a number of different ways. Firstly, in the case of **normative measures**, a test taker's score on a scale is converted to a standardised score, such as a **percentile** or a **sten**. Whereas the test taker's scale score (known as their **raw score**) does not tell the test user much about their level of a trait or resultant likely behaviour, standardised scores allow for comparison with a **norm group** of test takers who have previously completed the tool. This allows a test taker's score to be placed within the distribution of test takers, and inferences to be made about their level of a trait in comparison to others. The other key aspect of standardisation of psychometric measures aims to make them more objective and less prone to random and systematic error variance. The instructions provided to test takers is always the same, and responses to items are scored in a strict and highly procedural way. Indeed, this principle of minimising error variance is a central one in the practice of psychometrics. To understand this further, however, it is necessary to understand something about their theoretical foundations.

BOX 3.1: Types of Reliability and Validity Relevant for Scale Development

Internal Consistency: A measure of the degree to which all items within a scale are consistent in their measurement. For example, in a scale that demonstrates good internal consistency, a test taker that scores highly on one item would be expected to score highly on the other items as they are all measuring the same construct. The method by which internal consistency is assessed depends upon the response format of the scale. Scales with two or more response options (e.g. Likert-type scales) are most frequently assessed using **Cronbach's α** (Alpha). Scales with dichotomous response formats (e.g. 'true/false', 'yes/no', correct/incorrect) tend to be assessed using **Kuder-Richardson Formula-20** (KR-20).

Content Validity: A measure of the degree to which the items within a scale *appear* to measure the construct of interest, as judged by subject matter experts. Traditionally a qualitative assessment, there has been a relatively recent move towards more robust, quantitative assessments of content validity.

Construct Validity: An overall assessment of whether the items within a scale measure the construct of interest. Construct validity takes a number of forms, all of which are assessed quantitatively using some form of factor analysis procedure such as PAF or CFA. **Structural validity** is an examination of whether the underlying latent factor structure of the tool is consistent with that which is theoretically expected. In single-construct measures, this tends to equate to the degree to which the items within the scale are unifactorial. **Convergent validity** represents the degree to which the items within a scale measure the same construct as items from another scale that has been designed to measure the same (or a very similar) construct. Conversely, **discriminant** (or **divergent**) **validity** is the degree to which a scale's items are unrelated to items from other scales that have been designed to measure different constructs.

BOX 3.1 (continued)

Criterion-related Validity: A measure of the degree to which scores on a scale are related to important outcomes. There are three kinds of criterion-related validity, all of which are assessed by computing correlation coefficients between scale scores and outcome variables. **Predictive validity** represents the degree to which a scale predicts an outcome in the future. **Concurrent validity** represents the relationship between scale scores and an outcome in the present. **Postdictive validity** refers to the relationship between scale scores and an outcome in the past. Some outcome variables in consumer psychology that are frequently used as criteria for these purposes are discussed below in Step 6.

Measurement Theory

Psychometric practice is heavily influenced by theory. The earliest attempt to formalise measurement theory was in the form of Classical Test Theory. First proposed by Novick (1966), Classical Test Theory revolves around the principle that psychological constructs are **latent** (that is to say that they are 'hidden', and are not directly observable). As a result, they cannot be measured in the same way as one might a person's height or weight. Rather, the way that we measure these hidden constructs is by measuring a person against indicators of that construct. Classical Test Theory makes the assumption that a test taker has a 'true' level of a trait (their 'true score'), but that measurement of this true score is necessarily obscured by random error in the process of measurement against these indicators. Therefore, the equation used to describe the relationship between a test taker's true score and their observed score according to Classical Test Theory is described by the following equation:

$$X = T + E,$$

where X represents a test taker's observed scale score, T represents their true (i.e. 'hidden') score, and E represents measurement error. Therefore, the key principle of Classical Test Theory is that, in order to be able to measure psychological constructs with any degree of accuracy, we must use a range of indicators that adequately samples behaviour relevant to the construct of interest, while, at the same time, minimises measurement error.

A relatively modern alternative to classical test theory is **Item Response Theory** (IRT). Whereas the psychometric focus of Classical Test Theory is based upon the properties of the test as a whole, the principles of item response theory (IRT) rely on the use of individual items, or a sub-set of items from within the test. One of the principle tenets of IRT is that the probability of a particular response to an item depends upon a number of parameters, most notably the 'person parameter', represented by Greek letter theta (θ). θ represents the true level of a trait (i.e. the construct of interest) of a particular test taker.

A central concept within IRT is that of item and test **information**. Information is analogous to reliability in Classical Test Theory, with one notable difference: It is recognised that the reliability of an item or scale may vary according to the level of θ of each individual test taker. For example, a scale that measures extraversion may be extremely reliable when administered to a test taker with a moderate level of extraversion, but may be much less reliable when measuring test takers with extremely high or extremely low levels of extraversion. IRT techniques take this contingent reliability into account when describing an item or test's information, representing it as an **item information curve** (IIC) or **test information curve** (TIC). Examining item and test information in this way can be a helpful additional way of identifying problematic items as a part of the scale development process alongside more traditional methods (see Steps 5 & 6 later in this chapter).

The Application of Psychometrics in Consumer Psychology

Psychometric tools have been used extensively in consumer psychology. Utilising psychometric methods and models allows for reliably measuring an individual's abilities, their attitudes, and their personality traits, and, thus, more confidently predicting patterns of behaviour. Therefore, the successful prediction of consumer behaviour has many applications within consumer psychology, where psychometric testing adds to the validity of research.

The technique of testing a sample of a population in order to predict behaviour generalisable to a more global level forms the basis of much research in consumer behaviour. However, conducting research does not necessarily require that a new instrument be constructed. A huge range of psychometric tools are available that a researcher can draw upon to measure variables relevant to consumer psychology. For example, Klein, Ettenson and Krishnan (2005) studied ethnocentrism from a viewpoint of consumers' preferences for foreign products. To investigate this phenomenon, they tested participants with a previously developed instrument, the CETSCALE (Netemeyer, Durvasula, & Lichtenstein, 1991). The authors of the CETSCALE had rigorously tested its factor structure and internal consistency estimates across four countries. As well as this, they examined discriminant validity by investigating the sample's general attitudes towards their home country, exploring their perception of their home country's people, practices and values. Furthermore, as there should be a link between countries for this consumer behaviour, the scale's **nomological validity** was also examined, to find out whether the constructs measured by the CETSCALE could be applied across different countries. For Klein and colleagues, the purpose of using this soundly psychometrically tested scale gave them an instrument to extend the knowledge around this construct further, by using it to test in different, previously untested countries from the original four, thus strengthening the validity and predictive value of the tool.

Psychometrically testing the characteristics and behaviours that form the constructs of human personality is well established. However, Aaker (1997) discovered that by applying the same psychometric principles to brands, one could identify a brand's 'personality'. Aaker defined this construct as "the set of human characteristics associated with a brand" (pp. 347). This may seem like an odd thing to want to examine, but, once one understands these perceived symbolic characteristics, they can be used to better understand the personality-based drivers behind purchasing decisions and brand loyalty. The steps taken to uncover the traits that exist within a coherent model of brand personality were done sequentially and rigorously, to ensure the properties remained reliable, valid and, thus, generalisable. Factor analysis conducted on the item set indicated that brand personality was comprised of five factors: Sincerity, Excitement, Competence, Sophistication, and Ruggedness.

