

## Chapter 2

# Norms for Test Construction

**Abstract** In the subsequent chapters, different scales have been developed and validated on Indian population. The current chapter aims to give an overview of various norms that were followed while constructing the scales. In literature, several methods for scale development and validation exist however, in this chapter only those methods are discussed more, which have been used in the later chapters. The current chapter gives an overview of definition of psychological test, steps of test construction, norms for sample size, preliminary data analysis, exploratory factor analysis, confirmatory factor analysis, etc.

**Keywords** Psychological test • Test construction • Norms • Preliminary data analysis • Exploratory factor analysis • Confirmatory factor analysis

## Introduction

### *What Is a Psychological Test?*

Psychological test is a set of objective and standardized self-report questions whose responses are then scored and aggregated to attain a composite score (Zumbo et al. 2002). The main components of a psychological test are (i) series statements to which an participant responds and (ii) a composite score that arises from scoring of statements that can be obtained either as binary scores that are dichotomous in nature or on a Likert type scale with grading statements, for example five points such as strongly agree to strongly disagree. The items in a test are indicators of the phenomenon under study and hence a composite score is also an indicator of the phenomenon and not the phenomenon itself (Zumbo et al. 2002).

## ***Test Construction***

Psychological tests are often subject-centered measurements and follow certain strict guidelines for construction, administration, scoring and interpretation. The main goal of developing a new scale is to create a valid and reliable measure of an existing construct. The following steps are followed during construction of a test (Crocker and Algina 1986):

1. Identification of primary purpose for which the test scores will be used
2. Identify behaviors that represent the construct or define the domain
3. Prepare a set of test specification, delineating the proportion of items that should focus on each type of behavior identified in step 2
4. Construct an initial pool of items
5. Review of items
6. Pilot test of the revised items
7. Modification of items (if any from the pilot study)
8. Field test the items on a large sample representative of the examinee population for whom the test is intended
9. Determine statistical properties of item scores and when appropriate, elimination of items that do not meet pre-established criteria
10. Design and conduct reliability and validity studies for the final form of the test
11. Develop guidelines for administration, scoring, and interpretation of test scores (Matlock-Hetzel 1997).

In the forthcoming chapters on positive psychology scales, the above guidelines were followed. An in-depth and exhaustive literature review was undertaken with respect to test construction and validation of each construct. Based on literature, an exhaustive list of domains was prepared followed by item pool generation under each domain. Then the content validity of the pooled items was established with the help of five experts who held doctoral degrees in Psychology and were well versed with scale construction. They independently reviewed the items in the context of their clarity, readability level and their suitability for the purpose on a four point rating scale with 1 as least relevant to 4 as most relevant. The items that were rated by all experts as relevant (3) or most relevant (4) were retained. The remaining items were deleted. After content validity, data were collected and then scrutinized for item statistical properties. At this stage too few items wherever required were deleted (Visser et al. 2000; Steger et al. 2006).

Exploratory factor analysis (EFA) was employed with an aim to reduce the number of items and retain most relevant items only. Post EFA, data were collected again on a different set of sample. The second set of data were split into one-third and two-third based on recommended methodologies (Guadagnoli and Velicer 1988; MacCallum et al. 1996). On one-third data, EFA was employed whereas on two-third data confirmatory factor analysis (CFA) was employed. The confirmed factor validation structure resulted in the final list of domains and items.

For example, construction of resilience scale was undertaken after exhaustive literature review of various established resilience scales such as The Adolescent Resilience Scale (ARS, Oshio et al. 2003), Resilience Scale for Adults (RSA, Friborg et al. 2003), Connor-Davidson Rating Scale (CD-RISC, Connor and Davidson 2003), Brief Resilience Scale (BRS, Smith et al. 2008), The Resilience Scale (RS, Wagnild and Young 1993), Dispositional Resilience Scale Revised 15 (DRS-R15, Bartone et al. 1989) and so on. The existing scales were reviewed for their factor structure consistency and validity and internal consistency. Later, item pool was generated by developing new items and items from existing scales which were subjected to content validity. The pooled items were rated by subject experts on parameters of relevance, cultural fairness, etc. According to the reviewer's suggestions during content validity, some more new items that were culturally relevant were added. Data were collected on the exhaustive list of items and later subjected to item analysis and EFA with an aim to reduce items and explore factor structure. This was followed by data collection on retained item list in the next phase and once again items analysis; EFA as well as CFA were employed. The confirmed factor structure items were then translated into Hindi and the CFA was employed to confirm the factor validation on Hindi version. Same procedure was followed for all constructs of interest in this book.

It is observed that once established scales show generally acceptable norms during revalidation except factor solution. Often researchers develop a pool of test items from existing scales and modify a few items. However, at times the scales lack rigorous validation process which has become the call of the hour in test construction. EFA and CFA are the cynosure of test validation. Almost all researchers use these techniques for analysis. However, several problems arise with using and reporting these techniques. It is necessary to focus on few parameters such as creation of item pool, basic principles of item writing, choice of format, sample size for validation, norms for EFA and CFA. Few other guidelines too were followed while developing the scales in the next chapters as outlined by Clark and Watson (1995).

### ***Principles of Item Writing***

Writing good items is a precursor to developing a good psychometric test. A researcher needs to do a thorough literature survey before commencing item writing (e.g., Angleitner and Wiggins 1985; Comrey 1988; Kline 1986: cited in Clark and Watson 1995). According to Clark and Watson (1995) the language should be unpretentious, upfront, suitable and simple in nature. The language should also be of the reading level of target population. A researcher should avoid using trendy expressions, idioms, other language forms that vary widely with age, ethnicity, region, gender, and items that virtually everyone misinterprets (e.g., "Sometimes I am better-off than at other times") or no one (e.g., "I am always energetic")

unless they are intended to assess invalid responding, avoid using complex or double-barrelled items (Clark and Watson 1995).

In the current book chapters, utmost care was taken by authors while item writing for the scales. All the items were evaluated by reviewers for the above discussed precursors for good item writing. All the above mentioned points were followed during developing items for constructs of interest in the next chapters.

### ***Format of Items***

The test developer also has to choose the format in which the items would be written. The two dominant response formats are dichotomous responding (e.g., true–false and yes–no) and Likert type rating scales with three or more options. Checklists, forced-choice, and visual analogy measures also have been used over the years, but for various reasons are not in favor (Clark and Watson 1995).

In the forthcoming chapters, the scales were designed as five point Likert type rating scales. The authors decided this format for construction based on literature review of the constructs and advantages of Likert type of scale.

The central goal of this stage is to represent thoroughly all content that is potentially relevant to the target construct. Loevinger (1957) stated that “The items of the pool should be chosen so as to sample all possible contents which might comprise the putative trait according to all known alternative theories of the trait” (p. 659). The two key propositions of Loevinger (1957) statement are (a) the initial pool of items should be extensive and more widespread than the theoretical view of the target construct and (b) should include content that ultimately will be shown to be tangential or even unrelated to the core construct. The aim should be that the resultant psychometric analyses can identify weak, unrelated items that should be dropped (Watson 2012) from the emerging scale. While creating an item pool inclusiveness is always better than excluding items related to any aspect of the construct (Clark and Watson 1995; Watson 2012).

It is also important that the scale developer must include an adequate sample of items within each of the major content areas comprising the broadly conceptualized domain (Clark and Watson 1995; Watson 2012). If one fails to do so then there is a chance of underrepresentation of items in the final scale. To ensure that each important aspect of the construct is assessed adequately, it is recommended that formal subscales be created to assess each major content area (Watson 2012). Loevinger (1957) recommended that the proportion of items dedicated to each content area should be proportional to the importance of that aspect in the target construct (Clark and Watson 1995).

Most researchers employ deductive method for scale construction. Good scale construction process involves several periods of item writing and conceptual and psychometric analysis (de Barros 2014). The psychometric analyses sharpen the understanding of the nature and structure of the target domain and also aid to

identify deficiencies in the initial item pool. For example, if factor analysis suggests that scale can be further divided into several subscales, but at item generation stage enough similar items were not pooled then reliability of the pooled items cannot be assessed (Clark and Watson 1995). Hence, new items would have to be rewritten and items would need to be subjected to item analysis once again. Alternatively, analyses may suggest the conceptualization of the target construct (Clark and Watson 1995).

In the preceding chapters, the authors have followed the above mentioned item writing process. An exhaustive list of items was generated by considering all domains for each of the construct. The list was then given to five experts who possessed doctoral degree in the subject and were well versed with scale construction. Subject experts independently evaluated the items in the context of their clarity readability level and their relevance for the construct and only the items which were rated relevant (3) or most relevant (4) by all experts were retained.

### *Sample Size for Validation*

Most of the times, the bigger question researchers face is “how to determine the sample size?” for validation. McQuitty (2004) suggested that it is important to determine the minimum sample size required in order to achieve a desired level of statistical power with a given model prior to data collection. Schreiber et al. (2006) mentioned that although sample size needed is affected by the normality of the data and estimation method that researchers use, the generally agreed-on value is 10 participants for every free parameter estimated (cited in Hoe 2008). Minimums of 5 or 10 cases per measure have typically been recommended (Comrey and Lee 1992; Gorsuch 1983; Zhao 2009). Tinsley and Kass (1979) recommended a minimum of five participants per variable whereas as a general rule of thumb for factor analysis is 300 cases (Tabachnick and Fidell 1996). However, on a more lenient note 50 participants per factor is acceptable (Pedhazur and Schmelkin 1991; Osborne and Costello 2004). Comrey and Lee (1992) stated that 50 as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1,000 as excellent sample size for factor analysis (Osborne and Costello 2004). Researchers (Sivo et al. 2006; Garver and Menter 1999; Hoelter 1983) have proposed a “critical sample size” of 200. Any number above 200 is considered to provide sufficient statistical power of data analysis (Hoe 2008). Hence, it can be stated that larger sample size is appreciable.

In our studies, the sample size was above 200 for EFA and 300 for CFA. It was ensured that at least five participants were recruited per item. Both data sets were mutually exclusive. The sample was divided into one-third and two-third and counterbalanced for gender. On one-third data EFA was employed and two-third data CFA was employed. After validating scales in English, these scales were translated in Hindi and established their validation through CFA.