BOX 3.2: A Case Study focusing on Positive Affectivity

The role that affect plays on consumer judgement and decision-making is well documented in the consumer behaviour literature (Cohen, Tuan Pham, & Andrade, 2008). Positive affect and the facial expression of joy – namely smiling – has been shown to increase brand loyalty and repeat purchasing behaviour (Jacoby & Kyner, 1973), customer satisfaction (Söderlund & Rosengren, 2008), and extends to product advertisement (Schmitt, 1999). However, recent investigations into gelotophobes, individuals with the fear of being laughed at, showed how these individuals misperceive displays of positive affect (Ruch, Hofmann, & Platt, 2015; Hofmann, Platt, Ruch, & Proyer, 2015). These individuals do not experience positive emotional contagion from hearing laughter or seeing smiling, but irrationally feel that they are being ridiculed. In these situations, most people will mirror the enjoyment and begin to smile also, but gelotophobes react to them by displaying facial expressions of contempt.

With up to 13% of the population experiencing some form of gelotophobia, any marketing strategy utilising positive affect should also consider its potential aversive impact upon their target market, in that the same stimuli could have both a positive or aversive impact on consumer behaviour such as purchasing decisions. A deeper understanding can be obtained using psychometric testing, which can circumvent any problems. Gelotophobes are less aversive to low arousal displays of positive emotion. Psychometrically testing the perceived level of arousal of the positive affect embedded within the stimuli could benefit the marketing campaign, gaining all the well documented benefits of using the positive affective states, while having none of the disadvantages of it becoming distressing to those sensitive to the laughter of others.

BOX 3.2 (continued)

Conversely, as eliciting negative emotion can also positively influence consumer behaviour, having psychometric tools that can measure when emotion elicitation becomes aversive can guide marketing.

Research projects in consumer psychology – as is the case in other branches of psychology – routinely include questions relating to demographic information. However, a recent big data trend is towards the use of psychographics to complement demographic information about consumer populations (e.g. Lin, 2002). Whereas demographic questions gather information on the characteristics of the population being measured, such as gender, age, education, race/ethnicity, occupation, income level and marital status, psychographics start with the very sensible proposition that it would be naïve to assume that these demographic groups are homogeneous in other respects. If one is able to understand particular consumer groups in terms of their differences on key psychological constructs, a greater depth of understanding of the consumer can be achieved, and the efficacy of things like marketing practices can be enhanced. Psychographics seeks to do this by enhancing market segmentation processes. Whereas traditional approaches to market segmentation (Smith, 1956) seek to understand groups within a market in terms of shared behaviour and needs, psychographic segmentation attempts to classify consumer groups according to similarities in their psychometric profiles (Kotler, 1997). This approach provides marketers with greater insight into the nature of sub-markets within demographic market segments, allowing for more effective, targeted sub-marketing strategies (Lin, 2002).

The Scale Design and Development Process

While there are a huge amount of psychometric measures – both freely available academic scales and commercially-focused tools marketed by test publishers – available for the purposes of research in consumer psychology, it may occasionally be the case that a researcher is unable to obtain a suitable scale to measure the psychological construct in which they are interested. This may be for a number of reasons. It may be the case that the researcher wishes to explore a new construct that has not been well researched at that point in time. Equally, measures may exist that measure the construct, but they may be of poor quality, prohibitively expensive, or may be dated and may not reflect our current understanding of the construct's nature. In any of these cases, a researcher might have to consider the possibility that he or she develops a new scale to measure the construct from scratch.

The process of scale design and development is a long and relatively labour-intensive one, and certainly not to be taken lightly. This reflects the nature of psychometric scales in that they have been carefully designed to ensure reliability and validity of measurement. Over the years, a great many extensive academic papers, book chapters, and entire textbooks have been written on the specifics of scale design and development (e.g. DeVellis, 2016; Schweizer & DiStefano, 2016; Hinkin, 1995; 2005; Churchill, 1979; Nunnally, 1978), and it is not the intention of the authors to present a comprehensive account of the single best practice approach within this chapter. Rather, the authors aim to synthesise and condense the wider scale design and development literature to provide a broad overview of the process that a researcher might use as a starting point in the development of a scale development project.

In the following steps, the authors will describe common approaches to the design of a scale. For the sake of simplicity, in the main, the process describes the design of a single, unifactorial scale. If one were seeking to design a multifactorial tool such as a personality measure, this process would be replicated for each construct to be measured by the final tool.

Step One: Defining the construct

The first – and in many respects the most critical – step in the process of scale design is to establish a clear definition of the construct of interest. This may seem like a fairly obvious point to make, but, in the authors' experience, students frequently express an interest in conducting scale development projects with no clear idea of what it is they want to measure before they begin the process. In all branches of psychology, research is continually moving forward, and, as a result, our understanding of the psychological constructs in which we are interested will slowly (or, occasionally, rapidly) evolve. The foundation of a good, robust psychometric measure is a coherent link between the construct being measured and the content of the items that comprise it.

There are two broad approaches to scale design. The approach taken is directly informed by current understanding of the construct of interest. If a construct is well understood and clearly defined within the literature, the **deductive approach** to scale design should be taken. The deductive approach involves a thorough search of the published literature relevant to the construct of interest in order to develop a definition of the construct that covers its full breadth and depth. It may be possible that an adequate working definition of this construct already exists, though, in many cases, existing definitions may need to be modified somewhat to accurately capture current understanding of it.

However, it is sometimes the case that no clear definition of the construct exists in the literature (perhaps because the construct is relatively new), or that the understanding of the construct has evolved over time to the point where definitions in the literature no longer accurately reflect its breadth and detail. In these instances, the **inductive approach** to scale design should be taken. The inductive approach relies on the use of field-gathered, observational information, most frequently in the form of interviews with **subject matter experts** (SMEs). SMEs can be anyone who has a good working knowledge of the area in which the construct is based, so can be either academics or practitioners specialising in that area, or a mixture of both. The inductive approach seeks to identify patterns of behaviour that can then be used to guide the development of new theoretical models. However, as Gioia, Corley, and Hamilton (2012) highlight, this approach must "apply systematic conceptual and analytical discipline that leads to credible interpretations of data" (pp. 15). Gibbert and Ruigrok (2010) argue that even with the lack of control of the context of this research method, it is possible to ensure qualitative rigor that also has both reliability and validity in the way the research is conducted. This can be achieved by addressing considerations such as triangulating sources of evidence, used here to mean looking at the same phenomenon but using different data collection strategies and source materials.

One such systematised information collection process that is particularly useful for the inductive approach is the Grounded Theory method, first proposed by Glaser and Strauss (1967). This method involves stages in the grounded theory process and, although there has since been divergence reported in the literature in terms of how the processes should be conducted (e.g. Heath & Cowley, 2003), the principles remain the same. The first is *data gathering* from an initial research question, which is then posed to SMEs via interviews, observations, diary keeping or focus group discussion. The second stage is *note taking* on the specifics of the collected data. The third stage is the *coding* of the information into categories, which then develop into the new construct or constructs. Gathering more data will do one of two things at this point. It will either yield new information, which can be grouped into an additional, new concept, or it will be repeated information, which relates to an older concept that has already been obtained and categorised. This process of collecting, analysing, and interpreting is continued until the point of information saturation is reached, at which point no new information can be gleaned from the data analysis. This is where the fourth and fifth stages of the process begin: *sorting*, followed by *writing up*. The advantage of the Grounded Theory approach is that it is designed to develop clear definitions of the construct or constructs of interest from first principles, untainted by any existing preconceptions that the researcher may have going into the process.

Step Two: Generating the initial item set and deciding upon response format

The next step in the process is to generate an initial set of items. This item set will undergo several stages of iterative refinement before the final item set that makes up the scale is identified, so it stands to reason that substantially more items than are intended to go into the final scale will need to be generated in this step.

The broad approach to scale design taken in the previous step – whether it be the deductive or inductive approach – will have an impact upon the process of item generation in this step. This is where the advantages of the deductive approach become apparent, as it is generally the case that the initial item set will be made up of fewer items than that for the inductive approach (Burisch, 1984). A rule of thumb for the deductive approach is to generate around twice as many items as are intended to make up the final item set (Hinkin, 1998), so, if a researcher was intending to design a relatively short scale of 10 items, at least 20 items should be generated. Other authors have recommended a more cautious approach to item generation, recommending that as many as three or four times the number of items required for the final scale are generated at this stage (DeVellis, 2016).

Regardless of the approach taken, the actual generation of items in this step is a relatively straightforward process. In many respects, the quality of items does not matter as long as sufficient items are generated, as the process of content validity and subsequent development will separate the weak items from the stronger ones. A good place to begin writing items is to attempt to paraphrase the construct as it has been defined (DeVellis, 2016). From there, item wording can be changed to express similar ideas in different ways. However, Hinkin (2005) identifies a number of guiding principles to the writing of items. Firstly, the perspective – whether behaviourally based or affective – of the items should be consistent across the whole item set. Secondly, each item should only reflect a single issue. Thirdly, items should not be phrased in the form of leading questions, neither should they be on topics that all test takers would be likely to endorse. Finally, every item should be written so that it can be easily understood by all test takers, so the researcher needs to carefully consider the target population for the tool and their likely level of literacy. As a rule, shorter items tend to be easier to read (DeVellis, 2016).

A key decision to be made at this stage is that of the response format of the items. There are many potential response format options that may be chosen for the tool, each of which have both strengths and weaknesses (see Table 1 for a summary). The nature of the response format chosen necessarily impacts upon later stages of the scale development process in terms of the kind of analyses that can – and should – be conducted upon the data collected using them. It is important to note that, whichever response format is decided upon, the chosen format should be the same across all items within a scale. While it is possible to include multiple response options within the same tool (e.g. Saville's Wave personality inventory; Kurz & McIver, 2008), for the purposes of development, multiple different response formats should be treated as different scales.

One particular distinction to draw is between normative and **ipsative** measures. Whereas normative measures seek to compare test takers to others (e.g. that a particular test taker is more outgoing than most people), ipsative measures draw comparisons between traits within an individual. Ipsative items force test takers to choose between mutually exclusive response options, allowing scales to differentiate between traits or behaviours in terms of their strength or importance relative to one another. For example, an ipsative item might ask a test taker to rank a set of statements in order of the degree to which they feel each statement applies to them, or might ask them to select one statement as being 'most like me' and another as 'least like me'. While ipsative measures do not allow comparison between test takers, they have a distinct advantage over normative measures in that they can be designed in such a way as to minimise **socially desirable responding** (SDR; see below) behaviour (Bäckström, Björklund, & Larsson, 2009). However, the process for scale development of ipsative measures is much more complex than development of normative measures (Hicks, 1970). For this reason, and for the purposes of clarity, the remaining steps of the scale development process described here will make the assumption that a normative scale is being developed.

Table 1.

Summary of response formats in psychometric scales.

Measure Type	Response Format	Example Response Options	Example Tools	Strengths	Weaknesses
Normative	Likert Scale	1 = Strongly disagree 2 = Disagree 3 = Neither Agree nor Disagree 4 = Agree 5 = Strongly Agree	NEO PI-3 (McCrae, Costa & Martin, 2005)	Allows for precise measurement; allows for comparison between test takers	Takes longer to complete
	Likert-type Scale	1 = Does not apply at all … 10 = Totally applies	Albert and Valette-Florence's (2010) Brand Love Scales		
	Dichotomous/ binary	True False	Hogan Personality Inventory (HPI; Hogan, 1995)	Quick to complete	Loss of data fidelity (i.e. accuracy of measurement)
Ipsative	Forced Choice; Rank Order	Most like me Least like me	PAPI (Lewis & Anderson, 1998)	Allows comparison of relative strength and importance of traits and behaviours within an individual; Less threat of socially desirable responding (SDR) behaviour	Cannot compare test takers

One particular issue of contention within the literature is whether any items should be generated that are **negatively-keyed**. Negatively-keyed items are worded in such a way that *not* endorsing them is associated with a higher score on the construct of interest than endorsing them. For example, in a scale to measure extraversion, the item *"I don't like being in large crowds of people"* would be endorsed more by test takers with lower levels of extraversion. Inclusion of negatively-keyed items presents a number of benefits to a scale, as they combat response set bias (Price & Mueller, 1986), and encourage test takers to pay more attention to the content of the items. However, their inclusion has the potential to affect the psychometric properties of the final scale (Harrison & McLaughlin, 1991, cited in Hinkin, 2005).

Step Three: Establishing content validity

Once the initial item set is generated, the process of iteratively reducing this item set down to make the resultant scale as robust as it can be can begin. The first stage in this process – and one that is frequently overlooked – is to establish content validity. It may appear, at least to the designer of the item set – that the items that have been generated are all perfectly fit for purpose, but this is a dangerous assumption to make. Once the process of item trialling (see next step) has begun, there is no going back – or, at least, not without substantial additional effort – and the discovery that one's item set does not behave in the way expected necessitates returning to the item generation stage and discarding any data collected to that point. Clearly, this a scenario that should be avoided, and the establishment of content validity is a critical step in minimising the risk of this happening.

To explore the content validity of the item set, the researcher will need to draw upon the expertise of some subject matter experts (SMEs), as is the case for the process of construct development in the inductive approach, albeit using different methods to do so. Generally speaking, the number of SMEs needed for this step is much smaller than the number of participants required in the later item trialling steps, as the analyses conducted on their responses tends to be less sophisticated, statistically speaking. For the most basic of content validity processes, feedback from between five and ten SMEs should be sufficient. Exploration of content validity tends to be more qualitative in nature, though there has been a relatively recent move towards imposing more quantitative frameworks upon the process (Hinkin, 2005).

Typically, this process is made up of two components. In the first, SMEs are asked to make a judgement (by using, for example, ratings of 1 to 5) on the degree to which each item taps into the construct of interest, based upon the definition with which the researcher has supplied them. In the case of tools that are made up of multiple scales, it is helpful to have SMEs rate the degree to which each item taps into every construct measured, not just the construct measured by the scale to which the item belongs. In the second, SMEs are asked to make comments upon individual items, and upon the scale in general. They are asked to comment on any items that they feel could be better worded (or should be removed altogether), and asked to highlight whether there is anything missing from the item set in terms of the breadth and depth of coverage of the construct.