## ***Preliminary Data Analysis***

### **Item Analysis**

It is essential that researchers employ preliminary data analysis. This covers screening the data for item means, standard deviation, skewness, kurtosis and item-total correlation. This would also determine which items are to be eliminated from or retained in the item pool. As a rule of thumb, the criterion suggested by Jang and Roussos (2007) that items with means less than 2 and more than 4 are to be rejected and by Jackson (1970) that item standard deviations less than 1 SD ( $< 1$ ) are to be eliminated. It is also essential to check the data for normality. Finney and DiStefano (2006a, b) stated that skewness values should not exceed an absolute value of 3 and kurtosis values not exceed an absolute value of 8 (Barry and Finney 2008) whereas (Curran et al. 1996) recommended level of skewness  $< 2$  and kurtosis  $< 7$ . Moreover, item-total correlation should not be less than 0.25 Likert (1932) or less than 0.2 or 0.3 (Field 2005a, b).

This stage is most important especially when the test developer is seeking to create a theoretically based measure of a construct. It is essential to be precise which appears to be a difficult task and often poorly understood by test developers. The most obvious problem is the ubiquitous misconception that the attainment of this goal can be established by demonstrating that a scale shows an acceptable level of internal consistency (alpha reliability), as estimated by an index such as coefficient alpha (Cronbach 1951, in Clark and Watson 1995) even though there are contrasting parameters to accept the coefficient alpha (Clark and Watson 1995). For example Nunnally (1978) recommended minimum standards of reliability  $r = 0.80$  and  $r = 0.90$  for basic and applied research, respectively. However, contemporary researchers consider reliabilities in the 0.60 and 0.70 s as good or adequate (Dekovic et al. 1991; Holden et al. 1991, cited in Clark and Watson 1995). According to Kline (1998), Cortina (1993a, b) internal consistency of 0.90 and above is excellent, 0.70–0.90 is good, 0.60–0.70 is acceptable, 0.50–0.60 is poor and below 0.50 is unacceptable. However, when a scale measures several domains, the acceptable value of 0.50 for eight item domains is deemed fit (Costa and McCrae 1992) as reliability is affected by number of items in the domain (Field 2005a, b). Thus, from literature it can be seen that domain Cronbach alpha with lower value is acceptable depending upon the number of items measuring a domain. It is essential to determine the internal consistency since it would determine the degree to which the items that make up a scale are inter-correlated (Clark and Watson 1995). Another view point of acceptability for Cronbach's alpha is that coefficients of 0.70 or higher and mean inter-item correlations in the 0.20–0.40 range indicate good reliability (Clark and Watson 1995; Nunnally 1978). A scale cannot be homogeneous unless all of its items are interrelated or correlated (Clark and Watson 1995). If the items are correlated then it can be concluded that the scale is measuring the same concept.

In the forthcoming chapters, all the above discussed norms have been followed very strictly. Items with mean beyond the range of 2–4, SD lesser than 1, skewness

greater than 2, kurtosis greater than 7, item-total correlation below 0.25, were deleted in all the constructs. After item analyses the retained items were subjected to exploratory factor analysis.

## ***Exploratory Factor Analysis***

Factor analysis is employed to analyze interrelationships among a large number of variables. The technique also explains these variables in terms of their common underlying dimensions. It involves condensing the scale items into a smaller set of dimensions (factors) with a minimum loss of information. In other words, EFA is a data reduction procedure. Researchers often face the issue of “how to extract factors” while employing EFA (Russell 2002). The classic factor analysis equation specifies that a measure being factored can be represented by the following equation:

$$x_1 = w_{11}F_1 + w_{21}F_2 + \dots w_{n1}F_n + w_{1U}U_1 + e_1,$$

where the  $F$ s represent the common factors

$U$ s represent factors

$w$ s represent loadings of each item

$e$ s reflect random measurement error

Each measured variable has its own unique factor, reflecting systematic variance in the item or measure that is not shared with the other measures being analyzed. On the basis of this equation, the variance in the measure being factored (i.e.,  $\sigma^2 x$ ) can be separated into three part

- A part of the variance in the measure that reflects the influence of the common factors, termed the communality of the variable (Russell 2002). However, communalities below 0.40 are not interpreted as evidence of poor fit so much as evidence that variables analyzed have little in common with one another and hence are not ‘reliable’ measures of the proposed factor solution (Russell 2002).
- A part of the variance that reflects the influence of the factor unique to that measure (Russell 2002).
- Random error variance, i.e., there is no measurement error in its ability to distinguish among cases/individuals (Russell 2002).

Two common methods employed for extracting factors in context of EFA are principal component analysis and principal axis factoring. Principal components analysis (PCA) and principal axis factoring (PAF) differ on the estimation of communalities for the measured variables, or in other words; the variance that each measured variable shares with the other measured variables. In PCA, the communalities for the measures are set at 1.0. It assumes that all of the variance in a measure is potentially explainable by factors that are derived. PCA extracts the factors based on correlations among the measures. On the other hand, in PAF the

estimate of the communality reflects the variance in each measure due to the influence of the factors (Russell 2002). PAF extracts factors using a reduced correlation matrix, where the 1.0 values on the diagonal of the correlation matrix are replaced by these initial communalities estimated. In PAF, the analysis of data structure is focused on shared variance and not on sources of error that are unique to individual measurements (Russell 2002).

PCA is applied to a single set of variables to discover which variables in the set form coherent subsets that are independent of one another. It also provides a unique solution, so that the original data, the covariance or correlation matrix can be constructed from the results. Furthermore it looks at the total variance among the variables so the solution generated will include many factors or components as there are variables, although it is unlikely that all of them will meet the criteria for retention (Russell 2002). PCA is characteristically exploratory in nature. Even before one proceeds for PCA, it is essential to know “*if the data is worth reducing.*” This can be determined from the Kaiser–Meyer–Olkin Measure of Sampling Adequacy (KMO-MSA) and Bartlett’s test of sphericity (Floyd and Widaman 1995).

The KMO-MSA is a statistic that indicates the proportion of variance and your variables that might be caused by common underlying factors. It is an index for comparing the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients (Kaiser 1974). High values (close to 1.0) indicate that a factor analysis may be useful with the data. If the value is less than 0.50, the results of the factor analysis may not be useful (Floyd and Widaman 1995; Russell 2002). The Bartlett’s test of sphericity tests the hypothesis that your correlation matrix is an identity matrix that would indicate the variables are unrelated and therefore unsuitable for structure detection (Floyd and Widaman 1995). A significant Bartlett’s test results indicate that EFA can be employed on the items since the null hypothesis would be rejected.

While employing PCA or Factor reduction in SPSS, one would look at extraction method tab and rotation tab.

Extraction Methods (Cudeck and O’Dell 1994; Fabrigar et al. 1999; Finch and West 1997)

- Principal (Axis) Factors method indicates that estimates of communalities are diagonal. It removes the unique and error variance. In this extraction the solution depends on quality of initial communality estimates.
- Maximum Likelihood is an intensive method for estimating loadings that maximize the likelihood of sampling the observed correlation matrix from a population.
- Unweighted least squares extraction method minimizes off diagonal residuals between reproduced and original matrix.
- Generalized (weighted) least squares too minimizes the off diagonal residuals and gives more weight to variables with larger communalities in analysis.
- Alpha factoring maximizes reliability of factors.
- Image factoring minimizes “unique” factors consisting of essentially one measured variable (Russell 2002).



Rotation Methods (Cudeck and O'Dell 1994; Fabrigar et al. 1999; Finch and West 1997).

Rotation is a pattern of loadings where items load most strongly on one factor and weakly on others. It serves to make the output more understandable. Rotations can be orthogonal or oblique allowing factors to correlate.

- Varimax rotation is the most popular orthogonal rotation. It cleans up the factors and it makes large loadings larger and small loadings smaller (Russell 2002).
- Quartimax cleans up the variables; each variable loads mainly on one factor, and works on rows of loading matrix. This is an orthogonal alternative which minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree. It is not used as often since the goal is not to simplify variables. Such a factor structure is usually not helpful to the research purpose (Suhr 2005).
- Equamax is a hybrid of varimax and quartimax criteria and is not popular (Suhr 2005).
- Direct Oblimin is a non-orthogonal (oblique) solution that allows factors to be correlated. This results in higher eigenvalues but diminished interpretability of the factors (Suhr 2005).
- Promax is an alternative non-orthogonal (oblique) rotation method which is computationally faster than the direct oblmin method and therefore sometimes used for very large datasets (Suhr 2005).

Researchers are often confused to know which factor solution is suitable for their dataset. Often researchers are faced with the problem of “*how many items should be there in a factor extracted by EFA*”. This is the issue of identification, or having a sufficient number of measures that load on each factor to be able to adequately operationalize the factor. At least three items per factor are required for a factor model to be identified; more items per factor results in over identification of the model. A number of writers (Comrey and Lee 1992; Fabrigar et al. 1999; Gorsuch 1983) recommend that minimum of three items and optimum of four or more items per factor be included in the factor analysis to ensure an adequate identification of the factors. MacCallum et al. (1999) found that in addition to the communality of the items, the results were more accurate if there were more items per factor. Therefore, it appears wise to test over identified factor models where the researcher includes four or more items per factor in the analysis (Comrey and Lee 1992; Fabrigar et al. 1999; Gorsuch 1983; Russell 2002).

Using one or more methods listed below, the researcher can determine an appropriate range of solutions to investigate. One needs to be cautious while determining the factor solution since methods may not agree. For example Kaiser criterion may suggest 5-factor solution and Scree test may suggest 2-factor, hence the researcher would have to look at 3- and 4-factor solution too (Russell 2002). However, the most important factor is to also select a more theoretically meaningful factor solution (Suhr 2005).