Guidance on how the researcher then interprets and utilises this data varies within the literature. The researcher may look over the items and remove any item which is poorly worded (for which a simple rewording of the item isn't an option), and/or which does not appear to tap strongly into the construct of interest. Alternatively, a more quantitative approach may be taken, in which the degree of agreement between SMEs' ratings of each item is calculated (using, for example, Fleiss' Kappa, a statistical measure of the consistency of rating across different raters), discarding any item for which consensus has not been reached. However, care should be taken when making decisions about whether to discard an item on the basis of a poor Fleiss' Kappa value (see Troubleshooting Tips). Content validity on multifactorial tools, in which SMEs are asked to rate the degree to which each item relates to each construct of interest, allows for a degree of quantitative comparison between ratings of relevance of an item across constructs. Using an ANOVA, a researcher can examine which of the items are rated as significantly more relevant to their intended construct than to other constructs in the tool, discarding those that are not (Hinkin, 2005). This is a much more robust and objective approach to the establishment of content validity, though it requires the collection of data from substantially more SMEs than the approaches described above in order to achieve adequate statistical power.

Step Five: Initial item trialling

Once the item set has been refined and the researcher is happy that the scale to this point demonstrates acceptable content validity, the next step involves the further development of the scale by examining its psychometric properties, and identifying any items that may be negatively affecting its reliability or validity. This process involves examining the scale's internal consistency and structural validity, and removing items from it iteratively, one at a time, until an item set can be identified that demonstrates robust psychometric properties.

To do so requires some data. The scale items are administered to a relatively large number of participants, chosen to represent the group for whom the final scale is intended. These participants complete all items that remain in the scale. The only data required for this process are the item responses themselves, so any other data collected at this stage, such as demographic data, will not be used for any analyses. The key determinant of the sample size required for this phase of item trialling depends upon the number of items within the item set. As some form of factor analysis (see below) is going to be the most complex analysis used in this step, the normal rules should be followed for establishing minimum sample size for these procedures. Though guidance in the literature varies, the most liberal estimate offered of the ratio of number of participants : number of items is 10:1 (Nunnally, 1978). That is to say, if the item pool entering this stage of analysis was made up of 20 items, approximately 200 participants would be needed.

First, the item set is checked for internal consistency. This is achieved by computing Cronbach's α for the scale as a whole, and for each individual scale if the tool is made up of a number of separate scales. The common rule of thumb is that a scale with a value of Cronbach's α above .7 shows acceptable internal consistency, though the closer to 1.0 this value is, the better, generally speaking. Certainly, a value of Cronbach's α below .7 is cause for concern, indicating that at least one of the items is adversely affecting the scale's reliability. It is worth pointing out that, if the response format of a scale's items is dichotomous in nature, the value of KR-20 for the scale should instead be calculated. The way in which KR-20 functions is very similar to Cronbach's α, however, so the rest of this section will refer exclusively to the calculation of α values.

Next, the tool's structural validity is examined. To achieve this, a dimension reduction technique such as Principal Axis Factoring (PAF), or another form of Exploratory Factor Analysis (EFA) is used. There is a degree of debate in the literature as to the 'correct' or 'best' method to use to explore the underlying structure of the tool, and, at least to some degree, the choice of analytic approach is influenced by the nature of the data obtained.

Hinkin (2005) suggests that the rotation method chosen for EFA depends upon whether the latent factors underlying the data are expected to be correlated (oblique rotation) or uncorrelated (orthogonal rotation), but that, to be on the safe side, it is prudent to run factor analysis with each and compare the solutions obtained. A point of much greater contention is the extraction method that should be used. It has been argued that, generally speaking, a maximum likelihood estimator (ML) is the best choice when data are normally distributed, whereas PAF is the best choice for non-normally distributed data (Costello & Osborne, 2005). However, purists might argue that, since Likert-type scales are ordinal in nature rather than being scale-level data, a diagonally-weighted least squares (WLSMV) estimator is more appropriate for factor analysis (Bandalos, 2014). To further muddy the water, in the case of dichotomous data, it appears that an unweighted least squares (ULSMV) estimator appears to produce the best results (Parry & McArdle, 1991). In practice, however, most forms of factor analysis techniques yield similar results, particularly with larger sample sizes. However, one approach that should never be used for the purposes of scale development is Principal Components Analysis (PCA). PCA is not, in actual fact, a form of factor analysis at all, and, as such, cannot be used to identify the latent structure underlying a dataset. Lee and Hooley (2005) lament that a great deal of scale development documented in the marketing literature wrongly uses PCA, and speculate that this may be because it is the default extraction option within SPSS.

Whichever method is chosen, and whether the tool is made up of one or a number of scales, the broad approach is the same: The structure of the data is examined to determine the number of factors that underlie it. There can be something of an art to determining the correct number of factors to extract, and traditional methods, such as the Kaiser Criterion (examining the number of factors with eigenvalues above 1.0), or examining the point of inflexion on the scree plot (Cattell, 1966), can provide wildly different solutions. The most robust way of determining the most likely number of factors underlying your data is to conduct **parallel analysis** (e.g. O'Connor, 2000). Parallel analysis works by simulating a random dataset that contains the same number of variables (i.e. items) and cases (i.e. participants) as the data set under examination. The eigenvalues for the random dataset are computed and are then compared to those of the factors extracted from the real data. The point at which the eigenvalues of the simulated data become larger than those of the real data indicates that no further meaningful factors underlie the data. Clearly, any indication that the number of factors underlying the data differs from the number of constructs the tool is intended to measure suggests that there is a need to examine the items more deeply to identify those that are problematic.

Assuming that the previous two analyses have uncovered some room for improvement in either the internal consistency or structural validity of your tool, the next step is to identify which items are causing the problem. Indeed, it is very unlikely that the tool cannot be improved in some way with the removal of specific items, so the researcher should expect to have to eliminate some items at this stage, highlighting the necessity for generating more items than are required for the final scale. The process by which items are identified as problematic is through use of the previous two statistical techniques, though their focus is somewhat different to that described above. The process revolves around four statistical checks for each item, upon failure of any of which the item is flagged as problematic. Each scale within the tool is examined separately for these analyses. In the first check, Cronbach's α is computed for the scale, along with the value of α should each individual item be deleted. Any item that shows a substantial increase in the value of Cronbach's α means that that item should be flagged for deletion. The most problematic item is removed, and α is computed for the remaining scale items. As before, if any item or items appear problematic, the worst offender is removed. This process repeats until no further substantial increase in α can be achieved.

The remaining checks are conducted using exploratory factor analysis techniques. In contrast to the guidance on factor analysis provided above, in these analyses, orthogonal rotation such as varimax rotation should be used in order to achieve maximum differentiation between the construct of interest represented by the primary extracted factor and any nuisance factor (Hinkin, 2005). In the second check, a two-factor solution is forced. As the aim is for a unifactorial scale, any item that loads substantially (i.e. with a factor loading greater than .30; Hair et al., 1998) upon the second, nuisance factor is flagged. As before, the analysis is then rerun with this item removed, and the loadings of the remaining items examined. This process continues until no further items are shown to load substantially upon the second extracted factor. For the third and fourth checks, factor analysis is again conducted, but this time forcing a one-factor solution. Here, any item that shows either a low loading on the single extracted factor (less than .3; Hair et al., 1998), or shows low communality with all the other items in the scale (less than .3) is flagged for deletion. Having removed the worst performing item, this analysis is repeated as before until no further items remain.

Once these checks have been completed, the researcher is then faced with the decision of which items to retain and which items to delete from the final scale. The number of flags obtained by each item will largely inform this decision, but may also be influenced by the nature of the tool. If the number of items required for the final scale is small relative to the number of items in the analyses, the researcher can afford to be more choosy, and can impose stricter criteria for inclusion in the final scale. However, Lee and Hooley (2005) urge caution when removing items on the basis of low communality, so any decision made on the basis of this criterion should be made in conjunction with evidence provided as part of the other checks. Once the decision has been made of which items to retain, Cronbach's α is computed and factor analysis is conducted (preferably supported by parallel analysis) to check that the final item set demonstrates acceptable internal consistency and is acceptably unifactorial.

Step Six: Second item trial

At this point, the researcher has designed and developed a scale that demonstrates (assuming that the previous steps have been at least partially successful) good psychometric properties, in that it reliably measures a single construct. However, the process of scale development is not complete at this stage. A key question remains over whether the construct that it measured by the scale is, in actual fact, the construct it is intended to measure. To address this, it is necessary to examine the scale's convergent and discriminant validity, which requires the identification of suitable other published psychometric measures. In order to explore convergent validity, a scale needs to be identified that measures the same construct. This shouldn't be a problem if the deductive approach has been taken, as there should be an abundance of scales that exist that measure the construct of interest. However, this can be challenging if the inductive approach has been taken. For discriminant validity, an existing scale needs to be identified that measures a construct different to the one that is the focus of the scale. If a personality measure is being designed, a useful resource to help in this process is the International Personality Item Pool (Goldberg et al., 2006). IPIP is a huge repository of robust, published, free-to-use personality scales that covers a very wide range of constructs.

At this stage, it is also a sensible idea to assess the scale's criterion-related validity to establish whether the scale is predictive of important outcomes, such as consumer behaviour.

As such, some sort of data should be collected to aid in this. The precise nature of the criterion data gathered will depend upon the nature of the scale and its intended use. Some examples of the kind of criterion data that is typically collected for the validation of scales within consumer psychology are things such as the prediction of brand associations (e.g. Berry, 2000), consumer decision making behaviour (e.g. Kidwell, Hardesty & Childers, 2007), and purchasing decisions (e.g. Nenkov et al., 2009).

Once the measures that will be used for this round of analyses have been identified, they are administered alongside the remaining items within the scale to a new set of participants. The number of participants required for this step again varies according to the size of the scale, but is determined largely by the number of cases needed for model identification in the analyses based on **confirmatory factor analysis** (CFA). Unlike for EFA, there is little clear guidance on this in the literature, as model identification is dependent on the specific nature of the model, the number of indicators (i.e. items), and the number of latent factors specified. Attempts to provide approximations of number of cases needed for model identification have been found to vary wildly (e.g. Wolf et al., 2013). Suffice to say, if the specified model has any degree of complexity, it is likely to require a sample size in the hundreds, or even the thousands for very complex models.

The scale's internal consistency and structural validity is first checked using Cronbach's α and factor analysis. The results of these are likely to be slightly different to those obtained in the previous step, but this is not a cause for concern as long as the scale still demonstrates acceptable reliability and validity. Once this has been checked, CFA is conducted on the scale items, specifying a single latent factor onto which each item is made to load. As for EFA in the previous step, the estimator used for CFA will vary according to the nature of the data. In the case of ordinal data such as Likert-type items, a robust diagonally-weighted least squares estimator is preferable (Schweizer & DiStefano, 2016), assuming that the statistical package you are using to conduct CFA supports computation of polychoric correlations (see Software section). The CFA model's fit indices are then examined to determine whether or not the data are acceptably unifactorial. The best fit indices to use for this purpose are, again, a source of some debate in the literature. The most commonly used measures of CFA model fit are shown in Table 2, along with recommended values by which to judge the adequacy of model fit based on guidance from Hu and Bentler (1999). If model fit is poor, special attention should be paid to the model's **modification indices**. These provide an indication of the improvement to the model if its structure is changed in specific ways. This is very helpful when conducting CFA on the items within a multifactorial measure, as it allows the researcher to examine model fit when item cross-loadings (i.e. when an item loads upon more than one construct) are taken into account.

Table 2.

Common measures of fit used in CFA (from Hu & Bentler, 1999)

Measure	Values for quality of model fit
Chi Squared (χ^2)	Extremely variable. Lower values indicate better model fit.
Comparative Fit Index (CFI)	.9 = acceptable .95 = good
Tucker-Lewis Index (TLI)	.9 = acceptable .95 = good
Root Mean Squared Error of Approximation (RMSEA)	.05/.06 = good .08 = acceptable >.12 = poor
Standardized Root Mean Squared Residual (SRMR)	.08 = good

Once the factor structure of the tool has been confirmed, convergent and discriminant validity can be examined. In the past, the common procedure for examining convergent and discriminant validity was to generate large matrices of Pearson correlations between the scale items and those within the additional existing scales that were identified prior to this round of data collection. This correlation matrix would then be examined to establish patterns of intercorrelations that suggested convergent and discriminant validity (i.e. strong correlations between items designed to measure the same construct and weak or no correlations between items designed to measure different constructs). However, modern approaches to investigating convergent and discriminant validity are somewhat more robust that this, drawing upon CFA procedures. In this approach, competing CFA models are examined and their fit indices compared. For convergent validity, a model is specified in which the scale's items and those from the existing scale designed to measure the same construct are all made to load onto a single latent factor. In the competing model, the scale's items are made to load onto a single latent factor, and the other scale's items onto a second latent factor, which is correlated with the first. To demonstrate acceptable convergent validity, it should be the case that the first model should be a better fit to the data than the second. The procedure for the examination of discriminant validity using CFA is very similar. In this case, the model in which the scale's items and those within the existing scale designed to measure a separate construct are made to load onto separate, correlated latent factors should be a better fit to the data than that in which both scale's items are made to load onto a single latent factor.

The examination of criterion-related validity is, statistically speaking, much more straightforward than the previous set of analyses. To establish criterion-related validity, scale scores are first computed for each participant by adding their scores on each item. Pearson correlations are then computed between these scale scores and the criterion data that has been collected. These correlation coefficients can then be judged according to their strength. Though the strength of these relationships can vary substantially according to the nature of the specific criterion used, general rules exist within the literature for the judgement of the adequacy of evidence to support criterion-related validity. Coefficients of .50 or above are considered to be excellent evidence for a measure's criterion-related validity, correlations above .35 to be good evidence, above .2 to be adequate, and .2 and below to be inadequate (Hemphill, 2003).

Step Seven: Norming the final tool and developing scoring procedures

If all has gone to plan, the researcher will have, at this point, a robust tool that measures the construct for which it was intended. The final step of the process it to norm the tool. The process of norming the tool adds meaning to specific scale scores, allowing users of the tool to understand the behavioural implications for consumers based on how they respond to the each scale's items. Developing a norm for each scale it is a relatively straightforward process. From the two item trialling phases in the previous two steps, there should be plenty of data that the researcher can use to construct a norm group, against which future test takers' scores can then be compared. Most statistical packages feature an option that allows the researcher to generate the percentile score that corresponds with each participant's scale score. Alternatively, sten scores for each participant can be calculated by first computing their z-score, then by applying the formula below.

$$Sten = 2z + 5.5$$

This will give each participant in the norm group a sten score between 1 and 10. Sten ranges can then be banded as being representative of the below average range of the trait (stens 1 to 3), the average range (4 to 7) or the above average range (stens 8 to 10). The researcher can then attached narratives to each of these ranges that can be used to make inferences about a consumer’s likely preferences for behaviour on the basis of their score.