## A Rule of Thumb for EFA

- Comprehensibility is purely subjective criterion to retain factors whose meaning is comprehensible (Suhr 2005).
- Kaiser criterion rule is to drop all components with eigenvalues under 1.0. However, this is default in SPSS. The eigenvalues refer to the amount of variance explained by a factor and are computed by squaring the loadings on a factor and summing them together. Although SPSS uses this criterion no matter what technique is used to extract the factors, in fact this criterion should only be used when principal components analysis (with communalities fixed at 1.0) is used as the extraction procedure (see discussion by Gorsuch 1983, cited in Russell 2002).
- In variance explained criteria, researchers use the thumb rule of keeping enough factors to account for 90 % of variation (Suhr 2005). However, when the goal is to emphasize on parsimony, variance may be as low as 50 %. Fabrigar et al. (1999) stated that the eigenvalue  $< 1.0$ , criterion often leads to extracting too many factors. For example, if one is to factor a set of 5 measures versus a set of 20 measures. If cutoff for eigenvalue is 1.0 then for five measures it would account for 20 % of total variance whereas for 20 measures it would explain just 5 % of the total variance. Hence, when one is factoring a large set of items, it is more likely that using this criterion will lead to extracting factors that account for only a small amount of the total variance (Russell 2002).
- Scree test plots the components as the  $X$  axis and the corresponding eigenvalues as the  $Y$  axis. Toward the right of the plot, the eigenvalues drop. When the drop ceases and the curve makes an elbow toward less steep decline, scree test says to drop all further components after the one starting the elbow. One then looks for a break in the values, where there is the last substantial drop in the eigenvalues. The number of factors prior to this drop represents the number of factors to be extracted (Russell 2002).

In the forthcoming chapters, EFA has been employed twice on all scales with an aim of item reduction. All parameters such as KMO above 0.80, significant Bartlett test at  $p = 0.05$  or  $p = 0.01$ , loading of items with cutoff at 0.40 factor loading, minimum of at least three items in every factor and more than 50 % of total variance was followed. Apart from this varimax rotation with principal component analysis was initially employed. To ensure robustness of factor solution, promax with maximum likelihood was employed. This ensured that the factor solution was near to accuracy and a theoretically sound factor model emerged.

## *Confirmatory Factor Analysis*

Factor analysis is also used to confirm a priori hypotheses. Researchers have often been able to generate hypotheses regarding the factors that should be represented in a given domain of inquiry. The hypotheses may be based on theory or results from

previous empirical studies. As a confirmatory procedure, factor analysis is primarily a method for assessing the construct validity of measures and it is not a means for data reduction unlike EFA. Construct validity is supported if the factor structure of the scale is consistent with the constructs the instrument wants to measure (Floyd and Widaman 1995). This approach allows for testing the relative fit of competing factor models. This approach is primarily useful for confirmation of theories as with other applications of structural equation modeling that suggest alterations in proposed factor structure. Therefore, CFA can be used to revise and refine instruments and their factorial structure. CFA is most effective when it is used to assess whether the existing factors structure adequately fits the data and if the structure fits as well as and as parsimoniously as other models (Suhr 2005; Floyd and Widaman 1995). These days, researchers use computer programs such as LISREL (Joreskog and Sorbom 1986), AMOS and EQS (Bentler 2002) to estimate and evaluate the structural portion of the model. The raw data for the variables are input into the software to generate the iterations, goodness-of-fit indices and standardized paths. The various variables are usually summated scales where the attributes measuring a common underlying construct are summed and divided by the number of items (Hoe 2008).

While performing CFA there are two important terms that a researcher should be aware of (i) Measured variables are those variables that are observed directly by physical measurement or on a scale such as a written test (measuring resilience, spirituality, flow, etc., on a five point scale). The scale items are measured variables. (ii) Latent variables are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured). To repeat, a factor or latent trait is an unseen construct, such as intelligence or anxiety or depression, that is responsible for the correlations among the measured variables that we do see (Norman and Streiner 2003). Hence, resilience is a latent variable. Latent variables reduce dimensionality of data. In other words, it can state that large numbers of observable variables are aggregated in a model that enables to understand the data easily. Latent variables represent ‘shared’ variance, or the degree to which variables “move” together. Variables that have no correlation cannot result in a latent construct based on the common factor model (Kline 2010; Hoe 2008; Norman and Streiner 2003).

The next step is to understand the output of CFA. There are various fit indices that are generated in a CFA output. The next section focuses on these indices and the recommended cutoff’s suggested by various researchers.

## ***Fit Indices***

There are abundant indicators of goodness-of-fit and researchers recommend evaluating the models by observing more than one of these indicators (Bentler and Wu 2002; Hair et al. 1998). Fit indices are categorized as absolute fit indices and relative fit indices (Hoe 2008).

**Absolute fit indices** determine how well, a priori model fits, or reproduces the data (McDonald and Ho 2002; Hooper et al. 2008a, b). Absolute fit indices include, but are not limited to, the Chi-Squared test, RMSEA, GFI, AGFI, RMR, and SRMR (Hooper et al. 2008a, b)