Further Issues in Scale Development

Measurement bias and its detection

A frequently neglected issue within scale development is that of measurement bias. Measurement bias, in its simplest form, occurs when a scale functions (i.e. measures the construct of interest) differently for different groups of test takers. The overwhelming majority of psychological research has been conducted on participants from Western, educated, industrialised, rich, and democratic societies (so-called WEIRD participants; Henrich, Heine, & Norenzayan, 2010), and scale development is no exception. It would be naïve to assume that all psychometric scales functioned in exactly the same way for all groups of participants, particularly those from different countries with different cultural artefacts and values. Therefore, it is prudent to consider the impact that any possible measurement bias might have on research findings when using scales in different cultural contexts.

Measurement bias is most frequently assessed using techniques based on IRT methodology (see above). The most common set of techniques aim to detect **differential item functioning** (DIF), which contributes to **differential test functioning** (DTF), an antecedent of measurement bias. Though these techniques can be extremely useful in establishing the validity of a piece of research that has been conducted in a context dissimilar to the one in which a psychometric it uses was developed, they are all either very time consuming, mathematically complex, or, most frequently, both. Most DIF/DTF procedures require specialist statistical software to run (see section on software below).

Socially Desirable Responding Behaviour

As with any self-report measure in psychology, an ever-present issue is that of socially desirable responding (SDR) behaviour. Previously known as faking, SDR occurs in instances in which it is tempting for test takers to present a more favourable portrayal of themselves than might actually be the case. While it is seen as a more serious problem for the use of psychometrics when they are used in high-stakes situations such as job selection (Hogan, Barrett & Hogan, 2007), SDR is, nevertheless, an issue for the use of psychometrics in lower-stakes situations, such as those that are the focus of consumer psychology research.

There are a number of solutions that have been formulated to address SDR behaviour. One somewhat contentious approach is the inclusion within the tool of an inbuilt SDR scale. SDR scale items are those that it would be very unlikely that a participant would endorse were they not trying to appear more socially desirable than was the case. For example, an item within an SDR scale might read *"I have never been late for an appointment"*. Punctuality is a socially desirable trait to have. However, almost without exception, everyone who has reached adulthood has been late for at least one appointment in their lifetime, so it is extremely unlikely that a test taker would endorse this if they were being completely honest in their responses. It should be clear, then, that as the number of items endorsed within an SDR scale increases, the probability that the test taker is responding genuinely becomes vanishingly small.

The concept of SDR scales sits uneasily with many researchers and practitioners. Many see such scales – quite rightly – as a form of deception, and, therefore, as being unethical to use. A somewhat more palatable approach to addressing SDR is for psychometrics to establish an

honesty contract within their instructions (e.g. Bartram, 2009). An honesty contract describes the necessity for honest responding from the test taker, and informs them that the authenticity of their responses can be checked, the hope being that they will be compelled to answer honestly.

Whichever of these approaches seems more appropriate, the issue of SDR behaviour is one that should not be ignored. SDR necessarily impacts upon the validity of measurements provided by psychometric scales (Hogan et al., 2007), thus eroding the validity of research findings that depend upon them.

Translating Scales into Other Languages

The final issue to be discussed on the subject of scale development is something of a special case, and one with which many researchers will never be faced. The translation of scales may seem like a relatively straightforward process, but it is fraught with difficulty. One cannot simply translate existing scale items and expect the whole scale's meaning to be preserved, as a sentence is more than the sum of its parts.

One common approach to ensuring that meaning is preserved when a scale is translated is the process of **translation-backtranslation** (van der Vijver & Leung, 1997). In this approach, a scale's items are first translated into the language intended for the final tool. The items are then translated again by a second researcher back into their original language. The initial item set and these translated-backtranslated items are then compared to ensure that the meaning of each item is preserved. Harkness (2003) takes this process a step further, describing a five-step process for the translation of scales according to best practice. This process hinges upon three key roles, each of which should be fulfilled by separate researchers. In the first step, the items are translated by a *translator*. Next, a *reviewer* produces several different versions of the translated items. Following this, an *adjudicator* selects which of these versions to use for each item in the translated version of the tool. Once this has been completed, the scale undergoes pretesting. Finally, any problems with specific items and decisions about solutions are documented. As should be clear, the translation of scales is an undertaking that should not be entered into lightly. That said, translating scales allows consumer psychologists to gain a much deeper understanding of consumers within different cultures in the global marketplace.

Troubleshooting Tips

The scale design and development process is rarely one that goes entirely according to plan. As such, there are myriad roadblocks that a researcher may face. It is important to understand what the issue is and how best to address it when these problems arise. To that end, the authors will attempt to address some of the more common problems encountered.

My CFA model is not identified

There are a couple of reasons why this might be the case. It may be that your model is misspecified, or it may lack a sufficient number of cases for identification. Try constraining additional model parameters first, as this may rectify the problem. If not, it may be that two or more of your latent factors which you thought to be orthogonal are actually correlated, in which case you should adjust the model accordingly. If neither of these approaches work, it may simply be the case that you need to collect more data.

When assessing content validity, my value of Fleiss' Kappa is very low

This generally indicates poor agreement between your raters, but it is not necessarily the case. Fleiss' Kappa assesses the degree to which your raters agree on specific scale numbers. As such, if two raters rated an item as being related to your construct of interest, one providing a rating of 4 and the other of 5, Fleiss' Kappa would treat this as disagreement. Clearly, this is nonsensical. In this situation, try collapsing the points on your rating scale into a smaller number, incorporating, for example, points 4 and 5 into an overall rating. This should address much – if not all – of the issue.

In EFA, I cannot get to the point at which my scale's items are unifactorial

Unfortunately, if you've followed the process described in Step 5 and you're still having problems, it may be an issue with either the wording of your items, the construct definition, or both. Scrutinise your items and your construct closely. Does it look like your construct might reflect two (or more) related-yet-different constructs? Do your items all adequately sample the construct? If the answer to either of these is yes, you may need to return to the earlier stages of the process.

I don't have enough items in my initial item set, but I can't think of any more

Try playing around with the phrasing of words, and try restating some of the existing items you have so that they focus on slightly different aspects of the construct. Perhaps also try writing some negatively-keyed versions of the items you have. If you're really stuck, it is perfectly allowable to consult some SMEs to ask for their input to the process. Scale design needn't be a solitary pursuit.

Software

In this section, the authors will discuss some of the software packages available to aid in the scale development process. While some of these packages will be familiar to the reader, it may be the case that they are not aware of some of the shortcomings of particular software packages. Any researcher wishing to carry out a scale development project should be aware of these pitfalls, as they have the potential to influence the results of the analyses carried out in the process of scale development.