1. Chi-squared ( $\chi^2$ ) test indicates the difference between observed and expected covariance matrices. Values closer to zero indicate a better fit; smaller difference between expected and observed covariance matrices and it can also be used to compare the fit of nested models (Gatignon 2010; Hooper et al. 2008a, b). One obscurity with the chi-squared test of model fit is that researchers may fail to reject an inappropriate model in small sample sizes and reject an appropriate model in large sample sizes (Gatignon 2010; Hooper et al. 2008a, b). Therefore, other measures of fit have been developed. A low  $\chi^2$  value indicates a good fit because chi-square test is used to assess actual and predicted matrices. On the other hand non-significance means that there is no significant difference between the actual and predicted matrices (Hair et al. 1998, cited in Hoe 2008). Therefore, low  $\chi^2$  values, which result in significance levels greater than 0.05 or 0.01, indicate that actual and predicted inputs are not statistically diverse. The significance levels of 0.1 or 0.2 should exceed before non-significance is confirmed (Fornell 1983, cited in Hoe 2008). However,  $\chi^2$  is highly sensitive to sample size especially when the observations are greater than 200. Hence an alternate method is used to evaluate the  $\chi^2$  statistic. The ratio of  $\chi^2$  to the degrees of freedom (df) for the model is calculated (Joreskog and Sorbom 1993; Hoe 2008). A small  $\chi^2$  value relative to its degree of freedom is indicative of good fit. Kline (1998) suggested that a  $\chi^2/\text{df}$  ratio of 3 or less is a reasonably good indicator of model fit (Hoe 2008).
2. **Root mean square error of approximation (RMSEA)** an extremely informative criterion in evaluating model fit. The RMSEA index measures the discrepancy between the observed and estimated covariance matrices per degree of freedom (Steiger 1990; Hoe 2008). It measures the discrepancy in terms of the population and not the sample. Furthermore, it avoids issues of sample size by analyzing the discrepancy between the hypothesized model, with optimally chosen parameter estimates, and the population covariance matrix (Hooper et al. 2008a, b). Thus, the value of this fit index is expected to better approximate or estimate the population and not be affected by sample size. The RMSEA ranges from 0 to 1, with smaller values indicating better model fit. A value of 0.06 or less is indicative of acceptable model fit. Values less than 0.05 indicate good fit, values up to 0.08 reasonable fit and ones between 0.08 and 0.10 indicate mediocre fit (All et al. 2013; Hu and Bentler 1999).
3. **Root mean square residual (RMR) and standardized root mean square residual (SRMR)** are the square root of the discrepancy between the sample covariance matrix and the model covariance matrix (Hooper et al. 2008a, b). The RMR may be somewhat difficult to interpret, however, as its range is based on the scales of the indicators in the model (this becomes tricky when you have multiple

indicators with varying scales; e.g., two questionnaires, one on a 0–10 scale, the other on a 1–3 scale) (Kline 2010). The standardized root mean square residual removes this difficulty in interpretation, and ranges from 0 to 1, with a value of 0.08 or less being indicative of an acceptable model (Hu and Bentler 1999).

4. **Goodness-of-fit index and adjusted goodness-of-fit index** GFI is a measure of fit between the hypothesized model and the observed covariance matrix. The adjusted goodness-of-fit index (AGFI) corrects the GFI, which is affected by the number of indicators of each latent variable. The GFI and AGFI range between 0 and 1, with a cutoff value of 0.9 generally indicating acceptable model fit (All et al. 2013; Baumgartner and Hombur 1996).

**Relative fit indices** are also called “incremental fit indices” (Tanaka 1993) and “comparative fit indices” (Bentler 1990) that compare the chi-square for the hypothesized model to one from a “null,” or “baseline” model (McDonald and Ho 2002). Relative fit indices include the normed fit index and comparative fit index (Hooper et al. 2008a, b).

1. **Normed fit index (NFI)** analyzes the discrepancy between the chi-squared value of the hypothesized model and the chi-squared value of the null model (All et al. 2013) and tends to be negatively biased (Bentler and Bonett 1980). **Non-normed fit index (NNFI)** (also known as the Tucker-Lewis index, as it was built on an index formed by Tucker and Lewis, in 1973) resolves some of the issues of negative bias, though NNFI values may sometimes fall beyond the 0–1 range (Bentler 1990). Values for both the NFI and NNFI should range between 0 and 1, with a cutoff of 0.95 or greater indicating a good model fit (All et al. 2013; Hooper et al. 2008a, b; Hu Bentler and Hoyle 1995).
2. **Comparative fit index (CFI)** was developed by Bentler (1990) as a non-centrality parameter-based index to overcome the limitation of sample size effects. It analyzes the model fit by examining the discrepancy between the data and the hypothesized model, while adjusting for the issues of sample size inherent in the chi-squared test of model fit (Gatigon 2010), and the normed fit index (Bentler 1990). CFI values range from 0 to 1, with larger values indicating better fit; a CFI value of 0.90 or larger is generally considered indicating acceptable model fit (Hu and Bentler 1999).

### *Parameters for Accepting a Model*

Geuens and Pelsmacker (2002, cited in Pandey and Saxena 2012) used the following six criteria for examining the model fit

- a. The goodness-of-fit index (GFI) greater than 0.80,
- b. The adjusted goodness-of-fit index (AGFI) greater than 0.90,

- c. The root-mean-square error of approximation (RMSEA) less than 0.08 (Cole 1987),
- d. The ratio of maximum likelihood chi-square to the degrees of freedom ( $X^2/df$ , Bartone et al. 1989) less than five,
- e. Tucker and Lewis non-normed fit index (TLI) greater than 0.90, and
- f. Bentler's comparative fit index (CFI) greater than 0.90 (see Pandey and Saxena 2012).