A statistical package that will be familiar to most – if not all – students and researchers in psychology is SPSS (IBM, 2016). For the most part, SPSS is able to conduct the analyses relevant to scale development. SPSS will allow you to compute Cronbach's α to examine a scale's internal consistency, to run PAF and/or another form of EFA to test its structural validity, and to compute the correlations used to explore its criterion-related validity and – at least to an extent – its convergent and discriminant validity. However, SPSS is not without its shortcomings. The first, and most striking, of these is that it is unable to run CFA for the purposes of confirmation of factor structure in the second item trialling phase of scale development, or for the more modern approaches to establishing convergent and divergent validity. To achieve this, a more specialist statistical package such as those described below is required. A more serious shortcoming – at least from an academic point of view – is in the way in which SPSS treats categorical data within PAF and other dimension reduction procedures. Even though SPSS allows the user to manually flag variables as being nominal, ordinal, or scale data, the algorithms it uses to produce its solution treats all variables as scale-level (Basto & Pereira, 2012). As such, any dimension reduction procedures run in SPSS are based on Pearson correlations, as opposed to the tetrachoric correlations suitable for dichotomous data, or polychoric correlations for ordinal data. Holgado–Tello et al. (2010) have demonstrated that using Pearson correlations in instances such as these tends to produce less accurate factor solutions than with polychoric or tetrachoric correlations. However, one potential way to address some of these shortcomings is through the use of a free-to-use plugin that allows the use of the R programming language (The R Foundation, n.d.) within SPSS. Although fairly complex for the average user to negotiate, R allows users to conduct analysis such as CFA in SPSS, and for tetrachoric and polychoric correlations to be computed for dichotomous and ordinal data respectively.

SPSS Amos (Analysis of moment structures; Arbuckle, 2016), is a statistical software package that is able to perform CFA and Structural Equation Modelling path analyses. The key advantage of Amos, particularly for those who are relatively experienced in conducting CFA, is that it includes a graphical interface that allows the user to draw out their structural models. This is a hugely attractive quality, and one that sets it apart from other CFA software. However, although a powerful statistical programme with a user-friendly interface, one of the shortcoming of Amos is that it cannot generate tetrachoric or polychoric correlations for CFA.

One alternative to both SPSS and Amos for the more experienced researcher is Mplus (Muthén & Muthén, 2017). Mplus is a very powerful and versatile statistical package that is able to run CFA as well as all the other procedures that SPSS will do. The key advantage of Mplus is that it allows the user to define which variables within a dataset are categorical variables. Having done so, it will then base subsequent analyses on polychoric and/or tetrachoric correlations, depending on the nature of the variables defined as categorical. In addition to this, a set of add-ons is available for the basic Mplus software package that allow it to be used to run IRT-based analyses, such as to detect DIF/DTF for the exploration of possible measurement bias, or for the generation of item and test information curves to aid in the item trialling process. The disadvantage of Mplus is that it is based largely on syntax, so can be off-putting for students who are used to SPSS' menu system or the graphical system used by Amos.

One little-known statistical program that is of particular note for scale development is FACTOR (Lorenzo-Seva & Ferrando, 2006). FACTOR offers a number of alternative procedures for parallel analysis to the traditional one described by O'Connor (2000). In particular, traditional methods tend not to handle items with dichotomous response formats particularly well (Tran & Formann, 2009). FACTOR is able to run parallel analysis based on tetrachoric correlations such as minimum rank factor analysis (PA-MRFA), which provides much more reliable indications of the likely number of factors underlying an item set (Timmerman & Lorenzo-Seva, 2011). FACTOR is freeware, so it costs nothing to download and use.

C-OAR-SE: A Critique of the Psychometric Approach

One relatively recent response to the psychometric approach to scale development within marketing that has received considerable attention comes in the form of C-OAR-SE (Rossiter, 2002; 2011). The C-OAR-SE method revolves around the principle that the only real scale development of any value is that of the initial establishment of content validity. C-OAR-SE itself is an acronym that reflects the six stages of scale development it recommends, namely construct definition, object representation, attribute classification, rater-entity identification, scale selection, and enumeration (Rossiter, 2011). Rossiter reasons that, if the content validity of a scale's items and response scale is effectively established, then there is no need to explore construct validity, nor criterion-related validity.

As one might expect, several critiques have been made by leading psychometricians and experts in marketing research of C-OAR-SE and of Rossiter's assumptions in developing it (Lee & Cadogan, 2016; Ahuvia, Bagozzi & Batra, 2013; Rigdon et al., 2011). Most of the criticisms levelled at C-OAR-SE say that, while encouraging good content validation practices is to be applauded, the wholesale rejection of the subsequent steps in the scale development process sets a dangerous precedent, one that has the potential to return scale development in consumer psychology and related fields to the dark ages described by Churchill (1979). Rather unsurprisingly, as psychometricians, the authors agree with this assessment.

Summary

The field of consumer psychology has made extensive use of tools designed according to the psychometric approach. They have allowed consumer psychologists to gain insights into the nature of consumers and their behaviour that could not be uncovered by any other means as accurately or effectively. While the psychometric approach within this area has had its critics in recent years, it remains the dominant model for scale development within the field. The scale development process described in this chapter is likely to continue to be used for the foreseeable future for the development of new measures, which will allow future researchers to measure new constructs in the most robust way available to them.

Exercises

Exercise One: Item Generation and Content Validity

Think about a psychological construct that is well understood and well researched. Search the literature for a brief, clear definition of this construct. Try writing ten items that are designed to tap into this construct, then investigate your item set's content validity by asking five friends or family members to rate – from 1 to 5 – the degree to which each item taps into the construct as you have defined it. Do any items look as though they might not be suitable to be included in a scale to measure this construct?

Exercise Two: Rigorous research design

Advertisers need to find out the potential problems they are facing when they try to sell what could be perceived to be an embarrassing product. What process would you choose in order to find out about the product's level of embarrassability and whether this has an impact on the consumer's willingness to buy it? How could you ensure that the way you gather the initial information and beliefs around this product is systematic and has rigor? How could you ensure the items you build to measure this phenomenon are psychometrically sound?

References

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research, 34*(3), 347-356. doi: 10.2307/3151897

Albert, N., & Valette-Florence, P. (2010). Measuring the love feeling for a brand using interpersonal love items. *Journal of Marketing Development and Competitiveness, 5*(1), 57–63.

Arbuckle, J. (2016). IBM® SPSS® Amos™ 24 User's Guide. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/amos/Manuals/IBM_SPSS_Amos_User_Guide.pdf.

Rigdon, E. E., Preacher, K. J., Lee, N., Howell, R. D., Franke, G. R., & Borsboom, D. (2011). Avoiding measurement dogma: a response to Rossiter. *European Journal of Marketing, 45*(11/12), 1589–1600.

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, *43*(3), 335–344. doi:10.1016/j.jrp.2008.12.013

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 102−116. doi: 10.1080/10705511.2014.859510

Bartram, D. (2009). The international test commission guidelines on computer-based and internet-delivered testing. *Industrial and Organizational Psychology*, *2*(1), 11–13.

Basto, M., & Pereira, J. M. (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software*, *46*(4), 1–29.

Bauer, H. H., Reichardt, T., Barnes, S. J., & Neumann, M. M. (2005). Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study. J*ournal of Electronic Commerce Research*, *6*(3), 181–191.

Berry, L. (2000). Cultivating service brand equity. J*ournal of the Academy of Marketing Science*, *28*, 128–137. doi: 10.1177/0092070300281012

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*(3), 214–227. doi: 10.1037/0003-066X.39.3.214

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.doi:10.1207/s15327906mbr0102_10

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16,* 64–73. doi: 10.2307/3150876

Cohen, J.B., Pham, M.T., & Andrade, E. B. (2008). The nature and role of affect in consumer judgement and decision making. (pp.297-348). In C.P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology*. Mahwah, NJ: Erlbaum.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1–9.

DeVellis, R. F. (2016). *Scale development: Theory and applications* (Vol. 26). London: Sage.

Gattol, V., Sääksjärvi, M., & Carbon, C. C. (2011). Extending the implicit association test (IAT): assessing consumer attitudes based on multi-dimensional implicit associations. *PLoS ONE, 6*(1), e15849. doi:10.1371/journal.pone.0015849.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory*. Hawthorne, NY: Aldine Publishing Company.

Gibbert, M., & Ruigrok, W. (2010). The "what" and "how" of case study rigor: Three strategies based on published work. *Organizational Research Methods, 13*(4), 710–737. doi: 10.1177/1094428109351319

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*(1), 84–96.doi:10.1016/j.jrp.2005.08.007

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998). *Multivariate data analysis* (5th Edition). London: Prentice-Hall.

Harkness, J. A. (2003). Questionnaire translation. *Cross-cultural Survey Methods, 1*, 35–56.

Heath, H., & Cowley, S. (2004). Developing a grounded theory approach: a comparison of Glaser and Strauss. *International Journal of Nursing Studies, 41*(2), 141–150. doi: 10.1016/S0020-7489(03)00113-5

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78–80. doi10.1037/0003-066X.58.1.78

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*(2-3), 111–135. doi: 10.1017/S0140525X10000725

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological bulletin, 74(*3), 167–184. doi:10.1037/h0029780

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*(5), 967–988. doi: /10.1177/014920639502100509

Hinkin, T. R. (2005). Scale development principles and practices. In R. A. Swanson & E. F. Holton III (Eds.) *Research in organizations: Foundations and methods of inquiry*. San Francisco, CA: Berrett-Koeller Publishers Inc.

Hofmann, J., Platt, T., Ruch, W., & Proyer, R. T. (2015). Individual Differences in Gelotophobia predict responses to joy and contempt. *Sage Open*, 1–12. doi: 10.1177/2158244015581191

Hogan, R. (1995). *Hogan personality inventory*. Hogan Assessment Systems.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*(5), 1270–1285.doi:10.1037/0021-9010.92.5.1270

Holgado–Tello, F. P., Chacón–Moscoso, S., Barbero–García, I. & Vila–Abad, E. (2010). Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables. *Quality & Quantity, 44(*1), 153–166. doi: 10.1007/s11135-008-9190-y

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55. doi: 10.1080/10705519909540118

IBM (2016). IBM SPSS Statistics 24 Documentation. Retrieved from http://www-01.ibm.com/support/docview.wss?uid=swg27047033.

Jacoby, J., & Kyner, D. B. (1973). Brand loyalty vs. repeat purchasing behavior. *Journal of Marketing Research, 10*(1), 1–9. doi: 10.2307/3149402

Kidwell, B., Hardesty, D. M., & Childers, T. L. (2007). Consumer emotional intelligence: Conceptualization, measurement, and the prediction of consumer decision making. *Journal of Consumer Research, 35*(1), 154–166. doi: 10.1086/524417

Klein, J.G., Ettenson, R., & Krishnan, B.C. (2005). Extending the construct of consumer ethnocentrism: When foreign products are preferred. *International Marketing Review, 23*(3), 304–321. doi: 10.1108/02651330610670460

Kurz, R., & MacIver, R. (2008). Coaching with Saville Consulting Wave™. In J. Passmore (Ed.) *Psychometrics in coaching: Using psychological and psychometric tools for development.* London: Kogan Page.

Lee, N., & Cadogan, J. (2016). Welcome to the desert of the real: Reality, realism, measurement, and C-OAR-SE. *European Journal of Marketing, 50*(11), 1959–1968. doi: 10.1108/EJM-10-2016-0549

Lee, N., & Hooley, G. (2005). The evolution of "classical mythology" within marketing measure development. *European Journal of Marketing, 39*(3/4), 365–385. doi:10.1108/03090560510581827

Lewis, C., & Anderson, P. (1998). *PAPI technical manual.* London: Cubiks.

Lin, C. F. (2002). Segmenting customer brand preference: demographic or psychographic. *Journal of Product & Brand Management, 11(*4), 249–268. doi:10.1108/10610420210435443

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88–91. doi: 10.3758/BF03192753

Maraz, A., Eisinger, A., Hende, Urbán, R., Paksi, B., Kun, B., Kökönyei, G., Griffiths, M.D. & Demetrovics, Z. (2015). Measuring compulsive buying behaviour: Psychometric validity of three different scales and prevalence in the general population and in shopping centres. *Psychiatry Research, 225,* 326–334. doi: 10.1016/j.psychres.2014.11.080

McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*(3), 261–270.doi:10.1207/s15327752jpa8403_05

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide: Statistical analysis with latent variables: User's guide* (8th Ed.). Los Angeles, CA: Muthén & Muthén.

Nenkov, G. Y., Morrin, M., Ward, A., Schwartz, B., & Hulland, J. (2009). Re-examination of maximization: psychometric assessment and derivation of a short form of the maximization scale. *ACR North American Advances, 36,* 734–735. doi:10.1177/0013164413495237

Netemeyer, R. G., Durvasula, S., & Lichtenstein, D. R. (1991). A cross-national assessment of the reliability and validity of the CETSCALE. *Journal of Marketing Research, 320–327*. doi:10.2307/3172867

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. doi 10.1002/j.2333-8504.1965.tb00132.x

Nunnally, J. (1978). *Psychometric theory* (2nd Ed.). New York : McGraw-Hill.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers, 32,* 396–402. doi: 10.3758/BF03200807

Price, J. L., & Mueller, C.W. (1986). *Handbook of organizational measurement.* Marshfield, MA: Pitman.

The R Foundation (n.d.). What is R?. Retrieved from https://www.r-project.org/about.html.

Rigdon, E. E., Preacher, K. J., Lee, N., Howell, R. D., Franke, G. R., & Borsboom, D. (2011). Avoiding measurement dogma: a response to Rossiter. *European Journal of Marketing, 45*(11/12), 1589–1600. doi:10.1108/03090561111167306

Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing, 19*(4), 305–335.doi: 10.1016/S0167-8116(02)00097-6

Rossiter, J. R. (2011). Marketing measurement revolution: The C-OAR-SE method and why it must replace psychometrics. *European Journal of Marketing, 45(*11/12), 1561–1588. doi: 10.1108/03090561111167298

Ruch, W., Hofmann, J., & Platt, T. (2015). Individual differences in gelotophobia and responses to laughter-eliciting emotions. *Personality and Individual Differences, 72,* 117–121. doi: 10.1016/j.paid.2014.08.034

Schmitt, B. (1999). Experiential marketing. *Journal of Marketing Management, 15(*1-3), 53–67. doi: 10.1362/026725799784870496

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*(5), 1178–1197. doi: 10.1037//0022-3514.83.5.1178

Schweizer, K., & DiStefano, C. (2016). *Principles and methods of test construction.* Oxford: Hogrefe.

Söderlund, M., & Rosengren, S. (2008). Revisiting the smiling service worker and customer satisfaction. *International Journal of Service Industry Management, 19*(5), 552–574. doi: 10.1108/09564230810903460

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209–220. doi: 10.1037/a0023353

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement, 69*, 50–61. doi: 10.1177/0013164408318761

van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Thousand Oaks, CA: Sage.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913–934. doi:10.1177/0013164413495237