Here, it is worth mentioning that although we are using the criteria of testing the model fit as used by Geuens and Pelsmacker (2002) to have parity, the recent recommendations for some of the said fit indices are slightly different. For example, now a days a good fit is inferred if the GFI and AGFI are greater than 0.95,  $\chi^2/df < 2$  and RMSEA is less than 0.05 (Hooper et al. 2008a, b, cited in Pandey and Saxena 2012). In all the forthcoming chapters, the models have been accepted or rejected based on the parameters mentioned above.

The new scales were constructed in English. However, it is a known fact that Hindi is spoken by approximately 41 % of Indian population (Census 2011). Therefore, a need was felt to translate the scales into Hindi. All the scales were translated into Hindi and data were collected from approximately above 500 participants. The translation was done by bilingual experts who translated the original English version into Hindi language. The test was later back-translated into English to verify the content similarity to the original scale and to ensure that translated tests were true copy of the original tests. The discrepancies were resolved and the test was once again verified by the authors and bilingual experts. On the Hindi scales, only CFA was employed. It was observed that CFA models for the test confirmed for all the scales.

Furthermore, in the forthcoming chapters concurrent validity too was evaluated by correlating the scales with other positive psychology constructs/scales. For establishing concurrent validity, well-established scales were used. For example, the newly constructed resilience scale's concurrent validity was evaluated by correlating the new scale with other well-established scales such as Mental Health Continuum Scale-Short Form (MHC-SF, Keyes 2009), Scale of positive and negative experiences and Flourishing Scale (SPANE and FS, Diener et al. 2010). Similarly, concurrent validity of the new spirituality scale was established by correlating it with Vedic Personality Inventory (VPI, Wolf 1998) and Flourishing Scale (FS, Diener et al. 2010)

In the future chapters, the cutoffs and statistical recommendations mentioned in this chapter have been used. If otherwise, the statistical recommendations have been mentioned in appropriate place. We hope that this chapter would enable the readers to understand the scale construction procedure followed in the forthcoming chapters.

## References

- All, C., Mahdi, C. G., & Isaksson, V. (2013). The effects of charismatic clients on auditors objectivity.
- Angleitner, A., & Wiggins, J. S. (1985). *Personality assessment via questionnaires*. New York: Springer-Verlag.
- Barry, C. L., & Finney, S. J. (2008). A psychometric investigation of the college self-efficacy inventory. Center for Assessment and Research Studies, James Madison University. Retrieved May 10th 2013, from [http://www.psyc.jmu.edu/assessment/research/pdfs/BarryFinneyCSEI\\_NERA.pdf](http://www.psyc.jmu.edu/assessment/research/pdfs/BarryFinneyCSEI_NERA.pdf)
- Bartone, P. T., Ursano, R. J., Wright, K. W., & Ingraham, L. H. (1989). The impact of a military air disaster on the health of assistance workers: A prospective study. *Journal of Nervous and Mental Disease*, 177, 317–328.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107 (2), 238–246.
- Bentler, P. M., & Wu, E. J. C. (2002). *EQS 6 for windows user's guide, encino*. CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309.
- Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, 56, 754–761.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Connor, K. M., & Davidson, J. R. (2003). Development of a new resilience scale: The Connor-Davidson resilience scale (CD-RISC). *Depression and Anxiety*, 18(2), 76–82.
- Cortina, J. M. (1993a). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104.
- Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, 13(6), 653–665.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Cudeck, R., & O'Dell, L. L. (1994). Applications of standard error estimates in unrestricted factor analysis: Significance tests for factor loadings and correlations. *Psychological Bulletin*, 115, 475–487.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16.
- de Barros, M. T. (2014). Brand relationships and corporate brand identity: A structural modelling approach. Unpublished Thesis.
- Dekovic, M., Janssens, J. M. A. M., & Gerris, J. R. M. (1991). Factor structure and construct validity of the Block Child Rearing Practices Report (CRPR). *Psychological Assessment*, 3, 182–187.
- Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., et al. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97, 143–156.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272–299.
- Field, A. (2005a). *Discovering statistics using SPSS* (2nd ed.). London: Sage.



- Field, A. (2005b). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, 31(4), 439–485.
- Finney, S. J., & DiStefano, C. (2006a). Nonnormal and categorical data in structural equation models. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269–314). Greenwich, CT: Information Age.
- Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation models. In G.R. Hancock & R.O. Mueller (Eds.), *A second course in structural equation modeling* (pp. 269–314). Greenwich, CT: Information Age.
- Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, 7(3), 286.
- Fornell, C. (1983). Issues in the application of covariance structure analysis: A comment. *Journal of Consumer Research*, 9(4), 443–448.
- Friborg, O., Hjemdal, O., Rosenvinge, J. H., & Martinussen, M. (2003). A new rating scale for adult resilience: What are the central protective resources behind healthy adjustment? *International Journal of Methods in Psychiatric Research*, 12(2), 65–76.
- Garver, M. S., & Mentzer, J. T. (1999). Logistics research methods: Employing structural equation modeling to test for construct validity. *Journal of Business Logistics*, 20(1), 33–57.
- Geuens, M., & Pelsmacker, P. D. (2002). Validity and reliability of scores on the reduced emotional intensity scale. *Educational and Psychological Measurement*, 62(2), 299. 315.
- Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, 103(2), 265.
- Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). New Jersey: Prentice-Hall.
- Hoe, L. S. (2008). Quantitative research methods. *Journal of Applied Quantitative Methods*, 3(1), 76–83.
- Hoelter, D. R. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research*, 11, 325–344.
- Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment*, 3, 111–118.
- Hooper, D., Coughlan, J., & Mullen, M. (2008a). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1), 53–60.
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008b). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods*, 6, 53–60.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, 6(1), 1–55.
- Hu, L. T., Bentler, P. M., & Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Evaluating model fit, 76–99.
- Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 61–96). New York: Academic Press.
- Jang, E. E., & Roussos, L. A. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, 44(1), 1–21. doi:10.2307/20461840.
- Joreskog, K. G., & Sorbom, D. (1986). *LISREL6 - computer program*. IN, Scientific Software: Mooreville.
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36.
- Keyes, C. L. M. (2009). Brief description of the mental health continuum short form (MHC-SF). Retrieved from <http://www.sociology.emory.edu/ckeyes/>. [Online, retrieved January 15, 2010]
- Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methuen.



- Kline, R. B. (1998). *Principles And Practice Of Structural Equation Modeling*, New York, Guilford Press.
- Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.
- Loevinger, J. (1957). Objective test as instruments of psychological theory. *Psychological Reports*, 3, 635–694.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130.
- Matlock-Hetzel, S. (1997). Basic concepts in item and test analysis. Texas A & M University. Retrieved May 2nd, 2013 from <http://ericae.net/ft/tamu/Espy.htm>
- McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting statistical equation analyses. *Psychological Methods*, 7(1), 64–82.
- McQuitty, S. (2004). Statistical power and structural equation models in business research. *Journal of Business Research*, 57(2), 175–183.
- Norman, G. R., & Streiner, D. L. (2003). Chapter 17: Path analysis and structural equation modeling. *PDQ Statistics* (3rd edn., Vol. 156, p. 177). London: BC Decker Inc.
- 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 & Evaluation*, 9(11), 8.
- Oshio, A., Nakaya, M., Kaneko, H., & Nagamine, S. (2003). Development and validation of an adolescent resilience scale. *Japanese Journal of Counseling Science*, 35, 57–65.
- Pandey, R., & Saxena, P. (2012). Validation of dimensionality of affect intensity using the hindi version of the emotional intensity scale. *Europe's Journal of Psychology*, 8(1), 139–158.
- Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum.
- Russell, D. W. (2002). In search of underlying dimensions: the use (and abuse) of factor analysis. *Personality and Social Psychology Bulletin*, 28(12), 1629–1646.
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6), 323–337.
- Singh, K., Ruch, W., & Junnarkar, M. (2014). Effect of the demographic variables and psychometric properties of the personal well-being index for school children in India. *Child Indicators Research*, 7(3), 1–15. doi:10.1007/s12187-014-9264-4.
- Sivo, S. A., Fan, X. T., Witta, E. L., & Willse, J. T. (2006). The search for 'optimal' cutoff properties: Fit index criteria in structural equation modeling. *The Journal of Experimental Education*, 74(3), 267–289.
- Smith, B. W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: Assessing the ability to bounce back. *International Journal of Behavioral Medicine*, 15(3), 194–200.
- Steger, M., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, 53(1), 80–93.
- Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180.
- Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis. *SUGI 30 Proceedings*, 203, 230.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd edn.). New York: HarperCollins.
- Wilkinson, L. (1999) Task force on statistical inference, APA board of scientific affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Tanaka, J. S. (1993). Multifaceted conceptions of fit in structure equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

- Tinsley, H. E., & Kass, R. A. (1979). The latent structure of the need satisfying properties of leisure activities. *Journal of Leisure Research*, 11(4), 278.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.
- Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 223–252). New York: Cambridge University Press.
- Wagnild, G. M., & Young, H. M. (1993). Development and psychometric evaluation of the resilience scale. *Journal of Nursing Measurement*, 1, 165–178.
- Watson, D. (2012). Objective tests as instruments of psychological theory and research. In H. Cooper, P. M. Camic, D. L. Long, A.T. Panter, D. Rindskopf, K.J. Sher (Eds.), *APA handbook of research methods in psychology* (Vol 1). Foundations, planning, measures, and psychometrics (pp. 349–369). Washington, DC, US: American Psychological Association, xlvii, 744 pp. <http://dx.doi.org/10.1037/13619-019>
- Wolf, D. B. (1998). The Vedic personality inventory: A study of the Gunas. *Journal of Indian Psychology*, 16, 26–43.
- Zhao, N. (2009). The minimum sample size in factor analysis. Retrieved May 5th 2013, from, <https://www.encorewiki.org/display/~nzhao/The+Minimum+Sample+Size+in+Factor+Analysis>
- Zumbo, B. D., Gelin, M. N., & Hubley, A. M. (2002). The construction and use of psychological tests and measures. In *Encyclopedia of life support systems*. France: United Nations Educational, Scientific, and Cultural Organization Publishing (UNESCO-EOLSS Publishing).

Measures of Positive Psychology

Development and Validation

Singh, K.; Junnarkar, M.; Kaur, J.

2016, X, 215 p. 19 illus., Hardcover

ISBN: 978-81-322-3629-0